

# Collecting and analysing two- and three-dimensional MRI data for Swedish

Olov Engwall and Pierre Badin\*

\* Institut de la Communication Parlée, UPRESA CNRS 5009, INPG-Université Stendhal, Grenoble, France

## Abstract

*MRI (Magnetic Resonance Imaging) data have been collected for a male speaker of Swedish producing sustained vowels and consonants in VCV-context. The resulting database consists of one 3D set of 43 Swedish articulations, covering the entire vocal tract, and one midsagittal set of 85 articulations. The three-dimensional vocal tract shape has been reconstructed for 26 of the articulations, and the result for the Swedish fricatives and a vowel subset is reported on. The midsagittal image set has been analysed by applying Principal Component Analysis on the vocal tract contours to extract articulatory control parameters. The results of the analysis are presented together with findings on articulation strategies for the subject. A number of articulatory measures have been determined and the co-articulatory influence on these measures has been investigated.*

## Introduction

The development of three-dimensional talking heads (e.g. Badin et al., 1998a; Cohen et al., 1998; Lundeberg & Beskow, 1999) and vocal tract models (Engwall, 1999a) calls for more full three-dimensional data of the vocal tract. A number of MRI studies have been carried out to analyse 3D vocal tract shapes for a few subjects in a few languages. Observations have dealt with vowels<sup>1</sup> and groups of consonants<sup>2</sup>: nasals, rhotics, fricatives and laterals. Tiede (1996) studied pharyngeal volume contrasts between English and Akan and found substantial differences, indicating language dependence of vocal tract shapes. We know of no systematic MRI study for Swedish. Foldvik et al. (1988, 1993 and 1995) has reported on MRI acquisition of Norwegian, but these studies however had other aims than the study presented here. Foldvik et al. (1993 and 1995) focused on modeling a time-evolving vocal tract for the transition of /a/ to /i/ and consequently only acquired a large number of repetitions (120 and 200, respectively) of the sequence /pai/ (1993) and /ai/ (1995). The

corpus of Foldvik et al. (1988) on the other hand was larger, consisting of 7 vowels and 8 consonants, including dental and palatal laterals, palatal plosives and a retroflex nasal. The study was however limited to the midsagittal plane. Three-dimensional measurement data for Swedish is hence of great interest, as Swedish involves a larger variety of vowels and fricatives compared to the languages in the three-dimensional MRI studies cited above (mainly English, French, Japanese or Tamil).

Most vocal tract models, both for use in articulatory synthesis and for studies on human speech production, rely on a representation of the vocal tract in the midsagittal plane (e.g. Rubin et al., 1981; Maeda, 1990). These models are based on the midsagittal distances, and the assumption that an approximate relationship can be found between these distances and the corresponding cross-sectional areas (cf. Bear et al. 1991 or Beautemps et al., 1995 for an overview). To model the vocal tract, not only in the midsagittal plane, but as a whole, including the articulatory organs, more detailed 3D data is needed. Knowledge of the vocal tract shape and of the area functions depends on such data, but this is all the more true for the complex three-dimensional shape of the tongue.

<sup>1</sup> Vowels: e.g. Bear et al. (1991), Yang & Kasuya (1994), Story et al. (1996), Demolin et al. (1996).

<sup>2</sup> Nasals: Matsumura et al. (1994), rhotics: Alwan et al. (1997), fricatives: Narayanan et al. (1995) and laterals: Narayanan et al. (1997).

The general lack of studies of full 3D vocal tract shapes is due both to the difficulties of making the measurements and to the fact that full 3D measurements still has to be performed on artificially sustained configurations. Direct vocal tract measurement studies are scarce: measures have been made from casts of live subjects (Ladefoged et al., 1971; Sundberg et al., 1987) or from cadavers (Heinz and Stevens, 1964).

For measures under speech-like conditions, a compromise between spatial and temporal resolution has to be found. Good temporal resolution can be obtained using X-ray (e.g. Fant, 1964, Sundberg, 1969, Branderud et al., 1998 for Swedish, Beutemps et al., submitted), X-ray microbeams (Kiritani, 1986), ultrasound imaging (Stone, 1990, 1996) or electropalatography (Nguyen et al., 1996; Mair et al., 1996). However, these methods provide only measurements of the vocal tract in two dimensions or in very restricted regions.

Stone (1996) used ultrasound imaging to measure and reconstruct tongue shapes for sustained articulations with an acquisition time of 10 seconds. The two methods providing full 3D data of the entire vocal tract existing today, Computed Tomography (CT) based on X-rays (Sundberg et al., 1987) and Magnetic Resonance Imaging (MRI) require yet much longer scanning times. They both present additional disadvantages, such as the unavoidable supine position and the noisy environment in MRI. Moreover, X-ray CT can not be practically used for speech studies, as even one single phoneme measurement would result in an X-ray dose exceeding tolerated limits for non-medical use. MRI, being non-invasive for the subject, is thus the only method for 3D measures available at present.

Fortunately, acquisition times for MRI have been drastically reduced during the past few years, from several minutes (Baer et al., 1991; Story et al., 1996) to around 40 seconds to cover the entire vocal tract (Badin et al., 1998b). This encouraged us to initiate the present study, and to record by MRI a Swedish subject on a rather large corpus consisting of both one set of full 3D measurements and one set of midsagittal images. The MRI data were collected at the Centre Hospitalier Régional Universitaire de Grenoble (CHRUG), France, in a joint effort between CTT and l'Institut de la Communication Parlée (ICP), Grenoble, using the set-ups and protocols developed by Badin et al. (1998b) for a French subject.

The resulting database of 2D and 3D vocal tract shapes for Swedish has been exploited to

extract articulatory measurements and vocal tract geometry for use in vocal tract modeling. This article describes the different steps of the study: image acquisition, 2D and 3D vocal tract shape reconstruction and articulatory measurements. It also includes an analysis of the articulatory strategies of the subject and a first linear midsagittal articulatory model.

## **The subject and the corpus**

### **The subject**

When considering using measures from a database to control an articulatory model, two approaches can be used. Either collecting data from a large number of speakers, trying to find speaker-independent characteristics, or using a reference subject for the acquisition (e.g. Beutemps et al., 1996; Story et al., 1996).

The advantage of the first approach is that the result is a general model representing standard articulations. The main drawback of this method is that the amount of data, as well as the complexity of the analysis, increases dramatically. The corpus size is generally limited as a consequence; Yang and Kasuya (1994) and Matsumura et al. (1994) both used a small corpus (the five Japanese vowels and the Japanese vowels plus /s/, respectively) when examining three subjects. Moreover, if the variability of articulatory strategies between subjects is great, there is an apparent risk that the averaged characteristics are unable to reproduce natural articulations.

Choosing to use a reference subject simplifies the data collection and analysis, but may introduce non-representative, speaker-specific articulations. The advantages are however important. The corpus size can be increased significantly; Story et al. (1996) collected data for 22 speech sounds and Badin et al. (1998b) for 34. Moreover, with a reference subject, the modeled vocal tract shape can be compared directly to measurements. The subject's natural speech also provides a direct reference that can be used when trying to interpolate between area functions to generate "dynamic" speech from static vocal tract shapes (Story et al., 1996). A model based on one reference subject also has larger possibilities of successfully combining data from different acquisition methods. This is an advantage especially if data from different sources only cover small parts, as the successful attempt by Badin et al. (1997) to reproduce tongue shapes from the positions of

Table 1. The midsagittal and 3D corpora

		Midsagittal	Full 3D
Vowels	Long vowels	/ɑ:, e:, ε:, æ:, i:/ /y:, u:, ʊ:, o:, œ:, ø:/	/ɑ:, e:, æ:, i:/ /y:, u:, ʊ:, o:, œ:, ø:/
	Short vowels	/a, ɪ, ʊ, ʏ, θ, ɔ/	/a, θ, ɔ/
Consonants	Context	/a, ɪ, u, ɔ/	/a, ɪ, u /
	Stops	/p, t, k/	/p, t, k/
	Liquids	/l, r/	/l, r/
	Fricatives	/f, s, ʃ, ʂ, ç, ʝ/	/f, s, ʃ, ʂ, ç/
	Nasals	/m, n, ŋ/	
	Retroflexes	/ʈ, ʡ, ɳ/	

Shaded slots indicate vowel context for consonants in the midsagittal and the 3D set, respectively.

only three tongue pellets measured with an articulograph.

In this study, the latter approach was opted for. The subject was a 27-year-old and 193 cm tall male native speaker of Swedish from the Stockholm area (the first author of this paper), with a mid-Swedish dialect close to standard. He has received no formal phonetic training and has no record of speech or voice disorders.

### The corpus

Two slightly different corpora were recorded for the full 3D image set and the midsagittal set. Due to the overall long image acquisition time, the 3D corpus was limited to the subset of 43 sounds in the midsagittal set that were considered most important. Both sets included reference configurations with upper and lower incisors aligned and touching.

The consonants were produced in VCV sequences, having the subject make the initial VC transition prior to the acquisition, then sustaining the articulation during the scan and finally making the CV transition.

#### The midsagittal corpus

The midsagittal corpus consisted of all Swedish long vowels, the short vowels with articulation differing from the long version for the subject, unvoiced fricatives and stops, liquids, nasals and the retroflex sounds /ʈ, ʡ, ɳ/.

Table 2. Reference words for vowel pronunciation

Unrounded vowels		Rounded vowels	
IPA	Word	IPA	Word
i:	vit	y:	byt
ɪ	vitt	ʏ	bytt
e:	vet	u:	bo
ε:	säl	ʊ	bott
æ:	här	ʊ:	hus
a	matt	θ	hund
ɑ:	mat	ø:	hö
		œ:	hör
		o:	gå
		ɔ	gått

Shaded slots: vowels included in the midsagittal set only.

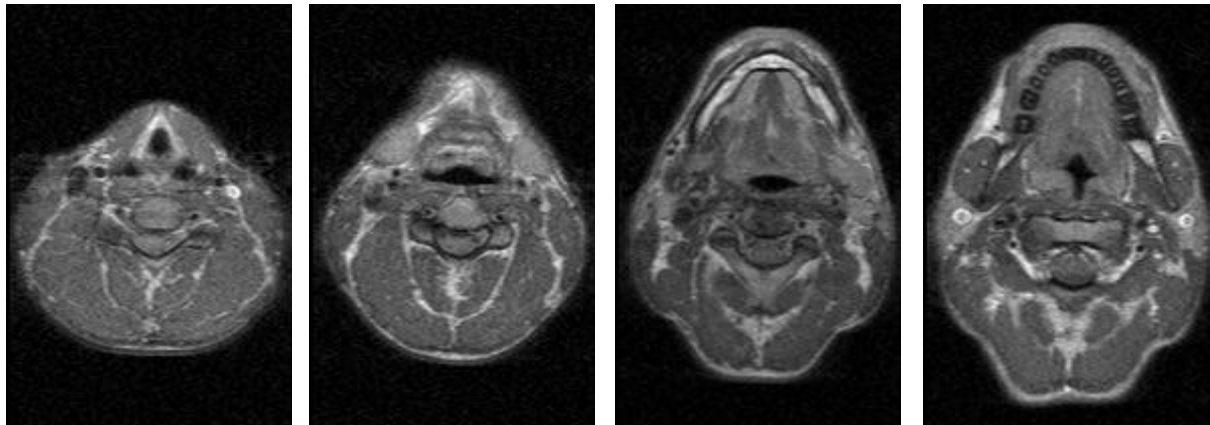
The vowels were produced as they would be pronounced in the words of Table 2 (Elert, 1989). A few vowels are however diphthongized (e.g. /e:/ and /o:/) in mid-Swedish in that context. The subject avoided the diphthongs in those cases, trying to sustain stable vowels.

The consonants in the midsagittal plane were measured in VCV context, with V = /a, ɪ, u, ɔ/. The fourth vowel /ɔ/ was added to the three point vowels to clarify context influence of a maximally retracted back vowel. Altogether 88 articulations were measured in the midsagittal plane.

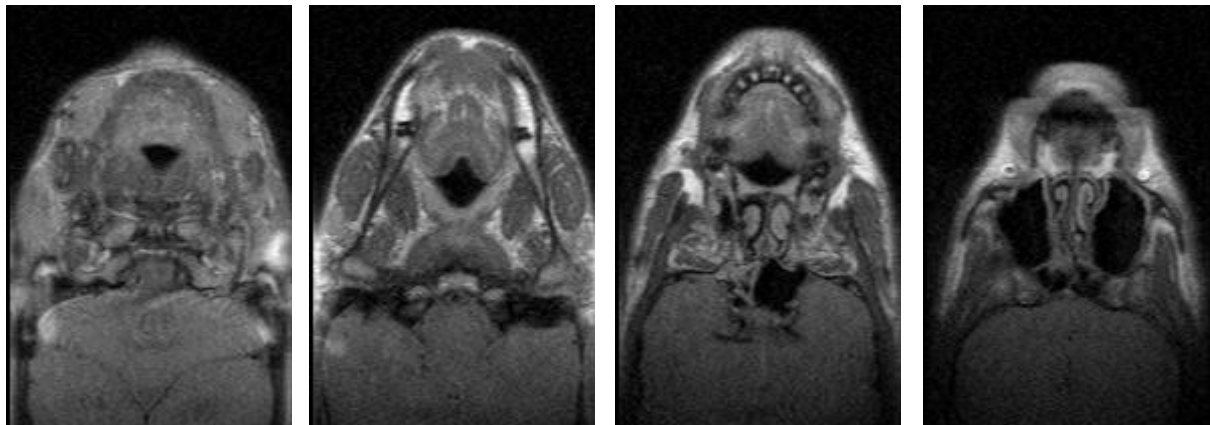
#### The 3D corpus

The 3D vowel corpus consisted of the vowels in the midsagittal corpus that were judged to differ significantly from each other and could be sustained throughout the acquisition time. The vowel corpus thus included ten long and three short vowels, pronounced in the reference context given in the unshaded part of Table 2, but with diphthongation suppressed.

