Aalto University

Speech communication technology research at Aalto University

Paavo Alku
Department of Signal Processing and Acoustics
Aalto University

9.11.2019

Contents

1. Aalto University and its speech research teams in a nutshell
2. Statistical parametric speech synthesis
3. Analysis of voice production based on multi-channel recordings of vowel production
4. Speech-based biomarking of human health
5. Summary

1. Aalto University and its speech research teams in a nutshell

Aalto University

- Founded in 2010 by merging Helsinki University of Technology, the Helsinki School of Economics, and the University of Art and Design Helsinki

- Consists of six schools (*e.g.* School of Electrical Engineering)

- About 11 000 students and 390 professors

Aalto University (in Espoo, about 10 km from Helsinki city center….and about 3500 km from Porto)

Department of Signal Processing and Acoustics

- Personnel: ca. 90 (11 professors)

- Budget: 5 M€

- Research areas

  - Acoustics
  - Measurement technology
  - Signal processing
  - **Speech**

Speech research teams

- Paavo Alku (Professor): Speech communication technology

- Tom Bäckström (Associate Professor): Speech coding, privacy issues in speech

- Mikko Kurimo (Associate Professor): Automatic speech recognition

Speech communication technology team

- Staff (1.11.-19): 1 professor, 3 post docs, 3 PhD students, 1 MSc student
- Our work has been highly **interdisciplinary** and conducted together with colleagues from psychology, neuroscience, medicine, phonetics, and mathematics
- Emphasis on **basic research,** yet also a long history in collaboration with **ICT industry**

Research topics

- Analysis of voice production
- Brain functions in speech perception
- Quality and intelligibility improvement of speech in mobile phones
- Robust speech processing
- Speech-based biomarking of human health
- Statistical parametric speech synthesis

From the team's research topics, the following three are discussed

- **Statistical parametric speech synthesis**

- **Analysis of voice production based on multi-channel recordings of vowel production**

- **Speech-based biomarking of human health**

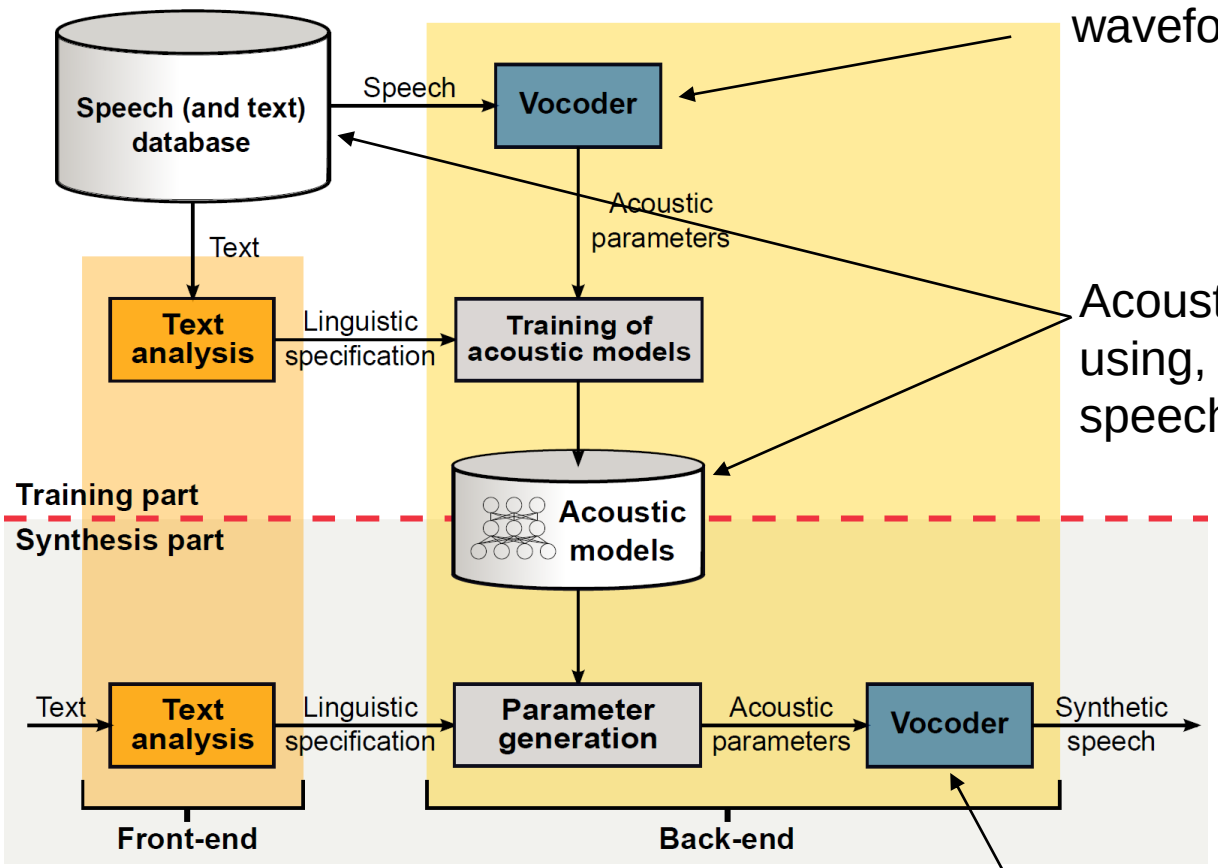
1. Statistical parametric speech synthesis

- Text-to-speech synthesis (TTS): converting text input to natural sounding speech based on a statistical and parametric approach (i.e. not concatenative synthesis)

- TTS has been studied by the team since 2008 by focusing particularly on vocoding

- TTS has progressed extensively in the past 5 years due to industry interest and conventional vocoders are increasingly replaced with neural vocoders (e.g. WaveNet)

