A Comparative Study of 3D and 1D Acoustic Simulations of the Higher Frequencies of Speech

Rémi Blandin, Simon Stone *Member, IEEE*, Angélique Remacle, Vincent Didone, Peter Birkholz *Member, IEEE*

Abstract—**Articulatory synthesis generates speech sounds by simulating the physical phenomena involved in speech production. The accuracy of the physical modelling is expected to affect the naturalness of the synthesis: the more realistic the description is, the greater the naturalness is expected to be. In this work, the accuracy of acoustic wave propagation in the vocal tract was evaluated with two perceptual experiments. Sustained vowels generated using a one-dimensional acoustic model, a three-dimensional acoustic model and an artificial bandwidth extension algorithm (without a physical basis) were compared. Since the difference between the acoustic methods tested affects mainly the frequencies above 4 kHz, we ensured that the low frequency part of the stimuli, up to 4 kHz, was similar. Thus, the participants' responses were based only on the differences at high frequency. The first experiment was a pair comparison, in which the participants had to select the more natural sounding stimuli. In the second experiment, the participants had to rate the naturalness of the stimuli on a linear scale. The results confirmed that a more accurate physical modeling leads to greater naturalness. However, this was limited to the phonemes /o/ and /u/, for which transverse resonances in the anterior vocal tract may play an important role that only a 3D acoustic simulation can accurately represent. It was also found that male stimuli were perceived as significantly more natural than female ones. However, voice quality did not affect naturalness.**

Index Terms—**Articulatory synthesis, wideband speech, multimodal method**

I. INTRODUCTION

A. General background

Articulatory synthesis is a useful tool for speech research [1] and has great potential for applications requiring natural and expressive speech synthesis. It relies on the description of the physical phenomena involved in speech production [2]–[6]. It simulates sound generation due to vocal fold oscillation and the aeroacoustic sound sources generated by turbulent flow. These sound generation mechanisms interact with sound propagation in the vocal tract. Vocal tract resonances enhance some parts of the radiated spectrum called formants, which convey information such as the phoneme pronounced, speaker characteristics or emotions.

Sound propagation is often simulated using the transmission line model (TLM), which relies on the assumption that only plane waves propagate along the vocal tract [2]. It allows researchers to ignore the curvature and the precise cross-sectional shape: a simple one-dimensional (1D) description of area variation along the vocal tract (area function) is sufficient. Another common simplifying assumption is that sound generation and propagation mechanisms are independent. This is the basis of the source-filter model. In reality, these two mechanisms interact with each other [7]–[9], but this simplification is commonly used as it makes it possible to use simple glottal flow models such as the Liljencrants-Fant (LF) model [10].

B. Limits of the TLM and potential impacts on the naturalness

The TLM is inaccurate above about 4 kHz (sometimes even 3 kHz) because, on one hand, it cannot take into account the precise three-dimensional (3D) vocal-tract shape, and on the other hand, it cannot describe the 3D aspects of the acoustic field such as transverse resonances or the curvature at areas where discontinuities are found. This leads to inaccurate resonance frequencies, amplitudes and bandwidths. These inaccuracies increase toward high frequencies (HF), amounting to about 5% for the frequencies of the first four resonances [11]. From 4 or 5 kHz on, the transverse resonances induce additional peaks and troughs in the transfer function (TF) which cannot be predicted by TLM [12].

More accurate acoustic models can account for the 3D aspect of the acoustic field. Such models are based on finite elements [13]–[15], finite differences [16], the multimodal method [17] and 3D waveguide meshes [18]. However, these methods have a much higher computational cost than TLM, which explains why their application to speech synthesis has been limited to isolated phonemes such as vowels, diphthongs and consonants.

Since 3D acoustic models are more realistic than TLM, they are expected to generate synthetic speech with greater naturalness. The greatest impact is expected for HF above 4 kHz, where the difference between 3D acoustic models and TLM is greatest. The lower frequency differences, consisting mainly of small formant frequency deviations, are not expected to have a significant impact on naturalness. The accuracy of acoustic modeling has been examined by

Rémi Blandin is with the Institute of Acoustics and Speech Communication, TU Dresden, Dresden 01062, Germany (e-mail: remi.blandin@tu-dresden.de)

Simon Stone was with the Institute of Acoustics and Speech Communication, TU Dresden, Dresden 01062, Germany (e-mail: contact@simon-stone.com)

Angélique Remacle is with the Research Unit for a Life-Course Perspective on Health and Education, Faculty of Psychology, Speech and Language Therapy, and Educational Sciences, University of Liège, Belgium and the Center For Research in Cognition and Neurosciences, Faculty of Psychological Science and Education, Université Libre de Bruxelles, Brussels, Belgium (email: angelique.remacle@uliege.be)

Vincent Didone is with the Psychology and Neuroscience of Cognition Research Unit (PsyNCog), Quantitative psychology, University of Liège, Liège, Belgium (email: vdidone@uliege.be)

Peter Birkholz is with the Institute of Acoustics and Speech Communication, TU Dresden, Dresden 01062, Germany (e-mail: peter.birkholz@tudresden. de)

very few investigations so far, and none have specifically targeted HF. To our knowledge, the only work investigating this question is Gully's PhD thesis [19]. Gully compared synthetic diphthongs generated with TLM, a two-dimensional (2D) and a 3D waveguide mesh methods using a perceptual rating of naturalness. She found that the 3D stimuli were perceived as significantly more natural than the TLM and 2D waveguide mesh stimuli. However, she reported that her method had limitations regarding the modelling of losses at HF because of the use of time-domain simulations. The version of the TLM used may also not have been optimized for naturalness, in contrast to articulatory synthesizers such as VocalTractLab [4]. Krug et al. [20] showed that state-of-the-art articulatory synthesis achieves a naturalness comparable to other non-commercial synthesis systems, but is significantly outperformed by commercial synthesis systems. Therefore, the acoustic model's accuracy is explored as a potential factor to further improve the naturalness of articulatory synthesis.

From 4 or 5 kHz on, 3D models differ substantially from the TLM. HF (above 4 kHz) have been shown to be important for naturalness [21]–[24]. Therefore, the accuracy of the acoustic modeling of HF can have a strong impact on naturalness. However, hearing sensitivity and the capacity to discriminate frequencies reduce toward HF. Thus, it is not clear to what extent the accuracy of acoustic modeling matters in this frequency range, or even if an acoustic model is needed at all to make stimuli sound natural. In fact, speech HF can be generated from narrow-band speech signals such as telephone signals using Artificial Bandwidth Extension (ABE) without any physical model as a basis [25], [26]. These methods are based on linear prediction and/or training of a deep neural network on wide-band speech recordings.

The accuracy of acoustic modeling affects some phonemes more than others. In [11], for example, more differences between 1D TLM and a 3D multimodal method (MM) were observed for /u/ than for /a/.

Since the differences due to acoustic modeling are expected to affect mainly HF, the HF content of the sound source may influence how they are perceived. Because a pressed voice has more HF than a modal voice, more differences may be perceived with pressed voice. Similarly, since a female voice contains more HF than a male voice [27], more differences may be perceived with a female voice.

C. Objective and outline

To investigate how the degree of physical accuracy of the generation of HF affects the naturalness of synthesized vowels, we generated synthetic vowels using ABE method, TLM and 3D simulations. The ABE was used as a reference alternative for HF generation without physical insight. Since the objective was to study HF, we made sure that the stimuli had the same frequency content up to 4 kHz. The pressed voice was used in addition to the normal voice because its stronger HF content could better highlight the potential differences between the different HF generation methods. The perceived naturalness of the different synthesis methods was compared using a pair comparison paradigm and evaluated using a metric scale.

