Perceptual evaluation of the naturalness of broadband articulatory speech synthesis using a 1D versus a 3D acoustic model

Rémi Blandin[1], Vincent Didone[2], Peter Birkholz[1], Angélique Remacle[3,4]

[1]*Institute of Acoustics and Speech Communication, TU Dresden, Dresden, 01062, Germany*
[2]*Psychology and Neuroscience of Cognition Research Unit (PsyNCog),*
Quantitative psychology, University of Liège, Liège, Belgium
[3]*Research Unit for a Life-Course Perspective on Health and Education,*
Faculty of Psychology, Speech and Language Therapy, and Educational Sciences, University of Liège, Liège, Belgium
[4]*Center For Research in Cognition and Neurosciences, Faculty of Psychological Science and Education,*
Université Libre de Bruxelles, Brussels, Belgium
remi.blandin@tu-dresden.de

Abstract

Keywords: Speech, acoustics, perception, naturalness, articulatory synthesis

Articulatory synthesis is a useful tool to explore the relationship between the speech production and perception processes. However, including the high frequencies (HF, above about 5 kHz) requires a three-dimensional (3D) acoustical model for realistic simulations. In this frequency range, one-dimensional (1D) acoustic models fail to predict additional resonances and anti-resonances related to the 3D properties of the acoustic field. While articulatory synthesis based on 3D acoustic models is nowadays achievable for isolated phonemes, the impact of such models on the perception by human listeners remains largely unknown. The objective of this work was to determine whether a more realistic computation of transfer functions with a frequency domain approach results in phonemes perceived as more natural. For this purpose, a perception experiment using a 4-points Likert scale was conducted to evaluate the naturalness of seven static phonemes, /a, e, i, ə, f, s, ʃ/, synthesized with a 1D and a 3D models. No significant influence of the acoustic model was found, however, significant differences between the phonemes were perceived.

1. Introduction

Articulatory synthesis relies on a description of the physical phenomena involved in speech production. It uses a geometrical description of the speech production apparatus and models the sound generation and propagation mechanisms.

A very common simplifying assumption is to consider that the acoustic propagation is unidimensional, i.e. it depends only on the cross-sectional area along the vocal tract (Sondhi and Schroeter 1987). However, this assumption is increasingly unrealistic toward HF. The divergence with realistic models first appears as shifts in resonance frequencies due to the curvature of the acoustic field at changes in cross-sectional area. At HF, above about 4-5 kHz, the higher order modes generate additional resonances unpredicted by 1D models (Blandin, Arnela, Laboissière, et al. 2015). These phenomena can be properly described by 3D models, such as finite elements (Arnela et al. 2019), finite differences (Takemoto, Mokhtari, and Kitamura 2010), the multimodal method (Blandin, Arnela, Félix, et al. 2022) or waveguide mesh models (Gully, Daffern, and Murphy 2017).

So far, articulatory synthesis based on 3D acoustic models has been achieved for isolated phonemes (Gully, Daffern, and Murphy 2017; Arnela et al. 2019; Dabbaghchian et al. 2021). One can expect that using more realistic acoustic models for articulatory synthesis would result in a greater resemblance to actual human speech, and that it would be perceived as more natural. However, the hearing sensitivity toward HF reduces both in terms of sound pressure level (SPL) and frequency discrimination. Thus, this increase of realism, which happens mostly at HF, may not substantially impact the perceived naturalness. This implies the necessity to evaluate the perceptual impact of such models.

Prior to our study, to our knowledge, only one study addressed this question using a perceptual test. Gully (2017) found that diphthongs generated with a 3D waveguide mesh were perceived as more natural than diphthongs generated with a 2D waveguide mesh and a Kelly-Lochbaum 1D model. However, the 3D simulation method used, waveguide mesh, is non standard and not very well proven, so the increase of realism can be questioned. The use of a time-domain method reduced the quality of the simulations above 5 kHz, and the observed difference was mainly due to differences below 5 kHz. Thus, to investigate the perceptual impact of HF, a better modelling of these frequencies, and particularly of the loss mechanisms is necessary.

