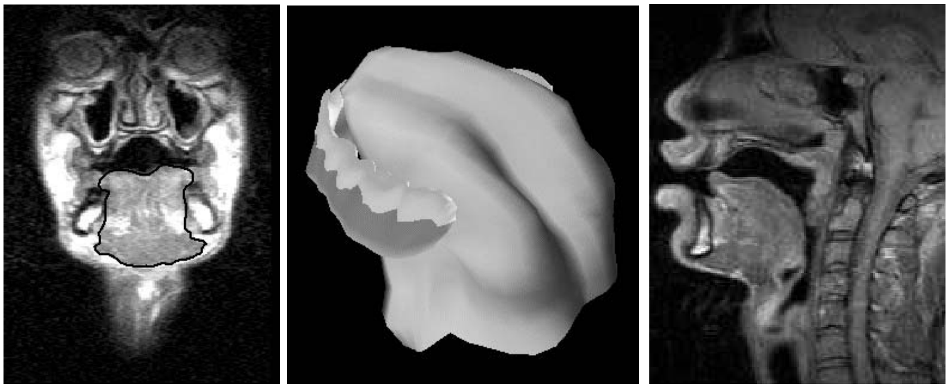
KUNGL
TEKNISKA
HÖGSKOLAN

Royal Institute of Technology
Speech, Music and Hearing

Tongue Talking

Studies in Intraoral Speech Synthesis



Olov Engwall

Doctoral Dissertation
Stockholm 2002

KUNGL
TEKNISKA
HÖGSKOLAN

Tongue Talking

Studies in Intraoral Speech Synthesis

Olov Engwall

Stockholm 2002

Doctoral Dissertation
Royal Institute of Technology
Department of Speech, Music and Hearing

Akademisk avhandling som med tillstånd av Kungl Tekniska Högskolan framlägges till offentlig granskning för avläggande av teknisk doktorsexamen fredagen den 14 juni 2002 kl 10.00 i Kollegiesalen, Administrationsbyggnaden, Kungl Tekniska Högskolan, Valhallavägen 79, Stockholm.

ISBN 91-7283-258-9
TRITA-TMH 2002:04
ISSN 1104-5787

Universitetsservice US AB, Stockholm 2002

Abstract

This thesis presents work in three-dimensional multimodal intraoral articulatory speech synthesis. A 3D model of the vocal tract has been developed with the aim of synthesizing speech in two modalities: both visual and acoustic. The model is controlled by parameters to make articulatory changes such as lowering the jaw or raising the tip of the tongue. The model and its parameters have been defined on articulatory measurements of one reference subject. No single existing technique is able to measure real-time intra-oral articulations in three dimensions and the data for the model has hence been collected using different methods that combined provide the needed information.

Magnetic Resonance Imaging (MRI) was used to measure the vocal tract three-dimensionally for a large number of Swedish articulations, allowing to reconstruct the shape of the articulators. Based on these reconstructions a set of parameters was defined to replicate the observed articulations as closely as possible. The acquisition time of MRI required the subject to sustain the articulations for 43 seconds, and the MRI data must hence be complemented with other measurements in order to model running speech. Firstly, such measurements are needed to get information on the timing and the transitions between the static articulations. Secondly, as the articulations measured with MRI were artificially sustained, real-time measurements are needed to assess the static data to ensure that it is representative of running speech and to adjust the model to account for the differences between static and dynamic articulations. Electromagnetic articulography (EMA) and electropalatography (EPG) were used in this project for assessment, tuning and to provide information on the kinematics.

The possible domains of applications for the vocal tract model are visual feedback in speech training and acoustic articulatory speech synthesis. Visual feedback is an important tool in pronunciation training for hearing-impaired children and second language learners, using the visual channel to illustrate differences that the user is unable to discern aurally. Articulatory acoustic speech synthesis is the most flexible of the currently existing synthesis methods, as the physiology of the model can be changed to adjust the synthesis to mimic different speakers.

Keywords: Articulatory modeling, speech synthesis, three-dimensional, vocal tract, intra-oral, multimodal, visual feedback, articulatory synthesis, Magnetic Resonance Imaging (MRI), electromagnetic articulography (EMA), electropalatography (EPG).

Contents

Included papers

This dissertation consists of a summarising overview and the following papers:

Paper I Engwall, O. (1999)
Vocal tract modeling in 3D.
TMH-QPSR1-2/1999, 31-38.

Paper II Engwall, O. & Badin, P. (1999)
Collecting and analysing two- and three-dimensional
MRI data for Swedish.
TMH-QPSR3-4/1999, 11-38.

Paper III Engwall, O. & Badin, P. (2000)
An MRI study of Swedish fricatives: coarticulatory effects.
Proceedings of the 5th Speech Production Seminar, 297-300.

Paper IV Engwall, O. (2000a)
Dynamical aspects of coarticulation in Swedish fricatives
– a combined EMA & EPG study.
TMH-QPSR 4/2000, 49-73.

Paper V Engwall, O. (2000b)
Are static MRI data representative of dynamic speech?
Results from a comparative study using MRI, EMA and EPG.
Proceedings of the 6th ICSLP, I:17-20.

Paper VI Engwall, O. (submitted1)
Combining MRI, EMA & EPG measurements
in a three-dimensional tongue model.
Submitted to *Speech Communication.*

Paper VII Engwall, O. (submitted2)
Concatenative Articulatory Synthesis
Submitted to *Journal of Phonetics.*

Paper VII Engwall, O. (submitted3)
Evaluation of a System for Concatenative Articulatory Visual Synthesis
Submitted to *the 7th ICSLP, Denver, Colorado, September 16-20, 2002.*

Paper IX Engwall, O. (2001d)
Synthesizing static vowels and dynamic sounds using
a 3D vocal tract model.
Proceedings of the 4th ISCA workshop on Speech synthesis, 81-86.

The papers will be referred to by the Roman numerals given above.

List of Figures and Tables

Figures

Tables

Glossary of abbreviations

2D	Two-dimensional
3D	Three-dimensional
CI	Coarticulation index of the EPG contact pattern
CP	Constriction place, measured with EPG
COG	Centre of gravity of the EPG contact pattern
CW	Constriction width, measured with EPG
CTT	Centrum för talteknologi (Centre for Speech Technology)
EMA	Electromagnetic articulography
EMG	Electromyography
EPG	Electropalatography
EUROSPEECH	European Conference on Speech Communication and Technology
fps	frames per second
ICP	Institut de la Communication Parlée, Grenoble, France
ICPhS	International Congress of Phonetic Sciences
ICSLP	International Conference on Spoken Language Processing
JawAdv	Jaw advance measurements
JawHei	Jaw height measurements
JH	Jaw height parameter
KTH	Kungliga Tekniska Högskolan (Royal Institute of Technology)
LarHei	Larynx height measurements
LAT	Lateral asymmetry index in EPG contact patterns
LCA	Linear Componenet Analysis
LipHei	Measurements of the distance between the upper and lower lip
MRI	Magnetic Resonance Imaging
NUTEK	Närings- och teknikutvecklingsverket (Swedish National Board for Industrial and Technical Development)
PCA	Principal Component Analysis
ProBot	Measurements of lower lip protrusion
ProTop	Measurements of upper lip protrusion
QPSR	Quarterly Progress and Status Report, TMH, KTH
RMS	Root Mean Squared
VCV	Vowel-Consonant-Vowel triplett
VT	Vocal tract
T1-T3	First (tip/blade), second (body) and third (dorsum) EMA tongue coils
TA	Tongue advance parameter
TB	Tongue body parameter
TD	Tongue dorsum parameter
TMH	Institutionen för tal, musik och hörsel (Department of Speech, Music and Hearing)
TngAdv	Tongue advance measurements
TngTip	Tongue tip measurements
TT	Tongue tip parameter
TW	Tongue width parameter
VCV	A Vowel-Consonant-Vowel word
VelHei	Velum height measurements
VINNOVA	Verket för innovationssystem (The Swedish Agency for Innovation Systems)

Acknowledgements

I would like to thank my supervisor Björn Granström for giving me the opportunity to work in this fascinating field and for his support during my doctoral studies.

My deep gratitude goes to my co-author of Papers II and III, Pierre Badin, for his guidance and friendship during my stay at ICP in Grenoble 1999. His prior knowledge in the field of MRI acquisition and analysis was invaluable and the MRI data set collected with his and Christian Segerbarth’s assistance is the basis on which the vocal tract model rests.

I am also grateful to Jonas Beskow for paving the grounds with his work on synthetic faces and the framework for the visual speech synthesis. His programming skills solved tricky situations on innumerable occasions and his enthusiasm for the field has been very contagious for me as a room mate at TMH.

I have received many encouraging words from Professors Gunnar Fant and Björn Lindblom during the work, for which I am grateful, as this support has been a great source of inspiration.

The floorball players at the Department of Speech, Music and Hearing (TMH) deserve a special thanks for contributing to one of the highlights of the working week. All the friends of Fysikalen (Anna, Arun, BOS, Dennis, Fredrik, Gitte, Göran, Hedvig, Johan, Maria, Magnus, Mats, Mats, Mats, Mikko, Måns, Olle, Olof, Wille, Ylva, ...) deserve another special thanks for all the highlights in the *non-working* time during the past decade.

My collegues at TMH have made it a very pleasant working place and I would in particular like to mention the administrative staff, Caroline Bergling, Cathrin Dunger, Markku Haapakorpi, Niclas Horney and Ebi Rohani-Makvandi, for making the department run smoothly and always providing cheerful assistance. My room mate Magnus Nordstrand has also contributed to making TMH merrier by sharing my love for spex (students’ farce) and Hammarby IF.

David House deserves to be mentioned as well for proof-reading both Paper I and the summary part with the eyes of a native speaker. All remaining errors are entirely my responsability, but thanks to David, a number of ’that’s could be corrected into ’which’s, some hair-raising typos could be removed and a few neglected grammatical rules could be applied.

Most of all, my thoughts go to my parents who have been my main educational support from kindergarten and onwards, always encouraging and inspired at sharing the marvels of science and research. I would also like to thank them for providing me with the reference subject for the measurements in Papers II-VII and giving him such good teeth, thus avoiding interference of dental fillings in the MRI acquisition.

The picture in Fig. 1.1 is used by kind permission of Lucsfilm Ltd and it may not be reproduced without the prior written consent from the copyright holder.

This work was carried out within the Centre for Speech Technology supported by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organizations.

The tongue is a little member, and boasteth great things.
(Ja 3:5,6)

1. Introduction

Talking machines have a long tradition in science fiction literature and films, both in the shape of humanoids (classical examples are the female robot in Fritz Lang's *Metropolis*, 1926, and C-3PO in George Lucas' *Star Wars*, 1977) and as talking computers (almost synonymous to HAL in Stanley Kubrick's *2001: A Space Odyssey*, 1966).

With the focus on speech synthesis there is one major difference between the humanoid robot C-3PO and the talking computer HAL: the speech production and presentation method, as summarized in Fig. 1.1.

HAL is a black box, outputting his speech sounds through loudspeakers, without giving any information on how the speech was produced. HAL's synthetic speech is of very natural sounding quality, but in other respects the human-computer dialogue is quite unnatural, with HAL being ubiquous in the spaceship, rather than located to one physical body. Talking to HAL is hence somewhat like talking right out into thin air. From the user's point of view the interaction is unimodal, with speech only, whereas the computer has multimodal information from both microphones and cameras.

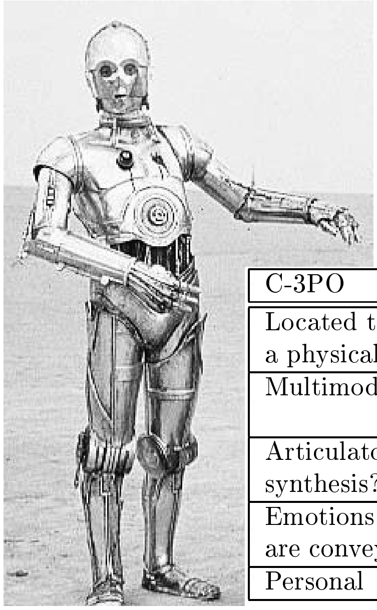
The dialogue is also quite unequal in HAL's favour, as the computer sees all the facial expressions and hears the emotional variations in the human users' utterances, whereas the computer's feelings are not reflected in his speech, which is always very calm and neutral (this was a very conscious choice from the makers of the film, and the actor that was originally choosen to read HAL's lines, Martin Balsam, was in fact replaced because his voice was considered too emotional). This makes HAL frightful and unhuman; not only is he an omnipotent computer that sees and hears everything and has acquired the human property of having feelings, but he also has the power to suppress them from his speech.

C-3PO, on the other hand, has a humanlike body (cf. Fig. 1.1), and either produces his speech mechanically or creates the impression that the speech sounds were created this way, by outputting the speech through the mouth. C-3PO's speech has a distinct synthetic quality, but in other respects his dialogues mimic human-to-human multimodal interactions rather closely, with his face and body being used to complement or underline the information conveyed by the acoustic speech.

Unlike HAL, C-3PO's feelings are very clearly reflected in his speech and gestures. In fact, he is quite an hysterical robot, much more emotional than his human friends in the films, and this makes him more human-like and less frightening than HAL.

As speech synthesis moves from science fiction over science to everyday reality some of these distinctions remain, concerning naturalness, multimodality and emotions.

At one edge of the spectrum there is concatenative synthesis, more and more common in commercial systems, such as e.g. automated timetable request services over the telephone. As concatenative synthesis uses prerecorded natural speech units, good concatenative synthesizers can sound quite natural, but at the cost of being both much less flexible and less penetrable for the user. The synthesis produced is restricted in variation, regarding voice quality, speaking style and emotions, to the database that was recorded. As the recorded databases are almost exclusively of neutral speech (the few existing emotional databases are summarized and discussed in Campbell, 2000) and changing the emotion of concatenated speech is difficult (cf. Schröder, 2001 for a review), concatenative synthesis runs the risk of conveying the same emotionally cool and neutral feeling as HAL, regardless of the content



C-3PO	HAL
Located to a physical body	Ubiquous
Multimodal	Unimodal for the user
Articulatory synthesis?	?
Emotions are conveyed	Always neutral
Personal	Unpersonal

Image of C-3PO from *Star Wars: A New Hope* courtesy of Lucasfilm Ltd.
All rights reserved. Used under authorization.

Figure 1.1. The humaniod C-3PO from George Lucas' *Star Wars* (1977) and his features regarding speech production compared to the talking compouter HAL in Stanley Kubrick's *2001: A Space Odyssey* (1966).

of the utterance. Moreover, like HAL, the synthesis is a black box, from which speech comes out, but where most of the parameters controlling the speech are hidden.

Formant synthesis allows its users to "look into the box" and gain access to the parameters. The speech sounds less natural than with concatenative synthesis, but is more flexible, as the user can have direct control of the sound, with the possibility to change a large number of acoustic parameters to adjust voice quality as well as emotional states (e.g. Carlson *et al.* , 1992; Murray & Arnott, 1996). Formant synthesis can thus be preferable to concatenative synthesis in some aspects, as long as no reliable conversion of emotional states exists for the concatenated units. Formant synthesis has however only indirect coupling to natural speech production, as the parameters control the sound and not how it is produced.

At the other end of the spectrum there is articulatory synthesis, with a large flexibility and intuitive control of the model for sound changes, just as for C-3PO. Articulatory synthesizers can, in principle, adapt all aspects of the speech, for instance both voice quality and emotional state. Moreover, as articulatory synthesis is often based on a representation of the vocal tract (cf. section 2.2 for an overview of such models), the synthesizer has multimodal potentials of producing both acoustic and visual speech with the same set of parameters. The quality of the state-of-the-art acoustic articulatory synthesis is however much lower than for concatenative or formant synthesis, due to the complexity of the modeling task and the fact that comparatively little effort has been put into articulatory synthesis research.

The situation is similar concerning visual speech synthesis. The frame-by-frame manually controlled animations used in film industry provide very lifelike sequences that can actually fool the viewer that the sequence really has taken place in the physical world, but at the cost of flexibility: the animated sequence does not generalize to any other scenes and every frame has to be generated with large efforts.

The avatars and animated characters that have started to appear in products reading news (e.g. http://www.ananova.com) or email (e.g. http://www.lifefx.com) more or less automatically from text have much more flexibility, while trying to maintain beliveability rather than naturalness. The aim is then not to give the impression that the agent is a real person, but rather that the user accepts the agent as being the speaker.

These commercial faces are to a large extent based on superficial observations of facial movements rather than detailed measurements of speech articulation. Furthermore, only the outer, facial parts can be observed directly, while the interior articulators must be analysed using different kinds of articulatory measurements.

It is hence here that basic research can contribute, creating physiologically correct models, based on human speech production, aiming at a representation where both visual and acoustic speech is produced in a more humanlike way. This thesis is intended as a contribution to that research: articulatory data has been collected, analysed and combined in a three-dimensional model of the vocal tract and methods for intra-oral visual speech synthesis have been developed and evaluated. The synthetic faces have hence been extended to the intra-oral parts and a basis for further work with three-dimensional acoustic synthesis has been generated.

The history of articulatory synthesis goes back to the late 18^{th} century, when Wolfgang von Kempelen constructed his speaking machine, in which the sound output could be controlled by changing the resonance properties of the machine's mouth, using the left hand as the moving articulator. The machine, described in *Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine* from 1791, had many of the properties of modern articulatory synthesizers and some that the current ones do not have, such as using air flow to generate the acoustic output. Some of the features that it shared with today's most advanced articulatory synthesizers were that it was controlled by a few articulatory parameters (the movement of the hand), that it was three-dimensional and that the synthesis was made in real-time. It was however neither automatic, allowing the synthesizer to be controlled by a text string or a predefined set of parameter values, nor corresponding visually to human speech production, i.e. its parts had no direct correspondence to the articulators, and the user's control of the machine did not replicate speech movements.

The last point is addressed by Glovetalk (Fels, 1994), a modern articulatory synthesizer controlled by hand movements, that does, to some extent, use articulatory defined gestures to produce speech. The Glovetalk user is fitted with a cyberglove, a tracker, a keyboard and a foot-pedal, that are used to control the 10 parameters of a parallel formant speech synthesizer. Hand movements in the vowel space result in different vowels, and specific gestures, partially corresponding to the place and manner of articulation, define the consonants. Glovetalk is hence closer to what is commonly defined to be an articulatory synthesizer, but it is not automatic and does only mimic the articulators partially.

Both von Kempelen's machine and Glovetalk also differ in that they are models in the real world, whereas most modern articulatory synthesizers are computerized virtual models.

An exception is the talking robot, Waseda Talker-1, WT-1 (Nishikawa *et al.* , 2000), that is closer to modern articulatory modeling than von Kempelen's machine in that it includes models of the articulators (lips, teeth, tongue, nasal cavity and velum) and that its speech synthesis is based on articulatory gestures corresponding to human speech production, just as the virtual models. In the remainder of this thesis, the terms 'articulatory synthesizer' and 'articulatory model' will nevertheless refer to computer models. It is worth noting this distinction as three-dimensional articulatory modeling is a prime notion throughout the thesis alluding to a three-dimensional representation in the computer rather than to a 3D model in the real world.

1.1 Overview of the thesis

Regardless of whether the modeling takes place in the real world or in the computer, articulatory synthesis needs measurements of human speech production and methods to transfer the data to the model. As human speech production is a very complex process, all models use simplified representations of the articulators and the choice of the representation will affect the model's performance when used for visual or acoustic speech synthesis. The task can hence be divided into two, firstly measuring articulations and secondly choosing the parametrisation of the data to create a model. The following background chapter surveys measurement methods that can be used for articulatory measurements and representations that have been chosen in previous work on articulatory modeling. This will give an overview of previous work in the field, but also introduce the considerations behind the modeling in the current project.

The project itself is covered in chapter 3 which is divided into four parts dealing with the main aspects of the articulatory modeling: model construction, measurements, data parametrisation and applications. Each part is described in one or several of the included papers and these are therefore summarized in the corresponding part.

The first part of chapter 3 is an introduction to the KTH 3D Vocal Tract model. It deals with the basic characteristics of the model, regarding surface generation and parameter definition and further discusses the benefits and potential applications of the model. In doing so it is also a summary of Paper I.

The second part, comprising Papers II-V, describes the articulatory measurements that were made, both in terms of acquisition and analysis of the data. A combination of MRI and EMA-EPG was employed to collect three-dimensional and kinematic data, respectively. The data from the different sources were used for two coarticulation studies of Swedish fricatives (Papers III-IV). The objective of the coarticulation studies was two-fold. Firstly they provided data on articulatory variation that was then implemented in the model. Secondly, doing the coarticulation study with both sustained (MRI) and dynamic articulations (EMA/EPG), permits the assessment of the MRI data regarding if the artificially sustained articulations can be considered as good approximations of articulations in running speech.

The third part focuses on the parametrisation of the data in the model that was dealt with in Paper VI. It describes the creation of articulator surfaces from geometrical data, the definition of parameters from a statistical analysis of variations in the corpus, the kinematic control of the parameters and the physical constraints that are needed.

In the fourth part, three applications of the model, concatenative articulatory visual synthesis, (Papers VII-VIII), three-dimensional articulatory acoustic synthesis (Paper IX), and speech production training are discussed and evaluated.

The chapter ends with some conclusions and perspectives of future development of multimodal articulatory synthesis.

Finally, the included papers are appended.

2. Background: Articulatory measurements and modeling

2.1 Measurement methods

Articulatory modeling necessitates measurements on which the model can be based, both of the three-dimensional shape and the kinematics in the articulatory gestures. Ideally one would wish that both three-dimensional and real-time data could be measured simultaneously, using a method that does not interfere with the subject's normal articulation. Currently, no such method exists, however; all methods require the subject to produce the articulations under more or less unnatural circumstances and no method measures the full 3D geometry and kinematics at the same time, even if promising recent advances in the MRI technique make real-time 3D measurements conceivable (Demolin *et al.* , 2000). The 3D and the kinematic measurements are hence made separately, using different methods, that are summarized in the following sections, with their advantages and disadvantages (Webb, 1996). The aim of this survey is to illustrate why the measurement methods opted for in the KTH 3D Vocal Tract modeling were selected.

