JASA ARTICLE

Modelling microprosodic effects can lead to an audible improvement in articulatory synthesis

Paul Konstantin Krug,[1,a)] Branislav Gerazov,[2] Daniel R. van Niekerk,[3] Anqi Xu,[3,b)] Yi Xu,[3] and Peter Birkholz[1]
[1]*Institute of Acoustics and Speech Communication, Technische Universität Dresden, Germany*
[2]*Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University in Skopje, Republic of North Macedonia*
[3]*Department of Speech, Hearing and Phonetic Sciences, University College London, United Kingdom*

ABSTRACT:
When pitch is explicitly modelled for parametric speech synthesis, microprosodic variations of the fundamental frequency f_0 are usually disregarded by current intonation models. While there are numerous studies dealing with the nature and the origin of microprosody, little research has been done on its audibility and its effect on the naturalness of synthetic speech. In this work, the influence of obstruent-related microprosodic variations on the perceived naturalness of articulatory speech synthesis was studied. A small corpus of 20 German words and sentences was re-synthesized using the state-of-the-art articulatory synthesizer VocalTractLab. The pitch contours of the real utterances were extracted and fitted with the Target-Approximation-Model. After the real microprosodic variations were removed from the obtained pitch contours, synthetic variations were applied based on a microprosody model. Subsequently, multiple stimuli with different microprosody amplitudes were synthesized and evaluated in a listening experiment. The results indicate that microprosodic variations are barely audible, but can lead to a greater perceived naturalness of the synthesized speech in certain cases. © *2021 Acoustical Society of America*.
https://doi.org/10.1121/10.0005876
(Received 19 February 2021; revised 13 July 2021; accepted 22 July 2021; published online 17 August 2021)
[Editor: Paavo Alku] Pages: 1209–1217

I. INTRODUCTION

Apart from intentional intonation (*macroprosody*), human speech also exhibits small, systematic perturbations of its fundamental frequency f_0. In the literature, these variations are often referred to as *microvariations*, *micromelody*, or *microprosody* due to their relatively small amplitude of about 1 to 2 semitones[1,2] superimposed on the macroprosodic pitch contour and their relatively short timescale of about 30 to 150 ms.[2] While systematic f_0 variations were first observed in the German language,[3] the effects were found to occur across all languages and speakers, regardless of gender, although the exact extent of the f_0 variations is strongly context and speaker dependent.[1]

Usually a distinction is made between vowel-related f_0 perturbations, whereby on average, higher vowels have a higher f_0 than lower vowels,[4,5] and consonant-related f_0 perturbations,[4–9] whereby the voicing property of the consonant determines the character of the f_0 variation.[4–11] In the literature these two types of microprosody are sometimes referred to as "*VF0*" and "*CF0*";[12] however, this kind of classification and naming could be confusing since authors also speak of *intrinsic* f_0 (IF0) and *co-intrinsic* f_0 (CF0).[13] The former describes the f_0 perturbation that happens during the articulation of a phone (therefore intrinsic to the phone), while the latter describes the f_0 variation that an obstruent induces into its succeeding phone. While vowel-height-related microprosody is an IF0 effect, the situation is more complicated in the case of obstruent-related variations. It is well known that a vowel preceded by a voiceless consonant usually starts with an increased f_0, whereas a vowel preceded by a voiced consonant usually starts with a decreased f_0.[4,14] Nevertheless, it was found that the latter observation is actually related to an intrinsic drop of f_0 that happens during the articulation of the voiced consonant.[11] Hence, consonant-related f_0 variations due to voiced obstruents can actually be considered an IF0 effect. Since this analysis is about obstruent-related microprosodic effects only, the term IF0 will therefore refer to the intrinsic drop of f_0 during voiced consonants, and the term CF0 will refer to the co-intrinsic raise of f_0 at the beginning of vowels preceded by a voiceless consonant.

The exact mechanisms responsible for the microprosody have not yet been clarified beyond doubt. While some studies indicate that microprosody originates from bio-mechanical effects related to vocal fold tension, other studies propose aerodynamic effects as a possible explanation.[15–20] Even though there are numerous studies that provide explanations to the origin of microprosody, the proposed explanations are often incompatible or even contradictory.[11,21] Some authors also provided arguments for microprosody to be a controlled mechanism that is used to enhance speech.[21] However, with regard to the current state of research, it seems likely that microprosodic effects are not actively controlled, but rather produced unintentionally as a side-effect of articulation.[1]

[a)]Electronic mail: paul_konstantin.krug@tu-dresden.de
[b)]ORCID: 0000-0002-4331-6676.

Regardless of the mechanism that causes microvariations of f_0 on the fundamental level, the question of whether microprosodic effects have an impact on the perceived naturalness of synthetic speech arises. While there is evidence that f_0 can help listeners to differentiate between voiced and unvoiced consonants[16,22,23] and therefore increases the intelligibility of speech, the only study that tested the influence of the f_0 microvariations on the naturalness of speech in a listening experiment, performed by Rao *et al.*,[24] failed to observe a significant effect. Nevertheless, in their analysis, the authors only considered voiced vowel-consonant-vowel structures and neither real words nor continuous speech. Furthermore, the pitch-manipulated stimuli presented to the listeners were processed and synthesized with the help of the WORLD vocoder,[25] a technique that is not loss-free and may even introduce unnatural artifacts by itself.[26]

With the present study, we extend the current state of research by the following contributions:

(1) We re-synthesized 20 real German utterances (ten single words and ten short sentences) based on recordings of natural human speech, using the state-of-the-art articulatory synthesizer VocalTractLab[27,28] version 2.3 (VTL). The synthetic pitch contour was derived from the recordings and subsequently manipulated to include synthetic microprosodic effects according to a microprosody model based on previous work by Birkholz and Zhang.[2]
(2) We performed a perception experiment in order to evaluate whether German native speakers preferred the synthetic speech samples with microprosody over samples without the microprosodic effects or vice-versa.

