# Computational speech modelling

Jarmo Malinen

Aalto University,

Dept. of Mathematics and Systems Analysis

Lappeenranta, 18.10.2011.

# What is COMSPEECH?

Multidisciplinary research project combining phonetics, medicine, technology, and modelling.

Theoretical, computational, and experimental research.


$$
\begin{cases}
\Phi_{tt} = c^2 \Delta \Phi & \text{for } (\mathbf{r}, t) \in \Omega \times \mathbb{R}, \\
\Phi = 0 & \text{for } (\mathbf{r}, t) \in \Gamma_1 \times \mathbb{R}, \\
\frac{\partial \Phi}{\partial \nu} = 0 & \text{for } (\mathbf{r}, t) \in \Gamma_2 \times \mathbb{R}, \text{ and} \\
\Phi_t + c \frac{\partial \Phi}{\partial \nu} = 2 \sqrt{\frac{c}{\rho_0}} u & \text{for } (\mathbf{r}, t) \in \Gamma_3 \times \mathbb{R},
\end{cases}
$$

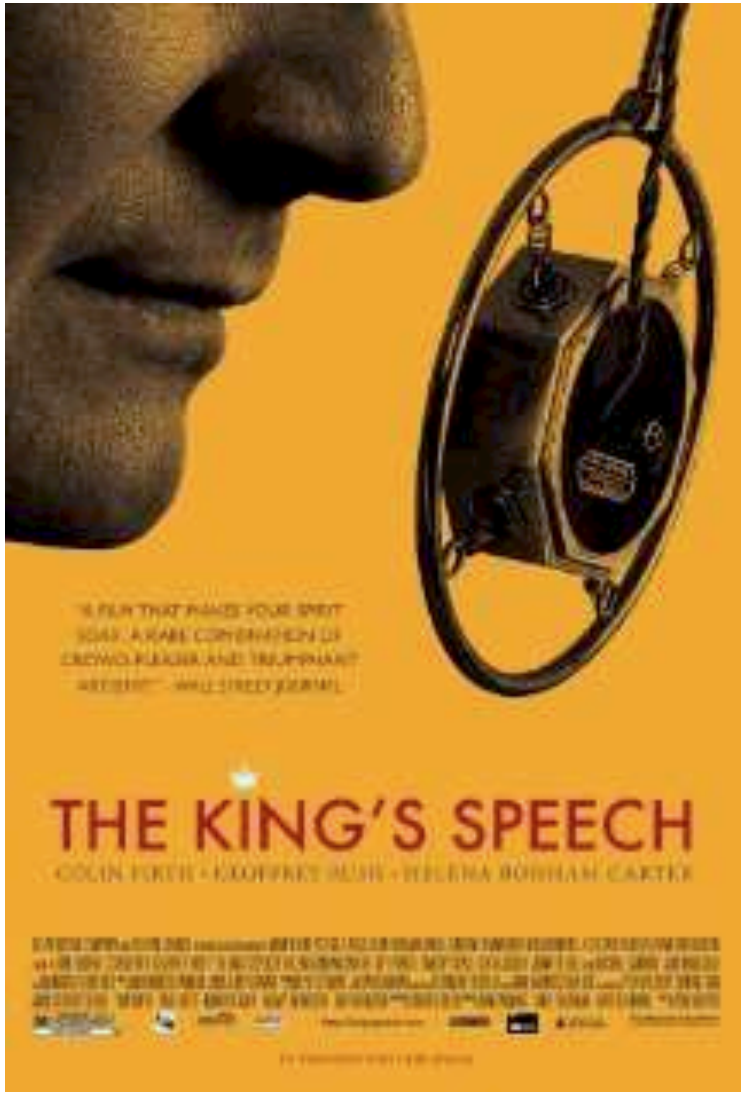
The long term aim is a design tool for use in oral and maxillofacial surgery as well as rehabilitation.

# A few facts from medicine...

In industrialised countries, $\approx$ 200 000 surgical operations in the head and neck area are carried out each year:

- Innate abnormalities or those acquired as a result of distorted growth,
- cancers in the head and neck area, and
- traumas.

The operations have an effect on speech through, e.g., the acoustics and flow mechanics of the vocal tract.


THE KING'S SPEECH

# ... and one more fact:

This can be **modelled computationally**.

