PERCEPTUAL CUES FOR SMILED VOICE - AN ARTICULATORY SYNTHESIS STUDY

Simon Stone, Pia Abdul-Hak, Peter Birkholz

Technische Universität Dresden
simon.stone@tu-dresden.de

Abstract: The present study investigated the perceptual cues in continuous speech for a smiling speaker. Most previous studies analyzed natural speech samples and acted smiles and thus were not able to separate the effects of different parameters. Here, we followed an analysis-by-synthesis paradigm and used articulatory synthesis to create a set of synthetic utterances with controlled variation of the most likely cues as identified by a comprehensive literature review (the mean fundamental frequency and the vocal tract length). These stimuli were then rated in a listening experiment to examine the influence of each parameter separately and jointly. This study is the first such effort using complete sentences, while previous work in this line of research has only focused on isolated vowels. The responses of the experiment's 57 participants confirmed that the analyzed cues can indeed individually convey the impression of a smiling speaker, most strongly so if both are present at the same time.

1 Introduction

Human speech expresses much more than just semantics. One of the many dimensions of speech is the expression of emotions [1]. As somewhat of a special case of emotional expression, the notion of a "smiled voice" refers to the characteristic sound of speech that is produced by a smiling speaker and auditorily detectable by listeners (as shown by, e.g., [2, 3]). A smile is not always an indication of positive affect, however: Some authors make a distinction between felt (Duchenne) smiles and unfelt (non-Duchenne) smiles [4, 3]. A felt smile, in this framework, is a smile that involves the activation of the muscles around the eye (causing "crow's feet") while an unfelt smile does not. Other classification schemes differentiate more than 50 diferent kinds of smiles based on facial measurements [5, p. 127]. At least some of these variations can not only be detected but also discerned by the listener [4]. Furthermore, a "smile" is not even a clearly defined category, but exists on a spectrum that extends all the way to laughter [6, 7]. Due to this vast diversity among all these expressions that are collectively called "smiles", the literature offers a similarily diverse set of auditory cues for them. Table 1 summarizes the findings from ten selected studies. The inclusion criteria for the studies in this overview were: Published in 2010 or later OR among the most frequently cited papers in the field (according to Google Scholar) OR involving German language (because German was also the target language in this study).

2 Methods

The articulatory synthesizer `VocalTractLab` allows the parametric synthesis of utterances based on models of the vocal folds or glottis, the 3D vocal tract shape geometry, their dynamic control, and the aero-acoustics involved in the interplay of all these components. The

Study	f_0	F_1	F_2	F_3	LP	LxH	Stimuli
Tartter (1980) [8]	+	+	+	+	−		natural, nonsense syllables, acted smile
Tartter & Braun (1994) [2]	∘		+		−		natural, nonsense syllables, acted smile
Lasarcyk & Trouvain (2008) [9]	+	+	+		−	+	synthetic, isolated vowels
Drahota *et al.* (2008) [4]	∘				−		natural sentences, provoked smile
Erickson *et al.* (2009) [10]	+	−	+	−	−		natural words and phrases, acted smile
Fagel (2010) [11]	+	∘	+	∘	−	+	natural words, acted smile
Torre (2014) [12]		+	∘	+			natural, conversational speech, authentic smile
Keough *et al.* (2015) [13]	+	∘	∘		−	∘	natural, vowels, acted smile
El Haddad *et al.* (2015) [14]	+	∘	+		+		natural, sentences, acted smile
Ponsot *et al.* (2018) [15]		+	+	∘			manipulated natural, isolated vowel

Table 1 – Overview of the findings from selected studies on cues for smiled speech: Fundamental frequency f_0, formant frequencies F_1, F_2, F_3, lip protrusion LP and larynx height LxH. A '+' denotes an observed increase of the respective parameter in smiled speech compared to neutral speech, a '-' denotes a decrease, '∘' means *no change*, and an empty cell denotes *not analyzed*.