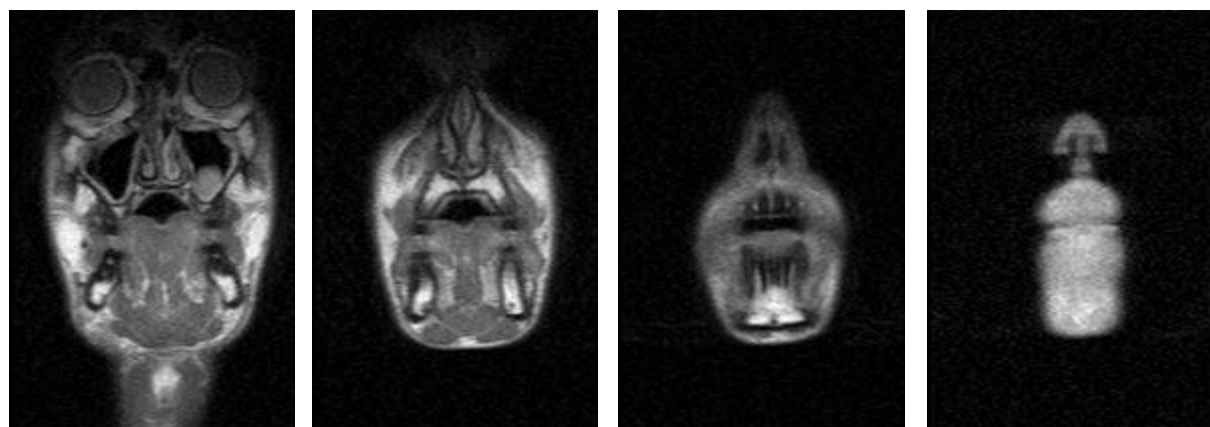
The 3D consonant corpus contained plosives /p, t, k/, liquids /l, r/, fricatives /f, s/ and the three different mid-Swedish sj-sounds /ʃ, ʂ, ç/. All consonants were produced in VCV context with the three point vowels V=/a, ɪ, u/. In total 47 different configurations were recorded in full 3D.



A) Pharyngeal set. Pictures 1, 7, 13 and 18 (cf. Figure 2 for the corresponding grid position). The jaw is in the upper part, the neck in the lower in each picture. Note the epiglottis that starts to appear in picture 7, the marrow of the teeth appearing in white in picture 13 and the coronas of the teeth in black in picture 18.



B) Tilted set. Pictures 19, 25, 31, 36. Orientation: Jaw at top, brain at bottom. Note the important tongue grooving in picture 25, the nasal cavity structure in pictures 31 and 36 and the lower lip appearing in picture 36.



C) Coronal set. Pictures 37, 41, 45 and 48. The bilabial closure is made evident in picture 48.

Figure 1. Excerpt from the 3D image set for /ɪpɪ/. Air passages and calcified structures appear in black, hydrogen rich parts, such as fat and marrow, in white.

## Acquisition set-up and method

MR images and speech sound were recorded quasi-simultaneously, in order to allow for verification of the coherence of the data.

### MR image acquisition

The measurements were realised in one three hour session with a 1 Tesla MRI Scanner Philips GyroScan T10-NT at the Grenoble Regional University Hospital (CHRUG), France.

As the entire corpus was acquired during one single session, no errors were introduced due to subject repositioning between sessions.

The subject lay in supine position in the MR machine with his head inside a Radio Frequency (RF) coil. To minimise head movements a padded crane support was used in the RF coil. The subject's head was hence not fixed, but movements were limited by the crane support and the coil provided a reference frame for keeping the head in position.

The acquisition time was approximately 43 seconds for the 3D set and 11 seconds for the 2D set. During this time, the subject held the articulation in full apnoea (vowels and stops) or breathing out very slowly (fricatives). The long acquisition time is of course a drawback, as the subject could not phonate naturally during image acquisition. It is however lower than in earlier studies (Baer et al., 1991; Matsumura et al., 1994; Yang & Kasuya, 1994; Story et al., 1996; Tiede, 1996).

It was also considered better to scan the entire vocal tract during one sustained production than having the subject repeat the articulation several times as done by (Narayanan et al., 1995, 1997, Alwan et al., 1997, Mohammad et al., 1997), since this could introduce artefacts due to intra-subject variability between acquisitions.

### The 3D set

For each phoneme in the 3D set, 54 256x256-pixel images were collected, with a final image resolution of 1 mm/pixel.

The images were taken as slices (Figure 1), 3.6 mm thick, and sampled every 4.0 mm, as shown in Figure 2.

As first proposed by Demolin et al. (1996) and then systematically used by Badin et al. (1998b) and Apostol et al. (1999), the acquisition was carried out using three stacks of parallel slices. Each stack was adjusted to

generate slices as orthogonal as possible to the midline of the vocal tract.

The first stack, denoted A in Figure 1-2, contains 18 horizontal slices of the pharynx. The second stack of 18 slices, B in Figure 1-2, is tilted 45° with respect to the horizontal. The last, coronal, stack consists of 18 vertical slices (denoted C in Figure 1-2) of the oral cavity.

Altogether 54 slices were obtained, sampling the vocal tract from the larynx to the lips in a partly overlapping manner.

Figure 1 shows four images from each stack, with the air passage and calcified structures (teeth and bones) in black, muscular tissue in grey and fatty tissues and bone marrow in white, indicating large amounts of free hydrogen atoms.

### The midsagittal set

The midsagittal images had a size of 256x256 pixels and a final resolution of 1 mm/pixel. As exemplified in Figure 3, they were acquired to cover the largest vocal tract length possible, with maximally lowered larynx and maximally protruded lips.

Collecting a midsagittal set of images permits on the one hand to get an overall view of the articulation from one image, and on the other hand to perform two-dimensional modeling directly on the midsagittal image set. Furthermore, since the acquisition time is much shorter, the vocal tract contours from the midsagittal set can be used to detect artefacts in the 3D set that are due to fatigue related to

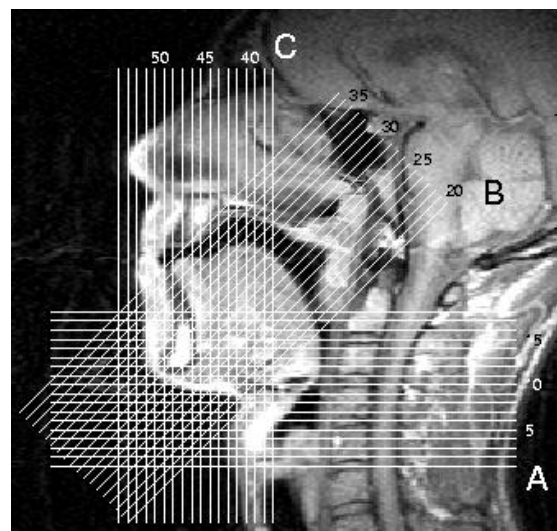


Figure 2. The 54 slices of the CHRUG grid used for the 3D acquisition. For reference the grid has been superimposed on the midsagittal picture of /ɪp/.

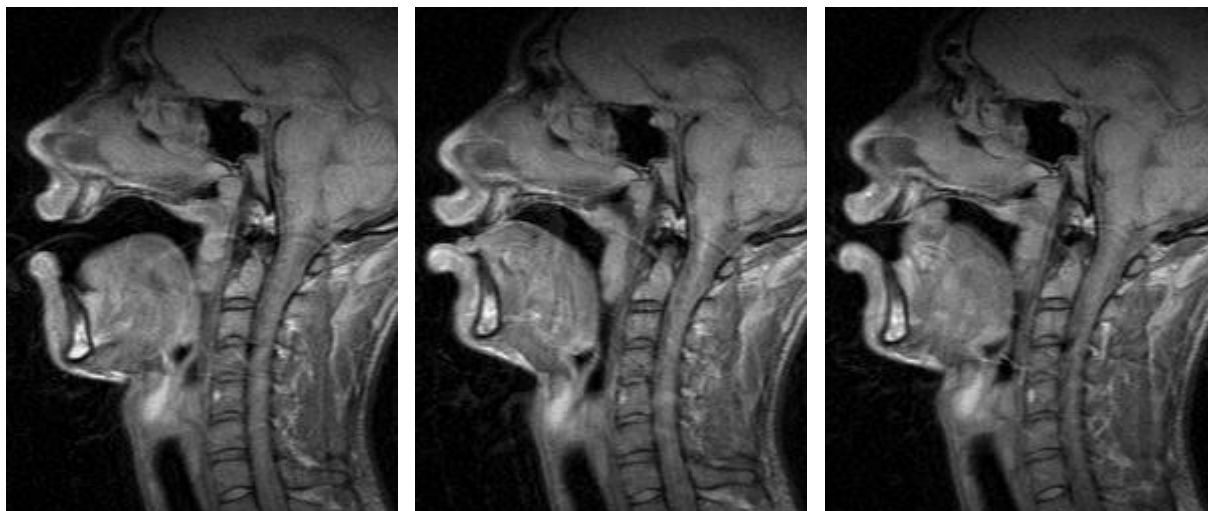


Figure 3. Midsagittal images of /œ:/, /unu/ and /ɔ]ɔ/. In all pictures, an artefact due to minor movements of the velum can be seen. Note the absence of the teeth in the pictures. The position of the incisors in the midsagittal plane could however be inferred from /unu/, where the subject protrudes his tongue tip between the teeth, leaving clear marks in the tongue.

the long duration of the acquisition. It should however be stated that the two image sets were scanned separately and thus do not show identical articulations.

### Dental cast

As calcified structures, such as teeth and bones, contain no free hydrogen atoms, they will not appear in the MR images.

In earlier studies, this problem has been resolved in various manners. The teeth and palate have been painted with mineral oil, a thin coat of paraffin wax or a paste containing gadolinium and barium sulphate in attempts to make the structures visible, but the results are often not satisfactory (Narayanan et al., 1995).

Wakumoto et al. (1996) used thin (a total thickness of 0.5 mm) plates for the dental crown made with two layers of thermoforming material. As a contrast medium for MRI was enclosed and sealed between the layers, the dental shape appeared with sufficient quality in the images for boundary extraction. The inconvenience of this method is the uncertainty in the positioning of the film (Wakumoto et al., 1996, adjusted the plate individually for each subject using plaster casts) and the possible interference with the subject's normal articulation.

Story et al. (1996) used Electron Beam Computed Tomography (EBCT) to generate a "cast" of the teeth, which was subsequently subtracted from the vocal tract shapes.

The solution opted for here was similar, as the teeth and hard palate were measured separately and introduced with correct position

and orientation when reconstructing the vocal tract (as e.g. Narayanan et al., 1995, 1997; Badin et al. 1998b).

Casts of the subject's palate and dentition were created from dental impression and were immersed in water to be scanned by the same MRI machine as the subject.

Two stacks of images, one coronal and one sagittal, were taken of the two casts; both sampled every 1.4 mm (cf. Figure 4). This resulted in sets of 35 coronal images for the palate and of 41 images for the jaw, as well as sets of 48 sagittal images for both parts.

The contours in each image were detected, using a method explained below in the description of vocal tract contour extraction, and the relevant parts of the casts were preserved. The midsagittal contours of the casts were positioned to fit the upper and lower incisors in a midsagittal MR image of the reference position<sup>3</sup>.

The geometry of the reconstructed casts is given in Figure 4, with the coronal slices indicated. The reconstructed casts have been exploited to introduce teeth and palate into the KTH vocal tract model. In that process, the casts were subsampled and made symmetrical and the teeth and palate were separated and given natural colours. For details on the adaptation to the 3D environment, refer to the description in Engwall (1999b) on how tongue shape data was incorporated.

<sup>3</sup> The reference position depicts the neutral vocal tract shape with upper and lower incisors touching and aligned.

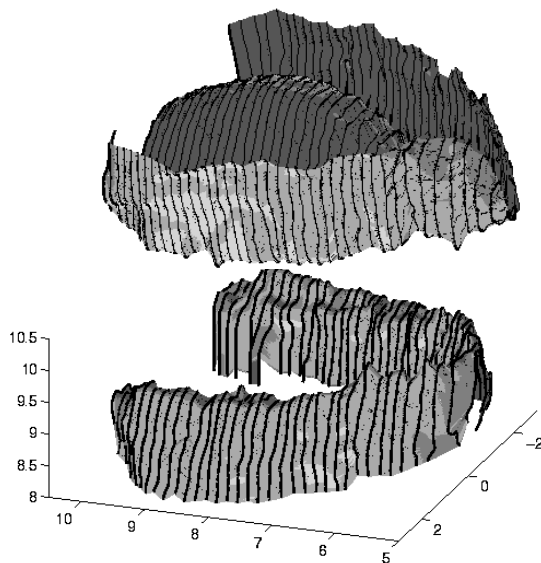


Figure 4. The reconstructions of casts of palate (above) and jaw (below), viewed from (200, 20), with spacing of coronal set slices.

### Sound recordings

Due to the noise level during image acquisition it was not possible to do any sound recording during the acquisition phase. However, to provide a speech signal reference, recordings were made of the subject producing a normal, voiced articulation immediately before and after each acquisition phase. The recordings can be analysed to detect differences in the sound characteristics prior to and after the image acquisition. Divergence could indicate that the subject had not been able to sustain the articulation throughout the acquisition, thus undermining the image accuracy.

The recordings were made with a DAT-recorder connected to a microphone with a screened cable. The microphone was placed inside the MRI tunnel as close as possible to the subject's head.

The recordings were judged too noisy for systematic evaluation of formants for a large number of sounds. The long vowels of the midsagittal set were however analysed with the speech analysis module Snack<sup>4</sup> (Sjölander et al., 1998) for comparison with the subject's normal vowel articulation. The aim of this evaluation is to assure that formant frequencies before and after acquisition are consistent, to detect large differences compared to the subject's formant frequencies under normal conditions or find divergent formant pattern of the speaker when compared to Fant (1969).

<sup>4</sup> Snack is an extension to the Tcl/Tk script language for sound applications, available at <http://www.speech.kth.se/SNACK>.

A normal articulation reference was recorded as the subject was sitting upright saying the words with normal duration. The recording was made with a different microphone, at a later occasion and in a different room. The measured formant frequencies may thus be slightly different due to these factors, but the conclusions of the evaluation are unaffected by the differences.

### Evaluation of formants

Fant (1969) showed that there are some differences in the formant values of the Swedish vowels sustained during 4 seconds (Fant, 1959) and those of normal duration (Fant, 1969). It is thus of interest to get an indication of frequency differences between the subject's normal vowels and those sustained over 12-13 seconds. It is however difficult to isolate differences due to vowel duration from those induced by the supine position, the acoustic environment (earplugs and a narrow MRI tunnel) or possible tension in the subject's voice because of the experiment set-up<sup>5</sup>. Influence of the supine position can to some extent be analysed by comparing the reference formant frequencies with those recorded under identical conditions but in supine position.

Table 3 shows the formant frequencies before and after acquisition compared to the subject's vowel formants under normal conditions. The formants before and after acquisition are clearly consistent, the difference generally being below 10% (which is a reasonable error level as the formants could only be extracted with an accuracy of 20 Hz in the noisy spectrogram). For a small number of formants, there are large (up to 31%) differences compared to the reference vowels, but for the majority the deviation is of an order that could be attributed to or masked by intra-subject variability. The vowels measured in the MRI study can hence be considered representative of the subject's vowels and deviations are not systematic.

Moreover, the supine position does not seem to have large influence on the vowel formants, as indicated by Table 4, other than a slight decrease of F1. Table 4 also indicates that the subject has overall higher vowel formants than the mean value found by Fant (1969) for 24 Swedish male students.

<sup>5</sup> In particular body immobility over several hours, the noise level during acquisition and the narrowness of the MRI tunnel.

Table 3. Formant frequencies before and after MR image acquisition compared to the subject's normal vowels.