Parameter	Modal		Pressed	
	female	male	female	male
Amplitude	300	300	300	300
Open quotient	0.84	0.84	0.78	0.7
Shape quotient	1.9	2.15	1.99	2.18
Spectral tilt	0.04	0.04	0.01	0.01
Noise level (dB)	-55	-65	-45	-50

TABLE I

PARAMETERS OF THE LF MODEL USED TO CREATE THE EXCITATION SIGNALS. THE OPEN AND SHAPE QUOTIENTS WERE TAKEN FROM [31]. THE SPECTRAL TILT AND NOISE LEVEL WERE DETERMINED MANUALLY TO CREATE ACCEPTABLY DISTINCT AND SUFFICIENTLY NATURAL SOUNDING MODAL AND PRESSED VOICE QUALITIES.

The generation of stimuli and the perceptual experiments are explained in the section II. The results are presented in section III and discussed in section IV.

II. METHODS

A. Stimulus generation

To investigate the research questions posed above, we needed a set of stimuli that contained a systematic variation of the factors of interest in this study (voice quality, gender, acoustic simulation method). To that end, we created excitation signals and appropriate vocal tract transfer functions based on both the MM and the TLM. By convolving the excitation signals with the transfer functions, we obtained the final synthetic speech samples. For reference, additional stimuli were produced using ABE on narrow-band versions of the MM stimuli. The Matlab script for the creation of the stimuli and all the necessary data are provided in the supplementary materials.[1]

1) Excitation: We created a male and a female excitation signal, each with a modal and a pressed voice quality. The excitation signals were created using the LF model [28], [29] to generate glottal flow pulses. To mimic the aspiration noise, these pulses were then superimposed with Gaussian noise, with a spectral slope of $-9.4\,\mathrm{dB\,kHz^{-1}}$ [30], obtained using a finite impulse response filter, and temporally gated by multiplying by the flow pulses amplitude. Table I summarizes the parameters used in the generation of the flow pulses. The f_0 contours and the temporal envelope of the flow signals were based on a natural reference utterance (male German native speaker, 36 years, no discernible accent) of the phoneme /a/. For the male excitation signal, the reference f_0 contour was used directly. For the female version, the f_0 was shifted up by one octave. The signals were calculated at four times the intended sampling rate and then downsampled to 44.1 kHz to avoid aliasing effects. All excitation signals were padded with 250 ms of silence on both ends.

2) Geometries: All transfer functions used in this study were based on the standard VocalTractLab version 2.3 (VTL) shapes for the vowels /a, e, i, o, u/. The speaker models in VocalTractLab are detailed 3D models of real speakers' vocal tracts based on MRI data [4]. These vowels were chosen

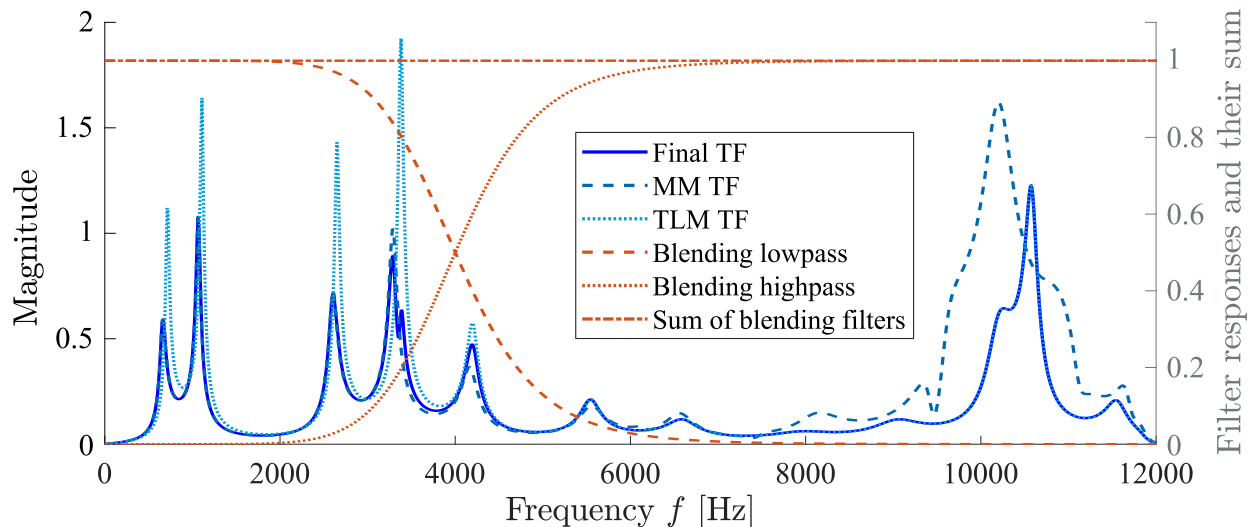[1] https://www.vocaltractlab.de/index.php?page=birkholz-supplements

Fig. 1. Transfer function blending (here: the male /a/). The blended TF is the sum of the lowpass-filtered MM TF and the highpass-filtered TLM TF. The responses of the lowpass and highpass filters add up to 1 (unity-gain crossover filter).

because they exist in many languages and their quality is not very different between French (the language of the study participants) and German (the native language of the speaker on whose vocal tract the shapes are based [4], [32]). Since the MM does not consider side branches, the piriform sinus was excluded from the transfer function calculation. This led to a shift in the formants of the standard vocal tract shapes. To offset this shift, the 3D vocal tract geometries were manually tweaked: the positions of the articulators were slightly moved so that the first two resonance frequencies matched those of the original geometries including the piriform sinus.

3) Transfer functions: For each vocal tract shape, two transfer functions were calculated: one using the MM [17] and one using the TLM [33], [34] both implemented in VocalTractLab3D.[2] In addition to the obviously large HF differences between the calculation methods, the transfer functions also had slightly different formants in the low-frequency range. At low frequencies (up to 4 or 5 kHz), small changes in formant frequencies (2%-10%) are generally perceivable, independently of the kind of simulation [35], [36]. Since the objective was to evaluate differences in the HF range, it was necessary to somehow ensure that the low-frequency part of the TFs was identical across all conditions. Therefore, we used a unity-gain crossover filter with a single crossover frequency of 4 kHz and crossover slopes of 48 dB oct^{-1} to blend the lower-frequency part of the MM TF and the high-frequency part of the TLM TF. Figure 1 shows an example of the process. The final transfer functions used for stimulus generation were therefore the MM TFs and the blended MM-TLM TFs.

4) Audio synthesis: The sounds were synthesized by convolving the excitation signals with the impulse response corresponding to the transfer functions and then shifting the result by half the impulse response length to correct the introduced lag. Before calculating the impulse responses, the transfer functions were lowpass-filtered with a cutoff frequency of 12 kHz, because the audiometric device to screen the par-

ticipants in the perception experiment was only validated up to 12.5 kHz. All sounds were loudness-normalized to -23 Loudness Unit Full Scale (LUFS) according to the EBU R 128 standard and saved as a mono, 16-bit pulse code modulation (PCM) wave file with a sampling rate of 44.1 kHz.

5) Artificial bandwidth extension: ABE is used to extend narrow-band speech signals (e.g., telephone speech) to a wide-band signal. These methods do not use an acoustic model to calculate the missing frequency components, but instead extrapolate from the narrow-band spectrum. This extrapolation can be based on (among other methods) linear prediction (i.e., using the narrow-band linear prediction coefficients to find the best-matching entry in a codebook of full-band linear prediction coefficients) [26], or on deep learning and pre-trained predictive modeling of the missing frequencies [25]. The ABE stimuli were created by first band-limiting the speech signals created with the MM TFs to 4 kHz and then extending this narrow-band signal first to 8 kHz [25] and then to 16 kHz [26]. The parameters for the ABE algorithms were adopted from the examples provided with the implementation and are listed in Table II. They were not optimized in any way because, as mentioned above, the study's focus was not on evaluating the quality of ABE. These samples were simply included as an alternative, signal-processing-driven way of generating the HF part of the speech signal with less physical correctness than with TLM. Some minor, audible artifacts in the otherwise silent sections of the stimuli were removed by forcing these sections to zero using a Tukey window (with $\alpha = 0.1$). Finally, the ABE stimuli were upsampled to 44.1 kHz (to match the sampling rate of the other stimuli). As with the synthesized stimuli, the ABE stimuli were also low-pass filtered at 12 kHz, loudness-normalized to -23 LUFS and saved as a mono, 16-bit PCM wave file.