2.1.1 Three-dimensional data acquisition

Mechanical measurement methods

Until about two decades ago, when 3D imaging methods, such as CT, ultrasound and MRI (see below), became current, three-dimensional measurement of the vocal tract was a difficult task indeed. The main problem, which remains the most severe disadvantage of the methods, was how to make measurements on a talking, or even living, subject.

Ladefoged *et al.* (1971) created a model based on dissections of the vocal tract of newly deceased and casts made on living subjects. The casts were made by pouring paste for making dental casts into the mouth of a subject who was producing the neutral vowel hanging upside down. The cast could then be extracted and sliced and the cross-sectional area measured.

Perkell (1974) used a cadaver tongue as an anatomical reference for his tongue model. Two tongue halves were frozen and cut into radially oriented wedge shaped slices, permitting the study of the geometry of the tongue and the orientation of its muscle fibres.

The above two studies provided important information on the anatomy and general geometry of the vocal tract and the tongue, respectively, but they could not give any substantial information on varied articulations. The study of Dart (1964) was somewhat more flexible in that respect, using live subjects and a custom made instrument, consisting of a scaled cantilever that was lowered into the pharynx, carrying a light source, a mirror to reflect the light down the pharynx and calipers with which the horizontal distance could be measured. Needless to say, the articulatory situation for the subject was not very natural, however, barring more extensive studies of the vocal tract for different speech sounds.

The fiberscope method used by Gauffin & Sundberg (1978) was a little bit less unnatural for the subject and the four sustained, and isolated vowels /uː, oː, ɑː, a/ were recorded for one male subject. The fiberscope emits and receives light through two bundles of fibres and by inserting the fibrescope tube (⌀=0.5 cm) into the nose and velar opening, cinefilm frames at 16 frames/sec were recorded by swinging the tip of the tube from the left to the right side of the pharynx. When the pictures were combined they provided a composite picture of the pharyngeal cross-sections and the location of the fibrescope tip was estimated by comparing the pictures with X-ray pictures of the same subject.

With the apparition of 3D imaging methods the need for direct measurements has been greatly diminished, but dissections may still be of large importance in physiological modeling, as exemplified by Takemoto (2000), who based 3D models of human and chimpanzee tongues on gross dissections and histological examination.

Computed Tomography (CT)

CT employs X-rays that are sent through the body and registered on a bent detector. The principle is the same as for X-rays: the radiation is absorbed or attenuated to different degrees depending on the tissue. Contrary to traditional cineradiography, the radiation source and the detectors are rotated a full 360° around the patient, taking typically 1,000 snapshots of the X-ray profile at different angles. Through a backward reconstruction of the profiles registered at each orientation, a two-dimensional image of the scanned slice can be created. Sundberg *et al.* (1983) used CT to image the cross-sectional area at four 2 mm thick layers of the pharynx for two subjects for the four vowels /œ, i, u, a/ and they concluded that "the image contrast between tissues and air is high, so that unambiguous records are obtained". In total 16 tomograms were taken for each subject during an acquisition time of 3.2 seconds. It is however typical that the tomograms were acquired at four levels only, as a restriction due to radiation limits.

With a restricted corpus, modern CT can however be used for volumetric imaging of the entire vocal tract and remain within the ionizing radiation limits. Tom *et al.* (1999) used electron beam computed tomography (EBCT) to collect 60 3 mm slices of the entire vocal tract of a subject who sang the sustained vowel /ɑ/ under eight phonatory conditions, varying voice register (falsetto/chest), pitch (B-flat$_4$, F$_4$, C$_4$, D$_4$, B-flat$_2$) and loudness levels (from mp to ff). The purpose was to study the differences in vocal tract shapes for the different phonatory conditions and EBCT was chosen over Magnetic Resonance Imaging (MRI) as it "yields images of higher resolution than MRI images. The air-tissue boundary is captured with greater accuracy and bony structures and teeth are clearly imaged. Using currently available EBCT scanners, /.../ the entire vocal tract can be scanned relatively quickly (40 seconds). This comparatively brief image acquisition time greatly reduces the potential for subject fatigue and associated movement artifact".

The above comparison was however made against the MRI technique of the early nineties, when the acquisition time was 4-5 minutes (Story *et al.* , 1996). The current MRI scanners can image the entire vocal tract as quickly as the EBCT scanner that Tom *et al.* (1999) used. The resolution was about twice as high in the EBCT study (512·512 pixel images, 0.410 mm/pixel), but it should be noted that only axial images can be collected with EBCT, which reduces the accuracy in the air-tissue boundary when the slice is far from being orthogonal to the vocal tract. The higher resolution may still be advantageous in some regions as is the fact that the teeth are imaged, but the restricted corpus that can be collected, due to the radiation, makes MRI a better alternative as a tool for speech production measurements, since images of comparable quality can be obtained without health hazards for the subject.

Magnetic Resonance Imaging (MRI)

Since MRI was first used to analyze the vocal tract by Rokkaku *et al.* (1986) and Baer *et al.* (1987), (1991) it has grown to be the dominating method for measuring speech production three-dimensionally in many different languages (cf. Table 2.1), and sometimes with a specific focus (cf. Table 2.2) such as one part of the vocal tract or the variation between speakers or languages.

The success during the past decade is based on image features and quality, subject-friendliness, technical advances and availability.

The basis for MRI is that the hydrogen atoms in the body can be aligned using a strong induced magnetic field. A radio frequency pulse is directed towards the area of the body that is to be examined and the proton of the hydrogen atoms absorbs energy that makes it spin in a different direction. Using pulses of a specific frequency, the Larmour frequency, the protons can be made to precess in a determined direction. Once the pulse is turned off, the protons return to their natural alignment in the magnetic field, and in doing so they release the surplus energy, which can be captured by the magnetic coil. The data of the energy release can then be converted into a picture using Fourier transforms. Moreover, using gradient magnets that are turned on and off very rapidly, the magnetic field can be altered in a small area, which means that MRI is able to collect data in slices of 2-5 mm at any orientation. These features allow two-dimensional images of (approximately) two-dimensional arbitrarily oriented slices to be collected (as opposed to X-rays that collects a 2D image of a 3D object, and CT, where the slices can be made in the axial plane only). The images are moreover nowadays of very good quality for studies of the physiology, as opposed to ultrasound (see below), where the image is blurred. The early MRI measurements (Rokkaku *et al.* , 1986), were also quite fuzzy, due to the inability of the subject to hold the articulation for the very long acquisition time needed, but the MR image quality has increased as the acquisition time has decreased.

MRI is subject-friendly in the sense that it has no known harmful side effects and no ethical constraints need hence to be put on the amount of data that can be collected. The technique has however several aspects that are still rather subject-unfriendly, as acquisition is made with the subject in supine position in a narrow tunnel of electromagnets that produce high amplitude noise. The noise is caused by the rising electrical current in the wires of the gradient magnets being opposed by the main magnetic field and its amplitude is proportional to the strength of the main field. The most severe disadvantage of the technique is the prolonged acquisition time, during which the subject must remain immobile, as even slight movements of the scanned body part cause distorted images.

In the study by Baer *et al.* (1991) 30 minutes were required to obtain the full set of images for a given vocal tract configuration and the subjects had to produce a sustained monotone for the 3.4 minutes it took to acquire each image, breathing in briefly every 15 seconds. The technical advances, that allow the acquisition of the entire vocal tract to be made in around 30 seconds, is hence a very important contributing factor to the success of the method. The acquisition times needed are still decreasing and as we will see in the next section, even sub-second MRI is now available.

The availability of MRI scanners at hospitals has increased dramatically during the last decade. MRI machines are still very expensive and speech production measurements still have to be done at hospitals at odd times when the machine is not occupied with medical imaging of patients and with the assistance of devoted hospital physicists. The fact that MRI scanners exist at regional hospitals as a standard measurement procedure is nevertheless a significant reason for the increased use of MRI in speech production measurements.

Table 2.1. *MRI studies in different languages.*

	vowels	fricatives
English	Story *et al.* (1996)	Narayanan *et al.* (1995)
	Tiede *et al.* (1996)	Jackson & Shadle (2000)
	Gick *et al.* (2000)	
French	Demolin *et al.* (1996)	Badin *et al.* (1998)
	Badin *et al.* (1998)	
German	Hoole *et al.* (2000)	Hoole *et al.* (2000)
	Kröger *et al.* (2000)	
Japanese	Matsumura *et al.* (1994)	Niikawa *et al.* (2000)
Korean	Yang (1996)	
Norwegian	Foldvik *et al.* (1993)	

	liquids	plosives
English	Bangayan *et al.* (1996)	Story *et al.* (1996)
	Narayanan *et al.* (1997)	
	Alwan *et al.* (1997)	
	Gick *et al.* (2000)	
French	Badin *et al.* (1998)	Badin *et al.* (1998)
German	Hoole *et al.* (2000)	Hoole *et al.* (2000)
Norwegian	Foldvik *et al.* (1988) (2D)	Foldvik *et al.* (1988) (2D)
Tamil	Narayanan *et al.* (1996)	

	nasals
English	Story *et al.* (1996)
German	Hoole *et al.* (2000)
Japanese	Yang & Kasuya (1994)
Norwegian	Foldvik *et al.* (1988) (2D)

Table 2.2. *MRI studies with a specific focus.*

Velum opening	Demolin *et al.* (1998)
Relation between tongue position and midsagittal pharynx shape	Whalen *et al.* (1999)
Relation between oral cavity shape and the larynx position	Honda & Tiede (1998)
Interlanguage contrasts between Akan and English	Tiede (1996)
Intersubject variability	Apostol *et al.* (1999) Yang & Kasuya (1994)
Acoustic to articulatory inversion	Mathie & Laprie (1997) Dang & Honda (2000a)
Whispering	Matsuda & Kasuya (1999)
Generation of parametric vocal tract models	Yehia & Tiede (1997) Badin *et al.* (2000a)

Ultrasound

Ultrasound can be used for either kinematic two-dimensional (at 30-60 Hz) or static three-dimensional measurements. Ultrasound uses a transducer probe containing piezoelectric crystals, that change shape rapidly when subjected to an electric current. As the crystals vibrate, sound waves are emitted, and conversely, when a sound wave is absorbed by the crystal, it emits an electric current that can be used to obtain an image. The ultrasound probe sends out a sound wave that is reflected against a boundary in the imaged tissue and information on the boundary can hence be obtained. This means that only the outer tongue body shape can be measured as the available boundary is that between the tissue and the air. Parts where there is also air underneath, such as the tongue tip and the lateral margins, do not show up. The measurements are hence often restricted to the tongue body as the tongue root is excluded as well, obscured by the hyoid bone.

A good introduction to the ultrasound technique, its theoretical principles and properties, is given in Kelsey *et al.* (1969), which was one of the first suggestions for using ultrasound in speech research. Methodological and technical questions in using ultrasound for tongue measurements are also addressed in Keller & Ostry (1983), regarding the system setup, transducer placement and aspects such as peak detection and measurement resolution. Considerations in the subsequent three-dimensional reconstruction of the tongue surface are reviewed in Lundberg & Stone (1999) and the authors show that the tongue surface can be reconstructed using only six coronal slices.

Both Minifie *et al.* (1971) and Sonies *et al.* (1981) compared the ultrasound measurements with recordings done with X-ray fluoroscopy and concluded that ultrasound gives measures comparable, or even remarkably similar, to those from X-ray. Minifie *et al.* (1971) measured the dorsal surface of the tongue for 4 adult males who maintained the articulations of /i, æ, ɑ, u, s, ʃ/ during the longitudinal scan of 3 seconds. Sonies *et al.* (1981) reported on the development of an ultrasonic imaging system and an evaluation of the technique as such, using 30 subjects who produced eleven phonetic sequences composed of various combinations of the phonemes /i, ɑ, k, t/.

Kinematic studies include e.g. Parush *et al.* (1983), who studied lingual coarticulation in VCV sequences at a rate of 1 kHz using a single pulsed-echo ultrasound transducer and Moody (1999) who analyzed the principal components in the production of repeated voiced stops /b, d, g/ in /ə/ context with a 2-4 MHz variable frequency ultrasound scanner.

Keller (1987) used real-time ultrasound to study kinematic variables in patients with speech impairements, caused by Parkinson's disease, senile dementia, adult stuttering or cranial traumatism. One ultrasound transducer was placed below the inferior mandible at an angle adjusted to measure the dorsal movements in syllables of the type /k/+vowel. The data was hence one-dimensional, measuring the distance between the transducer and the tongue dorsum every millisecond.

Stone & Lundberg (1996), on the other hand, used ultrasound at high frequency (5MHz) emitted from 128 ultrasound crystals to scan the three-dimensional tongue surface. 60 coronal slices, oriented radially in space and 1° apart, were collected in about 10 seconds, with each 90° sector being scanned in 33 ms. A contour tracking algorithm and a lengthwise smoothing was then applied to make reconstructions of the tongue body surface of 11 vowels and 6 consonants. The reconstructed surface shapes can provide important information on tongue body position and lateral variations, such as grooving, but much information on the remainder of the tongue is missing and ultrasound can hence not be used as the only method to measure the tongue three-dimensionally.

2.1.2 Real-time data acquisition

Dynamic MRI

As was stated in the previous section, the acquisition times for MRI have decreased drastically during the past years and methods to image the moving vocal tract are emerging. One possibility is to use many repetitions of a phoneme string and generate a real-time image sequence through post-processing. Foldvik *et al.* (1993 and 1995) did pioneering work of capturing movements with MRI by the means of cardiology image processing. The subject uttered the sequence /ɑi/ every 2 seconds for a period of 4 minutes, amounting to 120 repetitions in total. The vocal tract was represented by a mesh of finite elements that was deformed during the successive stages of the utterance, thus showing a time evolving vocal tract model. As the articulation varies slightly between the repetitions, the model showed an aggregate of all the repeated articulations, rather than the true articulation, and this lead to the model displaying some discontinuities in vocal tract shape over the sequence.

Mohammad *et al.* (1997) developed a new method for increasing the temporal resolution of the MR images. Instead of synchronizing the speaker, external audio timing information was used in the post-processing, which made the speech environment more natural for the subject. The method still relied on repetitions of a short sequence, /pasi/, that the subject produced 360 times, with 6 repetitions in each scan, that lasted 2.8 seconds. Through reconstruction, using spectrogram aligning and 2D Fast Fourier Transforms, 5 mm thick, 128·128 pixel images of the midsagittal plane were obtained. The images represented 25 sequential frames with a resolution of 21 ms, showing the articulatory movements in /pasi/.

The same approach was used for pseudo three-dimensional kinematic measurements by Shadle *et al.* (1999), where three slices: left, middle (corresponding to the midsagittal plane) and right were acquired. The subject produced 24 times 12 repetitions of /pasi/ during the repeated scans of 6.6 seconds each. The three slices were 5 mm thick with a resolution of 128·128 pixels and together covering a 22 mm thick section of the vocal tract. The lips, tongue, palate, front and back pharynx walls and incisor marrow were marked up manually and the extracted contours were put together to create a 39 image pseudo-time movie with 16 ms frame rate.

Similarily, Mathiak *et al.* (2000) achieved a time resolution of 120 Hz in a selected plane, using fast MRI and a stroboscopy-like procedure, where the subjects repeated the sequences /goŋ/, /giŋ/, /gɛŋ/ and /gaŋ/ 170 times, and assuming that the repetitions were identical, the images were reconstructed into one slice per time point.

Stone *et al.* (2000) proposed a method for real-time imaging of the internal tongue, using cine MRI measurements of the midsagittal plane in a 2D study and three symmetric sagittal slices in a 3D study. The speaker had to repeat the phoneme sequences (/kɑ/ 32 times in the 2D study and /ʃɑ/ 18 times in the 3D study), but the novelty of the method was that a grid of tags was overlayed on the images, so that the movement of 40 internal points in each slice could be tracked. The principal strains were then calculated and the muscle action was inferred from local contractions assuming local homogeneous deformations and the simplest possible kinematic interpretation of the internal tongue deformations. The method hence presented an alternative to electromyography to detect muscle activation and a possible statistical method to model the internal muscular movement of the tongue.

The large number of repetitions may introduce variability in the articulations and the development of sensitive encoding systems or ultra fast Turbo Spin Echo, allowing to capture several images per second (Demolin *et al.* , 1997), is hence a great advance in dynamic MRI, as real-time capturing of slowly produced sequences can be made with the technique. Demolin *et al.* (1997) acquired 4 images per second and used these to study articulatory compensation with a bite-block and coarticulation in /beben/, /tagy/, /tyga/, /iui/, /iai/ and /ieaou/ for 4 speakers. The midsagittal images were 6 mm thick, of rather low resolution and only a small area was covered (32·128 pixels), but it demonstrated sub-second MR imaging. The technique was further refined in Demolin *et al.* (2000), where 4-5 images per second were taken of the sequence /ieaou/ for two speakers. Mády *et al.* (2001) has reported on assessment of consonant articulation in glossectomy patients using dynamic MRI that captured 8 images per second. The acquisition was in the midsagittal plane only, but the images were of good quality and (Mády *et al.* , 2001, p. 143) "hope to achieve a maximum of 15 MR images per second soon, with unchanged resolution and slice thickness", and if the development continues, MRI will soon be a real alternative for real-time measurements of very slow speech. Moody (1999) further claims that the relatively slow temporal resolution of about 30 Hz in ultrasound "is not a serious limitation for many speech sounds because most muscle-induced speech movements have bandwidths below 15 Hz (Müller & McLeod, 1982), but aerodynamically influenced movements such as trills and plosives may not be captured at this rate". The MRI technique is hence approaching a time resolution where many articulatory movements can be studied in real time. The limit at 100 Hz, which is often considered as the lower limit for visual synthesis, is however still far away, and other real-time measurements will probably still be required.

Cineradiography (X-rays)

Cineradiography has traditionally been the main information source on real-time movement in the midsagittal plane, e.g. used in the influential studies by Fant (1960, 1964, 1983). The advantage of X-ray imaging is that it provides real-time measurements of the entire two-dimensional tongue contour in upright position, and its importance in speech research is indicated by the extensive bibliography compiled by Dart (1987), listing 282 X-ray studies, done in a large number of languages.

X-ray measurements have however been drastically reduced and restricted over the last decades, as the hazards for the subjects became apparent. Instead of collecting new X-ray data, Munhall *et al.* (1995) suggested that earlier studies should be distributed and shared within the research community. They put together a 1-hour videodisc containing 25 high-quality X-ray films of 14 native speakers of Canadian English or French, compiled from studies by C. Rochette and K. Stevens in the 1960s and 1970s. The speech samples are sentences and nonsense mono- or bisyllables. This database has been used by e.g. Tiede & Vatikiotis-Bateson (1994), who implemented a video technique to find air-tissue boundaries, reduce measurement noise and track the movements of the articulators automatically. Similar work on articulator extraction has been carried out by Thimm & Luettin (1999), using the same database. Another example of the use of early X-ray films is Gick (2002), who reexplored a X-ray motion pictures of four speaker of American English made 1960 at Haskins Labs to study the occurence of a midpharyngeal constriction in English schwa.

Lately, advances in the technique permit collection of databases where the subjects are exposed to an effective radiation dose as low as 0.1 mSv (i.e. a tenth of the background annual radiation dose), during a recording time of 20 seconds, as in Stark *et al.* (1999). In that study the image quality was moreover much improved by using a paste of barium sulphate as contrast enhancement on the lips and in the mouth. A copper wire was attached to the hard palate to indicate head movements and to clarify the palatal contour.

20 seconds is however a rather short total measurement time and the corpus is hence restricted. Cineradiography is thus well suited for assessment of a model, controlling the articulatory transitions of the model against those of the subject, but other measurements are needed as a basis for the modeling if the corpus is to be large enough.

X-ray microbeam

The ethical constraints of exposing subjects to X-rays for non-medical purposes still remain, and the radiation dose has to be limited as much as possible. Using a contrast medium, as Stark *et al.* (1999) did, is one possibility to reduce the necessary radiation, using the X-ray microbeam technique (Kiritani *et al.* , 1975; Fujimura, 1991) is another. Small metallic (lead or gold) pellets are glued on the tongue surface and they provide clear point-wise data of the movement of the fleshpoints where they are fastened. The main advantage over traditional X-ray measurement, apart from the reduction in the radiation dose for the subject, is that the amount of data is reduced from a continuous shadow to clearly defined discrete points, facilitating the data processing. This of course means that information on the remaining contour is lost, but it can be reconstructed through interpolation and combination with other data sources. Badin *et al.* (1997) used cineradiography to generate an articulatory model of the subject and the tongue contour was then estimated from the three X-ray microbeam pellet positions through inverse mapping of the data onto the model. Beaudoin & McGowan (2000) used principal component analysis of X-ray microbeam data to be able to recover the pellet positions from estimations of the talker's lip aperture and location and degree of linguopalatal contact.

If the data does not have to be collected on a specific subject, it is possible to use the X-ray microbeam database at the University of Wisconsin (Westbury, 1994), that contains approximately 200 different speakers reading excerpts from novels, number names, isolated words and vowels, vowel sequences, VCVs and DARPA/TIMIT sentences. Each subject was recorded for about 19 minutes and the total database is over 3200 tracking minutes.

Electromagnetic articulography (EMA)

Another point-wise midsagittal measurement method is EMA which employs alternating magnetic fields instead of X-rays. Two to three transmitters fastened on a light-weight helmet emit an electromagnetic field that is registered in small (about 2 mm long) receiver coils (refer to Figs. 2 and 5 of Paper IV for an illustration). The strength of the electromagnetic field depends on the distance from the transmitter coils, and the variance in the field can hence be used to measure the horizontal-vertical movement of the receiver coil, once it has been calibrated for the reference position. The theory and properties of EMA are described in more detail in Perkell *et al.* (1992), where questions about calibration, accuracy and possible error sources are addressed as well.

As for X-ray microbeam, small coils are fastened on relevant articulator points in the midsagittal plane using acrylitic glue. The number of coils used has to be based on a compromise between the wanted amount of data on the one hand and the naturalness of the speech situation for the subject and the number of channels on the articulograph on the other. The Carsten's AG100 system has 10 channels (Schönle *et al.* , 1987) and Movetrack 6 channels as standard, but it can be expanded to 24 (Branderud, 1985).

