Recording speech sound and articulation in MRI

Jarmo Malinen

Aalto University School of Science, Finland

BioDevices 2011, January 26 – 29, 2011, Rome

What?

- We simulate vowels in high precision by computationally modelling the acoustics of the vocal tract.
- Models require accurate anatomic data (including vocal tract geometry) during speech.
- For parameter estimation and model validation, the speech must be recorded simultaneously.

Why?

Each year, ca. 250 000 patients undergo oral or maxillofacial surgery that has an effect on speech:

Cancer and tumor surgery, orthognathic surgery, trauma surgery, reconstructive surgery, and surgery for dental impantology.

Sophisticated speech production models give novel tools for designing the surgery and patient rehabilitation.

Basic research in phonetics is quite interesting as well.

How?

Vowel production is modelled by Webster's equation

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

or, for higher precision, by 3D wave equation

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

with boundary conditions at glottis, mouth, and vocal tract walls.

And how it relates to the physical world?

The vocal tract geometry Ω must be obtained by MRI from test subjects (or patients).

The mouth and glottis boundary conditions involve separate models, and they contain empirical parameters.

The resolution (hence, the applicability) of the model depends crucially on the quality of the data.

All PDE's are numerically solved by FEM...
... but this is not really the subject of this talk.

Recording sound and articulation in MRI?

Let's review MRI technology, experimental arrangements, phonetical aspects, and sound recording in MRI.

MRI coils, sequences, and imaging times

The subject lies in a Siemens Avanto 1.5T -machine in supine position with a sound collector positioned in front of his mouth.

12-element Head Matrix Coil, combined with the 4-element Neck Matrix Coil, cover the speech organs.

GRAPPA technique is used with accel. factor 2.

3D VIBE is the most suitable MRI sequence.

MRI with 1.8 mm isotropic voxels takes **7.6 s**.

Resolution of 1.2 mm isotropic voxels requires **17 s**.

Experimental arrangement (1)

The subject is given instructions and a cue signal right before MRI starts. The cue signal gives the starting time and the desired constant pitch (110 Hz or 137.5 Hz).

The subject hears his own voice through earphones.

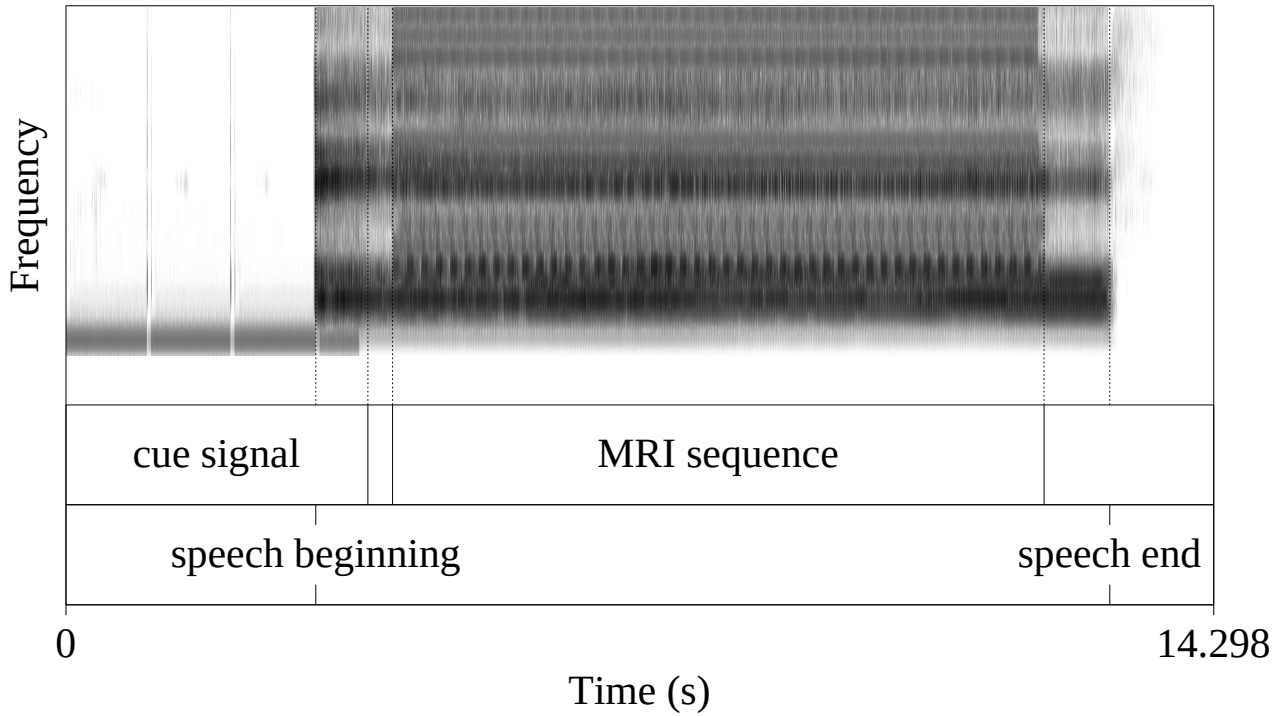
The speech production is longer than the MRI sequence to get clean samples before and after the MRI noise.

