PROSPECTS OF ARTICULATORY TEXT-TO-SPEECH SYNTHESIS

Simon Stone¹, Atilla Azgin², Sabrina Mänz², Peter Birkholz¹ ¹Technische Universität Dresden, Germany ²Aristech 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.

e con export synolese mi	oner inne stannese	Help											
ignals Vocal tract Time doma	in simulation Gestura	il score											
Gesture		0.00 0.10	0.20	0.30 0	40 0.50	0.60	0.70 0.1	80 0.90	1.00	1.10	1.20	1.30	
Ineutral/absent Value 0.00 16000.00 dPa	Words Phones	-											_
0.00 +	Main oscillo.		THE OWNER OF T	0 1	INC.			Min .					-
Store		1.000	AUL D. L.	1 d	iki	durch		i a	ul.m.				
4 N	Bill a		history by the	4-1	and Month			A March	and the second		لإزارار ويقق	ACCOLLEGE A	He.
Start time: 1.322 s				r 1		Bana der.		1 M M	il lotter.			21111111	
Duration: 105 10 ms		110000	*****		1000		11					dilline.	
Time constant: 113 ms	Main spectre.				1/XE	Contraction of the	W.	100	64214123		06.0311	al tat bits	
< > >		- tintik	NUMBER	6 1	m. a	MARINE'S	P 118	6. M	809009	1 1	1964.11	Loca and La	
and the second second		10000000	11111111111111	10 1	abbaba - Ull	WF BOA	1 100	in	SIM PRO 2		MERNI	aaann	μ.
ane and a la		1111111	1999191	i) i	100000000 ·····························	10/01 3/1		Mill - Chill	0000.000	- 1	111/1908	THEFT	Ľ
Mark pos.: 0.013 s		111111111	NAME OF ALL OF	10 J	1000	101112 1. 11	a	444.0	a set to be	. 11	MU CH	********	1
Display choice		2033030	and the second		TERLE A	1857.381	6 100	HILL SHE	ALC: NO.	1			
Show gestural score		human	all carros	No. 19		ensered.	NI. INT	Station -1	CANDEL		100000	*****	Ŵ
O Show motor program		1 manual and			1000	COLUMN TO A PARTY	1000						
Audio cineak		<			_	_	_	_		_	_	_	
	vower gestures	1		· · · · ·									
Di Show stratorem													
Show segmentation													
Show model F0 curve	Lin nextures	-							6.14	1			-
Show animation	Co protect												
									-				-
synthesize										-			
					B-alveolar-fric		2	t-atveciar-fric					
Acoustics (time-domain)	Tongue tip g.												
Acoustics (time-domain) Show vocal tract	Tongue tip g.				-			-	_				
Acoustics (time-domain) Show vocal tract	Tongue tip g.												-
Acoustics (time-domain) Show vocal tract Show glottis	Tongue tip g.												_
Acoustics (time-domain) Show vocal tract Show glottis	Tongue tip g. Tongue body g.								_				_
Acoustics (time-domain) Show vocal tract Show glottis Calculate F0	Tongue tip g. Tongue body g.				-						_		_
Acoustics (time-domain) Show yocal tract Show glottis Calculate F0 Calculate guilty	Tongue tip g. Tongue body g.				-						-		_
Acoustics (thme-domain) Show vocal tract Show glettis Calculate Fi Calculate Fi Calculate Fi Analysis results	Tongue tip g. Tongue body g. Velic pestures												
Acoustics (time-domain) Show vocal toot Show glettis Calculate RI Calculate RI Calculate RI Analysis results Analysis restings	Tongue tip g. Tongué body g. Velic pestures						<u> </u>						
