DNN-BASED PARAMETRIC SPEECH SYNTHESIS ENHANCED WITH ARTICULATORY INFORMATION

Anastasia Tsukanova* *Ioannis K. Douros**† *Yves Laprie**

* Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
† Université de Lorraine, INSERM U1254, IADI, F-54000 Nancy, France

The objective of this work is to build a DNN-based parametric French speech synthesizer augmented automatically with articulatory information: 10 articulatory parameters from real-time magnetic resonance imaging (rtMRI) mid-sagittal frames.

The source data was an rtMRI subset of Art-SpeechMRIfr corpus [1], with the mid-sagittal section of two speakers delivering prompted and spontaneous speech: 59.53 min. of speech for speaker S_1 (eventually split into 917 utterances) and 52.65 min. for S_2 (618).

We processed the images with bilateral filter smoothing and adaptive thresholding. The resulting steps are shown in Fig. 1. Then we identified image-dependent windows containing the lips and the velum (Fig. 2a, 2d).

The signals were aligned with phonetic and other linguistic information in force-aligned HTS labels [2, 3].

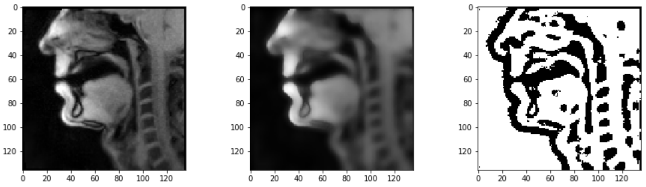
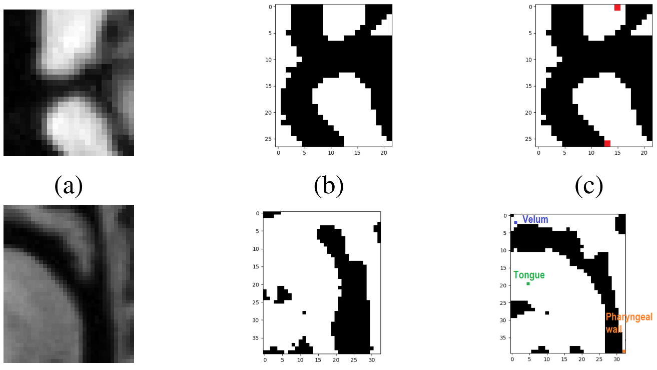


Fig. 1: Original, filtered and thresholded rtMRI.

We analyzed the (typically numerous and irregular) contours detected in the filtered windows Fig. 2 (topologically, Fig. 2c, and spatially) and extracted 4 values corresponding to the lips and 6 to the velum and its vicinity through spatial and topological analysis of the contours.

The lip parameters encode their opening (ls_dist), contact surface (ls_cont) and protrusion (up-, lw_l_protr).

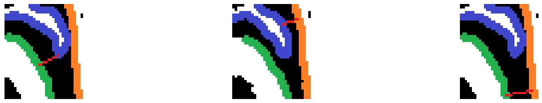
The tongue, velum and pharyngeal wall parameters reflect the constriction between the velum and the pharyngeal wall (v_w_dist for the distance, and boolean v_w_cont for the contact), which defines nasality; between the velum and the tongue (t_v_dist, t_v_cont), as place of articulation; and between the tongue and the pharyngeal wall (t_w_dist, t_w_cont), as circumstantial evidence of the front/back position of the tongue.



(a) (b) (c)

(d) (e) (f)

Fig. 2: Lips (2a) and velum (2d) windows of the frames; their filtering (2b, 2e); seed points (2c, 2f).

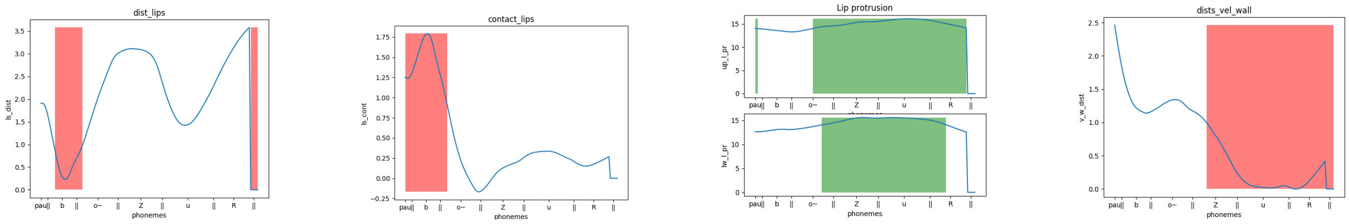


$v\_t\_dist = 7.62$ $v\_w\_dist = 5.10$ $t\_w\_dist = 10.05$

Fig. 3: v_t_dist, v_w_dist and t_w_dist computed as the minimal distances between the respective articulators.