# Computational models

# Short history of speech modelling

- H. von Helmholtz: Acoustic theory of vowels in 1863,
- G. Fant & al: Source–filter paradigm in practical speech synthesis in 1960's,
- J. Kelly and C. Lochbaum: First computer simulation by a scattering model in 1962, and
- R. Carré, M. Mrayati & al: Distinctive Regions Model relating anatomy to vowel characteristics in 1988.

Articulatory speech synthesis has been revolutionised by recent advances in computational possibilities.

# Acoustical equations

The acoustics of the vowel production can be simulated by using either Webster's equation

$$\Psi_{tt} = \frac{c^2}{A(s)}\frac{\partial}{\partial s}\left(A(s)\frac{\partial \Psi}{\partial s}\right) \quad \text{in} \quad [0, l] \times \mathbb{R}$$

or, for higher precision, 3D wave equation

$$\Phi_{tt} = c^2 \Delta \Phi \quad \text{in} \quad \Omega \times \mathbb{R}$$

for velocity potentials $\Psi$ and $\Phi$ that satisfy boundary conditions at the glottis, lips, and walls of the vocal tract $\Omega$. Here $A(\cdot)$ is the area of transversal sections of $\Omega$.

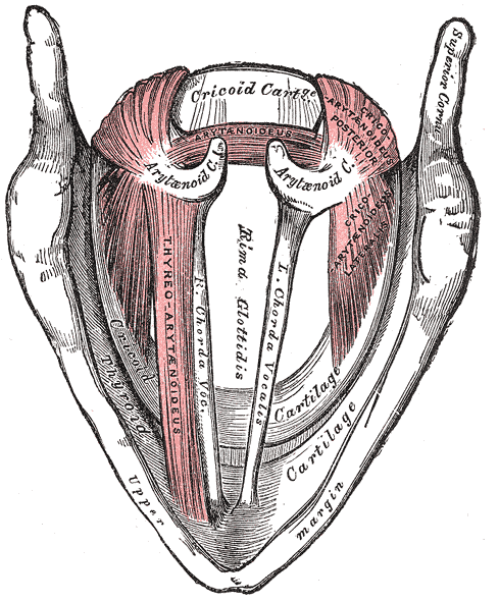
# Research model (1)

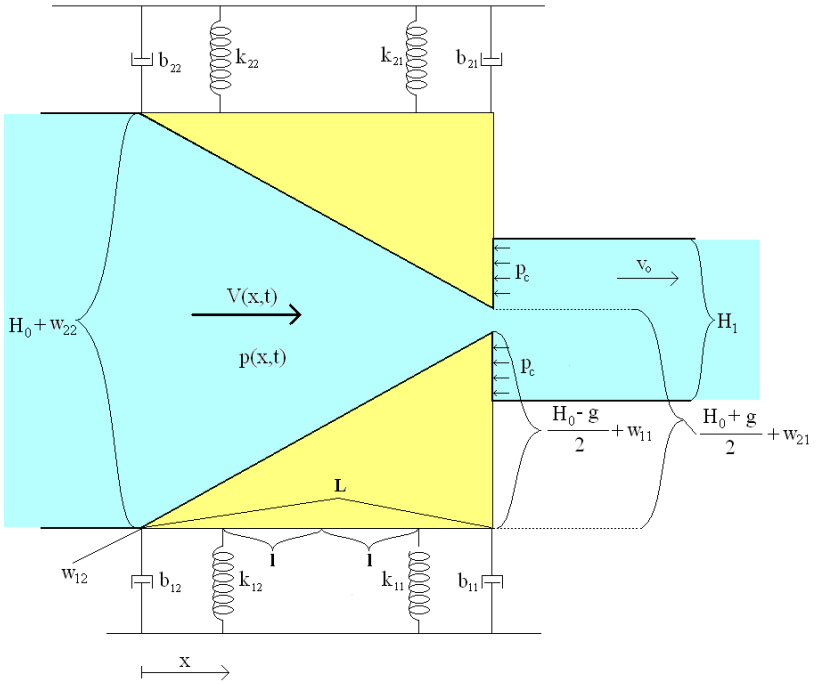
Our research model is a simulator that is based on

- an incompressible laminar Bernoulli air flow that produces an aerodynamic force...
- ...inducing vibrations in a damped mass-spring model of vocal folds which modulates the air flow as well as...
- ...produces sound pressure to a lossless acoustic resonator model of the vocal tract (henceforth, VT), realised using Webster's equation.

