JAIST Repository
https://dspace.jaist.ac.jp/

Title	Acoustic and articulatory analysis and synthesis of shouted vowels
Author(s)	Xue, Yawen; Marxen, Michael; Akagi, Masato; Birkholz, Peter
Citation	Computer Speech & Language, 66: 101156
Issue Date	2020-10-09
Type	Journal Article
Text version	author
URL	http://hdl.handle.net/10119/18074
Rights	Copyright (C)2020, Elsevier. Licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International license (CC BY-NC-ND 4.0). [http://creativecommons.org/licenses/by-nc-nd/4.0/] NOTICE: This is the author's version of a work accepted for publication by Elsevier. Changes resulting from the publishing process, including peer review, editing, corrections, structural formatting and other quality control mechanisms, may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in Yawen Xue, Michael Marxen, Masato Akagi, Peter Birkholz, Computer Speech & Language, 66, 2020, 101156, https://doi.org/10.1016/j.csl.2020.101156
Description	

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Acoustic and articulatory analysis and synthesis of shouted vowels

Yawen Xue[a)*], Michael Marxen[b)], Masato Akagi[a)], Peter Birkholz[c)]

[a)] *School of Information Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa, 923-1292, Japan*

[b)] *Department of Psychiatry and Neuroimaging Center, Technische Universität Dresden, Germany*

[c)] *Institute of Acoustics and Speech Communication, Technische Universität Dresden, Germany*

Abstract

Acoustic and articulatory differences between spoken and shouted vowels were analyzed for two male and two female subjects by means of acoustic recordings and midsagittal magnetic resonance images of the vocal tract. In accordance with previous acoustic findings, the fundamental frequencies, intensities, and formant frequencies were all generally higher for shouted than for spoken vowels. The harmonics-to-noise ratios and H1-H2 measures were generally lower for shouted vowels than for spoken vowels. With regard to articulation, all subjects used an increased lip opening, an increased jaw opening, and a lower tongue position for shouted vowels. However, the changes of vertical larynx position, uvula elevation, and jaw protrusion between spoken and shouted vowels were inconsistent among subjects. Based on the analysis results, a perception experiment was conducted to examine how changes of fundamental frequency, subglottal pressure, vocal tract shape, and phonation type contribute to the perception of stimuli created by articulatory synthesis as being shouted. Here, fundamental frequency had the greatest effect, followed by vocal tract shape and lung pressure, with no measurable effect of phonation type.

Keywords: Shouted speech, articulatory analysis, articulatory synthesis, Magnetic Resonance Imaging

1. Introduction

Currently, both speech recognition and synthesis technology mainly focus on the processing of "neutral" speech. Other speech modes like whispery, soft, loud, or shouted speech are rarely considered [1], although they may be highly relevant for future speech technology systems, e.g., for fully expressive speech synthesis. Among the various speech modes, shouted speech is arguably one of the most extreme modes and requires the highest dynamic change in vocal excitation [1, 2]. Most likely because of this, the performance of speaker verification and speech recognition systems often drops significantly when trained with speech of normal vocal effort and tested with shouted speech [3, 4, 5]. Therefore, a better understanding of shouted speech may benefit not only speech synthesis, but also speech recognition and speaker verification.

Closely related to shouted speech is Lombard speech [6, 7, 8], which is optimized to be understood in noisy environments. Shouted speech is not necessarily well to understand. Instead, its intelligibility can be drastically reduced at very high vocal efforts [9, 10]. However, both shouted and Lombard speech share the same general production mechanisms [11].

Previous studies of the acoustic differences between spoken and shouted speech revealed that shouted speech is characterized by higher fundamental frequency (f_0), increased sound pressure level, longer vowel duration, and a flattened spectral tilt [12, 13, 14, 15, 16, 17, 18], all of which are highly speaker-dependent. The reasons for the acoustic differences are rooted in the glottal excitation and the supraglottal articulation. In particular, the proportion of the duration of the closed phase in a glottal cycle is much higher for shouted than for normal speech [19, 20, 21], and many supraglottal articulators are displaced further in shouted than in normal speech, especially the lips and the jaw [22, 23, 24, 25, 26]. Echternach et al. [27] explored articulatory differences during singing at different loudness levels and found increased lip opening, increased pharynx width, and elevated vertical larynx position for increased loudness. However, little is known about the detailed articulatory differences of normal and shouted speech and the contribution of individual articulatory or acoustic features to the perception of shouted speech.

With regard to speech synthesis, the synthesis of Lombard speech or speech with high vocal effort [3, 28, 29, 30] has been previously examined using concatenative speech synthesis [31], statistical parametric speech synthesis [7, 32], and neural network-based speech synthesis [33]. Concatenative speech synthesis and the recently developed methods for neural end-to-end speech synthesis (e.g. Tacotron [34]) can currently achieve the best speech quality, but not without some limitations. Using for example concatenative synthesis, it is very difficult to

*Corresponding author

Email addresses: xue_yawen@jaist.ac.jp (Yawen Xue[a)]), michael.marxen@tu-dresden.de (Michael Marxen[b)]), akagi@jaist.ac.jp (Masato Akagi[a)]), peterbirkholz@gmx.de (Peter Birkholz[c)])

modify the pre-recorded speech material to sound like another speaking style or mode without a loss of quality. The neural end-to-end synthesizers have different problems like unstability and mispronunciations, and require huge amounts of training data. Furthermore, all of these synthesis methods have in common that they essentially neglect the speech production mechanisms. In contrast, articulatory speech synthesis simulates the process of speech production at the articulatory and acoustic levels to generate speech. Even though it allows the direct manipulation of all the articulatory parameters of interest, this method has so far not been used to explore the differences of normal vs. shouted speech.

In summary, while the *acoustic* differences between normal and shouted speech are rather well known, especially in terms of f_0 and formant frequencies, the study of *articulatory* differences was mainly limited to the jaw and the lips, which are well accessible. Much less is known about the differences of the *lingual* articulation between normal and shouted speech. The only exceptions are a study that used electromagnetic articulography (EMA) of the tongue during Lombard speech [25], and a study that used Magnetic Resonance Imaging (MRI) of the vocal tract to study articulatory differences in different loudness conditions in singing [27]. However, with regard to the latter, it is not clear to what extent the articulatory differences between soft and loud *singing* apply to normal and shouted *speech*. Furthermore, given the range of articulatory and acoustic features that discriminate normal and shouted speech, it is unclear which features contribute to what extent to the perception of speech as normal or shouted. This knowledge could help to improve methods for the parametric synthesis of more expressive speech, the recognition of expressive speech as well as speaker verification [4, 5]. Hence, the purpose of the present study was twofold: 1) to verify and supplement the previous observations of acoustic and articulatory differences between normal and shouted speech, and 2) to quantify the contributions of individual articulatory-acoustic features to the perception of speech as being shouted.

For the first aim, acoustic data and Magnetic Resonance Imaging (MRI) [35, 36] data of the vocal tract of spoken and shouted vowels of two male and two female subject were captured and analyzed with respect to a range of acoustic and articulatory measures. In contrast to the measurement methods used in previous studies on shouted speech, MRI data provide a *complete* picture of the vocal tract including lip, jaw, and tongue articulation. Articulatory measures were taken both from vocal tract contours in midsagittal MR images and from the cross-distance functions (similar to vocal tract area functions) of the vowels. Beyond the acoustic features f_0, sound pressure level, and formant frequencies, which have been analyzed before in the context of shouted speech, we also included the amplitude of the first harmonic relative to that of the second (H1-H2), and the harmonics-to-noise ratio (HNR). H1-H2 is known to correlate with properties of the voice source, most notably with the open quotient, and hence characterizes the glottal flow waveform [37]. HNR, on the other hand, quantifies the amount of noise in the voice signal and hence indicates the degree of hoarseness [38].

For the second aim, we used the results of the acoustic and articulatory analyses to create consonant-vowel syllables that differed with respect to f_0, lung pressure, vocal tract shape, and phonation type, using the articulatory speech synthesizer VocalTractLab 2.2 (`www.vocaltractlab.de`). In a perception experiment, the synthetic stimuli were then rated as sounding spoken or shouted in order to find out which features contribute to what extent to the stimuli sounding shouted.

