Manipulation of the Prosodic Features of Vocal Tract Length, Nasality and Articulatory Precision Using Articulatory Synthesis

Peter Birkholz[a,*], Lucia Martin[b], Yi Xu[c], Stefan Scherbaum[d], Christiane Neuschaefer-Rube[b]

[a] *Institute of Acoustics and Speech Communication, Technische Universität Dresden, 01062 Dresden, Germany*
[b] *Department of Phoniatrics, Pedaudiology and Communication Disorders, University Hospital Aachen and RWTH Aachen University, Pauwelsstr. 30, 52074 Aachen, Germany*
[c] *Department of Speech, Hearing and Phonetic Sciences, University College London, Chandler House, 2 Wakefield Street, London*
[d] *Department of Psychology, Technische Universität Dresden, 01062 Dresden*

Abstract

Vocal emotions, as well as different speaking styles and speaker traits are characterized by a complex interplay of multiple prosodic features. Natural sounding speech synthesis with the ability to control such paralinguistic aspects requires the manipulation of the corresponding prosodic features. With traditional concatenative speech synthesis it is easy to manipulate the "primary" prosodic features pitch, duration, and intensity, but it is very hard to individually control "secondary" prosodic features like phonation type, vocal tract length, articulatory precision and nasality. These secondary features can be controlled more directly with parametric synthesis methods. In the present study we analyze the ability of articulatory speech synthesis to control secondary prosodic features by rule. To this end, nine German words were re-synthesized with the software VocalTractLab 2.1 and then manipulated in different ways at the articulatory level to vary vocal tract length, articulatory precision and degree of nasality. Listening tests showed that most of the intended prosodic manipulations could be reliably identified with recognition rates between 77-96 %. Only the manipulations to increase articulatory precision were hardly recognized. The results suggest that rule-based manipulations in articulatory synthesis are generally sufficient for the convincing synthesis of secondary prosodic features at the word level.

Keywords: prosody, feature manipulation, articulatory synthesis

1. Introduction

Speech prosody encodes linguistic and paralinguistic information (Ladd, 2008; Grichkovtsova et al., 2009). Paralinguistic information includes for example the emotional state of the speaker (Schröder, 2001), speaker traits (Schuller et al., 2015), the speaking style (Yamagishi et al., 2005), and speech and voice disorders. The simulation of these paralinguistic aspects is still a challenging problem in speech synthesis technology. Each paralinguistic function (e. g., the expression of a certain vocal emotion) is implemented by the complex interplay of multiple prosodic features. To synthesize speech that conveys certain paralinguistic information, it is thus necessary to be able to manipulate the involved prosodic features. In this study we analyzed the potential of articulatory speech synthesis to individually control specific prosodic features that are highly relevant for the encoding of paralinguistic information but have rarely been controlled in speech synthesis so far.

In the following, we differentiate the "primary" prosodic features of pitch, duration, and intensity on the one hand, and "secondary" prosodic features like voice quality (Campbell & Mokhtari, 2003; Pfitzinger,

*Corresponding author
Email address: peter.birkholz@tu-dresden.de (Peter Birkholz)

2006), nasality (Scherer, 1978), vocal tract length (Chuenwattanapranithi et al., 2008), and articulatory precision (Burkhardt & Sendlmeier, 2000; Beller et al., 2008) on the other hand. The primary features are easy to manipulate with the prevailing concatenative speech synthesis methods (Hunt & Black, 1996) as well as easy to quantify and analyze in recordings of natural speech. Therefore, the relation between these features and many paralinguistic aspects has been well studied, e. g., in the context of vocal emotions (Scherer et al., 2003; Schröder, 2001; Scherer et al., 2015).

In contrast, most secondary prosodic features are more difficult to manipulate in the acoustic domain of concatenative speech synthesis, because rather simple changes at the articulatory level may have complex consequences in the acoustic domain. For example, a nasal voice quality can be produced by a very specific and localized action at the articulatory level (lowering of the velum) but strongly affects the speech spectrum (introduction of pole-zero pairs in the vocal tract transfer function, shift of resonances). Analogously, a change of articulatory effort could be considered as a simple change of the speed of the articulators in the articulatory domain, but results in a complicated change of the formant trajectories in the acoustic domain. Hence, in order to manipulate secondary prosodic features like nasality or articulatory effort for speech synthesis, it is most favorable to do it in the articulatory domain of an articulatory speech synthesizer.

There is ample evidence that secondary prosodic features are just as important as the primary features for the implementation of diverse paralinguistic functions. For example, with regard to the expression of vocal emotions, the role of the feature phonation type was found to be a major cue in the expression of anger or fear (Gobl & Ní Chasaide, 2003; Burkhardt, 2009; Campbell & Mokhtari, 2003; Airas & Alku, 2006; Birkholz et al., 2015). Also vocal tract length is an important feature in the expression of emotions (Chuenwattanapranithi et al., 2008). According to the bio-informational dimensions theory, speakers modify their vocal tract length to project a larger body size to appear dominant and a smaller body size to appear friendly (Xu et al., 2013). Furthermore, articulatory precision is related to certain vocal emotions (Burkhardt & Sendlmeier, 2000; Murray & Arnott, 1993). For example, precise articulation contributes to a joyful impression and an imprecise one reduces it (Burkhardt & Sendlmeier, 2000). With regard to other paralinguistic functions, a nasal voice quality was, for example, identified as a vocal cue for the expression of body complacency (Sendlmeier & Heile, 1998), extroversion (Scherer, 1978) and sarcasm (Gibbs, 1986).

To vary these features with concatenative speech synthesis, the synthesizer needs a database of speech units that cover the necessary variation. However, since humans are not used to control prosodic features individually, they cannot help but let them co-vary with other features when asked to perform a manipulation. For example, to create a concatenative speech synthesizer for the synthesis of vocal emotions, the speech corpus needs to be recorded with multiple emotions that contain the emotion-specific feature combinations (Black, 2003; Iida et al., 2003). However, this is not only very laborious but the coverage of the feature space remains rather limited with respect to the variety of possible feature combinations.