II. METHODS

A. Re-synthesis of natural utterances

Re-synthesis means to create a synthetic version of a natural utterance that matches the original utterance as closely as possible. From VTL version 2.3 on, articulatory re-synthesis can be done semi-automatically using a phoneme-to-speech conversion. The utterance to be re-synthesized must first be loaded as a reference into the software. There, it is possible to manually segment and annotate the loaded audio material (so called *segment sequence*). In this study, the segmentation and annotation of the audio material was done by one phonetic expert and verified by a second phonetic expert. Subsequently, VTL can automatically transform the segment sequence into a set of articulatory gestures (*gestural score*[29]) This gestural score determines the time variation of the vocal tract shape, from which the synthetic utterances will be generated. Once the gestural score is computed, the utterance can be synthesized by the computational aerodynamic-acoustic simulation[30] of sound wave propagation within a realistic vocal tract model based on MRI data[27] using a geometric vocal fold model.[31]

A re-synthesized utterance should also have a pitch contour similar to the one of the original utterance. In order to match the synthetic pitch contour as well, the original pitch trajectory must be extracted from the recorded utterance. In this analysis, the software Praat[32] was used for the purpose of pitch extraction. Thereafter, the signal processing continued as follows:

Let $N(t_a), N \in [-1, 1]$ be the time-discrete digital audio signal of a recorded natural utterance and let $R(t_a)$ be the corresponding, re-synthesised audio signal of the same utterance. Thereby, $a = 1, \ldots, A$, where A is the number of audio samples. Let $\{(t_1, p_1) \ldots (t_B, p_B)\}, t_b \in T, p_b \in P$ be a set of time-pitch pairs, that is extracted from the signal $N(t_a)$. Here, $b = 1, \ldots, B$, where B is the number of time-pitch pairs and T and P denote the sets of time and pitch instances, respectively. Let $f_0^N : T \rightarrow P; t_b \mapsto p_b$ be the corresponding discrete pitch contour function. Since VTL pitch trajectories are modelled in terms of *pitch targets* based on the Target-Approximation-Model (TAM),[33,34] the time discrete pitch contour $f_0^N(t_b)$ must first be transformed into a TAM representation $f_0^{\mathrm{TAM}}(t)$ to be imported into VTL. In order to obtain the corresponding pitch target parameters, a time-continuous fit of the discrete pitch contour was performed using the software TargetOptimizer (TO)[35] version 1.0. The TO allows the determined pitch targets to be exported as a gestural score that can be imported into VTL. As the resulting pitch trajectory is an approximation of the natural pitch, it may still include remnants of the microprosodic effects (even though the TAM acts as a low-pass filter with a smoothing effect). Since the f_0 microvariations were to be modelled, visible remnants of microprosody were first removed from the natural contour by one phonetic expert. For this purpose, the pitch contour $f_0^{\mathrm{TAM}}(t)$ was manually modified by changing the TAM properties, e.g., the target time constants, slopes, and offsets of the corresponding pitch gesture within VTL. This way, the microprosody-corrected fit $f_0^{\mathrm{PLAIN}}(t)$ was obtained that is used as the pitch contour of the microprosody-corrected re-synthesis $R(t_a)$. In the following, such samples will be referred to as *plain* samples and corresponding data (segment sequences, gestural scores, audio files, etc.) will be referred to as *plain* data.

The top and middle plots in Fig. 1 show the audio signals of the original and the re-synthesized version of the word "Ferrari," respectively. The bottom plot shows the corresponding pitch trajectories. Thereby, the dots represent the discrete pitch data that were extracted from the original recording. The solid line represents the microprosody-corrected fit $f_0^{\mathrm{PLAIN}}(t)$, that is based on the TAM. The dashed line represents the pitch contour $f_0^{\mathrm{MP}}(t)$, which additionally contains the modelled microprosodic effects. The next section describes how such a trajectory is obtained.

B. Modelling microprosodic effects

In order to create pitch-manipulated samples, the microprosodic variations were added to the plain pitch contour automatically via a Python script (available in the supplementary materials[36]). For this purpose, we first exported the plain gestural score as a *tract sequence* file within VTL.

https://doi.org/10.1121/10.0005876

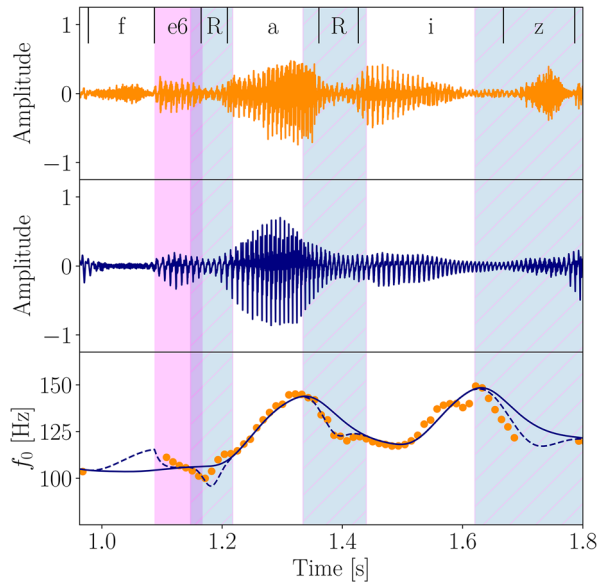

FIG. 1. (Color online) The word "Ferrari" (SAMPA: /fe6RaRi/) within the sentence "Sie fährt keinen Ferrari, sondern einen Maserati" (see utterance 19 in Table I for the IPA notation and translation). Top: The amplitude $N(t_a)$ of the digital audio signal of the natural speech recording. The aligned phoneme sequence is shown at the top of the plot. The corresponding CF0 and IF0 regions are drawn vertically across all subplots as solid (pink) and hatched (blue) bars, respectively. Middle: the audio amplitude $R(t_a)$ of the re-synthesized utterance. Bottom: The measured fundamental frequency data $f_0^N(t_b)$ of the original natural utterance is plotted as dots. The re-synthesized plain TAM trajectory $f_0^{\mathrm{PLAIN}}(t)$ and the manipulated TAM trajectory $f_0^{\mathrm{MP}}(t)$ with an amplitude factor $\mathcal{A} = 1.0$ are drawn as a solid line and a dashed line, respectively.

A tract sequence file contains the numerical values of all vocal tract and glottis parameters at each time step of the articulatory speech simulation performed by VTL. Thereby, a new state is evaluated every 110 samples of the resulting digital audio signal. Since VTL is using an audio sample rate f_S of 44.1 kHz, the internal state sample rate is therefore $f_I = 110^{-1} f_S \approx 400.91$ Hz. This is important because it means the pitch contour to be modified, is discretized at that rate. In order to apply the microprosodic manipulations at the correct instants in the signal, the corresponding time intervals, given by the obstruents that induce the pitch variation, must be transformed into tract-state intervals using the rate f_I as the conversion factor.

