Mathematical modeling of speech acoustics

D. Sc. Daniel Aalto
Ultimate goal

Predict the *speech* outcome of oral and maxillofacial surgery patients
How to reach the goal?

- By synthesizing the speech of a virtually operated vocal tract
This talk

- Orthognathic patients and model validation
- Objective measurement of speech acoustics
- Presentation of the computational vowel model
Orthognathic patient data I

- Anatomic changes are predictable (advancement of mandible and/or maxilla)
- Comparison of measured acoustics and computed speech characteristics from the MR images
- Comparison of the prediction and the real speech outcome
Orthognathic patient data II

- 20 (10 women, 10 men) patients undergoing an orthognathic treatment in Turku are enrolled
- so far: 6 pre-treatment recordings
- 70 sustained sound productions (10 s) for each patient
- 12 short sentences
- Measurable changes in formants (Niemi et al. 2006)
Orthognathic patient data III

- Phonetically rich sentences
- Vowels occur in phonetically controlled contexts
- Coarticulation and motility in sagittal videos
- Changes in vowel space are predicted
What is speech for?

Linguistic code:

Communicating the “thoughts” of a person through language e.g. “Mom, do pharaohs exist?”

Biological code:

The speaker reveals information about him/herself e.g. the speaker sounds assertive, healthy, happy, tired, 5-6yo boy, speaker of Finnish (Turku dialect)
Vocal tract resonances

Thanks to:
A. Suni, T. Raitio, P. Alku
Vowel space: Formants

- Size of the vowel space correlates with intelligibility
- In addition, vowel formants provide information of the speaker
Dico: modeling speech acoustics

An integrated vowel acoustics model:

- 3D-image is transformed to a sound!
- A mechanical model for vocal folds
- An acoustic model for the air vibrations inside the vocal tract
- Interaction between vocal folds and vocal tract
- Model = simplified reality!
Dico: Submodel for vocal folds

Vocal fold movement:

\[
\begin{align*}
M_1 \ddot{W}_1(t) + B_1 \dot{W}_1(t) + PK_1 W_1(t) &= -F(t), \\
M_2 \ddot{W}_2(t) + B_2 \dot{W}_2(t) + PK_2 W_2(t) &= F(t)
\end{align*}
\]

Air flow through glottis:

\[p_{sub} = \bar{C}_{inert} \ddot{v}_0(t) + \bar{C}_{visc} W_1(t) \dot{v}_0(t)\]

Vocal tract impact:

\[
\begin{align*}
\frac{\partial^2 \Psi(x,t)}{\partial t^2} &= \frac{\partial^2}{\partial x^2} \left( A(x) \frac{\partial \Psi(x,t)}{\partial x} \right), \quad x \in [0, LVT], \\
\Psi_x(0,t) &= -\dot{w}_0(t), \\
\Psi_x(LVT,t) + \theta v \Psi(LVT,t) &= 0, \\
p_0(t) &= \rho \Psi(0,t).
\end{align*}
\]
Dico: Acoustic equations

\[
\begin{aligned}
\Phi_{tt} &= c^2 \Delta \Phi \text{ on } \Omega, \quad \Phi = 0 \text{ on } \Gamma_1, \\
\frac{\partial \Phi}{\partial \nu} &= 0 \text{ on } \Gamma_2, \quad \Phi_t + c \frac{\partial \Phi}{\partial \nu} = 2\sqrt{\frac{c}{\rho_0}} u \text{ on } \Gamma_3,
\end{aligned}
\]

\(\Phi = \text{virtual sound}\)

\(\lambda = \text{Formant/resonance frequency}\)
Dico: Solving the equations

Computed standing waves
Validation details

- Are we solving the right equations?
- Possible error sources:
  - Instabilities in sustained vowel production
  - Spatial inaccuracies in MRI and image processing
  - Room acoustics, measurement equipment, noise cancellation
Vowel production stability

- Tongue tip advancement during a sustained [ae] with bite-block
- Three first formants of an [a] and an [æ]
Image processing

- Semi automatic volume detection
- Point Cloud Library
- Teeth superposition
- Surface model

Aalto, et al. 2013
Sound recording

- Compensation for sound deformations due to the wave guides
- Noise cancellation
- Room acoustics still challenging
First results (pilot data)
Next steps

- Model improvement: lip boundary
- Complete analysis of the data
- Predictions for the post-op conditions
- Extending the methods to other patient groups
- Mobility analysis based on mid-sagittal videos
- Sensitivity analysis based on 3D geometries
- Visualization by multicolored 3D-prints
Generalizability of the approach

- The model is based on anatomy and physics: potential for reliable simulation of the vocal tracts independent of age, sex, ethnic group
- Model focuses on the passage from anatomy to acoustics: feasible for most patient groups
- Here the data is Finnish but every language in the world contains several vowels
- Some consonants are also captured by the model
Future questions

- Combining precise anatomical images (MRI) with moving vocal tract videos, could we pre-synthesize the speech following any surgery?

- Could a speech synthesizer be personalized to sound like the patients’ own voice (e.g. cerebral palsy patients) based on the anatomy of the vocal tract?

- Could we predict the post-op intelligibility and/or ease of articulation of a patient undergoing oral and maxillofacial surgery?
Thanks!