All parameter sequences were filtered to exclude values too far from their recent predecessors; then, upsampled with interpolation. A value-label consistency test showed precision from 81.59% to 97.70%.

The speech was synthesized in two setups: `no art` without articulatory parameters and `full art` with them. It was done with a standard (`no art`) and modified (`full art`) “build your own voice” recipe in Merlin [2] using WORLD. `full art` transformed the acoustic model into an articulatory-acoustic one. Fig. 4



(a) ls_dist (b) ls_cont (c) $up\_$, lw_l_protr (d) v_w_dist

Fig. 4: Synthesis of "bonjour" /bɔ̃ʒuʁ/. The lip closure (4a, 4b in red) is consistent with the production of the labial stop /b/ and the narrowed labial opening for /u/ and with open lips throughout the rest of the utterance. The lips are correctly protruded (4c, green) at /ɔ̃/ and /u/. Nasality (4d, red for oral sound) is correct for all phonemes but /b/.

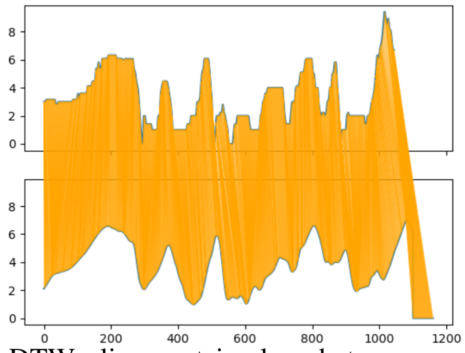


Fig. 5: DTW alignment is close between an original and generated ls_dist sequences (above and below correspondingly) "Il éblouit le veau et les pioupious qui sautaient à une encablure du Cher" , $DTW = 41.14$.

shows an example of generated articulatory parameters.

We evaluated the system acoustically, by labeling-value consistency and by aligning synthetic sentences never seen in training to their real recordings.

Acoustic evaluation (mean mel-cepstrum distortion, band aperiodicity prediction error, root mean square error of F0, F0 correlation coefficient, frame-level voiced/unvoiced error) shows that the acoustic `no art` model handled the additional articulatory parameters—forming `full art`—reasonably well, though in most cases, the `full art` setup gets slightly worse values.

The label-value consistency of the generated parameter sequences follows that of the corpus. The major issue is attaining articulatory contacts.

We trained 10 instances of the model with the input data where all samples containing a test sentence were left out, and checked the difference between the generated parameter sequences and those in the original corpus with dynamic time warping (DTW). Fig. 5 shows an example of aligned parameter sequences. While the DTW distances between the synthesized sequences and the original ones are higher than those between the original ones themselves, the variance is very high too.

To conclude, we presented a DNN-based articulatory speech synthesizer, where articulation was extracted completely automatically at a comprehensive rate and modeled jointly with the acoustics of the aligned signal. This added articulatory information fully describes nasality, the midsagittal behavior of the lips and partially the layout of the velar region. The risk to extract it incorrectly due to movement and reduced image quality at the back of the vocal tract was addressed by post-processing, and label consistency statistics suggest a reasonable accuracy.

This work is a step towards synthesizing articulation together with, and just like, acoustics in parametric speech synthesis, which has all the potential of traditional applications of articulatory speech synthesis and could help with coarticulation in purely acoustic speech synthesis. In the future, it would be useful to complete the study by experimenting with other neural network types, especially LSTMs and BLSTMs known to excel in speech synthesis thanks to their improved management of temporal relations, and fine-tuning them.

1. REFERENCES

[1] I. Douros, J. Felblinger, J. Frahm, K. Isaieva, A. A. Joseph, Y. Laprie, F. Odille, A. Tsukanova, D. Voit, and P.-A. Vuissoz, "A multimodal real-time MRI articulatory corpus of French for speech research," in *InterSpeech-20th Annual Conference of the International Speech Communication Association*, 2019.

[2] Zhizheng Wu, Oliver Watts, and Simon King, "Merlin: An open source neural network speech synthesis system," *Proc. SSW, Sunnyvale, USA*, 2016.

[3] S. Roekhaut, S. Brognaux, R. Beaufort, and Th. Dutoit, "eLite-HTS: Un outil TAL pour la génération de synthèse hmm en français," in *Démonstration aux Journées d'étude de la parole (JEP)*, 2014.