# Research model (2)

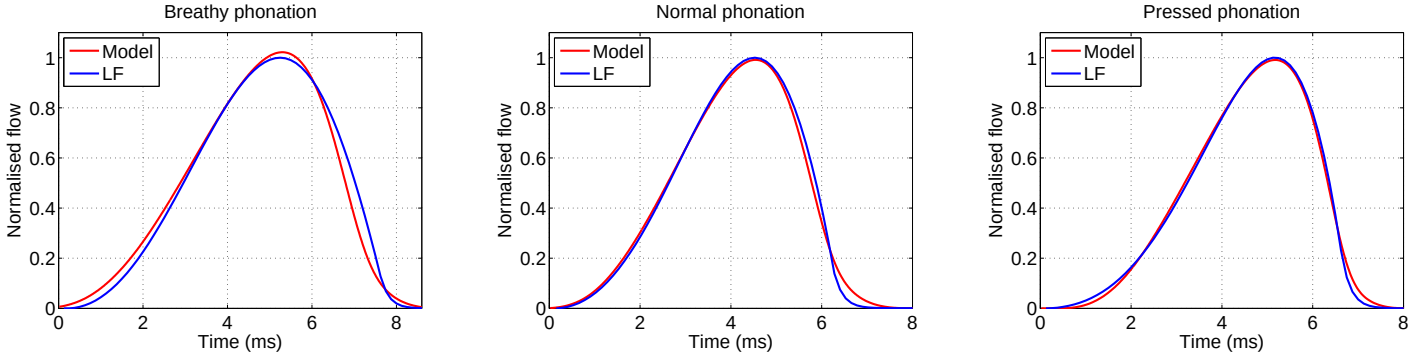
The anatomic geometry of the vocal folds in the model is highly idealised...



# Research model (3)

...but it compares well with measured data:

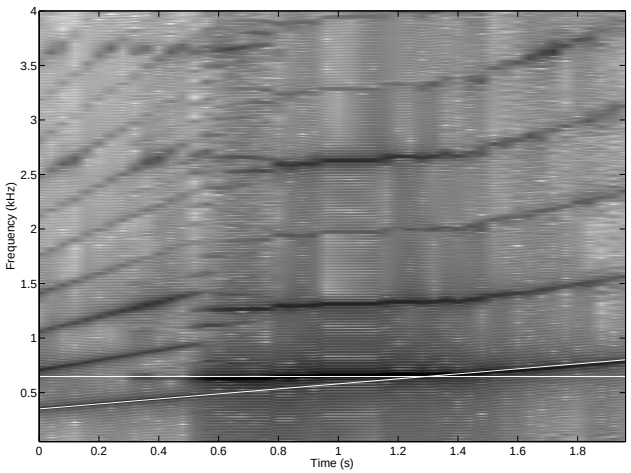


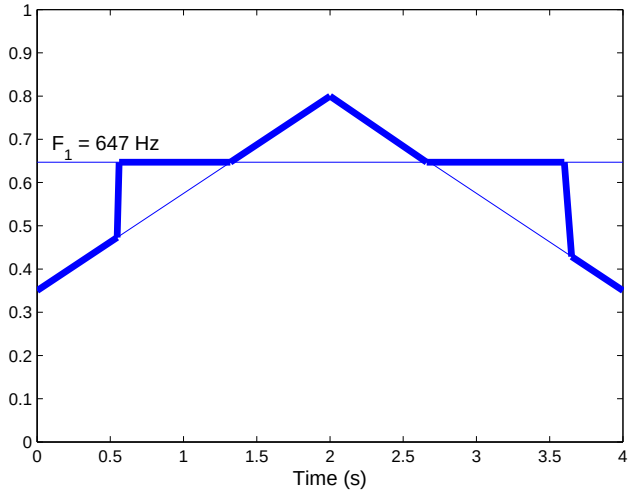

These are the glottal pulses in breathy, normal, and pressed phonation as inverse filtered from Prof. P. Alku.

Simulated in red, measured in blue.