Overview of statistical parametric speech synthesis

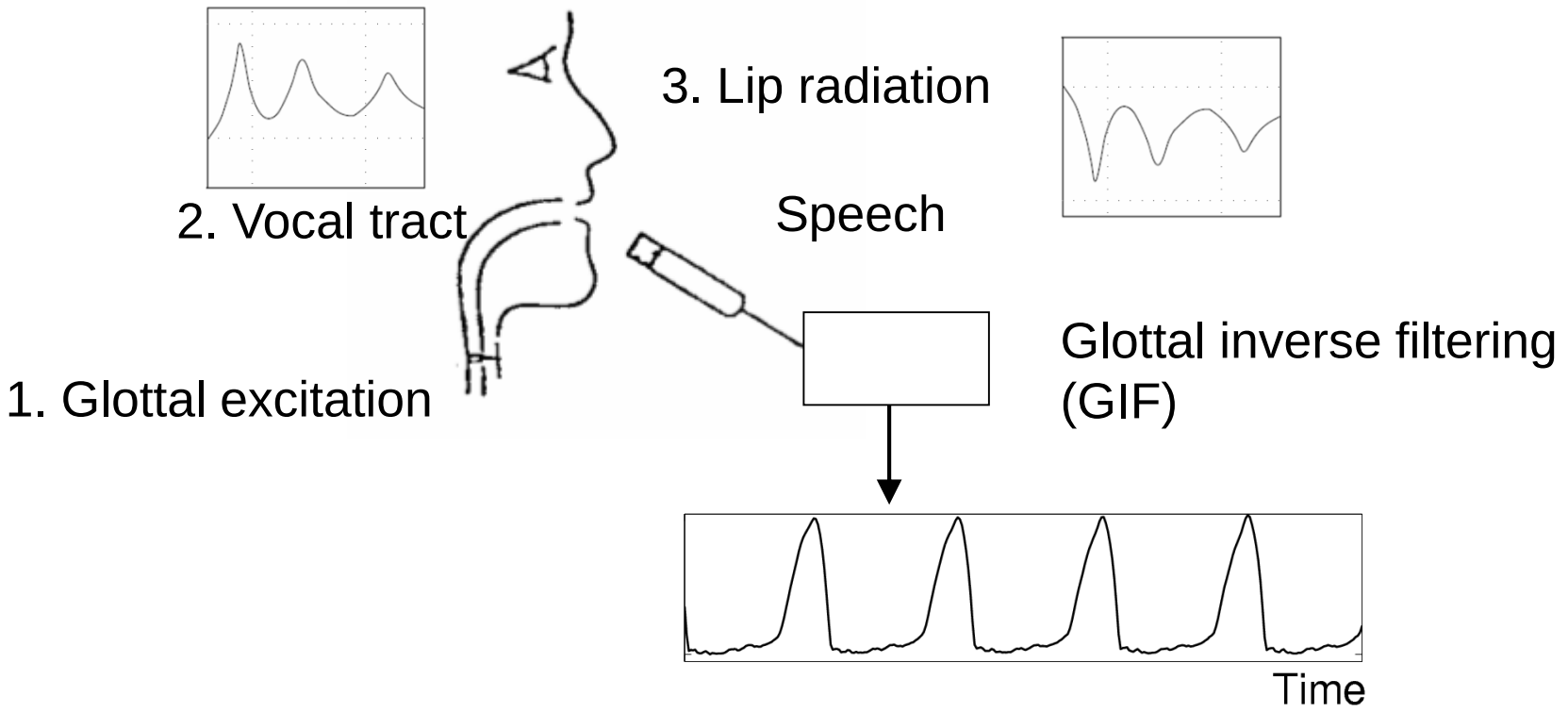


waveform =>parameters

Acoustic model is trained using, e.g., 2-3 hours of speech

parameters =>waveform

Our TTS studies have used extensively vocoding based on glottal inverse filtering (GIF)



Examples of studies in TTS:

- Introduction of GIF-based vocoding in parametric speech synthesis

  Raitio, Suni, Yamagishi, Pulakka, Nurminen, Vainio, Alku: *HMM-based speech synthesis utilizing glottal inverse filtering*. IEEE Transactions on Audio, Speech, and Language Processing, Vol. 19, No. 1, pp. 153-165, 2011

- Comparison of different vocoders

  Airaksinen, Juvela, Bollepalli, Yamagishi, Alku: *A comparison between STRAIGHT, glottal, and sinusoidal vocoding in statistical parametric speech synthesis*. IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol. 26, No. 9, pp. 1658-1670, 2018

- Neural vocoding of glottal flow

  Juvela, Bollepalli, Tsiaras, Alku: *GlotNet—A raw waveform model for the glottal excitation in statistical parametric speech synthesis*. IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol. 27, No. 6, pp. 1019-1030, 2019

- Adaptation of speaking style in TTS

  Bollepalli, Juvela, Airaksinen, Valentini-Botinhao, Alku: *Normal-to-Lombard adaptation of speech synthesis using long short-term memory recurrent neural networks*. Speech Communication, Vol. 110, pp. 64-75, 2019

Demonstration of TTS (based on Airaksinen et al. IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol. 26, No. 9, pp. 1658-1670, 2018)

2. Analysis of voice production based on multi-channel recordings of vowel production

- Data of three channels (acoustic speech, EGG, high-speed imaging of the vocal folds) were collected from 10 speakers at Helsinki University Central Hospital

- This data has been used for two purposes:
  - 2.1: To build an open environment for GIF evaluation
  - 2.2: To study glottal flow vs. glottal area in vowel production

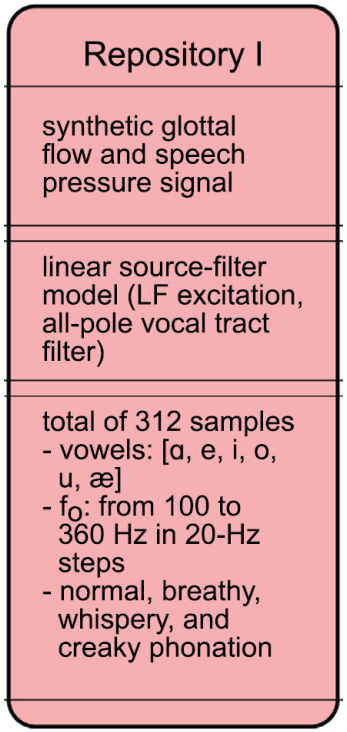
2.1 To build an open environment for GIF evaluation

- Background:
  - Performance of GIF methods is problematic to evaluate due to the lack of ground truth (i.e. no access to the glottal flow in natural production of speech)
  - To circumvent the above problem, different approaches (e.g. synthetic speech, EGG) have been used. However, the study area lacks an open, joint evaluation platform

