Comparison of tongue contour extraction methods from ultrasound images for use in Text-To-Speech synthesis

Tamás Gábor Csapó, Steven M. Lulich

csapot@tmit.bme.hu, slulich@indiana.edu

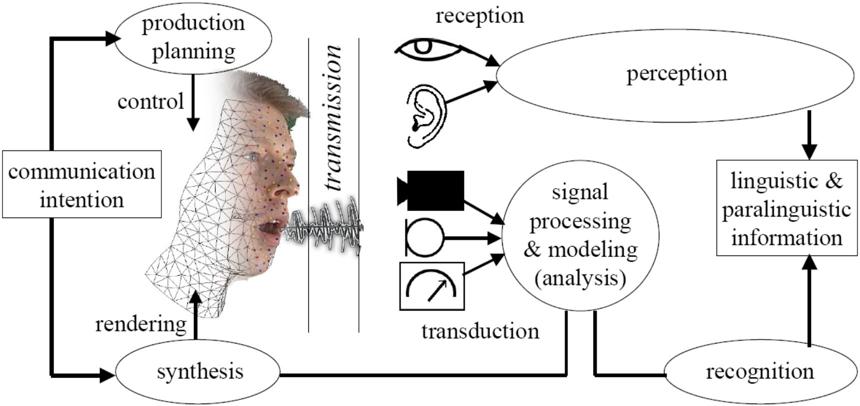
Inaugural Conference of the Hungarian Cultural Association
April 6, 2014

MŰEGYETEM 1782

INDIANA UNIVERSITY
BLOOMINGTON

1. Introduction
   - Text-To-Speech synthesis
   - Phonetic research with ultrasound
   - Goals of this study
2. Methods
   - Recordings
   - Manual tongue contour tracing
   - Automatic tongue contour tracking
3. Results
   - Analysis of manual tongue contour tracing
   - Analysis of automatic tongue contour tracking
4. Summary

Speech communication chain



[Fagel, 2007]

Text-To-Speech synthesis (TTS) I

- important in human-computer communication
- applications like talking robot, car speech interface
- helpful for the visually and speech impaired people to access and share information
- samples from state-of-the-art technique
  - English (🔊 click)
  - Hungarian (🔊 click)
- highly intelligible
- still far away from natural speech

[Németh and Olaszy, 2010, Zen et al., 2009, Tóth and Németh, 2010]

Text-To-Speech synthesis (TTS) II

Audiovisual TTS

- adding articulatory features might improve TTS quality
- tongue movement
- lip motion
- talking head (click)

[Ling et al., 2009, Schabus et al., 2014]

Speech research with ultrasound I

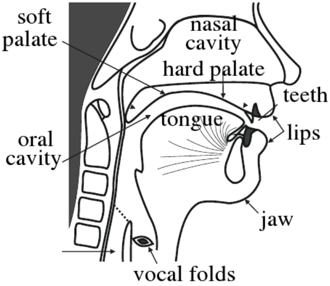
Ultrasound (US)

- used in speech research since early ’80s
- US transducer positioned below the chin during speech
- record video of tongue movement
- series of gray-scale images
- tongue surface has a greater brightness than the surrounding tissue and air

[Stone et al., 1983, Stone, 2005]

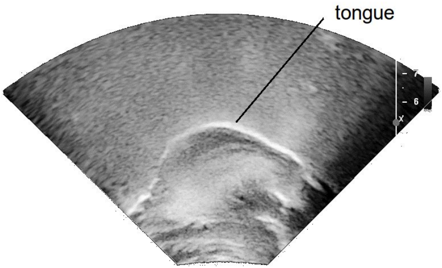
Speech research with ultrasound II

Vocal tract



[Németh and Olaszy, 2010]

Ultrasound sample



(click)

Speech research with ultrasound III

Phonetic research examples

- reconstruct tongue shape during sustained vowels
- investigate speech sounds of under-researched languages
- compare articulatory characteristics of vowels
- analyze tongue shapes for clinical purposes

First step is always the tongue contour tracking!