B. Perceptual experiment

1) Objectives and hypotheses: This perceptual experiment was carried out to answer the following general research

[2]https://vocaltractlab.de/index.php?page=vocaltractlab-download

ABE from 4 kHz to 8 kHz [25]		ABE from 8 kHz to 16 kHz [26]	
Past frames:	1	LP order:	16
Future frames:	1	FFT length:	1024
Input features:	logMFE, PCA	Window length:	25 ms
Input dimension:	10	Gain:	1
Output dimension:	10		
Model:	pretrained		

TABLE II
PARAMETERS USED FOR THE ABE ALGORITHMS (SEE [25], [26] FOR DETAILS).

question: How does the degree of physical accuracy of the generation of HF affect the naturalness of synthesized vowels perceived by young adults? The high-frequency part of the synthesized vowels was generated using three methods that produce different degrees of physical accuracy:

(a) a 1D TLM acoustic model,
(b) a 3D MM acoustic model, and
(c) an ABE algorithm.

Sixty audio stimuli (3 synthesis methods × 5 vowels (/a, e, i, o, u/) × 2 genders (male, female) × 2 voice qualities (modal, pressed)) were synthesized. The experiment comprised two tasks designed to assess the naturalness of the synthesized vowels:

(a) a paired comparison paradigm, and
(b) an evaluation of each vowel using a metric scale ranging from 0 to 100.

We hypothesized that, the more realistic the synthesis method is in terms of acoustical modeling, the more natural the vowels will be perceived as. More specifically,

(H1) the vowels generated with MM should be perceived as more natural than those generated with TLM;
(H2) the vowels generated with MM should be perceived as more natural than those generated with ABE;
(H3) the vowels generated with TLM should be perceived as more natural than those generated with ABE.

In addition to the effect of the synthesis method, we studied the effect of vowel, gender and voice quality.

2) Participants: The sample included 40 students (20 males, 20 females) recruited at the University of Liège. They were all native speakers of French, aged 19 to 24 years old (M = 21.9; SD = 1.5), and did not have any past or present hearing problems. All participants had hearing thresholds ≤ 20 dB HL bilaterally at octave frequencies between 0.5 kHz and 12.5 kHz. This was assessed with pure-tone audiometry using a MADSEN Itera II audiometer with Sennheiser HDA 300 headphones.

3) Listening condition: The perceptual experiment was conducted in an audiometry room in the University of Liège. During the experiment, the participants were seated in front of a table on which a computer allowed them to perform the perceptual tasks through interfaces programmed with Octave. The background noise in the experimental conditions (a computer and 2 persons present in the room) was measured at 36 dB SPL. The stimuli were played with a loudspeaker (XM6.D sn 07-3301 from FAR by ATD) placed 1 m in front of the participant's head. A loudspeaker offers better control over experimental conditions than headphones and better simulates a real speaker as the sound is radiated from a distant sound source. The level of the stimuli was adjusted to 70 dB SPL at the location of the participant's head. In order to control the stimuli presented to the participants in the experimental condition, the stimuli were recorded with a measurement microphone (G.R.A.S. 40CE, SN 217653) placed at the location of the participant's head. These recordings are provided in the supplementary materials. The participants were instructed to stay in the same position as much as possible during the experiment, and in particular to try to keep their head in the same position with respect to the loudspeaker to minimize potential disturbances due to directivity effects.

Figure 2 presents the spectra of the stimuli generated using the modal voice. As expected, the spectra of the MM and TLM stimuli are identical up to about 4 kHz. Above 4 kHz, they differ to various degrees depending on the phoneme. The MM and TLM HF spectra are almost identical for /e/, and significantly different for /o/ and /u/. The ABE stimuli start to differ from the other stimuli at a slightly lower frequency (from about 3.3 kHz). One can see a large overall difference between MM and TLM at HF, in particular above about 7 kHz, for the vowels /a/, /e/ and /i/. The differences are smaller for /o/ and /u/. The stimuli in a female voice tend to have more HF, as expected. However, this tendency is not observed with ABE. Except in a few specific frequency ranges, most of the energy of the stimuli is above the background noise. The same observations can be made with the pressed voiced quality (not shown): there is more energy at HF (see supplementary materials).

4) Preliminary tasks: The ethics committee of the Faculty of Psychology, Speech Therapy and Education Sciences (University of Liège, Belgium) approved this study. All participants provided written informed consent after receiving a complete description of the study.

The participants were tested individually in a single session lasting approximately 1 hour. After completion of the questionnaire and the audiometric screening, each participant performed two listening tasks. Participants were given a break between the two tasks to avoid fatigue.

Each task began with two practice trials using different phonemes from the tasks. These practice trials were later discarded from the statistical analysis. Within each task, audio stimuli were presented randomly across participants. To evaluate intra-judge reliability, each vowel was presented twice at different, randomly set times, referred to as test and retest. The participants' responses and the number of times they listened to each stimulus were saved in .csv format for further statistical analysis.

5) First perceptual experiment: This experiment was based on a paired comparison paradigm. The participants listened to each of the two stimuli as often as they wanted, by clicking on the corresponding button. For each pair of stimuli presented, participants had to select the sound they considered the most natural. Each pair was composed of the same vowel, the same gender, and the same voice quality (e.g., vowel [a] synthesized with a female pressed voice). Only the synthesis method
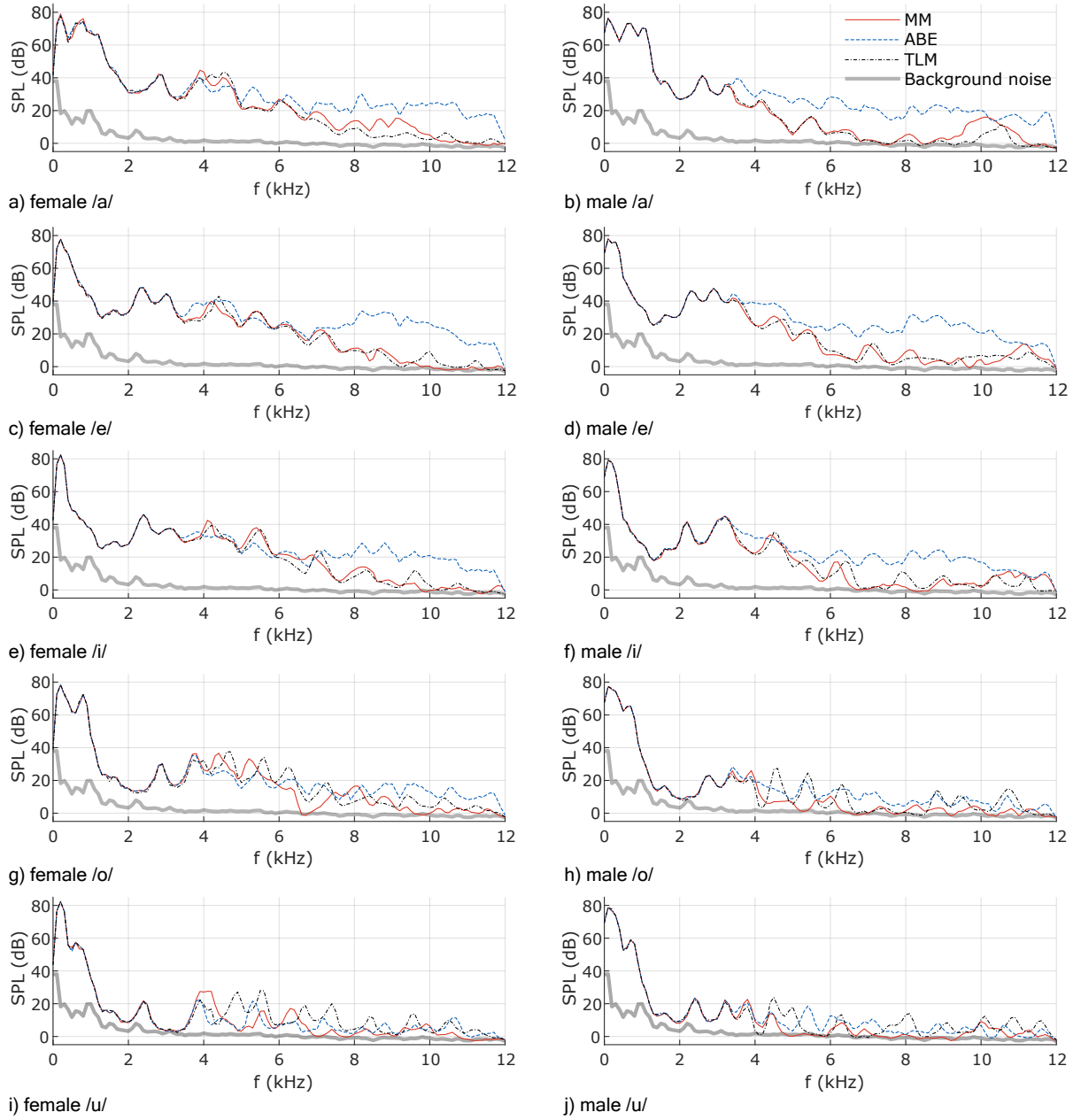
Fig. 2. Long-term average spectra of the stimuli generated using the modal voice recorded in the experimental conditions. A Hanning window of a length of 512 samples (sample rate of 51,200 Hz) and an overlapping rate of 50% was used.