# Modal locking to VT resonances (1)

We have recently simulated soprano singing near the first formant $F_1$ (i.e., the resonance of the VT):

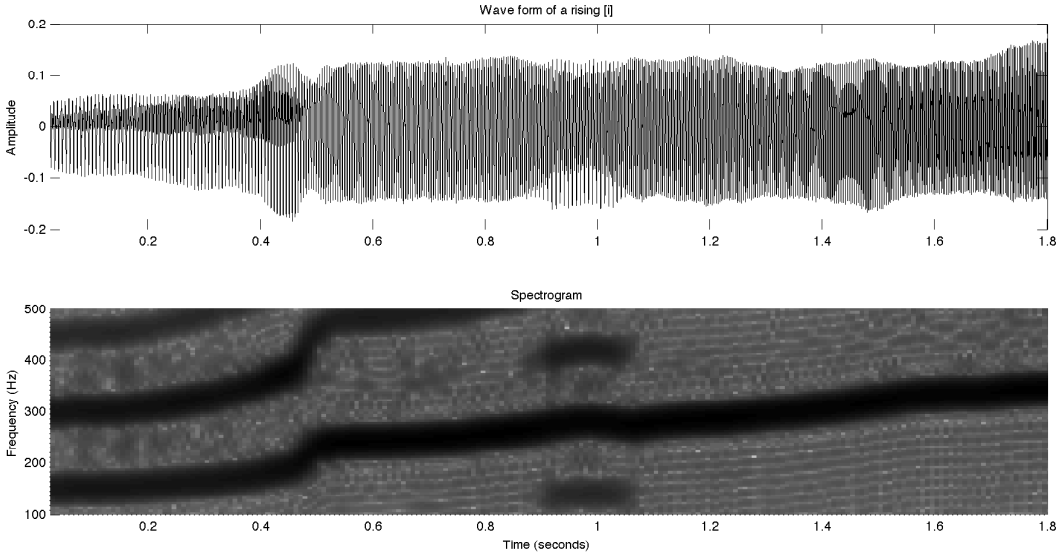



Extension of earlier work by I. Titze & al., in 2008.

# Modal locking to VT resonances (2)

The simulations are supported by experimental data:




Note the subphonation episode after the locking.

# Wave and Helmholtz equations

The wave/Helmholtz equations require more complicated boundary conditions than Webster's equation.

Cheap modelling of the exterior space?

The time-variant boundary condition at the glottis?

We have experimented with a Helmholtz solver with rudimentary boundary conditions, using vowel geometries and speech signals from Prof. O. Engwall, KTH.

There is a correspondence of sound and picture but a systematic error of about 3.5 semitones in the four lowest resonance frequencies.

# Why do we use multiple models?

Up to 4 kHz, the wave equation and Webster's equation on nonpathological human VT give approximately the same results (such as the formant positions).

At over 4kHz, the cross-mode resonances and bending of the VT produce phenomena that Webster's model cannot detect.

The "production model" will be based on the wave equation and the precise geometry of the VT.

The "research model" – based on Webster's equation – is developed further for use in numerically fast simulations and sanity checks.

# Experimental research

# How about the human experiments?

The vocal tract geometry $\Omega$ and the corresponding area function $A(\cdot)$ is acquired with MRI from test subjects (in clinical phase, from patients).

The boundary conditions at lips and glottis require special models that include experimental parameters.

The resolution (hence, the applicability) of the model depends crucially on the quality of the data, used for parameter estimation and validation.

**"No data, no research."**

# Data acquisition

Modelling requires simultaneous MRI and speech recording.


...but the devil is in the details.