[Stone and Lundberg, 1996, Mielke et al., 2011, Benus and Gafos, 2007, Zharkova, 2013]

Our goals

This study

- compare manual tongue tracings of several individuals
- compare automatic tongue contour extraction programs
- use 2D ultrasound at high frame rate

Long-term

- extend TTS with tongue contour data based on ultrasound
- include tongue movement in audiovisual speech synthesis (e.g. talking head)
- use real-time 3D ultrasound

Methods

Subjects

- two female and two male
- 3 speakers of American English
- 1 speaker of Hungarian

Speech material

- '*I owe you a yo-yo.*' sentence two times

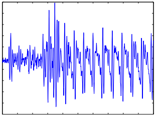
Recordings

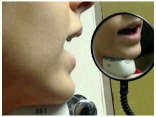
Location

- Speech Production Lab
- Dept. of Speech and Hearing Sciences
- Indiana University

Parallel recordings

- speech signal with a microphone
- video of the lips with a webcamera
- video of the tongue with an ultrasound device (Philips EpiQ-7G, xMatrix 6-1 MHz)





Recording setup

Ultrasound recordings

DICOM video

- 40–45 frames / second
- 800x600 pixels resolution
- 0.2 mm / pixel

JPG image sequence

- altogether 1 145 US tongue images
- (389, 275, 241 and 240 for the 4 speakers)

Manual tracings

Tracers

- 7 individuals (2 authors and 5 students)
- drag a computer mouse cursor
  from the root of the tongue (left)
  to the tip of the tongue (right)
- about 150–200 points per image
- about 5–10 seconds per image

Manual tracing website

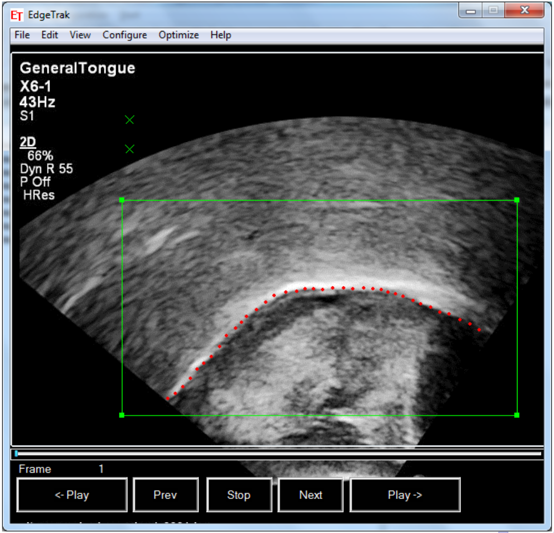
Automatic tongue contour tracking algorithms

4 freely available programs, baseline settings

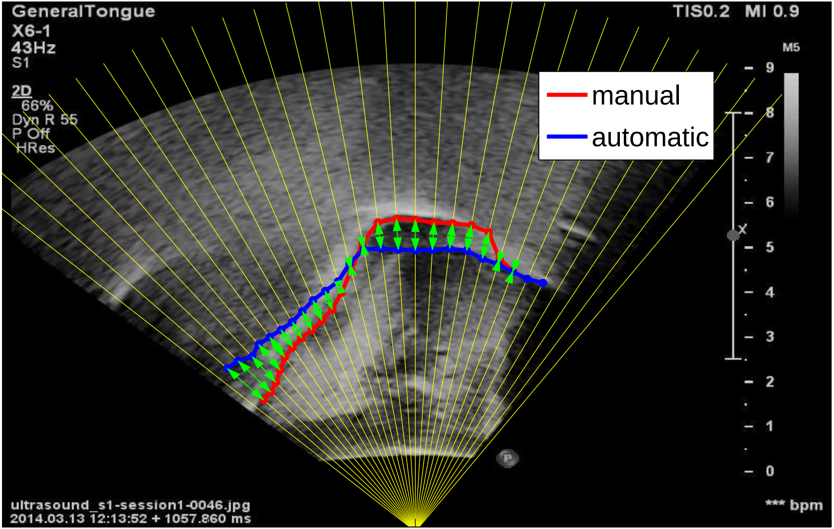
- EdgeTrak (University of Maryland, USA)
- Palatoglossotron (North Carolina State University, USA)
- TongueTrack (Simon Fraser University, Canada)
- Ultra-CATS (University of Toronto, Canada)

[Li et al., 2005, Baker et al., 2005, Tang et al., 2012, Bressmann et al., 2005]

EdgeTrak sample



Comparison of two tongue contours



Manual tracings

RMSE (Root Mean Squared Error) difference from mean

- Average: 7.11 pixel (1.42 mm)
- Std. dev.: 5.07 pixel (1.01 mm)
- depending on the speaker, tracer and image

US video samples

- speaker1 (click)
- speaker4 (click)

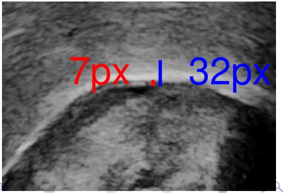
Automatic trackings

RMSE (Root Mean Squared Error)
difference from mean of manual tracing

- Average: 32.30 pixel (6.46 mm)
- Std. dev.: 29.06 pixel (5.81 mm)
- depending on the speaker, program and image
- (compare with: 7.11 pixel inter-tracer variability)