In particular, 500 ms time is reserved for the stabilisation of voice in the beginning of each experiment.

A phonetician follows speech quality in real time.

The MRI quality is inspected on the spot as well.

Experimental arrangement (2)



A spectrogram of a speech recording during MRI.

A list of things not allowed in MRI

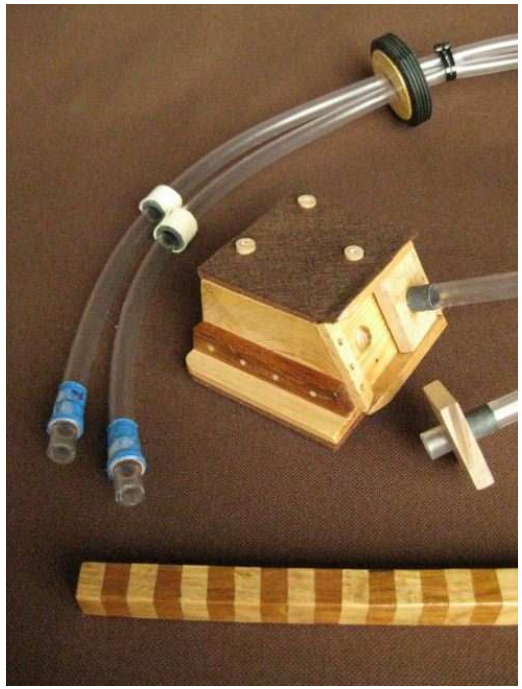
- No metal or electronics inside the MRI machine,
- no ferromagnetic material inside the MRI room,
- nothing to introduce artefacts to MR images,
- acoustic noise of 90dB (SPL) and a strong EM field at 81 MHz must be tolerated.

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

We use the latter approach.

Sound collector and wave guides

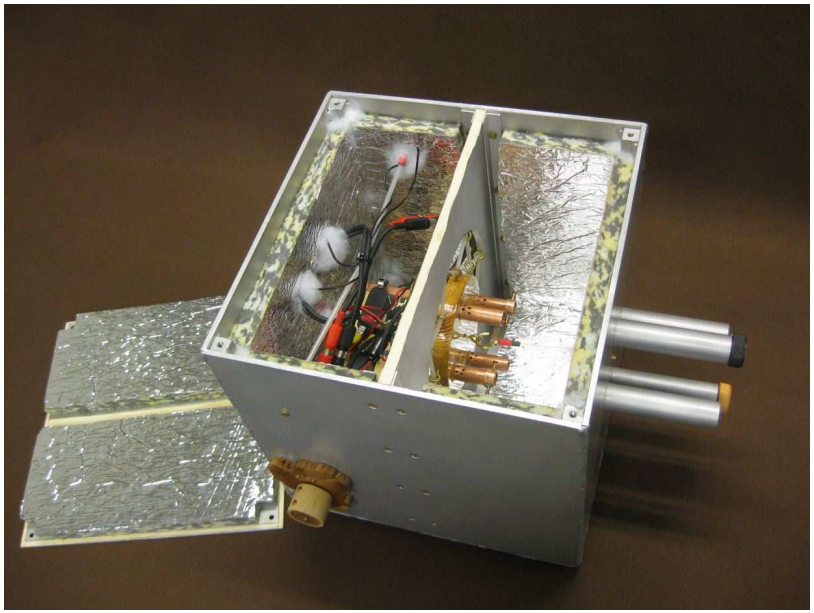
Separate channels for speech and noise samples:



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

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.