results are comparable in naturalness and intelligibility with other, even non-parametric synthesis systems [16]. Therefore, it is ideally suited to conduct analysis-by-synthesis studies to evaluate the effects of various phonetic and articulatory parameters on the perceived speech (see, e.g., [17, 18]). In this paradigm, instead of *analyzing* speech samples trying to identify the parameters that are used as cues for a particular impression, speech is *synthesized* using different parameter combinations and repeatedly evaluated to observe the effect of the varied parameters on the perception. Here, we analyzed the effect of the fundamental frequency f_0 and the vocal tract length (as determined by the degree of lip protrusion and the larynx height) on the perception of smiled speech by human listeners. As described in section 1, these parameters were repeatedly identified to have an effect in this context, albeit to different degrees and even in different directions. The formant frequencies F_1, F_2, and F_3 were not directly manipulated at the signal level but indirectly at the articulatory level by changing the vocal tract shape.

2.1 Stimuli generation

The stimuli generation consisted of two steps: (1) Creating variants of the static vocal tract shapes for the vowels with raised larynx and less protruded lips, and (2) selecting and synthesizing variants of a set of suitable sentences using the original vocal tract shapes, the manipulated shapes, and two different levels of the fundamental frequency f_0 (*neutral* and *raised*).

`VocalTractLab` offers a set of pre-defined vocal tract shapes for all canonic German speech sounds originally derived from magnetic resonance imaging of a human speaker [19]. These shapes are intended to convey a neutral expression of each sound. According to the literature reviewed in section 1, the most frequently observed changes to the vocal tract shape when smiling were an increased larynx height and a decreased lip protrusion, resulting in an overall shorter vocal tract. Therefore, changes were made to the neutral baseline vowel shapes to reflect these observations. In the `VocalTractLab`'s geometric vocal tract model [19], the corresponding parameters are the lip protrusion *LP* and the vertical hyoid position *HY*. The minimum value for *LP* is defined by the vocal tract model as -1 and the maximum value for *HY* as -3.5. A naive manipulation strategy would therefore simply set these parameters to those values in all vocal tract shapes. However, since changing the vocal tract geometry also changes the formants, this also affects the qualities of the sounds. This was also observed by Lasarcyk & Trouvain [9], but they intentionally did not try to compensate this formant change. However, while the literature also reports different formant frequencies for smiled and neutral speech, the changes are usually not as large as they would be based on the vocal tract length manipulations (in the

order of 10 % instead of up to 100 %). Furthermore, a retracted tongue dorsum was observed in smiled speech [10]. This may indicate that compensatory movements are made by the articulators (not necessarily just the tongue) to reduce the effect of the shortened vocal tract (which would be predicted by the motor equivalence theory [20]). The manipulation strategy in this study was therefore slightly more elaborate than in [9]:

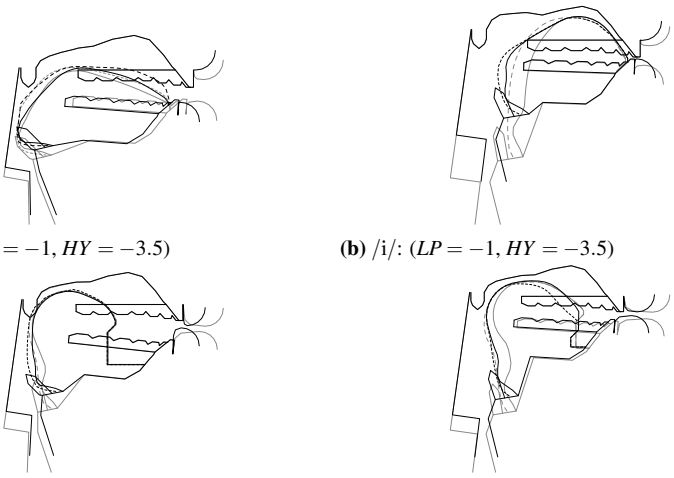
1. Set *LP* and *HY* to their extreme values (called the *LH* configuration).
2. Automatically optimize all *other* vocal tract parameters (except the velum parameters) until the first three formant frequencies are at least 10 % and at most 20 % higher than the expressively neutral baseline.
3. If this cannot be achieved, set *LP* and *HY* to the half-way point from their neutral to their extreme position (called the *lh* configuration) and optimize again.
4. Save the final vocal tract shape as the *smiled* variant of the baseline shape.