2. Method

2.1. Data acquisition

For four native German speakers (two male and two female), we captured the vocal tract shapes using Magnetic Resonance Imaging (MRI) and the acoustic signals during the productions of 15 vowels, both in a normal and a shouted speaking style. One female subject (F1) was a trained singer, while the other female subject (F2) and both male subjects (M1, M2) had no special vocal education. The corpus consisted of the German tense vowels /aː, eː, iː, oː, uː, ɛː øː yː/ and the lax vowels /a, ɛ, ɪ, ɔ, ʊ, ʏ, œ/.

The MR images were acquired on a Siemens 3T TIM Trio with a 12-channel head coil combined with additional neck elements. To image the vocal tract, we used a sagittal 3D volume interpolated gradient echo sequence (VIBE - fl3d_vibe) with 1.2 mm x 1.2 mm x 1.8 mm resolution, 16 slices, matrix size 192, field of view = $(230\ \text{mm})^2$, repetition time TR = 7.15 ms, echo time TE = 2.09 ms, flip angle 9 deg, Q-fatsat, 8 lines per shot, 7/8 phase partial Fourier, 6/8 slice partial Fourier, ipat factor 2 (PE only), 24 reference lines and a bandwidth of 220 Hz/pixel. The acquisition time for one volume (i.e., vowel) was 7.3 s. Each subject first produced the 15 vowels in speaking style, and then in shouting style. To retain a stable and natural "shouting" articulation during the rather long scanning periods of 7.3 s per vowel, the subjects were instructed to imagine that they produce each vowel as part of a presented name of a person who they call out to over a long distance (without any particular vocal emotion). The names were selected such that the vowels of interest were in the syllables with primary or secondary stress. In total, 120 volumes were acquired (4 speakers x 15 vowels x 2 styles).

Due to the loud noise in the MRI scanner, high-quality audio recordings of the sustained productions of the spoken and shouted vowels were made during a separate session in a soundproofed audio studio at the same day the MRI data were recorded. During these productions, the subjects lay down on the floor similar to their position in the MRI scanner. They were asked to produce the vowels as similar as possible as they did during the MRI recordings. The audio data were recorded using a measurement microphone (type MK250 by RFT VEB Mikrofontechnik with pre-amplifier MG MV220) connected to a USB audio interface (TASCAM). The microphone was placed at a distance of 30 cm in front of (i.e., above) the lips of the speakers. The audio format was 44.1 kHz, 16 bit, mono. The recording software was Audacity v2.0.2 running on a standard laptop computer.

2.2. Acoustic analysis

The speech data were analyzed with respect to sound pressure level (SPL), fundamental frequency (f_0), the first three formant frequencies (F_1, F_2 and F_3), the amplitude of the first harmonic relative to that of the second (H1-H2), and the harmonics-to-noise ratio (HNR). SPL, f_0 and the first three formant frequencies were determined using the software Praat version 5.4 [39]. H1-H2 and HNR were extracted with the software VoiceSauce version v1.34 [40, 41].

For SPL and f_0, the average values within a 50 ms interval during the stable period of the sound were calculated. Formant frequencies were determined using Praat's built-in LPC formant tracker. For each vowel, the number of LPC coefficients was carefully adjusted following the method outlined in [42]. However, even with optimal parameter settings, LPC-based formant estimates tend to get biased by the nearby harmonics, especially for high-pitched voices as the shouted samples in this study [43]. Hence, the results must be interpreted with this methodological issue in mind.

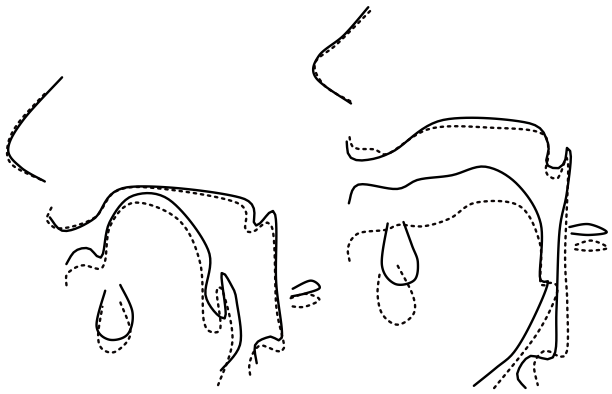

Figure 1: Vocal tract contours of spoken vowels (solid lines) and shouted vowels (dashed lines). Left: Vowel /iː/ of subject F1. Right: Vowel /aː/ of subject M1.

2.3. Articulatory analysis

In the MRI volume of each recorded vowel, the midsagittal image was selected for the further analysis. In each of these midsagittal images, the vocal tract contours were traced using spline curves in the custom-made software Image3D [44]. Besides the anterior-inferior contour and the superior-posterior contour, we also traced the contours of the nose, the jaw bone, and one vertebra. As examples, Figure 1 shows the tracings of the spoken vs. the shouted /iː/ of the female speaker F1, and the spoken vs. the shouted /aː/ of the male speaker M1. The complete set of tracings for all four speakers is available in the supplemental material at `http://www.vocaltractlab.de/index.php?page=birkholz-supplements`.

Based on the contours, a series of distance measures was taken to quantify the vocal tract shapes, similar to Echternach et al. [27, 45]. The distance measures are illustrated in Figure 2. First, two auxiliary lines L1 and L2 were defined (dashed lines). L1 was constructed as the connecting line of the highest point of the hard palate and the lowest part of the occipital bone. This

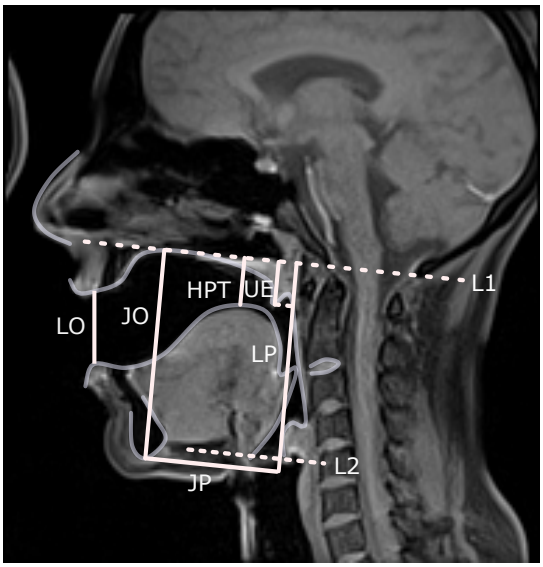

Figure 2: Measured distances in the midsagittal MR images.

line was found to be essentially constant across the vowels of the same speaker (the head was properly fixed in a constant position in the MRI scanner) and was hence a suitable reference line. The line L2 is parallel to L1 and runs through the center point of the vocal folds. The following distance measures were extracted:

- LO (lip opening): the shortest distance between the upper and lower lips.
- JO (jaw opening): the distance from the lowest point of the jaw bone to L1.
- HPT (highest point of tongue): the distance from the highest point of the tongue to L1.
- UE (uvula elevation): the distance from the lower tip of the uvula to L1.
- LP (larynx position): the vertical larynx position defined as the distance between L1 and L2.
- JP (jaw protrusion): distance between the jaw bone and the glottis as measured between the two lines for LP and JO.

