# DNN-BASED PARAMETRIC SPEECH SYNTHESIS ENHANCED WITH ARTICULATORY INFORMATION

*Anastasia Tsukanova*[⋆] *Ioannis Douros*[⋆†] *Yves Laprie*[⋆]

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

## ABSTRACT

We present a **DNN-based parametric French speech synthesizer augmented with articulatory information**. Developing it comprises automatic extraction of **10 articulatory parameters** from real-time magnetic resonance imaging **(RT-MRI)** mid-sagittal frames encoding the configuration of the **lips, nasality and constrictions between the velum and the tongue and the tongue and the pharyngeal wall**. After evaluating the parameters' consistency with phonetic labeling, we add them to a standard Merlin's "build your own voice" speech synthesis implementation trained on denoised RT-MRI speech recordings. The articulatory speech synthesizer is evaluated through the **comparison to an identically trained purely acoustic speech synthesizer** and through **dynamic time warping between original articulatory parameter sequences and the corresponding synthetic ones**. We conclude that the model copes with additional parameters well and that the generated articulatory parameter sequences match those from the corpus acceptably closely, though struggle more at attaining a contact between the articulators.

## NEEDS

For such applications like language learning and speech impairment treatment and for foundational studies of speech production, when adding articulation, we need:

## DATA. PROCESSING

- Comprehensive information about the vocal tract;
- Good temporal coverage;
- A solid relation between the position of the articulators and the produced sound.

The source data was an rtMRI subset of Art-SpeechMRIfr corpus [1]. We processed the images with bilateral filter smooth-ing and adaptive thresholding (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.

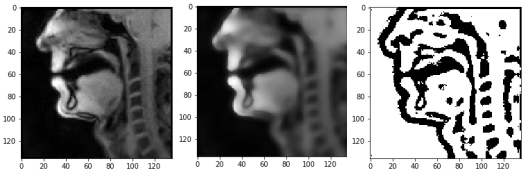


**Fig. 1**: An RT-MRI frame, from left to right: original, smoothed with a bilateral filter and adaptive thresholding.

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 param-eters 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*),

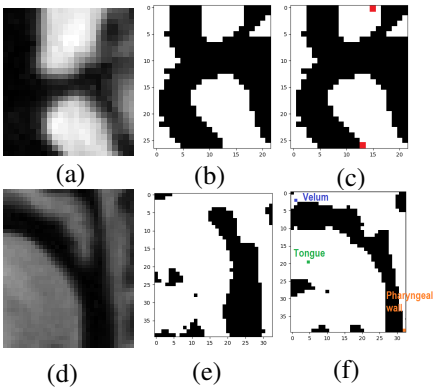



**Fig. 2**: Lips (2a) and velum (2d) windows of the frames; their filtering (2b, 2e); seed points (2c, 2f).

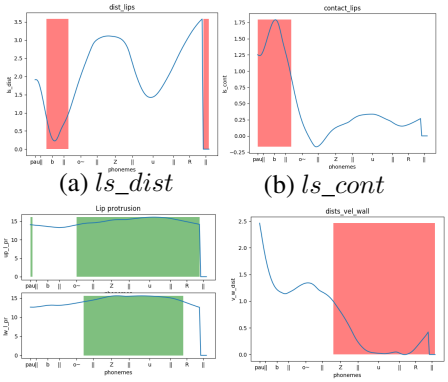
*v_t_dist* = 7.62 *v_w_dist* = 5.10 *t_w_dist* = 10.05

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

as place of articulation; and between the tongue and the pharyngeal wall (t w dist, t w cont), circumstan-tial evidence of the front/back position of the tongue.

All parameter sequences were filtered to exclude values too far from their recent predecessors; then, up-sampled 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 shows an example of generated articulatory parameters.




(a) *ls_dist* (b) *ls_cont*

(c) *up , lw_l_ protr* (d) *v_w_ dist*

**Fig. 4**: Synthesis of "bonjour". 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). Nasality (4d, red for oral sound) is correct for all phonemes but /b/.

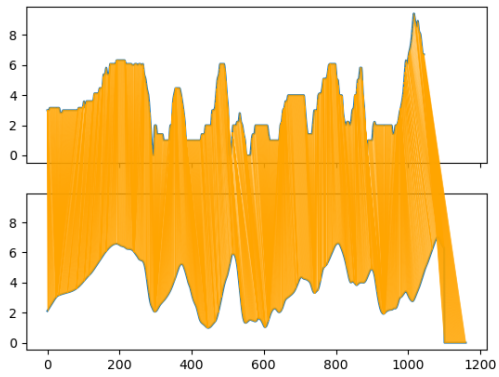


**Fig. 5**: DTW alignment is close between an original (above) and generated (below) *ls_dist* sequences "Iléblouit le veau et les pioupious qui sautaient à une enca-blure du Cher" , DTW = 41.14.

## EVALUATION

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

Acoustic evaluation shows that the acoustic no art model handled the additional articulatory parameters, making full art, reasonably well, though in most cases, the full art setup takes slightly worse values.

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

We trained ten instances of the model with the same input data, but all samples containing a test sentence out, and checked the difference between the generated pa-rameter sequences and those in the original corpus with dynamic time warping (DTW). Fig. 5 shows an exam-ple of aligned articulatory 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 ex-tracted fully automatically at a comprehensive rate and modeled jointly with the acoustics of the aligned sig-nal. This added articulatory information fully describes nasality, the midsagittal behavior of the lips and par-tially 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 articula-tion 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 with experimenting with other neural network types, especially LSTMs and BLSTMs known to excel in speech synthesis thanks to their improved manage-ment of temporal relations, and fine-tuning them.