The automatic optimization was done using a greedy algorithm implemented in `VocalTractLab` and described in [19]. As was to be expected, the unrounded vowels could be realized with the more extreme manipulations, while the rounded vowels (plus /ɔ/ and /ʊ/) required less extreme values for *LP* and *HY* to stay within the acceptable formant range (see Table 2). One notable exception to this systematic procedure was the /u/: Even when using the *lh* values, the formants were too different from the neutral baseline and the resulting sound was more /o/-like. In this one special case, *LP* and *HY* were moved back to their neutral baseline from their *lh* values until the formants were within the tolerance again. Some example shapes are shown in Figure 1.

Vowel	/ə/	/a/	/e/	/i/	/o/	/u/	/ɛ/	/ø/	/y/	/ɪ/	/ɔ/	/ʊ/	/œ/	/ʏ/	/ɐ/
Variant	*LH*	*LH*	*LH*	*LH*	*lh*	*lh**	*LH*	*lh*	*lh*	*LH*	*lh*	*lh*	*LH*	*lh*	*LH*

Table 2 – Degree of manipulations for the various vowels. *LH* denotes extreme values for the lip protrusion and larynx height, *lh* denotes values at the halfway-point between the baseline and the extreme values (* required even more protrusion to remain within the acceptable formant range).

The manipulated shapes were then used to synthesize a set of German sentences in different variations. The text material for these sentences was chosen from the *Berlin sentences* [21], which were designed to contain an average of five words per sentence and include all of the German phonemes and a large number of possible biphonematic combinations. Of these 100 sentences, a subset of 15 sentences was selected. The selection criteria were (a) no linguistic content potentially suggestive of a smile (e.g. “The sun is laughing.”) and (b) a phoneme distribution representative of the German language (based on internal statistics obtained on the Spoken Wikipedia Corpus [22], see Figure 2). Mind that the distribution was calculated based on the vocal tract shapes that were actually used to synthesize the sentences. Since `VocalTractLab` realizes the canonic phonemes in a way that minimizes the vocal tract shape inventory while at the same time maximizing the intelligibility of the sounds, this distribution is slightly different than the distribution of the phonemes based on the canonic transcription. Most notably, primary diphthongs are realized as a sequence of two monophthong shapes (hence no diphthong phonemes appear in Figure 2) and the vocalic /ʁ/ allophones are realized using two different vocal tract shapes /ɐ$_{\text{low}}$/ and /ɐ$_{\text{mid}}$/, depending on the context (see [23] for details). The total number of sentences was heuristically derived from the targeted duration of the listening experiment of approximately 10 minutes per participant. The list of sentences is shown in Table 3.



(a) /a/: ($LP = -1$, $HY = -3.5$) **(b)** /i/: ($LP = -1$, $HY = -3.5$)

(c) /o/: ($LP = -0.16$, $HY = -4.34$) **(d)** /u/: ($LP = 0.35$, $HY = -5.2$)

Figure 1 – Example original and manipulated vocal tract shapes. The gray dashed contours are the original shapes, the black contours are the *smiled* versions. The long-dashed lines mark the tongue's side elevation.

The selected sentences were then synthesized using `VocalTractLab`. Starting with version 2.3, `VocalTractLab` supports the automatic generation of a gestural score (the means to control the synthesis parameter trajectories over time) based on a sequence of phone labels and their acoustic durations. The first step was therefore to generate these sequences for each sentence. To that end, reference recordings of each sentence were made in a quiet office environment using a high-quality headset with a paired USB sound interface (Corsair Virtuoso RGB Wireless) and recording software Audacity. The recorded speaker was a male 36-year-old German native speaker originally from the region of Hanover with no notable accent who was instructed to read the sentences with a neutral expression. The phone labels and their corresponding durations were then annotated in the audio files using the software `Praat` [24], exported in TextGrid format, and converted into a segment-file (*.seg), which can be imported into `VocalTractLab`. In addition, the trajectory of the fundamental frequency f_0 was extracted from each recording using the `Praat` function *To Pitch...* with a pitch floor of 50 Hz and a pitch ceiling of 300 Hz and exported as a PitchTier-file. These trajectories were then parametrized using the software `TargetOpimizer 2.0` [25] to obtain sequences of pitch targets suitable for the Target Approximation Model used in `VocalTractLab`. Using the segment files and the corresponding pitch targets, the gestural scores for the non-smiled, regular f_0 baseline stimuli were automatically generated as described in section 7.6 of the `VocalTractLab` manual[1] and then manually optimized by tuning the closure durations of the stops to match the natural reference recordings. Next, all vowel vocal tract shapes in the baseline gestural scores were substituted with their *smiled* counterpart. These scores were used to generate the smiled, regular f_0 stimuli. Finally, two more sets of scores were generated by raising all pitch targets by 2 st in each score, resulting in the scores for the non-smiled, raised f_0 and smiled, raised f_0 stimuli. All scores were then synthesized using `VocalTractLab` 2.3 with default synthesis options. The final set therefore consisted of four conditions (permutations of smiled/not-smiled and regular/raised f_0) of 15 stimuli each for a total of 60 stimuli. The speaker file containing the baseline and manipulated vocal tract shapes, all recordings and their annotation files in TextGrid format, the corresponding segment files, the automatically generated baseline gestural