Regarding the obstruents, we adopted the convention of Birkholz and Zhang.[2] We defined IF0 time intervals $T_{\mathrm{IF0}} \ni \vec{t}^{\,i}_{\mathrm{IF0}} = [t^i_{\mathrm{I},S}, t^i_{\mathrm{I},E}]$, where $t^i_{\mathrm{I},S} = t^i_{\mathrm{I},1} - 0.4 \cdot |t^i_{\mathrm{I},2} - t^i_{\mathrm{I},1}|$ and $t^i_{\mathrm{I},E} = t^i_{\mathrm{I},2} + 0.2 \cdot |t^i_{\mathrm{I},2} - t^i_{\mathrm{I},1}|$ and $\{t^i_{\mathrm{I}}\}_{1,2}$ denote the start and end of the ith voiced obstruent of a given utterance, respectively. CF0 time intervals were defined as $T_{\mathrm{CF0}} \ni \vec{t}^{\,j}_{\mathrm{CF0}} = [t^j_{\mathrm{C},E}, t^j_{\mathrm{C},E} + 80\ \mathrm{ms}]$, where $t^j_{\mathrm{C},E}$ is the end time point of the jth voiceless obstruent. Thereby, T_{IF0} and T_{CF0} are the sets that contain all IF0 and CF0 intervals of an utterance, respectively. The corresponding index sets are denoted as $\mathfrak{I}$ and $\mathfrak{C}$, respectively, whereby $i \in \mathfrak{I}$ and $j \in \mathfrak{C}$. The manipulated pitch trajectory including the modeled microprosodic effects is then defined as

$$f_0^{\mathrm{MP}}(t) = f_0^{\mathrm{PLAIN}}(t) + \Delta f_0(t), \tag{1}$$

whereby the pitch manipulation term $\Delta f_0(t)$ is described by the following relation:

$$\Delta f_0(t) = \begin{cases} I\left(t, \vec{t}^{\,i}_{\mathrm{IF0}}\right), & \forall\, t \in \left\{\vec{t}^{\,i}_{\mathrm{IF0}}\right\}_{i \in \mathfrak{I}} \\ C\left(t, \vec{t}^{\,j}_{\mathrm{CF0}}\right), & \forall\, t \in \left\{\vec{t}^{\,j}_{\mathrm{CF0}}\right\}_{j \in \mathfrak{C}} \\ \Omega\left(t, \vec{t}^{\,j}_{\tilde{\mathrm{V}}}\right), & \forall\, t \in \left\{\vec{t}^{\,j}_{\tilde{\mathrm{V}}}\right\}_{j \in \mathfrak{C}} \\ 0, & \text{else}, \end{cases} \tag{2}$$

where $I(t, \vec{t}^{\,i}_{\mathrm{IF0}})$ and $C(t, \vec{t}^{\,j}_{\mathrm{CF0}})$ denote the IF0 and CF0 pitch contour functions, respectively. Note that, in case of overlapping intervals, the respective manipulations will be superimposed. The IF0 manipulations are given by

$$I\left(t, \vec{t}^{\,i}_{\mathrm{IF0}}\right) = -\mathcal{A} \cdot 10.75\ \mathrm{Hz} \cdot \exp\left(-\frac{(\tau - 0.49)^2}{0.26^2}\right), \tag{3}$$

where $\tau = (t - t^i_{\mathrm{I},S})/(t^i_{\mathrm{I},E} - t^i_{\mathrm{I},S})$. The CF0 pitch function is defined as an exponential function

$$C\left(t, \vec{t}^{\,j}_{\mathrm{CF0}}\right) = \mathcal{A} \cdot 11.1\ \mathrm{Hz} \cdot \exp\left(-\lambda\left(t - t^j_{\mathrm{C},E}\right)\right), \tag{4}$$

which decays to e^{-1} of its initial value after approximately 10 ms, since $\lambda = 95.9\ \mathrm{s}^{-1}$. In contrast to the work of Birkholz and Zhang,[2] here we additionally introduced the dimensionless amplitude factor $\mathcal{A}$ that allows for a scaling of the microprosodic effect. The reasons for this extension are described in Sec. II C. Evidently, Eqs. (3) and (4) are introducing discontinuities to the manipulated pitch contour. Regarding the IF0 effect, this is not a problem, since the boundary values are $I(t^i_{\mathrm{I},S}) = \mathcal{A} \cdot 0.31$ Hz and $I(t^i_{\mathrm{I},E}) = \mathcal{A} \cdot 0.23$ Hz. For $\mathcal{A} \in \{1.0, 1.5, 2.0\}$ (as used in this study), the f_0 discontinuity is therefore below 1 Hz, which is on the order of natural, random pitch fluctuations. However, the CF0 pitch manipulation causes a discontinuity between 11.1 and 22.2 Hz, depending on the amplitude $\mathcal{A}$. Even though the related obstruent is voiceless, such a jump in the f_0 trajectory might cause audible artifacts if the voice-onset time precedes the end of the consonant $t^j_{\mathrm{C},E}$; see Fig. 2. For this reason, a cosine interpolation was introduced to the manipulated pitch contour,

$$\Omega\left(t, \vec{t}^{\,j}_{\tilde{\mathrm{V}}}\right) = \frac{\mathcal{A}}{2} \cdot 11.1\ \mathrm{Hz} \cdot \left(1 - \cos\left(\pi \cdot \varphi\right)\right), \tag{5}$$

with $\varphi = (t - t^j_{\mathrm{C},S})/(t^j_{\mathrm{C},E} - t^j_{\mathrm{C},S})$. The cosine interpolation is defined on voiceless consonant intervals $T_{\tilde{\mathrm{V}}} \ni \vec{t}^{\,j}_{\tilde{\mathrm{V}}} = [t^j_{\mathrm{C},S}, t^j_{\mathrm{C},E}]$, where $t^j_{\mathrm{C},S}$ and $t^j_{\mathrm{C},E}$ are the start and end points of the jth voiceless obstruent and $T_{\tilde{\mathrm{V}}}$ is the set of relevant voiceless obstruents within a given utterance. Trivially, the index set of $T_{\tilde{\mathrm{V}}}$ is equal to $\mathfrak{C}$. After the plain pitch contour within the tract-sequence file is manipulated according to Eq. (1), the tract-sequence file can be directly turned into an audio file using VTL. Analogous

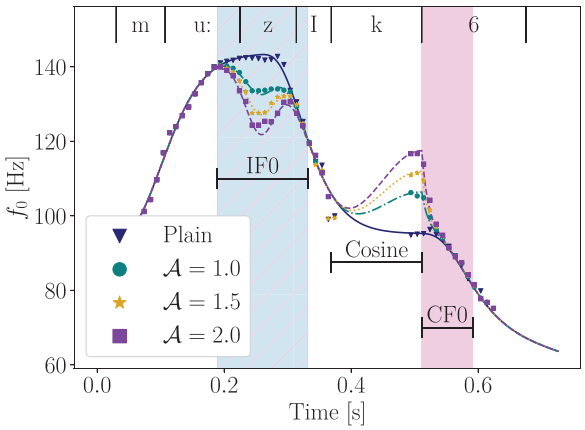

FIG. 2. (Color online) Different f_0 contours for the word "Musiker" (SAMPA: /mu:zIk6/, see utterance 8 in Table I for the IPA notation and translation). The plain TAM trajectory $f_0^{\mathrm{PLAIN}}(t)$ is drawn as a solid line. Manipulated TAM trajectories with amplitude factors $\mathcal{A} = 1.0, 1.5, 2.0$ are drawn as dashed-dotted, dotted, and dashed lines, respectively. The measured fundamental frequency data of the synthetic utterances are plotted as triangles (plain), dots ($\mathcal{A} = 1.0$), stars ($\mathcal{A} = 1.5$), and squares ($\mathcal{A} = 2.0$).

to the plain audio samples, we will refer to pitch-manipulated samples as *manipulated* samples.