Pilot experiments

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

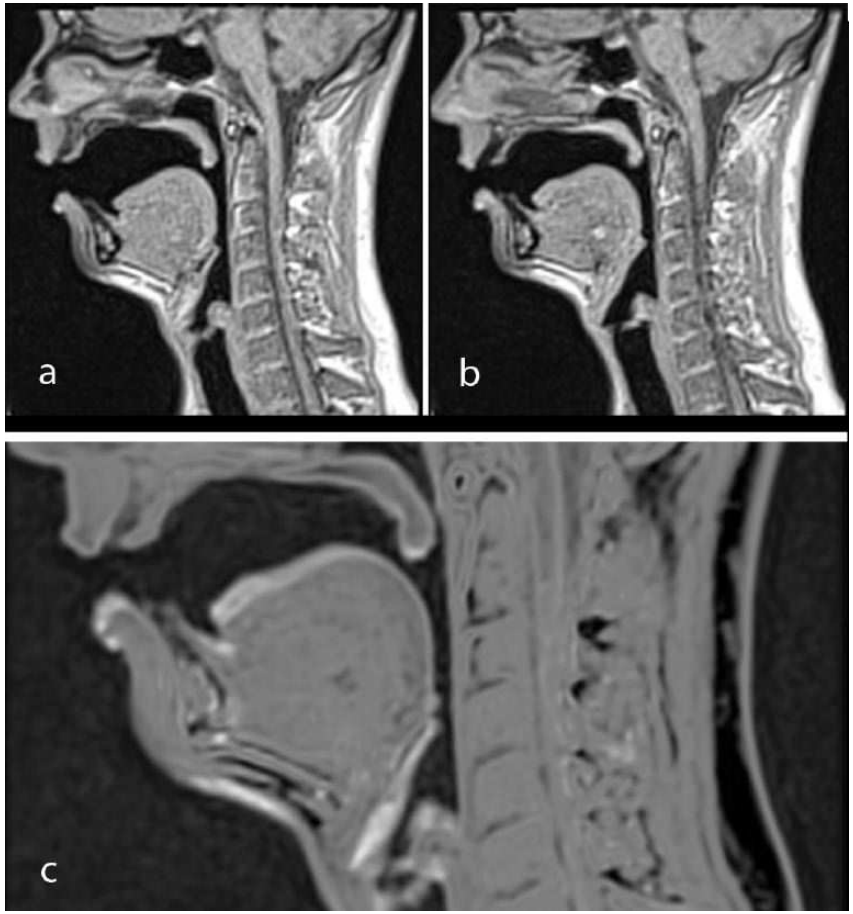
The subject is a 30 years old healthy male with background in speech sciences and music.

The vovels [ɑ], [e], [i], [o], [u], [y], [ae], and [oe] 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.

In addition, dynamical MRI was used to study the movement of the vocal tract during long vowels.