    =>

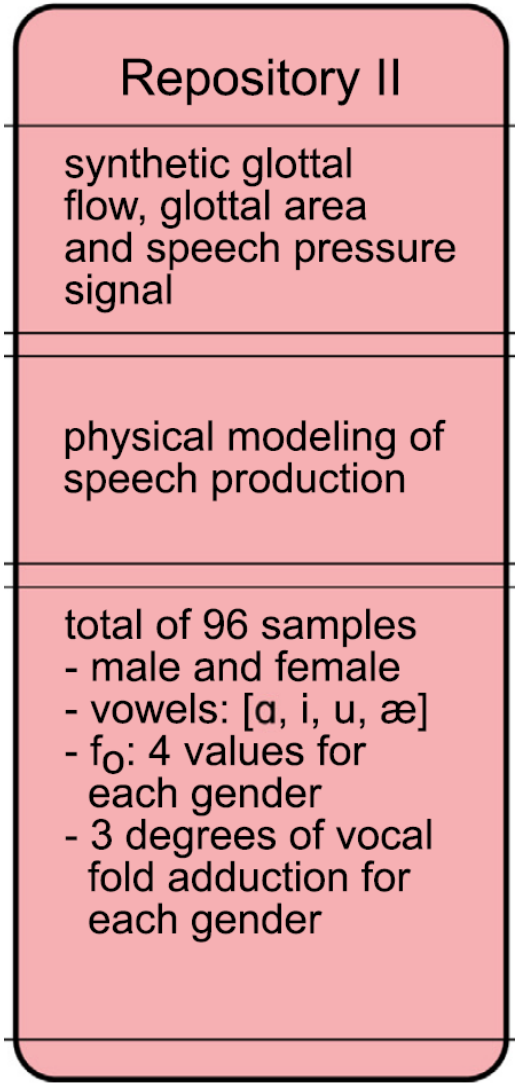
    The OPENGLOT environment was recently launched (Alku et al., *OPENGLOT - An open environment for the evaluation of glottal inverse filtering*. Speech Communication, Vol. 107, pp. 38-47, 2019)

- The OPENGLOT environment
  - Provides open data for evaluation of GIF (and formant estimation methods)
  - Data is organized into four repositories (I-IV):



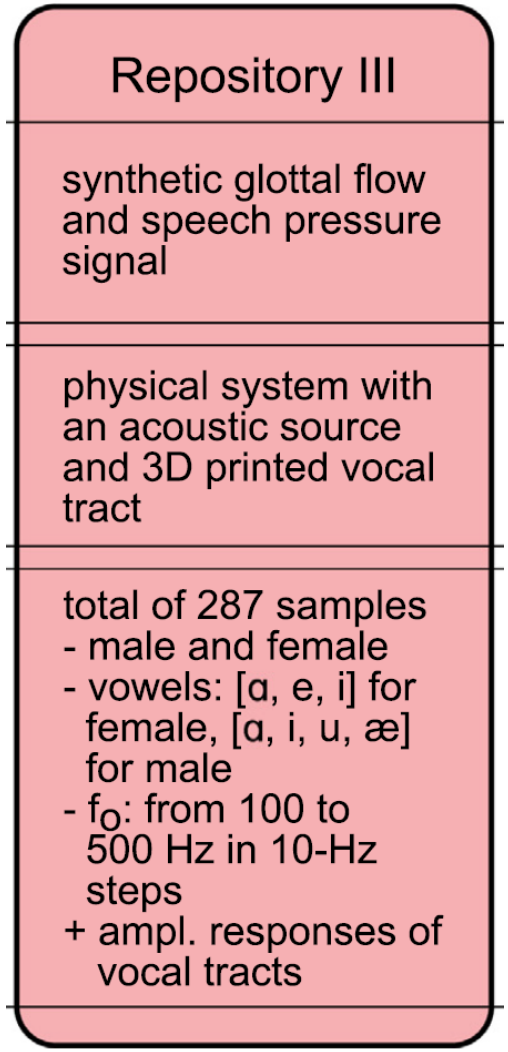
Repository I

- Synthetic vowels produced by a linear source-filter model (LF excitation, all-pole tract)



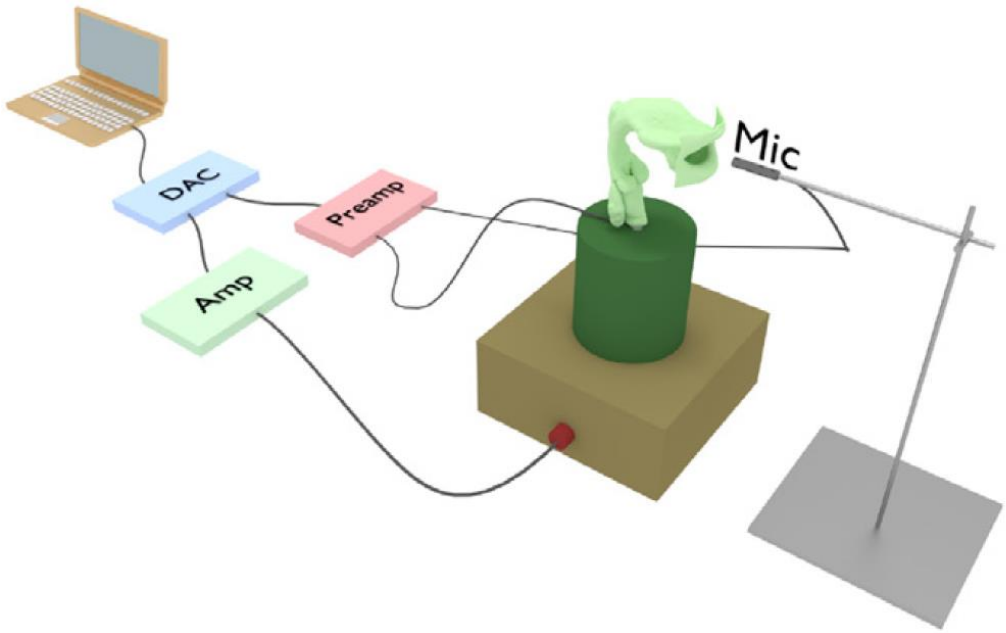
Repository II

- Synthetic vowels produced by a physical modelling approach (by Brad Story)
- Stimuli are created by modelling physical laws in sound production and transmission