C. Experimental design and stimuli preparation

The pitch manipulation is on the order of 10 Hz for an amplitude factor of $\mathcal{A} = 1.0$, which corresponds to approximately 1.5 to 2 semitones for a typical male voice with a fundamental frequency between 80 and 120 Hz. Such small pitch variations represent the minimal pitch discrimination limens for human speech.[37] Hence, an acoustic discrimination between the plain and manipulated versions was expected to be very difficult and the listening experiment was therefore designed as a pairwise comparison test.

Our choice of utterances was strongly related to the chosen design of the perception test. The fact that a participant would hear the synthetic utterances once with and once without microprosodic variations directly after another means two things: First, the chosen utterances should be rich in voiced and unvoiced obstruents so that many microvariations are present. At the same time, the stimuli should not be too long, so that the listener can remember his listening impression. As a consequence, ten simple words with three syllables were chosen. These were selected so that they include at least one IF0 and one CF0 obstruent. Additionally, ten short sentences with a length between five and seven words were constructed, so that they contain at least five obstruents. An overview of the chosen utterances is given in Table I.

In order to generate the stimuli, a 24 year old male German native speaker who uttered the selected words and sentences in a neutral manner of speaking, was recorded digitally with a sample rate of 44.1 kHz. Each recording was rendered as a 16-bit WAV file and loaded into VTL. Afterwards, the utterances were re-synthesized and manipulated as described in Secs. II A and II B, respectively. Thereby, a plain version as well as three manipulated versions with $\mathcal{A} = 1.0, 1.5, 2.0$ were generated for each of the utterances. This was done in order to test different strengths of microprosodic variations. It was expected that a stronger variation might increase the perceptibility of the effect and makes a more conscious decision possible, since amplitude factors of 1.5 and 2.0 correspond to a variation of around 3 to 4 semitones, which is the scale from which a larger number of listeners are able to perceive pitch variations.[37] Also, the amplitudes of the IF0 and CF0 effects, as determined by Birkholz and Zhang[2] were averaged over a broad range of phonemes and utterances. Thus, these amplitudes do not account for the large variance of the effect strength that is observed in natural speech. The mean Pearson correlation between the natural pitch contours and the TAM contours is 0.963 in case of the plain samples and 0.964, 0.960, and 0.953 in case of the manipulated versions with $\mathcal{A} = 1.0$, 1.5, and 2.0, respectively.

The stimuli were synthesized using the default speaker file of the VTL, *JD2*. All glottis parameters were kept at the default values, except for f_0 (which was modified as explained earlier) and the *flutter* parameter, which introduces pseudo-random f_0 fluctuations based on a model developed by Klatt and Klatt.[38] To avoid unpredictable interactions between the microprosodic and pseudo-random f_0 variations, the flutter parameter was always set to zero. All audio files were peak normalized to –1 dBFS. All used segment sequences and gestural scores can be found in the supplementary materials,[36] which also provide the necessary code to generate the tract sequence files and the audio stimuli.

D. Perception experiment

During the perception experiment, participants heard synthetic utterances once with and once without microprosodic variations in direct succession. The participants were allowed to play the audio of an utterance-pair as often as they liked. Subsequently, they had to decide which version they preferred. No stimuli pair was repeated in the course of the experiment in order to keep the experiment as short as possible. To ensure adequate statistical power, the listening experiment was conducted with 30 subjects (19 male, 11 female), which is in line with the recommendations by Wester *et al.*[39] The subjects were aged between 18 and 48 years with a median age of 30.5 years. As described earlier, the 20 chosen utterances were created with three different strengths of the microprosodic variation, so that each participant had to make a total of 60 decisions. In order to avoid a bias by intentionally corrupted results (e.g., participants choosing a single one of the answers all the time), each person was given a unique test in which the internal order of each pair ([A: plain, B: manipulated] or [A: manipulated, B: plain]) was determined randomly. The 60 utterances were split up in three blocks based on the microprosodic effect strength. The order of the blocks was randomized. The order of the 20 different pairs within each block was also randomized. The experiment took place in a

https://doi.org/10.1121/10.0005876

TABLE I. The used words and sentences in German, in the IPA notation and in English. The number of IF0 and CF0 deflections that happen during an utterance is given in the "Effect" column. The last column describes the measured relative frequency (see Sec. II E) for every utterance and separated into the three modes: ●, $\mathcal{A} = 1.0$, ★, $\mathcal{A} = 1.5$, ■, $\mathcal{A} = 2.0$. The combination of the three modes is denoted as ◆. Statistically significant results ($p < 0.05$) are denoted with *.