# A list of things not allowed in MRI

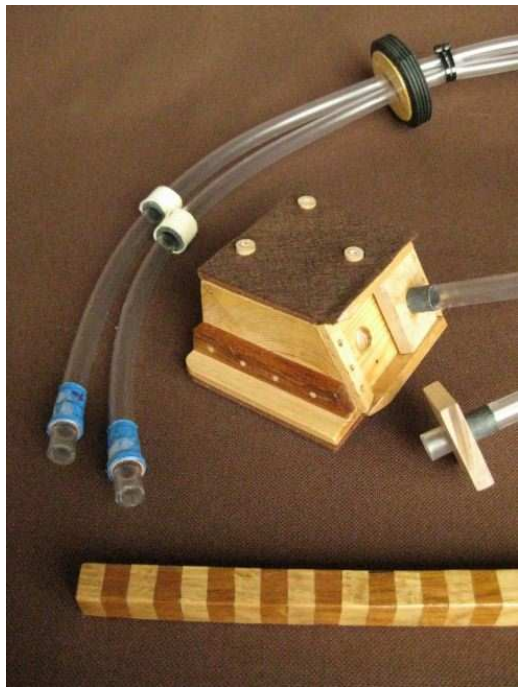
- No metal or electronics inside the MRI machine,
- no ferromagnetic material inside the MRI room, and
- nothing to introduce artefacts in MR images.
- In addition, acoustic noise of 90dB (SPL) and a strong EM field at 64 MHz must be tolerated.

Therefore, sound recordings must be carried out either by optical or acoustical arrangements.

We use the latter approach.

# Sound collector and wave guides (1)

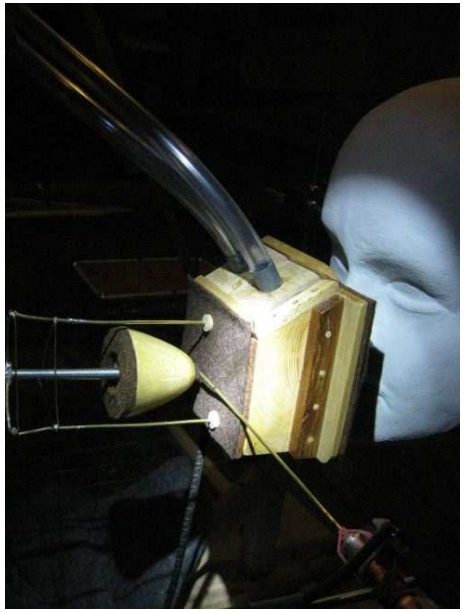
Separate channels for speech and noise samples:



Transmission properties of these channels are carefully matched to facilitate analogue noise cancellation.

# Sound collector and wave guides (2)

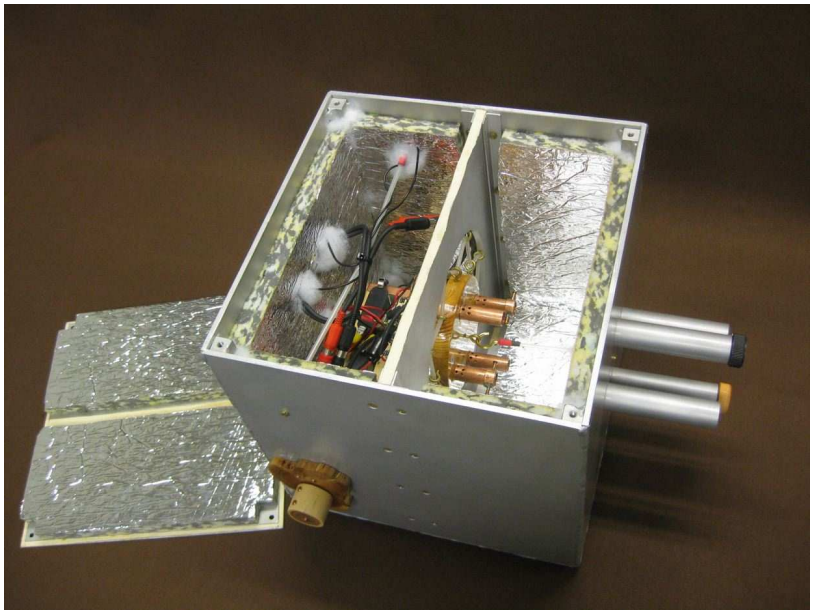
The relative position of the sound collector and the subject is shown in the laboratory picture on the left.



As shown on the right, the noise collector beam is widened by shadowing it with a paraboloid in front.

# Microphone assembly

The wave guides (of length 3.0 m) lead to microphones that are placed inside a sound-proof Faraday cage:



The speech and noise signals are then taken to an adjustable differential amplifier using RF-shielded cables.

# Data

# Pilot experiments in MRI

During three consecutive days in June 2010, 53 pilot experiments were carried out.