In addition to the distance measures, we determined the midsagittal cross-distance function for each vowel, i.e., the distance between the inferior-anterior and the superior-posterior contour lines as a function of the position along the vocal tract center line. For example, at the glottis, the cross-distance corresponds to the length of the vocal folds, and at the lips, the cross-distance is the distance between the upper and lower lips in the midsagittal plane. The center line and the cross-distance functions were automatically determined based on the contour tracings using the custom-made software Image3D as described in detail in Echternach et al. [46]. Note that we did not make any effort to calculate the cross-sectional area functions from the cross-distance functions, because these transformations are

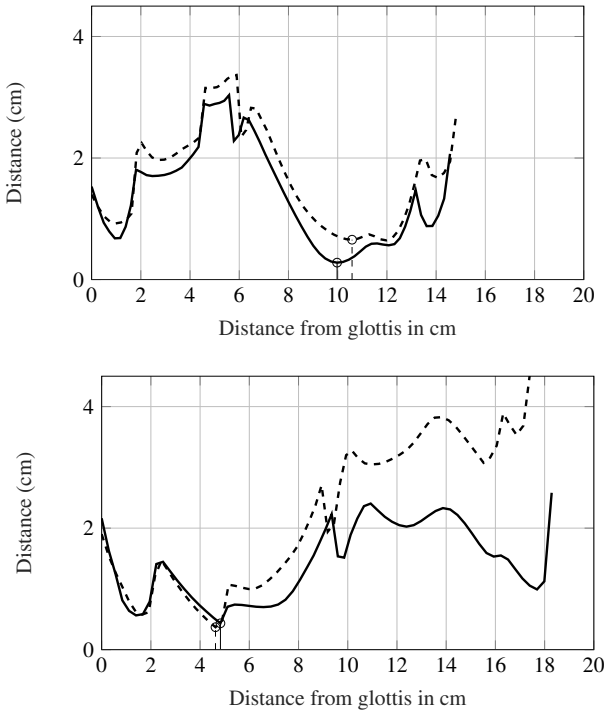

Figure 3: Cross-distance functions of the spoken (solid line) and shouted (dashed line) vowel productions of /iː/ of the speaker F1 (top), and of the spoken and shouted vowel productions of /aː/ of the speaker M1 (bottom). The vertical lines indicate the shortest cross-distances in the vocal tract.

rather vague and speaker-specific [47] and would not add new information to the present analysis. As examples, Figure 3 shows the cross-distance functions for the spoken and shouted /iː/ of the female subject F1, and for the spoken and shouted /aː/ of the male subject M1. The complete set of cross-distance functions is available as supplemental material.

For each cross-distance function, four measures were obtained:

- The vocal tract length in cm, i.e., the length of the vocal tract center line between the glottis and the lips;
- The area under the cross-distance function in cm^2 as an indicator for the total volume of the vocal tract;
- The minimum cross-distance in cm;
- The position of the minimum cross-distance on the center line from the glottis in cm.

2.4. Stimulus creation

To find out how the measured articulatory-acoustic differences of the spoken and shouted vowels contribute to the *perception* of a vowel as spoken vs. shouted, we used articulatory speech synthesis to prepare a series of stimuli for a perception experiment. Here we used the articulatory speech synthesizer VocalTractLab 2.2 [48] (`www.vocaltractlab.de`), which allows the synthesis of arbitrary utterances based on a geometrical 3D model of the vocal tract [49], an advanced self-oscillating bar-mass model of the vocal folds [50, 51], and an aero-acoustic simulation [52]. Utterances are manually created in terms of gestural scores [53, 54] based on the concepts of articulatory phonology [55].

Table 1: The feature settings used for the synthesis of the eight syllables for each of the speakers F1 and M2. The f_0 values without brackets refer to the female speaker F1, and the f_0 values in brackets refer to the male speaker M2.

#	VT shape	f_0	P_{lung}	Phonation type
1	speaking	230 (170) Hz	800 Pa	modal voice
2	speaking	230 (170) Hz	800 Pa	pressed voice
3	speaking	230 (170) Hz	1600 Pa	modal voice
4	speaking	230 (170) Hz	1600 Pa	pressed voice
5	speaking	360 (270) Hz	800 Pa	modal voice
6	speaking	360 (270) Hz	800 Pa	pressed voice
7	speaking	360 (270) Hz	1600 Pa	modal voice
8	speaking	360 (270) Hz	1600 Pa	pressed voice
9	shouting	230 (170) Hz	800 Pa	modal voice
10	shouting	230 (170) Hz	800 Pa	pressed voice
11	shouting	230 (170) Hz	1600 Pa	modal voice
12	shouting	230 (170) Hz	1600 Pa	pressed voice
13	shouting	360 (270) Hz	800 Pa	modal voice
14	shouting	360 (270) Hz	800 Pa	pressed voice
15	shouting	360 (270) Hz	1600 Pa	modal voice
16	shouting	360 (270) Hz	1600 Pa	pressed voice

As stimuli for the perception experiment, we synthesized each of the 8 syllables /baː/, /beː/, /biː/, /boː/, /buː/, /bɛː/, /bøː/, and /byː/ in 16 variants for both the speakers F1 and M2. Hence, for each speakers there were $8 \times 16 = 128$ stimuli. The 16 variants of each syllable consisted of all combinations of four binary articulatory-acoustic features, namely f_0, vocal tract shape, lung pressure P_{lung}, and phonation type [56] shown in Table 1. For each feature, there is one value that is assumed typical for speaking (e.g., P_{lung} = 800 Pa), and one value that is assumed typical for shouting (e.g., P_{lung} = 1600 Pa). For variant 1, all four feature settings are typical for speaking, and for variant 16, all four feature settings are typical for shouting. The variants 2-15 represent mixtures of characteristics of spoken and shouted speech, and hence represent a kind of "vocal effort mismatch". For example, variant 10 mixes the vocal tract shape and phonation type for shouting with the f_0 and lung pressure for speaking.

The two settings (for speaking and shouting) of the individual features were chosen on the following basis: As the two values for f_0, we used the average fundamental frequencies of the spoken and shouted vowels produced by each of the two reference speakers. For the feature "vocal tract shape", we re-created the vocal tract shapes of the female and male subjects obtained by MRI during the production of the spoken and shouted vowels with the 3D vocal tract model of the synthesizer. To this end, we manually adjusted the anatomical dimensions of the vocal tract model to fit the dimensions of the two subjects separately and then adjusted the articulatory parameters to fit the contour lines of the vocal tract model with those extracted from the MR images. As an example, Figure 4 shows

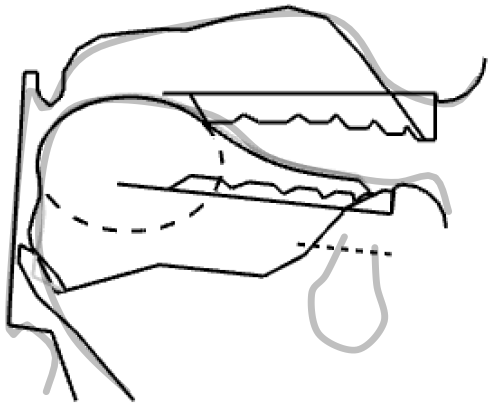

Figure 4: Midsagittal tracing of the vocal tract outline in the MR image (thick gray lines) and manually fitted shape of the vocal tract model (black) for the shouted vowel /a:/ of the speaker F1.

the MRI-based vocal tract contours as thick gray lines, and the fitted vocal tract model contours as thin black lines. For all spoken and shouted vowels of the two subjects, a good visual match of the contours was obtained.

For the features "lung pressure" and "phonation type" we did not measure any direct values for our subjects. Instead, we assumed lung pressures of 800 Pa and 1600 Pa as typical values for speaking and shouting, respectively, and that shouted speech is associated with pressed voice, and normal speech with modal voice. While a lung pressure of 800 Pa is widely assumed for "normal" phonation [57], there is little published data for shouting. Therefore, we decided to use twice the value of "normal" phonation for the synthesis of shouting based on a study on singing at different loudness levels [58]. In this study, loud singing was found to be produced with roughly twice the subglottal pressure of soft phonation. In the VTL synthesizer, increasing the lung pressure leads to wider oscillations of the vocal folds and hence to more sudden opening and closing events of the glottis. This in turn flattens the spectral tilt of the generated speech and increases the sound intensity.

The generation of modal and pressed voice was controlled in terms of the rest displacement of the vocal folds. The self-oscillating bar-mass model of the vocal folds [50] has two parameters to adjust the rest displacement from the glottal midline: x_{lower} for the lower bar mass, and x_{upper} for the upper bar mass. In this study we used the values $x_{\text{lower}} = 0.05$ mm and $x_{\text{upper}} = 0.0$ mm for modal phonation, and $x_{\text{lower}} = 0.0$ mm and $x_{\text{upper}} = -0.05$ mm for pressed phonation. The other control parameters of the vocal fold model were set to their default values for both phonation types (no posterior glottal chink and an aspiration noise factor of -40 dB).