	Utterance (German)	Canonical transcription (IPA)	Translation (English)	Effect		Rel. frequency [$\cdot 10^{-2}$]			
				IF0	CF0	●	★	■	◆
1	Badetag	ˈbaːdəˌtaːk	*Bathing day*	2	1	57	53	47	52
2	Bewirken	bəˈvɪʁkən̩	*To effect*	2	1	60	57	60	59
3	Butterweich	ˈbʊtɐˈvaɪ̯ç	*As soft as butter*	2	1	67	50	60	59
4	Giraffe	ˌgiˈʁafə	*Giraffe*	2	1	53	70*	53	59
5	Kakadu	ˈkakadu	*Cockatoo*	1	2	60	53	60	58
6	Karotte	kaˈʁɔtə	*Carrot*	1	2	57	47	43	49
7	Kassette[a]	kaˈsɛtə	*Cassette*	1	2	63	73*	70*	69*
8	Musiker	ˈmuːzɪkɐ	*Musician*	1	1	40	47	50	46
9	Vergessen	fɛɐ̯ˈgɛsn̩	*To forget*	1	2	53	47	57	52
10	Zigarre	t͡siˈgaʁə	*Cigar*	2	1	47	63	37	49
11	Aber sehen will sie ihn doch.	ˈʔaːbɐ ˈzeːn vɪl ziː ʔiːn dɔx	*But she wants to see him.*	5	0	60	70*	57	62*
12	Er sah viele bunte Regenbogen.	eːɐ̯ zaː ˈfiːlə ˈbʊntə ˈʁeːgn̩ˌ boːgn̩	*He saw many colourful rainbows.*	6	2	33	53	53	47
13	Chabos wissen wer der Babo ist.	ˈt͡ʃaːboːs ˈvɪsn̩ veːɐ̯ deːɐ̯ ˈbaːboː ʔɪst	*The boys know who the boss is.*	6	2	57	50	57	54
14	Das Telefon ist seit sieben Tagen kaputt.	das ˈteːləfoːn ʔɪst zaɪ̯t ˈziːbn̩ ˈtaːgn̩ kaˈpʊt	*The phone has been broken for seven days.*	5	5	33	57	53	48
15	Die Artikel waren wieder vorrätig.	diː ˌʔaʁˈtiːk ˈl ˈvaːʁən ˈviːdɐ ˈfoːɐ̯ ˌʁɛtɪç	*The products were in stock again.*	6	4	43	33	57	44
16	Die Soße ist viermal übergekocht.	diː ˈzoːsə ʔɪst ˈfiːɐ̯ maːl ˈʔyːbɐgəˌkoxt	*The sauce boiled over four times.*	4	3	50	47	57	51
17	Die Straßenbahn fuhr weiter geradeaus.	diː ˈʃtʁaːsn̩ˌbaːn fuːɐ̯ ˈvaɪ̯tɐ gəʁaːdəˈʔaʊ̯s	*The tram continued straight ahead.*	7	3	53	40	47	47
18	Diese Zeitung ist bereits veraltet.	ˈdiːzə ˈt͡saɪ̯tʊŋ ʔɪst bəˈʁaɪ̯t͡s fɛɐ̯ˈʔaltət	*This newspaper is already outdated.*	4	4	53	40	53	49
19	Sie fährt keinen Ferrari, sondern einen Maserati.	ziː fɛːɐ̯t ˈkaɪ̯nən feˈʁaːʁiː ˈzɔndɐn ˈʔaɪ̯nən mazəˈʁaːtiː	*She does not drive a Ferrari, but a Maserati.*	7	4	63	50	57	57
20	Benno gefällt die orange Vase.	ˈbɛno gəˈfɛlt diː ʔoˈʁaŋʒə ˈvaːzə	*Benno likes the orange vase.*	6	2	50	63	47	53

[a]The official transcription indicates an unvoiced /s/ instead of /z/, but some people, especially our speaker, pronounce the "ss" voiced. Since we have re-synthesized the sentences in as much detail as possible, we used the voiced consonant in the synthesis.

sound-insulated audio studio. As the headphone, a STAX SR-202, driven by a STAX SRM-212 pre-amplifier, was used. The driver unit was connected to a laptop that was running the listening experiment, using an AUREON XFIRE 8.0 HD audio interface.

The experiment itself was carried out using PRAAT. The initial silence duration (before the start of the stimulus sound) was set to 0.3 s. The inter-stimulus duration (between samples A and B of a given pair) was set to 0.25 s. In this way, a pleasant listening experience was possible without irritating the listener by playing the samples too quickly in succession. The following text was displayed to the participants (translated to English): "*This is a perceptual experiment with synthetic speech. You will hear the same phrase twice in a row and you should decide whether you prefer the first or second realization. If you hear little or no difference between both samples, try to decide by feeling. The stimuli can be repeated as often as desired.*"

The measured data collected in the hearing experiment include both the information on the selected sample from each of the 60 stimuli-pairs and the reaction times measured by PRAAT. The reaction time is defined as the span between the start of the stimulus sound, i.e., after the initial silence (counting from the last repetition, if the statement is played repeatedly by the participant) and the time point, at which the test person selects one of the two answer options. Note that the accuracy of the reaction times, as measured by PRAAT, is limited by the clock accuracy of the operating system, which is typically on the order of 10 ms. This limitation is further discussed in Sec. IV B.

E. Statistical data analysis

In order to evaluate the choices of the listeners, we used the number of events in which a manipulated sample was preferred over the corresponding plain sample, divided by the total number of events as an observable. In the following, this quantity will be referred to as *relative frequency*. The null hypothesis H_0 underlying this observable is mathematically governed by a binomial distribution with the probability $P = 0.5$ and it describes two indistinguishable scenarios:

(1) Participants were choosing answers randomly (either because no difference between plain and manipulated

samples was perceived, or because the test was not taken seriously).
(2) Participants perceived acoustic differences between the plain and manipulated samples; however, on average, none of the versions were preferred over the others.

In order to evaluate the reaction times, we defined the following two observables: T_R^{PLAIN} and T_R^{MP}, which describe the reaction time when a plain or manipulated sample was chosen as the preferred sample, respectively. Furthermore, we defined normalized reaction times $\hat{T}_R^{\mathrm{PLAIN}}$ and $\hat{T}_R^{\mathrm{MP}}$ as

$$\hat{T}_R = \frac{T_R - \alpha}{\alpha}. \tag{6}$$

Thereby, $\alpha = 2D + 0.25\,\mathrm{s}$ is the time shift produced by the duration of the audio information presented to the listener. The time shift is composed of the medial silence (0.25 s) and two times the duration D of audio sample A or B. Their duration is equal since the pitch manipulation does not change the length of the utterance. Due to the subtraction of α and the division by α, the distribution of T_R becomes normalized, so that –1.0 (0.0) marks the start (end) of the played audio information presented to the listener.

III. RESULTS

If H_0 would actually describe the underlying distribution of the measured data, individual results among different utterances and participants might be combined to a single statistic and mathematically treated as a single Bernoulli experiment. Figure 3 shows the results summed over all individual participants, separated into words, sentences, and the combination of both. Further, results are shown for the individual amplitude factors $\mathcal{A} \in \{1.0, 1.5, 2.0\}$ as well as for the combined statistic of the three modes. Additionally, the 68% and 95% confidence intervals around H_0 are shown as a solid and a hatched area, respectively. Evidently, manipulated words with the mode $\mathcal{A} = 1.5$ are preferred 56.0% of the time, a two-sided binomial test yields a p-value of $p < 0.05$. Combined over all three modes, manipulated words are preferred in 55.1% of the cases ($p < 0.01$). In case of the combined sentences, no significant deviation from the null hypothesis could be observed. Combining the statistics of all utterances and microposody modes $\mathcal{A}$, manipulated samples are preferred over plain samples 53.2% of the time ($p < 0.01$).