Our objective was to determine whether an articulatory synthesis based on a 3D acoustic model with a frequency domain approach results in phonemes perceived as more natural.

To that end, four vowels (/a, i, u/ and /ə/) and three consonants (/f, s, ʃ/) were synthesized for a male and a female speaker. For this purpose, we applied a source-filter approach in which the filter (vocal tract transfer function) was computed with both 1D and 3D acoustic models.

2. Methods

2.1. Stimuli generation

The stimuli were generated with the articulatory synthesizer VocalTractLab3D[1] (Blandin, Arnela, Félix, et al. 2022), which can synthesize speech sounds with a 1D or a 3D acoustic model. The vocal tract geometries used are predefined in VocalTractLab3D. They have been generated by fitting the parameters of the geometric vocal tract model implemented in VocalTractLab3D to magnetic resonance images (MRI) obtained for multiple phonemes produced by a male (Birkholz 2013) and a female (Drechsel et al. 2019) speaker.

The 3D simulation method implemented in VocalTractLab3D is a multimodal method which relies on a decomposition of the acoustic field $p(x, y, z)$ over the local transverse modes $\varphi_n(y, z)$:

$$p(x, y, z) = \sum_{n=0}^{\infty} p_n(x)\varphi_n(y, z), \tag{1}$$

where $p_n(x)$ descibes the amplitude of the transverse mode $\varphi_n(y, z)$ along the vocal tract.

A complete description of the method can be found in Blandin, Arnela, Félix, et al. 2022. Its main advantages are to be computationally efficient and to provide a better understanding of the physical phenomena involved. In the context of our study, another advantage is the possibility to tune the dimension of the model through the number of transverse modes used: using only one transverse modes makes a 1D simulation and using a correctly tuned number makes a 3D simulation. This tuning was done through convergence tests and comparison with finite elements simulations (Blandin, Arnela, Félix, et al. 2022).

Several vocal tract transfer functions were computed:

- for the vowels (/a, i, u, ə/), from the volume velocity at the glottis and from the acoustic pressure at a point about 2 cm downstream of the glottis to the acoustic pressure at a point located 1 m in front of the lips,
- for the fricatives (/f, s, ʃ/), from the acoustic pressure at a point in the sound generation area (teeth or hard palate) to the acoustic pressure at a point located 1 m in front of the lips. This point source was placed between the lips for /f/, at the downstream edge of the lower lips for /s/, and between the teeth for /ʃ/. Its location was fine tuned to reproduce properly the intended phonemes.

The vocal fold sound source signal was generated using the Liljencrants- Fant (LF) glottal pulse model (Fant, Liljencrants, Lin, et al. 1985) implemented in VocalTractLab3D. The fundamental frequency was set to a target of 120 Hz and 210 Hz for the male and female voices, respectively. To increase the naturalness of the stimuli, small variations of fundamental frequency were generated with a "flutter" as proposed in Eq. (1) in Klatt and Klatt (D. Klatt and L. Klatt 1990). An open quotient of 0.5, a shape quotient of 3.0 and spectral tilt of 0.02 were used in order to generate a modal voice quality which corresponds to normal speech.

The noise sources present immediately downstream of the vocal folds for the vowels and in the vicinity of obstacles for the fricatives were generated by filtering Gaussian white noise with a first-order low-pass filter. Cut-off frequencies of 10 kHz for the vowels, 5 kHz for /f/, and 8 kHz for /s, ʃ/ were used. These values roughly create source spectra according to Shadle 1991. The gain of the sources was adjusted in such a way that the intensity of the produced noise at the different places of articulation closely matches real fricative intensities (Birkholz 2014).

To generate the stimuli, the source signals were convolved with the impulse responses of the transfer functions. In the case of the vowels, the amplitude p_s of the noise source was set proportional to the cube of the low frequency part of the vocal fold output particle velocity $\bar{v}$, $p_s \propto |\bar{v}^3|$ as proposed by Stevens (Stevens 2000). Applying the principle of superposition of linear acoustics, the signals from the noise source attenuated by 30 dB and the vocal fold were then added to form the radiated sound. In total, 28 stimuli were generated: 2 acoustic models (1D or 3D)×7 phonemes×2 genders.

