PROSPECTS OF ARTICULATORY TEXT-TO-SPEECH SYNTHESIS

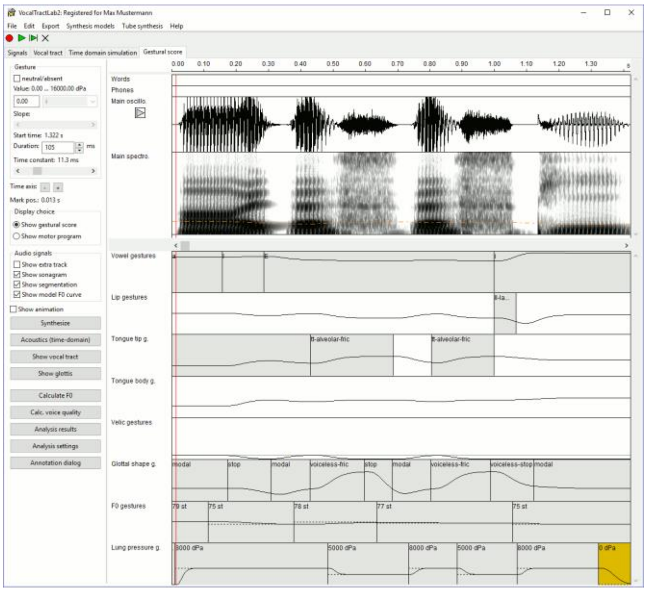
Simon Stone[1]*, Atilla Azgin*[2]*, Sabrina Mänz*[2]*, Peter Birkholz*[1]
[1]*Technische Universität Dresden, Germany*
[2]*Aristech GmbH, Heidelberg, Germany*
simon.stone@tu-dresden.de

Abstract: Text-to-Speech (TTS) synthesis has come a long way from the first (English) system by Matsui et al. [1], which used a cascade of analyses and models to get from a text representation to the acoustic speech signal, to the currently popular end-to-end systems (e.g., Tacotron [2]), which use a single large model (usually a recurrent neural network) to map input character sequences directly to the corresponding acoustic waveform. However, all current TTS systems use pre-recorded speech snippets (at least during the training process), which means that they are limited to the available units (in unit-selection systems) or at least to the voice and speaking style of the training material (in end-to-end systems). Any para-linguistic expressiveness, e.g., emotion, can only be "applied" after the "neutral" signal is synthesized using morphing techniques, which introduce audible signal-processing artifacts and can lower the perceived naturalness. In this paper, we therefore present our ongoing work on a TTS pipeline that uses an articulatory-based synthesis backend instead of a signal-based synthesizer. We use the VocalTractLab*'s [3] vocal tract, articulatory, and acoustic models to perform the synthesis. The input modality for the articulatory control is the so-called gestural score [4]: a sequence of parallel articulatory targets that are approximated at adjustable speed by the individual articulators (see Figure 1). While this kind of control over the vocal tract model is very powerful, flexible, and expressive, offering arguably the same degrees of freedom as the human vocal tract itself, it is also very time consuming to arrange the score for even a basic utterance and requires expert knowledge of the phonetic and articulatory intricacies. Therefore, articulatory synthesis has so far been confined to a niche in speech research despite its many advantages over unit-selection or waveform synthesis.

In this paper, we outline our efforts to bring articulatory synthesis closer to the mainstream by introducing a preprocessing pipeline that generates a gestural score (and thus, by extension, the speech signal) from plain, orthographic text. The first stage in an articulatory TTS pipeline is similar to, e.g., unit selection: The transliteration of the orthographic text into a syllabified, phonetic representation (grapheme-to-phoneme conversion). Currently, this stage uses a proprietary web-service by our partner Aristech GmbH using a hand-transcribed lexicon and a Finite-State-Transducer to handle out-of-vocabulary words. After this stage, the pipeline diverges from the usual procedure, because articulatory synthesis requires very low-level information: Every articulatory gesture requires a target position, a duration, and a time constant (controlling the speed of the articulatory target approximation). The sequence of target positions can be derived from the phonetic transcription, since every phoneme corresponds uniquely to a vocal tract configuration. For the timing information, the acoustic duration of each syllable is determined using a Deep Neural Network and a combination of phonetic and linguistic features (details in the full paper). However, the mapping was trained using acoustic sound durations, because no training data to predict the gestural durations directly is currently available. Therefore, these durations are only used to initialize the gestures and their true timing is found in an iterative way to match the simulated articulatory landmarks with the given acoustic landmarks (phone boundaries). Finally, the gestural score is generated and used to synthesize