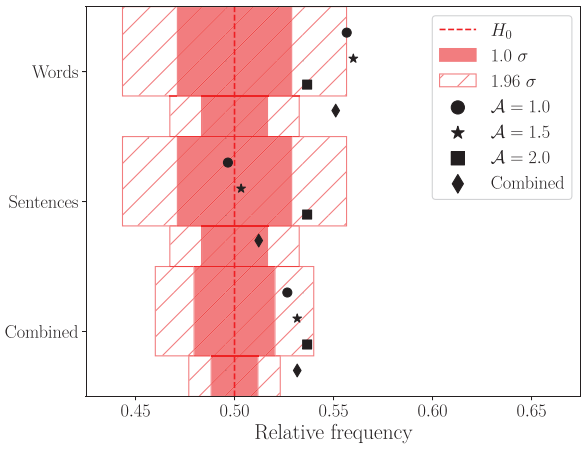

FIG. 3. (Color online) The relative frequency of events in which the manipulated sample was preferred are shown for all words, sentences, and the combination of both. Results are shown for the three different manipulation modes $\mathcal{A} = 1.0$ (circles), $\mathcal{A} = 1.5$ (stars), $\mathcal{A} = 2.0$ (squares), and the combined statistic (diamonds).

Figure 4 shows the distributions of obtained relative frequencies across all participants. The distributions are shown as estimated Gaussian kernel densities. Additionally, the respective boxplots are superimposed to the density plots. Two-sided Kolmogorov-Smirnov (KS) tests yield significant differences between the observed distributions and the null hypotheses in case of the single words ($p < 0.01$ in case of $\mathcal{A} = 1.0$, $\mathcal{A} = 1.5$ and the combination of all three modes, $p < 0.05$ in case of $\mathcal{A} = 2.0$). No significant differences were observed in case of the sentences. For the combination of words and sentences, the distributions for the mode $\mathcal{A} = 1.5$ and the combination of all modes were observed to be significantly different from the null hypotheses ($p < 0.05$). In the latter case, three participants preferred the manipulated samples beyond the 95% confidence interval (0.75 participants expected, given H_0), while none of the participants significantly preferred the plain samples (0.75 expected).

Figure 5(a) shows the distributions of T_R^{PLAIN} and T_R^{MP}, combined across all participants, as a box plot. In the case of the words, it was observed that the mean (median) reaction time decreased by 100 ms (37 ms) from T_R^{PLAIN} to T_R^{MP}. The absolute values of the means and medians are given in Table II. In case of the sentences, the mean (median) reaction time decreased by 137 ms (121 ms) from T_R^{PLAIN} to T_R^{MP}. While the reaction times for the sentences are significantly different from each other based on a two-sided KS test ($p < 0.05$), no significant difference was observed in the case of the single words ($p = 0.15$). The normalized reaction times,

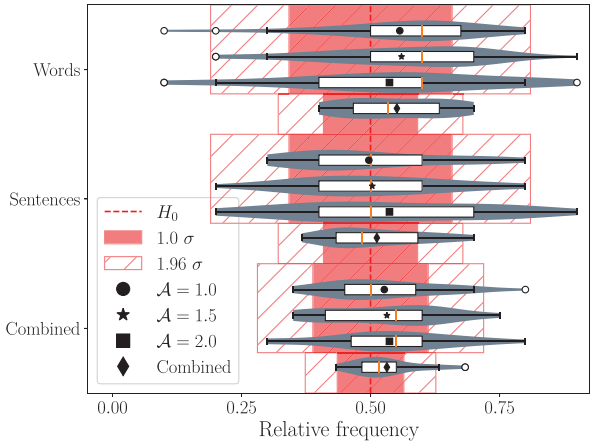

FIG. 4. (Color online) The relative frequencies of events in which the manipulated sample was preferred by a single participant are shown as distributions, separated into words, sentences, and the combination of both. Results are shown for the three different manipulation modes $\mathcal{A} = 1.0$ (mean, circles), $\mathcal{A} = 1.5$ (mean, stars), $\mathcal{A} = 2.0$ (mean, squares), and the combined statistic (mean, diamonds). Each median is indicated by a vertical line within the respective box.

https://doi.org/10.1121/10.0005876

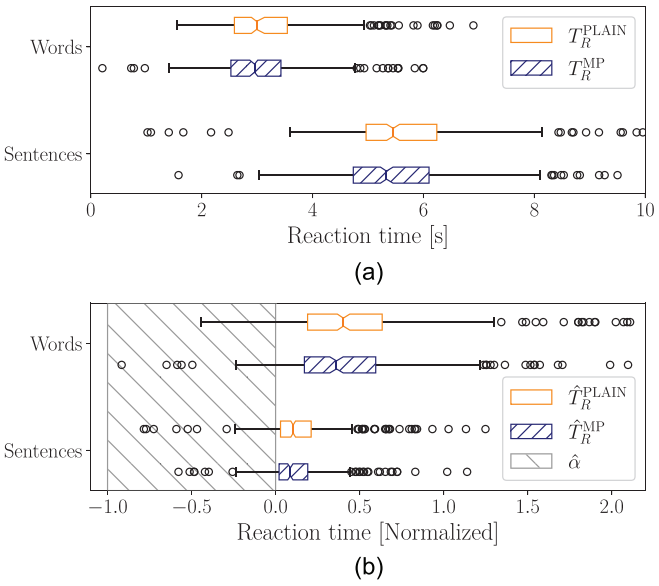

FIG. 5. (Color online) (a) The reaction times T_R^{PLAIN} and T_R^{MP} and the normalized reaction times $\hat{T}_R^{\mathrm{PLAIN}}$ and $\hat{T}_R^{\mathrm{MP}}$, separated into the individual distributions for words, sentences, and the combination of both, are shown as box plots in (a) and (b), respectively. The median of each distribution is indicated by the vertical line within each box. The 95% confidence interval of each median is indicated by the corresponding notch. The hatched area in (b) represents the (normalized) time span $\hat{\alpha}$ during which the stimulus is played.

separated into words and sentences, are shown in Fig. 5(b). Apparently, relative to the length of the utterance, listeners react much faster to sentences compared to single words. When it comes to plain (manipulated) sentences, listeners made their decision while the audio was still playing in 14.3% (16.1%) of the cases. For single words, this happens in only 2.9% (5.2%) of the cases. Further, this difference is evident not only on the relative, but also on the absolute timescale: Given a median duration of 0.90 s in case of the words and 2.29 s in case of the sentences, the median normalized reaction time measured from the end of the stimulus sound, is on the order of approximately 0.81 and 0.46 s (averaged over the plain and manipulated samples), respectively. Note that for the decisions made while the audio was playing, no utterance-specific or participant-specific pattern was found, as it happened across all utterances and among 22 out of 30 participants.