One interference with the subject's natural speech is the helmet and the newly developed AG500 (Zierdt *et al.* , 2000) overcomes this by replacing the helmet by a 'cage' on which six transmitter coils are fastened. AG500 hence allows for free head movements and the sensors can further be positioned outside the midsagittal plane and in all orientations. This means that the EMA measurements have gone from being two-dimensional and point-wise to three-dimensional, but still point-wise. The three-dimensional EMA system is still very much under development and midsagittal EMA remains the standard articulography method. Two-dimensional EMA has been used in a number of studies to explore tongue movements, such as Fitzpatrick & Ní Chasaide (1999) and Nguyen-Trong *et al.* (1991). EMA studies relevant to the present work are further described in Paper IV.

Electropalatography (EPG)

Electropalatography provides point-wise binary information on the contact between the tongue and the palate, using a subject-specific acrylitic mm-thin synthetic palate on which electrodes have been placed (cf. Fig. 1 of Paper IV for an illustration of the EPG palate).

The number of electrodes differ depending on the make. The most common, the Reading palate used in Paper IV has 62, placed in rows and columns, wheras the Kay Elemetrics palate has 96, placed on semi-ellipses with varying axes lengths. Fougeron *et al.* (2000) did a comparative analysis of the two palates and concluded that the Kay palate provided more information on the subtle variations of the linguopalatal contact and had better precision, but that the Reading palate was also able to reflect the articulatory variations.

The information given by EPG may seem to be a quite limited aspect of the articulation, but it gives data on the place and manner of articulation for many consonants and closed vowels, and further on articulatory timing, lateral asymmetries and coarticulation. Examples of articulations studied with EPG are coronal consonants in Hindi (Dixit, 1999), English (Mair *et al.* , 1996) and Swedish (Lindblad & Lundqvist, 1999), sibilants in Swedish (Lindblad & Lundqvist, 1995; Engstrand, 1989) and English (Gibbon & Hardcastle, 1994), lingual stops in English (Nicolaidis *et al.* , 1995) and Shockey (1991) even measured conversational speech

Additional information can be gathered with the alternative EPG palate developed by Matsumura *et al.* (2000). It is able to estimate the linguopalatal pressure in the contact, from the force applied to each of the 15 cantilever type force sensors that are placed on the palate instead of the electrodes. The force sensors are larger than the normal electrodes (3 mm·5mm) and placed in four rows with 1, 3, 5 and 6 sensors in each.

EPG has been used in a cross-language investigation of acoustic-articulatory correlations in coarticulatory processes in the EUR-ACCOR project (Marchal & Hardcastle, 1993), where a common corpus, of VCV sequences, real words and 14 short sentences containing assimilations, weak forms etc, was recorded in 7 European languanges. Palatography is nowadays also a commonly used tool both for evaluation (Gibbon & Wood, 2001) and treatment (Hartelius *et al.* , 2001) of speech disorders.

Another aspect of EPG measurements is how the data should be interpreted, as the contact pattern is not primarily what is of interest, but rather what this contact pattern means for the tongue position and the air stream passing between the tongue and the palate. Hardcastle *et al.* (1991) presented data reduction methods and discussed what these meant for studies of coarticulation. The reduction methods consist in trying to summarize as much of the EPG pattern as possible in a few indices that capture the interesting qualities of the pattern, such as the main contact area, the asymmetry or the coarticulatory influence. The indices that are of importance in this project are described in detail in Paper IV in conjunction with their use in the study to evidence coarticulation.

The conclusions that can be drawn from the EPG data has been evaluated by Tabain (1998) and Fitzpatrick & Ní Chasaide (2000) using locus equation and EMA data, respectively. Such assessments allow for interpretations of the tongue shape for parts that are not mapped by the EPG patterns.

On the other hand, EPG *is* restricted, as meaningful results are obtained only for phonemes that have clear linguopalatal contact, and EPG is therefore often used in combination with other methods, either simultaneously or to complement other data. EPG has been used by e.g. Alwan *et al.* (1997), Narayanan *et al.* (1996-1997) to complement MRI data, by Fougeron & Keating (1996) in combination with airflow measurements and by Cohen *et al.* (1998) together with ultrasound data. The standard combination is however EPG and EMA as these measurements can be made at the same time and they complement each other, in that both are two-dimensional, but in almost orthogonal planes: the EMA data is collected in the midsagittal plane, whereas the EPG data gives information in the curved surface spanned by the palate. Combined EMA and EPG has been used in several studies of fricatives and stops, for example by Nguyen *et al.* (1998), Hardcastle *et al.* (1996) and Ellis & Hardcastle (2000).

Optopalatography (OPG)

The main disadvantage of EPG is that it measures linguopalatal contact only, a deficit that optopalatography aims at overcoming. The idea of the optopalatograph is to use a device similar to EPG, but that has the capability of measuring distance using light. This thought was first proposed by Chuang & Wang (1978), but was further developed by Wrench *et al.* (1996, 1997 and 1998). The device consists of an acrylitic palate of the same thickness as EPG palates, i.e. 0.5-2 mm, but with the EPG electrodes replaced by 8 to 20 sensors, consisting of one individual infra-red LED (light emitting diode) and one photodiode each. The intensity of the reflected light at the receiver is measured and the distance h of the tongue surface from the receiver can be inferred, as the intensity is proportional to the distance h and to variables whose values are known (the angle of divergence of the beam, the angle of inclination, the distance between the transmitter and the receiver and the reflectivity of the surface). The tongue surface can then be approximated through a Bézier curve fitting. The frame rate in the measurements is of 100 Hz, as for traditional electropalatography, and the reported error is ±0.25 mm, ±5%. The availability of the device is still scarce, as it is a research prototype, but its features are promising for future speech production measurements.

As the device by Matsumura *et al.* (2000), OPG can give an indication of the linguopalatal pressure, but on a coarser scale. The optopalatograph can indicate whether the contact is made with a light or heavy pressure, as the tongue is semi translucent and light is also reflected from below the surface.

Choosing method

MRI is now the dominating measurement method for three-dimensional imaging, with features that no other method can compete with today and with acceptable disadvantages in the acquisition. It was hence quite natural that MRI was selected as the basis for the 3D modeling and that the kinematics consequently had to be measured using another method. The length of the acquisition time is a compromise between the image quality in each slice and the number of slices on the one hand and the naturalness of the articulation and the possibility for the subject to sustain it on the other. In the present project a rather prolonged acquisition time was used, for reasons given in section 3.2.1. Concerning the naturalness of the articulations, it is evaluated in section 3.3, to ensure that the prolonged acquisition time did not make the articulations distorted.

EMA and EPG were chosen for the real-time measurements as these methods complement the MRI data well and the data from these sources can be used directly in the modeling to control the articulatory parameters (section 3.3.4) and for tuning of parameter values (section 3.3.3). The facts that the two methods could be used simultaneously, their relative ease of use and interpretation of data, combined with them being non-hazardous for the subject were other important factors for the choice.

2.2 Articulatory modeling

Modern articulatory synthesis takes as its role to generate synthetic speech using methods that as closely as possible mimic human speech production, while making simplifications that are needed for the synthesis to be feasible with the computational power at hand. In the state-of-the-art synthesizers these simplifications include considering the vocal tract as a straight cylindrical tube with varying cross sectional area, thus neglecting the bending and the asymmetries of the vocal tract. Other simplifications can be categorized by classifying the models with respect to four different properties: 2D *vs.* 3D, real-time *vs.* batch mode synthesis, physiological *vs.* statistical and tube-like *vs.* articulator modeling.

Many projects limit the model to a 2D representation in the midsagittal plane and restricting the synthesis to static sounds, hence relieving the synthesizer of the real-time constraint. With ever-increasing computer power, tasks that were not long ago impossible are now becoming reality on standard PCs and the realization of real-time three-dimensional articulatory speech synthesizers is approaching. At least for statistical models; physiological models still require quite extended computation times, as will be shown below.

The present model belongs to the category of 3D, real-time, statistical, articulator models and the following description of previous work is a review with the present model as viewpoint, to trace the path leading to it rather than an exhaustive review of articulatory modeling up until today. It should be noted that this survey concentrates on models representing the vocal tract anatomically, as opposed to merely numerically, where the input to the synthesizer is the area function directly. Such synthesizers, e.g. Tracttalk (Lin, 1990), Båvegård (1996), Fant (1992), Childers & Ding (1991), Meyer *et al.* (1989), Sondhi & Schroeter (1986), are hence omitted. Taking the present model as the viewpoint further means that the survey focuses on models of the tongue or the entire vocal tract, and hence exludes models that deal only with e.g. the velum (Wrench, 1999), the larynx (Lobo & Malley, 1996) or the jaw (Vatikiotis-Bateson & Ostry, 1995; Westbury, 1988).

The following sections on 2D and 3D modeling have been divided into parts, where similar models are grouped. This is made to increase the clarity of the presentation, but it should be noted that the division is not complementary, as some models could be classified into several groups, e.g. both statistical and for visual speech synthesis.

2.2.1 Two-dimensional models

The first computer implemented articulatory models saw light in 1966 with a dynamic model at MIT (Henke, 1966) and a functional model at Bell Labs (Coker & Fujimura, 1966).

Henke's midsagittal model (Henke, 1966) had a feature that is of interest for its similarity with the implementation in the present model, since it specified articulatory goals for only a part of the tongue. As that part moved towards the goal, the remaining parts of the tongue simply followed the motion. This resembles the parameter definition in the KTH model, with prototypes and targets, as explained in 3.1. Anticipatory coarticulation was handled by a look ahead operator and equations of motion controlled the dynamics of the model.

Functional models

The Coker-Fujimura model (Coker & Fujimura, 1966) introduced independently controllable articulators that set the geometry of the model using only seven parameters. This concept is still used in present models, and even if the articulatory parameters are slightly different, Coker's and Fujimura's intuitive parametric description remains valid. The parameters controlled the opening and protrusion at the lips, the lowering of the velum, the horizontal and vertical movement of the tongue body and the movement of the tongue tip parallel to or orthogonal to the tongue surface. It is worth noting that the model was able to produce highly intelligible speech and that it was even used in a text-to-speech system.

The Coker-Fujimura model was functional in the sense that elementary articulatory gestures were used to control the vocal tract shape, rather than the individual muscles. The standpoint is that speech production is governed by articulatory commands at a higher level and that it is not necessary to model the activity at the muscular level to replicate human speech production convincingly. This standpoint, advocated by many speech researchers, is based both on the difficulties of data collection at the muscular level and on computational efficiency in the modeling. The standpoint that the articulator movements can be explained without modeling lower level control mechanisms is however not undisputed, refuted e.g. by Payan & Perrier (1997b), as indicated below.

Another argument is that the activation of different articulator muscles is covariant, so that it can be difficult to decompose the movement into activation of several muscles, and that the decomposition might not be a one-to-one mapping, but that several different sets of muscle activation can give the same movement. The tension in all the muscles of interest would hence have to be measured, using e.g. Electromyography (EMG), if the model's movements should be based on muscle activation.

The present model belongs to the functional category, for the reasons stated above, but it is nevertheless instructive to consider the physiological models, as the definition of several of the articulatory parameters is coupled to the activity of the large muscles or muscle groups.

Physiological models

Perkell (1974) represented the tongue as a simplified muscular structure, where each muscle was a line element, modeled by lumped springs and a damper. The articulation was changed by modifying the stiffness of an active spring, and volume conservation and boundary collisions were handled using mechanical forces.

Finite element modeling (FEM) has been used to define the tongue structure by several researchers, starting with Kiritani *et al.* (1976), who proposed to represent static tongue structures with tetrahedral elements and deform the tongue shape using elastic stresses within each element (see further section 2.2.2).

Another implementation with isoparametric FEM was made by Payan *et al.* (1995, 1997a), who based the stress within the elements of the model on measurements of EMG activation in the muscles. Using the model to simulate vowel-to-vowel transitions, Payan & Perrier (1997b) claimed that kinematic measurements on speech signals can only be accounted for by using a full neurophysiological and mechanical model of human speech production, as the kinematics is influenced by morphological and dynamical properties of the vocal tract, and the speech production control can only be fully understood by considering the underlying properties.

Honda *et al.* (1994) defined a complete speech production model, including a finite element representation of the tongue and mass-spring models for the rigid structures such as the skull, spine, mandible and hyoid bone. In addition to studies of interaction of different articulators, the model also produced synthetic speech through area function estimations based on the midsagittal vocal tract shape and three-dimensional MRI data.

Davis *et al.* (1996) and Stone *et al.* (2000) used a basically kinematic model to study biomechanical features, including strain distribution, muscle stretch and needed velocity. Employing tagged MRI (cf. Dynamic MRI in section 2.1.2) and continuum mechanics criteria, the distribution and magnitude of the deformation in different parts of the tongue were investigated and interpreted in terms of muscle activation.

Motor task modeling

Another approach to modeling the vocal tract shape is the motor task concept, where the articulator movement is determined through model-defined trajectories (Bailly *et al.* , 1991). Rather than controlling the articulators by muscle activation, the movement is defined through the sequence of phoneme-specific features that the model should achieve, the motor task. The phoneme-specific features are extracted from articulatory or acoustic data, using the criterion that the features should be invariant across different utterance conditions. The motor tasks can be represented both acoustically, e.g. with the formant frequencies, and articulatorily, using e.g. the location and degree of the vocal tract constriction.

The motor task models deal with the articulator kinematics rather than the geometry, but the simulated movements result in vocal tract shapes that can be used for synthesis purposes in the same manner as the other models presented here.

Kaburagi & Honda (1996) defined *state variables*, the active movement relative the neutral reference and *tract variables*, the coordinative movement of the lips and tongue as a function of the state variables of the jaw, lips and tongue. With this definition, the model relates quite closely to pointwise data of specific articulator points, such as EMA or X-ray microbeam, and Kaburagi & Honda (1996) used EMA to define the motor tasks through least-mean-square-error fitting to the measurements.

Geometrical models

The model proposed by Mermelstein (1973) used parameters to control the jaw (a rigid rotation), the hyoid bone (horizontal and vertical movement), the tongue body (a circle with fixed radius and the center moving in the midsagittal plane), the tongue blade (a rotation relative the tongue body), the lips (opening and protrusion), the maxilla and rear pharyngeal wall (that were fixed in the synthesis). The definition of the parameters was made on a purely theoretical basis and the model's ability to replicate X-ray tracings was controlled visually.

The Mermelstein (1973) model was developed further by Rubin *et al.* (1981) in *ASY – Articulatory Synthesis program*, where six parameters from the original model were used and could be controlled either numerically from a datafile or interactively by the user. Rubin *et al.* (1996) proposed additional features to *ASY* in *CASY – Configurable Articulatory Synthesis*, in which the user could set the midsagittal shape of the vocal tract by tracing its contour superposed on a sagittal image from e.g. MRI. The CASY model could hence be configured to mimic articulations of individual speakers both in the midsagittal plane and in the relation between midsagittal distance and cross-sectional area. Further improvements that were made consisted in new models of the tongue tip and the velum and changing the relation between the position of the velum and nasality.

The Mermelstein (1973) model was also used by Boersma (1998), with the interesting modification that the vocal tract shape was controlled by muscular models rather than geometrical deformations. Simulations of air pressure, damping, coupling, collision and tension were also included to introduce an additional step in the synthesis; i.e. a part ”from muscular activity to vocal tract shape” was added to the usual ”from vocal tract shape to sound”.

Statistical models

The alternative to the above theoretical definition of parameters (i.e. top-down) is statistical models that are strictly data based (i.e. bottom-up). The latter approach is used in the present model, in the tradition of Madea (1988) and it is hence worthwhile to survey this analysis method. Maeda generated a midsagittal model from simultaneous X-ray and labio-films of 10 French sentences produced by two subjects. The tracings of the midsagittal vocal tract outline were then sampled with a semi-polar grid with 30 gridlines. This step allowed the tongue and vocal tract wall shapes to be represented as vectors in each frame, with the distance from the inner part of the grid as a function of the gridline number.

These vectors can then be used to create a linear component model, such that the vector in each frame is described as the weighted sum of the parameters and the mean value for all the data frames. The analysis of the measurement data consists in determining the weighting coefficients and then the value of the parameters. Liljencrants (1971) was the first to propose such a description, using harmonic series with three cosine components to specify the tongue vector. The sinusoidal components have however no articulatory meaning and the most common is to use a factor analysis without any constraint on the pattern instead, such as Principal Component Analysis (PCA) or PARAFAC (Harshman *et al.* , 1977). These methods have the benefit of explaining the tongue shapes optimally with as few parameters as possible, but they have a main weakness for articulatory modeling in that they do not guarantee that the extracted components represent elementary articulatory gestures. Maeda hence instead proposed the arbitrary factor analysis, meaning that the tongue surface was decomposed using PCA, but only after the effect of the jaw position had been removed using a linear regression. This resulted in four articulatory parameters interpreted as jaw position, front-back tongue body position, arching-flattening of the dorsal shape and raising-lowering of the tongue blade to describe the tongue shape. These findings on the main components generalize to other subjects (Beautemps *et al.* , 2001; Engwall & Badin, 1999) and to 3D analysis (Badin *et al.* , 1998). Perrier *et al.* (2000) suggested that the main components can be explained by the biomechanics of the tongue, mainly in the activation of the genioglossus and the styloglossus muscles.

An articulatory synthesis model for Swedish, with slightly different aims than in the present model, APEX, has been developed previously by Stark *et al.* (1996). The goal of the APEX project is to create a tool for the examination of apical sounds and speaker-dependent articulatory behavior (Stark *et al.* , 1999). The model is based on X-ray images to determine the midsagittal contour and 3D MR images for calibration of the distance-to-area conversion. The tongue body model is controlled by two parameters: position and deviation from the neutral for the most constricted part (Lindblom & Sundberg, 1971) and the apex is modeled using the two parameters protrusion and elevation from the neutral.

The models presented this far were all two-dimensional, representing the vocal tract by its shape in the midsagittal plane and then applying an empirical conversion (linear, polynomial and power function) from midsagittal distance to cross-sectional area, if the model was used for synthesis purposes. The conversion principle is further discussed in Paper I, pp. 5-6, and the different forms of transformations to generate cross-sectional areas from midsagittal measurements were evaluated in Soquet *et al.* (2002), using midsagittal and cross-sectional MR Images.

The study confirmed that the transformations were subject-specific and could only be used accurately for the original subject and further concluded that the power transformation seemed the most accurate, but that all three transformations only capture the general properties of the relationship and that the vocal tract has to be considered three-dimensionally in speech production studies and articulatory-acoustic synthesis.

2.2.2 Three-dimensional models

Physiological models

Kiritani *et al.* (1976) divided the tongue body into 14 units of elementary shape in a symmetrical linear FEM model with isotropic elastic properties and elastic stresses caused by passive elastic reaction to outside forces or active contraction in the unit.

Wilhelms-Tricarico continued work in finite element modeling with a very ambitious physiological fully three-dimensional model of the tongue (Wilhelms-Tricarico, 1997) and the mouth floor (Wilhelms-Tricarico, 2000), where the muscle fibers were modeled on data from the Visible Human Project[1]. The movements and deformations of the structures were computed as the solutions to non-linear second-order differential equations that approximate the energy functions of the structures. The elasticity of the tongue was modeled with an isotropic exponential strain energy function and the incompressibility was taken care of in each element solving a system for computing Lagrange multipliers. Wilhelms-Tricarico (1997) hence proposed a quite complex mathematical model that simulated the characteristics of the tongue through mechanical modeling of the viscosity and incompressibility. Simulations demonstrated the physiological validity of the model, but the amount of computations needed made the model very slow.

Dang & Honda (1998) created a physiological articulatory model that was quasi-three-dimensional, in that it consisted of only three sagittal planes; the midsagittal and one plane on each side, displaced 2 cm laterally from the midsagittal. The model represented the tongue, mandible, hyoid bone and vocal tract wall, reconstructed from MR Images of a male Japanese speaker. The dynamics of the tongue were improved using X-ray microbeam data for vowel and VCV sequences for 11 Japanese speakers (Dang & Honda, 2000b). All the structures were modeled by mass-points connected with viscoelastic springs of varying stiffness. The tongue consisted of 11 muscles, of which the four extrinsic muscles were used to produce vowel sequences. The computations involved are however time-consuming, and the synthesis was realized at 50 times real-time. The model can hence not be used for real-time speech synthesis, and it still requires the area function to be estimated from partial information in the model, but it is nevertheless interesting in that it reduces the computational time in the physiological model greatly compared to the finite element method most often used for physiological modeling.

Takemoto (2000) presented a very promising and impressive physiological model of the muscle fibers in the tongue, but it has this far only been used for morphological analyses contrasting human and chimpanzee tongues, rather than for speech production modeling.

Tube-like models

Physiological modeling of all articulators in 3D is very heavy and it might be unnecessary if the object of the model is to generate synthetic speech, rather than the deformation properties of the articulators. For sound generation, all that matters is really the surface delimiting the air channel, and there is no need to divide this surface into separate articulators. This approach is taken by the tube-like models and since the amount of tissue to model is diminished, the computational load to model the vocal tract with FEM is drastically decreased, allowing for a finer mesh spacing and consequently more accurate deformation modeling.

Niikawa *et al.* (2000) used about 30,000 tetrahedral elements with 7,000 nodes to model the vocal tract shape of the fricatives /s/ and /ʃ/ based on MRI measurements. The air

[1] The data consists of transverse CT, MRI and cryosection images of a representative male and female cadaver at an average of one millimeter intervals. Detailed information on the project can be obtained from http://www.nlm.nih.gov/pubs/factsheets/visible_human.html.

was supposed to flow into the vocal tract vertical to the glottis and the flow rate was then calculated throughout the vocal tract. The authors showed that the air speed was highest at the constricted passage. An electric circuit model was used to produce acoustic synthesis by cascading circular tubes and with noise as the sound source; an auditory test showed that "the generated sounds were intelligible".

Matsuzaki *et al.* (1994) and Matsuzaki & Motoki (2000) used 3D FEM to model the vocal tract as a bent tube. The first of the two studies tested non-homogeneous wall impedance; the distribution of the sound pressure in the tube was determined and experiments were made on the radiational effect. The second focused on asymmetrical vocal tract shape, where the asymmetries had been measured using MRI data. In both cases the system was evaluated by calculation of the transfer functions.

Motoki *et al.* (2000) modeled the vocal tract as a FEM structure of asymmetrically connected rectangular tubes, and Sinder *et al.* (1996) used fluid dynamics to study the air flow velocity and the resulting frequency spectrum.