changed (i.e., ABE, TLM, MM). Based on this principle, the following combinations of pairs were presented to participants to test the three hypotheses set out above:

(H1) Two stimuli (same vowel, same gender, same voice quality) generated with the MM and TLM methods (e.g., Sound 1: [a] synthesized with a female pressed voice using the MM method; Sound 2: [a] synthesized with a female pressed voice using the TLM method).

(H2) Two stimuli (same vowel, same gender, same voice quality) generated with the MM and ABE methods (e.g., Sound 1: [e] synthesized with a female pressed voice using the MM method; Sound 2: [e] synthesized with a female pressed voice using the ABE method).

(H3) Two stimuli (same vowel, same gender, same voice quality) generated with the TLM and ABE methods (e.g., Sound 1: [i] synthesized with a male modal voice using the TLM method; Sound 2: [i] synthesized with a male modal voice using the ABE method).

Each pair of stimuli was presented twice (test and retest). Experiment 1 included a total of 120 pairs to evaluate (3 combinations × 5 vowels × 2 genders × 2 voice qualities × 2 times).

The following instructions were given to participants orally (in French) before the experiment: *"You're going to listen to two sounds and then you must indicate which one seems more natural to you. To listen to the sounds, click on 'sound 1' or*

'sound 2.' Then choose the sound that seems more natural to you by clicking on *'sound 1 is more natural'* or *'sound 2 is more natural.'* To move on to the next pair, click on *'next pair.'* However, once you've clicked on *'next pair,'* you will no longer be able to go back. First, we're going to do a trial run with two pairs of sounds. During the experiment, you'll have 120 pairs of sounds to compare. After the trial run, don't hesitate to ask any questions you may have."

6) Second perceptual experiment: Participants had to assess the naturalness of 120 audio stimuli (3 synthesis methods × 5 vowels × 2 genders × 2 voice qualities × 2 times) presented in isolation. The participants listened to each stimulus as often as they wanted. Then they evaluated its naturalness using a metric scale ranging from 0 (not natural at all) to 100 (totally natural).

The following instructions were given to participants orally before the experiment: *"You're going to listen to 120 sounds, one after the other. To listen to a sound, click on the 'listen' button. For each sound, indicate its naturalness by moving the cursor on the scale ranging from 0 'not natural at all' to 100 'totally natural.' 'Not natural at all' means that the sound you're listening to resembles an artificial voice, very different from a real voice. 'Totally natural' means that the sound you're listening to resembles a real voice. Once you've clicked on the 'next' button, you will no longer be able to go back. We will do a trial run with two sounds, and after that you can ask any questions you may have."*

C. Statistical analysis

The participants' responses in the paired comparisons experiment were fitted using loglinear Bradley-Terry models (LLBT) [37]–[41]. Conceptualized by Bradley and Terry [42], LLBT was specifically developed to analyze paired comparisons designs and has been commonly used in several research domains [43]–[45]. LLBT provides preferences values (worth parameters) and the associated estimated probability of being preferred for all items. Thus, the worth parameters π of each stimulus were estimated quantifying the relative positions of the stimulus on a standardized latent scale from 0 to 1. A higher worth value is indicative of that stimulus's greater preference relative to another stimulus (the sum of all values is 1.0). All analyses were performed using the software R 4.2.0 [46], with the prefmod [47] and gnm [48] R packages. Due to the experimental design, all factors (phonemes, gender, voice quality) were analyzed separately.

The participants' responses to the evaluation of the naturalness of the stimuli were fitted using a linear mixed-effects model (LME). As recommended by many authors [49]–[51], we preferred to perform mixed-effects models instead of repeated measures analyses of variance in the case of repeated-measures data. LME have numerous advantages. For example, LME can properly account for correlation between repeated measurements on the same subject; they provide flexibility and can model individual characteristics. For the specification of the model, a random effect of participant was used and the fixed effects were the acoustic model (ABE, TLM or MM), the type of phoneme (/a, e, i, o, u/), the speaker's gender (female
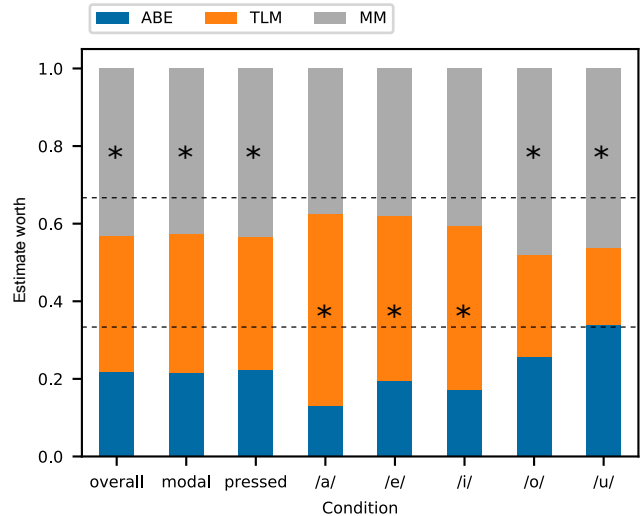

Fig. 3. Stacked estimate worth of the different methods in the pairwise comparison (experiment 1) for different conditions. The estimate worth is similar to the probability of choice, and therefore, the sum of the estimate worth of all the methods is 1. The larger the estimate worth, the more likely it is that the corresponding method will be selected. The method that dominates the preference score is indicated by an * symbol. The values 0.33 and 0.66 are indicated by dashed lines.

or male), the voice quality (modal or pressed) and the time of the test (two times: test and retest). The model included each main factor, and all the interactions between the factors. Contrasts were performed between the levels of the factors and interactions that were significant in the analysis of the models. The Holm method of alpha adjustment was used to correct for multiple testing. These analyses were performed using the R packages car [52], tidyverse [53], lme4 [54], lmerTest [55], emmeans [56] and multcomp [57].