In contrast to concatenative synthesis, parametric synthesis methods can manipulate the features of the voice source and the vocal tract independently. The main parametric synthesis methods are formant synthesis (Klatt, 1980), HMM-based synthesis (Zen et al., 2009) and articulatory synthesis (van den Doel et al., 2006; Birkholz, 2013a; Aryal & Gutierrez-Osuna, 2016). Formant synthesis has long been the first choice for the synthesis and analysis of (secondary) prosodic features for, e. g., emotions (Murray & Arnott, 1995; Burkhardt & Sendlmeier, 2000). More recently, HMM-based synthesis has been applied to modify secondary prosodic features for modeling different speaking styles and emotions (Yamagishi et al., 2005) or hypo- and hyperarticulated speech (Picart et al., 2014). However, both formant synthesis and HMM-based synthesis model speech in the temporal and spectral domain instead of the articulatory domain.

Articulatory speech synthesis can in principle vary all prosodic features directly at the articulatory and physiological level. Therefore, this kind of synthesis is generally considered as the best choice for research on paralinguistic effects like emotions (Schröder et al., 2010). However, despite considerable progress in the recent years, articulatory speech synthesis still sounds somewhat less natural than unit-selection synthesis, and the articulatory and acoustic models are rather time consuming. Hence, articulatory synthesis is not yet at a level of development where it is competitive for text-to-speech synthesis, but it is very well suited for analysis-by-synthesis experiments as in the present study. The effectiveness of articulatory synthesis in such an experiment for the analysis of phonation type in vocal emotions was recently demonstrated in Birkholz et al. (2015). However, the articulatory synthesis of further secondary prosodic features has so far not been

Manipulated feature	Kind of manipulation	Notation of stimuli
None	Re-synthesis of natural words	Standard stimuli
Vocal tract length	Larynx lowered by 1 cm, lips protruded by 1 cm, adjustments for natural vowel quality	Stimuli with longer vocal tract
	Larynx raised by 1 cm, lips retracted by 1 cm, adjustments for natural vowel quality	Stimuli with shorter vocal tract
Articulatory precision changed by effort	Half speed of articulatory target approximation	Stimuli with lower effort
	Double speed of articulatory target approximation	Stimuli with higher effort
Articulatory precision changed by centralization	Centralization of vocal tract targets by 25 %	Slightly centralized stimuli
	Centralization of vocal tract targets by 50 %	Very centralized stimuli
Creation of a permanent nasal leak	Velo-pharyngeal port opening of 10 % (0.2 cm^2) during all vowels and approximants	Nasalized stimuli

Table 1: Overview of the articulatory manipulations. Each of the 9 basis words was manipulated in each of the 8 ways in the middle column of the table (72 stimuli in total).

demonstrated in a systematic way. In this study we therefore examined different ways for the variation of the prosodic features of vocal tract length, nasality and articulatory precision using articulatory speech synthesis. It is shown that rule-based articulatory manipulations suffice for the perceptually convincing generation of these features.

2. Method

Nine German words (Banane [banaːnə], Birne [bɪrnə], Blaubeere [blaʊbeːʀə], Himbeere [hɪmbeːʀə], Mandarine [mandaʀiːnə], Melone [məloːnə], Mirabelle [mirabɛlə], Orange [oʀanʒə], Rosine [ʀoziːnə]; engl.: banana, pear, blueberry, raspberry, mandarin, melon, mirabelle, orange, raisin) were spoken in a neutral way by a male German native speaker and used as basis words for this study. These words were then re-synthesized as accurately as possible using the articulatory speech synthesizer VocalTractLab 2.1 (Birkholz, 2013b). In addition, each word was re-synthesized again in seven variants, where for each variant one of the features in the middle column of Table 1 was exclusively manipulated (leaving all other prosodic features unchanged). In a listening experiment, all 72 stimuli (9 words × 8 variants) were rated by 16 subjects with regard to naturalness and the perceptual identification of the intended manipulations. All stimuli are available for download from `http://www.vocaltractlab.de/index.php?page=birkholz-supplements`. The set of basis words was limited to nine items because the articulatory re-synthesis of natural words is currently still very laborious, and because we wanted to restrict the time for the perception experiment to a duration of about 30 min.

2.1. Speech synthesizer

The articulatory synthesizer VocalTractLab 2.1 (VTL) includes a detailed geometrical 3D model of the vocal tract (Birkholz, 2013a), an advanced self-oscillating bar-mass model of the vocal folds (Birkholz

et al., 2011a,c), an aero-acoustic simulation based on a time-varying branched tube system of the vocal apparatus (Birkholz & Jackèl, 2004; Birkholz, 2005) and a gestural score for articulatory control (Birkholz, 2007; Birkholz et al., 2011b). The gestural score contains eight tiers to define sequences of eight types of articulatory gestures (Figure 1). Each gesture controls the movement of the participating articulators (in terms of vocal tract or vocal fold model parameters) towards a target configuration. The upper five tiers define the supraglottal articulation with tract-forming gestures for vowels, lip gestures, tongue tip gestures, and tongue body gestures to form consonantal constrictions or closures, and velic gestures to control velum movements. The lower three tiers define the laryngeal articulation by glottal shape gestures, F0 gestures, and lung pressure gestures.

2.2. Stimulus creation

The gestural scores for the stimuli were manually created using the graphical editor in VTL (Figure 1) and the synthetic speech signals were saved as 16 bit, 22050 Hz WAV files for the perception experiment.