* http://www.vocaltractlab.de

the speech signal. Based on this neutral gestural score, many manipulations can be introduced to add para-linguistic information (e.g., changing the pitch offset and range, the articulatory precision, the phonation type, the subglottal pressure, and many more). In contrast to morphing-based signal processing techniques to turn a neutral speech *signal* into an expressive speech signal, manipulations at the *articulatory* level are much more realistic sounding and can focus on individual features without unintentional side effects. Some audio examples of such manipulations are available at http://www.vocaltractlab.de/index.php?page=birkholz-supplements.



Gesture
☐ neutral/absent
Value: 0.00 ... 16000.00 dPa
0.00
Slope:
Start time: 1.322 s
Duration: 105 ms
Time constant: 11.3 ms
Time axis:
Mark pos.: 0.013 s
Display choice
● Show gestural score
○ Show motor program
Audio signals
☐ Show extra track
☑ Show sonagram
☑ Show segmentation
☑ Show model F0 curve
☐ Show animation

Figure 1: Gestural score for the utterance "ISSP" (/aI ?Es ?Es pi:/).

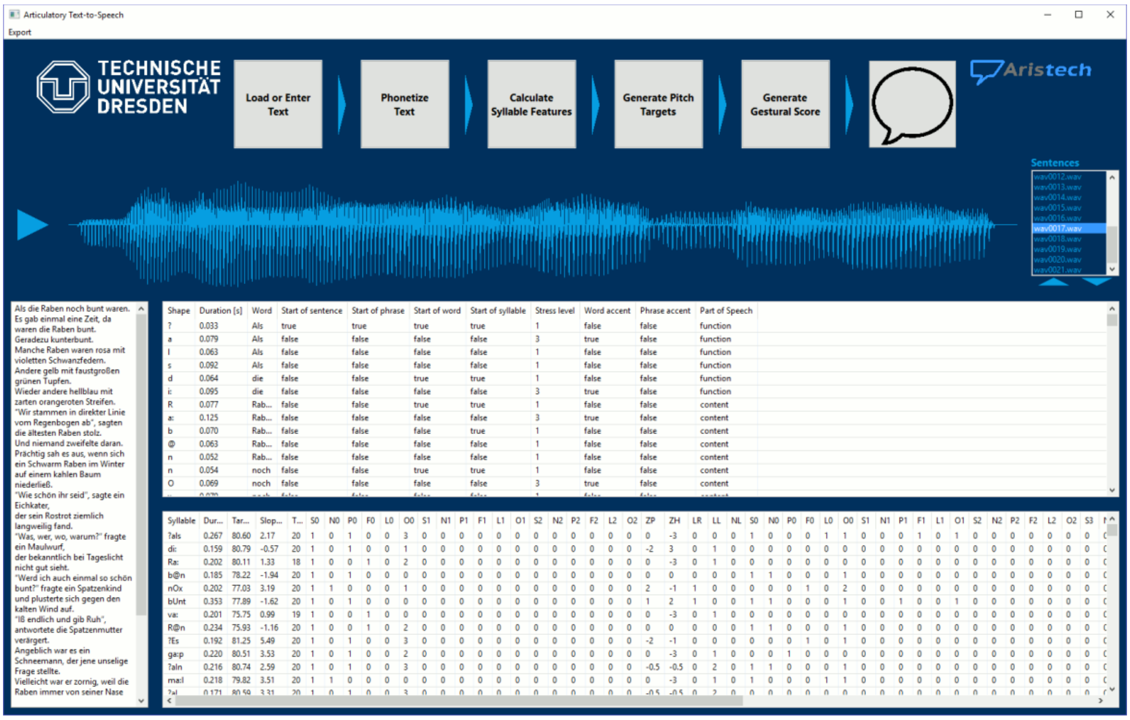
References

[1] Umeda, Noriko, E. Matsui, Torazo Suzuki, and Hiroshi Omura. "Synthesis of fairy tales using an analog vocal tract." In *Proceedings of the 6th International Congress on Acoustics*, pp. B159-162. 1968.