2.2. Perception experiment

Naturalness was evaluated by 31 participants aged between 21 and 28 years old (4 males and 27 females), all native French speakers without past or present hearing problems. They all had hearing thresholds $\leq$ 20 dB hearing level (HL) bilaterally at octave frequencies between 500 and 8000 Hz (audiometric screening with pure-tone audiometry using a MADSEN Itera II audiometer with TDH-39 earphones). The experiment took place in a listening booth where the stimuli were played through a loudspeaker placed one meter in front of the participants. The choice of a loudspeaker instead of headphones was motivated by the better control over the listening conditions that it offers and the fact that it is closer to a real life listening condition. In addition, it eliminates the problem of achieving the same HF response for all participants, which is challenging with headphones. The gain of the amplifier of the loudspeaker was adjusted so that the level of the stimuli at the location of the head of the participants was 70 dB SPL. Participants listened to each stimulus as many times as they wanted and were asked to rate it on a 4-points Likert scale ranging from 0 (not at all natural) to 3 (completely natural). The stimuli were presented in a randomized order and each stimulus was rated twice at random times.

2.3. Statistical analysis

Participants' responses were analyzed with an ordinal cumulative logistic regression model using the "ordinal" R packages (Christensen 2015). A random effect of the participant was used and the fixed effects were the acoustic model (two conditions: the 1D and 3D models), the type of phoneme (/a, i, u, ə, f, s, ʃ/), the gender of the speaker (female and male) and the moment of the test (two moments: test and retest). The model included each main factor, the interactions between the model and the phoneme, and the interaction between the model and the gender. The significance of the main effect (phoneme) and the interactions were assessed using a likelihood-ratio test. Contrasts (or comparisons) were made between the levels of the factors and interactions that were significant in the analysis of the models using the R packages emmeans (Lenth et al. 2019) and multcomp (Jiang and Nguyen 2007). The Holm method of alpha adjustment was used to correct for multiple testing. Inter-rater reliability was assessed using the Intraclass Correlation Coefficient (ICC) (Shrout and Fleiss 1979).

3. Results

Figure 1 shows the average rating for each phoneme synthesized with both acoustic models. The level of inter-rater relia-

[1] VocalTractLab3D is freely available at: https://vocaltractlab.de/index.php?page=vocaltractlab-download

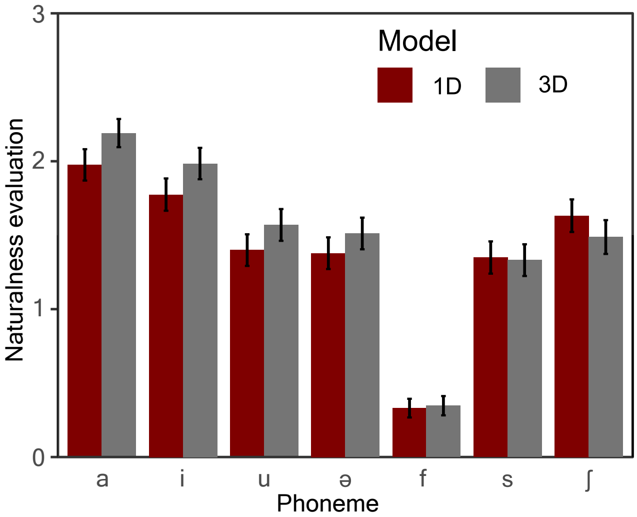

Figure 1: Average ratings for the phonemes synthesized with the 1D and 3D acoustic models in the naturalness rating task using a Likert scale from 0 (not at all natural) to 3 (completely natural).

bility can be regarded as good to excellent with ICC = 0.9 (with 95% confident interval = 0.86 - 0.94 and p < .0001). There was no significant effect of the acoustic model (χ^2 (1) = 2.96, p = 0.085) nor the gender (χ^2 (1) = 1.13, p = 0.288). The interaction between the model and the gender was non-significant (χ^2(1) = 0.021, p = 0.885), as well as the interaction between the model and the phoneme (χ^2(6) = 6.82, p = 0.337). However, a significant effect of the phoneme was found (χ^2 (6) = 464, p < 0.001).