Models for visual synthesis

While being adequate for sound generation, tube-like models are hard to interpret visually as the surfaces of different articulators blend together and the model disregards all parts where there is no air boundary. The second simplification to modeling the articulators physiologically takes the visual aspect into account but defines the articulatory parameters statistically based on a large number of observed articulator surface shapes.

Badin *et al.* (1998, 2000b), used so called Guided Principal Component Analysis, combining pure PCA with Linear Component Analysis, where the factors to extract are imposed by the modeler, based on articulatory measurements in MR Images, to define models of the vocal tract and the tongue. The method was developed further in Badin *et al.* (2000a), when applied to video images of the face. Several aspects of the modeling and the parameter definition in the KTH model were inspired from this work, as stated in the different papers of chapter 3.

Cohen *et al.* (1998) used a similar approach to model the surfaces of the intraoral structures. Three-dimensional ultrasound data from Stone & Lundberg (1996) was used to measure the geometry of the tongue and an algorithm for handling boundary collisions with the teeth and the palate was proposed. The correction method is of interest as a source of inspiration for the correction algorithm described in section 3.3.2. Cohen *et al.* (1998) further proposed to use EPG patterns observed in human speech as a method to tune the tongue model, but did not report on any actual tuning. This idea was employed in the current project, as reported in section 3.3.3. The similarities of the two projects stretch to the visual synthesis applications as well, where the aim is to provide a good instruction environment, e.g. for children with hearing impairment or for second language learners. The issues posed in multimodal speech instruction thus apply to both models and these issues will be covered in section 3.4.3.

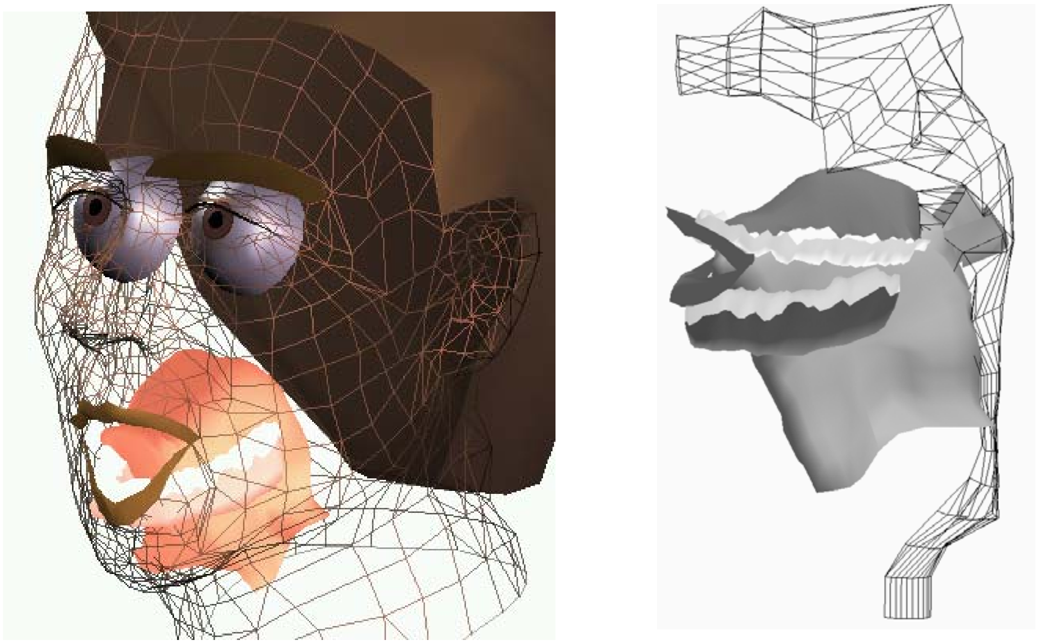
The tongue in Sams *et al.* (2000) is used as a part of a Finnish audio-visual talking head, where the tongue is visible through the mouth opening only. The shape of the tongue is therefore not very anatomically realistic, but the model is nevertheless of interest as it uses the X-ray microbeam database of speech production (Westbury, 1994) to set the values of the parameters that control the model. As the database is of American English, the parameter values were subsequently adjusted based on a Finnish visual speech database. The tongue model in Sams *et al.* (2000) is hence an example of the possibility to extract parameter values from an existing articulatory database of point-wise measurements, a method that was also employed in the present model, as described in section 3.4.1.

3. The 3D Vocal Tract model

This chapter summarizes the included papers and in doing so it gives a general overview of the KTH vocal tract model (Fig. 3.1a-b), starting with the fundamentals and aims of the project (Paper I) and continuing with a description of the articulatory measurements and the data processing (Papers II–V) that are the basis of the model.

The generation of the different parts of the model is described mainly in Paper VI, where the models of the tongue, palate and the jaw are presented, and in Paper IX, where the vocal tract walls are introduced linked to their role in the acoustic synthesis. Paper VI also describes the approaches taken to combine data from different sources to create the 3D kinematic model and to introduce physiological constraints.

One of the applications, concatenative articulatory visual speech synthesis, which starts from the ideas in Paper VI of combining data from different sources, but develops them further, using a database of articulatory measurements, is explained in Paper VII and it is evaluated in Paper VIII. Another application, the 3D-geometry-to-speech module and the acoustic output generated from the model, is presented in Paper IX. Finally future development in multimodal articulatory speech synthesis is discussed.



(a) The tongue and jaw models in the frame of a face

(b) Side view

Figure 3.1. The KTH 3D vocal tract model, based on MRI data of one reference subject. The shaded part between the vocal and nasal tract in b) is the velum.

3.1 3D modeling of the vocal tract: Paper I

Engwall, O. ***"Vocal tract modeling in 3D"***. TMH-QPSR1-2/1999, pp. 31-38.

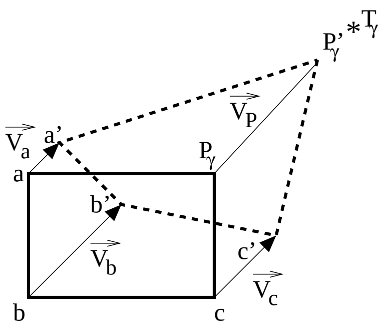
This article is an extended version of the paper presented at Eurospeech 1999 in Budapest, entitled "Modeling of the vocal tract in three dimensions" (Engwall, 1999).

The paper describes the first implementation of the KTH 3D Vocal Tract model along with its goals and potential applications. The vocal tract model is an intra-oral extension of the visual speech synthesis at KTH (Beskow, 1995), modeling the vocal tract by the same means as used for the synthetic faces. This means that the model is built up of deformable wireframe meshes, cf. Fig. 3.1a-b, that can be controlled by a set of parameters.

These parameters use the movement of a prototype towards a target and a weight function for the remaining vertices, to define each deformation, as illustrated in Fig. 3.2 (taken from Paper VI). The parameter γ deforms a rectangle using a translation defined for the prototype P_γ towards the target T_γ and the weights $(W_\gamma(a), w_\gamma(b), w_\gamma(c))$ stating the influence on the remaining vertices (a, b, c).

The model itself, as well as the implementation of the articulatory parameters, has undergone substantial changes as new articulatory data has been introduced. Some of the articulatory parameters introduced in this paper have even been excluded as the separate raising and lowering of the edges relative the midsagittal plane was automatically handled by other parameters when defining these based on a statistical analysis of measurement data. Other parameters have been completely redefined from the geometrical approach inspired by Mermelstein (1973) to a linear component model following Madea (1988). This means that Table 1 and Fig. 4 of Paper I are obsolete and that the description in the section "Articulatory parameters" is valid as a general overview of the parameters, but not in the details on the definitions. The paper is nevertheless still relevant in the description of the present model as an introduction to the deformable polygon mesh, the area function calculation and the aim of the modeling.

The advantages of using a three-dimensional model as opposed to 2D models for intraoral visual speech synthesis and articulatory speech synthesis are introduced. The increase in naturalness when going to three dimensions could be exploited in the visual feedback



P_γ = prototype vertex for parameter γ
T_γ = target vertex for parameter γ
P'_γ = deformed position for P_γ
a, b, c, P_γ = rectangle corners
a', b', c', P'_γ = deformed rectangle corners

$$P'_\gamma = P_\gamma + \overrightarrow{V_P} = P_\gamma + A_\gamma \cdot \overrightarrow{P_\gamma T_\gamma}$$

$$a' = a + \overrightarrow{V_a} = a + A_\gamma \cdot W_\gamma(a) \cdot \overrightarrow{PT}$$

A_γ = activation of the parameter γ
$W_\gamma(a)$ = influence of γ on vertex a

Figure 3.2. The definition of translational deformations in the model exemplified for a rectangle. The solid rectangle contour $abcP_\gamma$ is deformed into the dashed quadrangle $a'b'c'P'_\gamma$, due to differences in the influence of the parameter on the four corners. In this example A_γ=0.8, $W_\gamma(a)$=0.25, $W_\gamma(b)$=0.75 and $W_\gamma(c)$=0.5.

presented to hearing-impaired persons or second language learners. In articulatory speech synthesis the 3D model has the benefit of including all information explicitly in the model, whereas models that represent the vocal tract in the midsagittal plane firstly need a relation to convert from midsagittal distance to cross-sectional area and secondly are unable to model lateral variations in the tongue shape.

Fig. 3.1 shows the current version of the vocal tract model, as a part of a synthetic face or separately. The model consists of an asymmetric tongue and symmetric vocal and nasal tract walls, velum, palate, jaw and lips. The different parts of the model are described in more detail in later sections of this chapter.

3.2 Data acquisition and analysis: Papers II–V

The four papers summarized in this section describe the data acquisition, divided into the three-dimensional, but static, data, collected with MRI and the real-time, but two-dimensionally point-wise, data, acquired with EMA and EPG.

The measurements are described separately in sections 3.2.1-3.2.2 and they are investigated for agreeing and conflicting results in section 3.2.3.

3.2.1 3D measurements with MRI: Paper II & Paper III

Engwall, O. & Badin, P. ***"Collecting and analysing two- and three-dimensional MRI data for Swedish"****.* TMH-QPSR 3-4/1999, pp. 11-38.
Engwall, O. & Badin, P. ***"An MRI study of Swedish fricatives: coarticulatory effects"****.* In Proceedings of 5th Speech Production Seminar, 297-300.

These two articles were co-written with Pierre Badin, Institut de la Communication Parlée (ICP), INPG, Grenoble, France as the result of the first author's three month stay at ICP 1999 to collect an MRI database of Swedish. Badin defined the protocol for the acquisition, implemented a majority of the software used in the analysis and assisted in the acquisition, the analysis and the writing of the papers, whereas the first author was the subject in the acquisition and did the major part of the data analysis and the writing.

Paper II consists mainly of three parts. The first describes the acquisition, the second the analysis of the 3D set and the third the midsagittal measurements and the 2D linear articulatory model. The three-dimensional set is used for the present articulatory model, whereas the midsagittal set serves as reference both in the image analysis and in the articulatory modeling.

The role of the MRI measurements was to obtain three-dimensional data of high enough accuracy for the modeling, which meant (i) high-resolution images, with (ii) small inter-slice spacing (iii) covering the entire vocal tract. 54 images of each articulation were collected in the 3D set, each 3.6 mm thick and with a 4.0 mm centre-to-centre inter-slice spacing. The images were 256·256 pixels and had a resolution of 1 mm/pixel, and this required an acquisition time of 43 seconds with the 1.0 T Philips Gyroscan that was used.

The 3D corpus consisted of all Swedish long vowels, i.e. /ɑː, eː, æː, iː, yː, uː, ʉ̟ː, oː, øː, œː/ and the three short vowels /a, ɵ, ɔ/ in isolation and the voiceless consonants /p, t, k, l, r, f, s, ɕ, ʂ, ɧ/ in symmetric /a ɪ ʊ/ context, while the midsagittal corpus included some additional short vowels, nasals and retroflexes as well. The selection of the articulations to include in the corpus was based on a compromise between an aim of collecting as complete a corpus as possible and the time limit imposed by the availability of the MRI scanner and subject fatigue. All long vowels were included to cover a large part of the subject's articulatory space and the consonants were chosen to cover all types and places of articulation. It was assumed that the voiced consonants could be modeled on their voiceless counterpart. This assumption has not been validated other than by using the conclusions from earlier studies (Narayanan *et al.* , 1995 found that voiced fricatives had slightly larger areas immediately behind the constriction, significant tongue root advancement and slightly higher posterior tongue region, but the over all correspondence between voiced and unvoiced fricatives was large).

The subject was a 27-year-old male native speaker of Swedish, the present author, who has no dental fillings that could distort the MR Images. The reasons for choosing one reference subject instead of a larger group were both theoretical (that intersubject variability may obscure interphoneme differences in a model based on the subjects' mean articulations, as outlined in Paper II) and practical (as the measurements were made at several different occasions and locations, one important constraint was that the same subject should be available for all measurements. Taking the author as reference subject was hence the simplest solution).

The paper further describes the analysis of the 3D data set, with the contour extraction, mesh construction and the resulting area functions. The method is evaluated for measurement artifacts using both synthesis from the area functions and inspection of images and reconstructed vocal tract shapes.

The use of a dental cast measured separately to overcome the problem that the teeth do not show up in MRI measurements is discussed. It is worth noting that while this meant that the dental structures needed to be introduced later on in the reconstruction, as opposed to methods (Wakumoto *et al.* , 1996) where the border is introduced using a film with contrast medium for MRI, the dental casts also allowed models of the palate and jaw to be made from the same measurements, as will be described in section 3.3.3.

The midsagittal set was used for articulatory measures that were evaluated against earlier studies to assure that the subject's articulations were representative. Correlations between the measures were further calculated to indicate which articulatory parameters the model needed and how these should be defined. A 2D articulatory model using 9 parameters was defined using guided PCA (Beautemps *et al.* , 2001) and the model was evaluated in terms of its ability to explain the variation observed in the corpus. The linear midsagittal articulatory model is important for the three-dimensional modeling that followed, not because the midsagittal model was used directly, but because it guided the parameter extraction and definitions in the 3D set.

Paper III describes the evaluation of the sub-corpus consisting of the Swedish fricatives in different vowel contexts. Its main focus is to control if coarticulation can be evidenced in the artificially sustained articulations in the MRI data. The fricatives were measured in [a ɪ ʊ] context in the 3D set and in [a ɪ ʊ ɔ] context for the midsagittal set and the different articulations were hence expected to be influenced by coarticulation. It is however not evident if, and what, coarticulatory effects are captured when the articulation is static, as in the MRI acquisition. Paper III shows that coarticulatory effects could be found for the place of constriction, the lip protrusion, the jaw and larynx height and the area function. A synergetic movement of the lips and the larynx moreover increased the coarticulatory influence by lengthening the vocal tract at both ends in rounded vowel context. Synergetic coupling was also found between the jaw advancing and the lip protrusion.

Fig. 3.3 complements section 3 of Paper III with an illustration of the coarticulatory influence on the place of constriction. The coarticulation on the place of articulation is null for /f/, small for /ç, ʂ/, somewhat larger for /s/ and more important for /ɧ/. Fig. 3.3 is worth considering in relation to the discussion in Paper V of the static *vs.* kinematic lingual

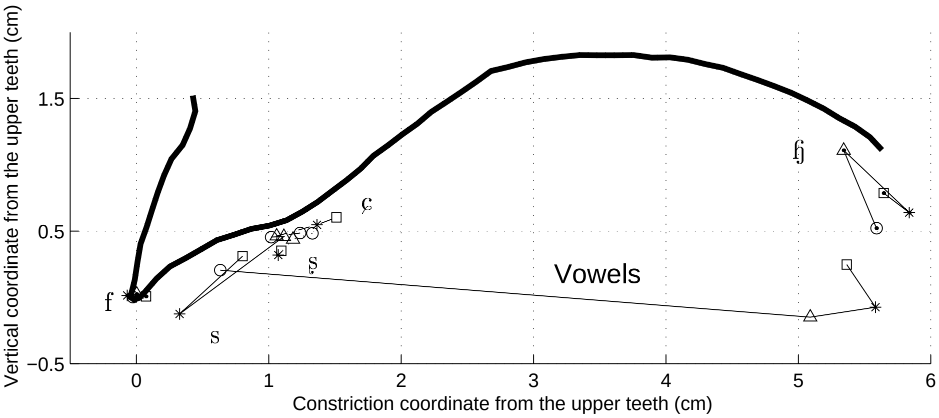


Figure 3.3. The coarticulatory influence on the place of constriction, measured as the point on the tongue contour with minimal distance to the palatal outline. Context caption: ○ a–ɑː, △ ɪ–iː, ✱ ɔ–oː and □ ʊ–uː. The short vowels refer to vowel context for the fricatives and the long to the isolated vowels.

coarticulation. /ɧ/ allows for more lingual coarticulation, whereas the places of articulation for /s, ç, ʂ/ are so close that little variation can be allowed, agreeing with the findings of the EPG analysis of Paper IV.

3.2.2 Kinematic measurements with EMA & EPG: Paper IV

Engwall, O. ***"Dynamical aspects of coarticulation in Swedish fricatives – a combined EMA & EPG study".*** TMH-QPSR 4/2000, 49-73.

MRI is well suited for the 3D measurements, but it is yet unable to measure the speech production kinematics directly and other methods are needed to collect this data. The choice fell upon combined EMA and EPG, for reasons given in section 2.1.2, and the study in Paper IV focused on Swedish fricatives in different vowel contexts.

Fricatives were studied based on mainly four different considerations. Firstly, the fricative group has a large interval of places of articulation, from the labiodental /f/ to the velar /ɧ/ and the measured transitions hence cover a large part of the kinematic articulatory movements that the tongue model should be able to replicate. Secondly, the palatal contact is important for the fricative articulations (except for /f/) which makes the EPG measurements relevant. Thirdly, as small articulatory changes result in large acoustic changes, especially for the distinction between /s, ç, ʂ/, it is of interest to investigate the fricatives for the allowed variability due to coarticulation in the production. Finally, coarticulation in the Swedish fricatives had been studied using MRI previously (section 3.2.1) and choosing the same corpus hence made assessment of the static MRI data possible, as described in section 3.2.3.

For the EMA measurements, the Movetrack electromagnetic articulography system (Branderud, 1985) was used to determine the articulatory movement in the midsagittal plane of the five fleshpoints where the receiver coils were fastened. Measures of the jaw position, lip protrusion and the tongue body were made at the onset and the offset of the fricative to detect the articulatory movement in the fricative. The stable articulatory phase in the fricative was further measured to evidence coarticulatory effects.

EPG was employed in much the same fashion; measuring the contact pattern at the fricative onset, offset and within the fricative to conclude on coarticulatory effects and the movement of the tongue.

The main contributions of the study described in Paper IV is that it allows for assessment of the MRI data (section 3.2.3), that the EPG data provides a good alternative for parameter tuning in the model (section 3.3.3), that the EMA data can be used to control the kinematics of the tongue model (section 3.3.4) and that the data on the timing between different articulators can guide the adjustment of filters in the KTH text-to-visual-speech synthesis to get a correct timing of the articulations (section 3.3.4). These contributions are much more important than the findings on the coarticulation as such, even if this study appears to be the first where Swedish fricatives have been studied using EMA measurements.

3.2.3 Data assessment: Paper V

Engwall, O. ***"Are static MRI data representative of dynamic speech? Results from a comparative study using MRI, EMA and EPG"****.* In Proceedings of the 6^{th} ICSLP, I:17-20.

One of the main disadvantages of MRI acquisition is that it requires the subject to sustain the articulation artificially; for 11 seconds for the midsagittal and 43 seconds for the 3D set in the present database. As the MRI data is used to model running speech it is of great importance to decide whether or not the data can be considered as a true representation of speech at normal speaking rates and where modifications are needed. Paper V evaluates the naturalness of the MRI data compared to the kinematic EMA–EPG measurements.

The subset of the corpus used in Papers III and IV, consisting of fricatives in vowel context, was employed and articulatory measurements of the positions of the jaw, the lips and the tongue were compared to indicate similarities and differences in the two data sets. As the corpus included three different vowel contexts, the data could be evaluated both with respect to fricatives and to coarticulation.

The conclusion of the comparison was that the coarticulatory aspects found in the MRI data (Paper III) were also found in the EMA and EPG data (Paper IV) and that the relation between different fricatives is quite consistent between the two studies. Substantial differences were nevertheless also found, in that the MRI measurements showed more extreme articulations and less coarticulatory influence on the tongue contour. The static articulations had larger differences in the lip protrusion and the jaw height between contexts, but smaller on the tongue position. This indicates that the static articulations were hyper-articulated in the MRI measurements, which is explained by the fact that the subject aimed at producing as prototypic articulations as possible and then holding them for a prolonged period. The aspects on which the fricative articulation allow for greater variation to adjust to the articulation of the surrounding vowel, such as lip protrusion and jaw height, were more coarticulated as a result of the subject's aim to produce as clear instances of the articulation as possible. On the other hand, the aspects that are fricative specific, such as the tongue contour, were less coarticulated and more typical of the fricative in the MRI data, as the subject always had ample time to produce the articulations before the acquisition started, and the articulatory targets of the fricative were hence always reached in the MRI acquisition, leading to less influence of the vowel.

The consequence of this conclusion is that the MRI data can be used for the kinematic three-dimensional model, but that the data needs to be complemented with real-time measurements that can control the movements of the model and tune the articulatory parameters to levels used in running speech.

3.3 Combining data in the model: Paper VI

Engwall, O. ***"Combining MRI, EMA & EPG measurements in a three-dimensional tongue model".*** Speech Communication

Paper VI presents the three-dimensional tongue and teeth models that were generated from the MRI data, the tuning of the tongue model with EPG data and its kinematic control using EMA data. The algorithm to handle contacts between the tongue and the intraoral structures is also described.

3.3.1 The 3D linear component tongue model

The tongue contour was first extracted manually from each MR Image and the tongue surface was reconstructed for each articulation, using a semi-polar grid. The tongue shape for every articulation was then re-sampled, so that each tongue shape consisted of an equal number of vertices spanning a regular polygon mesh.