[2] Wang, Yuxuan, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Zongheng Yang Jaitly, Ying Xiao et al. "Tacotron: Towards End-to-End Speech Synthesis." In *Proc. of the Interspeech*, Stockholm, Sweden, pp. 4006-4010. 2017.

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

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



Als die Raben noch bunt waren.
Es gab einmal eine Zeit, da waren die Raben bunt.
Geradezu kunterbunt.
Manche Raben waren rosa mit violetten Schwanzfedern.
Andere gelb mit faustgroßen grünen Tupfen.
Wieder andere hellblau mit zarten orangeroten Streifen.
"Wir stammen in direkter Linie vom Regenbogen ab", sagten die ältesten Raben stolz.
Und niemand zweifelte daran.
Prächtig sah es aus, wenn sich ein Schwarm Raben im Winter auf einem kahlen Baum niederließ.
"Wie schön ihr seid", sagte ein Eichkater,
der sein Rostrot ziemlich langweilig fand.
"Was, wer, wo, warum?" fragte ein Maulwurf,
der bekanntlich bei Tageslicht nicht gut sieht.
"Werd ich auch einmal so schön bunt?" fragte ein Spatzenkind und plusterte sich gegen den kalten Wind auf.
"Iß endlich und gib Ruh",
antwortete die Spatzenmutter verärgert.
Angeblich war es ein Schneemann, der jene unselige Frage stellte.
Vielleicht war er zornig, weil die Raben immer von seiner Nase

Shape	Duration [s]	Word	Start of sentence	Start of phrase	Start of word	Start of syllable	Stress level	Word accent	Phrase accent	Part of Speech
?	0.033	Als	true	true	true	true	1	false	false	function
a	0.079	Als	false	false	false	false	3	true	false	function
I	0.063	Als	false	false	false	false	1	false	false	function
s	0.092	Als	false	false	false	false	1	false	false	function
d	0.064	die	false	false	true	true	1	false	false	function
i:	0.095	die	false	false	false	false	3	true	false	function
R	0.077	Rab...	false	false	true	true	1	false	false	content
a:	0.125	Rab...	false	false	false	false	3	true	false	content
b	0.070	Rab...	false	false	false	true	1	false	false	content
@	0.063	Rab...	false	false	false	false	1	false	false	content
n	0.052	Rab...	false	false	false	false	1	false	false	content
n	0.054	noch	false	false	true	true	1	false	false	content
O	0.069	noch	false	false	false	false	3	true	false	content

Syllable	Dur...	Tar...	Slop...	T...	S0	N0	P0	F0	L0	O0	S1	N1	P1	F1	L1	O1	S2	N2	P2	F2	L2	O2	ZP	ZH	LR	LL	NL	S0	N0	P0	F0	L0	O0	S1	N1	P1	F1	L1	O1	S2	N2	P2	F2	L2	O2	S3
?als	0.267	80.60	2.17	20	1	0	1	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	-3	0	0	0	1	0	0	0	1	1	0	0	0	1	0	1	0	0	0	0	0	0	0
di:	0.159	80.79	-0.57	20	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	-2	3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ra:	0.202	80.11	1.33	18	1	0	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	-3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
b@n	0.185	78.22	-1.94	20	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
nOx	0.202	77.03	3.19	20	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	2	-1	1	0	0	0	0	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
bUnt	0.353	77.89	-1.62	20	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	1	0	0	1	1	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0
va:	0.201	75.75	0.99	19	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R@n	0.234	75.93	-1.16	20	1	0	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
?Es	0.192	81.25	5.49	20	1	0	1	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	-2	-1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ga:p	0.220	80.51	3.53	20	1	0	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	-3	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
?aIn	0.216	80.74	2.59	20	1	0	1	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	-0.5	-0.5	0	2	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ma:l	0.218	79.82	3.51	20	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-3	0	1	0	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 2: Interface for the Articulatory Text-to-Speech Synthesis.