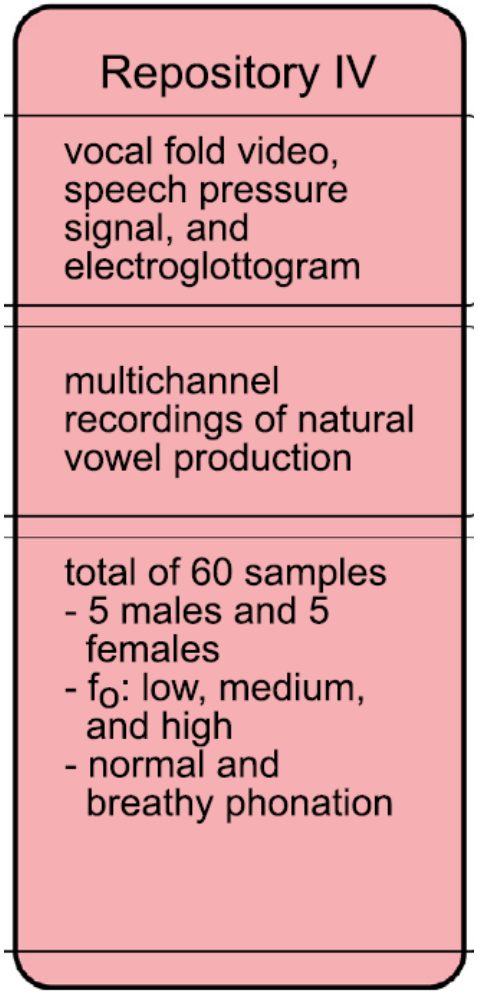
Repository III

- Physical production of vowel stimuli with a loudspeaker and a plastic 3D replica of the vocal tract
- Sweep measurements for the vocal tract



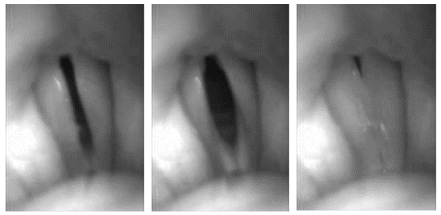
Left: Stimulus generation system

Right: 3D-printed vocal tract (of the vowel /a/)



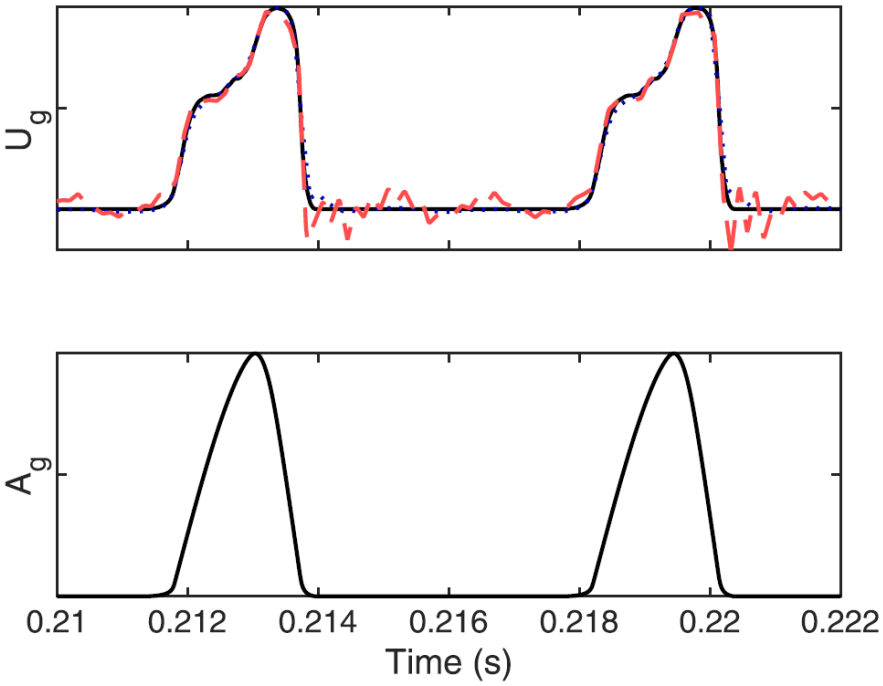
Repository IV

- Data from multi-channel recordings of natural vowel production: acoustic speech signal, EGG, high-speed imaging of the vocal folds



Left: Simultaneous recording of speech, electroglottography (EGG) and high-speed digital imaging of the vocal folds. **Right:** Three examples of still images from the recorded video. The images depict the vocal folds at different phases of the glottal cycle.

Example of the use of OPENGLOT (Repository II)

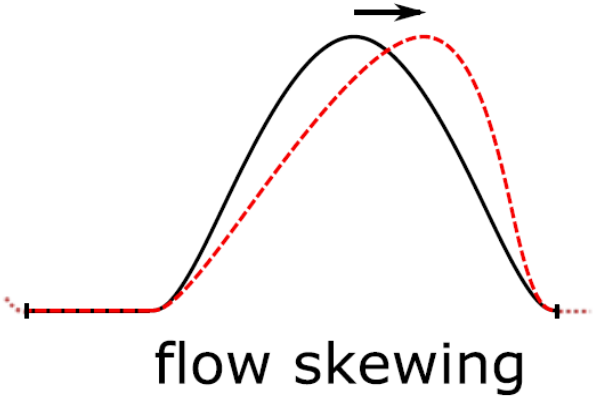


Top panel: Reference glottal flow (solid), as well the glottal flows estimated using two GIF methods (dashed, dotted). Bottom panel: Area of the glottal opening.