III. RESULTS

A. Pair comparison (experiment 1)

This section describes the results obtained with the pair comparison. Figure 3 shows the estimate worth of the different methods. The estimate worth is similar to the probability of choice of one of the methods.

One can see that the hypotheses are generally verified, since overall the stimuli synthesized with MM are most preferred, whereas the stimuli generated with ABE are overall least preferred. This effect is significant for the three types of pairs tested:

- H1 (MM more natural than TLM)
  $z = 5.02$, $p < 0.0001$
- H2 (TLM more natural than ABE)
  $z = 11.15$, $p < 0.0001$
- H3 (MM more natural than ABE)
  $z = 15.89$, $p < 0.0001$

When the analysis is restricted to the modal and pressed voices, we find very similar distributions of estimate worth, and the effect remains significant. For the modal voice:

- H1, $z = 3.04$, $p = 0.0024$
- H2, $z = 8.54$, $p < 0.0001$
- H3, $z = 11.40$, $p < 0.0001$

	/a/		/e/		/i/		/o/		/u/	
Hypothesis	z-value	p-value	z-value	p-value	z-value	p-value	z-value	p-value	z-value	p-value
H1 (MM > TLM)	**-2.869**	**0.0041**	-1.117	0.26	-0.4699	0.64	**6.433**	**< 0.0001**	**8.714**	**< 0.0001**
H2 (TLM > ABE)	**12.42**	**< 0.0001**	**7.958**	**< 0.0001**	**9.094**	**< 0.0001**	0.1854	0.85	**-5.636**	**< 0.0001**
H3 (MM > ABE)	**10.13**	**< 0.0001**	**6.938**	**< 0.0001**	**8.683**	**< 0.0001**	**6.606**	**< 0.0001**	**3.337**	**0.0008**

TABLE III

Z-SCORE AND p-VALUE OF THE HYPOTHESIS TESTED FOR EACH PHONEME; SIGNIFICANT VALUES ARE HIGHLIGHTED IN BOLD.

Fig. 4. Average naturalness rating per method and phoneme; the error bars indicate the standard error of measurement.
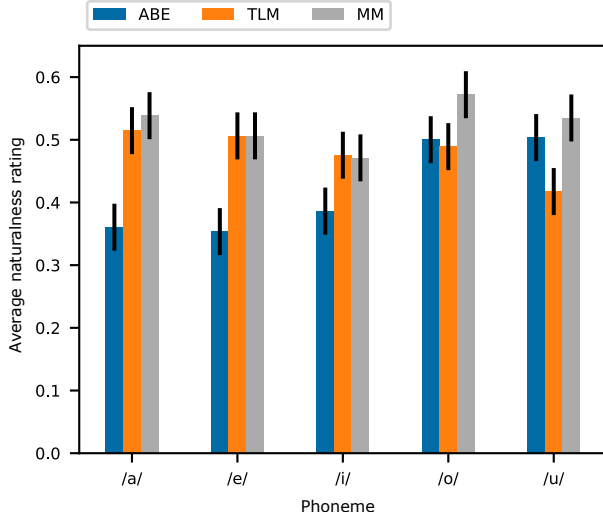

Fig. 5. Overall average naturalness rating per gender and phoneme; the error bars indicate the standard error of measurement.

And for the pressed voice:

- H1, $z = 4.05$, $p < 0.0001$
- H2, $z = 7.23$, $p < 0.0001$
- H3, $z = 11.08$, $p < 0.0001$

Note that there is a slightly greater difference between MM and TLM with the pressed voice.

When the analysis is done for each specific phoneme, we find different distributions of the preferences, which do not systematically verify the hypotheses. The z-score and the p-values obtained for each method and each phoneme are presented in Table III. For /a/, TLM is considered more natural than MM, contrary to expectations (H1). There are no significant differences between TLM and MM for /e/ and /i/, so hypothesis H1 cannot be verified or contradicted. For /o/, there is no significant difference between TLM and ABE, so H2 (TLM more natural than ABE) cannot be verified or contradicted. H2 is contradicted for /u/, as ABE is considered more natural than TLM.

B. Naturalness rating (experiment 2)

The average naturalness ratings normalized between 0 and 1 are presented in Figs. 5 and 4 for different categories.

The time of the test (test or retest) did not significantly impact the naturalness rating ($F(1, 4641) = 0.0947$ $p = 0.758$).

On the other hand, the method significantly impacted the naturalness rating ($\chi^2 = 52.203$, p < 0.0001). Similarly to the pair comparison experiment, hypotheses H1, H2 and H3 are generally verified since the average naturalness rating for MM is higher than the rating for TLM, which itself is higher
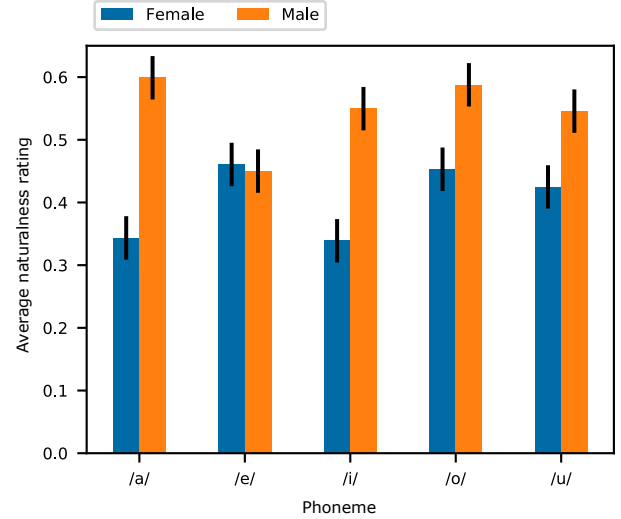
than the rating for ABE. These differences between methods are significant, as illustrated by the linear contrasts:

- H1, $z = 5.67$, $p < 0.0001$
- H2, $z = 7.69$, $p < 0.0001$
- H3, $z = 13.36$, $p < 0.0001$

Gender also significantly impacted the naturalness rating ($\chi^2 = 37.60$, $p < 0.0001$): on average, the male stimuli were rated higher than the female ones (see Fig. 5). This is confirmed by a significant linear contrast between the male and female stimuli ($z = 22.43$, $p < 0.0001$). However, there was no significant interaction between gender and method ($\chi^2 = 5.85$, $p = 0.054$). There was also no significant interaction between the listener's gender and the gender of the stimuli ($\chi^2 = 0.683$, $p = 0.409$). More generally, no significant effect was associated with the listener's gender.

Voice quality had no significant impact on the naturalness rating ($\chi^2 = 0.72$, $p = 0.39$). However, listener had a significant impact ($\chi^2 = 584.99$, $p < 0.0001$). This corresponds mainly to variations in the amount of differences between the naturalness rating of the different categories: for example, the difference between ABE and TLM can be small for some subjects and bigger for others. In some cases, the ordering of the values differs: for example, some subjects rate ABE higher than TLM.