The deviations are given as procentual difference of the frequency after compared to that prior to acquisition ( $\Delta_{A-B}$ , columns 4, 8 and 12) and deviation of frequencies measured during the MRI session compared to normal vowel frequencies ( $\Delta_{Ref}$ , columns 6, 10 and 14). Before and after for the reference formants refer to the beginning and end of the vowel. The vowel diphthongation in the reference was suppressed by considering start- and endpoints of the stable vowel part for diphthongated vowels.

IPA		F1 (kHz)				F2 (kHz)				F3 (kHz)			
			$\Delta_{A-B}$ (%)	Ref.	$\Delta_{Ref}$ (%)		$\Delta_{A-B}$ (%)	Ref.	$\Delta_{Ref}$ (%)		$\Delta_{A-B}$ (%)	Ref.	$\Delta_{Ref}$ (%)
u:	Before	0.36	5.6	0.38	-5.3	0.66	9.1	0.74	-11	2.20	1.8	2.20	0
	After	0.38		0.34	11.8	0.72		0.74	-2.7	2.24		2.18	2.8
o:	Before	0.38	5.3	0.40	-5.0	0.66	7.6	0.74	-10.8	-	-	2.10	-
	After	0.40		0.40	0	0.71		0.74	-4.0	-		2.12	-
ɑ:	Before	0.64	0	0.60	6.7	1.02	-3.9	1.00	2.0	2.72	2.2	2.64	3.0
	After	0.64		0.62	3.2	0.98		0.98	0	2.78		2.58	7.8
æ:	Before	0.74	-2.7	0.72	2.8	1.62	2.5	1.46	11	2.46	2.4	2.46	0
	After	0.72		0.70	2.9	1.58		1.48	6.8	2.52		2.44	3.3
ɛ:	Before	0.66	0	0.62	6.5	1.74	0	1.74	0	2.46	0.81	2.46	0
	After	0.66		0.66	0	1.74		1.74	0	2.48		2.40	3.3
e:	Before	0.38	5.3	0.38	0	2.34	0.85	2.20	6.4	2.66	0	2.62	1.5
	After	0.40		0.50	-25	2.36		2.14	10	2.66		2.62	1.5
i:	Before	0.38	11	0.32	19	1.86	8.6	1.52	22	2.76	-1.4	2.90	-4.8
	After	0.42		0.32	31	1.70		1.68	1.2	2.72		3.08	12
y:	Before	0.38	-5.3	0.36	5.6	1.72	1.2	1.56	10	2.44	11	2.88	-15
	After	0.36		0.36	0	1.74		1.56	12	2.72		2.94	-7.5
ɯ:	Before	0.38	0	0.40	-5.0	1.66	1.2	1.72	-3.5	2.40	0	2.40	0
	After	0.38		0.34	12	1.68		1.66	1.2	2.40		2.34	2.6
ø:	Before	0.62	-3.2	0.52	19	1.12	0	1.08	3.7	2.68	1.5	2.38	13
	After	0.60		0.50	25	1.12		1.08	3.7	2.72		2.36	15
œ:	Before	0.60	3.3	0.58	3.5	1.70	-1.2	1.52	12	2.32	-1.2	2.38	-2.5
	After	0.58		0.58	0	1.68		1.50	12	2.38		2.36	0.85

Table 4. Compared formant frequencies of the subject's vowels in supine (Sup., columns 2, 6, 10, 14) and normal position (Ref., columns 3, 7, 11, 15). Formant frequencies found by Fant (1969) for 24 subjects are given as a reference of Swedish standard pronunciation. The deviations are given as relative difference of the frequency in supine position compared to normal vowel frequencies ( $\Delta_{Ref}$ , columns 4, 8 and 12).

IPA	F1 (kHz)				F2 (kHz)				F3 (kHz)				F4 (kHz)			
	Sup.	Ref.	$\Delta_{Ref}$ (%)	Fant	Sup.	Ref.	$\Delta_{Ref}$ (%)	Fant	Sup.	Ref.	$\Delta_{Ref}$ (%)	Fant	Sup.	Ref.	$\Delta_{Ref}$ (%)	Fant
u:	0.34	0.38	-11	0.29	0.80	0.74	8.1	0.60	2.32	2.20	5.5	2.33	3.20	3.04	5.3	3.30
o:	0.40	0.40	0	0.39	0.74	0.74	0	0.69	2.40	2.10	14	2.42	3.14	3.00	4.7	3.15
ɑ:	0.56	0.60	-6.7	0.60	0.94	1.00	-6.0	0.93	2.74	2.64	3.0	2.54	3.24	3.34	-3.0	3.29
æ:	0.76	0.72	5.6	0.63	1.34	1.46	-8.2	1.72	2.44	2.46	-0.8	2.50	3.60	3.42	5.3	3.40
ɛ:	0.60	0.62	-3.2	0.51	1.62	1.74	6.9	1.94	2.46	2.46	0	2.54	3.54	3.56	0.56	3.42
e:	0.34	0.38	-11	0.35	2.10	2.20	-4.5	2.25	2.60	2.62	-0.8	2.85	3.52	3.52	0	3.40
i:	0.32	0.32	0	0.26	1.58	1.52	3.9	2.19	3.04	2.90	4.8	3.15	3.72	3.46	7.5	3.40
y:	0.30	0.36	-17	0.26	1.54	1.56	-1.3	2.06	2.84	2.88	-1.4	2.68	3.50	3.50	0	3.30
ɯ:	0.36	0.40	-10	0.29	1.72	1.72	0	1.64	2.54	2.40	5.8	2.25	3.28	3.26	0.61	3.31
ø:	0.50	0.52	-3.8	0.38	1.06	1.08	-1.9	1.73	2.48	2.38	4.2	2.29	3.24	3.02	7.3	3.39

## Analysis of the 3D set of data

This section describes the steps involved in the reconstruction of the vocal tract in three dimensions. Area functions are also derived from the set of vocal tract contours extracted from the 3D set.

### Contour extraction

The contour of the vocal tract was extracted from each image using edge detection in binary image mode, obtained by thresholding. The output value for each pixel of the image in the thresholding was set according to its grey level as

$$p_{out} = \begin{cases} 0, & \text{if } p_{in} < T \\ 1, & \text{otherwise} \end{cases} \quad \text{with } T = 0.2188$$

The boundary was then detected by a chained pixel search and the contour was coded by Bézier points controlling interpolated splines.

Every contour was checked, and if necessarily, corrected manually in a Matlab editing software developed at ICP, as image brightness can vary slightly from image to image, causing the edge detection algorithm to misinterpret some edges.

In this study only the main vocal tract channel was considered, i.e. the sinus piriforms and sublingual cavities were ignored if not connected with the main air passage, and thus only one contour was determined for each image, as shown in Figure 5.

The result of this first step consisted of up to 54 planar vocal tract contours that were then re-aligned by translations and rotations in the midsagittal plane, from the reference frame of the MRI system, shown in Figure 2, to the semi-polar grid system defined below.

### Determining the reconstruction grid

Following Badin et al. (1998b), the vocal tract shape is defined as planar contours specified by the semi-polar partially dynamic grid of Figure 6 (Beautemps et al., submitted). The grid consists of a dynamically variable larynx set at gridlines 1-6, a traditional semi-polar grid (Maeda, 1988) at gridlines 7-22 and another dynamic set of gridlines 23-28 to follow tongue tip movements, as defined in Beautemps et al. (1996). A third dynamically adjustable set has been added as grid lines 29-33, equally spaced between the tongue tip and the upper incisors lower edge. Finally, the

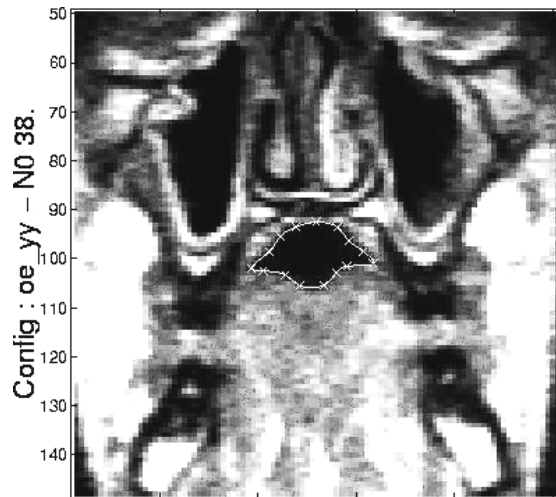


Figure 5. Vocal tract contour extraction with automatic boundary detection of thresholded image and subsequent handediting of the contour using Bézier points.

spacing of gridlines 34-37 are dynamically adjusted to follow lip protrusion. The protrusion is determined using the protrusion of the upper lip, ProTop; of the lower lip, ProBot, and of the lip corners, CnrAdv (refer to Figure 7 for definitions).

These measures were determined from the subset of images of the lips. The degree of protrusion was defined in a one decimal 'image number co-ordinate', judging the value from interpolation between the last image where the lip corner or lip was present and the following, where it was not. As the position of each image slice and the distance to the next is

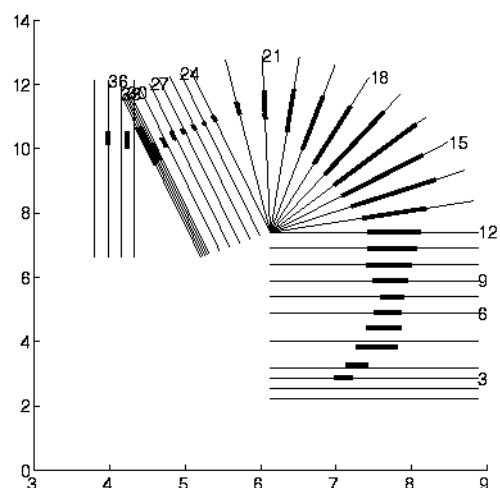


Figure 6. The semi-polar, partially dynamic grid of /ucu/ and its intersections with the vocal tract. Gridline 1-6 (for larynx movements), 23-28 (tongue blade and tip), 29-33 (alveolar cavity) and 34-37 (lip protrusion) are dynamic.

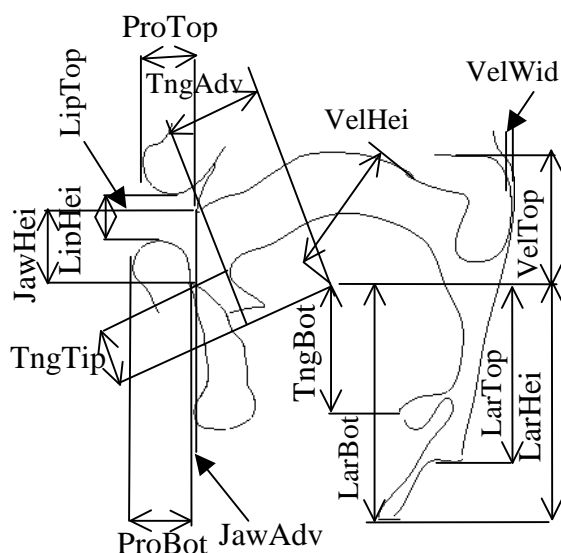


Figure 7. The definitions of articulatory measurements in the midsagittal plane (with vocal tract contour for /æ:/).

known, the absolute protrusion, measured from the lower part of the upper incisor (cf. Figure 7), is uniquely determined from this image co-ordinate.

The remaining measures needed to specify the grid: the extremities of the larynx, LarTop and LarBot, the tongue tip, TngAdv, and the tongue root, TngBot, were determined from a midsagittal section being reconstructed from the original 3D image set using NIH Image<sup>6</sup>.

Furthermore, the same reconstructed midsagittal image was used to correct for minor head movements between acquisitions. Midsagittal profiles of the palate and the jaw, extracted from the dental casts, were fitted into place on the midsagittal image for each phoneme using a sequence of translations and rotations: rototranslations.

These rototranslations provide a measure of the subject's head movements between the measurements. As the reference point of the semi-polar grid is the lower end of the upper incisor, the rototranslations can be used directly to fit the vocal tract contours to the 3D reconstruction of the palate. Every vocal tract reconstruction is made relative to the palate that is positioned and oriented identically: the lower edge of the upper incisor placed at ( $x=5$  cm,  $y=10$  cm) and its maxillary occlusal

plane<sup>7</sup> making an angle of  $-5.6^\circ$  with the horizontal. The resulting 3D vocal tract shapes were hence reconstructed as if produced in identical position.

### Three-dimensional reconstruction

The centroid was determined for each cross-section and the regression line for these centres was considered as the midline of the vocal tract. The contours were centred on this midline to avoid unnatural discontinuities.

This centring is based on the assumption that the subject's articulations are symmetric, which is approximately, but not totally true, as can be seen in Figure 8. The lateral deviation from the median plane is of a few millimetres for all gridlines, slightly less than found by Story et al. (1996). However, for some configurations a large (of the order of 1 cm) deviation can be found for gridlines 1-3, due to the asymmetry in level of connection between the main vocal tract and the sinus piriformis at the left and right side. As only one contour was extracted from each image, an isolated sinus piriformis was not taken into account, whereas one that was connected to the main vocal tract was. This caused the centroid of the contour to be shifted towards the side of the included sinus. Hence, care had to be taken to centre the contours of these configurations on the

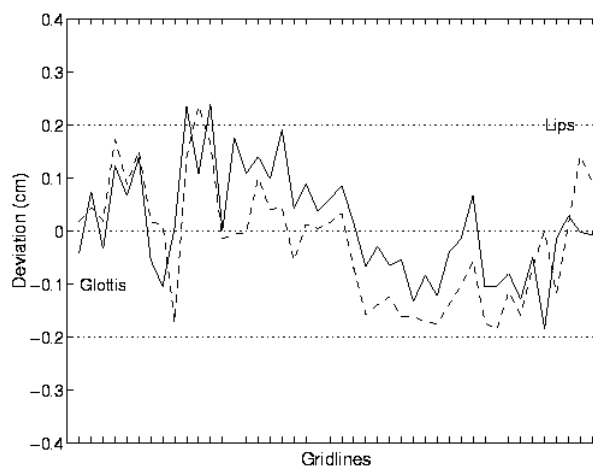


Figure 8. Off-axis deviations of the vocal tract centreline (in cm) as a function of gridline number from the glottis to the lips. Solid line for /asa/, dashed for /utu/. The lateral off-axis variation is plotted as deviation from the median value of the centrelines lateral co-ordinates.

<sup>6</sup> NIH Image is a public domain software developed at the U.S. National Institute of Health, available at <http://rsb.info.nih.gov/nih-image>.