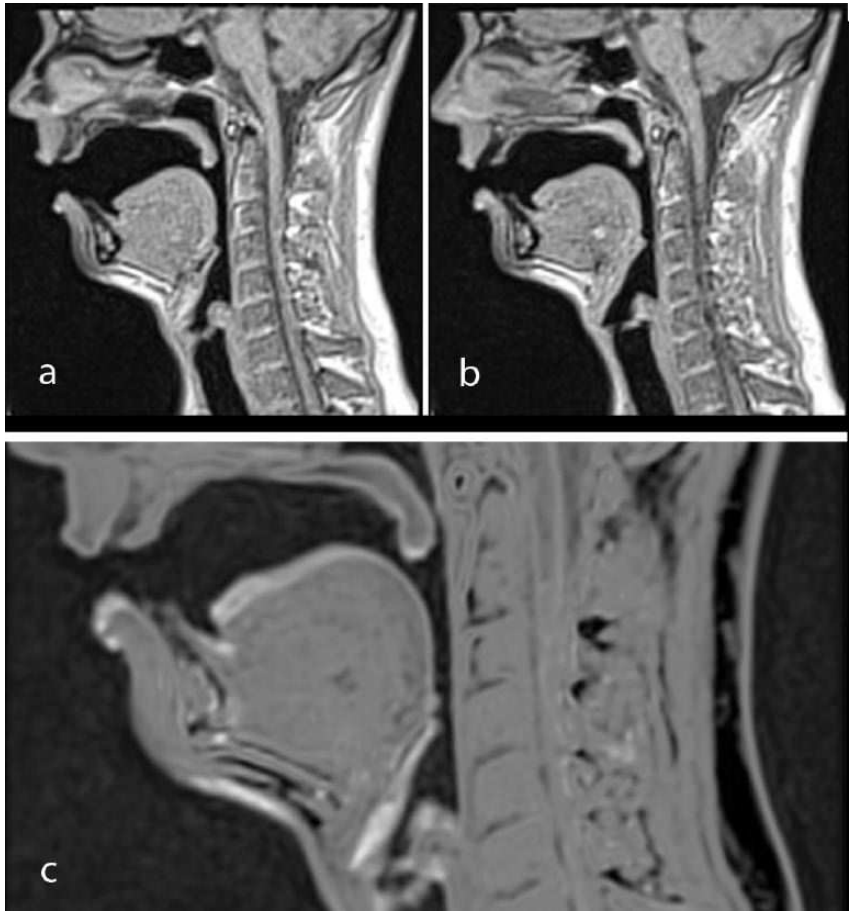
The subject is a 30 years old healthy male with background in speech sciences and music.

The vovels [ɑ], [e], [i], [o], [u], [y], [æ], and [œ] were produced at 110 Hz and 137.5 Hz, using 8 s scans.

The same set of vowels was produced at 110 Hz, using 18 s scans.

The experiments were repeated in anechoic chamber to get comparable sound data.