Based on these meshes, the deviation from the neutral tongue shape was calculated for each articulation and a set of articulatory parameters was defined through a linear component analysis to explain the variance in the corpus. Articulatory measures of the jaw height and the tongue contour in different regions were used to guide the analysis, in which the six parameters jaw height, tongue body, tongue dorsum, tongue tip, tongue advance and tongue width were defined. The definition of the parameters and their influence on the tongue shape are presented in the paper. Figs. 3.4-3.5 show the range of activation for the six parameters defined in Paper VI.

The first five factors were able to explain 88% of the variance in the midsagittal plane and 78% of the total variation in sagittal coordinates and the RMS error between the measured tongue shapes and the model's was 0.13 cm.

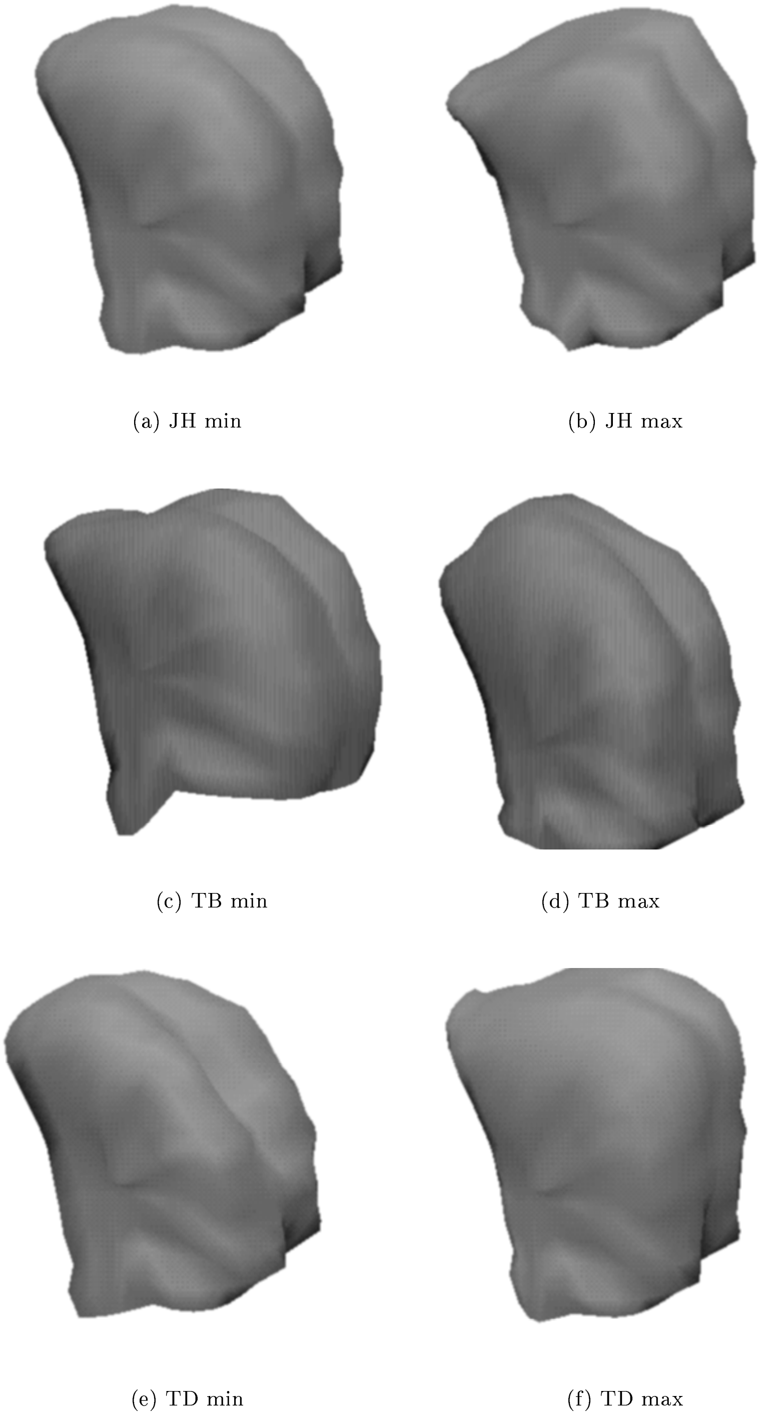
The linear model is thus able to replicate the tongue shapes in the corpus quite accurately, and the linear parameters handle lateral differences, such as tongue grooving and asymmetries, automatically. The paper also includes an evaluation of the error as a function of phoneme and vowel context, that shows that the reconstruction error is clearly affected by the openness of the vowel.

3.3.2 Introducing constraints on realism

The wireframe meshes in the model are not physical surfaces and this means that basic physical constraints that come automatically for the real tongue have to be introduced as rules or corrections in the model. These constraints are volume conservation of the tongue, that only certain tongue shapes are physically possible and that the surfaces are impenetrable so that boundary collisions need to be taken care of.

The first two constraints are handled implicitly as the model is based on measurements that define statistically how the articulatory parameters are combined for the articulations in the corpus. Making other combinations of the parameters may result in tongue shapes that are in conflict with the incompressibility of the tongue or the distribution of the muscles, but as long as the transitions between different articulations are within the articulatory space spanned by the articulations of the corpus, this risk is minimized.

The third constraint needs to be introduced explicitly, however. Perkell (1974) introduced an impenetrability mechanism where a force was applied to the fleshpoint to bring it to a physically proper position. The force was proportional to the stiffness of the wall K_I and the distance d that the fleshpoint had reached beyond the impenetrability threshold,



(a) JH min

(b) JH max

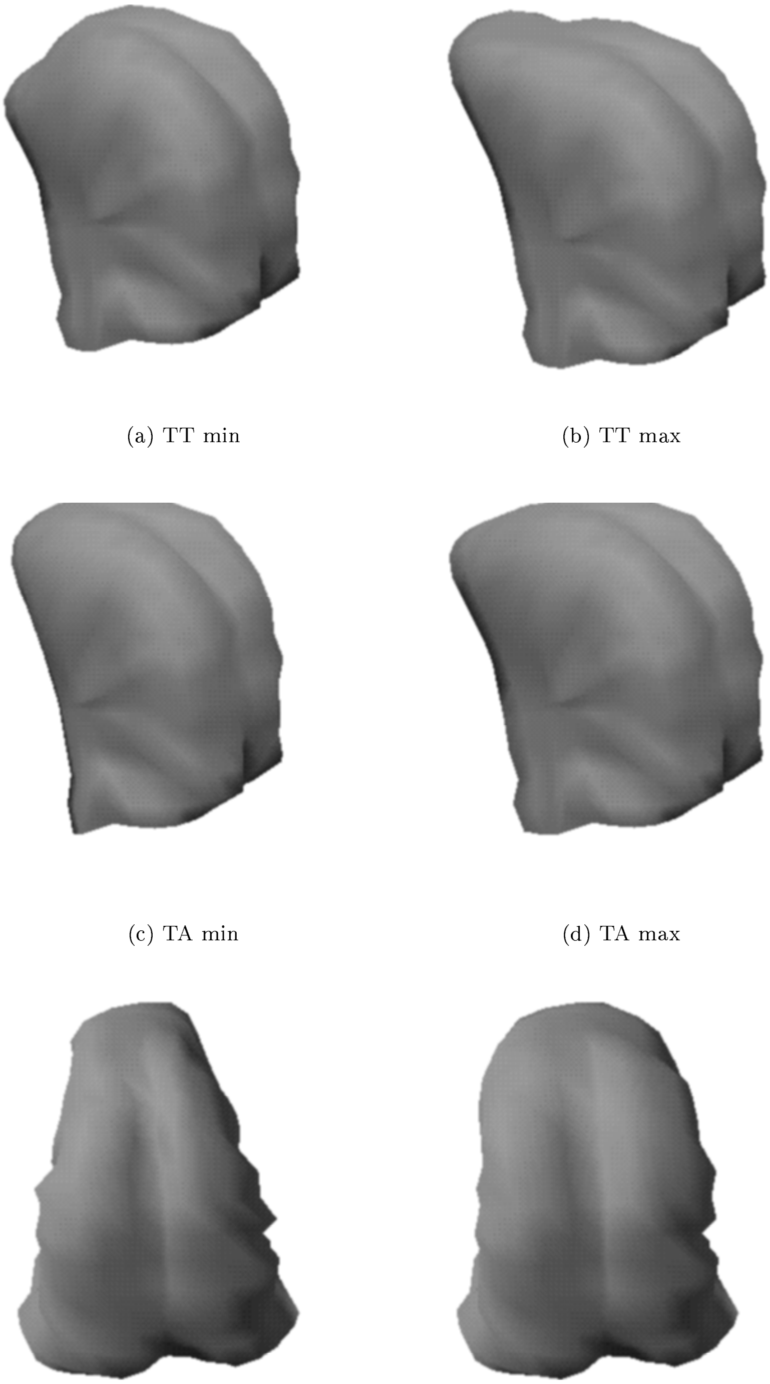
(c) TB min

(d) TB max

(e) TD min

(f) TD max

Figure 3.4. The range of variation for the tongue deformation caused by the parameters JH, TB and TD.



(a) TT min

(b) TT max

(c) TA min

(d) TA max

(e) TW min

(f) TW max

Figure 3.5. The range of variation for the tongue deformation caused by the parameters TT, TA and TW.

$|\overrightarrow{F_I}| = K_I \cdot d$. Dang & Honda (1998) used a similar, but more complex, approach, calculating the energy and momentum $m_0 v_0$ of a tongue node m_0 to determine the reaction forces from the wall that bring m_0 to an equilibrium position.

The approach chosen in the current project is substantially simpler, as all surfaces are treated as unyielding and the impenetrability is achieved by mapping fleshpoints that otherwise would reach beyond a surface onto that surface. The method bears several similarities to the algorithm proposed by Cohen *et al.* (1998), but the implementation is optimized to the KTH tongue model to speed up the detection and correction of unphysical fleshpoints. The key feature of the algorithm is that the detection and correction is carried out against a finer, regular mesh of sagittal and lateral lines, created from interpolation and sampling of the original boundary. This regular boundary surface is ideal for the boundary collision handling for several reasons:

i) The identification of the polygon closest to the fleshpoint to check is simplified. As the coordinates of the corners of every polygon in the boundary mesh are integer millimeters, the polygon is identified directly from the coordinates of the fleshpoint. For example, if the fleshpoint has the coordinates (12.456, 3.932, 8.128), then the closest polygon is the one that has corners at (12, 2, z_1), (12, 4, z_2), (14, 4, z_3) and (14, 2, z_4), where the values of the z_i-coordinates are determined by the boundary mesh.

ii) The sagittal and coronal lines are ordered with increasing x-coordinate as the first criterion and increasing y-coordinate as the second. This means that the coordinates of the polygon corners in i) can be found using a fast search through the vertices: Scan the vertices until the first vertex with matching x-coordinate is found, then continue the scan amongst the following vertices (that have the same x-coordinate value) looking for the one that has a matching y-coordinate. That is the first corner of the polygon searched for, and its z-coordinate is taken directly from the list.

iii) The correction is carried out in the same step as the detection, by mapping a violating fleshpoint upon the closest vertex in the boundary mesh. This speeds up the correction substantially, as no search for the closest polygon and the projection onto it is made.

iv) As the boundary mesh is much finer than the tongue surface mesh, the risk of clustering of tongue fleshpoints through the correction is small.

The correction algorithm used takes less than 10 ms/frame, which means that the model can be displayed at an average frame-rate of 100-130 fps. It is hence a simple method that gives approximatively correct articulations while maintaining the real-time generation of the articulatory movements, which is important for the applications of the model.

3.3.3 Adjusting the model to running speech

As was evidenced by the study in Paper V, the tongue model defined in section 3.3.1 represents hyper-articulated articulations and real-time data is needed to tune the parameters of the model to values that generate articulations used in running speech. One possibility is to use the EPG data on the natural linguopalatal contact and compare it to the synthetic contact patterns of the vocal tract model. The parameters of the model can then be adjusted to replicate the natural contact patterns as closely as possible.

This part of Paper VI describes how a synthetic EPG palate was generated to allow for comparisons with the natural contact patterns and it further reports on the new palate and jaw models that were created for the KTH Vocal Tract model.

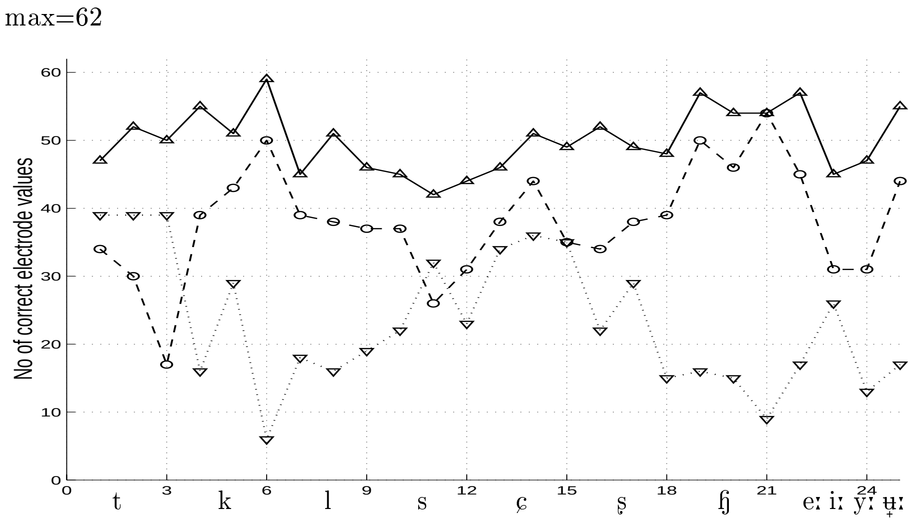


Figure 3.6. The increase of correct electrode values from the initial (◯) to the adjusted (△) model and the number of contacted electrodes in the natural pattern (▽). Vowel context from left to right for the consonants: /a, ɪ, ʊ/.

The results of the parameter tuning are discussed in Paper VI, but Fig. 3.6 summarizes the results in a slightly different manner, showing the increase in the number of correct contacts, rather than decrease in wrong. It also shows the number of contacted electrodes in the natural pattern, to indicate the relation between the modeling result and the amount of linguopalatal contact. The general relation is that articulations with fewer contacts are modeled better, which is natural as the contacted electrodes often are more homogeneously grouped in these articulations and it is easy to decrease the linguopalatal contact by lowering the jaw. There is also a weaker tendency that the total amount of correct electrode values is higher for articulations with large linguopalatal contact. As can be seen from Fig. 3.7 there is no direct connection between the number of contacted electrodes and the tuning result, even if the graph shows a local minimum for articulations that have about an equal number

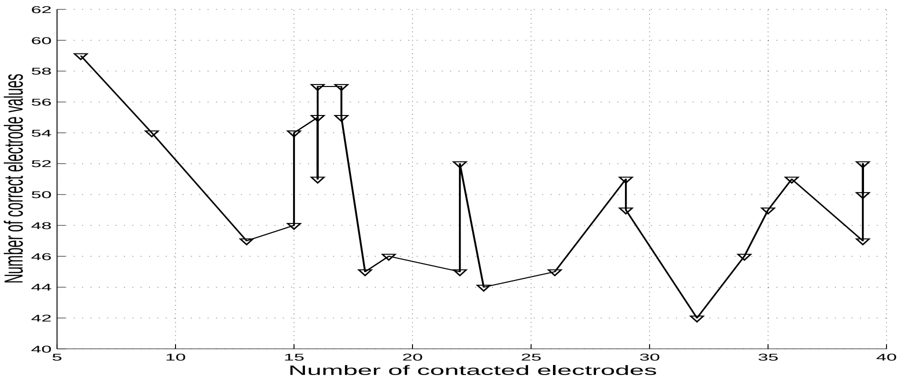


Figure 3.7. The relation between the number of contacted electrodes in the natural pattern and the result in the model tuning.

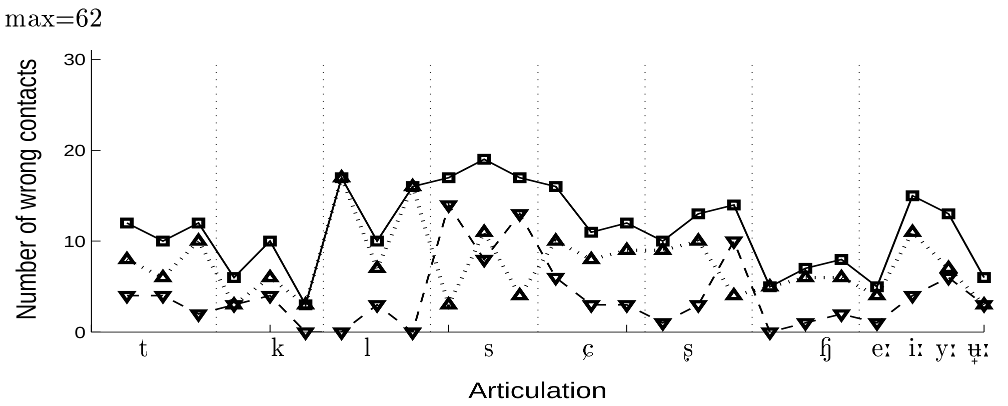


Figure 3.8. The type of remaining error for the tuned synthetic articulations compared to the natural contact pattern. □ = total error, ▽=false contact and △=missing contact. Vowel context from left to right for the consonants: /a, ɪ, ʊ/.

of contacted and non-contacted electrodes. There must hence be articulation specific effects that affects the outcome of the tuning, as discussed in Paper VI.

Fig. 3.8 shows the remaining error in the model, with the contribution of false and missing contacts. As could be expected from the tuning method, where false and missing contacts are penalized equally, the number of false contacts in Fig. 3.8 is largest for one of the phonemes, /s/, where the correct contacts contribute the most to the total amount of correct contacts. On the other hand, the other two such phonemes, /t, ʂ/, have about the same amount of false contacts as the other articulations and the fear that aiming at maximizing the number of correct contacts would lead to a large amount of erroneous contacts seems to be unfounded.

Fig. 3.9 illustrates one of the main problems of the tongue model compared to the natural data that was discussed in Paper VI, i.e. that the lateral asymmetry is not handled to the full extent. The synthetic patterns did not achieve the same amount of left asymmetry that the natural patterns had and the variation is much larger in the synthetic patterns. A solution to this problem, using optopalatography to tune the weight functions of the parameter deformations, is discussed in Paper VI.

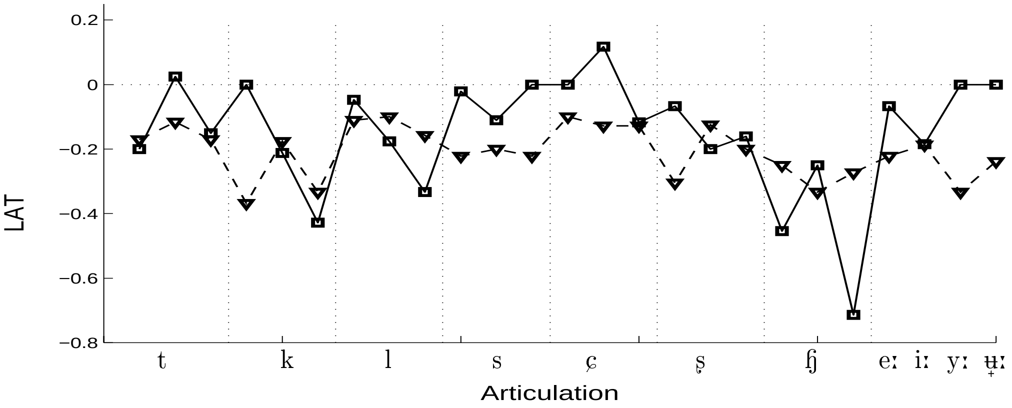


Figure 3.9. The lateral asymmetry in the tuned synthetic (□) and the natural (▽) contact patterns. Vowel context from left to right for the consonants: /a, ɪ, ʊ/.

3.3.4 Combining real-time and 3D data

Some kinematic control is needed for the transition between the static articulations created in section 3.3.1. This part of Paper VI suggests that EMA data can be used either directly or as a basis to generate synthetic kinematic control sequences. This suggestion is founded on the fact that the EMA coil's were placed on flesh-points whose moves match the definitions of the articulatory parameters for jaw height, tongue body, tongue dorsum and tongue tip, respectively. As discussed in Paper VI, this is a simplification, as there is no exact correspondence between the physical points where the EMA coils are fastened on the real tongue and any one of the points on the tongue surface measured with MRI. The approximation is nevertheless relevant as this is how EMA data could be used in a future possible application of the model in speech production training, where the patient has to get feedback in real time of his or her tongue movements that are measured with EMA.

The EMA data is well-suited to use in the parameter control as the output from each receiver is given as a function of time of the coil's Euclidean deviation (horizontally and vertically) from the reference position. This output is quite close to the definition of parameter activation in the model, with the exception that the deviation in the model is given as the percentage of a movement towards a predefined target rather than in centimeters. If the movement of the EMA coil can be considered to be towards such a target, then the coils' output can be transformed to the activation level in the model by dividing the output with the maximal deviation in that direction and using the obtained activation function directly in the model, as shown in Fig. 3.10.

One alternative to using the entire data output directly is to take measures at relevant points of the phoneme sequence and generate a sinusoidal activation function that passes through these points (onset, middle and offset of the fricative and the middle of the vowel). This activation function leaves the model with more flexibility as the transitions between the measured points can be generated through rules in the model. Fig. 3.11 shows the activation functions for the Swedish sentence ”Fiffig Orsa-farsas vassa chassi.” (”The ingenious Orsa daddy's sharp chassis”, Orsa being a town name) with the transcription /fɪfːɪ ʊʂa faʂas vasːa ɧasːɪ/. The sentence includes only phoneme transitions that were measured in the study in Paper IV and rules for sinusoidal interpolation were used to determine the trajectory between the measured articulatory values for each parameter.