All four features (vocal tract shape, f_0, P_{lung} and phonation type) could be individually varied in the VocalTractLab software to synthesize the 16 variants of each syllable. As described above, P_{lung}, f_0, and the degree of glottal abduction (for the phonation type) are control parameters of the vocal fold model [50], while the vocal tract shapes are defined with the parameters of the vocal tract model [49].

All synthesized signals were peak-normalized, because we were interested in the perceptual effects of the articulatory-acoustic features independently from the volume setting of the playback device. After normalization, the sound pressure levels of the stimuli were very similar to each other, with an average difference of less than 1.5 dB between the extreme variants 1 and 16 of the syllables. The greatest difference between the variants 1 and 16 was 4.84 dB for the syllable /be:/ of speaker M2. All normalized signals were saved as 16 bit, 22050 Hz WAV files for the perception experiment.

2.5. Perception experiment

Fifteen adult, native German subjects (12 male, 3 female) were invited to assess the synthesized stimuli. They evaluated the synthesized stimuli of the female subject in one session, and the stimuli of the male subject in a different session another day. The task of the subjects was to rate how spoken or shouted each stimulus sounded. Each subject was individually seated in an audio studio in front of a computer screen to conduct the experiment. The 128 stimuli were presented in an individually randomized order per subject over high-quality headphones (type AKG K240) connected to a laptop computer. After the presentation of each stimulus, the subject was asked to assess whether the stimulus was 1-"spoken", 2-"rather spoken", 3-"rather shouted", or 4-"shouted", by clicking on one of four buttons with the according label. The subjects could replay each stimulus as often as they liked. The whole experiment took about 10 minutes per subject.

In the second session, the listeners performed a second experiment where they were asked to rate the naturalness of the stimuli. To this end, all stimuli were played again in a new randomized order per subject under the same conditions as for the first experiment. However, this time the listeners were asked to rate the naturalness of the individual stimuli using a four-point Likert scale with the options 1 - "Unnatural", 2 - "Rather unnatural", 3 - "Rather natural", and 4 - "Natural". This experiment was performed to detect potential differences in the naturalness of the stimuli for different articulatory-acoustic feature combinations that might have biased the spoken-shouted ratings in the previous experiment.

3. Result

3.1. Acoustic analysis results

The results of the acoustic analyses are shown in Table 2 and Figure 5. Due to the limited number of subjects, comparisons between the measures obtained from the spoken and shouted samples were made individually for each subject, and two-tailed Student's *t*-tests were used to assess the significance of differences between the two conditions. For all subjects, f_0 and SPL were significantly higher for the shouted vowels than for the spoken vowels ($p < 0.001$). The average increase of f_0 from the spoken to the shouted vowels was 56 %, 28 %, 93 % and 58 % for subject F1, F2, M1 and M2, respectively. The SPL

Table 2: Mean and 95% confidence intervals of fundamental frequency (f_0), sound pressure level (SPL), H1-H2, and harmonics-to-noise ratio (HNR) of the spoken and shouted vowel productions of all subjects.

		f_0 in Hz	SPL in dB	H1-H2 in dB	HNR in dB
M1	spoken	112.3 ± 8.0	46.3 ± 1.3	6.9 ± 1.5	68.4 ± 2.7
	shouted	221.1 ± 15.8	65.5 ± 2.0	0.3 ± 2.3	60.4 ± 4.3
M2	spoken	171.2 ± 5.8	59.2 ± 1.3	4.2 ± 1.3	62.3 ± 3.1
	shouted	269.7 ± 7.8	72.1 ± 1.1	6.9 ± 3.6	54.4 ± 2.2
F1	spoken	231.1 ± 2.7	53.7 ± 2.4	7.0 ± 3.1	75.2 ± 3.2
	shouted	361.3 ± 1.6	67.0 ± 3.3	4.3 ± 3.1	55.6 ± 4.2
F2	spoken	234.5 ± 5.7	51.8 ± 1.5	9.2 ± 2.4	70.9 ± 3.8
	shouted	301.2 ± 5.2	65.8 ± 2.4	5.7 ± 3.4	64.9 ± 2.9

of the shouted vowels was on average 13.3 dB, 14 dB, 19.2 dB, and 13 dB higher than for the spoken vowels for subjects F1, F2, M1, and M2, respectively. These results are in line with previous studies finding that the SPL of shouted speech is usually considerably more than 10 dB higher than that of normal speech, and that the f_0 of shouted speech is 56% higher than that of normal speech [12, 15].

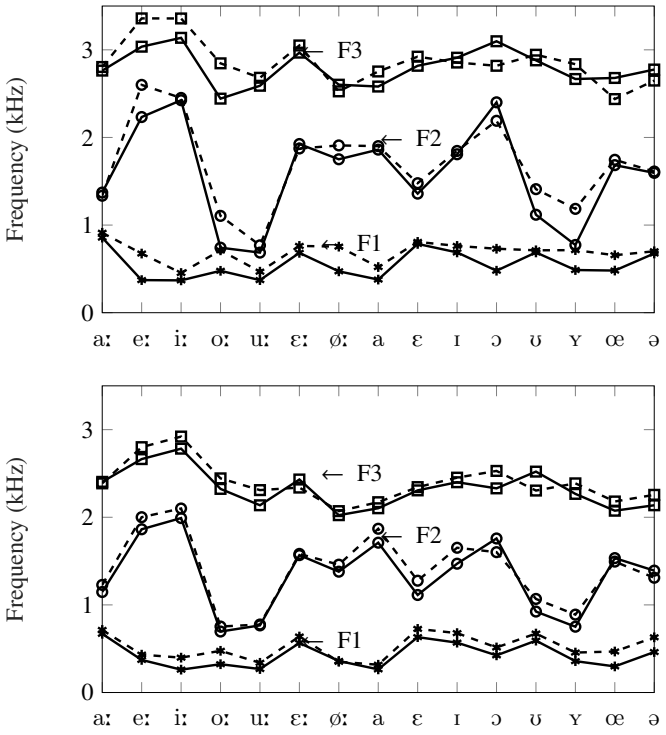

Figure 5: Formant frequencies of the spoken vowels (solid lines) and shouted vowels (dashed lines) of the female speaker F1 (top) and the male speaker M1 (bottom).

Figure 5 shows the formant frequencies determined for the spoken (solid lines) and shouted (dashed lines) vowels for the speakers M1 and F1. This figure shows that the first two formant frequencies were mostly higher for the shouted vowels. Across all four speakers, F_1 of the shouted vowels was on average 25 %, 23 %, 22 %, 19 % higher than F_1 of the spoken vowels for the speakers F1, F2, M1 and M2 respectively. These differences were statistically significant for speaker F1 ($p < 0.05$), but not for the other speakers. Similarly, F_2 of the shouted vowels was on average 7 %, 4 %, 2 % and 6 % higher for the speakers F1, F2, M1 and M2, respectively, but without being significant for any speaker. For F_3, there was no consistent difference between the spoken and shouted vowels across all four speakers.

With regard to H1-H2, we found higher average values for the spoken vowels compared to the shouted vowels for all speakers except M2 (see Table 2), but none of these differences was significant ($p > 0.05$).

The HNR was significantly higher for the spoken vowels than for the shouted vowels for all speakers ($p < 0.001$). Hence, the speakers produced more noise relative to the harmonic components during shouting. This was most probably caused by a greater subglottal pressure during shouting that led to a greater average airflow through the vocal tract and hence greater turbulence.

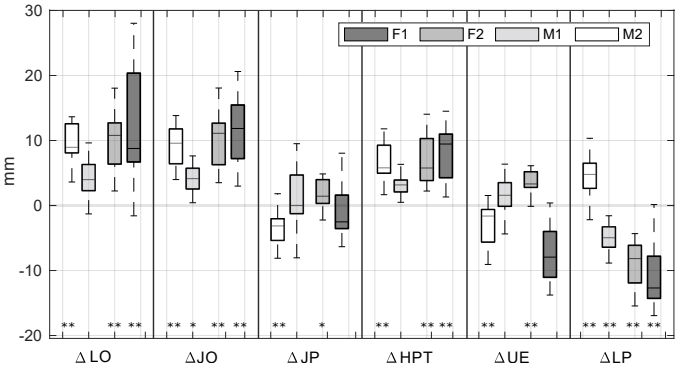

Figure 6: Differences between the articulatory distance measures of the shouted and spoken vowel productions. Values greater than zero indicate that the distance was greater for the shouted vowels ($^{*}p < 0.05$, $^{**}p < 0.01$).