Acoustics (time-domain) Show vocal tract Show youths Calculate F0 Calc. voice quality Analysis results Analysis results Analysis settings Anotation dislog	Tongue top g. Tongue body g. Velic pestures	modal	kite	Inda	leiceisas fir	kite	Inedal	kojceliusa fre	Unication	s-strelmo	dal		
Acoustics (time-domain) Show yocal tract Show yocal tract Calculate F0 Calc. voice quality Analysis results Analysis results Analysis results	Tongue top g. Tongue body g. Velic gestures Gliottal shape g.	model	stop	medal	loiceless fric	\$100	modał	voiceless-Pic	voiceles	is-stop mo	dal		
Accounts: (time-domain) Serve vocal tract Show glattis Calculate R0 Calculate R0 Calculate R0 Calculate R0 Calculate R0 Analysis results Analysis results Analysis results	Tongue top g. Tongue body g. Velic pestures Gliottal shape g.	modal	stop	medal	voiceless fric	stop	Indat	kokelass-fic	voice(e)	s-stop/mo	dal		
Acoustics (time-domain) Show vocal tract Show glottis Calculate R0 Calculate R0 Calculate R0 Calculate R0 Calculate R0 Calculate R0 Calculate R0 Calculate R0 Analysis results Analysis results Analysis results	Tongue top g. Tongue body g. Velic gestures Gliottal shape g.	modal	stop	modal	voiceless-fric	5100	modal	voiceness Fic	roiceles	s-stop mo	dal	~	
Acoustics (time-domain) Servivocal text Show glottis Calculate R0 Calculate R0 Calculate R0 Analysis results Analysis retrings Annotation dialog	Tongue top g Tongue body g Velic gestures Glottal shape g FØ gestures	modal 79 st 79	stop 5 st	medal	triceless-fric	8100	modal	voice##ss-fric	viceles	s-stop mo	dəl		
Acoustics (three domain) Show vocal theat Draw glattis Calculate R0 Calculate R0 Calculate R0 Analysis results Analysis results Analysis estings Annotation dialog	Tongue top g. Tongue body g. Velic pestures Glottal shape g. FØ gestures	modal 79 st 71	stop 5 st	medal	loiceless-fric	Stop 77 s	modal	voicelless #rc	voceles	s-stop mo	dəl		
Accountics (three-domain) Show vpcal thack Show glattilis Calculate FD Calculate FD	Tongue top g. Tongue body g. Velic gestures Glottal shape g. FO gestures	modal 70 st 71	jstep 5 st	medal p	loiceless filt	8100 77 s	modal	voice#ss=#ic	- POCENIES	s-stop mo	dəl		
Accounts (three-doward) Shear-yocal third Shear-yocal third Shear-yocal third Calculate for Analysis results Analysis results Analysis results Analysis results	Tongue top g. Tongue body g. Velic gestures Dicital shape g. F0 gestures	modal 79 st 77	jstop 5 st	medal 7	loicetess filt:	8100 77 1	inodal a	koiceress tric	voiceles	s-stop mo	dəl		
Acoustic (three-doward) One vocal funct Seew glottis Calcuette 10 Calcuette 10 Ca	Tongue top g. Tongue body g. Velic pestures Glottal shape g. F0 gestures Lung pressure g.	79 st 71	bite 5 st	medal	Voiceles Tric	×100 77 1	medal I	Voicèliess-Ric	Pa	s-stop mo 75 st	dəl		
Acouting time downer) Siew synch hart Siew grints Calculate H0 Calculate H0 Calculate H0 Calculate H0 Calculate H0 Analysis results Analysis results Analysis results	Tongue to p. Tongue body g. Velic gestures Glottal shape g. FO gestures Lung pressure g.	modal 79 st 71 8000 dPa	bstep 5 st	modal	Voicetees fric	\$100 77 1 Pa	medal it	Voiceiess fit	Pà	5-stop mo	dəl	0 97	