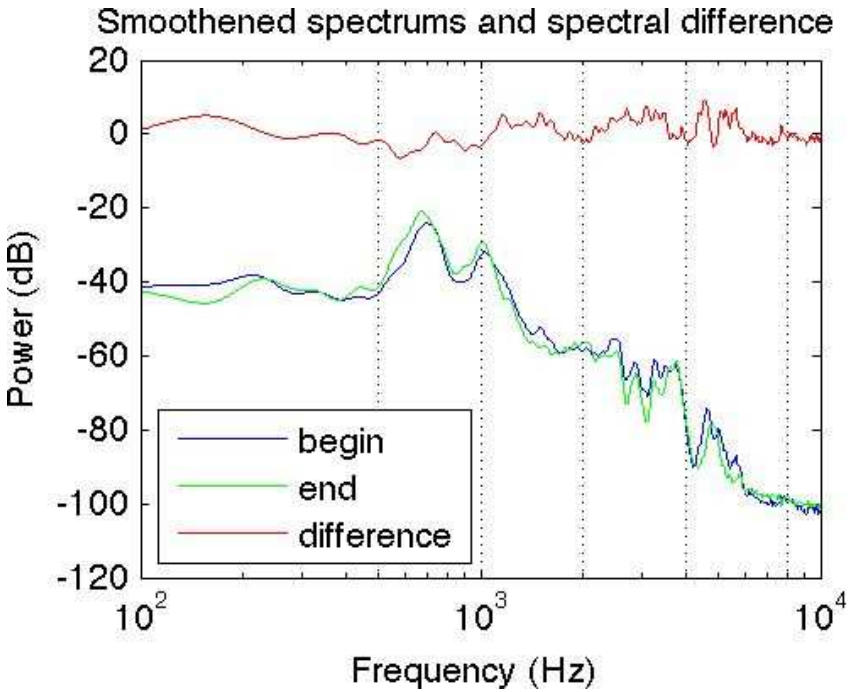
MRI materials

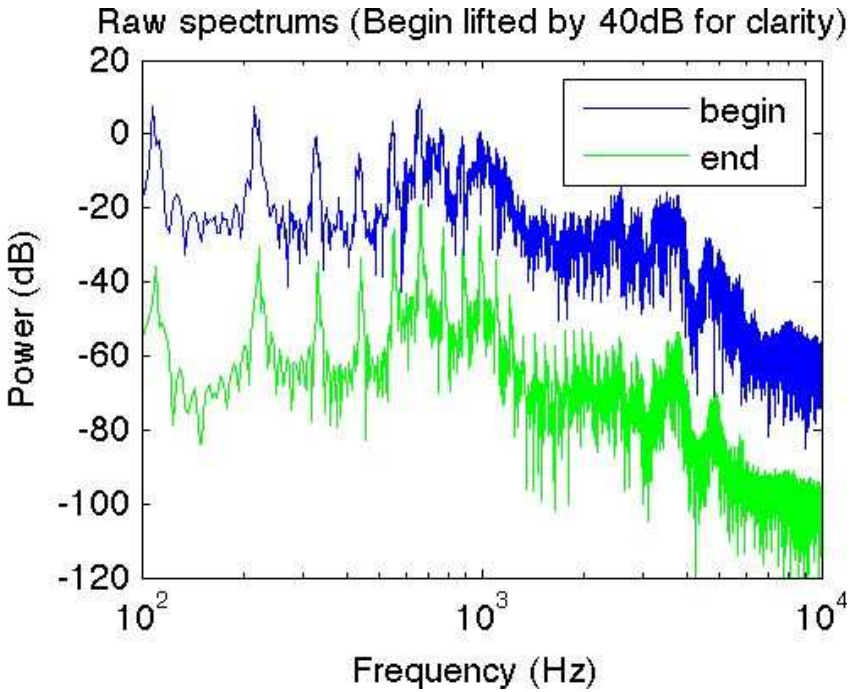


An example: Anatomy of [ɑ] at 110 Hz and 137.5 Hz.

Quality of sound data (1)