[1] `https://www.vocaltractlab.de/download-vocaltractlab/VTL2.3-manual.pdf`

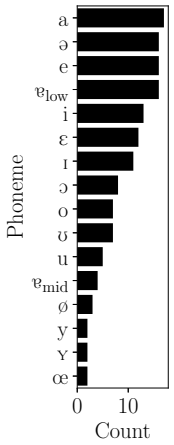


Figure 2 – Phoneme distribution in the selected subset of the Berlin sentences.

Index	Orthographic	Phonetic
002	Am blauen Himmel ziehen die Wolken.	ʔam blaʊn hɪml t͡siːn diː vɔlkŋ
007	Jetzt sitzen sie beim Frühstück.	jɛt͡st zɪt͡sn ziː baɪm fʁyːʃtʏk
013	In der Mitte steht der Brötchenkorb.	ʔɪn də mɪtə ʃteː deːɐ̯ bʁøtçnkɔ̯ɐp
019	Überquere die Straße vorsichtig!	ybɐkfveːʁ diː ʃtxʁaːsə foːɐ̯zɪçtɪç
026	Öl fehlte wohl auch.	ʔøːl feːltə voːl ʔaʊ̯x
039	Wir wollen heute spazieren gehen.	viːɐ̯ vɔln hɔʏtə ʃpat͡siːʁngeːn.
040	Da möchte ich gerne mit.	daː mœçt ɪç gɛɐ̯nə mɪt
041	Zuvor müssen wir uns stärken.	t͡suːfoːɐ̯ mʏsn viːɐ̯ ʊns ʃtɛɐ̯kŋ.
042	Die Kartoffeln gehören zum Mittagessen.	diː kaʁtɔfəln gəhøʁn t͡suː mɪtaːgəsn
043	Zum Schnitzel gibt es Erbsen.	t͡sʊm ʃnɪt͡səl giːpd əs ʔɛɐ̯psn
061	Auf dem Brett leuchten bunte Tulpen.	ʔaʊ̯f deːm bʁɛt lɔʏçtn bʊntə tʊlpm
066	Der Bahnhof liegt sieben Minuten entfernt.	deːɐ̯ baːnhoːf liːkt ziːb minuːtn ɛntfɛɐ̯nt
075	Der junge Zugbegleiter pfeift zur Abfahrt.	dɛ jʊŋə t͡suːkbəglaɪtɐ pfaɪ̯ft t͡suːɐ̯ ʔapfaːɐ̯t
083	Es gehört zu einer Feldscheune.	ʔəs gəhøɐ̯t t͡suː ʔaɪ̯nɐ fɛlʧɔʏnə
087	In der Dämmerung kommen wir heim.	ʔɪn deːɐ̯ dɛmərʊŋ kɔm viːɐ̯ haɪ̯m

Table 3 – Selected sentences and their phonetic realizations in `VocalTractLab`.

score files, the manually manipulated gestural score files, and the synthesized speech files are available at `https://www.vocaltractlab.de/index.php?page=birkholz-supplements`.