2.2 To study glottal flow vs. area in vowel production

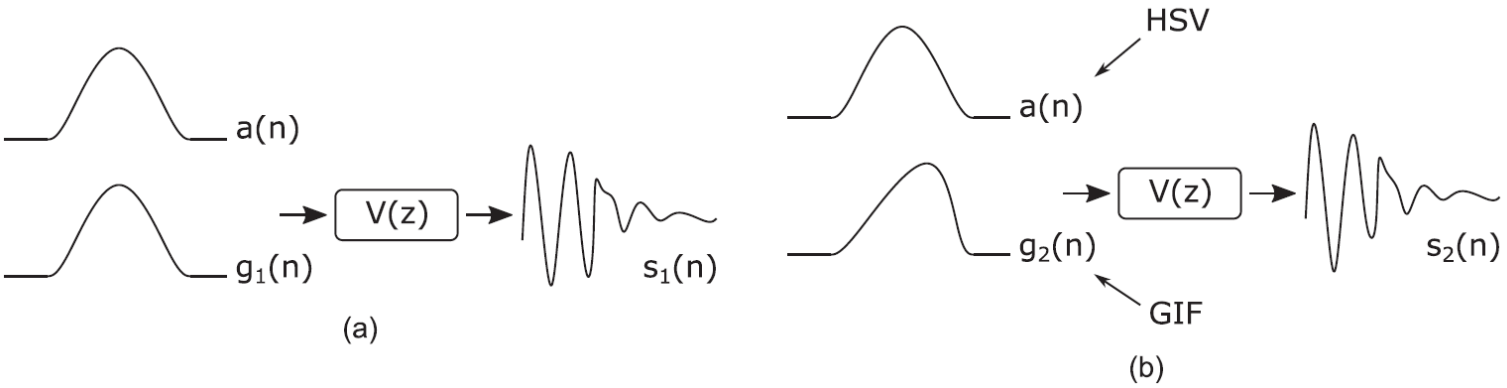
- Multi-channel data has been used, for example, in studying the following topics
  - Using GIF to extract parameters for a physical model of voice production (Murtola, Alku, Malinen, Geneid: *Parameterization of a computational physical model for glottal flow using inverse filtering and high-speed videoendoscopy*. Speech Communication, Vol. 96, pp. 67-80, 2018)
  - Studying phonation onsets in vowel production (Murtola, Malinen, Geneid, Alku: *Analysis of phonation onsets in vowel production, using information from glottal area and flow estimate*. Speech Communication, Vol. 109, pp. 55-65, 2019)

- Studying glottal flow skewing glophonation onsets in vowel production (Alku, Murtola, Malinen, Geneid, Vilkman: *Skewness of the glottal flow with respect to the glottal area measured in natural production of vowels*. Journal of the Acoustical Society of America, Vol. 146, No. 4, pp. 2501-2509, 2019)



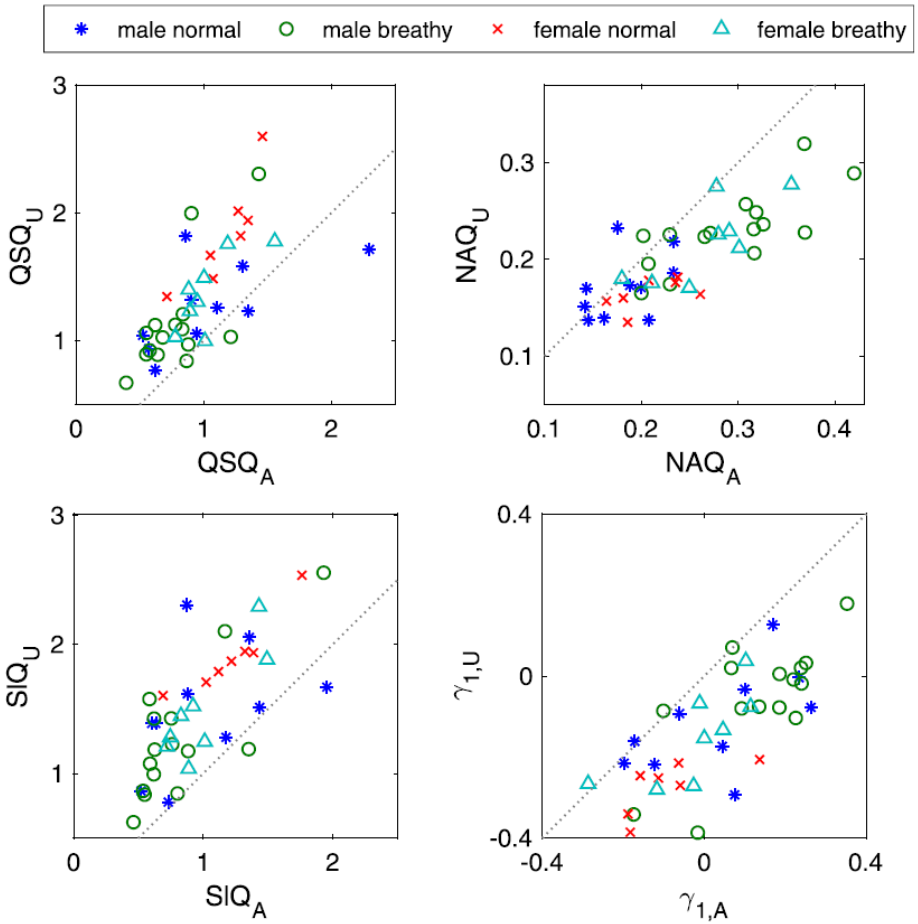
Skewing of the glottal flow (red curve) with respect to the glottal area (black curve).

- Flow skewing is caused by the inertia of the air column of the vocal tract
- Flow skewing is a known phenomenon that has been studies previously (already in the 80’s) using analog circuits and later using computer simulations
- Studies using natural speech are, however, sparse and they contain little quantitative data
- Potential reasons for the lack of studies in natural production of speech: (1) difficult to simultaneously measure glottal flow and area, (2) the phenomenon involves nonlinearity

