Abstract for ESSV 2010 (by Eva Lasarcyk): Acoustics vs. articulation in articulatory speech synthesis: One vocal tract target configuration has more than one sound.

The goal of this contribution is to illustrate the importance of the acoustic settings of articulatory speech synthesis when using it for perception/validation experiments regarding the relationship between articulation and fine phonetic detail in the acoustic domain. In our case, we focus on details of vowel height, lip rounding, and horizontal tongue position.

Articulatory synthesis is capable of imitating or modeling fine phonetic detail and has been applied to various aspects of articulation, such as e.g. nasal coupling and vowel height (Krakow et al. 1985), displacement of the tongue body center and formant values (Perkell & Nelson 1985), acoustic and perceptual characteristics of smiled vowels (Lasarcyk & Trouvain 2008), or impressions of body size vs. anger and joy (Xu & Chuenwattanapranithi 2007).

Articulatory synthesizers are thus a tool for phonetic experiments to systematically investigate articulatory patterns in idealized, i.e. synthetic, vocal tracts. One advantage of using artifical vocal tracts is that, in principle, the results are reproducable because the computer program can be executed again.

A common working assumption seems to be that a specific target configuration of the vocal tract always yields the same acoustic signal and reproducable auditory percept. Therefore, reports on the methods used in studies often omit a detailed description of the acoustic settings for generating the speech samples. We argue that this may underspecify the experimental descriptions and may be a reason to render them non-reproducable for others. The underspecification may lead to different sounding acoustic samples stemming from identical vocal tract target configurations. This may happen without explicit awareness even of the experimenter if one is not aware of all relevant influences on the generated sound wave and does not document in sufficient details the settings that were used.

In our case, we imitated the acoustics of several tense German vowels with the synthesizer VocalTractLab (Birkholz 2006) by designing appropriate vocal tract target configurations (work in progress). These were then used to generate perception test vowel stimuli. They were to be transcribed by a number of phoneticians to validate that the theoretical acoustic formant properties really created the correct vocalic phoneme impression.

At first, we did not pay attention to the acoustic settings of the articulatory synthesizer and simply used the default settings for generating gestural scores that would utter the desired vowels. However, the resulting vowel sounds were different from the ones generated internally during the process of designing the appropriate vocal tract targets (target configurations for each vowel). The differences were exlusively located in variables *outside* the per-vowel defined target configuration. They included the excitation model, the intonation contour of the vowel, the control of the position of the velum, and certain other aerodynamic-acoustic properties.

Due to the conspicuous differences in the audio, we generated both sets of vowels and had all of them evaluated systematically by phonetic transcription. The results showed systematic differences in all three articulatory dimensions, i.e. vowel height, horizontal tongue position, and lip rounding. Since one and the same underlying vocal tract target configuration even yielded different phonemes (not only within-phoneme changes), it seems essential to be aware of the acoustic settings used. If, by accident, these are not consistent, this may result in low intra-study consistency and low inter-study comparability.

To make a point, if a phonetic study aims at imitating fine phonetic details and validates the articulatory patterns by auditive assessment and transcription, it is essential to assure that the experimental variables are not influenced unexpectedly by settings of the acoustic rendering of the vocal tract target configuration. Otherwise, this may foil the study and hide links between articulation and acoustics that otherwise would have been transparent.

Peter Birkholz (2006). *3D-Artikulatorische Sprachsynthese*. Berlin: Logos

Rena A. Krakow, Patrice S. Beddor, Louis M. Goldstein, and Carol A. Fowler (1985). Effects of contextual and noncontextual nasalization on perceived vowel height. *J. Acoust. Soc. Am.* 77, S8.

Eva Lasarcyk and Jürgen Trouvain (2008): Spread Lips + Raised Larynx + Higher F0 = Smiled Speech? - An Articulatory Synthesis Approach. *Proc. 8th International Speech Production Seminar (ISSP)*, Strasbourg, December 8-12. 345-348.

Joseph S. Perkell and Winston L. Nelson (1985). Variability in production of the vowels /i/ and /a/. *J. Acoust. Soc. Am.* 77 (5), 1889-1895.

Yi Xu and Suthathip Chuenwattanapranithi (2007). Perceiving anger and joy in speech through the size code. In *Proc. 16th International Congress of the Phonetic Sciences (ICPhS)*, Saarbrücken, August 6-10. 2105–2108.