2.2 Experimental design

A listening experiment was designed to evaluate, which of the described manipulations contributed to the perception of a smile by human listeners. Due to the ongoing COVID-19 pandemic at the time of the study, the experiment was conducted online using the `webMUSHRA` experiment software [26]. Although most of the experimental environment cannot be controlled in an online experiment, the participants were asked to use headphones and a quiet environment to minimize such influences. The experimental design was as follows: Each participant listened to all 60 stimuli in an individually randomized order. After listening to each stimulus, the participant was asked to rate the stimulus as *smiled* or *neutral* by clicking on a correspondingly labeled button with a stylized smiling or neutral face. In total, 57 participants (36 male, 20 female, 1 diverse; age range 12 to 60, mean 26, standard deviation 8 years) completed the experiment. None of the participants reported any hearing or speech impairment or any kind of empathy deficit disorder. All participants gave informed consent and volunteered their time without compensation.

3 Results

The experiment was evaluated in terms of the proportion of the participants that have rated a particular stimulus as *smiled*. Table 4 summarizes the results.

By ordering the proportions from lowest to highest, a relative ranking of the effect of the various conditions on the perceived *smiledness* could be determined. Since each participant rated all stimuli, McNemar's test for paired nominal data was used to estimate the significance of the pair-wise ranking. As Table 4 shows, both the raised f_0 and the manipulated vocal tract shapes individually significantly increased the perception of a smile compared to the baseline. Between the two kinds of manipulations, the vocal tract manipulations had a slightly larger effect on the smile perception than the raised f_0, but the difference was not significant. However, both manipulations combined achieved a once again significant increase in perceived smiled-

Sentence index	00	01	10	11	Ranking (* significant at $\alpha = 0.05$)
002	7.9 %	20.2 %	30.7 %	53.5 %	$00 \overset{*}{<} 01 < 10 \overset{*}{<} 11$
007	25.4 %	48.2 %	47.4 %	71.1 %	$00 \overset{*}{<} 10 < 01 \overset{*}{<} 11$
013	23.7 %	42.1 %	39.5 %	62.3 %	$00 < 10 < 01 < 11$
019	37.7 %	64.0 %	64.9 %	81.6 %	$00 \overset{*}{<} 01 < 10 < 11$
026	25.4 %	41.2 %	47.4 %	70.2 %	$00 < 01 < 10 \overset{*}{<} 11$
039	31.6 %	56.1 %	26.3 %	68.4 %	$00 < 10 \overset{*}{<} 01 < 11$
040	16.7 %	36.0 %	30.7 %	58.8 %	$10 < 00 < 01 \overset{*}{<} 11$
041	20.2 %	43.0 %	36.8 %	57.0 %	$00 \overset{*}{<} 10 < 01 \overset{*}{<} 11$
042	18.4 %	24.6 %	47.4 %	58.0 %	$00 < 01 \overset{*}{<} 10 < 11$
043	28.1 %	50.9 %	29.8 %	65.8 %	$00 < 10 \overset{*}{<} 01 < 11$
061	19.3 %	41.2 %	47.4 %	61.4 %	$00 \overset{*}{<} 01 < 10 \overset{*}{<} 11$
066	23.7 %	28.1 %	44.7 %	57.9 %	$00 < 01 < 10 < 11$
075	15.8 %	32.5 %	50.9 %	65.8 %	$00 \overset{*}{<} 01 \overset{*}{<} 10 < 11$
083	9.6 %	25.4 %	28.9 %	48.2 %	$00 < 01 < 10 \overset{*}{<} 11$
087	9.6 %	18.4 %	23.7 %	41.2 %	$00 < 01 < 10 \overset{*}{<} 11$
all	20.9 %	38.1 %	39.8 %	61.4 %	$00 \overset{*}{<} 01 < 10 \overset{*}{<} 11$

Table 4 – Results of the listening test. The conditions are coded as follows: baseline (00), raised f_0 (01), manipulated vocal tract shape (10), both raised f_0 and manipulated vocal tract shape (11). The reported percentages are the proportion of participants that have rated the respective stimulus as *smiled*. The significance of the relative ranking was calculated using McNemar's test for paired nominal data.

ness compared to their individual contributions.