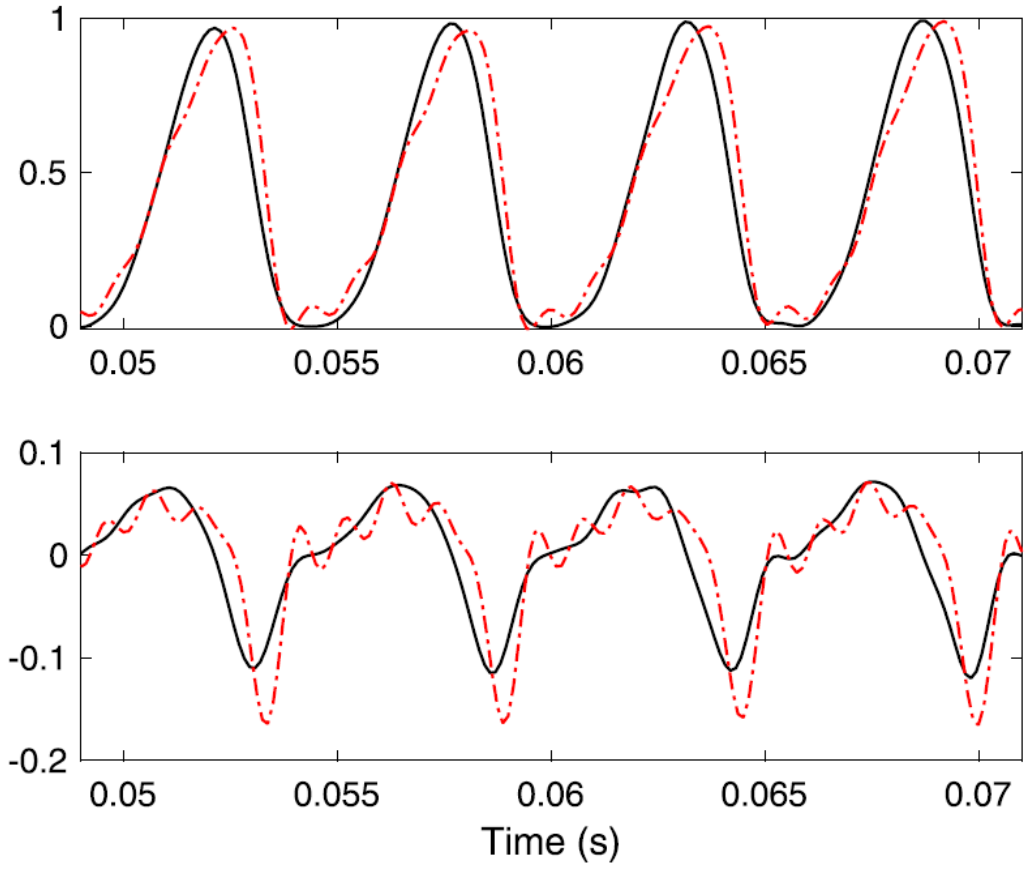


A simplified chart of vowel production in the case of a non-interactive (a) and interactive (b) vocal tract.

- Glottal area pulse, denoted by a(n), is the same in (a) and (b).
- Due to source–filter interaction, glottal flow $g_2(n)$ in (b) is more skewed to the right than glottal flow $g_1(n)$ in (a).
- Skewing of glottal flow is measured by estimating the skewed glottal pulse from $s_2(n)$ using GIF and by measuring area pulse a(n) from simultaneously recorded high-speed imaging of the vocal folds.



Values of four skewness parameters, extracted from both the glottal area (subscript A) and flow (subscript U) waveforms. Line y=x is shown for reference. Flow skewing manifests itself in dots that are located **above** the line in the parameters on left and **below** the line in the parameters on right.



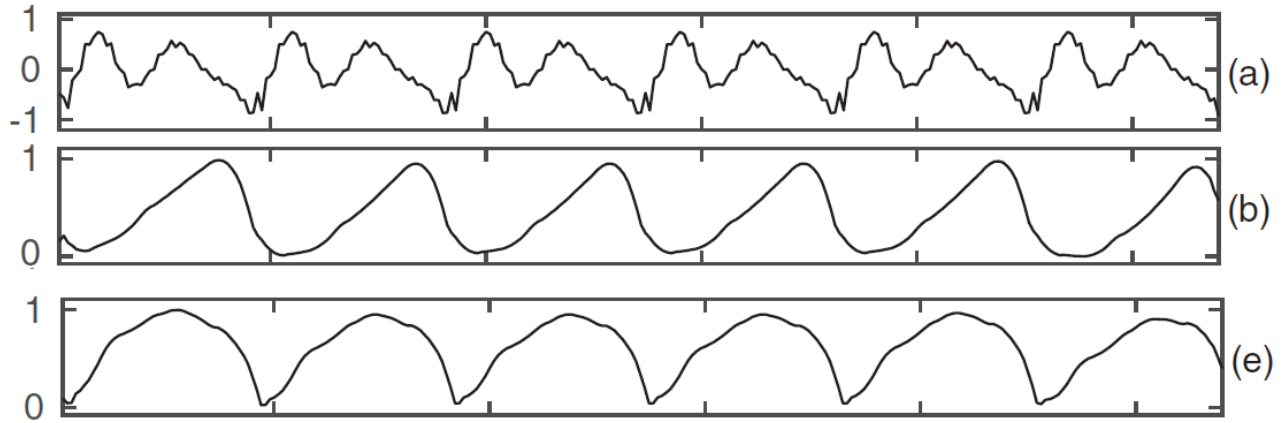
Examples of glottal area A (black curve) and glottal flow U (red curve) waveforms (upper panes) and their corresponding derivatives (lower panes) demonstrating skewness.