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

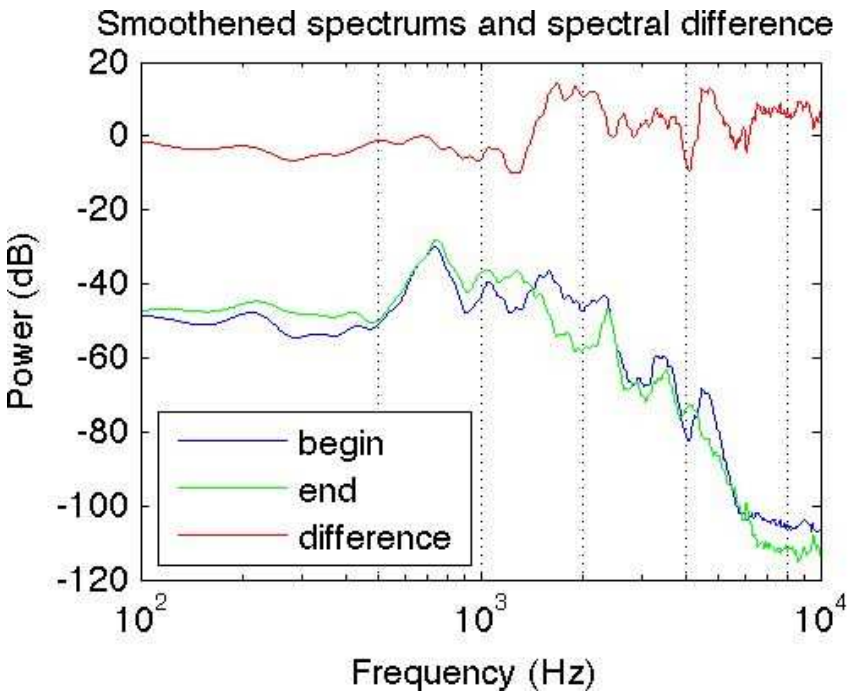


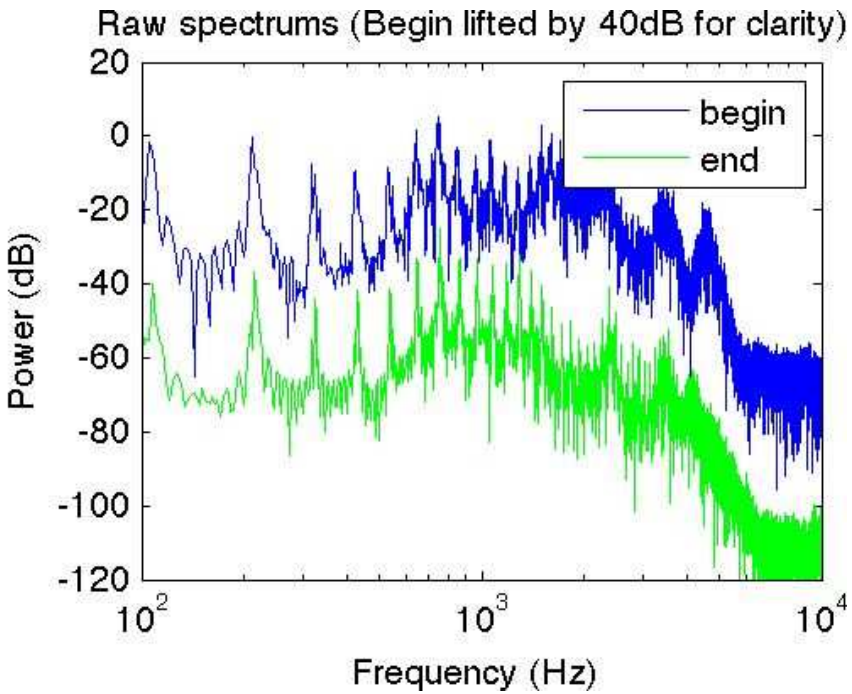


Spectra of vowel [ɑ] at 110 Hz, 8 s scan.

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)