4 Discussion

The results show that the sentences had a wide range of baseline smiledness ranging from 7.9 % in sentence 002 to 37.7 % in sentence 019. The highest rating for each stimulus significantly correlates with this baseline (Pearson correlation coefficient $\rho = 0.9$, $p < 0.001$). The sentences had no discernible biased content: The sentence "Be careful when you cross the road!" for example had the highest baseline rating and the sentence "They are having breakfast now." had the lowest. The baseline rating had no significant correlation with the (uncontrolled) mean f_0 ($\rho = 0.27,\ p > 0.3$) or its standard deviation ($\rho = -0.04,\ p > 0.8$) in the neutral sentences. There is also no correlation between the baseline rating and the number of rounded vowels ($\rho = 0.08,\ p > 0.7$), the number of unrounded vowels ($\rho = -0.26,\ p > 0.35$), or the number of spread vowels ($\rho = 0.12,\ p > 0.6$) in a sentence. It appears that there are as-of-yet unknown confounding factors in the perception of smiled speech that require further investigation.

5 Conclusion and outlook

We conducted an analysis-by-synthesis study on the effect of a shortened vocal tract and a raised f_0 on the perception of a smile in synthetic speech. The study was able to confirm the findings of Lasarcyk & Trouvain [9] for connected utterances: Spreading the lips, raising the larynx, and raising f_0 increases the perception of a smile. Both manipulations can contribute individually but have the strongest effect when combined. The results revealed that there are more confounding factors that may influence the listener's impression. Future work should therefore incorporate additional likely parameters like the phone durations [2], the intensity

[4], the voice quality or the spectral slope [10], and include more levels of the manipulated parameters.

References

[1] SCHERER, K. R., R. BANSE, and H. G. WALLBOTT: *Emotion inferences from vocal expression correlate across languages and cultures*. *Journal of Cross-cultural psychology*, 32(1), pp. 76–92, 2001.

[2] TARTTER, V. C. and D. BRAUN: *Hearing smiles and frowns in normal and whisper registers*. *The Journal of the Acoustical Society of America*, 96(4), pp. 2101–2107, 1994.

[3] SCHRÖDER, M., V. AUBERGÉ, and M.-A. CATHIARD: *Can we hear smile?* In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*. Sydney, Australia, 1998.

[4] DRAHOTA, A., A. COSTALL, and V. REDDY: *The vocal communication of different kinds of smile*. *Speech Communication*, 50(4), pp. 278–287, 2008. doi:https://doi.org/10.1016/j.specom.2007.10.001. URL `https://www.sciencedirect.com/science/article/pii/S0167639307001732`.

[5] EKMAN, P.: *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage (Revised Edition)*. WW Norton & Company, New York, NY, USA, 2009.

[6] TROUVAIN, J.: *Phonetic aspects of "speech-laughs"*. In *Oralité et Gestualité: Actes du colloque ORAGE*, pp. 634–639. L'Harmattan, Paris, France, Aix-en-Provence, France, 2001.

[7] SZAMEITAT, D. P., C. J. DARWIN, A. J. SZAMEITAT, D. WILDGRUBER, and K. ALTER: *Formant characteristics of human laughter*. *Journal of Voice*, 25(1), pp. 32–37, 2011. doi:https://doi.org/10.1016/j.jvoice.2009.06.010. URL `https://www.sciencedirect.com/science/article/pii/S0892199709001088`.

[8] TARTTER, V. C.: *Happy talk: Perceptual and acoustic effects of smiling on speech*. *Perception & psychophysics*, 27(1), pp. 24–27, 1980.

[9] LASARCYK, E. and J. TROUVAIN: *Spread lips + raised larynx + higher f0 = smiled speech? - An articulatory synthesis approach*. In *Proc. of the 8th International Seminar on Speech Production (ISSP)*, pp. 43–48. Strasbourg, France, 2008.

[10] ERICKSON, D., C. MENEZES, and K.-I. SAKAKIBARA: *Are you laughing, smiling or crying?* In *Proc. of the Annual Summit and Conference Asia-Pacific Signal and Information Processing Association (APSIPA ASC)*, pp. 529–537. Asia-Pacific Signal and Information Processing Association, Sapporo, Japan, 2009.