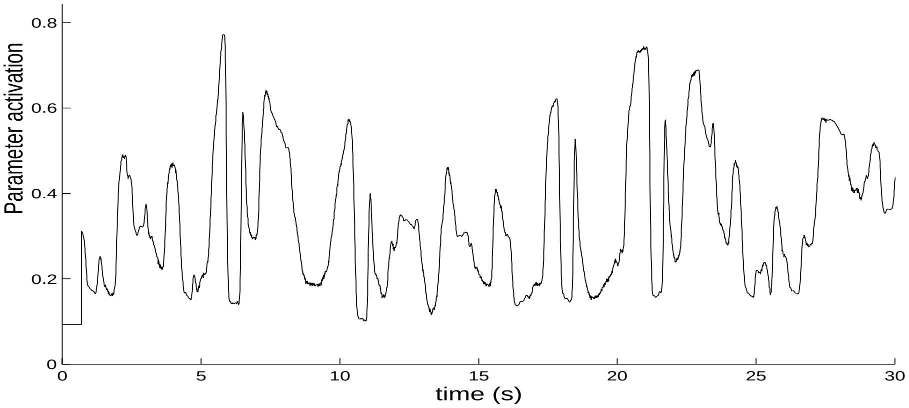
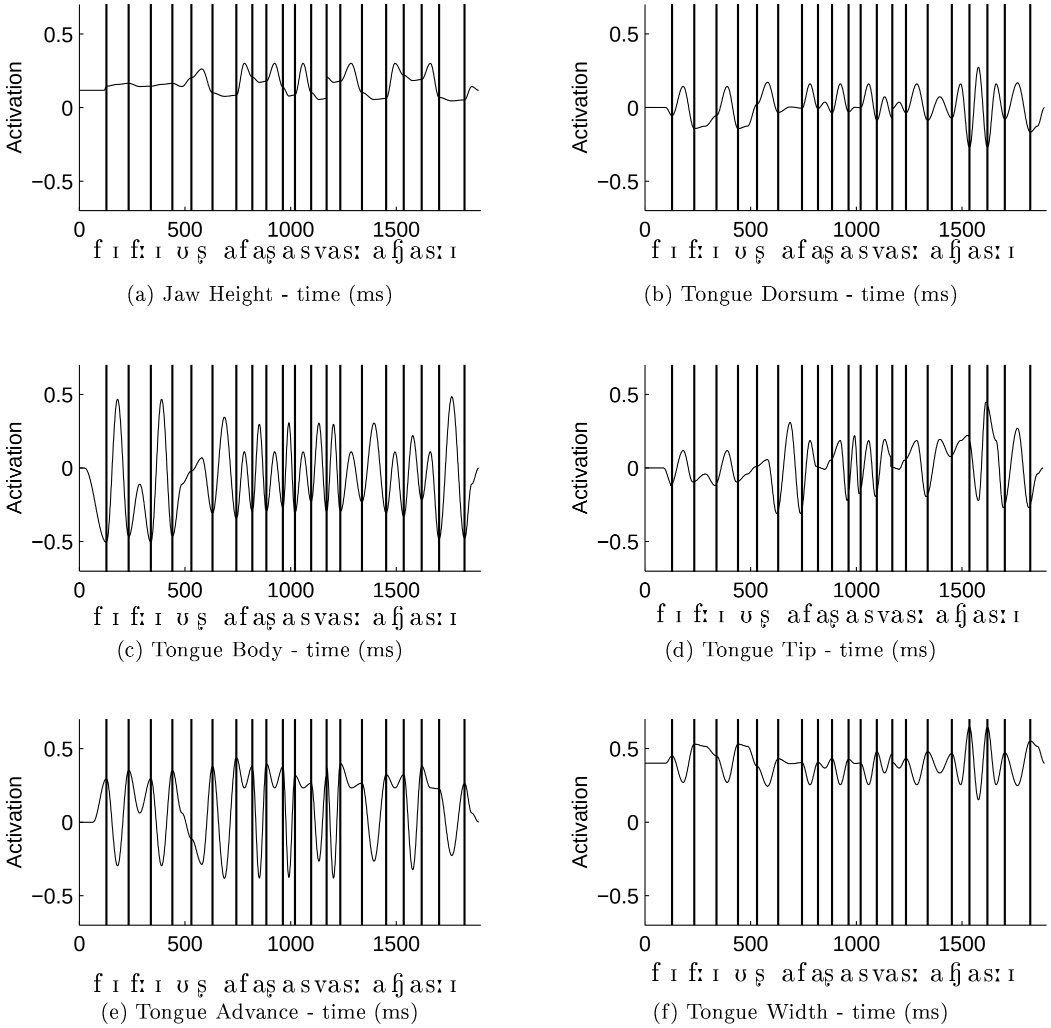


Figure 3.10. The control sequence for a sequence of 16 VCV units for the parameter JawHeight, created by scaling the articulatory measures of the movement of the incisor coil J in the y-direction to the model's parameter activation level.



(a) Jaw Height - time (ms)

(b) Tongue Dorsum - time (ms)

(c) Tongue Body - time (ms)

(d) Tongue Tip - time (ms)

(e) Tongue Advance - time (ms)

(f) Tongue Width - time (ms)

Figure 3.11. Rule-generated activation functions for the phoneme sequence /fɪfːɪʊʂafaʂasvasːaɧasːɪ/ using articulatory measures and sinusoidal interpolation.

This approach is still limited to phoneme sequences that have been measured and another solution is called for in the general text-to-visual-speech application, where phonemes may be combined in (almost) any order. A text-to-intra-oral-visual-speech synthesis module has hence been developed as an extension to the KTH text-to-visual-speech synthesis (Beskow, 1995), specifying the prototypic values for each articulatory parameter in every phoneme and then leaving the transition control to the model.

The multimodal TTS module, described schematically in Fig. 3.12, is based on the RULSYS text-to-speech rule synthesis framework (Carlson *et al.* , 1982), where the orthographic text is transformed to strings of phonemes (for the audio output) and visemes (for the visual modality). The visual synthesis has been extended to include intraoral articulations as well, by adding three parameters: tongue body, tongue dorsum and tongue width to the parameters defined for the synthetic face, and replacing jaw rotation by jaw height, apex by

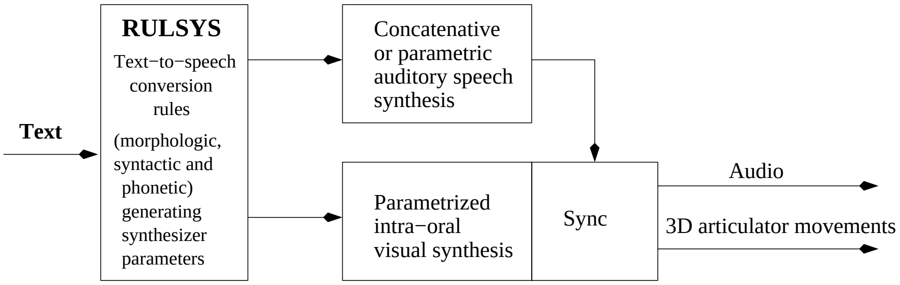

RULSYS
Text−to−speech conversion rules
(morphologic, syntactic and phonetic) generating synthesizer parameters
Text
Concatenative or parametric auditory speech synthesis
Parametrized intra−oral visual synthesis
Sync
Audio
3D articulator movements

Figure 3.12. Schematic overview of the multimodal text-to-speech system (after Beskow, 1995).

tongue tip and tongue length by tongue advance to correspond to the articulatory measures of the tongue. A stand-alone application, taking a text string provided by the user and generating the corresponding multimodal synthesis, has been developed. The tongue parameters have been assigned different time constants using time variable filters in RULSYS to get the right kinematic properties. These time constants have partly been based on the EMA study described in section 3.2.2 to replicate the observed differences in the timing of the articulators.

3.4 Applications

3.4.1 Concatenative Articulatory Synthesis: Papers VII & VIII

Engwall, O. ***"Concatenative Articulatory Synthesis".*** Preprint submitted to Journal of Phonetics, 2002.

Engwall, O. ***"Evaluation of a System for Concatenative Articulatory Visual Synthesis".*** Submitted to the 7^{th} ICSLP, Denver, Colorado, September 16-20, 2002

Paper VII starts from the propositions made in Paper VI, that EMA data can be used to control the articulatory parameters in the 3D tongue model. In this case, rather than using a small database of the reference subject's production, a large, freely available articulatory database was used to generate less restricted synthesis.

The choice fell upon an EMA database, MOCHA-TIMIT, collected at the University of Edinburgh to train an articulatory based recognizer (Wrench & Hardcastle, 2000), because it contains 460 phonetically balanced sentences, and it is hence well suited for concatenative synthesis. This term is usually considered to signify concatenative *acoustic* synthesis, when previously stored waveforms of phonemic transitions are glued together to create speech sentences. Paper VII proposes to use the same ideas for articulatory synthesis, i.e. to store articulatory transitions in a database, subsequently used to create articulatory synthesis by concatenation. MOCHA-TIMIT includes EMA measurements of coils placed on the incisors and the tongue, and the data from these coils can be used to control the articulatory parameters, in the same manner as outlined in Paper VI.

Articulatory transitions from the middle of one phoneme to the middle of the next, given the name *diart* in analogy with diphones, are used. The scaling of EMA data, the creation of the diart database and the text-to-concatenative-articulatory synthesis algorithm are described, and the discussion contains views on the potentials and evaluation of concatenative articulatory synthesis.

The question of evaluation of the system is further addressed in Paper VIII. MOCHA-TIMIT is not the only database available; other databases have used other measurement methods, and could be used to validate the intraoral visual synthesis based on the MOCHA-TIMIT database.

Four of these databases, the ultrasound database (Stone & Lundberg, 1996) of surface shapes for isolated English vowels and consonants, the X-ray Film Database for Speech Research (Munhall *et al.* , 1995), the X-ray microbeam database at the University of Wisconsin (Westbury, 1994) and the EUR-ACCOR database (Marchal & Hardcastle, 1993), have already been mentioned in sections 2.1.1-2.1.2.

The EPG data could be used in an evaluation of the linguopalatal contact in the manner proposed in Paper VI, and the ultrasound data to assess the model's tongue shapes. The ultrasound evaluation would however be restricted both in that the corpus is isolated phonemes and that the acquisition is non-real time. Real-time ultrasound studies have often been restricted in corpus and focus, as e.g. to cricothyroid space in Vilkman *et al.* (1997) or tongue dorsum movement in speech impaired subjects as in Keller (1987).

In Paper VIII, two methods of evaluation were used, against the MOCHA-TIMIT database itself and against the X-ray data of three other speakers included on the videodisc distributed by Munhall *et al.* (1995).

For the MOCHA-TIMIT evaluation, the original movement of the EMA coils was compared to the movement of the corresponding synthetic coils placed on the tongue model. The synthesis was made using the entire diart database, except for the original sentence that the synthesis was to be compared to. This evaluation method provides a good and easy measure of the quality of the synthesis method, as the trajectories of the natural and synthetic coils can be compared in a one-to-one mapping and a simple error measure of the distance between target and synthetic positions can be applied.

The above intrasubject evaluation is however only point-wise, of the points where the EMA coils are fastened. This means that differences in the detailed movement of the natural and synthetic coils will appear, but that the overall quality in tongue shape and articulatory transitions may be overlooked.

A more general evaluation, comparing the midsagittal tongue shape of the model with X-ray films of tongue movements of three Canadian speakers was hence carried out. The same sentences as in the films were synthesized and the articulatory postures and transitions were compared to the tongue contour segmented from the X-ray film.

The evaluation showed that while the movements of the model, controlled by concatenative articulatory visual synthesis, were similar to those in the cineradiographic films, there still is room for improvement, concerning both the movements of the tongue tip and root, which move less than in the natural data, and the concatenation process, as rare diarts may cause bad joins, if there are few realizations to choose from and these realizations differ substantially from the target.

The question of the length of the units to concatenate is addressed briefly in the discussion in Paper VII, and with the results of the evaluation in Paper VIII in mind, it is probable that the concatenative articulatory synthesis would be improved with syllable-sized units, in accordance with the proposition that the unit for articulatory organization is the syllable (Krakow, 1999).

3.4.2 Three-dimensional articulatory acoustic synthesis: Paper IX

Engwall, O. ***”Synthesizing static vowels and dynamic sounds using a 3D vocal tract model”.*** In Proceedings of 4th ISCA workshop on Speech synthesis.

The acoustic synthesis has been a secondary goal in the project and no effort has yet been put into modeling of tissue absorption, radiation loss, lateral asymmetries or glottal source, that would improve the acoustic synthesis and make use of the inherent advantages of the three-dimensional model. Paper IX describes an implementation of the 3D acoustic synthesis based on an area function to formants algorithm. The interest is not mainly in the sound generation as such, but in the method to determine the area function and in the evaluation of the model's ability to replicate the reference subject's formants.

The novelty in the version of the model that Paper IX describes is that vocal tract walls, modeled on MR Images of the reference subject, have been introduced together with parameters to control the terminations, i.e. larynx height, lip rounding and lip protrusion. These additions complete the vocal tract part of the model, which means that calculations of the tube geometry can be made and the sound output determined.

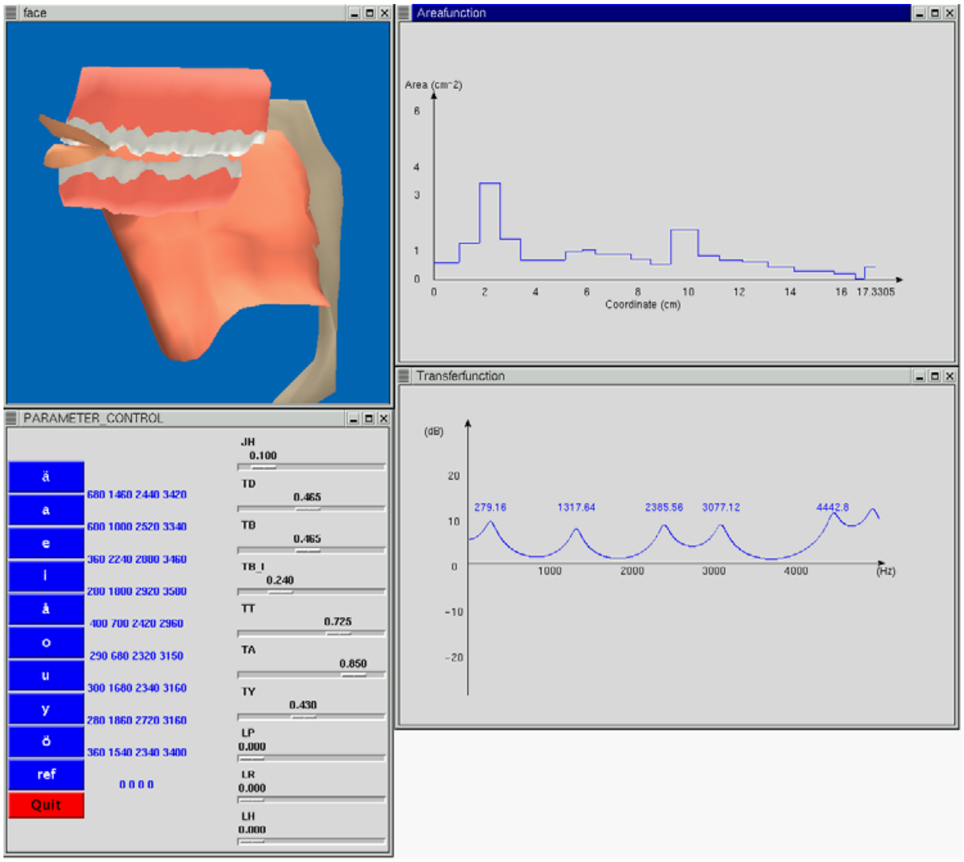
The wall model is generated in much the same way as the tongue, by contour extraction in each image and reconstruction in a semi-polar grid (Beautemps *et al.* , 2001). The use of the semi-polar grid in the reconstruction is of importance, as this allows the following calculation of the vocal tract tube cross-sections to be simplified. The area function is determined through sampling at the gridplanes of the semi-polar grid, where the vocal tract contours were extracted. The area of each cross-section is calculated with a polygon area summation formula.

The area function is affected by the model's parameters and by changing the parameters manually, using the display shown in Fig. 3.13, the model can be configured to replicate measured formants of the reference subject on which the model is based. The tuning, made with slide bars (lower left vindow), aimed at matching the first two formant frequencies in the bottom right window to the measured values, while keeping the higher frequencies in an acceptable range from the targets. The area function (upper right window) and the position of articulators (upper left window) were used to guide the parameter changes based on knowledge of standard area functions and the highest tongue point.

The sound generation was evaluated from three different viewpoints: i) the absolute difference in formants between the model and the target, ii) put into relation to results reported for other models and iii) using a perception test. The perception test confirmed the conclusions of the numeric evaluations, but it also suggested some differences between experienced and unexperienced listeners regarding synthetic speech. The differences are described in Paper IX, and one aspect is summarized in Table 3.1, where the differences between the responses from the two groups are presented.

The differences are generally small, except for /iː, yː, u̟ː, øː/. The reasons for this are discussed in Paper IX; it is for instance plausible that the naive listeners' hesitation concerning the classification of the model's /iː/ affected the result for /yː/, whereas the experienced listeners were more confident in their classification. The expert users knew what the /iː, yː/ were supposed to sound like and rejected the model's /iː/-stimuli that differed too much from the target and accepted the model's /yː/-stimuli, whereas the naive listeners became more confused by the decreased difference between the two stimuli. The other point to note is the difference in the role of /u̟ː/, which seems to be the preferred choice for the naive listeners whenever they hesitated within the front rounded vowel group.

Naive users are, in conclusion, more vulnerable to deviations from the standard synthetic speech, and this is an issue that should be considered if the model should be used for



680 1460 2440 3420
600 1000 2520 3340
360 2240 2800 3460
280 1000 2920 3500
400 700 2420 2960
290 680 2320 3150
300 1680 2340 3160
280 1860 2720 3160
360 1540 2340 3400
0 0 0 0

JH 0.100
TD 0.465
TB 0.465
TB_I 0.240
TT 0.725
TA 0.850
TY 0.430
LP 0.000
LR 0.000
LH 0.000

Figure 3.13. The graphical interface for the interactive articulatory synthesis, consisting of one window for the vocal tract model, one for the parameter and phoneme control, one for the area function and one for the transfer function.

pronunciation training. The synthesis of the current technology is not of high enough quality for language tutoring and the above results suggest that this is all the more true for naive users. The synthesis presented in this section should hence not be considered as an end product, but as an evaluation of the model's correctness vis-a-vis the reference subject.

Table 3.1. *The difference between naive and expert listeners in the confusion matrix for the model's vowels. Stimuli horizontally and response vertically.*

	iː	eː	æː	ɑː	ɔː	uː	ʉː	yː	øː	*Total*
iː	**6**							3	1	*10*
eː	1	**-2**								*-1*
æː			**-3**							*-3*
ɑː		1		**-1**						*0*
ɔː				1	**0**					*1*
uː						**0**	1			*1*
ʉː							**-2**	8	4	*10*
yː	-7							**-11**		*-18*
øː		1	3				1		**-5**	*0*

3.4.3 Three-dimensional multimodal speech instruction

The issues concerning the model's use for visual speech training has not yet been investigated within the current project, but they merit consideration at least briefly.

The question behind all issues is what mode of presentation that should be used to maximize the information that the user is able to retrieve from the model. The first issue concerns which structures that should be made visible and how they should be presented. It is quite clear that showing almost all structures is not optimal, as large parts of the tongue are then occluded by other surfaces. It is however neither the best solution to present the tongue alone, as the lack of reference frame will make the tongue movements harder to interpret. The best presentation mode probably lies somewhere between these two extremes and it is quite possible that the best solution depends on the training task at hand and the user. One strength of the implementation of the visual speech synthesis environment (Beskow, 1995) is the flexibility in the mode of presentation. Not only can the model be viewed from all different angles and at different degrees of magnifications; the visibility and the surface properties of each surface can easily be changed by the user, making quick adjustments to the preferred presentation mode possible. This means that no "best mode of presentation" need to be fixed, but that the display can be adapted to the subject and the training session.

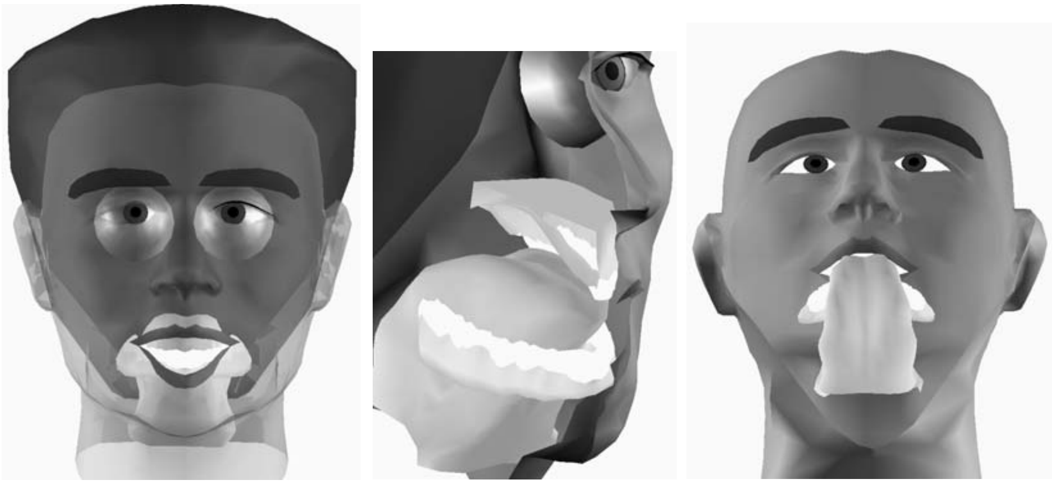
Two standard set-ups can nevertheless be envisaged as being beneficial for different tasks. The tongue together with the jaw and the palate would provide the basic information on the intra-oral articulation. The jaw is needed as a reference frame to illustrate the jaw movement whereas the palate indicates both the linguopalatal contact and distance, which are two of the most important articulatory features. One general presentation issue however applies to the palate (as for the face in cases where it is visible) and that is how it should be rendered to maximize the benefit for the user, as a compromise between naturalness of the palatal surface and visibility of the tongue has to be made. The three alternative modes of presentation for the palate is semi-transparent, in wireframe or in half, all three allowing the tongue to been seen either through the palate or by removing the occluding part.

The second set-up is to add the face to the above setting to complement the information with the visemes, but this requires a good strategy of making the facial surface semi-transparent alternatively to remove the hair and present the tongue from behind. The question is from which viewpoint the user perceives his or her own tongue movements and hence which view of the tongue model is the most natural for the user: straight face-to-face, which is the most common in human-human interaction and thus the most habitual view of the face, from the side, which is the view that makes the front- and backward movement of the tongue clearest and is the view that can be compared directly to midsagittal images from X-ray or MRI, or finally from behind, as if seen from inside the synthetic head.