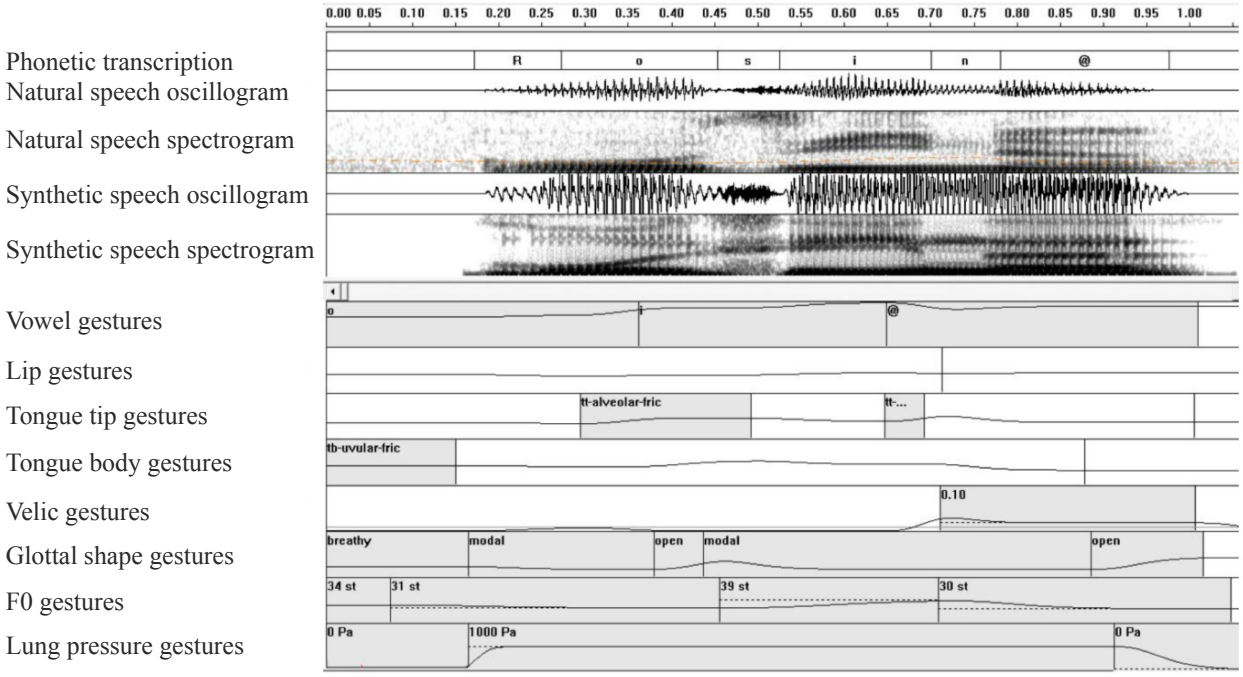

Figure 1: Gestural score editor to re-synthesize natural utterances with the articulatory speech synthesizer (lower part). The upper part shows the waveforms and spectrograms of the natural and synthetic utterances for the word "Rosine".

2.2.1. Re-synthesized stimuli

In the first step, we re-synthesized the natural utterances as accurately as possible in terms of phone durations, pitch contour, and voice quality. As articulatory targets for the phones we used the set of German phone targets provided for the standard speaker of VTL. Each gesture is associated with a time constant that controls how fast the corresponding articulatory target is approximated (Birkholz et al., 2011b). The smaller the time constant, the faster the target is reached. To reduce the number of manually tunable gestural parameters, we fixed the time constants for most types of gestures, so that the manual work was mostly restricted to the adjustment of start and end times of gestures. The time constants were set to values that reflect the typical differences of articulator velocities. According to the survey by Stevens (1998), lip movements are usually faster than tongue tip movements, and tongue tip movements are faster than tongue body movements. To account for these differences, the following time constants were used: 10 ms for lip

gestures, 15 ms for tongue tip gestures, 20 ms for tongue body gestures, and 15 ms as an intermediate value for tract-forming vowel gestures.

The voice quality (phonation type) of the original utterances, as measured by the peak slope parameter (Kane & Gobl, 2011), was manually reproduced in terms of appropriate glottal shape gestures. These gestures mainly control the degree of glottal abduction and affect the voice quality along the continuum from breathy to pressed phonation (Birkholz et al., 2011c). The pitch contour of the original utterances was closely reproduced in terms of F0 gestures analogously to Birkholz et al. (2015). The model underlying the F0 contour is the target approximation model by Prom-on et al. (2009). The subglottal pressure was set to a constant value of 1 kPa in all synthetic stimuli.

2.2.2. Stimuli with manipulated vocal tract length

For each of the nine re-synthesized stimuli, we created one variant with an elongated vocal tract and one with a shortened vocal tract. Therefore, we modified the vocal tract shapes associated with the articulatory gestures. To elongate the vocal tract, speakers have the possibilities to lower the larynx and to protrude the lips. To shorten the the vocal tract, they can raise the larynx and spread the lips. However, these actions are not independent from the rest of the vocal tract, because the articulators are bio-mechanically coupled. Furthermore, a speaker has to make sure that the "acoustic signature" of a phone is preserved when the vocal tract length is altered. It is not exactly clear how phonemes differ acoustically when they are produced by the same speaker with a "normal", an elongated and a shortened vocal tract. Therefore, we considered female-male differences in formant frequencies as reference point for acoustic differences of long and short vocal tracts. The vocal tract of men is on average about 20 % longer than that of women (Simpson, 2001). Vowel formant frequencies of men are therefore on average lower than those of women. However, the scaling is nonuniform and requires different scale factors for each formant and vowel category. Fant (1975) conducted an extensive study on female-male formant differences across populations of speakers of six languages, where he obtained vowel-dependent scale factors k_1, k_2 and k_3 for F_1, F_2 and F_3. These factors, transferred to the German phoneme set, are shown by the solid lines in Figure 2. The factors connected by the solid lines in the upper half of the figure are the inverse of the factors in the lower half, i. e., the factors for shortening and lengthening the vocal tract, respectively.

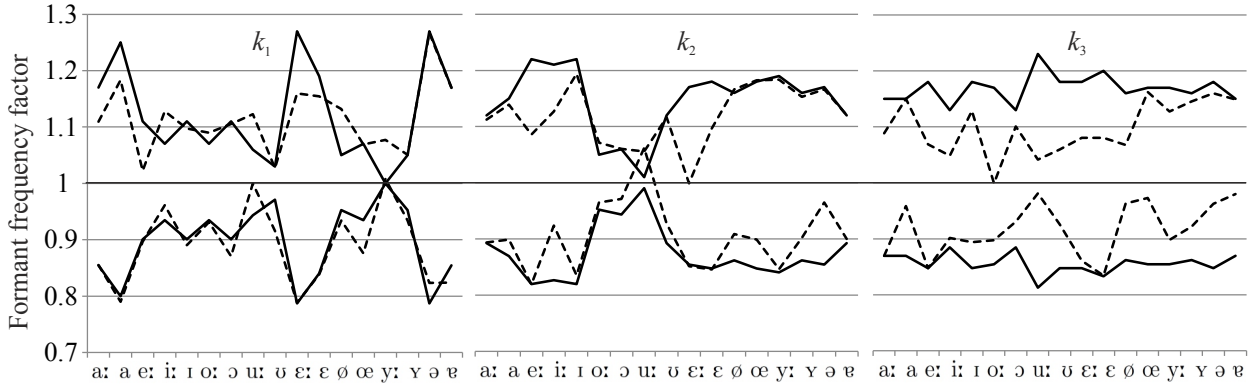

Figure 2: Solid lines: Vowel-dependent formant frequency factors to convert vowels to a shorter vocal tract (upper half) and to a longer vocal tract (bottom) according to Fant (1975). Dashed lines: Factors realized after the acoustic optimization in VTL.