[11] FAGEL, S.: *Effects of Smiling on Articulation: Lips, Larynx and Acoustics*, pp. 294–303. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. doi:10.1007/978-3-642-12397-9_25. URL `https://doi.org/10.1007/978-3-642-12397-9_25`.

[12] TORRE, I.: *Production and perception of smiling voice*. In *Proceedings of the First Postgraduate and Academic Researchers in Linguistics at York (PARLAY)*, pp. 100–117. York, UK, 2014.

[13] KEOUGH, M., A. OZBURN, E. K. MCCLAY, M. D. SCHWAN, M. SCHELLENBERG, S. AKINBO, and B. GICK: *Acoustic and articulatory qualities of smiled speech*. *Canadian Acoustics*, 43(3), 2015.

[14] EL HADDAD, K., S. DUPONT, N. D'ALESSANDRO, and T. DUTOIT: *An HMM-based speech-smile synthesis system: An approach for amusement synthesis*. In *Proc. of the 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 5, pp. 1–6. IEEE, Ljubljana, Slovenia, 2015.

[15] PONSOT, E., P. ARIAS, and J.-J. AUCOUTURIER: *Uncovering mental representations of smiled speech using reverse correlation*. *The Journal of the Acoustical Society of America*, 143(1), pp. EL19–EL24, 2018.

[16] KRUG, P. K., S. STONE, and P. BIRKHOLZ: *Intelligibility and naturalness of articulatory synthesis with vocaltractlab compared to established speech synthesis technologies*. In *Proc. of the 11th ISCA Speech Synthesis Workshop (SSW 11)*, pp. 102–107. 2021. doi:10.21437/SSW.2021-18.

[17] BIRKHOLZ, P., L. MARTIN, K. WILLMES, B. J. KRÖGER, and C. NEUSCHAEFER-RUBE: *The contribution of phonation type to the perception of vocal emotions in German: An articulatory synthesis study*. *The Journal of the Acoustical Society of America*, 137(3), pp. 1503–1512, 2015. doi:10.1121/1.4906836.

[18] XUE, Y., M. MARXEN, M. AKAGI, and P. BIRKHOLZ: *Acoustic and articulatory analysis and synthesis of shouted vowels*. *Computer Speech & Language*, 66, p. 101156, 2021. doi:https://doi.org/10.1016/j.csl.2020.101156.

[19] BIRKHOLZ, P.: *Modeling consonant-vowel coarticulation for articulatory speech synthesis*. *PLOS One*, 8(4), pp. 1–17, 2013. doi:10.1371/journal.pone.0060603.

[20] PERRIER, P. and S. FUCHS: *Motor equivalence in speech production*. *The Handbook of Speech Production*, pp. 225–247, 2015. doi:10.1002/9781118584156.ch11.

[21] PÄTZOLD, M. and A. P. SIMPSON: *Acoustic analysis of German vowels in the Kiel Corpus of Read Speech*. *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung Universität Kiel*, 32, pp. 215–247, 1997.

[22] KÖHN, A., F. STEGEN, and T. BAUMANN: *Mining the spoken Wikipedia for speech data and beyond*. In *Proc. of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pp. 4644–4647. 2016.

[23] STONE, S. and P. BIRKHOLZ: *Articulatory synthesis of vocalized /r/ allophones in German*. *IEEE Transactions on Audio, Speech and Language Processing*, In review.

[24] BOERSMA, P. and D. WEENING: *Praat: Doing phonetics by computer*. 2021. URL `http://www.praat.org/`. [Computer program] Version 6.1.48.

[25] KRUG, P. K., S. STONE, A. WILBRANDT, and P. BIRKHOLZ: *TargetOptimizer 2.0: Enhanced estimation of articulatory targets*. In S. HILLMANN, B. WEISS, T. MICHAEL, and S. MÖLLER (eds.), *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung*, pp. 145–152. TUDPress, Dresden, Berlin, Germany, 2021.

[26] SCHOEFFLER, M., S. BARTOSCHEK, F.-R. STÖTER, M. ROESS, S. WESTPHAL, B. EDLER, and J. HERRE: *webMUSHRA — A comprehensive framework for web-based listening tests*. *Journal of Open Research Software*, 6(1), 2018.