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

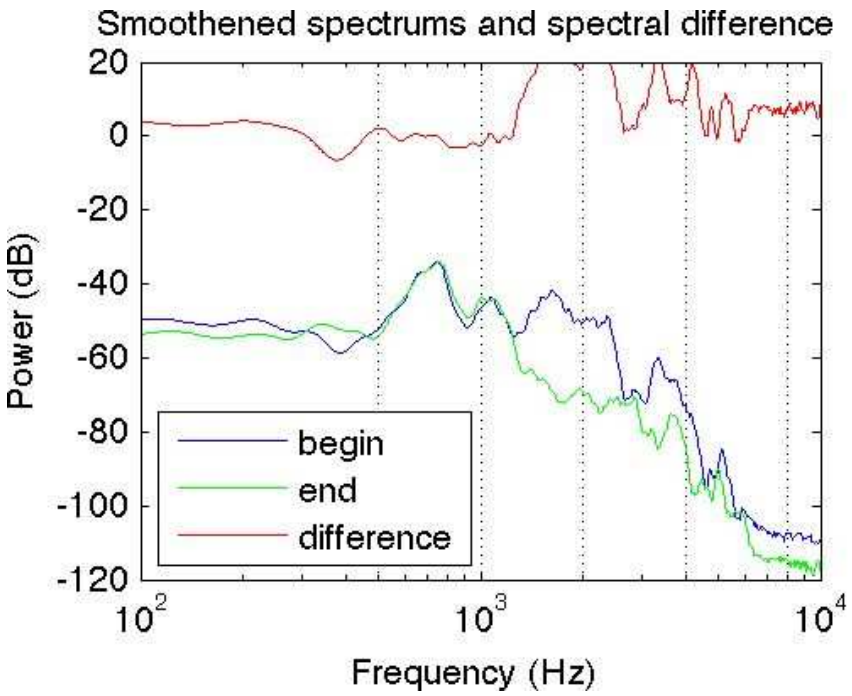
Out of 8 vowels at 137.5 Hz, 8 s scan, 2 are excellent and 6 are satisfactory.

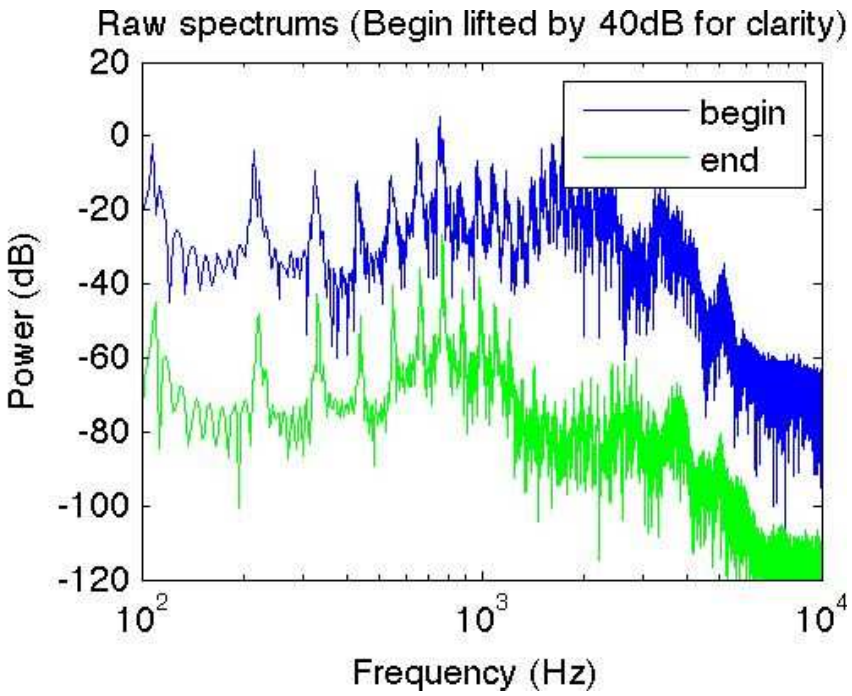
The sound data during longer scans (18 s) is typically worse than satisfactory.

The formant (i.e., the vocal tract resonance) extraction with standard LPC settings by Praat is not reliable with this data.

A real-time data rejection criterium is needed.