Unfortunately, these factors cannot be used to quantify the acoustic differences of consonants produced by the same speaker with different vocal tract length, and there seems to be no other obvious method to describe the differences. Therefore, we refrained from adapting the articulation of consonants for an elongated or shortened vocal tract in this study and made the adaptation for vowels only.

The adapted vocal tract shapes of vowels were obtained as follows. To create a vowel target with a longer vocal tract, the "standard" target for this vowel (as provided by VTL) was manipulated by protruding the lips by 1 cm and by lowering the larynx by 1 cm. For a shorter vocal tract, the larynx was raised and the lips

were retracted by 1 cm each. For some vowels with an inherently long or short vocal tract, theses changes were constrained by the ranges of the according vocal tract parameters. For example, in the standard target for /u/, the lips were already maximally protruded and the larynx could be lowered by only 5 more millimeters to elongate the vocal tract.

After these rough adaptations, the vocal tract shapes were acoustically optimized. The optimization method, a greedy coordinate descent algorithm (Birkholz, 2013a), automatically adjusted the vocal tract shapes to minimize the formant error

$$E = \sqrt{\frac{1}{3}\left(\left(1-\frac{F1}{F1'}\right)^2 + \left(1-\frac{F2}{F2'}\right)^2 + \left(1-\frac{F3}{F3'}\right)^2\right)}, \quad (1)$$

where F_1, F_2 and F_3 are the formant frequencies produced with the vocal tract shape, and F_1', F_2' and F_3' are the target formant frequencies. The target formants for the shorter/longer vocal tract were obtained by multiplying the formants of the standard shape with the factors in the upper/lower half of Figure 2, i. e., based on the formant differences between men and women. The changes in vocal tract shape due to the optimization were constrained to deviations of the midsagittal vocal tract contour of maximal 5 mm from the start shape to prevent unnaturally large changes. This is also the reason why we applied the gross changes of larynx height and lip protrusion before this optimization.

The formant factors for the shorter/longer vocal tract variants after optimization are connected by dashed lines in the upper/lower half of Figure 2, which are fairly close to the target formants for F_1 and F_2 for most vowels. As an example of the adjustments, Figure 3a shows the vowel [e:] with normal and longer vocal tract. Here we see that the optimization algorithm mainly affected the tongue shape and preserved the initial length changes by means of the lowered larynx and the protruded lips.

In the gestural scores using the adapted vocal tract target shapes, the timing of the gestures was carefully re-adjusted to ensure that the phone durations remained equal to the durations in the standard stimuli.

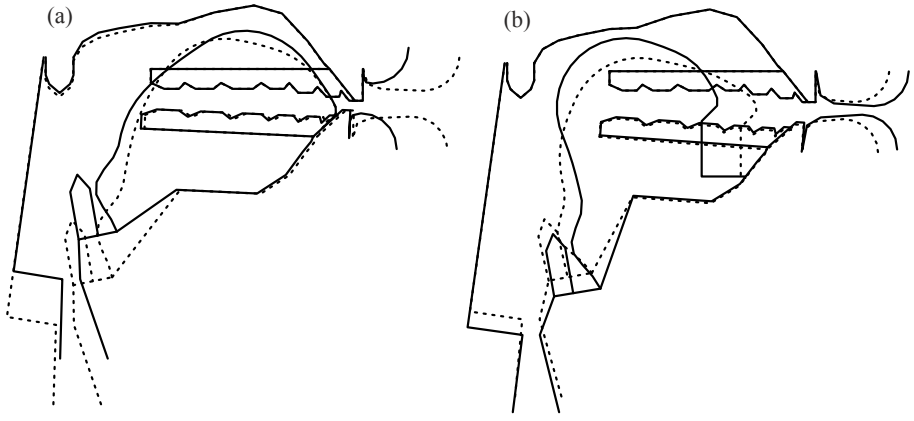

Figure 3: a) Articulatory targets for [eː] with normal (solid line) and long (dashed line) vocal tract length. b) Articulatory targets for [uː] with normal (solid line) and very centralized (dashed line) articulation.

2.2.3. Stimuli with manipulated articulatory precision

Articulatory precision refers to how well a phonetic segment is produced to resemble its canonical form. Deviations from the canonical form may result from undershoot, overshoot or centralization of articulatory targets (Fourakis, 1991). Undershoot can occur due to lack of time (Lindblom, 1963; Cheng & Xu, 2013) or effort (Lindblom, 1990), and centralization is a proposed process that moves weakly articulated vowels towards a more neutral place in the vowel plane, i.e., towards schwa (van Bergem, 1993). In the present study, we therefore tested two options for the manipulation of articulatory precision, namely centralization and variation of articulatory effort, each in two degrees.

Hence, the first kind of manipulation involved the centralization of all vowels and consonants towards a more neutral (schwa-like) configuration. When $\vec{x}_{\text{standard}}$ and $\vec{x}_{\text{schwa}}$ denote the vectors of vocal tract parameters for the "standard" shape of a certain vowel and schwa, respectively, then the centralized variant of the vowel was calculated as the weighted sum

$$\vec{x}_{\text{centralized}} = \alpha \cdot \vec{x}_{\text{standard}} + (1 - \alpha) \cdot \vec{x}_{\text{schwa}} \tag{2}$$

with $0 \leq \alpha \leq 1$. The slightly centralized vowels were obtained with $\alpha = 0.75$, and the very centralized vowels with $\alpha = 0.5$. As an example, Figure 3b shows the very centralized version of [u:] (dotted line).

Consonants in VTL are specified in terms of three vocal tract target shapes each. These represent the articulation of the corresponding consonant in the context of the corner vowels /a/, /i/ and /u/ and so represent its range of coarticulatory variation. The actual target for a consonant in a given vowel context is realized by bilinear interpolation between these corner targets (Birkholz, 2013a). For the generation of the centralized stimuli, the predefined consonantal corner targets were centralized analogously to Equation (2) as the weighted sum of the original targets and a maximally centralized target (equivalent to schwa). When the original corner targets of a consonant in the context of /a/, /i/ and /u/ are denoted as $\vec{x}_{\text{a}}$, $\vec{x}_{\text{i}}$ and $\vec{x}_{\text{u}}$, the maximally centralized target was calculated as

$$\vec{x}_{\text{max-centralized}} = (\vec{x}_{\text{a}} + \vec{x}_{\text{i}} + \vec{x}_{\text{u}})/3. \tag{3}$$

For the second kind of manipulation, i.e., the manipulation of articulatory effort, the standard phone targets were used, but the speed of articulatory target approximation was varied for all gestures in the gestural scores. For the stimuli with lower effort, the time constants of all vocalic and consonantal gestures were doubled compared to the values used for the re-synthesis (Sec. 2.2.1). For the stimuli with higher effort, the time constants were correspondingly halved (e. g., from 10 ms to 5 ms for tongue tip gestures). The change of a time constant and hence the velocity of the articulatory transition between two phones A and B may cause the durations of A and B to change. For an increased velocity, the target for B is reached earlier, such that the duration of phone A decreases and the duration of B increases (and vice versa). However, because we wanted to analyze the effect of target approximation velocity exclusively without any confounding effect due to phone duration changes, we compensated the phone duration changes manually by adjusting the starting times of the transitions (gestures).