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

References

- 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 Endto-End Speech Synthesis." In *Proc. of the Interspeech*, Stockholm, Sweden, pp. 4006-4010. 2017.
- [3] Birkholz, Peter. "Modeling consonantvowel 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.

	Articulatory Text-to-Speech																																				-		×
And of Raten noch burd states. State of parkers State of pylladie State of pylladi		ISC RSIT En	<u>H</u> E Ät	Le	oad or Te	r Ent xt	er			Phoi Te	netize ext)	Syll	Calcu able I	ilate Featur	es)	G	enerat Targ	e Pito ets	ch)	G	Ge Sestu	nerat Iral So	e :ore			\langle	~)	5	יכ	Ari [5 ente	ecl	ر ^
To grade control on solved And or brokes A	Als die Raben nach burt weren a					Stat			C.) Const												44444 11111														r	-	wav00 wav00 wav00 wav00 wav00 wav00 wav00 wav00	14.wav 14.wav 15.wav 16.wav 18.wav 19.wav 20.wav 21.wav	
a 0.07 Ab fale	Es gab einmal eine Zeit, da	Shape ?	Duration 0.033	n [s] N	Nord	Start	of sent	ence	Start o	t phrase	true	of wor	d Sta	t of syl	llable	Stress I	evel	Word a false	occent	false	e accer	nt Pa fu	art of S inction	peech															î
Mind Raber water role and role of the role	Geradezu kunterbunt.	a	0.079	1	Als	false			false		false		fals	e		3		true		false		fu	inction																
noise s 0.00 Main Faile	Manche Raben waren rosa mit	1	0.063	4	Als	false			false		false		fals	e		1		false		false		fu	inction																
guine guine faile faile <th< td=""><td>violetten Schwanzfedern. Andere gelb mit faustgroßen</td><td>s</td><td>0.092</td><td>1</td><td>Als</td><td>false</td><td></td><td></td><td>false</td><td></td><td>false</td><td></td><td>fals</td><td>e</td><td></td><td>1</td><td></td><td>false</td><td></td><td>false</td><td></td><td>fu</td><td>inction</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></th<>	violetten Schwanzfedern. Andere gelb mit faustgroßen	s	0.092	1	Als	false			false		false		fals	e		1		false		false		fu	inction																
Wieder anderse hallbal vm Austrantical Surgetiers is Opposite Fable F	grünen Tupfen.	d	0.064		die	false			false		true		tru			1		false		false		fu	inction																
Alter Control R O.07 Fab. fabe	Wieder andere hellblau mit	ii:	0.095	6	die	false			false		false		fals	e		3		true		false		fu	inction																
con-specifies faile faile faile faile faile faile content und promoty a 0.13 Bab faile faile <td>zarten orangeroten Streifen.</td> <td>R</td> <td>0.077</td> <td>F</td> <td>Rab</td> <td>false</td> <td></td> <td></td> <td>false</td> <td></td> <td>true</td> <td></td> <td>tru</td> <td></td> <td></td> <td>1</td> <td></td> <td>false</td> <td></td> <td>false</td> <td></td> <td>co</td> <td>ontent</td> <td></td>	zarten orangeroten Streifen.	R	0.077	F	Rab	false			false		true		tru			1		false		false		co	ontent																
Jeie Betracht Zuben Daben Bub. Rabe. Rabe. Rabe. Rabe. Rabe. Rabe. False False False Content Und ninnand Address Date Jaku, wenn sich 0.055 R.Bab false false false false false false content Practing and e Jaku, wenn sich 0.055 R.Bab false false false false content entent and method false false false false content entent and method false false false false content entent and method false false false false content entent and false false false false content entent and false false false content entent and false false false false content entent and false false	vom Regenbogen ab", sagten	a:	0.125	F	Rab	false			false		false		fals	e		3		true		false		ce	ontent																
Understand source/file dams. 00 0.051 Fabr. faile fabr. fa	die ältesten Raben stolz.	b	0.070	F	Rab	false			false		false		true			1		false		false		ce	ontent																
na Galaxam Rober mi Witem Wite	Und niemand zweifelte daran. Prächtig sah es aus wenn sich	۰	0.063	F	Rab	false			false		false		fals	e		1		false		false		co	ontent																
Inf encodering No No <td>ein Schwarm Raben im Winter</td> <td>n</td> <td>0.052</td> <td>F</td> <td>Rab</td> <td>false</td> <td></td> <td></td> <td>false</td> <td></td> <td>false</td> <td></td> <td>fals</td> <td>e</td> <td></td> <td>1</td> <td></td> <td>false</td> <td></td> <td>false</td> <td></td> <td>ce</td> <td>ontent</td> <td></td>	ein Schwarm Raben im Winter	n	0.052	F	Rab	false			false		false		fals	e		1		false		false		ce	ontent																
Operation Operation Minist M	auf einem kahlen Baum	n	0.054		hoch	false			false		true		tru					taise		false		ce	ontent																