Phoneme significantly affected the naturalness rating ($\chi^2 = 27.425$, $p < 0.0001$). The average naturalness ratings of the different phonemes for each method are presented in Fig. 4. /o/ has the highest average naturalness rating (0.52), whereas /i/ has the lowest (0.44). The linear contrasts highlight significant

	/a/		/e/		/i/		/o/		/u/	
	z-value	p-value	z-value	p-value	z-value	p-value	z-value	p-value	z-value	p-value
H1 (MM > TLM)	1.38	1.00	0.0032	1.00	-0.24	1.00	**4.77**	**0.00012**	**6.76**	**< 0.0001**
H2 (TLM > ABE)	**8.87**	**< 0.0001**	**8.79**	**< 0.0001**	**5.13**	**< 0.0001**	-0.63	1	**-4.96**	**< 0.0001**
H3 (MM > ABE)	**10.2**	**< 0.0001**	**8.80**	**< 0.0001**	**4.89**	**< 0.0001**	**4.14**	**0.0022**	1.8	1.00

TABLE IV
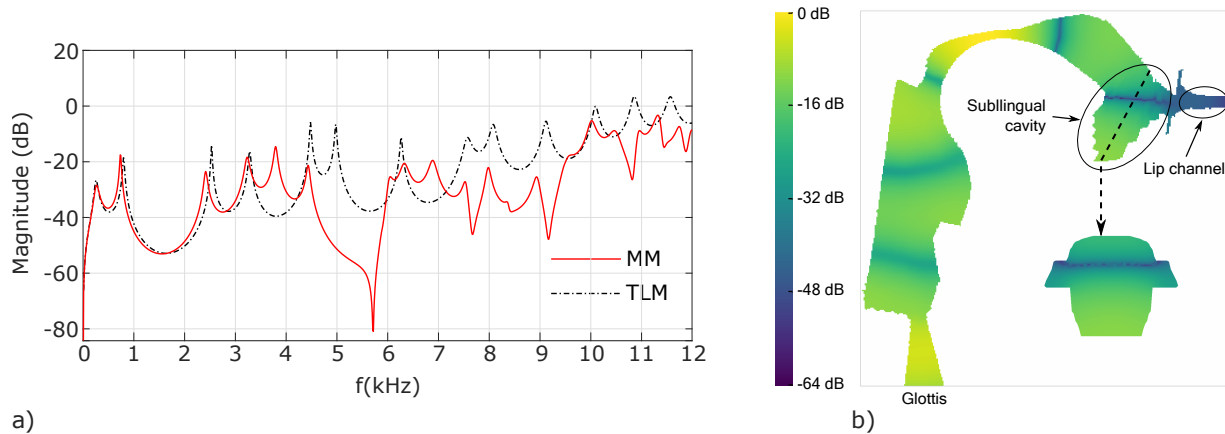LINEAR CONTRASTS BETWEEN THE METHODS FOR EACH VOWELS; SIGNIFICANT VALUES ARE HIGHLIGHTED IN BOLD.

Fig. 6. (a) Transfer functions of the male vowel /u/ computed with MM and TLM; (b) acoustic field computed with VocalTractLab3D in the mid-sagittal plane and a transverse plane in the middle of the sublingual cavity at the frequency of the first transverse resonance of the sublingual cavity corresponding to the deep trough at 5.71 kHz in the MM TF.

differences between most of the phonemes, except between /u/ and /a/, /e/ and /i/ and /e/ and /a/.

There is a significant interaction between phoneme and gender ($\chi^2 = 33.795$, $p < 0.0001$). This is illustrated in Fig. 5. The average naturalness rating is distributed differently for the male, female and overall stimuli. For male stimuli, the highest naturalness rating is obtained for /a/ (0.60), and the lowest for /e/ (0.45). In contrast, for female stimuli, the highest naturalness rating is obtained for /e/ (0.46) and the lowest for /i/ (0.34). However, the male stimuli are rated significantly higher than the female ones, except for /e/ for which no significant differences are observed ($z = 0.75$, $p = 1$).

There is also a significant interaction between phoneme and method ($\chi^2 = 51.740$, $p < 0.0001$). This is illustrated in Fig. 4. The linear contrasts between methods for each phoneme are presented in Table IV. The results are similar to those for the pair comparison experiment, except for

- /a/, for which no significant difference is found between TLM and MM;
- and /u/, for which no significant difference is found between MM and ABE.

In addition, hypothesis H2 (TLM more natural than ABE), is also contradicted for /u/ in this experiment.

IV. DISCUSSION

The hypothesis that more realistic acoustic modeling of HF results in greater naturalness is generally verified by both experiments. This is in agreement with the results of Gully [19], and the fact that our results were obtained using completely different methods and better modeling of HF losses reinforces this conclusion.

However, a closer look at the individual phonemes reveals that, in the comparison of TLM and MM, the better naturalness of MM is due only to the vowels /o/ and /u/. This may be related to the larger spectral differences observed for these phonemes (see Section II-B3 and Fig. 2). The HF amplitude is generally higher for TLM. In particular, peaks are present in the TLM spectra in the 4 to 6 kHz interval and absent or much smaller in the MM spectra. This might be related to a greater difference between the acoustic fields computed with a 1D simplifying assumption and a more realistic 3D method. In fact, transverse resonances occur at lower frequencies for /o/ and /u/ than for /a/, /e/ and /i/. This is due to the sublingual cavity, which is connected to the outside space by a narrow lip channel (see Fig. 6b). The lip channel acts like a side hole in a wind instrument: it induces a low acoustic pressure in the middle of the sublingual cavity, which favors transverse resonance at relatively low frequency. As illustrated in Fig. 6b for /u/, the nodal line (minimum acoustic pressure appearing in blue) of the transverse resonances created is aligned with the lip channel. This configuration considerably reduces the transmission of acoustic energy because the main direction of propagation is transverse to the lip channel. The consequence is a deep trough in the transfer function, as can be seen in the 4 to 6 kHz interval in Fig. 6a. There are other more complex transverse resonances at higher frequencies with a similar but smaller effect that reduces sound transmission overall. This explains why the TLM spectra have a generally higher amplitude at HF. Thus, the participants may perceive the presence of spectral peaks in the 4 to 6 kHz interval as unnatural, and the HF level of /o/ and /u/ generated by TLM as too high. Furthermore, people's hearing abilities are better in this frequency range than at higher frequencies, which induces

a higher perceptual impact. It may be possible to make an improved 1D model which could generate an effect similar to this transverse resonance. This could be achieved, for example, by adding a branch to the main tract as in [23].

In the first experiment, the vowel /a/ was perceived as more natural with TLM. This does not change the overall difference between methods and it was not observed in the second experiment. Thus, this effect is smaller than the differences observed for /o/ and /u/. However, it may be related to a higher spectral amplitude of MM in the 8 to 10 kHz range (see Figs. 2a and 2b). This might indicate that TLM is accidentally slightly more realistic than MM in this frequency range. Another explanation might be that proper modeling of the radiation is more important for the vowel /a/ as it corresponds to a larger mouth opening than the other vowels. Even though MM describes the internal acoustic field more accurately, the approximation made in the description of the radiation (no lips, a baffled exit, and no diffraction by the head and torso) may be more detrimental when a 3D model is used than with a simpler 1D model. However, according to Arnela *et al.* [58], the inclusion of the lips does not reduce the amplitude of HF; on the contrary, it increases it. Thus, the reason for this small difference is not clearly understood and may require a better modeling of all the phenomena involved in the radiation.

The lesser naturalness of ABE for /a/, /e/ and /i/ can easily be related to the much higher spectral amplitude at HF (see Figs. 2a-f). Distortions between 3.3 and 4 kHz may also play a role in some cases (male /a/, female /e/ and male /i/, see Figs. 2b, 2c and 2f). ABE has significantly better naturalness for /o/ and /u/, which is comparable to TLM for /o/ and MM for /u/. This might also be related to the HF spectral amplitude, which is comparable to the other models for these phonemes, and even lower than TLM for /u/. The absence of pronounced peaks in the range 4 to 6 kHz range for /u/ probably explains why it is more natural than with TLM. Thus, for some specific phonemes, ABE's naturalness can be as good as when a physical model is used to generate HF. The algorithms used in this study may not be the optimum state of the art since the motivation was to compare TLM and MM to a baseline without physical input and not to evaluate an ABE method. Better ABE implementations and/or methods may have even better naturalness compared to physics-based speech synthesis.