<sup>7</sup> The maxillary occlusal plane is "given by the tips of the central incisors and at least two other maxillary teeth on opposite side of the mouth" (Westbury, 1994),

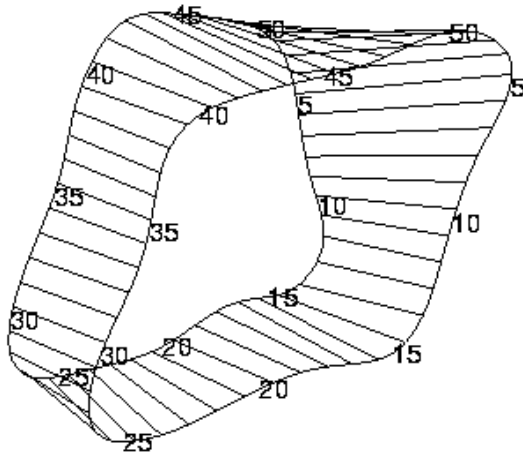


Figure 9. Connecting the planar contours as polygons with vertices  $(P_{i,j}, P_{i+1,j}, P_{i+1,j+1}, P_{i,j+1})$ , where  $i$  is the index of the contour and  $j$  the number of the point on the contour. In this example,  $i=16$  and  $N=50$  for /asa/.

centroid of the main vocal tract, in order to avoid this artefact.

The upper and lower teeth and the hard palate were first cut into slices by the grid associated with the MRI system (Figure 2). The jaw and palate contours for each grid plane were then superimposed on the vocal tract contour of that plane and the overlapping area was withdrawn, correcting for the overestimation of the area due to the fact that the teeth are invisible in MR images.

The corrected contours were then reconstructed three-dimensionally in the semi-polar grid (Figure 6), if necessary by choosing from which of the initial stacks the contour should be taken when stacks were overlapping. The choice was made as to make transitions between different stacks as smooth as possible when there was a discrepancy of contours from different sets. The overlapping and the possible ambiguities before transition smoothing are shown in the left columns of Figure 15.

To avoid minor reconstruction errors, the contours obtained from cutting the reconstructed three-dimensional vocal tract and dental casts with the semi-polar grid were then superposed and, when necessary, additional corrections of the vocal tract contours were performed to provide perfect fit against the dental cast.

The contours were then re-sampled with a set number of control points  $N$  ( $N=100$  in the present study), distributed evenly on each half of the contour ( $N/2$  points on each half, the points 1 and  $N/2$  being on the outer and inner

midsagittal contour). Each contour plane was connected with the following, using fibres linking every point in the plane with the corresponding point in the next. A mesh of polygons was thus formed of polygons having vertices  $(P_{i,j}, P_{i,j+1}, P_{i+1,j+1}, P_{i+1,j})$  as shown in Figure 9.

The purpose of this grouping is two-fold. The first reason is that grouping the points of each contour having the same number into fibres (Badin et al., 1998b) permits to analyse the vocal tract shape fibre by fibre. The most interesting fibres are naturally fibre 1 and  $N/2$ , providing an approximate midsagittal contour that can be compared to existing 2D models (e.g. Rubin et al., 1981, Maeda, 1990, Stark et al., 1996, Boersma, 1998, Beautemps et al., submitted) or with the 2D contours extracted from the midsagittal image set. The midsagittal contours, as generated by fibres 1 and  $N/2$ , of the Swedish fricatives are presented in Figure 10, as an example of how a midsagittal slice from the 3D reconstruction can be used to indicate the place of constriction.

The second purpose of the grouping is to form a mesh of polygons that can be directly introduced in the KTH 3D modeling environment (Beskow, 1995) as data for the vocal tract model (Engwall, 1999a). The number of points and polygons of the different

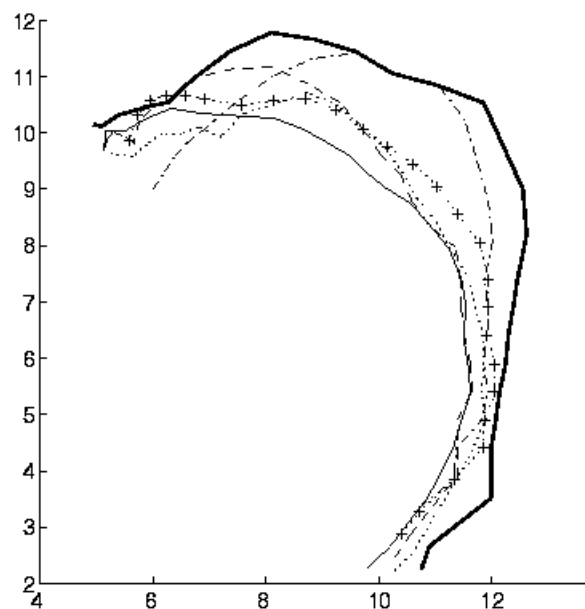


Figure 10. Midsagittal contours extracted from the 3D reconstruction of the Swedish fricatives /s/ (solid line), /f/ (dotted), /ɸ/ (dash-dotted), /ç/ (+-dotted) and /ʃ/ (dashed) in /v/ context. The outer contour is constructed from /s/, and is very close, but not identical, to outer contours of the other fricatives.

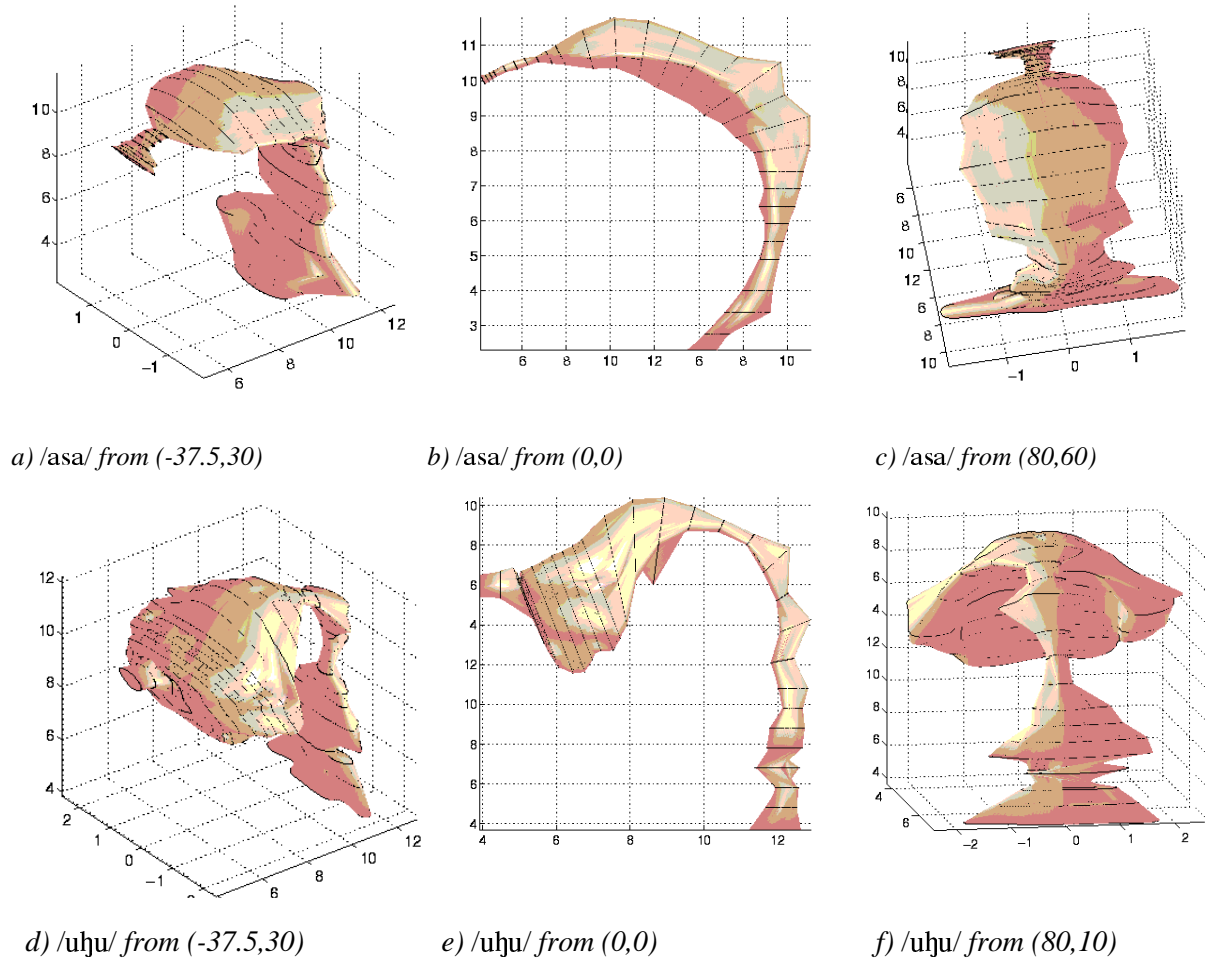


Figure 11. Reconstructed 3D shapes of the vocal tract for one coronal and one velar fricative viewed from three different angles.

vocal tract shapes being constant and the polygons evenly distributed, the polygon mesh of each configuration provides direct measures of the displacement of every vertex in the

mesh. The parameters, the activation factor and the parameter weights for each point can thus be modified to fit the model to the data using error-minimisation (Cohen et al., 1998).

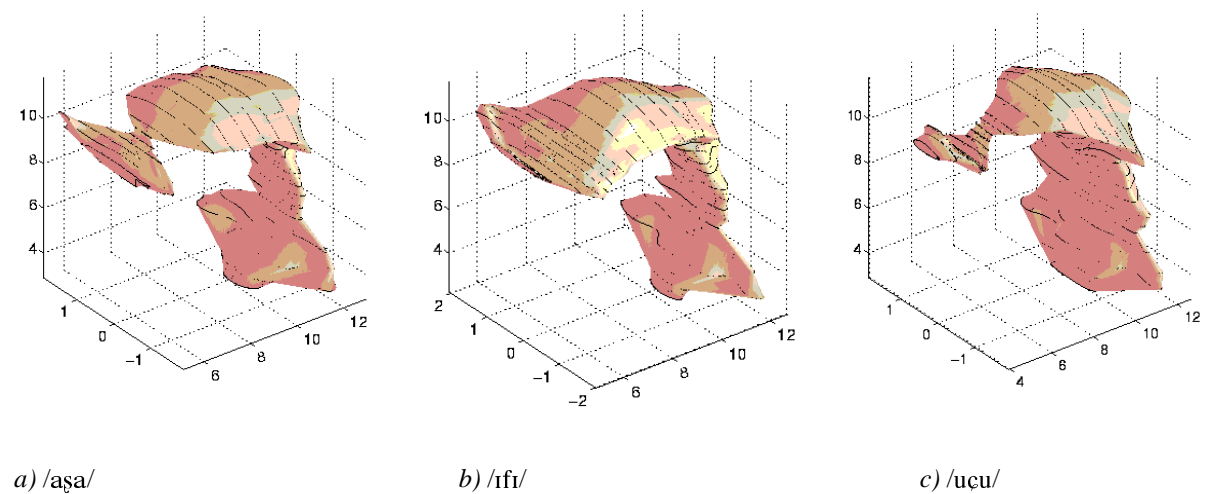


Figure 12. The palatal, labiodental and dental fricatives, viewed from (-37.5,30).

The point co-ordinates can also be used to establish 3D vocal tract models based on statistical Principal Component Analysis (PCA), as done for a French subject by Badin et al. (1998b).

Finally, the vocal tract is rendered with interpolated shading for viewing purposes, as exemplified in Figure 11-12.

The result is less detailed, but comparable to reconstructed vocal tract shapes for English fricatives found by Narayanan et al. (1995), the main difference being that the latter reconstructions included the sinus piriforms and isolated sublingual cavities. Figure 11(d-f) shows very important bilateral resonance cavities caused by the extremely lowered jaw and retracted and raised tongue body.

### Determining the area function

As explained in the previous section, the contours are defined as spline curves, passing through  $N$  Bézier points.

The co-ordinates of these points being known and the contours being non-self-intersecting, the area function can be determined straightforwardly considering the contour as a polygon with  $N$  edges and applying a summation formula (Goldman, 1991) over the vertices to calculate the area of each contour.

When  $N$  is large ( $N=100$  here) the points are spaced at a very small interval compared to the length of the contour. This, and the fact that the points are evenly distributed, assures

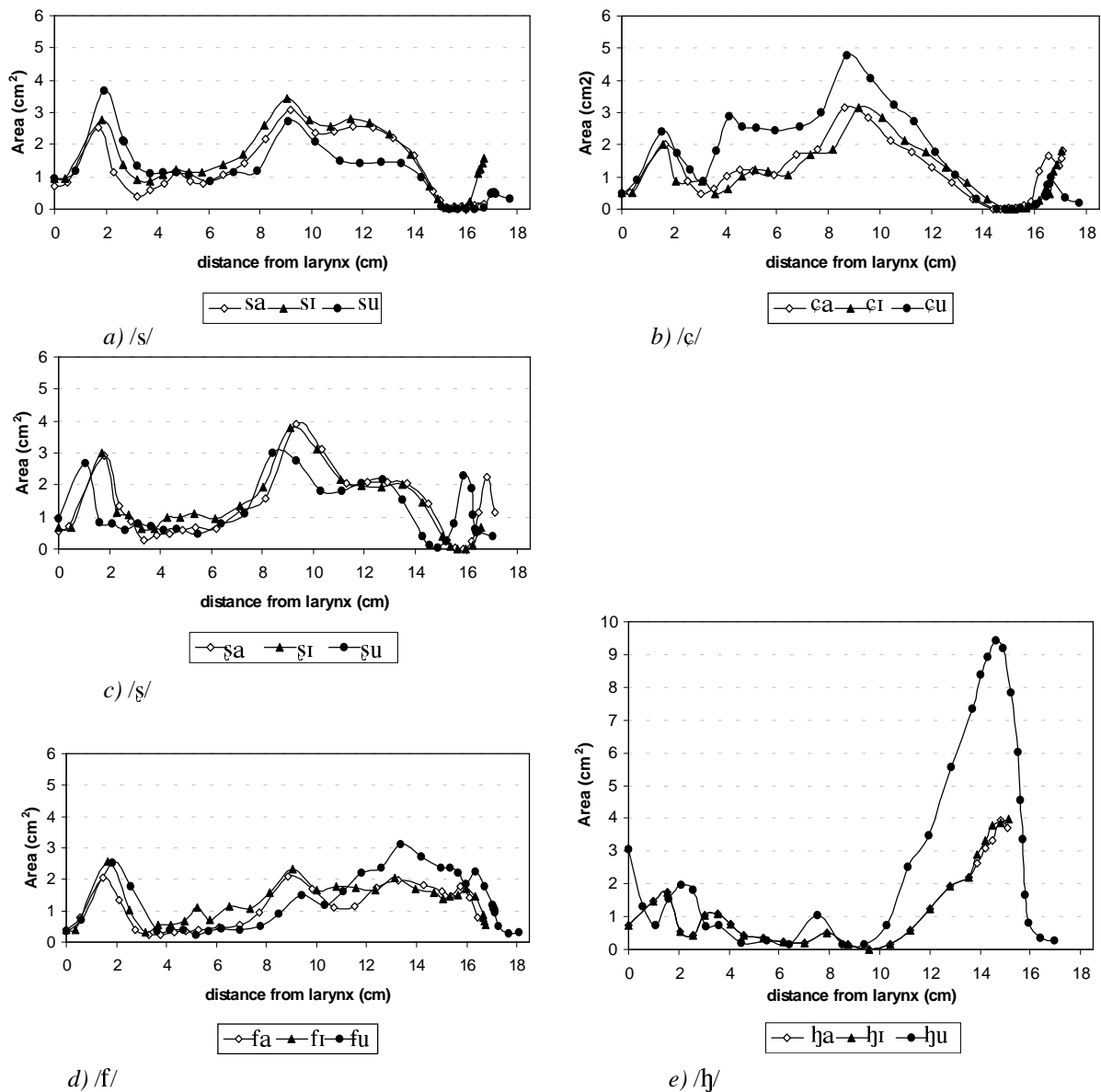


Figure 13. Area functions for the Swedish fricatives in the different vowel contexts /a/, /i/ and /u/.

that the error introduced by considering the section between adjacent points as a line is negligible.