3.2. Articulatory analysis results

The differences of the articulatory distance measures of the vocal tract shapes between shouted and spoken vowels, as obtained from the vocal tract contours, are summarized in Figure 6. Here, ΔLO is for example the lip opening of a shouted vowel minus the lip opening of the corresponding spoken vowel.

For each subject and parameter, a two-tailed t-test was performed to determine the level of significance of differences across the vowels (indicated by the stars in Figures 6 and 7). Among the six parameters, ΔLO, ΔJO, and ΔHPT show the same tendency for all four subjects, i.e., shouted vowels have an increased lip opening (LO), an increased jaw opening (JO), and a lower tongue position (HPT). With regard to jaw protrusion (JP), uvula elevation (UE), and vertical larynx position (LP), the production strategies differ across the four speakers. For shouted vowels, the subjects F2 and M1 had a protruded jaw and a higher uvula position, while the subjects F1 and M2 showed the opposite effects. Except speaker F1, all subjects used a higher larynx position for shouting than for speaking. The different strategy used by speaker F1, i.e., using a lower larynx position for shouting, might be due to her professional experience as a choir singer.

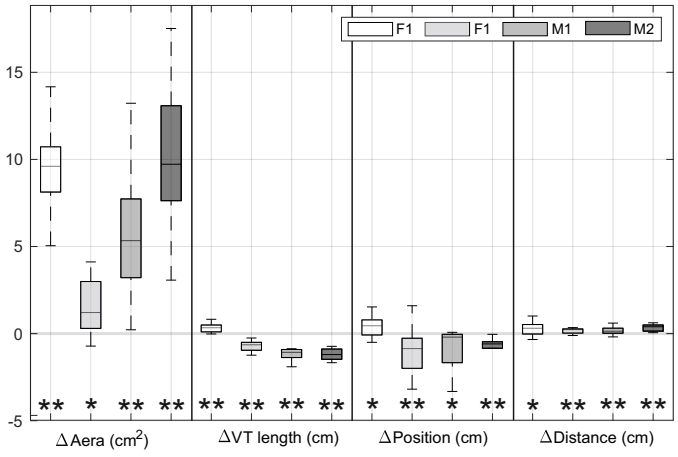

Figure 7: Differences between the measures obtained from the cross-distance functions of the shouted and spoken vowels. For example, ΔPosition is the position of the constriction in the cross-distance function of the shouted vowel minus the constriction position in the cross-distance function of the corresponding spoken vowel ($^{*}p < 0.05$, $^{**}p < 0.01$).

Figure 7 shows the differences between the measures obtained from the cross-distance functions of the shouted and spoken vowels as well as the statistical significance of these differences. Here, a difference value greater than zero means that the measure is greater for the shouted articulation than for the spoken articulation. Most notably, the area under the cross-distance function curves is greater for the shouted than for the spoken vowels (ΔArea > 0) for all subjects, indicating a larger overall volume of the vocal tract for shouted speech. Vocal tract length was significantly shorter for the shouted vowels of all speakers except for speaker F1. The latter is probably caused by the raised (instead of a lowered) larynx position during the shouted vowels for subject M1, as shown in Figure 6. The minimal cross-sectional distances were also significantly greater for the shouted vowels of all subjects. Finally, except subject F1, the speakers produced the main constriction in the vocal tract for the vowels slightly more posterior in the vocal tract for shouted vowels. Given that the overall vocal tract lengths changed by about the same amounts as the constriction positions, this means that shouting mainly affects the length of the posterior vocal tract cavity between the glottis and the constriction. Apart from the general differences between spoken and shouted vowels reported above, there are also vowel-specific differences. These are documented by the vocal tract contours and cross-distance functions contained in the supplemental material for all subjects and vowels.

3.3. Perception test results

The results of the perception experiment are shown in Figure 8. Each boxplot shows the distribution of listener responses to the stimuli for one particular combination of articulatory-acoustic features, with 120 responses per boxplot (8 syllables $\times$ 15 listeners). For example, the leftmost boxplot shows the responses for the items that were synthesized with modal phonation, a subglottal pressure of 1600 Pa, an f_0 that is typical for shouted speech (i.e., 360 Hz for speaker F1 and 270 Hz for speaker M2), and the vocal tract shapes that were measured for shouted vowels. The response "1" means that a corresponding speech item sounded like a spoken utterance, and "4" means that the item sounded like a shouted utterance. The results demonstrate that the speech items were perceived as "most shouted" for the high subglottal pressure, the f_0 typical for shouted speech, and for the vocal tract shapes that were measured for shouted vowels (leftmost two boxplots). In contrast, the speech items synthesized with the low subglottal pressure, the f_0 typical for spoken speech, and the vocal tract shapes that were measured for spoken vowels received the lowest scores and were hence perceived as "most spoken" (rightmost two boxplots). These results are as expected and demonstrate that the differences between spoken and shouted speech can be successfully simulated using the articulatory synthesizer.

To find out which articulatory-acoustic features contributed to what extent to the differentiation of shouting and speaking, we conducted an ANOVA with repeated measures for the four factors vocal tract shape (shouted and normal), f_0 (230 Hz and 360 Hz for subject F1, and 170 Hz and 270 Hz for subject M2), lung pressure (1600 Pa and 800 Pa) and phonation type (modal and pressed) for both reference speakers. For the female reference speaker F1, among the four factors, vocal tract shape ($F(1, 14) = 81.16, p < 0.01$), f_0 ($F(1, 14) = 300.61, p < 0.01$) and lung pressure ($F(1, 14) = 52.29, p < 0.01$) had a significant influence on the responses of the listeners. With regard to the factor phonation type, the responses did not differ significantly ($F(1, 14) = 2.46, p > 0.01$). To explore the contribution that each feature made to explain the observed differences, the effect size was calculated. The factor f_0 had the greatest contribution (partial $\eta^2 = 0.956$), followed by vocal tract shape (partial $\eta^2 = 0.853$) and lung pressure (partial $\eta^2 = 0.789$).

For the male speaker M2, among the four factors, vocal tract shape ($F(1, 14) = 31.57, p < 0.01$), f_0 ($F(1, 14) = 127.04, p < 0.01$) and lung pressure ($F(1, 14) = 57.64, p < 0.01$) had a significant influence on the responses of the listeners. With regard to the factor phonation type, the responses did not differ significantly ($F(1, 14) = 2.78, p > 0.01$). The factor f_0 had the greatest contribution (partial $\eta^2 = 0.901$), followed by lung pressure (partial $\eta^2 = 0.805$) and vocal tract shape (partial $\eta^2 = 0.729$).

A closer look at Figure 8 furthermore indicates a greater range of the response variable, i.e. the perceptual scores for subject F1 than for subject M2. In fact, the standard deviations of the scores across all stimuli variants and conditions are 0.93 for M2 and 1.0 for F1. This means that the synthesis based on the female model was able to generate a somewhat stronger contrast between spoken and shouted utterances than the synthesis based on the data of M2.

The results of the evaluation of the naturalness of the synthetic stimuli are shown in Figure 9. On average, the stimuli were rated as being between "rather unnatural" and "rather natural". The average score was 2.3. However, most importantly, the ratings did not significantly differ with regard to the settings of the articulatory-acoustic features. For example, when all ratings for stimuli created with the "shouted vocal tract shapes" are compared with all ratings for stimuli with "spoken vocal tract shapes", we get no significant difference ($p > 0.05$ based on Student's t-test. The same holds for all other groups of stimuli that differ in one feature (all "pressed voice" stimuli vs. all "modal voice" stimuli, etc.). Bonferroni correction was used to account for the multiple comparisons. This indicates that the synthesis quality was comparable for all feature combinations and should therefore not have biased the results of the spoken-shouted discrimination discussed above.