At HF, the difference between the modal and pressed voice consists mainly in a difference of global HF level. The absence of impact of the voice quality can be interpreted in two complementary ways:

- perception of naturalness is not simply related to the overall HF level but to finer spectral cues. This is in agreement with the better naturalness obtained for /o/ and /u/ with MM and ABE, which seems to be related to spectral peaks that are absent and/or less pronounced than with TLM.
- participants may adapt their natural speech reference to the voice quality that they hear.

The significantly poorer naturalness of the female voice may have several, potentially coexisting, causes:

- The main difference between the glottal flow model used for the male and the female voices is the fundamental frequency. Generating a proper female voice may require one to adapt other parameters, such as the intonation curve.
- The parameters of the articulatory model used for the female vocal tract shapes may have been less accurately fitted on the magnetic resonance images (MRI).
- Beyond the gender difference, what is observed could represent an inter-individual difference. In this regard, it might be interesting to conduct a similar experiment with more than two speakers.

The difference in the distribution of the naturalness rating over the phonemes for both genders indicates that the phoneme dependency of naturalness is potentially related to gender-specific or individual-specific vocal tract geometries, and/or voice source features. This could be clarified by conducting a similar experiment using the same voice source with different speaker geometries.

The significant effect of participant may indicate that participants based their judgements on different internal references for naturalness and/or that they had different concepts of naturalness.

Generally, the three methods tested were rated close to 0.5, which is far from perfectly natural. This may be due to multiple complementary factors:

- The material tested, isolated phonemes, is intrinsically not natural, as we are not used to hearing phonemes in isolation. In this regard, it would have been interesting to include recordings of actual human speech sounds as stimuli to evaluate how they would have been rated in this context and how the synthetic stimuli differ. This should be done in future work.
- Although MM describes the acoustic field more accurately, it still neglects and/or simplifies many other aspects of the physics of speech. This include the absence of side cavities, the imperfection of the radiation modeling, and the simplification of the sound generation mechanisms, which, among other phenomena, neglects the coupling between vocal folds and vocal tract.
- Although it was carefully fitted on MRI, the vocal tract geometry used may not be totally realistic. In particular, the asymmetries of real vocal tracts interact significantly with transverse modes [12]. Thus, it would be interesting to use vocal tract geometries obtained directly from MRI in future work.

It is also questionable whether the results would have been different if the low frequency differences had also been included. These differences consist mainly in small formant frequency deviations. It is clear that they can be perceived [35], [36]. However, we did not find any evidence in the literature that this affects the naturalness of the synthetic sounds. As a matter of fact, the formants of a given vowel in actual human production show substantial variation while sounding equally natural. Thus, we would not expect very different results if low frequency differences had been included.

It is also possible that the transfer function blending process

affects naturalness. Future work should therefore compare pure TLM stimuli with stimuli obtained with the blending process.

V. Conclusion

Our results confirm that HF contribute to speech naturalness and that the accuracy of the physical modeling of speech HF impacts the naturalness of synthetic speech. The most prominent differences were observed for the vowels /o/ and /u/, for which it appears to be important to account for the 3D aspects of the acoustic field. For these vowels, some transverse resonances present in the sublingual cavity probably significantly affect the naturalness. As a non-physical ground truth, ABE confirms that even a simplified 1D acoustic model improves naturalness. However, this is not the case for every phoneme tested, and better ABE implementations or methods may have improved naturalness.

Acknowledgments

References

[1] C. Shadle and R. Damper, "Prospects for articulatory synthesis: A position paper," in *4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis*, 2001.

[2] M. Sondhi and J. Schroeter, "A hybrid time-frequency domain articulatory speech synthesizer," *IEEE Trans. Audio Speech Lang. Process.*, vol. 35, no. 7, pp. 955–967, 1987.

[3] S. Fels, F. Vogt, K. Van Den Doel, J. Lloyd, I. Stavness, and E. Vatikiotis-Bateson, "Artisynth: A biomechanical simulation platform for the vocal tract and upper airway," in *International Seminar on Speech Production, Ubatuba, Brazil*, vol. 138, 2006.

[4] P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *PLOS ONE*, vol. 8, no. 4, p. e60603, 2013.

[5] Y. Laprie, M. Loosvelt, S. Maeda, R. Sock, and F. Hirsch, "Articulatory copy synthesis from cine x-ray films," in *InterSpeech–14th Annual Conference of the International Speech Communication Association-2013*, 2013.

[6] B. Story, "Phrase-level speech simulation with an airway modulation model of speech production," *Comput. Speech lang.*, vol. 27, no. 4, pp. 989–1010, 2013.

[7] D. Childers and C. Wong, "Measuring and modeling vocal source-tract interaction," *IEEE Trans. Biomed. Eng.*, vol. 41, no. 7, pp. 663–671, 1994.

[8] K. Stevens, *Acoustic phonetics*. Cambridge, MA: MIT Press, 2000, vol. 30.

[9] P. Birkholz, F. Gabriel, S. Kürbis, and M. Echternach, "How the peak glottal area affects linear predictive coding-based formant estimates of vowels," *J. Acoust. Soc. Am.*, vol. 146, no. 1, pp. 223–232, 2019.

[10] G. Fant, J. Liljencrants, Q. Lin *et al.*, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, no. 1985, pp. 1–13, 1985.

[11] R. Blandin, M. Arnela, S. Félix, J. Doc, and P. Birkholz, "Comparison of the finite element method, the multimodal method and the transmission-line model for the computation of vocal tract transfer functions,," in *Interspeech*, 2021, pp. 3330–3334.

[12] R. Blandin, M. Arnela, R. Laboissière, X. Pelorson, O. Guasch, A. Van Hirtum, and X. Laval, "Effects of higher order propagation modes in vocal tract like geometries," *J. Acoust. Soc. Am.*, vol. 137, no. 2, pp. 832–843, 2015.

[13] M. Arnela, S. Dabbaghchian, O. Guasch, and O. Engwall, "MRI-based vocal tract representations for the three-dimensional finite element synthesis of diphthongs," *IEEE Trans. Audio Speech Lang. Process.*, vol. 27, no. 12, pp. 2173–2182, 2019.

[14] M. Fleischer, S. Pinkert, W. Mattheus, A. Mainka, and D. Mürbe, "Formant frequencies and bandwidths of the vocal tract transfer function are affected by the mechanical impedance of the vocal tract wall," *Biomech. Model. Mechanobiol.*, vol. 14, no. 4, pp. 719–733, 2015.

[15] T. Vampola, J. Horáček, V. Radolf, J. Švec, and A. Laukkanen, "Influence of nasal cavities on voice quality: Computer simulations and experiments," *J. Acoust. Soc. Am.*, vol. 148, no. 5, pp. 3218–3231, 2020.

[16] H. Takemoto, P. Mokhtari, and T. Kitamura, "Acoustic analysis of the vocal tract during vowel production by finite-difference time-domain method," *J. Acoust. Soc. Am.*, vol. 128, no. 6, pp. 3724–3738, 2010.

[17] R. Blandin, M. Arnela, S. Félix, J. Doc, and P. Birkholz, "Efficient 3D acoustic simulation of the vocal tract by combining the multimodal method and finite elements," *IEEE Access*, 2022.

[18] A. Gully, H. Daffern, and D. Murphy, "Diphthong synthesis using the dynamic 3D digital waveguide mesh," *IEEE Trans. Audio Speech Lang. Process.*, vol. 26, no. 2, pp. 243–255, 2017.

[19] A. Gully, "Diphthong synthesis using the three-dimensional dynamic digital waveguide mesh," Ph.D. dissertation, University of York, 2017.

[20] P. Krug, S. Stone, and P. Birkholz, "Intelligibility and naturalness of articulatory synthesis with VocalTractLab compared to established speech synthesis technologies," in *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 102–107.

