Evaluating Features and Metrics for High-Quality Simulation of Early Vocal Learning of Vowels

Branislav Gerazov\textsuperscript{1}, Daniel van Niekerk\textsuperscript{2}, Anqi Xu\textsuperscript{2}, Paul Konstantin Krug\textsuperscript{3}, Peter Birkholz\textsuperscript{3}, and Yi Xu\textsuperscript{2}

\textsuperscript{1}Faculty of Electrical Engineering and Information Technologies, UCMS, Skopje, Macedonia
\textsuperscript{2}Department of Speech, Hearing and Phonetic Sciences, University College London, UK
\textsuperscript{3}Institute of Acoustics and Speech Communication, TU Dresden, Germany

\texttt{gerazov@feit.ukim.edu.mk}

1. Introduction

The way infants use auditory cues to learn to speak despite the acoustic mismatch of their vocal apparatus is a hot topic of scientific debate. The simulation of early vocal learning using articulatory speech synthesis offers a way towards gaining a deeper understanding of this process. One of the crucial parameters in these simulations is the choice of features and a metric to evaluate the acoustic error between the synthesised sound and the reference target. We contribute with evaluating the performance of a set of 40 feature-metric combinations for the task of optimising the production of static vowels with a high-quality articulatory synthesiser. Towards this end we assess the usability of formant error and the projection of the feature-metric error surface in the normalised F1-F2 formant space. We show that this approach can be used to evaluate the impact of features and metrics and also to offer insight to perceptual results.

1.1. Dataset

We generated in total 1 million vocal tract shapes (VTSs) for both models by random sampling of 17 of the 24 VTL vocal tract parameters in the parameter ranges of the speaker models: adult mimicry of babbles, which has yielded low vowel identification scores [9, 7, 10]. Some have successfully simulated vowel learning of syllables [11, 15].

1.2. Methodology

2.1. Dataset

Target vowel templates. The human speaker static vowel target templates comprise a single realisation of the five vowels: /a/, /e/, /u/, /o/, and /l/, as used in standard Macedonian, spoken by a native male speaker. This limited set provides ample coverage of the formant space as can be seen in Fig. 1.

Synthesised data. The data was synthesised using the VocalTractLab API.\textsuperscript{1} Two models were used in the analysis: i) an adult model based on MRI scans of a human subject, and ii) a prototype child model created as a scaled down version [21].

We generated in total 1 million vocal tract shapes (VTSs) for both models by random sampling of 17 of the 24 VTL vocal tract parameters in the parameter ranges of the speaker models: hyoid x and y position (HX, HY), jaw x and angle (JX, JA), lip protrusion and distance (LP, LD), velum shape (VS), tongue

\textsuperscript{1}VTL v.2.2 http://www.vocaltractlab.de/
centre, blade and tip x and y (TCX, TCY, TBX, TBY, TTX, TTY), and the four tongue side vertical positions (TS1, TS2, TS3, TS4) [19]. We generated the 500,000 VTSs for each model in 5 runs with 100,000 iterations each. All runs started from the neutral vocal tract position corresponding to a central schwa.

We prefiltered the VTSs based on the positional constraints for the tongue parameters and vocal tract closure. We then extracted F1 and F2 from the magnitude of the volume velocity transfer function using a peak picking algorithm and postfiltered the VTSs based on the expected F1 and F2 ranges [22].

Finally, we postfiltered the synthesised speech signals based on their low-frequency energy to include only VTSs that resulted with sustained phonation. This rigorous selection process resulted with 15,229 (3% of the original VTSs samples) for the adult model and 8,510 (1.7%) for the child model.\(^2\) Fig. 1 shows the formant spread for the two speaker models with the target vowel’s formant frequencies superimposed. We can see a well formed vowel triangle in both cases, with a larger spread for the child model in line with [22].

\(2\)Supplementary materials – http://www.homepages.ucl.ac.uk/~uclyyix/EVL/feature-metric.html

3. Results

3.1. Formant space analysis

Impact of high frequency emphasis. The obtained formant error when using HF emphasis aggregated across the vowels,
metrics, normalisation, and grouped by base feature for each model is shown in Fig. 2. We can see that the use of HF emphasis on average increases the error as measured by the distance to the target in the normalised F1-F2 space.

Impact of normalisation. The formant error results do not reveal a clear cut impact of normalisation in the optimisation task. Instead we investigate the error surface projections of MFCC12 MSE for /e/ and /u/ for the adult and child models shown in Fig. 3. We can see that the impact of normalisation is more pronounced for /a/. In deed, while it only leads to a loss of the pronounced minimum, for the child model the effects of normalisation are severe, shifting the global minimum to a different formant location altogether.

Impact of the metrics. The averaged impact of the metrics for all the vowels for the base features without HF emphasis and normalisation is shown in Fig. 4. We can see that although their performance is close, MSE has a slight advantage on average for both models.

Impact of the features. Fig. 5 shows the averaged impact of the base features without HF emphasis and normalisation for the different vowels. We can see that the different base features work consistently across the vowels and the two models. There are cases where MFCC12 work better (adult /a/ and /u/ and child /e/), but also worse (child /o/). This can be explained by the similarity of their error surfaces, as seen in Fig. 3.
If we examine the selected formant position for adult /e/ and compare it to the error surface shown in Fig. 3, we can see that it does not coincide with the expected global minimum. This is due to the variance of the binned errors around the calculated mean not shown here because of space limitations.

4. Conclusions

While formant error does not tell the whole story when it comes to the acoustic realisation of vowels, our findings show that normalised formant distance correlates well with perceptual scores of vowel quality. We have also shown that the projection of the error surface in the normalised F1-F2 space can serve to evaluate feature-metric pairs and predict their perceptual performance for the optimisation of vocal tract parameters in simulations of vocal learning. Moreover, these projections show wrong our intuition that there is a straightforward correspondence between error optimisation in the feature space and minimisation of formant error.