2.2.4. Nasalized stimuli

The nasalized stimuli were supposed to have a nasal voice quality. Therefore, the gestural scores of the standard stimuli were modified by opening the velo-pharyngeal port during all vowels and approximants. Prior to this study we experimented with synthesized vowels with different degrees of velo-pharyngeal port openings and found in informal listening tests that the degree of opening can hardly be distinguished as soon as it is above a certain threshold, e.g., the perceptual difference of nasalized vowels with 10 % and 30 % of the maximal possible opening is very small while the difference between 0 % (closed port) and 10 % (open port) is clearly perceivable. Therefore, we decided to set the port opening to a fixed value of 10 % of its maximum possible opening in the articulatory synthesizer, leading to an actual value of 0.2 cm^2. During obstruents, the velo-pharyngeal port opening was not manipulated, i. e., it was kept closed.

2.3. Subjects and experiment

Sixteen subjects (10 female, 6 male) between 19 and 49 years (mean: 27 years) participated in the perception experiment. All participants were native speakers of German. No participant was familiar with articulatory speech synthesis. The experiment was performed individually for each subject with high-quality closed headphones (type J88i by JBL) connected to a laptop computer. The experiment was created and conducted using the software Praat (Boersma & Weenik, 2014). It consisted of four tasks with a total duration of about 30 minutes and took place in a quiet room. A different randomized order of the stimuli (and stimulus pairs) was used for each participant. In every task, each stimulus (or stimulus pair) could be played a second time if requested by the user.

2.3.1. Task 1

The first task was designed to evaluate the naturalness of the generated stimuli. The 72 stimuli were presented to the subjects one after another. After hearing a stimulus the subjects were asked to rate the naturalness of the word on a Likert scale from 1 to 4 (1: *very unnatural*, 2: *rather unnatural*, 3: *rather natural*, 4: *very natural*).

2.3.2. Tasks 2-4

In tasks 2-4 stimuli were presented in pairs and the subjects had to click one of two buttons on the computer screen to select one stimulus of each pair in response to a question.

In task 2, the subjects were asked to decide which stimulus was spoken by a taller person. The task contained 27 pairings of stimuli with normal or modified vocal tract length, where each pairing was presented twice in the two possible orders of the two stimuli (54 pairs of stimuli in total):

- 18 pairs (9 words $\times$ 2 orders) of standard stimuli vs. stimuli with a longer vocal tract,
- 18 pairs of standard stimuli vs. stimuli with a shorter vocal tract,
- 18 pairs of stimuli with a longer vs. a shorter vocal tract.

In task 3, the subjects were asked to decide which stimulus was spoken more carefully. This task included 54 pairs of stimuli with normal or modified articulatory precision in two different orders (108 pairs of stimuli in total):

- 18 pairs of stimuli with higher vs. lower effort,
- 18 pairs of standard stimuli vs. stimuli with higher effort,
- 18 pairs of standard stimuli vs. stimuli with lower effort,
- 18 pairs of stimuli with very centralized vs. slightly centralized articulation,
- 18 pairs of standard vs. very centralized stimuli,
- 18 pairs of standard vs. slightly centralized stimuli.

In task 4, the subjects were asked to decide which stimulus was spoken with a more nasalized voice. This task contained 18 pairs of stimuli, i. e., the nasalized stimuli versus the standard stimuli of the nine words in two different orders.

3. Results and Discussion

Figure 4a shows the results of the first task where the subjects rated the naturalness of the stimuli. The standard stimuli, i. e., those without any further manipulations, were rated as most natural with a mean score of 2.96. Among the stimuli variants with prosodic manipulations, the stimuli with higher effort were rated best with a mean score of 2.84, and the very centralized stimuli received the lowest mean score of 2.21. As detailed in Appendix A, there was no statistically significant difference of the ratings across the nine basis words for the standard stimuli, but the basis words had an effect on the ratings for the feature-manipulated stimuli.

The results of the tasks 2-4 (discrimination tests) are shown in Figure 4b. Here, the bars show the percentages of "correct answers" (with respect to the intended prosodic features) added up across all raters and words. Using the binomial test we tested the hypothesis that the raters identified the correct stimuli in response to the questions more often than by chance (corresponding to 50 % correct answers). For all pairs the proportion of correct answers was significantly higher than 50 % with $p < .001$, except for the stimuli pairs "higher effort vs. standard" ($p = .216$).

The pairs of the second task including stimuli with varied vocal tract length and standard stimuli could be discriminated by the subjects with correct ratings between 78 % and 94 %. Here, the stimuli with a

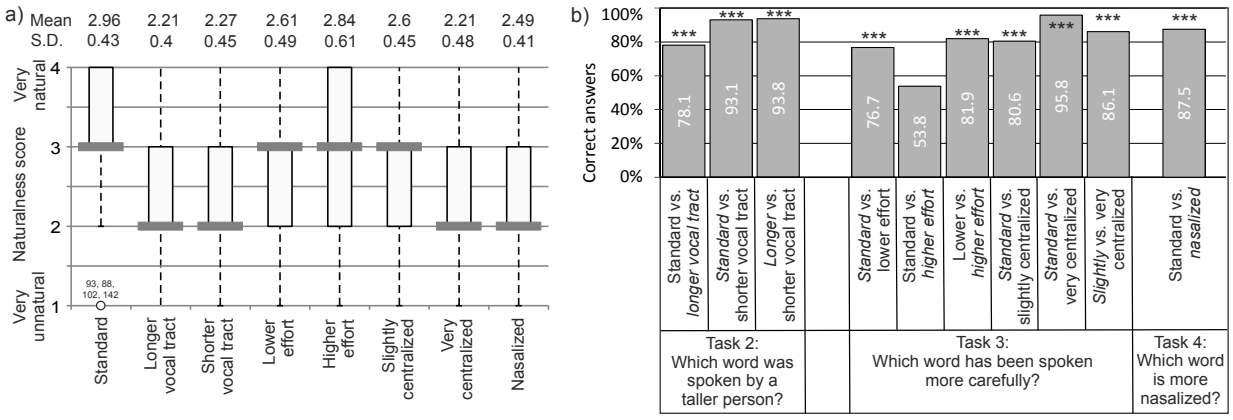

Figure 4: a) Boxplots for the naturalness scores of the re-synthesized and the feature-manipulated stimuli (presented analogously to (Clark et al., 2007)). The numbers given above the boxplots are the mean values and standard deviations (S.D.) of the ratings (1 = *very unnatural*; 4 = *very natural*) b) Percentage of correct answers in the discrimination tasks (binomial test, $\alpha = 0.05$, *** $p < .001$). The stimuli of the pairs in tasks 2-4 that were considered as correct answers to the questions are printed in italic letters.

