ARTICULATION-TO-SPEECH USING ELECTRO-OPTICAL STOMATOGRAPHY AND ARTICULATORY SYNTHESIS

Simon Stone and Peter Birkholz
Technische Universität Dresden
simon.stone@tu-dresden.de

Abstract: Articulation-to-Speech (ATS) is a topic of increasing interest and relevance in the speech research community. An ATS system consists of an articulatory data acquisition frontend, a parametric speech synthesis backend, and a mapping from the articulatory space to the parameter space of the synthesizer in between (see Figure 1). Many studies (e.g., [1, 2, 3]) used vocoders as the synthesizer and thus tried to find mappings from the articulatory data to the vocoder parameters (usually Mel-Frequency Cepstral Coeffficients, MFCCs). Two issues are inherent to that approach: The vocoder input space is rather high dimensional (e.g., 36 or even more dimensions) making the training of the mapping rather complex. The second issue is the different domain of the input and output of the mapping: the input features live in an articulatory domain and the output parameters live in an acoustic domain. Lastly, and most importantly, vocoder synthesis quality is generally mediocre and even under ideal circumstances reduces the attainable speech quality significantly: In [2], the authors report that the naturalness rating of natural speech, when vocoded with ideal parameters (not obtained from any articulatory data but from the original speech signal itself) dropped from 94.82 % (original) to 56.22 % (vocoded). Similarly, the authors report a naturalness rating of just slightly above 80 % in [3] for audio signals resynthesized with an ideally parametrized vocoder. These examples show that even small errors in the mapping may lead to ultimately unnatural sounding speech. In this paper, we therefore propose to use an articulatory synthesizer as the synthesis backend. For an articulatory synthesizer to work in this framework, it needs to be parametric, ideally with a low number of parameters to minimize the mapping complexity, precise enough to allow natural sounding, intelligible speech, and fast enough to enable real-time processing in a closed loop of measuring the articulation and synthesizing the corresponding speech output. While several articulatory synthesis systems exist (i.e., [4, 5]), none of these fulfill all of the above criteria. We therefore replaced the 3d vocal tract model in the articulatory synthesizer VocalTractLab (VTL) [5] with a recently proposed sparse 1d parametric vocal tract model [6]. Since most of the vanilla VTL synthesis time (real-time factor 2 on a moderately powerful desktop PC) is spent on the geometric dimensional reduction from 3d to 1d, modeling the 1d area function directly provides the necessary performance boost to run the synthesis close to real-time including the data acquisition. To gather the articulatory data, we used a measurement technique called Electro-Optical Stomatography (EOS), which is a combination of Electropalatography (EPG) and Optical Palatography (OPG) extended by an optical lip sensor (see Figure 2) [7]. We recorded EOS data from four subjects during the articulation of all voiced German speech sounds in various contexts (253 items) and trained four different families of regression models (Least-Squares, Support Vector Machines (SVM), Random Forests, and Gaussian Processes) to map the EOS data to the corresponding 1d vocal tract shapes. Table 1 shows the 5-fold cross-validated loss of all mappings. The overall precision of the parameter predictions was comparatively high for all vowel sounds resulting in clearly intelligible and natural sounding vowel sounds. The synthesis of consonantal sounds was also very natural sounding but much less intelligible (examples of both online at http://www.vocaltractlab.de/index.php?page=birkholz-supplements). Responsible were most likely coarticulation effects, which was supported by a t-SNE projection of the input data showing almost identity between,

e.g., the sensor data of a /v/ in an /a/ context and of an /a/ itself. More data and, more importantly, a sequence-to-sequence mapping (using, e.g., a Long Short-Term Memory network) could potentially improve the results. The full paper outlines ways to obtain the necessary vocal tract parameter trajectories to train this mapping.

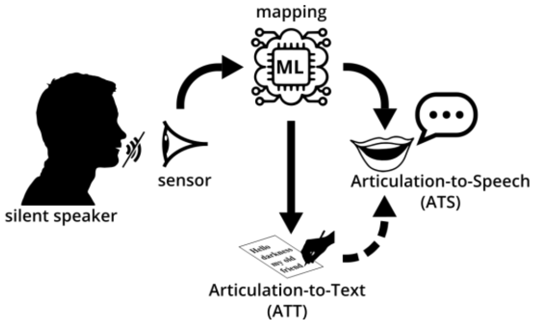

Figure 1: General framework of a Silent-Speech Interface.

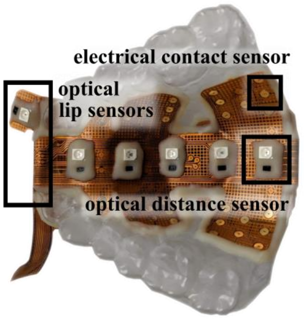

Figure 2: Example of an individually fitted pseudopalate of an EOS device.

	Least-Squares	SVM	Ensemble	GP
Subject 01	0.9794	0.8396	0.8226	0.8077
Subject 02	0.9378	0.8624	0.9378	0.7785
Subject 03	0.9313	0.8047	0.7276	0.9313
Subject 04	0.9078	0.6876	0.6783	0.6602

Table 1: Averaged 5-fold cross-validated loss for all vocal tract model parameter predictions made with Least-Squares regression, Support Vector Machines (SVM), Ensemble models (random forests), and Gaussian Process (GP) regression.

References

[1] M. Janke and L. Diener, “EMG-to-speech: Direct generation of speech from facial electromyographic signals,” IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 25, no. 12, pp. 2375–2385, 2017.

[2] T. G. Csapó, M. S. Al-Radhi, G. Németh, G. Gosztolya, T. Grósz, L. Tóth, and A. Markó, “Ultrasound-Based Silent Speech Interface Built on a Continuous Vocoder,” in Proc. Interspeech 2019, 2019, pp. 894–898. [Online]. Available: htp://dx.doi.org/10.21437/Interspeech.2019-2046

[3] J. A. Gonzalez, L. A. Cheah, A. M. Gomez, P. D. Green, J. M. Gilbert, S. R. Ell, R. K. Moore, and E. Holdsworth, “Direct speech reconstruction from articulatory sensor data by machine learning,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 12, pp. 2362–2374, Dec 2017.

[4] S. Fels, F. Vogt, K. Van Den Doel, J. Lloyd, I. Stavness, and E. Vatikiotis-Bateson, “Artisynth: A biomechanical simulation platform for the vocal tract and upper airway,” in Proc. of the 7th International Seminar on Speech Production, Ubatuba, Brazil, 2006.

[5] P. Birkholz, “Modeling consonant-vowel coarticulation for articulatory speech synthesis,” PLOS ONE, vol. 8, no. 4, p. e60603, 2013.

[6] S. Stone, M. Marxen, and P. Birkholz. "Construction and evaluation of a parametric one-dimensional vocal tract model." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26, no. 8 (2018): 1381-1392.

[7] S. Stone and P. Birkholz, “Cross-speaker silent-speech command word recognition using electro-optical stomatography,” in 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, submitted.