As depicted in Fig. 1, the phonemes /a/ and /i/ were rated as the most natural, with no significant difference between their ratings. /u, ə, s/ and /ʃ/ form another group with similar but lower naturalness. /f/ was rated the least natural, far below all the other phonemes, so it is mostly rated as "not at all natural".

4. Discussion and conclusion

In contrast to Gully (2017), our results do not show a significant influence of the 3D acoustic model on the perceived naturalness. This discrepancy between the two studies could be explained by differences in the simulation method, the phonetic material (isolated phonemes including consonants vs. diphthongs), the listening conditions (loudspeaker vs. headphones), or the experimental design (Likert scale vs. MUSHRA (Series 2014)). Additionally, the use of electrolaryngograph signals from human subjects for the sound source in the study of Gully might generate globally more natural sounding stimuli than the LF model.

In Fig. 1, the average naturalness of the vowels is slightly better for the 3D model compared to the 1D model. On the other hand, the p-value of the effect of the model (p = 0.085) is close to 0.05, which is the usual limit to consider an effect as significant. This suggests that a weak but significant effect might could be revealed using more participants, and/or different experimental design choices, such as a linear scale instead of a Likert scale. This tends to be confirmed in a subsequent study by Blandin, Stone, et al. 2023, showing significant differences using pair comparisons between 1D and 3D models, and a linear scale to rate the naturalness. However, only 5 vowels (/a, e, i, o, u/) were used and the frequencies up to 4 kHz were similar for each model. The perceived differences between 1D and 3D mainly concern the vowels /o/ and /u/.

As shown in Fig. 1, the highest average naturalness ratings are around 2 (rather natural), so none of the phonemes were rated as completely natural. This may be due to the material presented (isolated phonemes), geometric inaccuracies, limitations of the LF model, or remaining physical approximations (point sound source and simplified radiation).

Regardless of the acoustic model, there are significant differences of naturalness between the phonemes. This confirms that the perceptual experiment was successful in detecting variations of naturalness, but that the effect of the model, if existent, is probably too small to be observed this way. On the other hand, this also means that other phoneme-specific factors have more impact than the acoustic model.

Given the multiplicity of the phenomena involved, it is difficult to identify accurately which phenomenon is affecting naturalness the most for a specific phoneme. However, one can formulate hypotheses. For example, the sound generation is expected to take place in the vicinity of the lips for /f/. Therefore, the simplification of the lip shape as a flat opening may degrade the naturalness more for this specific phoneme. This may explain the particularly low rating for /f/. More generally, other causes may negatively affect the naturalness of the synthetic fricatives. The simplification of the aeroacoustic sound sources as a single point source may be a too rough approximation, their greater sensitivity to small geometric details may make them more sensitive to geometric inaccuracies, and the more directional radiation of the fricatives may be further degraded by the radiation simplifications.

Regarding the vowels, the source filter coupling (Titze 2008) was not taken into account in this study. The dependence of this phenomenon on the vocal tract shape may contribute to differences of the naturalness between the vowels (Birkholz et al. 2019): for vowels having a greater source filter coupling, not taking it into account may affect more their naturalness. This is in line with the results of Birkholz et al. 2019 who reported a stronger effect on close-mid to close vowels (/i, ə, u/) for which a lower naturalness was observed. In addition, the participants are not used to listening to the vowel /ə/ in isolation in natural speech. This may explain why it has the lowest naturalness among the vowels.

5. Acknowledgements

This study was supported by the German Federal Ministry of Education and Research (BMBF) within the project "Promise-AI", funding code 16SV8988.

6. References

Arnela, M, S Dabbaghchian, O Guasch, and O Engwall (2019). “MRI-based vocal tract representations for the three-dimensional finite element synthesis of diphthongs”. In: *IEEE Trans. Audio Speech Lang. Process.* 27.12, pp. 2173–2182.

Birkholz, P (2013). “Modeling consonant-vowel coarticulation for articulatory speech synthesis”. In: *PloS one* 8.4, e60603.

— (2014). “Enhanced area functions for noise source modeling in the vocal tract”. In: *10th International Seminar on Speech Production, Köln*, pp. 1–4.