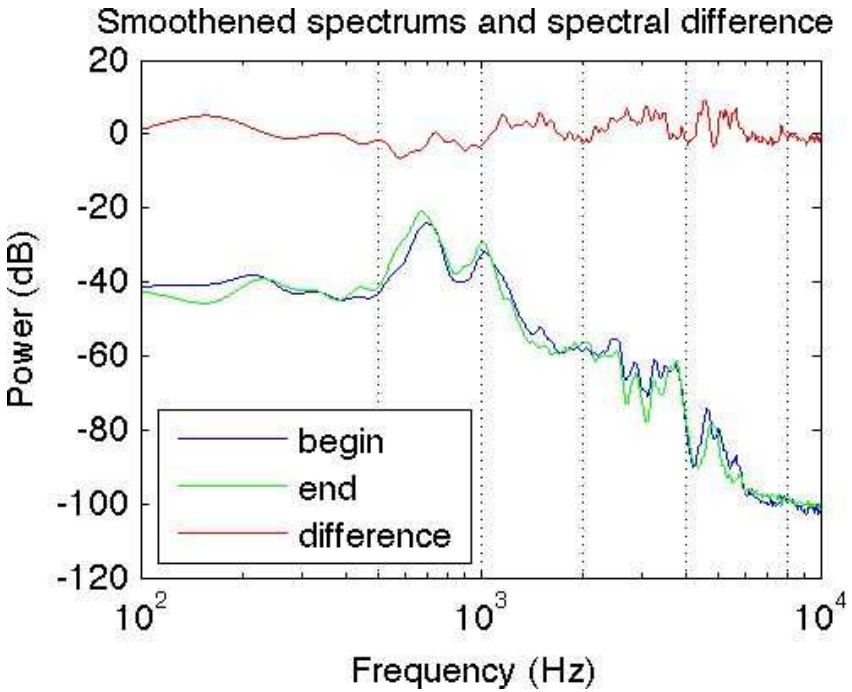
# MRI materials

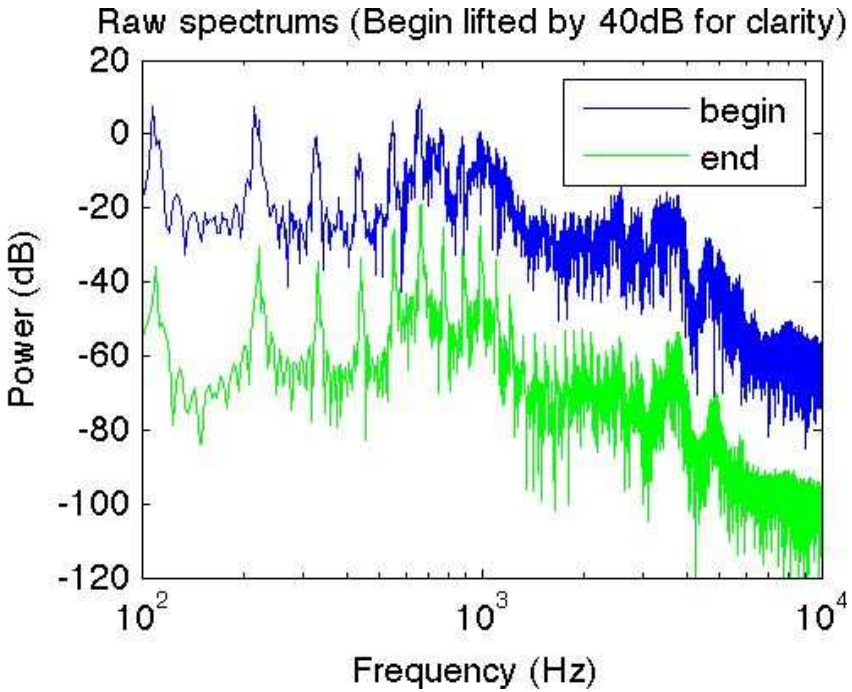



An example: Anatomy of [ɑ] at 110 Hz and 137.5 Hz.

# Quality of sound data (1)

The data is assessed by comparing the sound samples obtained immediately before and after the MRI noise.