The distance between a vocal tract contour and the next is set to the distance between their centroids. The result is an area function with varying distances between sampling planes, as exemplified by Figure 13, where the five Swedish fricatives are presented in three different vowel contexts. It should be stated that the purpose of this area function is to illustrate the variation of the cross-sectional area along the vocal tract and not to provide input for a speech synthesiser. In the latter case, an area function consisting of tubes with cross-sectional area equal to the average of two consecutive slices and two additional half tubes at the vocal tract ends, should be used instead.

The area functions of Figure 13 clearly indicate the difference in place of constriction and the large front cavity for /h/, especially in /u/ context. Narayanan et al. (1995) measured area functions for voiced and unvoiced English fricatives using MRI on four subjects. The results are similar, but more systematic comparisons are of little value because of the large inter-subject variations of that study.

### Measurement artefacts due to the method

It is clear that the subject could not produce the sounds in a natural way, due to the relatively long MRI acquisition duration, the supine position and the apnoea when producing the vowels.

The possible movements of the subject during image acquisition can induce image blurring (see comments on velum control below), and also in a discrepancy between the contours extracted from the three different stacks, as the stacks are not acquired simultaneously, but sequentially.

The supine position and the lack of phonation will cause the tongue body to move backwards, thus decreasing the passage in the pharynx. This is all the more the case for vowels, produced in apnoea, where no air pressure upstream the constriction (for fricatives) or the closure (for stops) helps the tongue to stay in place. The pharynx part of the reconstructed vocal tract is thus clearly too narrow for some vowels, as shown by the area functions in Figure 14.

The same, but milder, tendency for the lower part of the pharynx to have substantially

smaller areas than in Fant (1960) was found in an MRI study by Story et al. (1996).

The supine position also caused the head and thus the whole vocal tract to be inclined backwards, as can be seen in Figure 3. It is not clear to what extent this alters the vocal tract shape, mainly of the pharynx section, or the acoustic features, but the acoustic recordings suggest that this influence may be considered minor (Table 3-4).

There is also an artefact due to the lack of control over the velum during apnoea, causing the velum to be lowered, thus decreasing the passage in the velar area. This effect can be seen in Figure 3, where the movement of the velum is marked by the blurred outer contour (the sharper inner contour, corresponding to a static velum, was used for extracting the vocal tract contours in the right columns of Figure 15).

For a few sounds with extremely lowered larynx (particularly /h/), the vocal tract length is underestimated, as the lowering caused the lowest part of the pharynx to be below axial gridline 1. The entire vocal tract was however always covered by the midsagittal set (compare the two columns of Figure 15c-d).

Even with these artefacts in mind, the value of the study for three-dimensional modeling of the vocal tract is evident.

In order to evaluate the reliability of the 3D measurements, a vocal tract midsagittal outline was generated for every phoneme by determining the intersections between the plane contours extracted from the raw 3D MR images and the midsagittal plane (Figure 15, left). The comparison between the midsagittal contours from the 3D set and the corresponding contours extracted directly from the midsagittal image set (Figure 15, right) shows the overall coherence of the two sets of data. It also allows to detect the possible

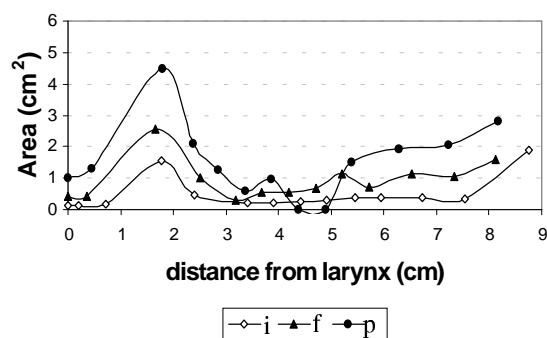


Figure 14. Measured area function for /i:/, /f/ and /p/ in the range  $d=0-9$  cm. Note the smaller areas for /i:/ compared to the fricative and plosive in corresponding vowel context.

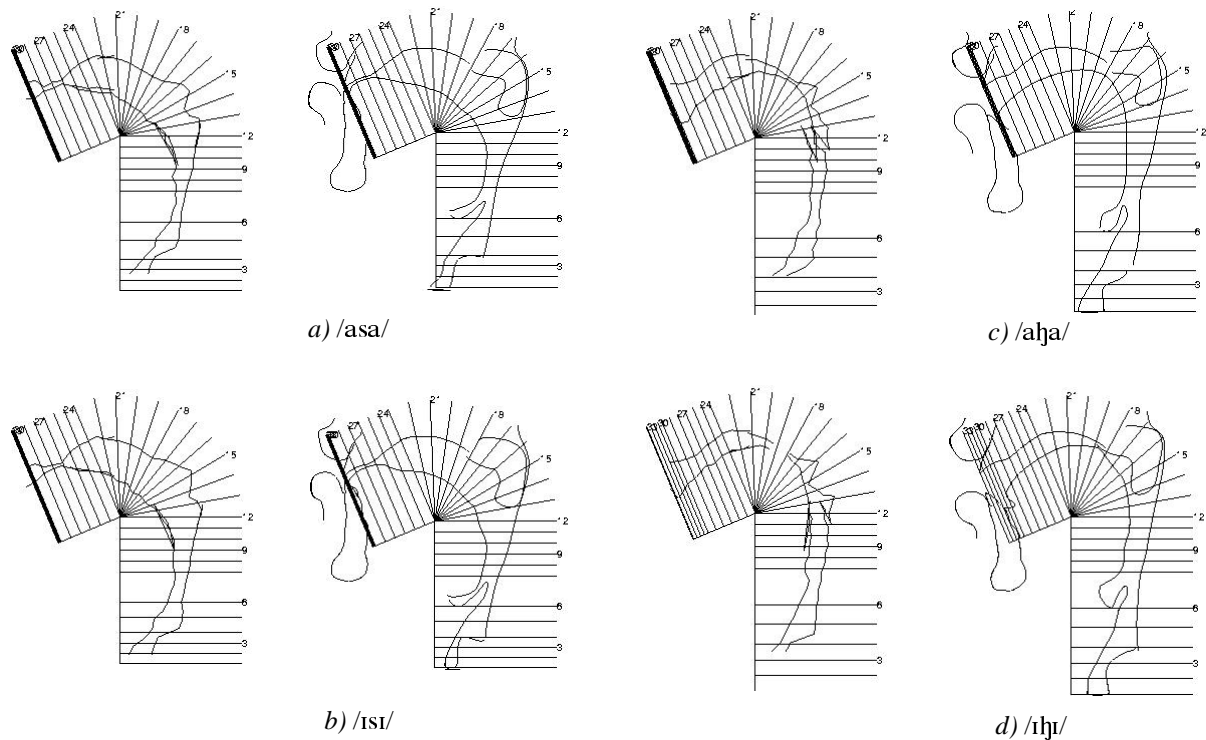


Figure 15. Midsagittal contours reconstructed from the 3D contours (left) and from the midsagittal set (right) for /s/ and /h/ in two different vowel contexts. The grid determined from the midsagittal image has been superimposed on the contour from the 3D set to facilitate comparisons. The midsagittal contour from the 3D set often has a substantially shorter larynx section because of measurement noise, thus reducing the actual vocal tract length. In the joints between the different 3D acquisition sets, the midsagittal contour is jagged, due to the overlapping of different sets. In some cases there are minor differences in contours acquired from different sets, indicating that the subject had not been able to sustain the vocal tract position exactly for the whole acquisition time.

discrepancies in the regions of partial overlap between the three stacks, and thus to assess the consistency between the three stacks, related to the subject's ability to sustain the articulation throughout the acquisition period. As can be seen from the joints in the left columns of Figure 15, there are minor differences in midsagittal contour from the different stacks, but there is an overall continuity.

Moreover it permits to compare the midsagittal outline to existing midsagittal articulatory models and thus evaluate the correspondence between midsagittal distance and the measured cross-sectional areas (cf. Badin et al., 1998b).

### Evaluation of the 3D reconstruction

The importance of acquiring the vocal tract three-dimensionally has been emphasised previously, but becomes even greater when considering co-articulation. As all consonants were acquired in VCV with the point vowels as vocalic context, the data set provides important information of co-articulatory

influence on vocal tract shape. This influence shows up to some extent in the area functions for fricatives described in the previous section, but is all the more clear when considering the vocal tract shape in itself. Figure 16 shows the context dependence of the vocal tract shape for /p/ at four planes in the oral cavity. The expected decrease of cross-sectional area in /ɪ/ context in comparison to /a/ and /u/ is realised both by tongue raising and increased grooving. The larger openness in /a/ context compared to /u/ is entirely due to tongue lowering in this case, without any flattening of the tongue. The context dependence of the vocal tract shape clearly illustrates the interest of a varied vowel context for consonant acquisition if a model based on this data should be able to replicate articulation of vowel-consonant series. Figure 16a) furthermore illustrates one important advantage of 3D measurements compared to midsagittal. Measuring the vocal tract shape only midsagittally would miss the co-articulatory effect almost entirely, as the midsagittal distance is almost the same for the three contexts, whereas the cross-sectional area

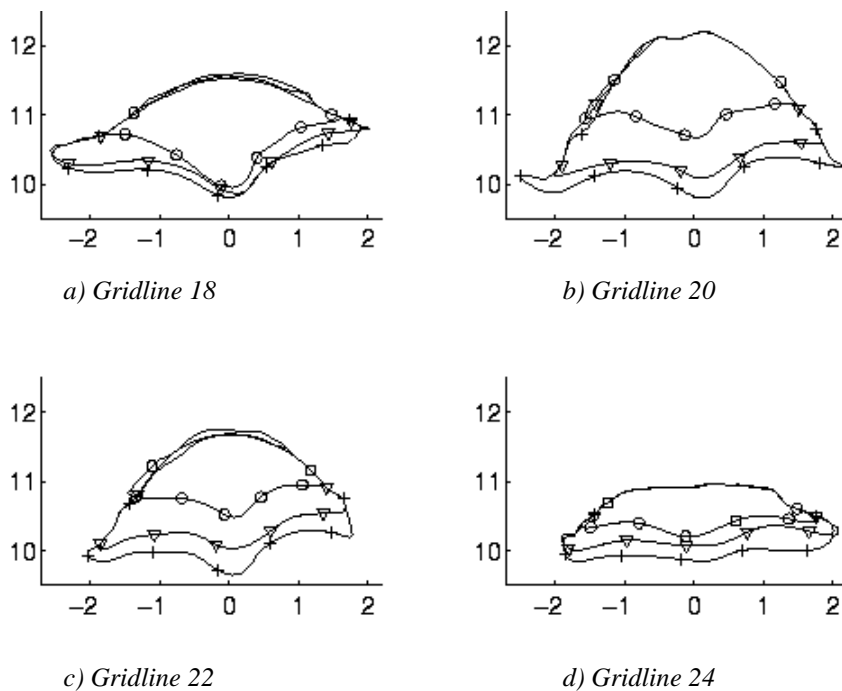


Figure 16. Coarticulatory influence on vocal tract shape for /p/. The effect of vowel context /a/ (+ marks), /i/ (o marks) and /u/ (∇ marks) for the oral cavity, seen from (-90,0).

is more than twice as large in /a/ context than in /i/ context.

A small study of five resynthesised vowels has been carried out to evaluate the ability to reproduce the target vowels from area functions given by the cross-sectional areas in the 3D set. In previous studies, vowels have been resynthesised directly from the area functions with varying success. Baer et al. (1991) were able to resynthesise some of the formants of the four vowels for their two subjects with acceptable error (11.6% overall mean error for F1), but failed for others (errors of 36% up to 71%). Formant frequency errors for American vowels in Story et al. (1996) ranged from 0.4% to 20%, with the majority slightly below 10%.

The above mentioned unnatural narrowing of the pharynx in this study for vowels lets suppose that vowels resynthesised directly from the area functions of the 3D set will not successfully reproduce the targets. To assess the quality of the 3D reconstructions and evaluate the importance of the narrowed pharynx, /ø:/, /i:/, /y:/, /u:/ and /ɑ:/ were synthesised from their area functions.

The area functions were generated considering the vocal tract as a number of uniform tubes of varying length, and the transfer functions were generated using a method for VT resonance frequency calculation proposed by Liljencrantz & Fant (1975). The target of the open vowel /ø:/ was

reached with small errors (Table 5). /i:/ and /y:/ should have a large back and a small anterior cavity, and can hence be expected to be affected more by the narrowed pharynx cavity (cf. Figure 17). The errors for /i:/ and /y:/ are in fact larger than for /ø:/, as indicated by Table 5, and in some instances higher than in Story et al. (1996). Considering that the sinus piriforms and sublingual cavities were ignored, they are however of quite acceptable order. The reconstructions from 3D data thus appear to be consistent with the recordings for these vowels. The underestimation of the vocal tract length due to measurement noise in the larynx region is, on the other hand, more critical for

Table 5. The first three formants of resynthesised vowels based on their area function calculated from the 3D reconstruction. The synthesised formants are compared to the subject's natural vowel frequencies recorded at the MRI session and the relative difference is indicated. The area function of /u:/ was adjusted by adding a larynx section compensating for underestimation of the vocal tract length due to measurement noise (cf. Figure 17).