From the evaluated feature-metric pairs we have demonstrated similarity in the formant space error surfaces, formant errors and perceptual scores between the MFCC12, MFCC22 and Log Mel base features. None of them has demonstrated superiority in the task of vowel production optimisation. The performance of the different metrics is also similar, with MSE giving slightly better average results. High frequency emphasis has shown to increase formant error and should not be used for the task of vowel learning. However, it might have a positive impact on consonant learning. Finally, normalisation has been shown to have a contradicting and severe impact on the error surface dynamics improving it for some vowels and degrading it for others.

5. Acknowledgements

This work has been funded by the Leverhulme Trust Research Project Grant RPG-2019-241: ‘High quality simulation of early vocal learning’. Formant analysis of target speaker was funded by the National Science Centre of Poland 2017/25/B/HS2/00760.
6. References

[1] P. K. Kuhl, “A new view of language acquisition,” *Proceedings of the National Academy of Sciences*, vol. 97, no. 22, pp. 11 850–11 857, 2000. [Online]. Available: https://www.pnas.org/content/97/22/11850

[2] M. M. Vihman and B. de Boysson-Bardies, “The nature and origins of ambient language influence on infant vocal production and early words,” *Phonetic*, vol. 51, no. 1-3, pp. 159–169, 1994.

[3] M. Pérez-Pereira and G. Conti-Ramsden, *Language development and social interaction in blind children*. Routledge, 2019.

[4] D. K. Oller and P. F. MacNeiagle, “Development of speech production: Perspectives from natural and perturbed speech,” in *The production of speech*. Springer, 1983, pp. 91–108.

[5] L. Ménard, C. Toupin, S. R. Baum, S. Drouin, J. Aubin, and M. Tiede, “Acoustic and articulatory analysis of french vowels produced by congenitally blind adults and sighted adults,” *The Journal of the Acoustical Society of America*, vol. 134, no. 4, pp. 2975–2987, 2013.

[6] W. T. Fitch and J. Giedd, “Morphology and development of the human vocal tract: A study using magnetic resonance imaging,” *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1511–1522, 1999.

[7] P. Messum and I. S. Howard, “Creating the cognitive form of phonological units: The speech sound correspondence problem in infancy could be solved by mirrored vocal interactions rather than by imitation,” *Journal of Phonetics*, vol. 53, pp. 125–140, 2015.

[8] C. Breazeal and B. Scassellati, “Robots that imitate humans,” *Trends in cognitive sciences*, vol. 6, no. 11, pp. 481–487, 2002.

[9] I. S. Howard and P. Messum, “Modeling the development of pronunciation in infant speech acquisition.” 2011.

[10] H. Rasilo and O. Räsänen, “An online model for vowel imitation learning,” *Speech Communication*, vol. 86, pp. 1–23, 2017.

[11] S. Prom-on, P. Birkholz, and Y. Xu, “Training an articulatory synthesizer with continuous acoustic data,” in *INTERSPEECH*, 2013, pp. 349–353.

[12] ———, “Identifying underlying articulatory targets of thai vowels from acoustic data based on an analysis-by-synthesis approach,” *IEEE Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, p. 23, 2014.

[13] B. Parrell, V. Ramaranarayanan, S. Nagarajan, and J. Houde, “The facts model of speech motor control: Fusing state estimation and task-based control,” *PLoS computational biology*, vol. 15, no. 9, p. e1007321, 2019.

[14] M. Murakami, B. Krüger, P. Birkholz, and J. Triesch, “Seeing [u] aids vocal learning: Babbling and imitation of vowels using a 3d vocal tract model, reinforcement learning, and reservoir computing,” in *2015 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. IEEE, 2015, pp. 208–213.

[15] A. Xu, P. Birkholz, and Y. Xu, “Coarticulation as synchronized dimension-specific sequential target approximation: An articulatory synthesis simulation,” in *Proceedings of The 19th International Congress of Phonetic Sciences (Melbourne)*, 2019.

[16] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey et al., “The kaldi book,” *Cambridge university engineering department*, vol. 3, p. 75, 2006.

[17] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz et al., “The kald ﬀ speech recognition toolkit,” in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[18] Y. Gao, S. Stone, and P. Birkholz, “Articulatory copy synthesis based on a genetic algorithm,” *Proc. Interspeech 2019*, pp. 3770–3774, 2019.

[19] P. Birkholz, “Modeling consonant-vowel coarticulation for articulatory speech synthesis,” *PloS one*, vol. 8, no. 4, 2013.

[20] P. Birkholz, D. Jackel, and B. J. Kröger, “Construction and control of a three-dimensional vocal tract model,” in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1. IEEE, 2006, pp. I–I.

[21] P. Birkholz and B. J. Kröger, “Simulation of vocal tract growth for articulatory speech synthesis,” in *Proceedings of the 16th international congress of phonetic sciences*, 2007, pp. 377–380.

[22] S. Lee, A. Potamianos, and S. Narayanan, “Analysis of children’s speech: Duration, pitch and formants,” in *Fifth European Conference on Speech Communication and Technology*, 1997.

[23] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, “Librosa: Audio and music signal analysis in python,” in *Proceedings of the 14th python in science conference*, vol. 8, 2015.

[24] N. Ryan, M. Slaney, M. Liberman, E. Shriberg, and J. Yuan, “Highly accurate mandarin tone classification in the absence of pitch information,” in *Proceedings of Speech Prosody*, vol. 7, 2014.

[25] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, “webmushra comprehensive framework for web-based listening tests,” *Journal of Open Research Software*, vol. 6, no. 1, 2018.