shorter vocal tract could be better discriminated from the standard stimuli than the stimuli with a longer vocal tract. This was probably due to the fact that the standard voice of the synthesizer is an adult male voice, which already has a long vocal tract (in comparison to female or child voices). The high recognition rates of all the stimuli with manipulated vocal tract length show that the re-synthesis of this prosodic feature was effective.

The results of the third task (articulatory precision) showed that the subjects could discriminate the stimuli in all these pairs significantly well, except for the pairs "higher effort vs. standard". For the latter, the percentage of correct answers was marginally above chance level (54 %). Thus, the increase of articulatory effort did not result in a reliable perception of "more careful speech", i. e., increased articulatory precision. Possibly, the articulatory precision was already rather high in the natural words used for the re-synthesis, so that there was little room for improvement. Picart et al. (2014) showed that also from an acoustic point of view, neutral and hyper-articulated speech (more careful speech) are closer together than neutral and hypo-articulated speech (less careful speech). Therefore they analyzed the size of the vowel space for these different ways of speaking and found a size of 0.201 kHz^2 for neutral speech, 0.059 kHz^2 for hypo-articulated speech, and 0.274 kHz^2 for hyper-articulated speech. Consistent with these findings, the stimuli pairs including the stimuli with lower effort and centralized articulation achieved very high recognition rates (between 77 % and 96 %). Therefore, to achieve the perceptual effect of reduced articulatory precision, both centralizing and lowering the articulatory effort are effective. However, if articulation is centralized too much, the naturalness of the stimuli drops (mean naturalness score of 2.21 for very centralized stimuli vs. 2.6 for slightly centralized stimuli, or median scores of 2 vs. 3).

The recognition of the nasalized stimuli in task 4 was high with a rate of 88 %. This demonstrates that the manipulation with a slightly opened velo-pharyngeal port is highly effective to produce a nasalized voice quality. For all kinds of prosodic manipulations, the basis words had a significant effect on the recognition rates (see Appendix A).

Despite the effectiveness of all but one of the proposed articulatory manipulations for the examined prosodic features, there are a couple of limitations of the present study that should be addressed in more detail in future work. Limitations are (1) that vocal tract length was only manipulated for vowels, (2) that the co-variation of articulatory features was not considered, (3) that the feature-manipulated stimuli were sounding less natural than the standard stimuli, (4) that we did not use natural speech stimuli as controls in the naturalness ratings, and (5) that we used a set of only nine basis words.

With regard to limitation (1), we are currently not aware of any study that indicates how *consonantal* articulation changes when we deliberately elongate or shorten the vocal tract. While protrusion or retraction

of lips and lowering or raising of the larynx may be perceptually adequate for some consonants like nasals, there are certain constraints on lip shape especially for fricatives. The real vocal tract adjustments for consonant production with a longer or shorter vocal tract should be investigated in a future study.

With regard to limitation (2), future studies should also investigate the joint manipulation of features. This could reveal, e. g., if the impression of reduced articulatory precision can be achieved in a more natural way when articulatory effort is reduced and articulatory targets are centralized at the same time.

With regard to limitation (3) it might be conjectured that the results of tasks 2-4 are partly confounded by the differences in naturalness of standard and feature-manipulated stimuli. Hence, if a person is asked to identify the one of two stimuli which is "spoken by a taller person", "spoken more carefully", or "more nasalized", he might tend to simply pick the stimulus which sounds more odd. However, it seems like this was not necessarily the case, because for some of the comparisons (e. g. "standard vs. shorter vocal tract") the correct answer was the standard stimulus (and not the manipulated stimulus with the shorter vocal tract), which was identified very well.

Due to limitation (4), i.e., the lack of natural control stimuli in task 1, we cannot make *absolute* claims about the naturalness of the synthetic stimuli. However, the results of task 1 do tell us to what extent a certain prosodic manipulation reduced the naturalness compared to the unmanipulated stimuli. In the context of the present study, this helps to evaluate the overall success of the proposed prosodic manipulations. That is, a certain manipulation can be considered as most successful in the case of high recognition rates of the corresponding stimuli in tasks 2–4 along with little decrease of naturalness due to the manipulation (task 1).

With regard to limitation (5), the small set of nine basis words does not cover all phones of the target language. So there is a certain probability that the results of the experiment would differ if different basis words were used. However, we think that the results are still generalizable to longer utterances with broader phonetic coverage, because there is evidence that prosodic features can already be detected from very short segments of speech, as demonstrated for examples in studies on vocal emotions (Waaramaa et al., 2008; Patel et al., 2011). Therefore, we would expect the recognition of the prosodic manipulations to improve with the length of the utterance. This is also supported by the additional analysis in Appendix A that shows the tendency that the prosodic manipulations are recognized better for longer words. However, it cannot be fully excluded that other potentially negative perceptual effects of the manipulations become evident only at the level of sentences or longer utterances.

4. Conclusions

This study demonstrated that the secondary prosodic features of vocal tract length, nasality and articulatory precision can be created by simple rule-based articulatory manipulations of neutrally-spoken re-synthesized words. The perceptual prosodic correlates of the articulatory variations were identified well by the subjects in a listening experiment. Only the articulatory manipulations made for increased articulatory precision were not successful. The studied prosodic features are potentially important factors for the expression of vocal emotions, speaker traits and other paralinguistic aspects of speech (Birkholz et al., 2015; Gibbs, 1986; Xu et al., 2013). It is now possible to examine the contribution of these (secondary) features for the implementation of paralinguistic functions, either alone or in combination, in greater detail using articulatory speech synthesis. In the longer term, articulatory speech synthesis could provide the platform for text-to-speech synthesis with the ability to convey a variety of paralinguistic information.