4. Discussion and conclusions

In this paper, we analyzed multiple articulatory and acoustic parameters of spoken and shouted vowels. With regard to the acoustic measures, shouted speech was found to have a significantly higher f_0, SPL, and F_1 than spoken speech. These findings are well in line with previous studies [12, 15, 14] and can be explained with a higher muscle tension, a higher subglottal pressure, and a lower tongue position during shouting. The increased F_1 of shouted vowels might also be partly explained with a stronger source-filter coupling because of the greater peak glottal area during shouting [42]. The second formant frequency was on average higher for the shouted vowels for all speakers, but without reaching a level of significance. An increased first formant frequency in shouted speech was also reported in [14] and [13]. However, in these studies F_2 and F_3 were found to be *lower* in shouted speech, indicating that these differences are speaker dependent [6].

With regard to vocal tract shape, all subjects had in common that they increased the lip opening, lowered the jaw, and lowered the tongue for shouting. These phenomena are in line with several studies related to loud or Lombard speech [22, 25]. As a result, the vocal tract size (as inferred from the cross-distance functions) was effectively greater for shouted than for spoken vowels.

With respect to the protrusion of the jaw, the height of the uvula/velum, and the vertical larynx position, subjects used different strategies during the production of spoken and shouted vowels. For the subjects F2, M1 and M2, shouted vowels had a higher larynx position, while the subject F1 showed the opposite effect. The subjects F1 and M2 had a lower uvula position and a more protruded jaw for shouted vowels, which is opposite to the subjects F2 and M1. Furthermore, in shouted speech, the overall vocal tract length was generally shorter, and the main constriction of the vowels was wider and slightly more posterior in the vocal tract.

With regard to the differences between the subjects, we assume that for shouting, the subject F1 used strategies learned for loud singing (she was a member of a professional choir), since her articulatory differences between spoken and shouted vowels reflect the morphometric differences that were found between soft and loud singing conditions [27]. The other subjects had no special vocal training. This indicates that professional singers and normal speakers might use different articulatory strategies to produce shouted speech. Consistent with results of [22, 59] and [25] for Lombard and loud speech, the jaw opening and lip opening were positively correlated with sound pressure level.

Finally, we conducted a perception experiment to examine the contribution of four articulatory-acoustic features to the articulatory synthesis of spoken vs. shouted speech. Here we found that an increase of f_0 most strongly shifted the perception of synthesized speech from spoken to shouted. Changing the vocal tract to a shape with a wide open mouth and increasing the subglottal pressure also contributed significantly to the perception of the synthesized speech as being shouted. Phonation type (modal vs. pressed) was not found to have an effect on the perception of speech as being spoken or shouted. With regard to articulatory speech synthesis, these results suggest that, besides changes of subglottal pressure and f_0, different vocal tract target shapes should be used for the individual speech sounds when spoken or shouted speech is supposed to be synthesized.

As a side note on the role of f_0, which was found to be the most essential parameter for shouting, a few previous studies have found that the rise of f_0 should be interpreted as a *passive* result of increasing subglottal pressure and vocal fold tension rather than as an active parameter used to generate loud speech [61, 62]. This phenomenon has also been corroborated by Cooke et al. [16] in speech intelligibility studies by showing that raising of f_0 (which automatically happens in Lombard speech) does not improve speech intelligibility.

Finally, we want to mention the limitations of the current study. Due to the small number of four subjects, it is not clear to what extent the conclusions generalize to bigger populations. To perform studies like this with more subjects, the methods for the analysis of the MRI data should be further automatized. More subjects would also allow to analyze gender differences both in the production of shouted speech and in the perception of synthetic speech as being shouted.

Another potential problem is that the noisy environment in the MRI scanner could have elicited the Lombard effect [60], so that the vocal tract shapes measured for the *spoken* vowels are rather characteristic of Lombard speech. This means that there is a potential mismatch between the articulatory data obtained in the MRI scanner and the audio recordings that were made in a quiet environment. However, if we assume that the MRI noise causes similar articulatory-acoustic changes for both spoken and shouted vowels, the articulatory *differences* between the spoken and shouted vowels might still reflect the corresponding acoustic differences obtained from the recordings in

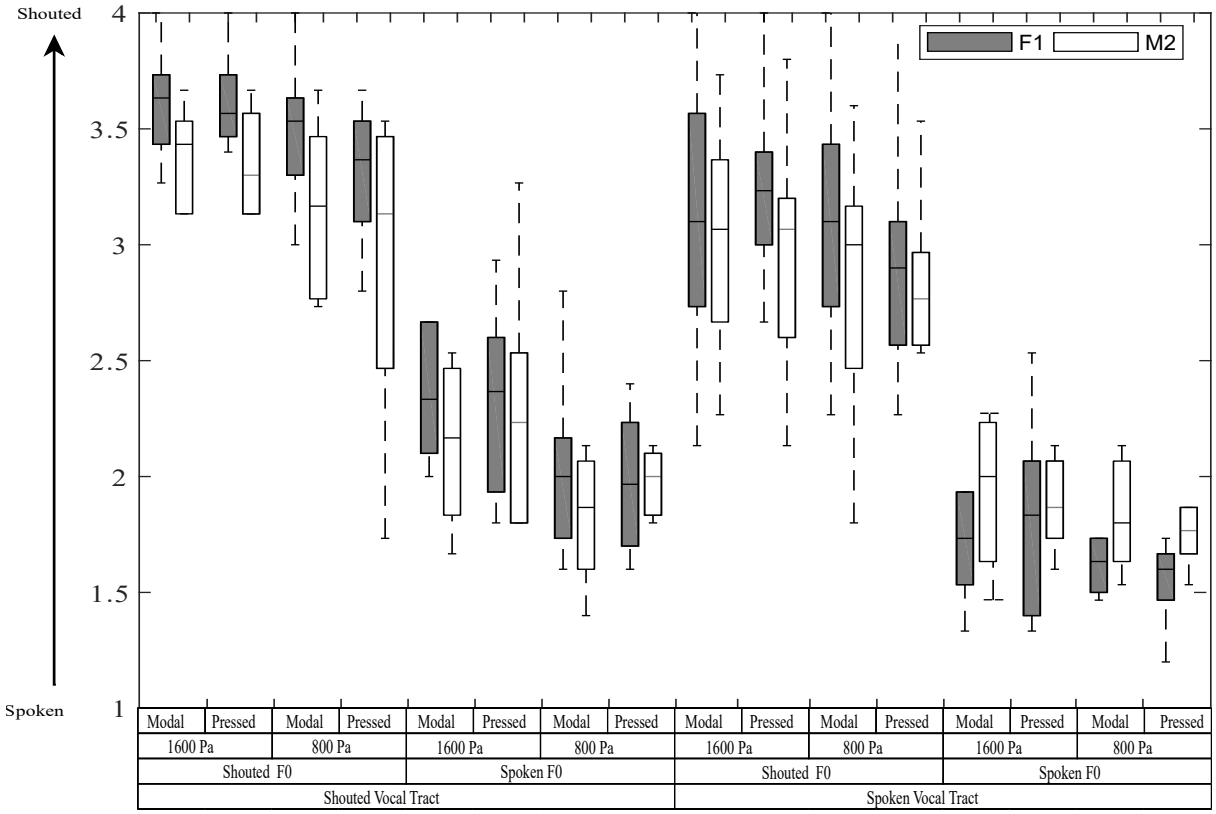

Figure 8: Ratings for the perception of the synthetic stimuli as spoken or shouted for the 16 different combinations of articulatory-acoustic features. Each box plot represents 120 responses (15 listeners × 8 syllables). Gray boxplots refer to the female reference speaker F1, and white boxplots refer to the male reference speaker M1.

the quiet condition.

Finally, the reason that we found no effect of the phonation type on the perception of the stimuli as shouted may be that the voice source settings for the two phonation types did not fully reflect the differences in real voices. As a subsequent analysis of the synthesized stimuli revealed, the HNR of the synthesized "shouted" stimuli was about the same as that of the synthetic "spoken" stimuli, but should have been lower according to the HNR values measured for the natural vowels (Table 2). This means that the noise component of the synthesized shouted stimuli was too low, which might have affected the perception. The realistic synthesis of different phonation types should be explored in more detail in future studies.

5. Acknowledgments

This study was supported by a Grant-in-Aid for Scientific Research (A) (No.25240026) and by JSPS KAKENHI Grant Numbers JP18J11207.