[21] B. Monson, E. Hunter, A. Lotto, and B. Story, "The perceptual significance of high-frequency energy in the human voice," *Front. Psychol.*, vol. 5, p. 587, 2014.

[22] H. Boyd-Pratt and J. D, "The perception and use of high-frequency speech energy: Clinical and research implications," *Perspect. ASHA Spec. Interest Groups*, vol. 5, no. 5, pp. 1347–1355, 2020.

[23] P. Birkholz and S. Drechsel, "Effects of the piriform fossae, transvelar acoustic coupling, and laryngeal wall vibration on the naturalness of articulatory speech synthesis," *Speech Communication*, vol. 132, pp. 96–105, 2021.

[24] B. Moore and C. Tan, "Perceived naturalness of spectrally distorted speech and music," *J. Acoust. Soc. Am.*, vol. 114, no. 1, pp. 408–419, 2003.

[25] P. Bachhav, M. Todisco, and N. Evans, "Exploiting explicit memory inclusion for artificial bandwidth extension," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada, 2018, pp. 5459–5463.

[26] ——, "Efficient super-wide bandwidth extension using linear prediction based analysis-synthesis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada, 2018, pp. 5429–5433.

[27] B. Monson, A. Lotto, and B. Story, "Analysis of high-frequency energy in long-term average spectra of singing, speech, and voiceless fricatives," *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1754–1764, 2012.

[28] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 26, no. 4, pp. 1–13, 1985.

[29] G. Fant, "The LF-model revisited. transformations and frequency domain analysis," *STL-QPSR*, vol. 36, no. 2–3, pp. 119–156, 1995.

[30] R. E. Hillman, E. Oesterle, and L. L. Feth, "Characteristics of the glottal turbulent noise source," *J. Acoust. Soc. Am.*, vol. 74, no. 3, pp. 691–694, 1983.

[31] P. Alku and E. Vilkman, "A comparison of glottal voice source quantification parameters in breathy, normal and pressed phonation of female and male speakers," *Folia Phoniatr. Logop.*, vol. 48, no. 5, pp. 240–254, 1996.

[32] S. Drechsel, Y. Gao, J. Frahm, and P. Birkholz, "Modell einer Frauenstimme für die artikulatorische Sprachsynthese mit VocalTractLab," *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019*, pp. 239–246, 2019.

[33] P. Birkholz and D. Jackel, "Influence of temporal discretization schemes on formant frequencies and bandwidths in time domain simulations of the vocal tract system," in *Eighth International Conference on Spoken Language Processing*, 2004.

[34] P. Birkholz, "3D-Artikulatorische Sprachsynthese," Ph.D. dissertation, Universität Rostock, 2014.

[35] J. Flanagan, "A difference limen for vowel formant frequency," *J. Acoust. Soc. Am.*, vol. 27, no. 3, pp. 613–617, 1955.

[36] D. Kewley-Port and C. Watson, "Formant-frequency discrimination for isolated English vowels," *J. Acoust. Soc. Am.*, vol. 95, no. 1, pp. 485–496, 1994.

[37] C. Sinclair, "GLIM for preference," in *GLIM 82: Proceedings of the international conference on generalised linear models*. Springer, 1982, pp. 164–178.

[38] R. Dittrich, R. Hatzinger, and W. Katzenbeisser, "Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings," *J. R. Stat. Soc., C: Appl. Stat.*, vol. 47, no. 4, pp. 511–525, 1998.

[39] R. Hatzinger and R. Dittrich, "Prefmod: An R package for modeling preferences based on paired comparisons, rankings, or ratings," *J. Stat. Softw.*, vol. 48, pp. 1–31, 2012.

[40] M. Cattelan, C. Varin, and D. Firth, "Dynamic Bradley–Terry modelling of sports tournaments," *J. R. Stat. Soc., C: Appl. Stat.*, vol. 62, no. 1, pp. 135–150, 2013.

[41] R. Dittrich and R. Hatzinger, "Fitting loglinear Bradley-Terry models (LLBT) for paired comparisons using the R package prefmod," *Psychological Test and Assessment Modeling*, vol. 51, no. 2, p. 216, 2009.

[42] R. Bradley and M. Terry, "Rank analysis of incomplete block designs: II. additional tables for the method of paired comparisons," *Biometrika*, vol. 41, no. 3-4, pp. 502–537, 1954.

[43] M. Tallon, M. Greenlee, E. Wagner, K. Rakoczy, W. Wiedermann, and U. Frick, "Assessing heterogeneity in students' visual judgment: Model-based partitioning of image rankings," *Front. Psychol.*, vol. 13, 2022.

[44] D. Snopková, L. De Cock, V. Juřík, O. Kvarda, M. Tancoš, L. Herman, and P. Kubíček, "Isovists compactness and stairs as predictors of evacuation route choice," *Sci. Rep.*, vol. 13, no. 1, p. 2970, 2023.

[45] N. Hall, F. Péron, S. Cambou, L. Callejon, and C. Wynne, "Food and food-odor preferences in dogs: a pilot study," *Chem. Senses*, vol. 42, no. 4, pp. 361–370, 2017.

[46] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[47] R. Hatzinger and M. J. Maier, *prefmod: Utilities to Fit Paired Comparison Models for Preferences*, 2022, R package version 0.8-35. [Online]. Available: https://CRAN.R-project.org/package=prefmod

[48] H. Turner and D. Firth, *Generalized nonlinear models in R: An overview of the gnm package*, 2022, r package version 1.1-2. [Online]. Available: https://cran.r-project.org/package=gnm

[49] C. Krueger and L. Tian, "A comparison of the general linear mixed model and repeated measures ANOVA using a dataset with multiple missing data points," *Biol. Res. Nurs.*, vol. 6, no. 2, pp. 151–157, 2004.

[50] R. Gueorguieva and J. Krystal, "Move over ANOVA: progress in analyzing repeated-measures data andits reflection in papers published in the archives of general psychiatry," *Arch. Gen. Psychiatry*, vol. 61, no. 3, pp. 310–317, 2004.

[51] R. Yang, "Towards understanding and use of mixed-model analysis of agricultural experiments," *Can. J. Plant Sci.*, vol. 90, no. 5, pp. 605–627, 2010.

[52] J. Fox and S. Weisberg, *An R Companion to Applied Regression*, 3rd ed. Thousand Oaks CA: Sage, 2019. [Online]. Available: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

[53] H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. L. Pedersen, E. Miller, S. M. Bache, K. Müller, J. Ooms, D. Robinson, D. P. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, and H. Yutani, "Welcome to the tidyverse," *J. Open Source Softw.*, vol. 4, no. 43, p. 1686, 2019. [Online]. Available: https://doi.org/10.21105/joss.01686

[54] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *J. Stat. Softw.*, vol. 67, no. 1, p. 1–48, 2015. [Online]. Available: https://www.jstatsoft.org/index.php/jss/article/view/v067i01

[55] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmertest package: Tests in linear mixed effects models," *J. Stat. Softw.*, vol. 82, no. 13, p. 1–26, 2017. [Online]. Available: https://www.jstatsoft.org/index.php/jss/article/view/v082i13

[56] R. Lenth, *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2022, R package version 1.7.3. [Online]. Available: https://CRAN.R-project.org/package=emmeans

[57] T. Hothorn, F. Bretz, and P. Westfall, "Simultaneous inference in general parametric models," *Biom. J.*, vol. 50, no. 3, pp. 346–363, 2008.

[58] M. Arnela, R. Blandin, S. Dabbaghchian, O. Guasch, F. Alías, X. Pelorson, A. Van Hirtum, and O. Engwall, "Influence of lips on the production of vowels based on finite element simulations and experiments," *J. Acoust. Soc. Am.*, vol. 139, no. 5, pp. 2852–2859, 2016.