Prospects of Articulatory Text-to-Speech Synthesis

Simon Stone, Atilla Azgin, Sabrina Mänz, Peter Birkholz

Articulatory synthesis

A powerful alternative

- Speech synthesis in commercial applications unit-selection or end-to-end synthesis
- Still, articulatory synthesis holds great potential to surpass the state-of-the-art [1]
- One major obstacle: accessibility for non-phonetics experts

The VocalTractLab

Synthesis from articulatory gestures

- Free and open-source software (www.vocaltractlab.de)
- Combines aero-dynamic, articulatory and acoustic models in a synthesis pipeline [2]
- Controlled in terms of articulatory trajectories using a gestural score [3]

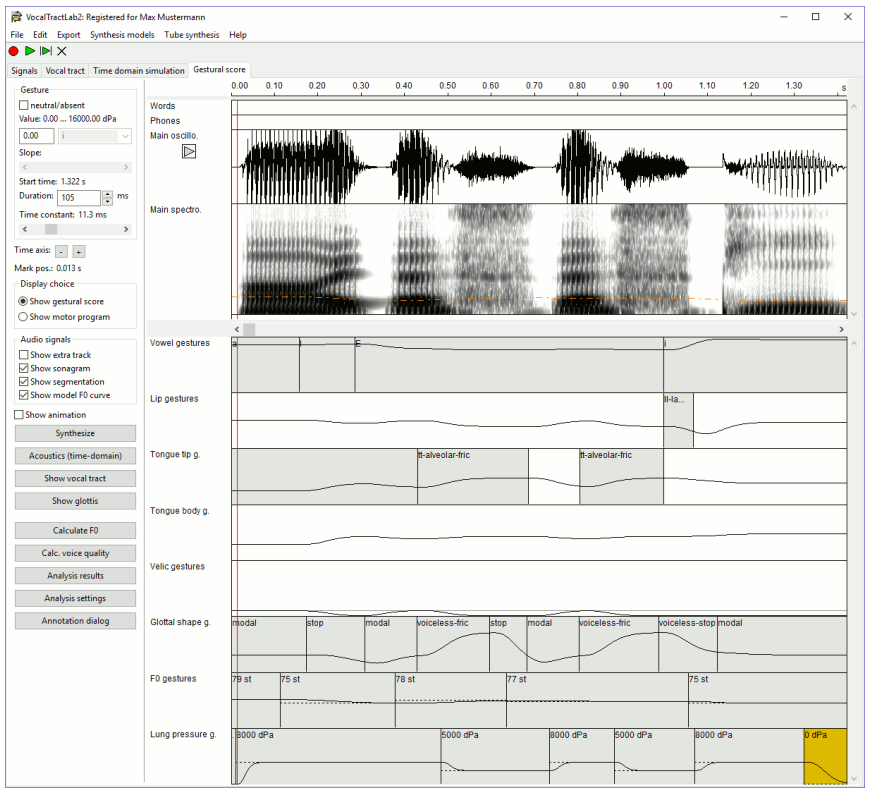

Figure 1: VocalTractLab's gestural score editor

- Gestural scores are extremely powerful and allow fine-grained control, but are somewhat obscure for untrained users
- Text-driven frontend very desirable
- Suggested workflow:
  - ✓ Initialize from text representation
  - ✓ Fine-tune at articulatory level

Articulatory Text-to-Speech

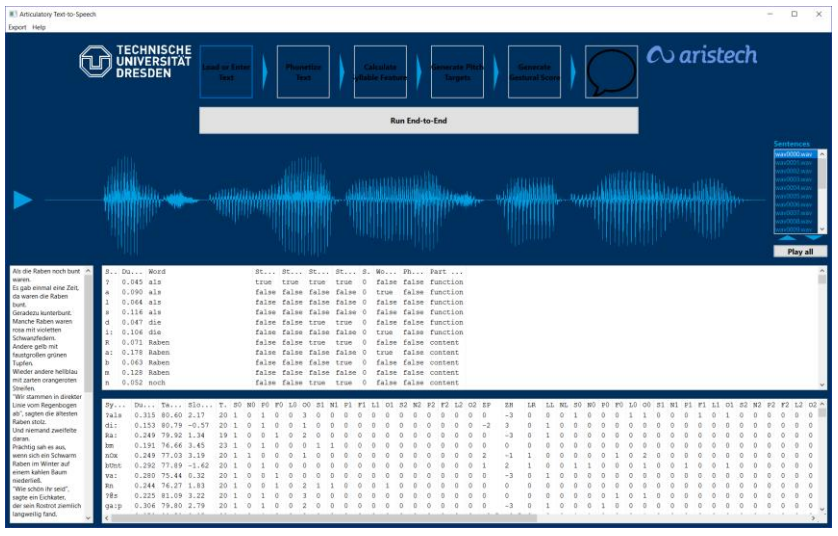

Figure 2: Articulatory Text-to-Speech

Step 1: Enter text

- Type in text or load from file

Step 2: Phonetic transcription

- Web interface by project partner Aristech GmbH provides grapheme-to-phoneme conversion, syllabification, Part-of-Speech tagging, stress marker insertion

Step 3: Calculate syllable features

- Intonation generation based on syllable structure
- Feature vector for each syllable including 70 phonetic, linguistic, and prosodic features

Step 4: Generate intonation

- Predict pitch target for each syllable
- Generate phone durations based on [4]

Step 5: Generate gestural score

Tiered approach to create score from segment sequence:

- First vowels, then fricatives, then stops, nasals, and the glottal fricative
- Timing of onset and offset in each tier carefully adjusted to match acoustic phone durations (≠ articulatory gesture durations!)

Step 6: Generate speech audio

- Using VocalTractLab synthesis backend

Outlook

- Replace proprietary components and integrate into VocalTractLab
- Include other languges (currently only German)
- Improve overall quality

References

[1] C. H. Shadle and R. I. Damper, "Prospects for articulatory synthesis: A position paper," in *Fourth ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis (SSW-4)*, Perthshire, Scotland, 2001.

[2] Birkholz, Peter. "Modeling consonant-vowel coarticulation for articulatory speech synthesis." *PloS one* 8, no. 4 (2013): e60603.

[3] Browman, Catherine P., and Louis Goldstein. "Articulatory phonology: An overview." *Phonetica* 49, no. 3-4 (1992): 155-180.

[4] Möbius, Bernd, and J. Von Santen. "Modeling segmental duration in German text-to-speech synthesis." In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, vol. 4, pp. 2395-2398. IEEE, 1996.