Acknowledgements

The authors would like to thank all volunteers for their participation in the perception experiments and the two anonymous reviewers for their valuable comments on an earlier version of the paper.

		Banane [banaːnə]	Birne [bɪrnə]	Blaubeere [blaʊbeːʀə]	Himbeere [hɪmbeːʀə]	Mandarine [mandaʀiːnə]	Melone [məloːnə]	Mirabelle [mirabɛlə]	Orange [oʀanʒə]	Rosine [ʀoziːnə]
Group	Tract length	85.4	92.7	90.6	94.8	87.5	87.5	84.4	82.3	89.6
	Effort	76.0	60.4	82.3	52.1	83.3	76.0	71.9	56.3	79.2
	Centralization	87.5	69.8	91.7	81.3	93.8	92.7	85.4	89.6	95.8
	Nasality	71.9	84.4	71.9	90.6	96.9	96.9	90.6	90.6	93.8
	Mean	81.9	75.3	86.6	77.5	89.1	86.6	81.6	77.5	88.8

Table 2: Recognition rates of the intended manipulations (divided into four groups) for the individual words (frequencies of correct answers in %). The last row provides the average recognition rate across all different manipulations per word.

Feature group	N	Cochran's Q	DOF	Asymp. sig
Vocal tract length	96	16.618	8	0.034
Effort	96	59.406	8	0.000
Centralization	96	48.068	8	0.000
Nasality	32	22.065	8	0.005

Table 3: Effect of the basis words on the recognition rates based on Cochran's Q test.

Feature group	N	Chi-Square	DOF	Asymp. sig
Vocal tract length	32	28.267	8	0.000
Effort	32	44.128	8	0.000
Centralization	32	50.671	8	0.000
Nasality	16	33.619	8	0.000
Standard	16	12.698	8	0.123

Table 4: Effect of the basis words on the naturalness scores based on Friedman's test.

Appendix A. Further analysis of results

For the analysis in Sec. 3, the recognition rates of the intended prosodic manipulations in tasks 2–4 were pooled across all 16 subjects and 9 words. In Table 2 we show in more detail the effect of the words on the recognition rates for the different manipulations. Here, the prosodic manipulations are divided into four groups, where the group "Vocal tract length" contains the stimuli with longer and shorter vocal tracts, the group "Effort" contains the stimuli with higher and lower effort, the group "Centralization" contains the slightly centralized and very centralized stimuli, and the group "Nasality" contains the nasalized stimuli. In each of the groups, there was a significant difference in the proportion of correct answers across the nine words based on Cochran's Q test ($p < 0.05$, see Table 3). On average, the word "Mandarine" was most effective to convey the intended manipulations, and the word "Birne" was least effective. The data show a moderate positive correlation between the average recognition rate of the intended manipulations and the word length in phonemes (Pearsons $r = 0.49$). However, due to the low number of nine words we do not know whether the correlation is statistically significant ($p = 0.183$).

With regard to task 1, we used Friedman's test to determine whether the basis words had a significant effect on the naturalness scores. Therefore, the feature-manipulated stimuli were again divided into four feature groups as above to detect a potential influence of the basis words. Except for the (unmanipulated) standard stimuli, the difference of the perceived naturalness across the basis words was highly significant ($p < 0.001$) in all feature groups (see Table 4).

References

Airas, M., & Alku, P. (2006). Emotions in vowel segments of continuous speech: Analysis of the glottal flow using the normalised amplitude quotient. *Phonetica*, *63*, 26–46.

Aryal, S., & Gutierrez-Osuna, R. (2016). Data driven articulatory synthesis with deep neural networks. *Computer Speech & Language*, *36*, 260–273.

Beller, G., Obin, N., & Rodet, X. (2008). Articulation degree as a prosodic dimension of expressive speech. In *Fourth International Conference on Speech Prosody*.

van Bergem, D. R. (1993). Acoustic vowel reduction as a function of sentence accent, word stress, and word class. *Speech Communication*, *12*, 1–23.

Birkholz, P. (2005). *3D-Artikulatorische Sprachsynthese*. Logos Verlag Berlin.

Birkholz, P. (2007). Control of an articulatory speech synthesizer based on dynamic approximation of spatial articulatory targets. In *Interspeech 2007 - Eurospeech* (pp. 2865–2868). Antwerp, Belgium.

Birkholz, P. (2013a). Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PLoS ONE*, *8*, e60603.

Birkholz, P. (2013b). VocalTractLab [software]. URL: `http://www.vocaltractlab.de`.

Birkholz, P., & Jackèl, D. (2004). Influence of temporal discretization schemes on formant frequencies and bandwidths in time domain simulations of the vocal tract system. In *Interspeech 2004-ICSLP* (pp. 1125–1128). Jeju, Korea.

Birkholz, P., Kröger, B. J., & Neuschaefer-Rube, C. (2011a). Articulatory synthesis of words in six voice qualities using a modified two-mass model of the vocal folds. In *First International Workshop on Performative Speech and Singing Synthesis (p3s 2011)*. Vancouver, BC, Canada.

Birkholz, P., Kröger, B. J., & Neuschaefer-Rube, C. (2011b). Model-based reproduction of articulatory trajectories for consonant-vowel sequences. *IEEE Transactions on Audio, Speech and Language Processing*, *19*, 1422–1433.

Birkholz, P., Kröger, B. J., & Neuschaefer-Rube, C. (2011c). Synthesis of breathy, normal, and pressed phonation using a two-mass model with a triangular glottis. In *Interspeech 2011* (pp. 2681–2684). Florence, Italy.

Birkholz, P., Martin, L., Willmes, K., Kröger, B. J., & Neuschaefer-Rube, C. (2015). The contribution of phonation type to the perception of vocal emotions in German: an articulatory synthesis study. *Journal of the Acoustical Society of America*, *137*, 1503–1512.

Black, A. W. (2003). Unit selection and emotional speech. In *Interspeech 2003* (pp. 1649–1652). Geneva, Switzerland.

Boersma, P., & Weenik, D. (2014). Praat: doing phonetics by computer [software]. URL: `http://www.praat.org/`.

Burkhardt, F. (2009). Rule-based voice quality variation with formant synthesis. In *Interspeech 2009* (pp. 2659–2662). Brighton, UK.