	/ø:/	/i:/	/y:/	/u:/ <sup>adj</sup>	/ɑ:/
F1 (Hz)	596	436	423	361	648
Δ (%)	-0.7	3.8	11.3	-5.0	1.2
F2 (Hz)	1186	1473	1667	919	1058
Δ (%)	-5.9	-13.4	-3.0	27.6	8.0
F3 (Hz)	2543	3162	2776	2232	2380
Δ (%)	-6.5	16.3	2.0	-0.3	-14.4

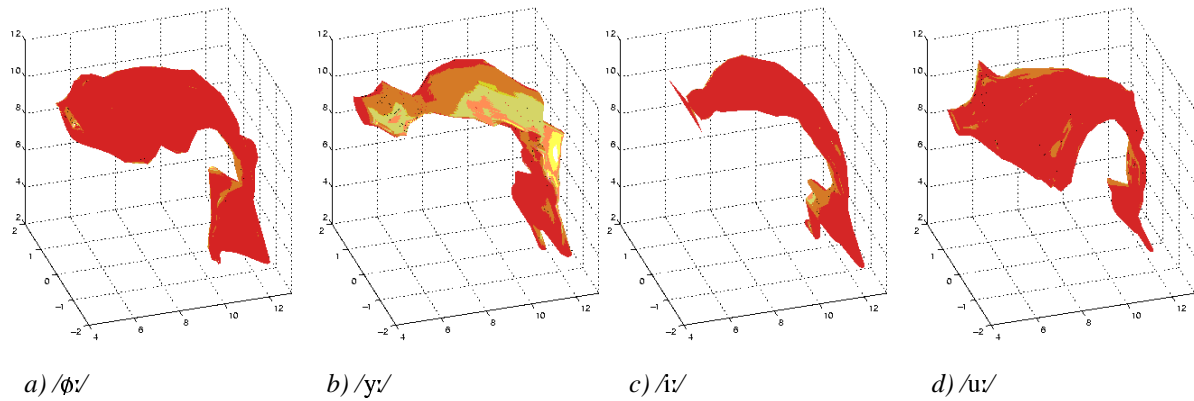


Figure 17. 3D reconstructions of four Swedish vowels. The vocal tract length of /u:/ is underestimated due to measurement noise in the larynx region, masking the air passage in the larynx completely.

/u:/, where the air passage in the larynx region is completely masked by noise. The vocal tract length for /u:/ is substantially shorter in the 3D reconstruction than in the midsagittal one, and even when compensating for this artefact by adding an extra larynx section to reach normal vocal tract length for /u:/, a large error remains in F2 (Table 5), whereas the targets for F1 and F3 are reached within a few percent. During reconstruction of /ɑ:/, three vocal tract contours had to be corrected manually at the velum, where the uncontrolled velum severely decreased the air passage. Artificially raising the velum to normal height in the reconstruction proved to be sufficient both for the vocal tract reconstruction and to resynthesise the vowel with less than 8% error in F1 and F2 (Table 5 and Figure 18), but with a larger remaining error in F3 (-14.4%).

The vowel space of the resynthesised vowels is decreased compared to the subject's natural vowel space, as indicated by Figure 18.

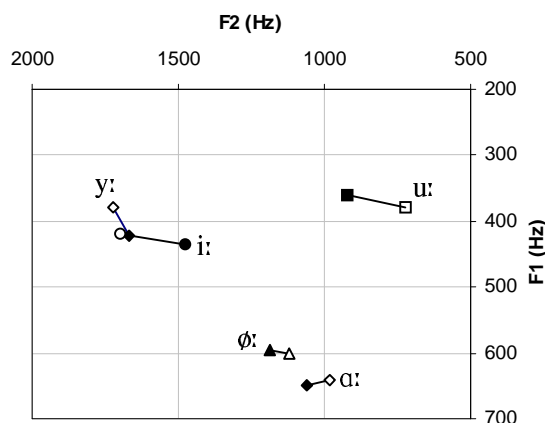


Figure 18. F1-F2 distribution and difference of natural and synthesised vowels of the four vowels in the evaluation test. Natural vowels are unfilled, synthesised filled.

All five resynthesised vowels are more central when considering the distribution of the first two formants of natural and synthesised vowel pairs. In the case of /i:/ vs. /y:/ there is an overlap of F1-F2 values for the natural and synthesised formants, but this ambiguity should however be resolved by the third formant, that is quite distinct for the two.

### Conclusions on the 3D data set

The evaluation of the 3D reconstructions shows that the data from the 3D set provide important information on co-articulation and that it furthermore is able to reach target frequencies for five analysed vowels with errors of the same order as earlier MRI studies of vowels (Baer et al., 1991, Story et al., 1996). The results of the preliminary resynthesis test further indicate that the results can be improved by combining the 3D reconstructions with data from the midsagittal set and other sources.

The 3D data set provides a unique mode of input to the three-dimensional vocal tract model as coherent data for all relevant modeled parts: vocal and nasal tract walls, tongue, teeth, velum and lips, has been collected simultaneously, for the same subject and for a large number of phonemes. The reconstruction representation, considering the organs as a mesh of data points further provides direct and complete input for sustained phonemes to every vertex of the model. Combined with real-time measures of speech dynamics, provided e.g. by ultrasound, video images and X-rays, the full 3D reconstruction should however be able to adequately reproduce dynamic vocal tract movements.

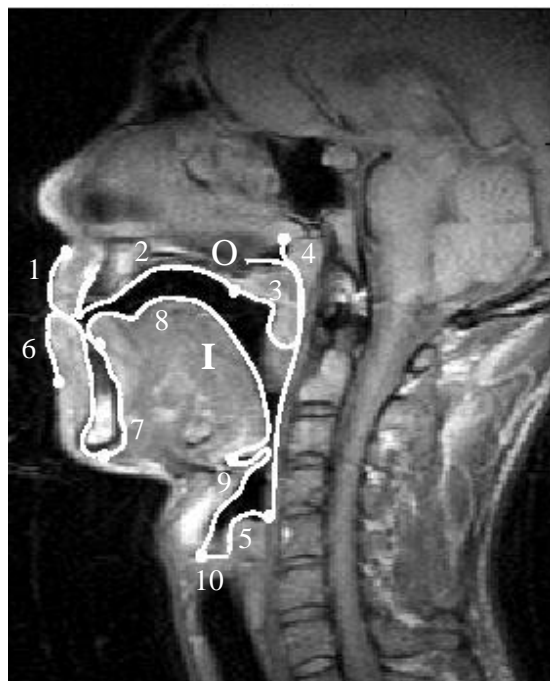


Figure 19. Midsagittal contours for /ɪpɪ/ extracted using automatic contour detection and hand editing: upper lip (1), palate (2), velum (3), pharynx (4), back larynx (5), lower lip (6), jaw (7), tongue (8), front larynx (9) and glottis (10).

## The midsagittal set

The entire midsagittal set, with exception of the retroflex phonemes /ʎ, ʟ, ɲ/, was analysed from the point of view of articulatory characteristics and measurements, and compared to other data available for Swedish. The articulatory degrees of freedom of the vocal tract midsagittal contours, and in particular of the tongue, were also analysed and represented in terms of linear articulatory modeling.

### Articulatory measurements

A vocal tract contour consisting of the 10 sub-contours *upper lip, palate, velum, pharynx, back larynx, lower lip, jaw, tongue, front larynx* and *glottis* was extracted from the midsagittal images using a Matlab version of the BtoC software (Maeda et al., 1993) adapted for MR images at ICP. The contours were generated by automatic contour detection in a thresholded image and subsequent hand editing. The palate and jaw were positioned by dragging and rotating the palate or jaw contours to fit the boundary in the MR image.

The result of contour extraction for /ɪpɪ/ is shown in Figure 19.

This section describes the articulatory measures that have been automatically extracted from the midsagittal contours, using the software adapted by Beautemps et al. (submitted) from Maeda et al. (1993). These articulatory measures, even if restricted to sustained articulations, provide a wealth of information on the articulatory control strategies used by the subject (i.e. coarticulation and synergy). Moreover, they will be used to develop a linear articulatory midsagittal model, as described below.

### The jaw

Jaw movements are analysed first, as the jaw influences the articulators attached to it, i.e. the lower lip and the tongue. This is the traditional approach (e.g. Lindblom & Sundberg, 1971), and different models have been proposed to account for jaw movements. Assuming that the movement is mainly restricted to the midsagittal plane for speech (Ostry & Vatikiotis-Bateson, 1995), the position of the jaw can be determined by a rotation with a flexible radius of rotation (Mermelstein, 1973), a two-dimensional translation or both (Edwards & Harris, 1990).

In the present analysis, the movement of the jaw is modelled as a two-dimensional translation, characterised by *JawHei*, the jaw

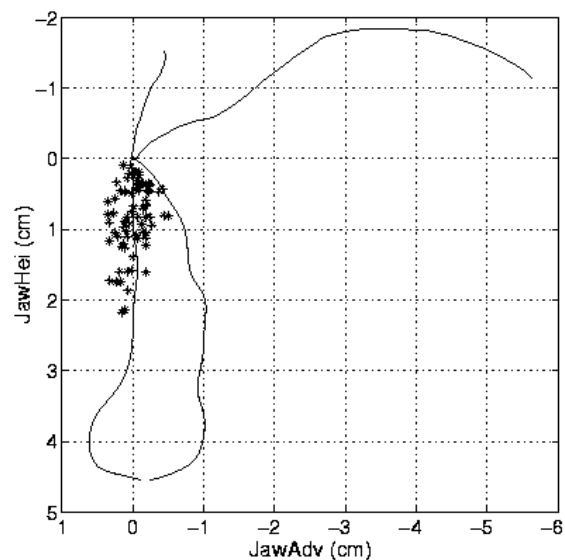


Figure 20. The jaw movement for the midsagittal corpus is shown as the dispersion of the lower incisor's upper edge. The contours of the hard palate and the lower incisor aligned in the occlusal plane are superposed for reference.

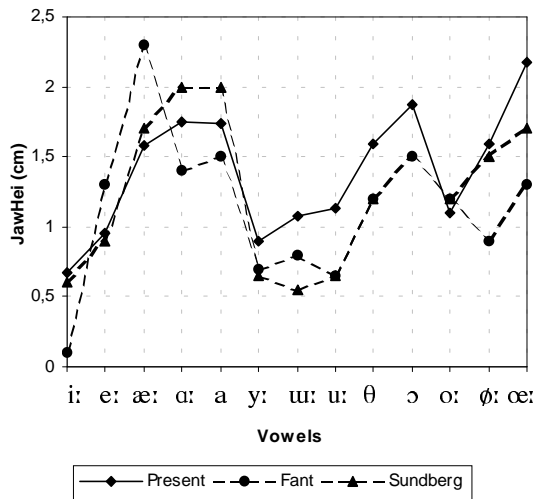


Figure 21. The jaw height measured from midsagittal images, compared with Fant (1993) and Sundberg & Lindblom (1971).

height, and JawAdv, the jaw advancing, as defined in Figure 7. Note that JawHei is measured as the vertical and JawAdv as the horizontal distance between the upper and lower incisors. Figure 20 shows the vertical (JawHei) and horizontal (JawAdv) components of the jaw position. The main jaw movement is explained by JawHei (standard deviation of 0.49) but JawAdv also contributes (standard deviation of 0.20). From Figure 20 and from the nomogram of JH in Figure 27 a small, but visible tendency for the subject to advance the jaw as it is lowered can be established (correlation 0.31, cf. Table 5), especially for the rounded open vowel /œ:/. The advancing can also be seen to some extent for rounded closed vowels, e.g. /u, u:, y/, where the jaw advances to contribute to the lip protrusion.

The jaw height measured for the vowels in this study is compared to values found by

Sundberg & Lindblom (1971) and Fant (1993) in Figure 21. The relative jaw height is similar in the three studies; the jaw height for unrounded vowels corresponding closely to values in the Sundberg & Lindblom study. The relative jaw height is quite similar also to values from Fant (1993), except for /æ:/ and /œ:/. The differences in jaw height between the two studies are likely to be a combination of inter-subject differences and sustained vs. dynamic articulation. The extremely open /œ:/ measured in this study is likely to be due to the latter effect.

The co-articulation effect on the jaw height for two groups of consonants, fricatives and plosives, is shown in Figure 22. Consonants with larger jaw heights are influenced more by co-articulatory effects and the overall relative jaw height is as expected from the vowel jaw height, i.e. largest for /ɔ/ and smallest for /i/. Some effects are more interesting, such as the jaw height for /ɔhɔ/ being even larger than for /ɔ/ and the effect of the protrusion of /u/ on the labiodental /f/, where the jaw height is decreased due to the protrusion.

The tongue

The tongue midsagittal contour (denoted I in Figure 19), covering a large part of the midsagittal vocal tract inner contour, can be represented in a sampled way by its intersections with the grid system depicted in Figure 6 and described above. Retroflex articulation is a special case that can not be dealt with using this representation and the retroflexes were hence excluded from this part of the study.

The tongue body as a whole is complex and no standard measures determine the entire contour. The most frequent is to indicate tongue body movements in the anterior-

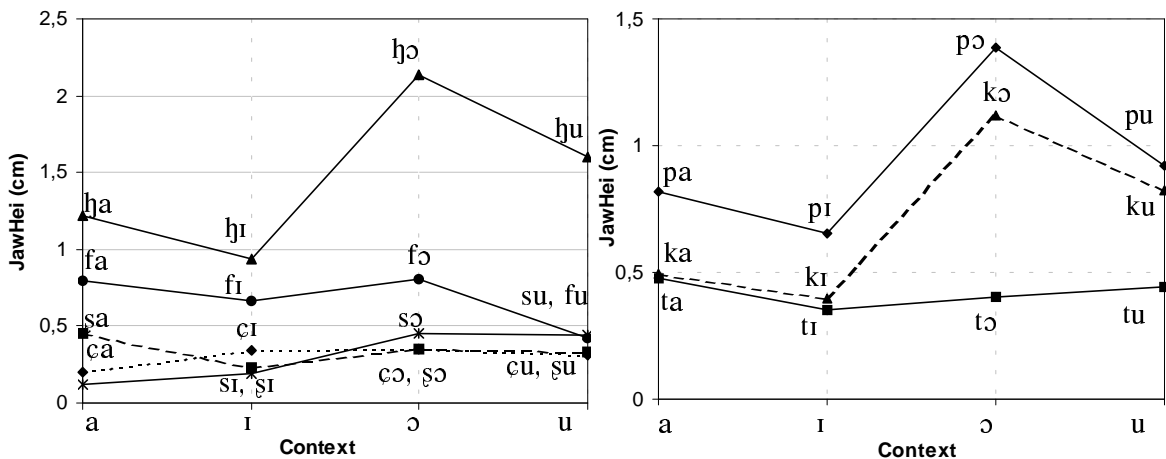


Figure 22. Context dependence of jaw height for fricatives (left) and stops (right).





Table 5: Correlation coefficients of measured articulatory parameters.

Correlations mentioned in the text are shown in bold.