Birkholz, P, F Gabriel, S Kürbis, and M Echternach (2019). “How the peak glottal area affects linear predictive coding-based formant estimates of vowels”. In: *J. Acoust. Soc. Am.* 146.1, pp. 223–232.

Blandin, R, M Arnela, S Félix, JB Doc, and P Birkholz (2022). “Efficient 3D acoustic simulation of the vocal tract by combining the multimodal method and finite elements”. In: *IEEE Access* 10, pp. 69922–69938.

Blandin, R, M Arnela, R Laboissière, X Pelorson, O Guasch, A Van Hirtum, and X Laval (2015). “Effects of higher order propagation modes in vocal tract like geometries”. In: *J. Acoust. Soc. Am.* 137.2, pp. 832–843.

Blandin, R, S Stone, A Remacle, V Didone, and P Birkholz (2023). “A Comparative Study of 3D and 1D Acoustic Simulations of the Higher Frequencies of Speech”. In: *IEEE Trans. Audio Speech Lang. Process.*

Christensen, RHB (2015). *Ordinal—regression models for ordinal data, 2015. R package version 2015.6-28.*

Dabbaghchian, S, M Arnela, O Engwall, and O Guasch (2021). “Simulation of vowel-vowel utterances using a 3D biomechanical-acoustic model”. In: *Int. J. Numer. Methods Biomed. Eng.* 37.1, e3407.

Drechsel, S, Y Gao, J Frahm, and P Birkholz (2019). “Modell einer Frauenstimme für die artikulatorische Sprachsynthese mit VocalTractLab”. In: *Konferenz Elektronische Sprachsignalverarbeitung*. TUDpress, Dresden, pp. 239–246.

Fant, G, J Liljencrants, Q Lin, et al. (1985). “A four-parameter model of glottal flow”. In: *STL-QPSR* 4.1985, pp. 1–13.

Gully, AJ (2017). “Diphthong Synthesis using the Three-Dimensional Dynamic Digital Waveguide Mesh”. PhD thesis. University of York.

Gully, AJ, H Daffern, and DT Murphy (2017). “Diphthong synthesis using the dynamic 3D digital waveguide mesh”. In: *IEEE/ACM Trans. Audio, Speech, Language Process.* 26.2, pp. 243–255.

Jiang, J and T Nguyen (2007). *Linear and generalized linear mixed models and their applications*. Vol. 1. Springer.

Klatt, DH and LC Klatt (1990). “Analysis, synthesis, and perception of voice quality variations among female and male talkers”. In: *J. Acoust. Soc. Am.* 87.2, pp. 820–857.

Lenth, R, H Singmann, J Love, P Buerkner, and M Herve (2019). “Emmeans: estimated marginal means, aka least-squares means (Version 1.3. 4)”. In: *Emmeans Estim. Marg. Means Aka Least-Sq. Means https://CRAN. R-project. org/package= emmeans*.

Series, B (2014). “Method for the subjective assessment of intermediate quality level of audio systems”. In: *International Telecommunication Union Radiocommunication Assembly*.

Shadle, CH (1991). “The effect of geometry on source mechanisms of fricative consonants”. In: *Journal of phonetics* 19.3-4, pp. 409–424.

Shrout, PE and JL Fleiss (1979). “Intraclass correlations: uses in assessing rater reliability.” In: *Psychological bulletin* 86.2, p. 420.

Sondhi, M and J Schroeter (1987). “A hybrid time-frequency domain articulatory speech synthesizer”. In: *IEEE/ACM Trans. Audio, Speech, Language Process.* 35.7, pp. 955–967.

Stevens, KN (2000). *Acoustic phonetics*. Vol. 30. MIT press.

Takemoto, H, P Mokhtari, and T Kitamura (2010). “Acoustic analysis of the vocal tract during vowel production by finite-difference time-domain method”. In: *J. Acoust. Soc. Am.* 128.6, pp. 3724–3738.

Titze, IR (2008). “Nonlinear source–filter coupling in phonation: Theory”. In: *J. Acoust. Soc. Am.* 123.5, pp. 2733–2749.