Lickhar Strate Data Tab Solution Tab Solution Solution <td>"Wie schön ihr seid" sante ein</td> <td></td> <td>0.009</td> <td></td> <td>10ch</td> <td>false</td> <td></td> <td></td> <td>faise faise</td> <td></td> <td>false</td> <td></td> <td>fals</td> <td></td> <td></td> <td>•</td> <td>_</td> <td>true</td> <td></td> <td>false</td> <td></td> <td></td> <td>antent</td> <td></td> <td>~</td>	"Wie schön ihr seid" sante ein		0.009		10ch	false			faise faise		false		fals			•	_	true		false			antent																~
def rain finances Serves Tan. Serves	Eichkater,					, and the second se																																	
marging marging <t< td=""><td>der sein Rostrot ziemlich</td><td>Syllable</td><td>Dur</td><td>Tar</td><td>Slop.</td><td> T</td><td>. SO</td><td>N0</td><td>PO FO</td><td>LO</td><td>O0 S1</td><td>N1</td><td>P1 F</td><td>1 L1</td><td>01</td><td>52 N</td><td>P2</td><td>F2</td><td>L2 (</td><td>D2 ZP</td><td>ZH</td><td>LR</td><td>u</td><td>NL S</td><td>0 N</td><td>P0</td><td>FO</td><td>L0 0</td><td>0 S1</td><td>N1</td><td>P1</td><td>F1 L1</td><td>01</td><td>S2</td><td>N2</td><td>P2 F</td><td>2 L2</td><td>02</td><td>^ 1 62</td></t<>	der sein Rostrot ziemlich	Syllable	Dur	Tar	Slop.	T	. SO	N0	PO FO	LO	O0 S1	N1	P1 F	1 L1	01	52 N	P2	F2	L2 (D2 ZP	ZH	LR	u	NL S	0 N	P0	FO	L0 0	0 S1	N1	P1	F1 L1	01	S2	N2	P2 F	2 L2	02	^ 1 62
and Madward: dc 0.19 0.70 -0.57 0.1 0 1 0<	"Was, wer, wo, warum?" fragte	?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	o (
Were drawtich bet lageslicit. Rar. 0.222 8.11 1.33 1.0 0 1 0 <th0< td="" th<=""><td>ein Maulwurf,</td><td>di:</td><td>0.159</td><td>80.79</td><td>-0.57</td><td>20</td><td>1</td><td>0</td><td>1 0</td><td>0</td><td>1 0</td><td>0</td><td>0 0</td><td>0</td><td>0</td><td>0 0</td><td>0</td><td>0</td><td>0 0</td><td>) -2</td><td>3</td><td>0</td><td>1</td><td>0 0</td><td>0</td><td>0</td><td>0</td><td>0 0</td><td>0</td><td>0</td><td>0</td><td>0 0</td><td>0</td><td>0</td><td>0</td><td>0 0</td><td>0</td><td>0</td><td>0 (</td></th0<>	ein Maulwurf,	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 (
Were fash acche small to child b@m 0.138 78.22 -1.54 1 0 <td>der bekanntlich bei Tageslicht nicht out rieht</td> <td>Ra:</td> <td>0.202</td> <td>80.11</td> <td>1.33</td> <td>18</td> <td>1</td> <td>0</td> <td>0 1</td> <td>0</td> <td>2 0</td> <td>0</td> <td>0 0</td> <td>0</td> <td>0</td> <td>0 0</td> <td>0</td> <td>0</td> <td>0 0</td> <td>0</td> <td>-3</td> <td>0</td> <td>1</td> <td>0 0</td> <td>0</td> <td>0</td> <td>0</td> <td>0 0</td> <td>0</td> <td>0</td> <td>0</td> <td>0 0</td> <td>0</td> <td>0</td> <td>0</td> <td>0 0</td> <td>0</td> <td>0</td> <td>0 (</td>	der bekanntlich bei Tageslicht nicht out rieht	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 (
bunt?"fragte ein Spatzenkind 🛛 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	"Werd ich auch einmal so schön	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 (
	bunt?" fragte ein Spatzenkind	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 C
und plusterte sich gesen den 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	und plusterte sich gegen den kalten Wind auf.	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 (
18 endlich und gib Ruh*, we 0.201 75.75 0.99 19 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0	"IB endlich und gib Ruh",	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 (
antwortet die Spatzenmutter R@m 0.234 7.593 -1.16 20 1 0 0 1 0 2 0 0 0 0 0 0 0 0 0 0 0 0	antwortete die Spatzenmutter	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 0
Verargenci. Tes ULS2 3AP 20 1 0 1 0 0 5 0 0 0 0 0 0 0 0 0 0 0 0 0	Angeblich war es ein	rbs	0.192	61.25	5.49	20	1	0	1 0	0	3 0	0	0 0	0	0	0 0	0	0		-2	-1	0	0	0 0	0	0	1	1	0	0	0	0 0	0	0	0	0 0	0	0	
Schneemann, der jene unselige 2010 0256 0021 202 201 0 1 0 0 0 0 0 0 0 0 0 0 0	Schneemann, der jene unselige	ga:p Zalo	0.220	90.74	2.50	20	1	0	1 0	0	2 0	0	0 0	0	0	0 0	0	0		05	-3	0	2	0 1	1	0	0	0 1	0	0	0	0 0	0	0	0	0 0	0	0	
Frage statile. The second seco	Frage stellte. Vielleicht war er zornig weil die	mail	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	-0.5	0	1	0 1	0	0	0	1 1	0	0	0	0 0	0	0	0	0 0	0	0	
	Raben immer von seiner Nase	?a ≮	0 171	80.59	3 31	20	1	0	1 0	n	3 0	0	0 0	0	n	0 0	ő	n	0 0	-0.5	-0.5	n	2	0 0	n	0	ñ	n n	n	0	0	0 0	ň	0	0	0 0	0	n	n jr Y

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