Burkhardt, F., & Sendlmeier, W. F. (2000). Verification of acoustical correlates of emotional speech using formant-synthesis. In *ISCA Workshop on Speech and Emotion 2000* (pp. 151–156). Newcastle, Northern Ireland, UK.

Campbell, N., & Mokhtari, P. (2003). Voice quality: the 4th prosodic dimension. In *The 15th International Congress of Phonetic Sciences* (pp. 2417–2420). Barcelona, Spain.

Cheng, C., & Xu, Y. (2013). Articulatory limit and extreme segmental reduction in Taiwan Mandarin. *Journal of the Acoustical Society of America*, *134*, 4481–4495.

Chuenwattanapranithi, S., Xu, Y., Thipakorn, B., & Maneewongvatana, S. (2008). Encoding emotions in speech with the size code – A perceptual investigation. *Phonetica*, *65*, 210–230.

Clark, R. A. J., Podsiadlo, M., Fraser, M., Mayo, C., & King, S. (2007). Statistical analysis of the Blizzard Challenge 2007 listening test results. In *Proc. Blizzard Challenge Workshop 3-2007 (in Proc. SSW6)*. Bonn, Germany.

van den Doel, K., Vogt, F., English, R. E., & Fels, S. (2006). Towards articulatory speech synthesis with a dynamic 3d finite element tongue model. In *7th International Seminar on Speech Production (ISSP '06)*. Ubatuba, Brazil.

Fant, G. (1975). Non-uniform vowel normalization. *STL-QPSR*, *16*, 1–19.

Fourakis, M. (1991). Tempo, stress, and vowel reduction in American English. *Journal of the Acoustical Society of America*, *90*, 1816–1827.

Gibbs, R. W. (1986). On the psycholinguistics of sarcasm. *Journal of Experimental Psychology: General*, *115*, 3–15.

Gobl, C., & Ní Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, *40*, 189–212.

Grichkovtsova, I., Morel, M., & Lacheret, A. (2009). Perception of affective prosody in natural and synthesized speech: Which methodological approach? In S. Hancil (Ed.), *The Role of Prosody in Affective Speech* (pp. 371–390). Peter Lang.

Hunt, A. J., & Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-96)* (pp. 373–376). Atlanta, Georgia.

Iida, A., Campbell, N., Higuchi, F., & Yasumura, M. (2003). A corpus-based speech synthesis system with emotion. *Speech Communication*, *40*, 161–187.

Kane, J., & Gobl, C. (2011). Identifying regions of non-modal phonation using features of the wavelet transform. In *Interspeech 2011* (pp. 177–180). Florence, Italy.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, *67*, 971–1000.

Ladd, D. R. (2008). *Intonational Phonology*. Cambridge University Press.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, *35*, 1773–1781.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle, & A. Marchal (Eds.), *Speech production and speech modeling* (pp. 413–415). Dordrecht, The Netherlands: Kluwer.

Murray, I. R., & Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustical Society of America*, *93*, 1097–1108.

Murray, I. R., & Arnott, J. L. (1995). Implementation and testing of a system for producing emotion-by-rule in synthetic speech. *Speech Communication*, *16*, 369–390.

Patel, S., Scherer, K. R., Björkner, E., & Sundberg, J. (2011). Mapping emotions into acoustic space: The role of voice production. *Biological Psychology*, *87*, 93–98.

Pfitzinger, H. R. (2006). Five dimensions of prosody: Intensity, intonation, timing, voice quality, and degree of reduction. In R. Hoffmann, & H. Mixdorff (Eds.), *Studientexte zur Sprachkommunikation: Speech Prosody Abstract Book* (pp. 6–9). TUDPress, Dresden.

Picart, B., Drugman, T., & Dutoit, T. (2014). Analysis and HMM-based synthesis of hypo and hyperarticulated speech. *Computer Speech & Language*, *28*, 687–707.

Prom-on, S., Xu, Y., & Thipakorn, B. (2009). Modeling tone and intonation in Mandarin and English as a process of target approximation. *Journal of the Acoustical Society of America*, *125*, 405–424.

Scherer, K. R. (1978). Personality inference from voice quality: the loud voice of extroversion. *European Journal of Social Psychology*, *8*, 467–487.

Scherer, K. R., Johnstone, T., & Klasmeyer, G. (2003). Vocal expression of emotion. In R. J. Davidson, H. Goldsmith, & K. R. Scherer (Eds.), *Handbook of the affective sciences* (pp. 433–456). Oxford University Press, New York and Oxford.

Scherer, K. R., Sundberg, J., Tamarit, L., & Salomão, G. L. (2015). Comparing the acoustic expression of emotion in the speaking and the singing voice. *Computer Speech & Language*, *29*, 218–235.

Schröder, M. (2001). Emotional speech synthesis: A review. In *Interspeech 2001* (pp. 561–564). Aalborg, Denmark.

Schröder, M., Burkhardt, F., & Krstulovic, S. (2010). Synthesis of emotional speech. In K. R. Scherer, T. Bänziger, & E. Roesch (Eds.), *Blueprint for affective computing* (pp. 222–231). Oxford University Press, London.

Schuller, B., Steidl, S., Batliner, A., & et al. (2015). A survey on perceived speaker traits: Personality, likability, pathology, and the first challenge. *Computer Speech & Language*, *29*, 100–131.

Sendlmeier, W. F., & Heile, A. (1998). Nasalität und Behauchung als Indikatoren für den stimmlichen Ausdruck körperlichen Wohlbehagens. In H. W. Wodarz (Ed.), *Forum Phoneticum 66* (pp. 1–14). FFM.

Simpson, A. P. (2001). Dynamic consequences of differences in male and female vocal tract dimensions. *Journal of the Acoustical Society of America*, *109*, 2153–2164.

Stevens, K. N. (1998). *Acoustic Phonetics*. The MIT Press.

Waaramaa, T., Laukkanen, A. M., Alku, P., & Väyrynen, E. (2008). Monopitched expression of emotions in different vowels. *Folia Phoniatrica et Logopaedica*, *60*, 249–255.

Xu, Y., Kelly, A., & Smillie, C. (2013). Emotional expressions as communicative signals. In S. Hancil, & D. Hirst (Eds.), *Prosody and Iconicity* (pp. 33–60). John Benjamins Publishing Co., Amsterdam.

Yamagishi, J., Onishi, K., Masuko, T., & Kobayashi, T. (2005). Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis. *IEICE Transactions on Information and Systems*, *88*, 502–509.

Zen, H., Tokuda, K., & Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication*, *51*, 1039–1064.