3. Speech-based biomarking of human health

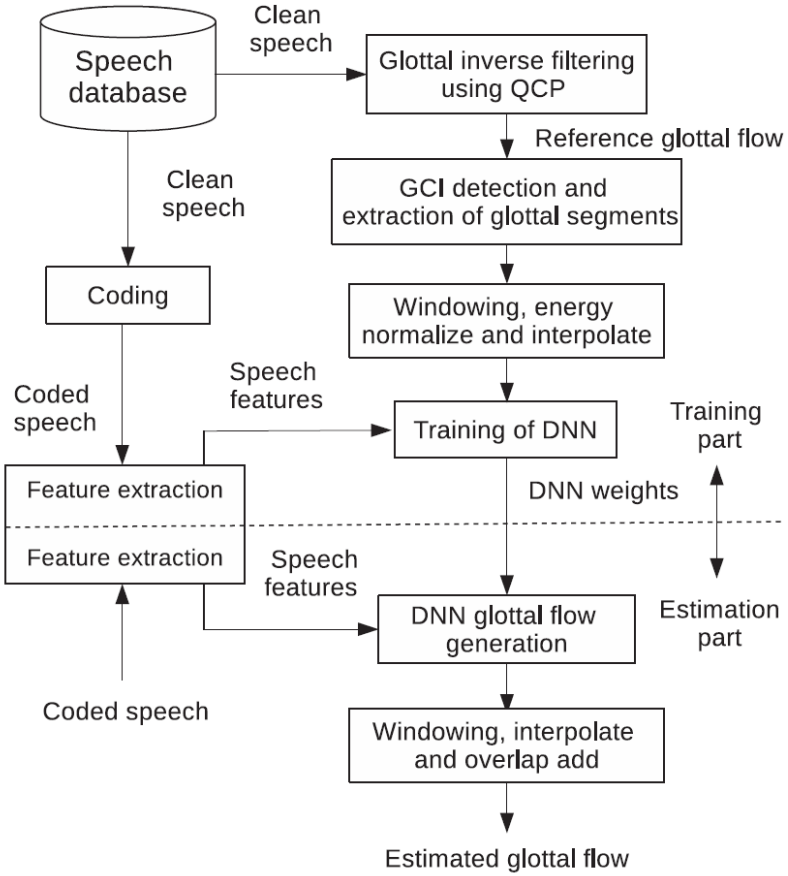
- Detection (and severity prediction) of diseases (e.g. Parkinson's disease) from speech is an established area of speech research that has gained momentum recently
- Progress in machine learning has made it possible to build systems of high performance using either (a) classical pipelines (front-end & back-end) or (b) end-to-end systems
- Our research has focussed on the use of glottal source, estimated with GIF, in classical pipeline systems

- A principle problem in using GIF with speech-based biomarking applications: GIF calls for high-quality recordings (e.g. linear phase microphones) which is not fulfilled, for example, in coded telephone speech

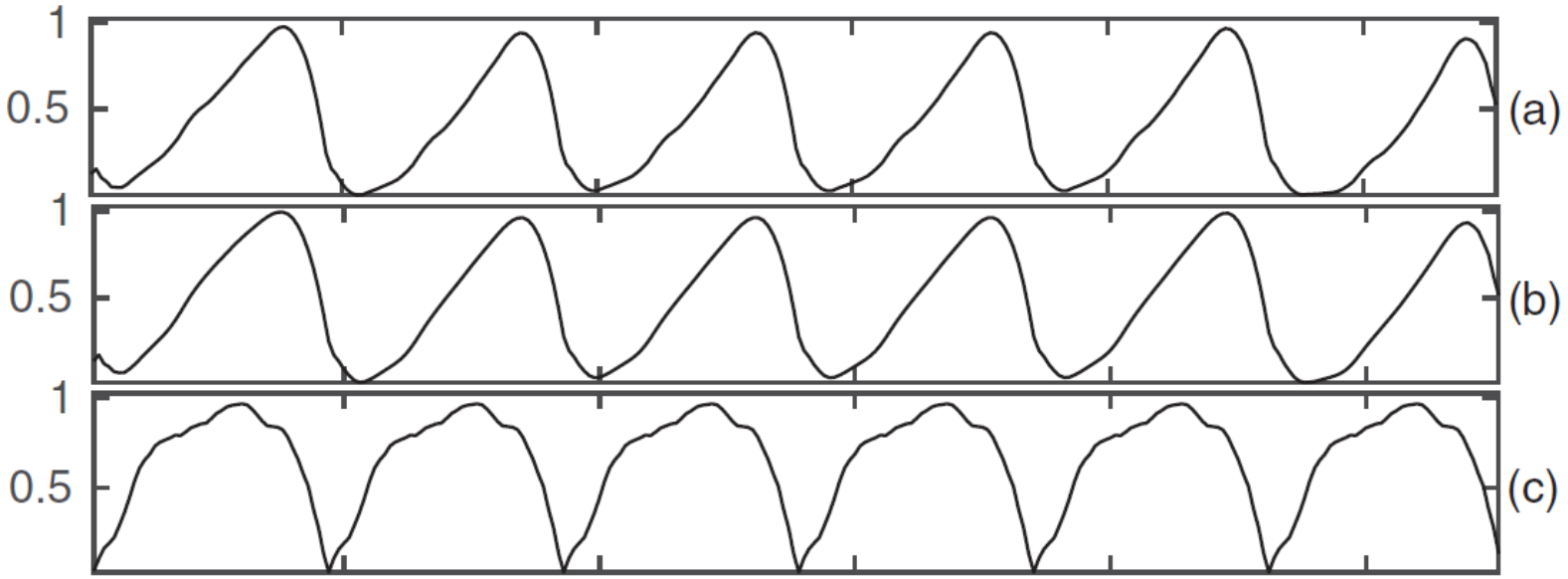
- To tackle the above problem, a deep neural network (DNN) –based GIF method was studied (Narendra, Airaksinen, Story, Alku: *Estimation of the glottal source from coded telephone speech using deep neural networks*. Speech Communication, Vol. 106, pp. 95-104, 2019)



Distortion of the glottal flow when the input is coded telephone speech: (a) speech, (b) glottal flow estimated from clean high-quality speech, (c) glottal flow estimated from telephone speech.

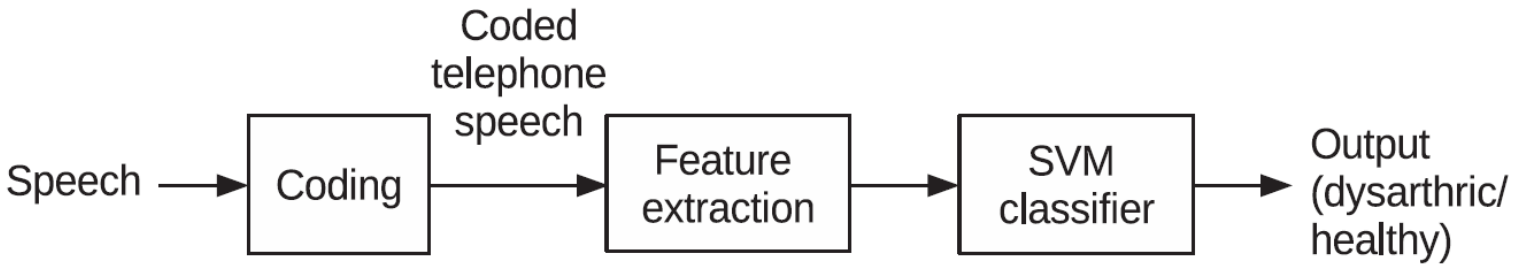


DNN-based estimation of the glottal flow from coded speech.