References

[1] C. Zhang and J. H. Hansen, "Analysis and classification of speech mode: whispered through shouted," *Interspeech*, pp: 2289–2292, Antwerp, Belgium, 2007.

[2] T. Ito, K. Takeda and F. Itakura, "Analysis and recognition of whispered speech," *Speech Comunication,* vol. 45, no. 2, pp: 139-152, 2005.

[3] Zelinka, P., Sigmund, M., & Schimmel, J., "Impact of vocal effort variability on automatic speech recognition," *Speech Communication,* vol. 54, no. 6, pp: 732-742, 2012.

[4] I. Shahin, "Speaker identification in the shouted environment using suprasegmental hidden Markov models," *Signal Processing,* vol. 88, no. 11, pp: 2700-2708, 2008.

[5] E. Jokinen, R. Saeidi, T. Kinnunen and P. Alku, "Vocal effort compensation for MFCC feature extraction in a shouted versus normal speaker recognition task," *Computer Speech & Language*, vol.53, pp: 1-11, 2019.

[6] J. C. Junqua, "The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex," *Speech Communication,* vol. 20, no. 1-2, pp: 13-22, 1996.

[7] T. Raitio, A. Suni, M. Vainio and P. Alku, "Analysis of HMM-based Lombard speech synthesis," *Interspeech,* pp: 2781-2884, Florence, Italy, 2011.

[8] M. Garnier, L. Ménard L and B. Alexandre, "Hyper-articulation in Lombard speech: An active communicative strategy to enhance visible speech cues?," *The Journal of the Acoustical Society of America,* vol. 144, no. 2, pp: 1059-1074, 2018.

[9] D. Rostolland, "Intelligibility of shouted voice," *Acta Acustica united with Acustica,* vol. 57, no. 3, pp: 103-121, 1985.

[10] J. M. Pickett, "Effects of vocal force on the intelligibility of speech sounds," *The Journal of the Acoustical Society of America,* vol. 28, no. 5, pp: 902-905, 1956.

[11] Z. S. Bond, and J. Moore. Thomas, "A note on loud and Lombard speech," *ICSLP,* pp: 969-972, Kobe, Japan, 1990.

[12] D. Rostolland, "Acoustic features of shouted voice," *Acta Acustica united with Acustica,* vol. 57, no. 3, pp: 118-125, 1985.

[13] D. Rostolland, "Phonetic structure of shouted voice," *Acta Acustica united with Acustica,* vol. 51, no. 2, pp: 80-89, 1982.

[14] J. Elliott, "Comparing the acoustic properties of normal and shouted speech: a study in forensic phonetics," *Proc. SST-2000: 8th Int. Conf. Speech Sci. & Tech,* pp: 154-159, 2000.

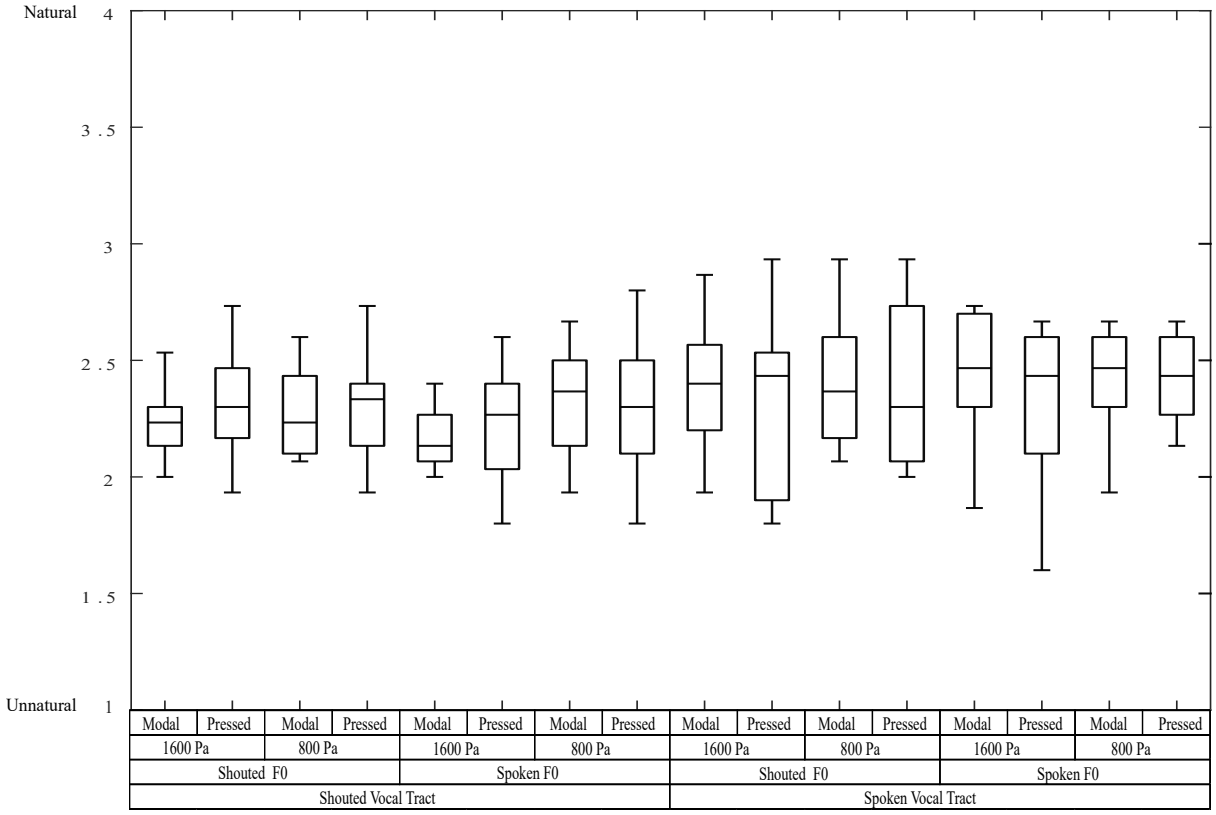

Figure 9: Ratings of the evaluation of naturalness of the stimuli created on the basis of the reference speaker M2.

[15] T. Raitio, A. Suni, J. Pohjalainen, M. Airaksinen, M. Vainio and P. Alku, "Analysis and synthesis of shouted speech," *Interspeech,* pp: 1544-1548, Lyon, France, 2013.

[16] Y. Lu and M. Cooke, "The contribution of changes in F0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Communication,* vol. 51, no. 12, pp: 1253-1262, 2009.

[17] JS. Liénard and MG. Di Benedetto, "Effect of vocal effort on spectral properties of vowels," *The Journal of the Acoustical Society of America,* vol. 106, no. 1, pp: 411-22, 1999.

[18] H. Traunmüller and A. Eriksson, "Acoustic effects of variation in vocal effort by men, women, and children," *The Journal of the Acoustical Society of America,* vol. 107, no. 6, pp: 3438-3451, 2000.

[19] V. K. Mittal and B. Yegnanarayana, "Effect of glottal dynamics in the production of shouted speech," *The Journal of the Acoustical Society of America,* vol. 133, no. 5, pp: 3050-3061, 2013.

[20] V. K. Mittal and B. Yegnanarayana, "Production features for detection of shouted speech," *Consumer Communications and Networking Conference (CCNC),IEEE,* 2013.

[21] M. Garnier, J. Wolfe, N. Henrich and J. Smith, "Interrelationship between vocal effort and vocal tract acoustics: a pilot study," *Interspeech,* pp: 2302-2305. Brisbane, Australia, 2008.

[22] R. Schulman, "Articulatory dynamics of loud and normal speech," *The Journal of the Acoustical Society of America,* vol. 85, no. 1, pp: 295-312, 1989.

[23] A. Geumann, "Vocal intensity: acoustic and articulatory correlates," *In 4th Conference on Motor Control,* Nijmegen, 2001.

[24] A. B. Wohlert and V. L. Hammen, "Lip muscle activity related to speech rate and loudness," *Journal of Speech, Language, and Hearing Research,* vol. 43, 1229-1239, 2000.