IV. DISCUSSION

A. Summary and conclusion

In the present study, we produced high quality synthetic speech using the articulatory synthesizer VTL, and manipulated the respective f_0 parameters to include microprosodic pitch deflections. In a listening experiment, 30 German native speakers had to decide whether they would prefer pitch-manipulated samples over plain samples or *vice versa*. From the combined results of that perception test, it was observed that manipulated words were preferred over their plain version in 55.1% of the cases, which is a significant deviation from the null hypothesis ($p < 0.01$). In case of the full sentences, the manipulated versions were preferred in 51.5% of the cases; however, this result is not significant. We interpret these results in the following way: Acoustic differences between plain and manipulated words were perceived by the listeners (either consciously or subconsciously) and, on average, test persons preferred the manipulated versions. Under the assumption that people prefer what they are used to and considering the fact that humans are used to hearing natural speech, it seems reasonable to conclude that the pitch manipulation according to the used microprosody model increased the naturalness of the re-synthesized words. However, in the case of the re-synthesized sentences, it was found that neither plain nor manipulated samples were consistently preferred. This means that either fewer differences were perceived compared to the situation with the single words, which would be conceivable if the f_0 variations are more obscured by the larger amount of acoustic information and the faster speaking style of the sentences compared to the single words, or, that a difference was perceived but the decision which version is preferred differed from subject to subject. We found evidence for the latter of the two explanations: If the decisions were indeed made more randomly, there is no reason why the distributions of the two reaction time observables T_R^{PLAIN} and T_R^{MP} should differ beyond random fluctuations. However, a significant difference was observed in case of the sentences. It seems likely that this difference is induced by participants that are more or less certain about their decision. We assume that an answer is selected faster if the listener is more certain that he prefers a certain version over the other. This effect is likely to be subconscious for the following reasons:

(1) Due to the randomized nature of the listening experiment, the participants had no control to intentionally modify the reaction times T_R^{PLAIN} and T_R^{MP} in different ways, without consciously discriminating between the plain and manipulated samples. If that condition was actually met, one would expect to observe a relative frequency of answers that deviates significantly from the null hypothesis.
(2) The results related to the normalized reaction times imply that the preferred sentences are rather selected intuitively since the decision is made while the stimulus is playing. In case of the single words, however, the decision is made significantly slower and happens after the end of the stimulus finished playing, which might imply that the listener tries to actively evaluate the listening impression.

TABLE II. The absolute mean (μ) and median (med) values of the reaction times and normalized reaction times.

	T_R^{PLAIN} [s]	T_R^{MP} [s]	$\hat{T}_R^{\mathrm{PLAIN}}$	$\hat{T}_R^{\mathrm{MP}}$
μ (Words)	3.175	3.075		
μ (Sentences)	5.618	5.481		
med (Words)	2.992	2.954	0.403	0.362
med (Sentences)	5.454	5.333	0.105	0.088

In summary, we conclude that modelled microprosodic effects in synthetic speech can be perceived by some listeners. In contrast to Rao *et al.*,[24] we found evidence that the modelled microprosodic effects can lead to a subtle increase in the naturalness of synthesized utterances.

B. Limitation and outlook

The following limitations have to be noted:

(1) The used microprosodic model is simplistic. It does not account for the variance that is observed among the microprosodic effect strengths. Birkholz and Zhang,[2] for example, reported largely different model parameters for the different obstruents. By averaging over a range of utterances and obstruents, the model represents an average that might rarely be represented by actual speakers and, in particular, the model might mismatch the microprosodic variations of the speaker that was recorded to obtain the natural pitch contours used for the re-syntheses. Additionally, the recorded speaker differs from the one on whose data the VTL vocal tract model is based. Although both persons were male, this difference together with the microprosody mismatch could lead to the contour sounding unnatural.
(2) Also, the quality of the speech annotations and the quality of the removal of the natural microprosodic effects may have an impact on the results. Even though both the annotations and the microprosody removal were performed and validated by phonetic experts, the removal of the microprosodic effects is subjective to some degree and might not be optimal.
(3) Last but not least, Kirby *et al.*[40] showed that the effect strength of obstruent-related microprosody differs among stressed and unstressed positions in a sentence, and therefore is context dependent. Future work should take this into consideration for a more realistic microprosody model.
(4) The listening experiment as carried out in this study is prone to random fluctuations. There is no guarantee that the sample of subjects is a good representation of the whole population of native speakers (and it is likely not to be the case). The number of participants is therefore critical, and a study of much larger scale would be an interesting opportunity for future research in this field.
(5) The fact that no significant trend has been observed among the different modes $\mathcal{A}$ is probably due to both, the insufficient sample size and the fact that a fixed strength of the microprosodic effect is not realistic. Future studies may also produce stimuli with amplitudes $\mathcal{A} > 2.0$ or $\mathcal{A} < 0.0$ in order to test if listeners prefer realistic ranges of the amplitude, or, whether it is just frequent f_0 changes that make utterances appear more preferable.
(6) The accuracy of the reaction times, as measured by Praat, is limited by the clock accuracy of the operating system, which is typically on the order of 10 ms. One can expect that such inaccuracies average out over a large amount of measured data. However, it is likely, that a more sophisticated measurement would be needed to further scrutinize the effects observed in the reaction time distributions. It is also clear that the interpretations of the reaction time measurements are quite speculative. Future studies should also consider recording the number of audio playbacks as a measure of the decisiveness of a participant. However, in order to actually pin down if decisions are made more or less consciously, a sophisticated measurement of neural activity in the brain might be necessary.

ACKNOWLEDGMENTS

This work has been funded by the Leverhulme Trust Research Project Grant No. RPG-2019-241: "High quality simulation of early vocal learning."