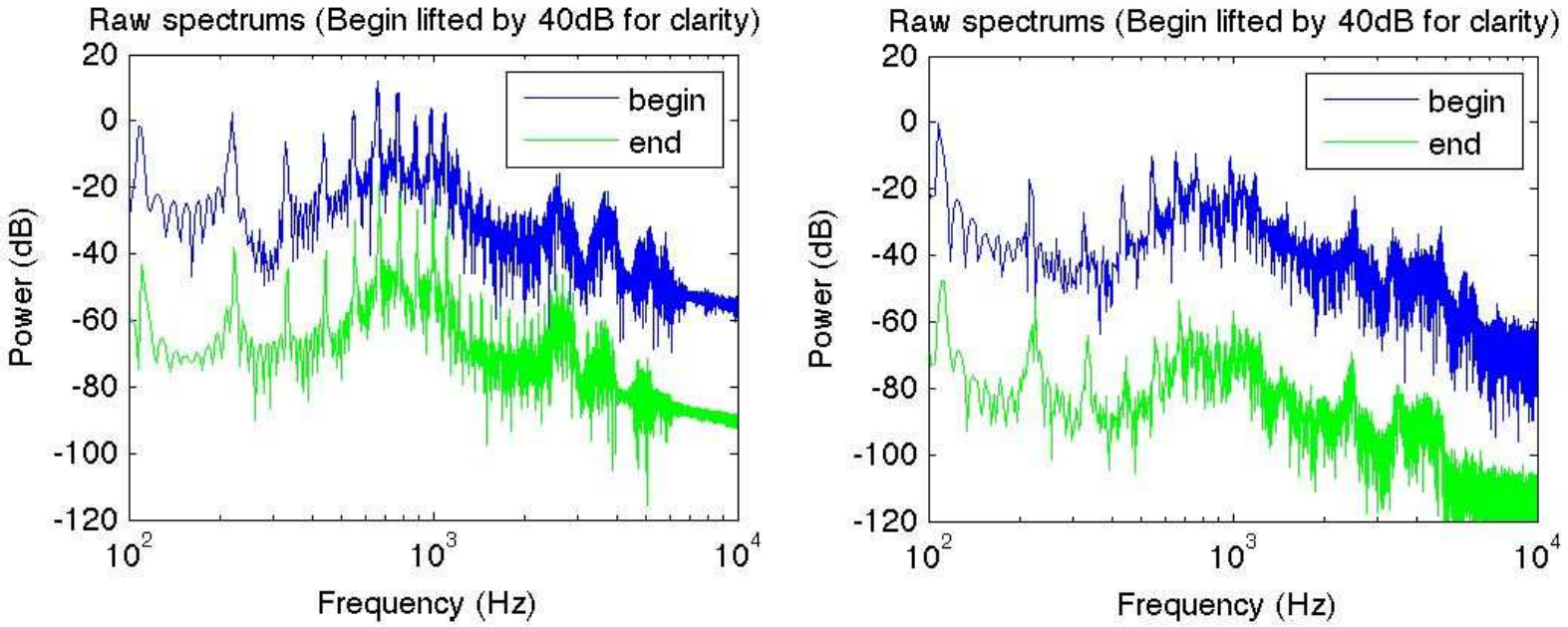
Observations (1)





Vowel glides from [ae] to [ɑ] during dynamical MRI.

Observations (2)



Normal and breathy phonation types of [a].

Conclusions

Quite satisfactory 3D, static MR image data can be obtained using relatively short scans of duration < 8 s.

During the MRI noise, we reach a positive S/N-ratio in most speech recordings by the real-time analogue noise cancellation alone.

After preprosessing by DSP, the frequency response is flat in the range $0.1 - 4.4$ kHz (and above).

First five formants can be recovered during MRI pauses.

The residual MRI noise can be substracted from the speech power spectra before formant extraction by LPC.

Problems and improvements (1)

Vocal organs always change shape during an 8 s MRI scan. The type and significance of the problem depends on the vowel.

Scanning times over 8 s cannot be recommended at all.

Any movement (or lack of it) would remain unobserved in a single, static MR image.

Some movement may remain unobserved in the speech recordings, too.

Problems and improvements (2)

Dentition is not visible in MRI at all.

The phonation type (breathy, normal, pressed) is difficult to standardise in the experiments.

The MRI sequence can be triggered externally (in sync with the cue signal) so as to introduce noiseless pauses.

The half-life of noise in MRI room is $\approx$ 20 ms, and hence the pauses must be longer than 100 ms.

Excellent data is possible using a healthy, phonetically trained subject.

Questions, please?

Involved in the 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. M. Vainio,

M.Sc. A. Aalto, M.Sc. A. Hannukainen, M.Sc. T. Murtola,

M.Sc. P. Palo, and B.Dent. J.-M. Luukinen.