	JawHei	JawAdv	LipHei	LipTop	ProTop	ProBot	TngTip	TngAdv	TngBot	LarHei
JawHei	1.0000	<b>0.3063</b>	0.3294	-0.4048	0.1389	0.0205	<b>-0.8593</b>	-0.0342	0.4592	0.2759
JawAdv	<b>0.3063</b>	1.0000	0.1639	-0.0857	0.0833	-0.1402	-0.1782	0.2117	-0.0218	-0.1748
LipHei	0.3294	0.1639	1.0000	<b>0.5536</b>	0.1097	0.0338	-0.2572	-0.0541	-0.0348	-0.0294
LipTop	-0.4048	-0.0857	<b>0.5536</b>	1.0000	-0.1472	-0.0482	0.3999	-0.0316	-0.2403	-0.2605
ProTop	0.1389	0.0833	0.1097	-0.1472	1.0000	<b>0.9203</b>	-0.1538	-0.1172	0.1879	<b>0.4078</b>
ProBot	0.0205	-0.1402	0.0338	-0.0482	<b>0.9203</b>	1.0000	-0.0130	-0.1340	0.2172	<b>0.3738</b>
TngTip	<b>-0.8593</b>	-0.1782	-0.2572	0.3999	-0.1538	-0.0130	1.0000	0.1448	-0.4376	-0.3656
TngAdv	-0.0342	0.2117	-0.0541	-0.0316	-0.1172	-0.1340	0.1448	1.0000	-0.0506	-0.2777
TngBot	0.4592	-0.0218	-0.0348	-0.2403	0.1879	0.2172	-0.4376	-0.0506	1.0000	<b>0.6318</b>
LarHei	0.2759	-0.1748	-0.0294	-0.2605	<b>0.4078</b>	<b>0.3738</b>	-0.3656	-0.2777	<b>0.6318</b>	1.0000

between larynx height, LarHei, and protrusion, ProTop (0.41) and ProBot (0.37), can not be explained bio-mechanically. It should likely be ascribed to the common strategy of articulatory synergy that consists in maximising total vocal tract length by lowering the larynx in rounded vowels, as evidenced by Hoole & Kroos (1998) for German, and in accordance with Figures 23-24.

### The linear midsagittal articulatory model

As pointed out by Vilain et al. (1998), linear articulatory models constitute useful tools for the understanding of coarticulatory strategies in speech production. The statistical analysis proposed by Harshman et al. (1977), or Maeda (1979), and recently refined by Beautemps et al. (submitted), was performed on the vocal tract geometry characterised by a constant number of points on the inner (I in Figure 19) and outer (O in Figure 19) contours. As already mentioned, the points are determined with a partially dynamical semi-polar intersection grid (cf. Figure 15, right) generated from the standard measures of TngAdv, TngTip, LarHei, LarTop, TngBot and VelHei (refer to Figure 7 for definitions), found from the contours.

The model's articulatory variables are dimensionless, centred and normalised. They are extracted through guided factor analysis (Maeda, 1990), consisting of an iterative subtraction of linear predictors of the vocal tract geometry in the midsagittal plane and classical Principal Component Analysis (PCA) applied to selected regions of the vocal tract contour (Beautemps et al., submitted).

### The Jaw: JH

Two parameters for the jaw height and the jaw advancing would uniquely model jaw movement. As pointed out above, the main jaw movement is described by the two parameters JawHei and JawAdv. However, when using guided component analysis, the contribution of JawAdv is negligible (explaining only 4.1% of the overall variation of the inner contour if applied as first factor in the PCA), except for gridlines 15-20, where it is small, but reaches 6.1%. In the following linear component analysis, JawAdv is thus not considered and the jaw movements are modeled with one degree of freedom, in agreement with an earlier study of three French subjects (Bailly et al., 1998). The predictor for the jaw movement, JH, is hence centred on the mean of the variable JawHei and normalised by the variable standard deviation. Its important influence on the anterior part of the vocal tract inner contour is shown in Figure 28, JH.

### The Tongue: TB, TD, TT, TA

The tongue movements are defined as the separate displacement of the tongue not induced by jaw movements. In reality, the active tongue movements are hard to separate from passive displacement due to jaw movements, as the tongue and the jaw are coupled very closely by their muscles (Perkell, 1974 & 1996). In addition, subjects tend to use articulatory control strategies where the tongue body and the jaw are actively moved in synergy, causing the tongue movements to be substantially larger than the associated jaw movement (Bailly et al., 1998). This possible synergy effect has however not been explicitly separated out in this study, and the tongue

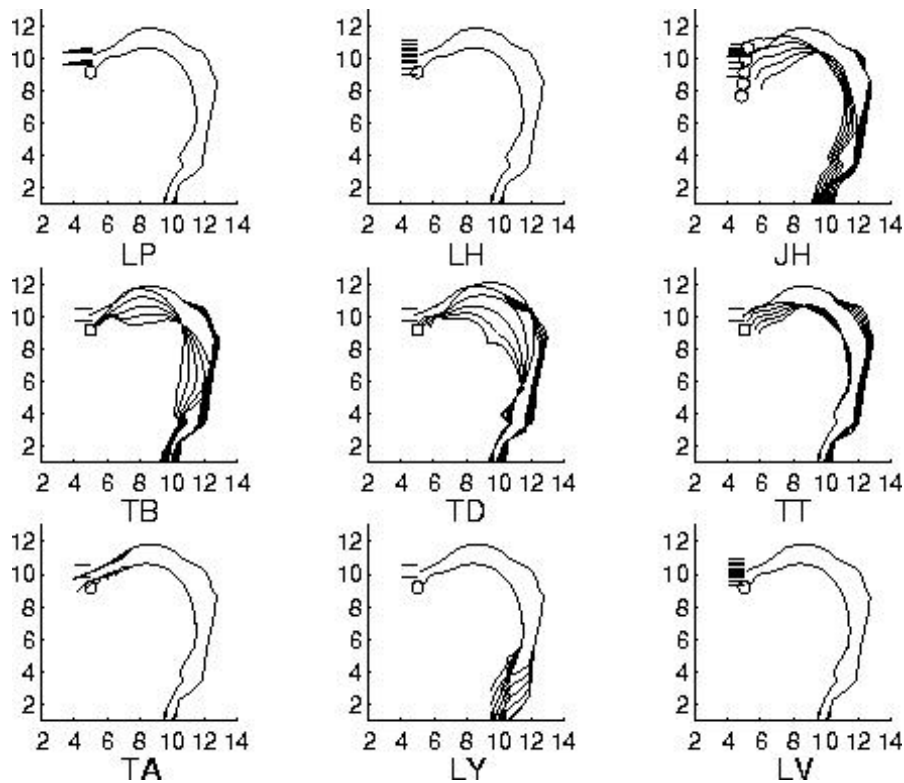


Figure 28. Articulatory nomograms. The variations of the midsagittal contours due to articulatory parameter variation from  $-3$  to  $+3$  with stepsize  $+1.5$ . LP = Lip Protrusion, LH = Lip Height, JH = Jaw Height, TB = Tongue Body, TD = Tongue Dorsum, TT = Tongue Tip, TA = Tongue Advance, LY = Larynx and LV = Lip vertical position.

contour thus has the parameter JH as first linear predictor. The influence of the jaw on the tongue shape is removed by considering the residual tongue shape, calculated as the difference between measured shape and that predicted by JH.

PCA is first applied to the residual of the tongue body (gridlines 7-24) excluding the tip (lines 24-27). Two predictors can be extracted: the tongue body predictor, TB, explains the front-back movement, while the tongue dorsum factor, TD, explains the flattening-arching movement. Their effects on the tongue shape can be observed in Figure 28, plots TB and TD, respectively. The factors TB and TD have been found as the projections of the centred and normalised residuals on the two principal eigenvectors of the cross-correlation matrix for the tongue body residual. The two principal eigenvectors are the highest two eigenvectors of this matrix.

The tongue tip region is predicted, in addition to the two factors TB and TD, by an extra predictor, TT, determined as the first factor found from the PCA of the residual of the tongue tip contour (gridlines 24-27) after removing JH, TB and TD. TT takes into account the possibility of raising and lowering

the tongue tip relatively to the tongue body, as can be seen in Figure 28, TT.

The tongue advance predictor, TA, is finally defined as the centred and normalised residual of the measured value for TngAdv (cf. Figures 7 and 20) when the contributions of JH, TB and TD have been removed.

#### *The Lips: LP, LH, LV*

Since upper and lower lip protrusions are very highly correlated (0.92, cf. Table 5), lip protrusion is modelled by one parameter LP only, calculated as the centred and normalised value of ProTop with the contribution of JH removed. Centring and normalising the value of LipHei after removing the JH contribution determines the factor LH. Finally, the effect of JH, LH and LP is removed from LipTop, and the centred and normalised residual determines LV, the vertical lip position. LV moves the upper and lower lips in conjunction vertically with respect to the upper incisor.

#### *The Larynx, LY*

The centred and normalised value for the larynx height, LarHei, provides the LY predictor. LY controls the laryngeal part of the

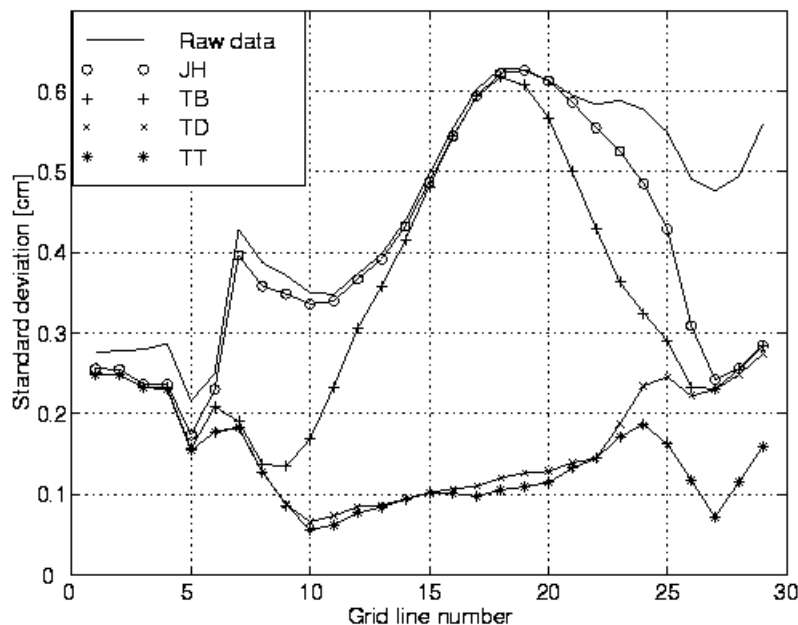


Figure 29. Standard deviation of the vocal tract inner contour (in cm) as a function of gridline number, when the contribution of the articulatory parameters is removed successively in the order JH, TB, TD and TT. The large standard deviation at the larynx is due to measurement noise in that region.

Table 7. Parameter influence on the midsagittal vocal tract contour. More than 90% of the contour variation can be explained with the four parameters for Jaw Height, Tongue Body, Tongue Dorsum and Tongue Tip.

Parameter	Explained variance
JH	20.05%
TB	20.47%
TD	45.66%
TT	4.10%
<b>Total</b>	<b>90.29%</b>

vocal tract contour by a vertical translation, as shown in Figure 28, LY.

### Evaluation of the midsagittal model

To evaluate the model and the relative contribution of each parameter, the variance explained by the predictors JH, TB, TD and TT was determined. Figure 29 shows the standard deviation of the inner contour of the vocal tract when the effect of each parameter is iteratively removed.

Over 90% of the variation is explained by these four factors, as shown in Table 7. This is superior to the value (88%) found by Beautemps et al. (1996) for the first *five* factors for their reference subject, but less than the 96% found in the more recent X-ray study (Beautemps et al., submitted). Large residual standard deviations remain in the larynx section, and the residual is larger by a factor of two compared to those found by Bailly et al. (1998) for three subjects examined with X-rays. The comparison of the present results with those of Bailly et al. (1998) shows significantly different contributions of the articulatory parameters. There is clearly an inter-subject variability, but there are indications that the variations are mainly caused by the difference between sustained articulations and normal speech. The most important dissimilarities are the lack of

influence of JH on the back region and the importance of TD compared to TB for tongue movements.

JH and TT explain much of the variations of tongue contour from gridline 20 and onwards. TB is the main contributor to the variations in the pharynx at gridlines 5-12 and the deviations from neutral shape in the central part of the tongue, gridlines 10-25, are almost exclusively due to TD.

### Correlations between articulatory parameters

The control parameters of the midsagittal model, JH, TB, TD, TT, TA, LY, LP, LH and LV, are not all orthogonal to each other, as pointed out in the correlation matrix of Table 8. These correlations are due to subject and corpus related control strategies. The most important correlations are found for LY: positive correlations with JH (0.28) and LP (0.37), and negative ones with TT (-0.24) and TA (-0.29). The effect of each control parameter is shown in the articulatory nomograms of Figure 28, generated by varying the control parameter from -3 to +3 in steps of +1.5.

The nomogram for JH in Figure 28 illustrates the coupling between jaw advancing and jaw lowering mentioned above. The

Table 8. Correlation coefficients of the articulatory control parameters of the midsagittal model. In bold: coefficients cited in the text. Empty slots indicate that the parameters are orthogonal.

	<b>JH</b>	<b>TB</b>	<b>TD</b>	<b>TT</b>	<b>TA</b>	<b>LY</b>	<b>LP</b>	<b>LH</b>	<b>LV</b>
<b>JH</b>	1.000					0.2759			
<b>TB</b>		1.000				-0.0223	-0.1358	-0.3230	-0.1529
<b>TD</b>			1.000			0.1239	0.0382	-0.0714	-0.0590
<b>TT</b>				1.000		-0.2370	-0.0601	0.0893	0.0786
<b>TA</b>					1.000	-0.2885	-0.1480	-0.1032	-0.0803
<b>LY</b>	0.2759	-0.0223	0.1239	-0.2370	-0.2885	1.000	0.3731	-0.1274	-0.0034
<b>LP</b>		-0.1358	0.0382	-0.0601	-0.1480	0.3731	1.000	0.0684	
<b>LH</b>		-0.3230	-0.0714	0.0893	-0.1032	-0.1274	0.0684	1.000	
<b>LV</b>		-0.1529	-0.0590	0.0786	-0.0803	-0.0034			1.000

comparison of this nomogram with those in earlier studies (Boë & Perrier, 1994; Beautemps et al., submitted) indicates that the influence of JH on the tongue contour in the pharynx region (gridline 7-13) is smaller but similar for the subject in the first and much less than for the subject in the latter.

The pivot point of TB (i.e. the point where the influence of TB on the tongue shape changes sign) is substantially retracted, and those of TD lowered (first) and advanced (second) compared to Boë & Perrier (1994) and to Beautemps et al. (submitted). This results in a large variation of TD in the velar region, and in a very small one in the apical region, leaving tongue tip movements to be modelled almost exclusively by TT. These differences can be ascribed to a number of factors: inter-subject variability, sustained vs. normal articulations, language and distribution of the corpus (the Swedish corpus containing significantly more back articulations). Reducing the corpus to the 26 corresponding configurations in the corpus of Beautemps et al., (submitted)<sup>9</sup>, does have the effects of approaching the pivot points of TD and TB to those of Beautemps et al., (submitted), but the inter-subject difference is still retained.