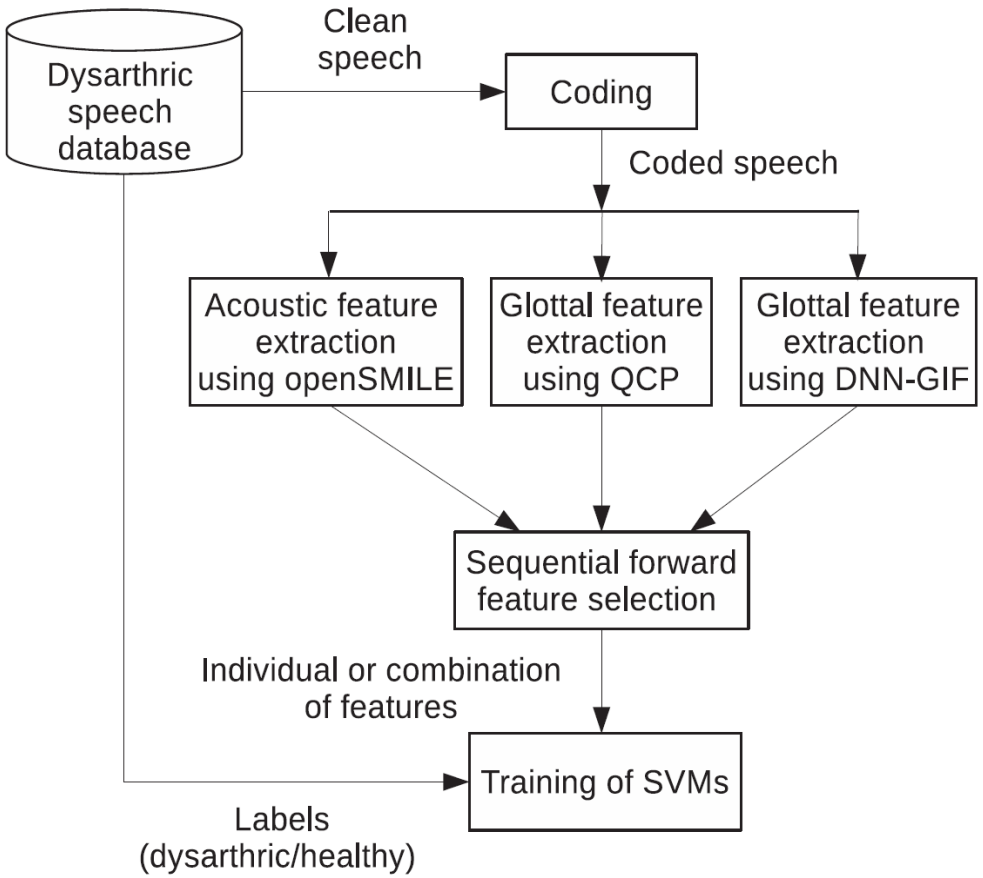


Examples of glottal flows:
(a) estimated from clean speech with a conventional GIF method
(b) estimated from telephone speech with the DNN-based GIF method
(c) estimated from telephone speech with a conventional GIF method

- Using glottal information in dysarthria detection from coded telephone speech (Narendra, Alku: *Dysarthric speech classification from coded telephone speech using glottal features*. Speech Communication, Vol. 110, pp. 47-55, 2019)



Binary classification using a classical pipeline with a support vector machine (SVM) classifier.



Training phase of the studied classification method.

Feature set (NB-coded)	Classification accuracy	
	Without feature selection (%)	With feature selection (%)
OpenSMILE-1	63.87	77.71
OpenSMILE-2	64.49	82.79
Glottal-1 (QCP)	43.52	64.12
Glottal-2 (QCP)	60.48	63.60
Glottal-1 (DNN-GIF)	54.01	72.76
Glottal-2 (DNN-GIF)	63.75	77.34
OpenSMILE-1 + Glottal-1 (QCP)	56.66	79.50
OpenSMILE-2 + Glottal-1 (QCP)	62.49	83.59
OpenSMILE-1 + Glottal-2 (QCP)	65.52	79.19
OpenSMILE-2 + Glottal-2 (QCP)	64.67	82.93
OpenSMILE-1 + Glottal-1 (DNN-GIF)	61.70	81.71
OpenSMILE-2 + Glottal-1 (DNN-GIF)	63.47	84.36
OpenSMILE-1 + Glottal-2 (DNN-GIF)	67.49	81.62
OpenSMILE-2 + Glottal-2 (DNN-GIF)	64.03	84.82

Feature set (NB-coded)	Classification accuracy	
	Without feature selection (%)	With feature selection (%)
OpenSMILE-1	90.42	91.18
OpenSMILE-2	95.11	95.25
Glottal-1 (QCP)	69.75	74.31
Glottal-2 (QCP)	67.60	68.58
Glottal-1 (DNN-GIF)	78.06	78.51
Glottal-2 (DNN-GIF)	81.49	80.54
OpenSMILE-1 + Glottal-1 (QCP)	87.66	91.70
OpenSMILE-2 + Glottal-1 (QCP)	94.40	95.64
OpenSMILE-1 + Glottal-2 (QCP)	88.39	91.22
OpenSMILE-2 + Glottal-2 (QCP)	94.85	95.58
OpenSMILE-1 + Glottal-1 (DNN-GIF)	89.63	91.82
OpenSMILE-2 + Glottal-1 (DNN-GIF)	95.17	95.81
OpenSMILE-1 + Glottal-2 (DNN-GIF)	88.03	91.99
OpenSMILE-2 + Glottal-2 (DNN-GIF)	95.20	96.07

Classification accuracy obtained for narrowband (NB) coded telephone speech using the TORGO database (left) and the UA-Speech database (right).

4. Summary

- Aalto University (Finland): 3 teams in speech research one of which in speech communication technology
- Three topics studied by the team were discussed: TTS, multi-channel recordings in vowel production, speech-based biomarking of human health…..all these three use modelling of voice production in one form or another
- Visitors and interns are welcome!