Examples of these three set-ups are shown in Fig. 3.14 and animations of the different views can be viewed from the project home page

http://www.speech.kth.se/multimodal/vocaltract.html.

The issue of how to present the model needs to be addressed in user studies as for the three issues that were brought up in Cohen *et al.* (1998). The first of these concerns the modality of the output, whether it should be visible speech only or multimodal, including the sound output from the model as an auditory feedback. Cohen *et al.* (1998, p. 202) "expect that the child could learn multimodal targets, which would provide more resolution than either modality alone." and this is the standpoint in the KTH project as well. The multimodality of the model is one of its key features and rather than merely presenting the acoustic or articulatory targets, the articulatory changes in the model should be reflected directly in the sound output through calculation of the transfer function from the model's



(a) Front view (b) Half face (c) From behind

Figure 3.14. Three different views of the intraoral parts of the KTH synthetic head that could potentially be used in pronunciation training, a) with a see-through face, b) with one half of the face removed and c) from behind, with the hair removed.

geometry (section 3.4.2). By this means the user can test directly how articulatory changes alter the acoustic output and this is beneficial not the least in second language learning where subtle pronunciation differences should be illustrated.

The second issue that Cohen *et al.* envisage is whether the articulation training should be made with static or dynamic presentations. They suppose before testing that "a combination of modes would be optimal" (p. 202) and this appears to be a very sound supposition. On the one hand the movement of the tongue between different targets is nearly as important as the targets themselves and much information is lost in the static representation. On the other hand, real-time tongue movements are so fast that it would be difficult for the user to grasp all features of the articulations included in the movement. The solution is rather to show both static and dynamic representations *and* to allow the movement to be slowed down to a very slow speaking rate, so that the articulatory transitions can be viewed clearly.

The last issue is the length of the instructional target; whether the user should be shown single syllable targets or larger units of words and phrases. Once again, the suggestion of Cohen *et al.*, that "several sizes of targets would be ideal" (p. 202) is shared within the current project. For the basic pronunciation training it can be assumed that it is beneficial to present the details of a difficult syllable in isolation, but to illustrate prosodic articulatory features as well as to maintain the user's interest for the training, words and phrases should be an important part of the instruction.

All the three issues are taken care of in Wavesurfer (Sjölander & Beskow, 2000) with the text-to-intra-oral visual speech synthesis plugin. It allows for large flexibility and freedom for the user, who can type in any word or phrase and then repeat it, or any part of it, as many times as wanted. The articulations can also be slowed down arbitrarily and the parameter activations altered to study the interaction of different parts of the tongue. It should be noted, however, that these changes are not yet reflected in the acoustic synthesis which remains the same.

3.5 Conclusions and perspectives

The present model is not yet at the goal of 3D multimodal articulatory synthesis, producing sound output for all types of phonemes in synchrony with the visual display. More work is needed before the model can complement human teachers as a language tutor that can show the intra-oral articulations and produce sound.

One remaining question to be solved is the quality of current speech synthesis, which is probably not good enough for pronunciation training, especially as naive users of speech synthesis were shown to be more vulnerable to deviations than expert users (cf. section 3.4.2).

In the language tutor application it would also be immensely beneficial to be able to do sound-to-gesture inversion, to contrast the student's and the tutor's pronunciation. This is a very difficult problem, as there is no unique one to one mapping, but increasing effort is dedicated to the matter, with studies such as Maeda (1994), Mathie & Laprie (1997), Moody (1999) and Dang & Honda (2000a).

The problem could also be simplified using a web cam, providing the system with some information of the student's jaw height, lip protrusion and, in lucky cases, larynx position. Vatikiotis-Bateson & Yehia (1997) showed that 77% of the variance in the vocal tract could be extracted from orofacial motion measured with an Optotrack system, where 12 markers were placed on the face. It should however be noted that the vocal tract variance that Vatikiotis-Bateson & Yehia measured was the movement of EMA coils on the jaw, upper and lower lip and four points on the tongue, all placed in the midsagittal plane, and the information on the entire vocal tract shape is hence more restricted than it first seems. Moreover, without the markers and with the poorer image resolution of a web cam it can not be hoped that image analysis can contribute as much in this case, but at least some ambiguities could be resolved, as the extra-oral information could diminish the possible articulatory combinations and hence facilitate the inversion. The model would however have to be adapted to different speakers, and one of the paths for future work is to determine simple parametric changes to scale the model from the reference subject to other speakers (e.g. warping such as thin-spline mapping, cf. Wu & Wilhelms-Tricarico, 1995). Another is to adapt the model to other languages, to make it possible to use it as a multilingual tutor.

The study in section 3.4.1 used a database of British English to control the tongue defined on Swedish articulations, without any explicit adaptation to English. The evaluation suggested that this multilinguality generally worked well, when comparing with the original articulations, but it did not focus specifically on possible errors due to articulatory differences between the two languages and to some of the articulations in the MOCHA-TIMIT database not being measured in the MRI acquisition. It was noted in the evaluation that the tongue model did not replicated the very important backward movement in "bel*ow* *ou*tside" and it is quite possible that this can be explained by the fact that these open back vowels were not part of the Swedish corpus. Future evaluation and development of the concatenative synthesis should look into this issue to define which articulatory changes are needed in the model when going from one language to another.

Work remains however even for the current language and the current reference subject. The 3D data in the model is quite exhaustive, but more real-time data might be needed, both for tuning of articulatory positions and transitions. The EPG data provided a rather good tuning for articulations with linguopalatal contact, but the same type of tuning is required for articulations without contact. The EPG tuning further suggested that the tongue parameters need some additional tuning for articulations with low tongue body and raised tongue tip, and for lateral asymmetries. Optopalatography is well suited for the lateral tuning, as it provides real-time lateral information on both linguopalatal contact and distance. The information on the distance from the sensor to the closest tongue point can be mapped on the current model, letting it interpolate between the measured flesh points, and then tune the vertex weights to get a smoothly varying tongue shape passing through the measured points.

Concerning articulatory transitions, there is a good covering of fricatives in VCV context, but speech of course contains many other movements and ideally a much more varied corpus should be collected to provide an adequate description of the articulatory behavior of the reference subject, in analogy with the MOCHA-TIMIT database presented in section 3.4.1. Assessment of the kinematics in the model should be made against continuous real-time data, such as X-rays or dynamical midsagittal MRI, as tested in section 3.4.1, but for Swedish and for a larger test corpus with the reference subject.

In the somewhat longer run, the ongoing technical advances leave hope for real-time 3D MRI in the years to come. This would allow for full three-dimensional real-time measurements, hence eliminating the need for different measurement sources and the fear that the 3D measurements are non-representative due to the artificial sustaining of the articulations. Another source of measurement error, the supine position, can also be eliminated with the development of upright MRI scanners, such as Fonar's Stand-Up MRI, released in 2000. The upright MRI scanners will allow for a much more natural speech situation, making it ideal for speech production research, if the acquisition time allows for running speech, so that the articulations need not be sustained without any movement for prolonged periods, which could be troublesome in upright position.

The technical development also gives increasing computational power, making real-time physiological modeling conceivable, which would allow for better impenetrability modeling, explicit volume conservation and more realistic acoustic synthesis, with asymmetries, radiation and non-homogeneous walls taken into account.

There is hence good hope for better articulatory measurements and models in the near future and 3D multimodal articulatory synthesis has potentials as a method for quality speech synthesis. The work presented in this thesis has taken several steps towards that goal, even if this is only the beginning of the path.

It is with a note of great expectations and confidence on behalf of 3D vocal tract modeling that I close this thesis, as the number of research teams working in the area is increasing. Apart from studies already mentioned within the thesis, several related projects are currently starting up, such as

> the Vocal Tract Modeling/Articulatory Synthesis Project at University of British Columbia, Vancouver, Canada
>
> realistic three-dimensional tongue models for training of medical students at Northern Ireland Technology Centre, Queen's University, Belfast, Great Britain
>
> and three-dimensional tongue models in a biofeedback loop for speech rehabilitation, responding in real-time to a subject's tongue movement at the School of Health & Rehabilitation Sciences, University of Queensland, Australia.

But that future story of intra-oral 3D modeling has to be told another time,
"break my heart, for I must hold my tongue."
W. Shakespeare, *Hamlet*, Act I, scene ii.

List of publication

The following papers have been published as a part of this thesis work. Articles marked with ⋆ are included in the thesis, those marked with • are not.

⋆ Engwall, O. (1999).
Vocal tract modeling in 3D.
In TMH-QPSR1-2/1999, 31-38.

• Engwall O (1999b).
Modeling of the vocal tract in three dimensions
In Proceedings of Eurospeech 1999, 113-116.

⋆ Engwall, O. & Badin, P. (1999).
Collecting and analysing two- and three-dimensional
MRI data for Swedish.
In TMH-QPSR3-4/1999, 11-38.

⋆ Engwall, O. & Badin, P. (2000).
An MRI study of Swedish fricatives: coarticulatory effects.
In Proceedings of the 5th Speech Production Seminar, 297-300.

⋆ Engwall, O. (2000a).
Dynamical aspects of coarticulation in Swedish fricatives
– a combined EMA & EPG study.
In TMH-QPSR 4/2000, 49-73.

⋆ Engwall, O. (2000b).
Are static MRI data representative of dynamic speech?
Results from a comparative study using MRI, EMA and EPG.
In Proceedings of the 6th ICSLP, I:17-20.

• Engwall, O. (2000c).
Replicating three-dimensional tongue shapes synthetically.
In TMH-QPSR 2-3 2000, 53-64.

• Engwall O (2000d).
A 3D tongue model based on MRI data.
In Proceedings of the 6th ICSLP, III: 901-904.

• Engwall, O. (2001a).
Using linguopalatal contact patterns to tune a 3D tongue model
In Proceedings of Eurospeech2001, volume 2, 1475-1478.

• Engwall, O. (2001b).
Making the Tongue Model Talk: Merging MRI & EMA Measurements
In Proceedings of Eurospeech2001, volume 1, 261-264.

• Engwall, O. (2001c).
Considerations in Intraoral Visual Speech Synthesis: Data and Modeling.
In Proceedings of the 4th International Speech Motor Conference, 23-26.

⋆ Engwall, O. (2001d).
Synthesizing static vowels and dynamic sounds using a 3D vocal tract model.
In Proceedings of the 4th ISCA workshop on Speech synthesis, 81-86.

⋆ Engwall, O. (submitted1).
Combining MRI, EMA & EPG measurements
in a three-dimensional tongue model.
Submitted to Speech Communication.

⋆ Engwall, O. (submitted2).
Concatenative Articulatory Synthesis
Submitted to Journal of Phonetics.

⋆ Engwall, O. (submitted3).
Evaluation of a System for Concatenative Articulatory Visual Synthesis
Submitted to the 7^{th} *ICSLP, Denver, Colorado, September 16-20, 2002.*

Bibliography

Alwan, A., Narayanan, S., & Haker, K. (1997). Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part II. The rhotics. *Journal of the Acoustic Society of America*, **1**01, 1078–1089.

Apostol, L., Perrier, P., Raybaudi, M., & Segebarth, C. (1999). 3D geometry of the vocal tract and inter-speaker variability. *Pages 443–446 of: Proceedings of the XIVth ICPhS*, vol. 1.

Badin, P., Baricchi, E., & Vilain, A. (1997). Determining tongue articulation: from discrete fleshpoints to continuous shadow. *Pages 47–50 of: Proceedings of Eurospeech '97*, vol. 1.

Badin, P., Bailly, G., Raybaudi, M., & Segebarth, C. (1998). A three-dimensional linear articulatory model based on MRI data. *Pages 249–254 of: Proceedings of the Third ESCA/COCOSDA International Workshop on Speech Synthesis.*

Badin, P., Bailly, G., M. Raybaudi, M., & Segebarth, C. (2000a). A three-dimensional linear articulatory model based on MRI data. *Pages 901–904 of: Proceedings of the* 6^{th} *ICSLP*, vol. III.

Badin, P., Borel, P., Bailly, G., Revret, L., Baciu, M., & Segebarth, C. (2000b). Towards an audiovisual virtual talking head: 3D articulatory modeling of tongue, lips and face based on MRI and video images. *Pages 261–264 of: Proceedings of the* 5^{th} *Seminar on Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling.*

Baer, T., Gore, J.C., Boyce, S., & Nye, P.W. (1987). Application of MRI to the Analysis of Speech Production. *Magnetic Resonance Imaging*, **5**, 1–7.

Baer, T., Gore, J.C., Gracco, L.W., & Nye, P.W. (1991). Analysis of vocal tract shape and dimensions using Magnetic Resonance Imaging: Vowels. *Journal of the Acoustic Society of America*, **9**0, 799–828.

Bailly, G., Laboissiere, R., & Schwartz, J. L. (1991). Formant trajectories as audible gestures: an alternative for speech synthesis. *Journal of Phonetics*, **19**, 9–23.

Bangayan, P., Alwan, A., & Narayanan, S. (1996). From MRI and Acoustic Data to Articulatory Synthesis: a Case Study of the Laterals. *Pages 793–796 of: Proceedings of the* 4^{th} *ICSLP.*

Båvegård, M. (1996). *Towards an Articulatory Speech Synthesizer: Model Development and Simulations.* Licentiate Thesis, KTH, Stockholm, Sweden.

Beaudoin, R., & McGowan, R. (2000). Principal Component Analysis of X-ray microbeam data for articulatory recovery. *Pages 225–228 of: Proceedings of the 5th Seminar on Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling.*

Beautemps, D., Badin, P., & Bailly, G. (2001). Degrees of freedom in speech production: analysis of cineradio-and labio-films data for a reference subject, and articulatory-acoustic modeling. *Journal of the Acoustic Society of America*, 2165–2180.

Beskow, J. (1995). Rule-based visual speech synthesis. *Pages 299–302 of: Proceedings of Eurospeech '95.*

Boersma, P. (1998). *Functional Phonology: Formalizing the interactions between articulatory and perceptual drives.* Ph.D. thesis, University of Amsterdam, The Netherlands.

Branderud, P. (1985). Movetrack – a movement tracking system. *Pages 113–122 of: Proceedings of the French-Swedish Symposium on Speech, Grenoble.*

Campbell, N. (2000). Databases of Emotional Speech. *Pages 34–38 of: ISCA workshop on Speech & Emotion.*

Carlson, R., Granström, B., & Hunnicut, S. (1982). A multi-lanuguage text-to-speech module. *Pages 1604–1607 of: Proceedings of ICASSP-Paris*, vol. 3.

Carlson, R., Granström, B., & Nord, L. (1992). Experiments with emotive speech – acted utterances and synthesized replicas. *Pages 671– 674 of: Proceedings of the* 2th *ICSLP.*

Childers, D. G., & Ding, C. (1991). Articulatory synthesis: nasal sounds and male and female voices. *Journal of Phonetics*, **19**, 453–464.

Chuang, C-K., & Wang, W. (1978). Use of optical distance sensing to track tongue motion. *Journal of the Acoustic Society of America*, **21**, 482–496.

Cohen, M., Beskow, J., & Massaro, D. (1998). Recent development in facial animation: an inside view. *Pages 201–206 of: Proceedings of AVSP'98.*

Coker, C., & Fujimura, O. (1966). Model for the specification of the vocal tract area function. *Journal of the Acoustic Society of America*, **40**, 1271.

Dang, J., & Honda, K. (1998). Speech production of vowel sequences using a physiological articulatory model. *Pages 1767–1770 of: Proceedings of the* 5th *ICSLP*, vol. 5.

Dang, J., & Honda, K. (2000a). Estimation of vocal tract shape from speech sounds via a physiological articulatory model. *Pages 233–236 of: Proceedings of the 5th Speech Production Seminar: Models and data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling.*

Dang, J., & Honda, K. (2000b). Improvement of a physiological articulatory model for synthesis of vowel sequences. *Pages 457–460 of: Proceedings of the* 6th *ICSLP*, vol. 1.

Dart, S. (1964). Replica of the vocal tract. *Working papers in Phonetics, UCLA Phonetics Laboratory Group*, **1**.

Dart, S. (1987). A bibliography of X-ray studies of speech. *Working papers in Phonetics, UCLA Phonetics Laboratory Group*, **66**, 1–97.

Davis, E., Douglas, A., & Stone, M. (1996). A continuum mechanics representation of tongue deformation. *Pages 788–792 of: Proceedings of the* 4^{th} *ICSLP.*

Demolin, D., Metens, T., & Soquet, A. (1996). Three-dimensional measurements of the vocal tract by MRI. *Pages 272–275 of: Proceedings of the* 4^{th} *ICSLP*, vol. 1.

Demolin, D., George, M., Lecuit, V., Metens, T., A., Soquet, & Raeymaekers, H. (1997). Coarticulation and articulatory compensations studied by dynamic MRI. *Pages 43–46 of: Proceedings of Eurospeech '97.*

Demolin, D., Lecuit, V., Metens, T., Nazarian, B., & Soquet, A. (1998). Magnetic Resonance measurements of the velum port opening. *Pages 425–428 of: Proceedings of the* 5^{th} *ICSLP*, vol. 2.

Demolin, D., Metens, T., & Soquet, A. (2000). Real time MRI and articulatory coordinations in vowels. *Pages 93–96 of: Proceedings of the* 5^{th} *Speech Production Seminar: Models and data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling.*

Dixit, P. (1999). Palatometric investigation of selected coronal consonants of Hindi. *Pages 439–441 of: Proceedings of the XIVth ICPhS*, vol. 1.

Ellis, L., & Hardcastle, W. (2000). Assimilation strategies in the production of alveolar to velar sequences: EPG and EMA data. *Pages 117–120 of: Proceedings of the* 5^{th} *Speech Production Seminar: Models and data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling.*

Engstrand, O. (1989). Towards an electropalatographic specification of consonant articulation in Swedish. *Perilus*, **X**, 115–156.

Engwall, O. (1999). Modeling of the vocal tract in three dimensions. *Pages 113–116 of: Proceedings of Eurospeech '99.*

Engwall, O., & Badin, P. (1999). Collecting and analysing two- and three-dimensional MRI data for Swedish. *TMH-QPSR*, **3**-4, 11–38.

Fant, G. (1960). *Acoustic theory of speech production.* Mouton, The Hague.

Fant, G. (1964). Formants and cavities. *Pages 120–140 of: Proceedings of the Vth ICPhS.*

Fant, G. (1983). Feature analysis of Swedish vowels - a revisit. *STL-QPSR*, **2**-3, 1–19.

Fant, G. (1992). Vocal tract area functions of swedish vowels and a new three parameter model. *Pages 807–810 of: Proceedings of the* 2^{nd} *ICSLP.*

Fels, S. (1994). *Glove-TalkII: Mapping Hand Gestures to Speech Using Neural Networks - An Approach to Building Adaptive Interfaces.* Ph.D. thesis, University of Toronto, Canada.

Fitzpatrick, L., & Ní Chasaide, A. (1999). Human speaker nomograms using EMA data. *Pages 2021–2024 of: Proceedings of the XIVth ICPhS*, vol. 3.

Fitzpatrick, L., & Ní Chasaide, A. (2000). Inferring tongue articulation from simultaneous EMA & EPG: Complementary and conflicting data. *Pages 205–208 of: Proceedings of the* 5^{th} *Speech Production Seminar: Models and data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling.*

Foldvik, A-K., Husby, O., & Kvaerness, J. (1988). Magnetic resonance imaging. *Pages 423–428 of: Proceedings of Speech'88.*

Foldvik, A-K., Kristiansen, U., & Kvaerness, J. (1993). A time-evolving three-dimensional vocal tract model by means of magnetic resonance imaging (MRI). *Pages 557–558 of: Proceedings of Eurospeech '93.*

Foldvik, A-K., Kristiansen, U., & Kvaerness, J. (1995). Three-dimensional ultrasound and magnetic resonance imaging: A new dimension in phonetic research. *Pages 46–49 of: Proceedings of the XIIIth ICPhS*, vol. 4.

Fougeron, C., & Keating, P. (1996). The influence of prosodic position on velic and lingual articulation in French: evidence from EPG and airflow data. *Pages 93–96 of: Proceedings of the 4^{th} Speech Production Seminar.*

Fougeron, C., Meynadier, Y., & Demolin, D. (2000). 62 vs. 96 Electrodes: A comparative analysis of Reading and Kay Elemetrics EPG pseudo-palates. *Pages 309–312 of: Proceedings of the 5^{th} Speech Production Seminar: Models and data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling.*

Fujimura, O. (1991). Recording and interpreting articulatory data – microbeam and other methods. *Pages 120–124 of: Proceedings of the XIIth ICPhS*, vol. 3.

Gauffin, J., & Sundberg, J. (1978). Pharyngeal constrictions. *Phonetica*, **3**5, 157–168.

Gibbon, F., & Hardcastle, W. (1994). Articulatory description of affricate production in speech-disordered children using electropalatography (EPG). *Pages 1191–1194 of: Proceedings of the* 3^{rd} *ICSLP*, vol. 3.

Gibbon, F., & Wood, S. (2001). Undifferentiated gestures and articulatory drift in the speech of children with articulation/phonological disorders. *Pages 22–56 of: Proceedings of the 4^{th} International speech motor conference: speech motor control in normal and disordered speech.*

Gick, B. (2002). An X-ray Investigation of Pharyngeal Constriction in American English Schwa. *Phonetica*, **5**9, 38–48.

Gick, B., Kang, M., & Whalen, D. (2000). MRI and X-ray evidence for commonality in the dorsal articulation of English vowels and liquids. *Pages 69–72 of: Proceedings of the 5^{th} Speech Production Seminar: Models and data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling.*

Hardcastle, W., Gibbon, F., & Nicolaidis, K. (1991). EPG data reduction methods and their implications for studies of lingual coarticulation. *Journal of Phonetics*, **1**9, 251–266.

Hardcastle, W., Vaxelaire, B., Gibbon, F., Hoole, P., & Nguyen, N. (1996). EMA/EPG study of lingual coarticulation in /kl/ clusters. *Pages 53–56 of: Proceedings of the 4^{th} Speech Production Seminar.*

Harshman, R. A., Ladefoged, P., & Goldstein, L. (1977). Factor analysis of tongue shapes. *Journal of the Acoustic Society of America*, **6**2, 693–707.