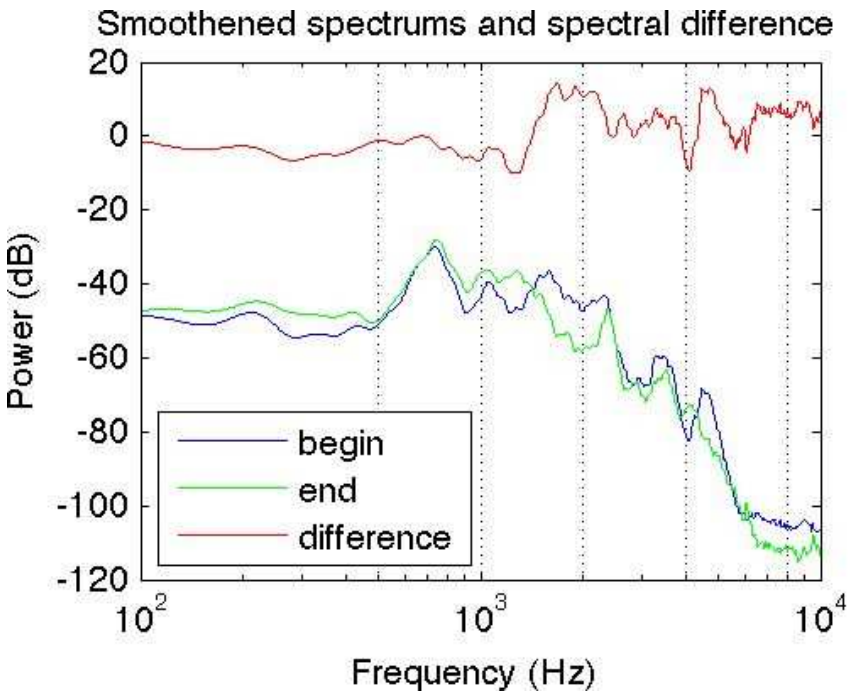


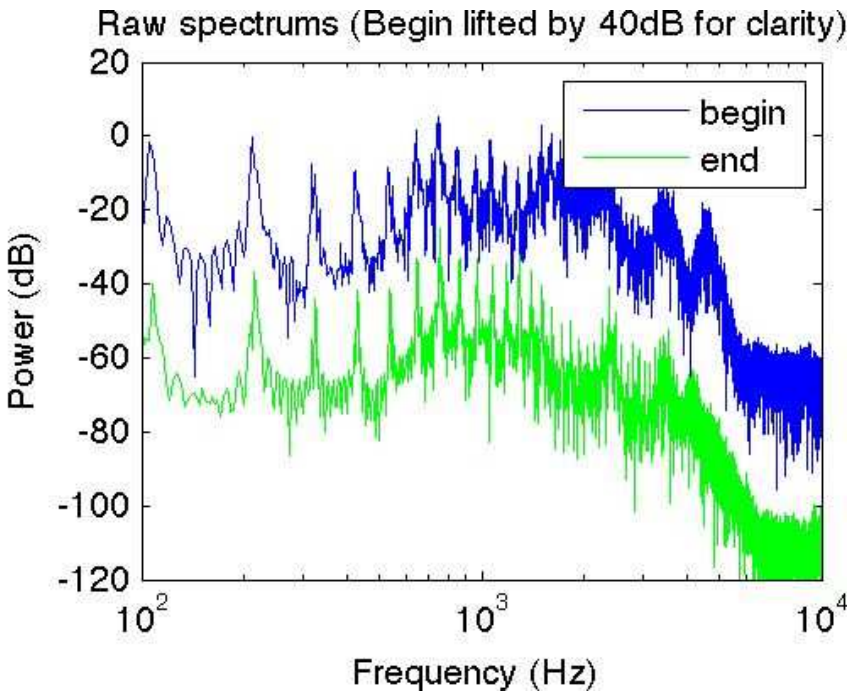




Spectra of vowel [ɑ] at 110 Hz, 8 s scan.

# Quality of sound data (2)

Sometimes the subject doesn't perform perfectly:




Spectra of vowel [ae] at 110 Hz, 8 s scan.

# Quality of sound data (3)

The movement of the speech organs during MRI is by far the dominant error source.

Out of 8 vowels at 110 Hz, 8 s scan, 4 are excellent and 4 are satisfactory.

Out of 8 vowels at 137.5 Hz, 8 s scan, 2 are excellent and 6 are satisfactory.

The sound data during longer scans (18 s) is typically worse than satisfactory.

A real-time data rejection criterium is needed!

# Dealing with the MRI noise

We use three approaches:

- arrange silent pauses during the MRI sequence using external control,
- cleaning up the power spectra from "spikes" of the MRI machine before formant extraction, and
- using advanced DSP such as sceptral analysis and the (modified) Richardson–Lucy deconvolution on spectrograms.

# What are we doing right now?

- Preparing the MRI of orthognathic patients, leading to a comprehensive MRI/sound data set,
- imaging of dentition and MRI markers on teeth, and
- developing automatic extraction of vocal tract boundary from the MRI data, FEM mesh generation, and area function estimation.

# That's all, folks. Questions?

Computational speech modelling for oral and maxillo-facial surgery.

Involved in the COMSPEECH -project:

Prof. O. Aaltonen, Prof. R.-P. Happonen, **Dr. J. Malinen,** Dr. D. Aalto, Dr. T. Lukkari, Dr. R. Parkkola, Dr. J. Saunavaara, Dr. T. Soukka, Dr. V. Turunen, Dr. M. Vainio, DI A. Aalto, DI A. Hannukainen, M.Eng. T. Murtola, DI P. Palo, and HLK J.-M. Luukinen.