US video samples

- speaker1 (click)
- speaker4 (click)



Average differences of automatic trackings from manual tracings

Table: Average RMSE differences (in pixels)

software	spkr1	spkr2	spkr3	spkr4	avg
EdgeTrak	15.10	9.01	12.00	32.19	17.08
Palatoglossotron	33.11	46.60	86.84	95.82	65.59
TongueTrack	14.86	14.73	20.48	19.50	17.39
Ultra-CATS	57.98	34.67	36.71	38.27	41.91

(compare with: 7.11 pixel inter-tracer variability)

Summary

This study

- ultrasound recordings with 4 speakers
- compared manual tongue tracings of 7 individuals
- compared 4 automatic tongue contour extraction programs

Future plans

- extend Hungarian / English Text-To-Speech with tongue contour data
- use 2D / real-time 3D ultrasound
- include tongue movement in audiovisual speech synthesis (e.g. talking head)

Acknowledgements

Support from

- Fulbright Hungary
- Hungarian Academy of Engineering

Thanks to

- Manual tongue contour tracings
    - Alexandra Abell, Sarah Janssen, Denice King, Rebecca Pedro, Schanna Schmutte
- Automatic tongue contour tracking
    - Adam Baker, Tim Bressmann, Jeff Mielke, Maureen Stone, Lisa Tang
- Manual tracing website
    - Hao Lu

References I

Baker, A., Mielke, J., and Archangeli, D. (2005).
Tracing the tongue with GLoSsatron.
In *Ultrafest III*, Tucson, Arizona, USA.

Benus, S. and Gafos, A. I. (2007).
Articulatory characteristics of Hungarian 'transparent' vowels.
Journal of Phonetics, 35(3):271–300.

Bressmann, T., Heng, C.-L., and Irish, J. C. (2005).
Applications of 2D and 3D ultrasound imaging in speech-language pathology.
Journal of Speech-Language Pathology and Audiology, 29(4):158–168.

Fagel, S. (2007).
Audiovisual Speech: Analysis, Synthesis, Perception and Recognition.
In *Proc. ICPhS*, pages 275–278, Saarbrücken, Germany.

Li, M., Kambhamettu, C., and Stone, M. (2005).
Automatic contour tracking in ultrasound images.
Clinical Linguistics & Phonetics, 19(6-7):545–554.

Ling, Z.-H., Richmond, K., Yamagishi, J., and Wang, R.-H. (2009).
Integrating Articulatory Features Into HMM-Based Parametric Speech Synthesis.
IEEE Transactions on Audio, Speech, and Language Processing, 17(6):1171–1185.

References II

Mielke, J., Olson, K. S., Baker, A., and Archangeli, D. (2011).
Articulation of the Kagayanen interdental approximant: An ultrasound study.
Journal of Phonetics, 39(3):403–412.

Németh, G. and Olaszy, G., editors (2010).
A MAGYAR BESZÉD; Beszédkutatás, beszédtechnológia, beszédinformációs rendszerek.
Akadémiai Kiadó, Budapest.

Schabus, D., Pucher, M., and Hofer, G. (2014).
Joint Audiovisual Hidden Semi-Markov Model-based Speech Synthesis.
IEEE Journal of Selected Topics in Signal Processing.

Stone, M. (2005).
A guide to analysing tongue motion from ultrasound images.
Clinical Linguistics & Phonetics, 19(6-7):455–501.

Stone, M. and Lundberg, A. (1996).
Three-dimensional tongue surface shapes of English consonants and vowels.
The Journal of the Acoustical Society of America, 99(6):3728–37.

Stone, M., Sonies, B., Shawker, T., Weiss, G., and Nadel, L. (1983).
Analysis of real-time ultrasound images of tongue configuration using a grid-digitizing system.
Journal of Phonetics, 11:207–218.

References III

Tang, L., Bressmann, T., and Hamarneh, G. (2012).
Tongue contour tracking in dynamic ultrasound via higher-order MRFs and efficient fusion moves.
Medical Image Analysis, 16(8):1503–1520.

Tóth, B. and Németh, G. (2010).
Improvements of Hungarian Hidden Markov Model-based Text-to-Speech Synthesis.
Acta Cybernetica, 19(4):715–731.

Zen, H., Tokuda, K., and Black, A. W. (2009).
Statistical parametric speech synthesis.
Speech Communication, 51(11):1039–1064.

Zharkova, N. (2013).
A normative-speaker validation study of two indices developed to quantify tongue dorsum activity from midsagittal tongue shapes.
Clinical Linguistics & Phonetics, 27(6-7):484–96.