[1]D. H. Whalen and A. G. Levitt, "The universality of intrinsic f0 of vowels," J. Phon. **23**(3), 349–366 (1995).
[2]P. Birkholz and X. Zhang, "Accounting for microprosody in modeling intonation," in *Proceedings of ICASSP 2020*, Barcelona, Spain (May 4–8, 2020), pp. 8099–8103.
[3]E. Meyer, "Zur Tonbewegung des Vokals im gesprochenen und gesungenen Einzelwort" ("On the tonal movement of the vowel in spoken and sung words"), Phonetische Studien (Beiblatt zu der Zeitschrift: Die neueren Sprachen) **10**, 1–21 (1896–1897).
[4]A. S. House and G. Fairbanks, "The influence of consonant environment upon the secondary acoustical characteristics of vowels," J. Acoust. Soc. Am. **25**(1), 105–113 (1953).
[5]I. Lehiste and G. E. Peterson, "Some basic considerations in the analysis of intonation," J. Acoust. Soc. Am. **33**(4), 419–425 (1961).
[6]W. A. Lea, "Segmental and suprasegmental influences on fundamental frequency contours," in *Consonant Types and Tone*, edited by L. M. Hyman (University of Southern California, Los Angeles, CA, 1973), Vol. 1, pp. 15–70.
[7]J.-M. Hombert, "Consonant types, vowel quality, and tone," in *Tone: A Linguistic Survey* (Academic Press, New York, 1978), Vol. 77, p. 112.
[8]K. J. Kohler, "F0 in the production of lenis and fortis plosives," Phonetica **39**, 199–218 (1982).
[9]K. Silverman, "F0 perturbations as a function of voicing of prevocalic and postvocalic stops and fricatives, and of syllable stress," in *Reproduced Sound: 1985 Autumn Conference, Windermere: Conference Handbook* (Institute of Acoustics, Windermere, UK, 1984), Vol. 6, pp. 445–452.
[10]H. M. Hanson, "Effects of obstruent consonants on fundamental frequency at vowel onset in English," J. Acoust. Soc. Am. **125**(1), 425–441 (2009).
[11]J. P. Kirby and D. R. Ladd, "Effects of obstruent voicing on vowel F0: Evidence from 'true voicing' languages," J. Acoust. Soc. Am. **140**(4), 2400–2411 (2016).
[12]J. Kingston, "Segmental influences on F0: Automatic or controlled?," in *Tones and Tunes* (Mouton de Gruyter, Berlin, Germany, 2007), Vol. 2, pp. 171–210.
[13]A. Di Cristo and D. J. Hirst, "Modelling French micromelody: Analysis and synthesis," Phonetica **43**(1–3), 11–30 (1986).
[14]J.-M. Hombert, J. J. Ohala, and W. G. Ewan, "Phonetic explanations for the development of tones," Language **55**, 37–58 (1979).
[15]M. Halle and K. N. Stevens, "A note on laryngeal features," in *MIT Quarterly Progress Report* (MIT, Cambridge, MA, 1971), Vol. 101, pp. 198–212.
[16]K. J. Kohler, "F0 in the perception of lenis and fortis plosives," J. Acoust. Soc. Am. **78**(1), 21–32 (1985).
[17]A. Löfqvist, T. Baer, N. S. McGarr, and R. S. Story, "The cricothyroid muscle in voicing control," J. Acoust. Soc. Am. **85**(3), 1314–1321 (1989).
[18]A. Löfqvist, L. L. Koenig, and R. S. McGowan, "Vocal tract aerodynamics in /aca/ utterances: Measurements," Speech Commun. **16**(1), 49–66 (1995).
[19]C. X. Xu and Y. Xu, "Effects of consonant aspiration on Mandarin tones," J. Int. Phon. **33**(2), 165–181 (2003).

[20]A. L. Francis, V. Ciocca, V. K. M. Wong, and J. K. L. Chan, "Is fundamental frequency a cue to aspiration in initial stops?," J. Acoust. Soc. Am. **120**(5), 2884–2895 (2006).
[21]J. Kingston and R. L. Diehl, "Phonetic knowledge," Language **70**(3), 419–454 (1994).
[22]O. Fujimura, "Remarks on stop consonants: Synthesis experiments and acoustic cues," in *Form and Substance: Phonetic and Linguistic Papers Presented to Eli Fischer-Jørgensen* (Akademisk Forlag, Copenhagen, Denmark, 1971), pp. 221–232.
[23]D. W. Massaro and M. M. Cohen, "The contribution of fundamental frequency and voice onset time to the /zi/-/si/distinction," J. Acoust. Soc. Am. **60**(3), 704–717 (1976).
[24]A. Rao MV, S. Victory J, and P. K. Ghosh, "Effect of source filter interaction on isolated vowel-consonant-vowel perception," J. Acoust. Soc. Am. **144**(2), EL95–EL99 (2018).
[25]M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," IEICE Trans. Inf. Syst. **99**(7), 1877–1884 (2016).
[26]W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: 2000-speaker neural text-to-speech," in *Proceedings of the ICLR*, Vancouver, Canada (April 30–May 3, 2018), pp. 214–217.
[27]P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," PloS One **8**(4), e60603 (2013).
[28]P. Birkholz, "VocalTractLab (version 2.3) [computer program]," https://www.vocaltractlab.de/ (Last viewed January 28, 2021).
[29]P. Birkholz, B. J. Kröger, and C. Neuschaefer-Rube, "Model-based reproduction of articulatory trajectories for consonant–vowel sequences," IEEE Trans. Audio Speech Lang. Process. **19**(5), 1422–1433 (2011).
[30]P. Birkholz, "Enhanced area functions for noise source modeling in the vocal tract," in *Proceeding of the ISSP*, Cologne, Germany (May 5–8, 2014), pp. 32–40.
[31]P. Birkholz, S. Drechsel, and S. Stone, "Perceptual optimization of an enhanced geometric vocal fold model for articulatory speech synthesis," in *Proceeding of Interspeech*, Graz, Austria (September 15–19, 2019), pp. 3765–3769.
[32]P. Boersma and D. Weenick, "Praat: Doing phonetics by computer (version 6.0.43) [computer program]," http://www.praat.org (Last viewed January 28, 2021).
[33]Y. Xu and Q. E. Wang, "Pitch targets and their realization: Evidence from Mandarin Chinese," Speech Commun. **33**(4), 319–337 (2001).
[34]S. Prom-On, Y. Xu, and B. Thipakorn, "Modeling tone and intonation in Mandarin and English as a process of target approximation," J. Acoust. Soc. Am. **125**(1), 405–424 (2009).
[35]P. Birkholz, P. Schmager, and Y. Xu, "Estimation of pitch targets from speech signals by joint regularized optimization," in *Proceedings of EUSIPCO*, Rome, Italy (September 3–7, 2018), pp. 2075–2079.
[36]See supplementary materials at https://github.com/TUD-STKS/Microprosody for the segment sequences, gestural scores, and the relevant code to produce the stimuli files.
[37]J. 't Hart, "Differential sensitivity to pitch distance, particularly in speech," J. Acoust. Soc. Am. **69**(3), 811–821 (1981).
[38]D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. **87**(2), 820–857 (1990).
[39]M. Wester, C. Valentini-Botinhao, and G. E. Henter, "Are we using enough listeners? No!—An empirically-supported critique of Interspeech 2014 TTS evaluations," in *Proceedings of Interspeech*, Dresden, Germany (September 6–10, 2015), pp. 3476–3480.
[40]J. Kirby, F. Kleber, J. Siddins, and J. Harrington, "Effects of prosodic prominence on obstruent-intrinsic F0 and VOT in German," in *Proceedings of the 10th International Conference on Speech Prosody*, Virtual Conference (May 23–24, 2020), pp. 210–214.