## Conclusion and perspectives

This study is, to the authors' knowledge, the first MRI study that has been conducted on Swedish. The studies of Foldvik et al. (1988, 1993, 1995) of Norwegian phonemes are the only previous studies of a closely related language. As pointed out in the introduction, the studies of Foldvik et al. were however

limited to the midsagittal plane (1988) or to repetitions of one vowel sequence (1993 and 1995). The midsagittal and full three-dimensional MRI data that have been acquired in this study for one subject uttering sustained vowels and consonants in VCV contexts thus offers a unique database for volumetric evaluations of vocal tract and articulator shapes for Swedish. This is especially true as the corpus was substantially larger than in earlier studies realised for other languages cited in this article.

Reconstructions have been made for both sets of midsagittal and 3D images using the same co-ordinate grid system aligned with the subject's hard palate, assuring convergence of the reconstructions. Data from the midsagittal set has then been used to describe and analyse some articulatory control strategies and coarticulation effects. Moreover, a linear midsagittal articulatory model has been created, and compared to other models developed according to the same method for other subjects.

Note that the subject was not chosen based on any proof of being a typical speaker of standard Swedish, and that it is unavoidable that a model based on one single speaker will be influenced by subject-specific control strategies. However, even though some of the subject-specific articulations might prove to be atypical, the present database constitutes a wealth of information on Swedish phonemes articulation. Further work with more subjects or with measures of dynamic speech for the same subject would naturally increase the interest of the present initial study.

As mentioned in the section on measurement artefacts, the pharynx cavity was very likely narrower than natural for vowels and this should be taken into account when using the data in the model. This unnatural

<sup>9</sup> Vowels: /a, æ, e, i, o, ɔ, u, y/

VCV: /apa, ipi, upu, ata, iti, utu, aka, iki, uku/  
/afa, ifi, ufu, asa, isi, usu, ala, ili, ulu/

narrowing seems to have been avoided in the midsagittal vowel set: a possible way of modeling the pharynx for vowels would thus be to combine tongue position for the vowels extracted from the midsagittal contour with pharynx properties found for consonants.

The evaluation of the midsagittal set proves that methods used for evaluation of dynamic speech captured with X-rays can be used for midsagittal MR data. Parameter values extracted from the midsagittal set can thus provide a basis for the three-dimensional modeling based on the 3D set. The results of the 2D analysis will be of great importance in the subsequent analysis of the 3D set and the three-dimensional modeling.

The present MRI study was first motivated as a means of providing data for both articulator shapes and parameter values for the existing KTH 3D model described in Engwall (1999a). On-going work concentrates on incorporating extracted vocal tract contours and parameter modeling in this 3D model.

Contours of the tongue have also been determined from the set of 3D images, and three-dimensional reconstruction of the tongue for the corpus is under way. Three-dimensional reconstruction of the tongue from MR images have previously been done for English fricatives (Narayanan et al., 1995), laterals (Narayanan et al., 1997) and rhotics (Alwan et al., 1997). The reconstructed tongue should then be incorporated in the 3D model and its parameters adapted to accord with the tongue shape measured from MR images. Massaro et al. (1998) used error minimisation to fit the tongue model of their talking head to articulatory data obtained from ultrasound and EPG (Stone & Lundberg, 1996). This method could be used in the future for the KTH model as well.

Finally, guided PCA could be applied to the 3D-tongue data to extract linear predictors, as has already been successfully attempted for three-dimensional vocal tract shapes (Badin et al., 1998b), as well as for the lips and the face (Borel, 1999, Badin et al., submitted), proving that it is a powerful alternative.

## Acknowledgement

This work is the result of a collaboration between the Centre for Speech Technology in Stockholm (CTT, sponsored by KTH, Nutek and Swedish industry), and the Institut de la Communication Parlée in Grenoble (ICP). A large part of the work was carried out at ICP during a stay of the first author, made possible

by grants from Telefonaktiebolaget LM Ericssons stiftelse för främjande av elektroteknisk forskning, Ragnar & Astrid Signeuls fond at KTH and the Swedish Institute.

The MRI data acquisition, realised at the Grenoble University Hospital (CHRUG), was supported by the French ARASSH (Association Rhône-Alpes pour les Sciences Humaines et Sociales).

The authors are very much indebted to Christoph Segebarth, INSERM Unit U438, Grenoble, France, for his help with the MRI acquisition.

## References

- Alwan A, Narayanan S, Haker K (1997). Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part II. The rhotics, *J Acoust Soc Amer*, 101: 1078-1089.
- Apostol L, Perrier P, Raybaudi M, Segerbarth C (1999). 3D geometry of the vocal tract and interspeaker variability. *Proc of ICPH'S'99*, 443-446.
- Badin P, Borel P, Bailly G, Revéret L, Raybaudi M, Segebarth C (submitted). Towards an Audio-visual Virtual Talking Head: 3D linear articulatory modelling of tongue, lips and face based on MRI and video images.
- Badin P, Baricchi E, Vilain A (1997). Determining tongue articulation: from discrete fleshpoints to continuous shadow, *Proc Eurospeech'97* 1: 47-50.
- Badin P, Bailly G, Boë L-J (1998a). Towards the use of a Virtual Talking Head and of Speech Mapping tools for pronunciation training, *Proc ETRW on Speech Technology in Language Learning*, 167-170.
- Badin P, Bailly G, Raybaudi M, Segebarth C (1998b). A three-dimensional linear articulatory model based on MRI data, *Proc 3<sup>rd</sup> ESCA/COCOSDA Intl Workshop on Speech Synthesis*, 249-254.
- Bailly G, Badin P, Vilain A (1998). Synergy between jaw and lips/tongue movements: consequences in articulatory modelling, *Proc ICSLP'94*, 5: 1859-1862.
- Baer T, Gore JC, Gracco LW, Nye PW (1991). Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels, *J Acoust Soc Amer*, 90: 799-828.
- Beautemps D, Badin P, Laboissière R (1995). Deriving vocal-tract functions from midsagittal profiles and formant frequencies: A new model for vowels and fricative consonants based on experimental data, *Speech Communication* 16: 27-47.
- Beautemps D, Badin P, Bailly G, Galván A, Laboissière R (1996). Evaluation of an articulatory-acoustic model based on a reference subject, *Proc 1<sup>st</sup> ESCA Tutorial and Research Workshop on Speech Production Modeling - 4<sup>th</sup> Speech Production Seminar*, 45-48.

- Beautemps D, Badin P, Bailly G (submitted). Degrees of freedom in speech production: analysis of cineradio- and labio-films data for a reference subject, and articulatory-acoustic modeling. Submitted.
- Beskow J (1995). Rule-based visual speech synthesis, *Proc Eurospeech '95*, 299-302.
- Boë L-J, Gabioud B, Schwartz J-L, Vallée N (1995). Towards the unification of vowel spaces. *Proc ICPhS'95* 4: 582-585.
- Boë L-J, Perrier P (1994). Simulations of the Macro-Variations of the Vocal-Tract Plant. In *From speech signal to vocal tract geometry, Speech Maps, Year 2 Report, Vol. III*.
- Boersma P (1998). Functional Phonology, LOT International Series 11, The Hague: Holland Academic Graphics. Pages i-ix, 1-493. *Doctoral thesis*, University of Amsterdam.
- Borel P (1999). Modélisation articulatoire linéaire d'un visage incluant des lèvres. Unpublished DEA report. Institut National Polytechnique de Grenoble.
- Branderud P, Lundberg HJ, Lander J, Djamshidpey, Wäneland I, Krull D, Lindblom B (1998). X-ray analyses of speech. Methodological aspects, *Proc Phonetik'98* 168-171.
- Cohen M, Beskow J, Massaro D (1998). Recent development in facial animation: An inside view, *Proc AVSP '98*, 201-206.
- Demolin D, Metens T, Soquet A (1996). Three-dimensional measurements of the vocal tract by MRI, *Proc ICSLP'96*, 1:272-275.
- Edwards J, Harris KS (1990). Rotation and translation of the jaw during speech, *J Speech Hear Res*, 33:550-562.
- Elert C-C (1989). *Allmän och svensk fonetik*, 30. Nordsteds förlag.
- Engwall O (1999a). Modeling the vocal tract in 3D, *KTH STL-QPSR* 1-2: 31-38.
- Engwall, O. (1999b). Modeling of the vocal tract in three dimensions, *Eurospeech '99*, 1: 113-116.
- Fant G (1959). Acoustic Analysis and Synthesis of Speech with Applications to Swedish, *Ericsson Technics* 15: 3-108.
- Fant G (1960). *The Acoustic Theory of Speech Production*. Mouton, The Hague.
- Fant, G (1969). Formant Frequencies of Swedish Vowels, *KTH-STL-OPSR* 4: 26-31.
- Fant G (1964). Formants and cavities, *Proc ICPhS'64*, 120-141.
- Fant G (1993). "A New Three-parameter Model of VT Area Functions". In *From speech signal to vocal tract geometry, Speech Maps, Year 1 Report, Vol. III*.
- Fant G, Båvegård M (1997). "Parametric model of the vocal tract area function: Vowels and consonants", *KTH-TMH-QPSR* 1: 1-20.
- Foldvik AK, Husby O, Kvaerness J (1988). Magnetic Resonance Imaging, *Proc 7th FASE Symposium*, 423-428.
- Foldvik AK, Kristiansen U, Kvaerness J (1993). A time-evolving three-dimensional vocal tract model by means of magnetic resonance imaging (MRI), *Proc Eurospeech'93*, 557-558.
- Foldvik AK, Kristiansen U, Kvaerness J (1993). Three-dimensional ultrasound and magnetic resonance imaging: A new dimension in phonetic research, *Proc ICPhS'95*, 4:46.
- Goldman R (1991). Area of planar polygons and volume of polyhedra. In: *Graphic Gems II*, 170-171, Academic Press.
- Harshman R, Ladefoged P, Goldstein L (1977). Factor analysis of tongue shape, *J Acoust Soc Amer*, 62: 693-707.
- Heinz JM, Stevens KN (1964). "On the derivation of area functions and acoustic spectra from cineradiographic films of speech", *J Acoust Soc Amer*, 36: 1037 (abs).
- Hoole P, Kroos C (1998). Control of larynx height in vowel production, *Proc ICSLP'98* 2: 531-534.
- Kiritani S, Sekimoto S, Imagawa H, Fujisaki H (1977). Parameter description of the tongue movements for the vowels, *Contribution papers of 9<sup>th</sup> ICA* 1: I13, 419.
- Ladefoged P, Anthony JFK, Riley C (1971). Direct measurement of the vocal tract, *UCLA Working Papers in Phonetics* 19: 4-13.
- Liljencrantz J, Fant G (1975). Computer program for VT-resonance frequency calculations, *KTH STL-QPST* 4:15-20.
- Lindblom B, Sundberg J (1971). Acoustical consequences of lip, tongue and jaw movements, *J Acoust Soc Amer*, 50: 1166-1179.
- Maeda S (1979). An articulatory model of the tongue based on a statistical analysis, *J Acoust Soc Amer*, 65: S22 (abs).
- Maeda S (1988). Improved articulatory models, *J Acoust Soc Amer*, 84: S146 (abs).
- Maeda, S (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In Hardcastle WJ, Marchal A (ed.), *Speech Production and Speech Modelling*, 131-149. Kluwer Academic publishers.
- Maeda S, Zerling JP, Simon P, Bothorel A, Wioland F (1993). Software for the digitalisation of labial and X-ray film data. In *From speech signal to vocal tract geometry, Speech Maps, Year 1 Report, Vol. III*.
- Mair S, Scully C, Shadle C (1996). Distinction between [t] and [tʃ] using electropalatography data, *Proc ICSLP'96*, 1597-1600.
- 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, *Proc ICSLP'94*, S12:13.1-13.4.
- Mermelstein P (1973). Articulatory model of speech production, *J Acoust Soc Amer*, 53: 1070-1082.
- Mohammad M, Moore E, Carter NJ, Shadle CH, Gunn SJ (1997). Using MRI to image the moving vocal tract during speech. *Proc Eurospeech'97*, 4:2027-2030.
- Narayanan S, Alwan A, Haker K (1995). An articulatory study of fricative consonants using magnetic resonance imaging, *J Acoust Soc Amer*, 98: 1325-1347.

- Narayanan S, Alwan A, Haker K (1997). Toward articulatory-acoustic models for liquid approximants based on MRI and EPG data. Part I. The laterals, *J Acoust Soc Amer*, 101: 1064-1077.
- Nguyen N, Marchal A, Content A (1996). Modeling tongue-palate contact patterns in the production of speech. *J of Phonetics* 24:77-97.
- Ostry D, Vatikiotis-Bateson E (1995). An analysis of the dimensionality of the jaw motion in speech, *J Phonetics*, 23: 101-117.
- Perkell J (1974). A physiologically-oriented model of tongue activity in speech production", *PhD thesis*, MIT.
- Perkell J (1996). Properties of the tongue help to define vowel categories: hypotheses based on physiologically-oriented modeling. *J of Phonetics* 24:3-22.
- Rubin P, Baer T, Mermelstein P (1981). An articulatory synthesizer for perceptual research, *J Acoust Soc Amer*, 70: 321-329.
- Sjölander K, Beskow J, Gustafson J, Lewin E, Carlson R, Granström B (1998). Web-based Educational Tools for Speech Technology, *Proc of ICSLP98*, 7:3217-3220.
- Stark J, Lindblom B, Sundberg J (1996). APEX: an articulatory synthesis model for experimental and computational studies of speech production, *KTH STL-QPSR* 2: 45-48.
- Stone M (1990). A three-dimensional model of tongue movement based on ultrasound and x-ray microbeam data, *J Acoust Soc Amer*, 87: 2207-2217.
- Stone M, Lundberg A (1996). Three-dimensional tongue surface shapes of English consonants and vowels, *J Acoust Soc Amer*, 99: 3728-3737.
- Story B, Titze IR, Hoffman E (1996). Vocal tract area functions from magnetic resonance imaging, *J Acoust Soc Amer*, 100: 537-554.
- Sundberg J (1969). On the problem of obtaining area functions from lateral x-ray pictures of the vocal tract, *KTH STL-QPSR* 1: 43-45.
- Sundberg J, Johansson C, Wilbrand H, Ytterbergh C (1987). From sagittal distance to area: A study of transverse, vocal tract cross-sectional area, *Phonetica*, 44: 76-90.
- Tiede M (1996). An MRI-based study of pharyngeal volume contrasts in Akan and English, *J Phonetics*, 24: 399-421.
- Vilain A, Abry C, Badin P (1998). Coarticulation and degrees of freedom in the elaboration of a new articulatory plant: Gentiane. *Proc of ICSLP98*, 7: 3147-3150.
- Wakumoto M, Masaki S, Honda K, Dang J (1996). Visualization of dental crown shape for MRI., *ATR Research Reports* 1996: 39.
- Westbury JR (1994). On coordinate systems and the representation of articulatory movements, *J Acoust Soc Amer*, 95: 2271-2273.
- Yang C-S, Kasuya H (1994). Accurate measurement of vocal tract shapes from magnetic resonance images of child, female and male subjects, *Proc ICSLP'94*, S12-14.1-S12-14