[25] J. Šimko, S. Beňuš, and M. Vainio, "Hyperarticulation in Lombard speech: Global coordination of the jaw, lips and the tongue". *The Journal of the Acoustical Society of America,* vol. 139, no. 1, pp: 151-162, 2016.

[26] D. Erickson, "Articulation of extreme formant patterns for emphasized vowels," *Phonetica,* vol. 59, no. 2-3, pp: 134-49, 2002.

[27] M. Echternach, F. Burk, M. Burdumy, L. Traser and B. Richter, "Morphometric differences of vocal tract articulators in different loudness conditions in singing," *PLOS ONE,* vol. 11, no. 4, pp: e0153792, 2016.

[28] J. Pohjalainen, P. Alku and T. Kinnunen, "Shout detection in noise," *Proc. IEEE Intr. Conf. Acoustics, Speech and Signal Processing (ICASSP),* pp: 4968-4971, Prague, Czech Republic, 2011.

[29] A. R. López, R. Saeidi, L. Juvela and P. Alku, "Normal-to-shouted speech spectral mapping for speaker recognition under vocal effort mismatch," *In Acoustics, Speech and Signal Processing, IEEE International Conference (ICASSP),* pp: 4940-4944, New Orleans, LA, USA, 2017.

[30] B. Shikha, K. Banriskhem, M. Prasanna and P. Guha, "Shouted/Normal speech classification using speech -specific features," *Region 10 Conference (TENCON), IEEE*, 2016.

[31] M. Schröder and M. Grice, "Expressing vocal effort in concatenative synthesis,"*Proc. 15th International Conference of Phonetic Science,* pp: 797-800, 2003.

[32] H. Zen, K. Tokuda and A.W. Black, "Statistical parametric speech synthesis," *Speech Communication,* vol. 51, no. 11, pp: 1039-1064, 2009.

[33] B. Bollepalli, L. Juvela, M. Airaksinen, C. Valentini-Botinhao and P. Alku, "Normal-to-Lombard adaptation of speech synthesis using long short-term memory recurrent neural networks," *Speech Communication*, vol. 110, pp.64-75, 2019.

[34] Y. Wang, RJ. Skerry-Ryan, D. Stanton D, Y. Wu, RJ. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q and Le Q, " Tacotron: Towards end-to-end speech synthesis," *arXiv preprint,* arXiv:1703.10135. Mar 29, 2017.

[35] S. Narayanan, K. Nayak, K, S. Lee, A. Sethy, and D. Byrd, "An approach to real-time magnetic resonance imaging for speech production," *The Journal of the Acoustical Society of America,* vol. 115, no. 4, pp: 1771-1776, 2004.

[36] P. Martins, I. Carbone, A. Pinto, A. Silva, A, and A. Teixeira, "European Portuguese MRI based speech production studies," *Speech Communication,* vol. 50, no. 11, pp: 925-952, 2008.

[37] H. M. Hanson, "Glottal characteristics of female speakers: Acoustic correlates," *The Journal of the Acoustical Society of America,* vol. 101, no. 1, pp: 466-481, 1997.

[38] E. Yumoto, W. J. Gould, and T. Baer, "Harmonics - to - noise ratio as an index of the degree of hoarseness," *The Journal of the Acoustical Society of America*, vol. 71, no. 6, pp :1544-1550, 1982.
[39] P. Boersma and D. Weenik, "Praat: doing phonetics by computer," [software]. <http://www.praat.org/>.
[40] Y.-L. Shue, The voice source in speech production: Data, analysis and models. UCLA dissertation, 2010.
[41] Y.-L. Shue, P. Keating , C. Vicenik, K. Yu, "VoiceSauce: A program for voice analysis, Proceedings of the ICPhS XVII", 1846-1849, 2011. <http://www.phonetics.ucla.edu/voicesauce/documentation/index.html.>
[42] P. Birkholz, F. Gabriel, S. Kürbis S and M. Echternach, "How the peak glottal area affects linear predictive coding-based formant estimates of vowels," *The Journal of the Acoustical Society of America,* vol. 146, no. 1, pp:223-32, 2019.
[43] P. Alku, J. Pohjalainen, M. Vainio, A. M. Laukkanen and B. H. Story, "Formant frequency estimation of high-pitched vowels using weighted linear prediction," *The Journal of the Acoustical Society of America,* vol. 134, no. 2, pp:1295-1313, 2013.
[44] P. Birkholz, Image3D [software]. <http://www.vocaltractlab.de/index.php?page=image3d-about>.
[45] M. Echternach, L. Trase and B. Richter, "Vocal tract configurations in tenors' passaggio in different vowel conditions- a real-time magnetic resonance imaging study," *Jounal of Voice,* vol. 28, no. 2, pp: 262.e1-262.e8, 2014.
[46] M. Echternach, P. Birkholz, J. Sundberg, L. Traser, J.G. Korvink and B. Richter, "Resonatory properties in professional tenors singing above the passaggio," *Acta Acustica united with Acustica,* vol. 102, pp: 298-306, 2016.
[47] A. Soquet, V. Lecuit, T. Metens, and D. Demolin, "Mid-sagittal cut to area function transformations: Direct measurements of mid-sagittal distance and area with MRI," *Speech Communication,* vol. 36, pp: 169-180, 2002.
[48] P. Birkholz, VocalTractLab, [software]. <http://www.vocaltractlab.de/index.php?page=vocaltractlab-download>.
[49] P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *Plos One,* vol. 8, no. 4, e60603, 2013.
[50] P. Birkholz, B. J. Kröger and C. Neuschaefer-Rube, "Synthesis of breathy, normal, and pressed phonation using a two-mass model with a modified two-mass model of the vocal fols," *Interspeech,* pp: 2681-2684, Florence, Italy, 2011.
[51] P. Birkholz, B. J. Kröger and C. Neuschaefer-Rube, "Articulatory synthesis of words in six voice qualities using a modified two-mass model of the vocal fold," *First International Workshop on Performative Speech and Singing Synthesis, Vancouver, BC, Canada, 2011.*
[52] P. Birkholz, "Influence of temporal discretization schemes on formant frequencies and bandwidths in time domain simulations of the vocal tract system," *Interspeech,* pp: 1125-1128, Jeju Island, Korea, 2004.
[53] P. Birkholz, "Control of an articulatory speech synthesizer based on dynamic approximation of spatial articulatory targets," *Interspeech,* pp: 2865-2868, Antwerp, Belgium, 2007.
[54] 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.* vol. 19, no.5, pp: 1422-1433, 2010.
[55] C. P. Browman and L. Goldstein, "Articulatory phonology: An overview." *Phonetica,* vol.49, no. 3-4, pp: 155-180, 1992.
[56] P. Birkholz, L. Martin, K. Willmes, B. J. Kröger and C. Neuschaefer-Rube, "The contribution of phonation type to the perception of vocal emotions in German: An articulatory synthesis study," *The Journal of the Acoustical Society of America,* vol. 137, no. 3, pp: 1503-1512, 2015.
[57] K. N. Stevens, Acoustic phonetics. MIT press, 2000.
[58] T. F. Cleveland and R. E. Stone (Jr), J. Sundberg and J. Iwarsson, "Estimated subglottal pressure in six professional country singers," *Journal of Voice*, vol. 11, no. 4, pp: 403-409, 1997.
[59] M. Garnier, L. Bailly, M. Dohen, M, P. Welby, P and H. Lavenbruck, "An acoustic and articulatory study of Lombard speech: Global effects on the utterance," *Interspeech*, pp: 1862-1866, Pittsburgh, PA, USA, 2006.
[60] A. J. Gully, P. Foulkes, P. French, P. Harrison and V. Hughes, "The Lombard effect in MRI noise," *Proc. of the 19th International Congress of Phonetic Sciences*, pp: 800-804, Melbourne, Australia, 2019.
[61] P. Gramming, J. Sundberg, S. Ternström, R. Leanderson and WH. Perkins, "Relationship between changes in voice pitch and loudness," *Journal of Voice,* vol. 2, no.2, pp: 118-26, 1988.
[62] P. Alku, J. Vintturi and E. Vilkman, "Measuring the effect of fundamental frequency raising as a strategy for increasing vocal intensity in soft, normal and loud phonation," *Speech Communication*, vol. 38, no. 3-4, pp: 321-334, 2002.