Hartelius, L., Mcaulifee, M., Murdoch, B., & Theodoros, D. (2001). The use of electropalatography in the treatment of disordered articulation in traumatic brain injury: a case study. *Pages 192–195 of: Proceedings of the 4^{th} International speech motor conference: speech motor control in normal and disordered speech.*

Henke, W. L. (1966). *Dynamic Articulatory Model of Speech Production Using Computer Simulation.* Ph.D. thesis, MIT, Cambridge, MA.

Honda, K., & Tiede, M. (1998). An MRI study on the relationship between oral cavity shape and larynx position. *Pages 437–440 of: Proceedings of the* 5^{th} *ICSLP*, vol. 2.

Honda, K., Hirai, H., & Dang, J. (1994). A physiological model of speech production and the implication of tongue larynx interaction. *Pages 175–178 of: Proceedings of the* 3^{rd} *ICSLP.*

Hoole, P., Wismueller, A., Leinsinger, G., Kroos, C., Geumann, A., & Minoue, M. (2000). Analysis of the tongue configuration in multi-speaker, multi-volume MRI data. *Pages 157–160 of: Proceedings of the* 5^{th} *Speech Production Seminar: Models and data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling.*

Jackson, P., & Shadle, C. (2000). Aero-acoustic modelling of voiced and unvoiced fricatives based on MRI data. *Pages 185–188 of: Proceedings of the* 5^{th} *Speech Production Seminar: Models and data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling.*

Kaburagi, T., & Honda, M. (1996). A model of articulator trajectory formation based on the motor tasks of vocal-tract shapes. *Journal of the Acoustic Society of America*, **99**, 3154–3170.

Keller, E. (1987). Ultrasound measurements of tongue dorsum movements in articulatory speech impairments. *Pages 93–112 of:* Ryalls, J.H. (ed), *Phonetic Approaches to Speech Production in Aphasia and Related Disorders.* College-Hill Press, San Diego, CA.

Keller, E., & Ostry, D. (1983). Computerized measurement of tongue dorsum movements with pulsed echo ultrasound. *Journal of the Acoustic Society of America*, **73**, 1309–1315.

Kelsey, C., Minifie, F., & Hixon, T. (1969). Applications of ultrasound in speech research. *Journal of Speech and Hearing Research*, **12**, 564–575.

Kiritani, S., Itoh, K., & Fujimura, O. (1975). Tongue-pellet tracking by a computer controlled X-ray microbeam system. *Journal of the Acoustic Society of America*, **48**, 1516–1520.

Kiritani, S., Miyawaki, K., & Fujimura, O. (1976). A computational model of the tongue. *Res. Instit. Logoped. Phoniatr. Annual Bulletin, University of Tokyo*, **10**, 243–252.

Krakow, R. (1999). Physiological organization of syllables, a review. *Journal of Phonetics*, **27**, 23–54.

Kröger, B., Winkler, R., Mooshammer, C., & Pompino-Marschall, B. (2000). Estimation of vocal tract area function from magnetic resonance imaging: preliminary results. *Pages 333–336 of: Proceedings of the* 5^{th} *Speech Production Seminar: Models and data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling.*

Ladefoged, P., Anthony, J., & Riley, C. (1971). Direct measurement of the vocal tract. *Working papers in Phonetics, UCLA Phonetics Laboratory Group*, **19**.

Liljencrants, J. (1971). Fourier series description of the tongue profile. *STL-QPSR*, **4**, 9–18.

Lin, Q. (1990). *Speech Production Theory and Articulatory Speech Synthesis.* Ph.D. thesis, KTH, Stockholm, Sweden.

Lindblad, P., & Lundqvist, S. (1995). The groove production of Swedish sibilants – an EPG analysis. *Pages 458–461 of: Proceedings of the XIIIth ICPhS*, vol. 2.

Lindblad, P., & Lundqvist, S. (1999). How and why do the tongue gestures of [t], [d], [l], [n], [s] and [r] differ? *Pages 417–420 of: Proceedings of the XIVth ICPhS*, vol. 1.

Lindblom, B., & Sundberg, J. (1971). Acoustical consequences of lip, tongue and jaw movements. *Journal of the Acoustic Society of America*, **50**, 1166–1179.

Lobo, A., & Malley, M. (1996). Towards a biomechanical model of the larynx. *Pages 279–282 of: Proceedings of the* 4^{th} *ICSLP.*

Lundberg, A., & Stone, M. (1999). Three-dimensional tongue surface reconstruction: Practical consideration for ultrasound data. *Journal of the Acoustic Society of America*, **106**, 2858–2867.

Madea, S. (1988). Improved articulatory models. *Journal of the Acoustic Society of America*, **8**4, S146.

Mády, K., Sader, R., Zimmermann, A., Hoole, P., Beer, A., Zeilhofer, H.F., & Hanning, C. (2001). Use of real-time MRI in assessment of consonant articulation before and after tongue surgery and tongue reconstruction. *Pages 142–145 of: Proceedings of the* 4^{th} *International speech motor conference: speech motor control in normal and disordered speech.*

Maeda, S. (ed). (1994). *SpeechMaps, WP2 – From speech signal to vocal tract geometry.* Vol. III.

Mair, S., Scully, C., & Shadle, C. (1996). Distinction between [t] and [tS] using electropalatography data. *Pages 1597–1600 of: Proceedings of the* 4^{th} *ICSLP.*

Marchal, A., & Hardcastle, W. (1993). ACCOR: Instrumentation and database for the cross-language study of coarticulation. *Language and Speech*, **36**, 137–153.

Mathiak, K., Klose, U., Ackerman, H., Hertrich, I., Kincses, W-E., & Grod, W. (2000). Stroboscopic articulography using fast magnetic resonance imaging. *Pages 97–100 of: Proceedings of the* 5^{th} *Speech Production Seminar: Models and data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling.*

Mathie, B., & Laprie, Y. (1997). Adaptation of Maeda's model for acoustic to articulatory inversion. *Pages 2015–2018 of: Proceedings of Eurospeech '97.*

Matsuda, M., & Kasuya, H. (1999). Acoustic nature of the whisper. *Pages 133–136 of: Proceedings of Eurospeech '99.*

Matsumura, M., Niikawa, T., Shimizu, K., Hashimoto, Y., & Morita, T. (1994). Measurement of 3D shapes of Vocal Tract, Dental Crown and Nasal Cavity using MRI: Vowels and Fricatives. *Pages 619–622 of: Proceedings of the* 3^{rd} *ICSLP*, vol. 2.

Matsumura, M., Niikawa, T., Torii, T., Yamasaki, H., , Hara, H., Tachimura, & Wada, T. (2000). Measurement of palatolingual contact pressure and tongue force using a force-sensor-mounted palatal plate. *Pages 893–896 of: Proceedings of the* 6^{th} *ICSLP*, vol. 3.

Matsuzaki, H., & Motoki, K. (2000). FEM analysis of 3-D vocal tract shape model with asymmetrical shape. *Pages 329–332 of: Proceedings of the* 5^{th} *Speech Production Seminar: Models and data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling.*

Matsuzaki, H., Miki, N., Nagai, N, Hirohku, T., & Ogawa, Y. (1994). 3D FEM analysis of vocal tract model of elliptic tube with inhomogenous-wall impedance. *Pages 635–638 of: Proceedings of the* 3^{rd} *ICSLP.*

Mermelstein, P. (1973). Articulatory model for the study of speech production. *Journal of the Acoustic Society of America*, **53**, 1070–1082.

Meyer, P., Wilhelms, R., & Strube, H. W. (1989). A quasiarticulatory speech synthesizer for German language running in real time. *Journal of the Acoustic Society of America*, **86**, 523–538.

Minifie, F., Kelsey, C., & Zagzebski, J. (1971). Ultrasonic Scans of the Dorsal Surface of the Tongue. *Journal of the Acoustic Society of America*, **49**, 1857–1860.

Mohammad, M., Moore, E., Carter, J., Shadle, C., & Gunn, S. (1997). Using MRI to image the moving vocal tract during speech. *Pages 2027–2030 of: Proceedings of Eurospeech '97.*

Moody, J. (1999). *Visualizing Speech with a Recurrent Neural Network Trained on Human Acoustic-Articulatory Data.* Ph.D. thesis, University of California, San Diego.

Motoki, K., Badin, P., Pelorson, X., & Matsuzaki, H. (2000). A modal parametric method for computing acoustic characteristics of three-dimensional vocal tract models. *Pages 325–328 of: Proceedings of the* 5^{th} *Speech Production Seminar: Models and data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling.*

Müller, E., & McLeod, G. (1982). Perioral biomechanics and its relation to labial motor control. *Journal of the Acoustic Society of America*, **71**, Suppl. 1, P8.

Munhall, K., Vatikiotis-Bateson, E., & Tokhura, Y. (1995). X-ray Film Database for Speech Research. *Journal of the Acoustical Society of America*, **98**, 1222–1224.

Murray, I.R., & Arnott, J.L. (1996). Synthesizing emotions in speech: is it time to get excited? *Pages 1816–1819 of: Proceedings of the* 4^{th} *ICSLP.*

Narayanan, S., Alwan, A., & Haker, K. (1995). An articulatory study of fricative consonants using Magnetic Resonance Imaging. *Journal of the Acoustic Society of America*, **98**, 1325–1347.

Narayanan, S., Kaun, A., Byra, D., Ladefoged, P., & Alwan, A. (1996). Liquids in Tamil. *Pages 797–800 of: Proceedings of the* 4^{th} *ICSLP.*

Narayanan, S., Alwan, A., & Haker, K. (1997). Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part I. The Laterals. *Journal of the Acoustic Society of America*, **101**, 1064–1077.

Nguyen, N., Wrench, A., Gibbon, F., & Hardcastle, W. (1998). Articulatory, acoustic and perceptual aspects of fricative/stop coarticualtion. *Pages 2371–2374 of: Proceedings of the* 5^{th} *ICSLP.*

Nguyen-Trong, N., Hoole, P., & Marchal, A. (1991). Articulatory-acoustic correlation in the production of fricatives. *Pages 1:18–21 of: Proceedings of the XIIth ICPhS.*

Nicolaidis, K., Waters, D., Hardcastle, W., & Gibbon, F. (1995). Variability of lingual stops in English – an EPG study. *Pages 456–458 of: Proceedings of the XIIIth ICPhS*, vol. 3.

Niikawa, T., Matsumura, M., Tachimura, T., & Wada, T. (2000). Modeling of a speech production system based on MRI measurement of three-dimensional vocal tract shapes during fricative consonant phonation. *Pages 174–177 of: Proceedings of the* 6^{th} *ICSLP*, vol. 2.

Nishikawa, K., Asama, K., Hayashi, K., Takanobu, H., & Takanishi, A. (2000). Development of a Talking Robot. *Pages 345–348 of: Proceedings of* 5^{th} *Seminar of Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling.*

Parush, A., Ostry, D., & Munhall, K. (1983). A kinematic study of lingual coarticulation in VCV sequences. *Journal of the Acoustic Society of America*, **7**4, 1115–1125.

Payan, Y., & Perrier, P. (1997a). Synthesis of V-V sequences with a 2D biomechanical tongue model controlled by the Equilibrium Point Hypothesis. *Speech Communiaction*, **2**2, 185–205.

Payan, Y., & Perrier, P. (1997b). Why should speech control studies based on kinematics be considered with caution? Insights from a 2D biomechanical model of the tongue. *Pages 2019–2022 of: Proceedings of Eurospeech '97*, vol. 4.

Payan, Y., Perrier, P., & Laboissire, R. (1995). Simulation of tongue shape variations in the sagittal plane based on a control by the Equilibrium-Point hypothesis. *Pages 474–477 of: Proceedings of the XIIIth ICPhS*, vol. 2.

Perkell, J. (1974). *A physiologically-oriented model of tongue activity in speech production.* Ph.D. thesis, MIT, Cambridge, MA.

Perkell, J., Cohen, M., Svirsky, M., Matthies, M., Garabieta, I., & Jackson, M. (1992). Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *Journal of the Acoustic Society of America*, **9**2, 3078–3096.

Perrier, P., Perkell, J., Payan, Y., Zandipour, M., Guenther, F., & Khalighi, A. (2000). Degrees of freedom of tongue movement in speech may be constrained by biomechanics. *Pages 162–165 of: Proceedings of the* 6^{th} *ICSLP.*

Rokkaku, M., Hashimoto, K., Imaizumi, S., Nimi, S., & Kirtani, S. (1986). Measurements of the Three-Dimensional Shape of the Vocal Tract Based on the Magnetic Resonance Imaging Technique. *Annual Bulletin of Research Institute of Logopedics and Phoniatrics*, **2**0, 47–54.

Rubin, P., Baer, T., & Mermelstein, P. (1981). An articulatory synthesizer for perceptual research. *Journal of the Acoustic Society of America*, **7**0, 321–328.

Rubin, P., Saltzman, E., Goldstein, L., McGowan, R., Tiede, M., & Browman, C. (1996). CASY and extensions to the task-dynamic model. *Pages 125–128 of: Proceedings of the* 1^{st} *ESCA Tutorial and Research Workshop on Speech Production Modeling –* 4^{th} *Speech Production Seminar.*

Sams, M., Kulju, J., Möttönen, R., Jussila, V., Olivés, J-L., Zhang, Y., Kaski, K., Majaranta, P., & Rih, K-J. (2000). Towards a High-Quality and Well-Controlled Finnish Audio-Visual Speech Synthesizer. *In: Proceedings of* 4^{th} *World Multiconference on Systemics, Cybernetics and Informatics and* 6^{th} *International Conference on Information Systems Analysis and Synthesis*, vol. 6.

Schönle, P.W., Gräbe, K., Wenig, P., Höhne, J, Schrader, J., & Conrad, B. (1987). Electromagnetic articulography: use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. *Brain and Language*, **3**1, 26–35.

Schröder, M. (2001). Emotional Speech Synthesis: A Review. *Pages 561–565 of: Proceedings of Eurospeech '01*, vol. 1.

Shadle, C., Mohammad, M., Jackson, P., & Carter, J. (1999). Multi-planar dynamic magnetic resonance imaging: New tools for speech research. *Pages 623–626 of: Proceedings of the XIVth ICPhS.*

Shockey, L. (1991). Electropalatography of conversational speech. *Pages 10–13 of: Proceedings of the XIIth ICPhS*, vol. 1.

Sinder, D., Richard, G., Duncan, H., Lin, Q., Flanagan, J., Krane, M., Levinson, S., Davis, D., & Slimon, S. (1996). A Fluid Flow Approach to Speech Generation. *Pages 203–206 of: Proceedings of the* 1^{st} *ESCA Tutorial and Research Workshop on Speech Production Modeling –* 4^{th} *Speech Production Seminar.*

Sjölander, K., & Beskow, J. (2000). WaveSurfer - an Open Source Speech Tool. *Pages 464–467 of: Proceedings of the* 6^{th} *ICSLP*, vol. IV.

Sondhi, M. M., & Schroeter, J. (1986). A nonlinear articulatory speech synthesizer using both time- and frequency-domain elements. *Pages 1999–2001 of: Proceedings of ICASSP-Tokyo.*

Sonies, B., Shawker, T., Hall, T., Gerber, L., & Leighton, S. (1981). Ultrasonic Visualization of Tongue Motion During Speech. *Journal of the Acoustic Society of America*, **7**0, 683–686.

Soquet, A., Lecuit, V., Metens, T., & Demolin, D. (2002). Mid-sagittal cut to area function transformations: Direct measurements of mid-sagittal distance and area with MRI. *Speech Communiaction*, **3**6, 169–180.

Stark, J., Lindblom, B., & Sundberg, J. (1996). APEX an articulatory synthesis model for experimental and computational studies of speech production. *TMH-QPSR*, **2**, 45–48.

Stark, J., Ericsdotter, C., Branderud, P., Sundberg, J., Lundberg, H.-J., & Lander, J. (1999). The apex model as a tool in the specification of speaker-specific articulatory behavior. *Pages 2279–2282 of: Proceedings of the XIVth ICPhS.*

Stone, M., & Lundberg, A. (1996). Three-dimensional tongue surface shapes of English consonants and vowels. *Journal of the Acoustic Society of America*, **9**9, 3728–3737.

Stone, M., Dick, D., Douglas, A., Davis, E., & Ozturk, C. (2000). Modelling the internal tongue using principal strains. *Pages 133–136 of: Proceedings of the* 5^{th} *Seminar on Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling.*

Story, B., Titze, I., & Hoffman, E. (1996). Vocal tract area functions from magnetic resonance imaging. *Journal of the Acoustic Society of America*, **1**00, 537–554.

Sundberg, J., Johansson, C., Wilbrand, H., & Ytterbergh, C. (1983). From Sagittal Distance to Area. *Phonetica*, **4**4, 76–90.

Tabain, M. (1998). Consistencies and inconsistencies between EPG and locus equation data on coarticulation. *Pages 1855–1857 of: Proceedings of the* 5^{th} *ICSLP.*

Takemoto, H. (2000). Morphological analys and 3D modeling of the tongue musculature in the human and chipmanzee. *Page 361 of: Proceedings of* 5^{th} *Seminar of Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling.*

Thimm, G., & Luettin, J. (1999). Extraction of articulators in X-ray image sequences. *Pages 157–160 of: Proceedings of Eurospeech '99.*

Tiede, M. (1996). An MRI-based study of pharyngeal volume contrasts in Akan and English. *Journal of Phonetics*, **2**4, 399–421.

Tiede, M., & Vatikiotis-Bateson, E. (1994). Extracting articulator movement parameters from a videodisc-based cineradiographic database. *Pages 45–48 of: Proceedings of the* 3^{rd} *ICSLP.*

Tiede, M., Yehia, H., & Vatikiotis-Bateson, E. (1996). A shape-based approach to vocal tract area function estimation. *Pages 41–44 of: Proceedings of the* 1^{st} *ESCA Tutorial and Research Workshop on Speech Production Modeling –* 4^{th} *Speech Production Seminar.*

Tom, K., Titze, I., Hoffman, E., & Story, B. (1999). 3-D Vocal Tract Imaging and Formant Structure: Varying Vocal Register, Pitch and Loudness. *Status and Progress Report, National Centre for Voice and Speech, Univeristy of Iowa*, **1**4, 101–113.

Vatikiotis-Bateson, E., & Ostry, D. (1995). An analysis of the dimensionality of jaw movement in speech. *Journal of Phonetics*, **2**3, 101–117.

Vatikiotis-Bateson, E., & Yehia, H. (1997). Unified physiological model of audible-visible speech production. *Pages 2031–2034 of: Proceedings of Eurospeech '97.*

Vilkman, E., Takalo, R., Maatta, T., Laukkanen, A-M., Nummenranta, J., & Lipponen, T. (1997). Ultrasonographic measurement of cricothyroid space in speech. *Pages 39–42 of: Proceedings of Eurospeech '97*, vol. 1.

Wakumoto, M., Masaki, S., Honda, K., & Dang, J. (1996). Visualization of dental crown shape for MRI. *ATR Research Report*, 39.

Webb, S. (ed). (1996). *The physics of medical imaging.* Institute of Physics Publishing, London.

Westbury, J. (1988). Mandible and hyoid bone movements during speech. *Journal of Speech and Hearing Research*, **3**1, 405–416.

Westbury, J.R. (1994). *X-ray microbeam speech production database user's handbook.* Tech. rept. Waisman Center on Mental Retardation and Human Developmnet, University of Wisconsin, Madison.

Whalen, D., Min Kang, A., Magen, H., Fulbright, R., & Gore, J. (1999). Predicting Midsagittal Pharynx Shape From Tongue Position During Vowel Production. *Journal of Speech, Language and Hearing Research*, **42**, 592–603.

Wilhelms-Tricarico, R. (1997). Physiological modeling of speech production: Methods for modeling soft-tissue articulators. *Journal of the Acoustic Society of America*, **97**, 3085–3098.

Wilhelms-Tricarico, R. (2000). Development of a tongue and mouth floor model for normalization and biomechanical modeling. *Pages 141–144 of: Proceedings of the 5th Speech Production Seminar: Models and data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling.*

Wrench, A. (1999). An investigation of sagittal velar movement and its correlation with lip, tongue and jaw movement. *Pages 435–438 of: Proceedings of the XIV ICPhS.*

Wrench, A., & Hardcastle, W. (2000). A multichannel articulatory speech database and its application for automatic speech recognition. *Pages 305–308 of: Proceedings of of the 5th Seminar on Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling.*

Wrench, A., McIntosh, A., & Hardcastle, W. (1996). Optopalatograph (OPG): A new apparatus for speech production analysis. *Pages 1589–1592 of: Proceedings of the* 4th *ICSLP.*

Wrench, A., McIntosh, A., & Hardcastle, W. (1997). Optopalatograph: development of a device for measuring tongue movement in 3D. *Pages 1055–1058 of: Proceedings of Eurospeech '97.*

Wrench, A., McIntosh, A., & Hardcastle, W. (1998). Optopalatograph: real-time feedback of tongue movement in 3D. *Pages 1867–1870 of: Proceedings of the* 5th *ICSLP.*

Wu, C-M., & Wilhelms-Tricarico, R. (1995). Tongue structural model: integrating MRI and anatomical structure information into a finite element model of the tongue. *Pages 490–493 of: Proceedings of the XIIIth ICPhS*, vol. 2.

Yang, B. (1996). Measurement and synthesis of the vocal tract of Korean monophtongs by MRI. *Pages 793–796 of: Proceedings of the* 4th *ICSLP.*

Yang, C-S., & Kasuya, H. (1994). Acurate Measurement of Vocal Tract shapes from magnetic resonance images of child, female and male subjects. *Pages 623–626 of: Proceedings of the* 3rd *ICSLP*, vol. 2.

Yehia, H., & Tiede, M. (1997). A parametric three-dimensional model of the vocal-tract based on MRI data. *Pages 1619–1622 of: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 3.

Zierdt, A., Hoole, P., Honda, M., Kaburagi, T., & Tillman, H. (2000). Extracting tongues from moving heads. *Pages 313–316 of: Proceedings of the 5th Speech Production Seminar: Models and data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling.*