Towards an articulatory-driven neural vocoder for speech synthesis

Marc-Antoine Georges[1,2], *Pierre Badin*[1], *Julien Diard*[2], *Laurent Girin*[1], *Jean-Luc Schwartz*[1], *Thomas Hueber*[1]

[1]Univ. Grenoble Alpes, CNRS, GIPSA-lab, 38000 Grenoble, France.
[2]Univ. Grenoble Alpes, CNRS, LPNC, 38000 Grenoble, France.

marc-antoine.georges@gipsa-lab.grenoble-inp.fr

1. Context

High-quality articulatory synthesis is increasingly required for both fundamental objectives, such as better understanding speech production and speech development, and applications that require to relate gestures and sounds, such as teaching, handicap remediation or augmented reality [1].

One possible approach to build such synthesizers is to exploit datasets containing "parallel" articulatory-acoustic data, i.e., speech sounds recorded simultaneously with the movements of the main speech articulators (tongue, lips, jaw, velum) using specific motion capture systems such as electro-magnetic articulography (EMA). The complex and non-linear relationship between the articulatory configuration and the spectral envelope of the corresponding speech sound is learned using supervised machine learning techniques such as Gaussian Mixture Models [2], Hidden Markov Models [3, 4], or Deep Neural Networks (DNN) [5]. The acoustic signal is finally synthetized by deriving an autoregressive filter (e.g., the MLSA vocoder) from the predicted spectral envelope, and exciting this filter with a source signal encoding the glottal activity.

We claim that an articulatory synthesizer built following this approach has two main drawbacks. First, its input control parameters (i.e., 2D or 3D coordinates of EMA-coils) are not articulatory parameters *per se* in the sense that they do not control explicitly the degrees of freedom of the vocal apparatus (i.e., the limited set of movements that each articulator can execute independently from the other articulators). Second, the synthesized speech sounds muffled. This is likely due to the vocoding process and the quality of the excitation signal.

This study aims at addressing these two issues. We propose a new approach for building a synthesizer driven by explicit articulatory parameters and able to produce high-quality speech sounds. This work relies on recent developments on so-called neural vocoders. A neural vocoder is a deep autoregressive neural network that synthesizes a sequence of time-domain signal samples. Neural vocoders such as WaveNet or LPCNet [6] have recently led to an impressive gain of performance in text-to-speech synthesis. Here, we propose to drive such a vocoder by a set of articulatory parameters. An overview of the proposed system is shown in Fig 1. The following paragraphs describe the different processing steps.

2. Method

Data The core of the proposed synthesizer is an articulatory-to-acoustic statistical model built from synchronized EMA and audio recordings of a "reference speaker." Articulatory data were recorded using the Carstens 2D EMA system (AG200). Six coils were glued on the tongue tip (*tip*), blade (*mid*) and dorsum (*bck*), as well as on the upper lip, the lower lip and the jaw (lower incisor). The recorded database consisted of 1,109 items in total (sustained vowels, VCV, words, sentences), which overall correspond to approximately 20 minutes of speech (after removing silences). EMA trajectories were low-pass filtered at 20 Hz and downsampled from 200 Hz to 100 Hz. Audio recordings were characterized using 20 features extracted using a 20-ms sliding analysis window with 10-ms frame shift, as done in the original implementation of the LPCNet vocoder [6]: a set of "spectral features" composed of 18 Bark-scale cepstral coefficients, and two "source features," which are the fundamental frequency f0 and a measure of the periodicity of the signal.

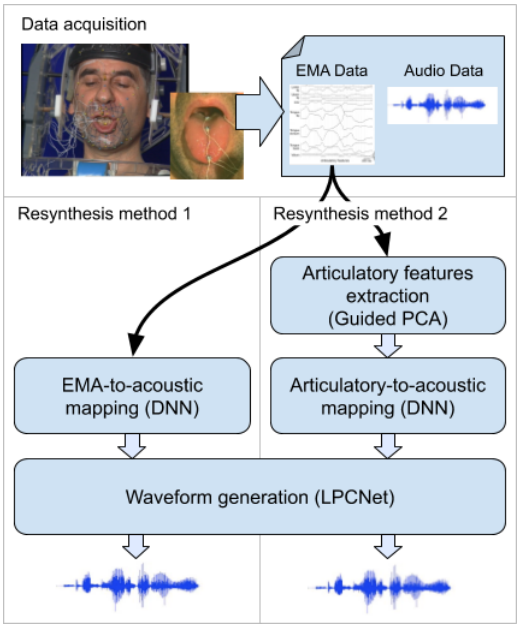

Figure 1: ***System overview.*** *Two methods for synthesis from articulatory data we investigated: Method 1 predicts a target cepstrum from raw EMA data, i.e., a 12-dimensional feature vector, while Method 2 first reduces it to a vector of 6 articulatory components (JH, TB, TD, TT, LH, LP).*

Articulatory features extraction First, a 6-parameter articulatory model was built from the 2D EMA coordinates over the whole corpus using guided Principal Component Analysis, as in [7]. For the tongue, the first component is controlled by parameter jaw height JH, extracted by PCA from the coordinates of the EMA coil attached to the lower incisor. Its contribution to tongue movements was then estimated using linear regression and removed from these data. The second and third component are controlled by parameters tongue body TB and tongue dorsum TD, extracted by PCA from the residue of the coordinates of the rear two EMA tongue coils. The contribution of TB and TD to the tongue tip movements was then estimated using linear regression of all the tongue coordinates against TB and TD, and removed from these data. The fourth component is controlled by the tongue tip parameter TT extracted by PCA on the residues of all the tongue coordinates. An example of extracted parameters is given in Fig 2.

A similar procedure was used to derive lip height (LH) and lip protrusion (LP) parameters to control the lips, in addition to the common jaw parameter. More details can be found in [7].

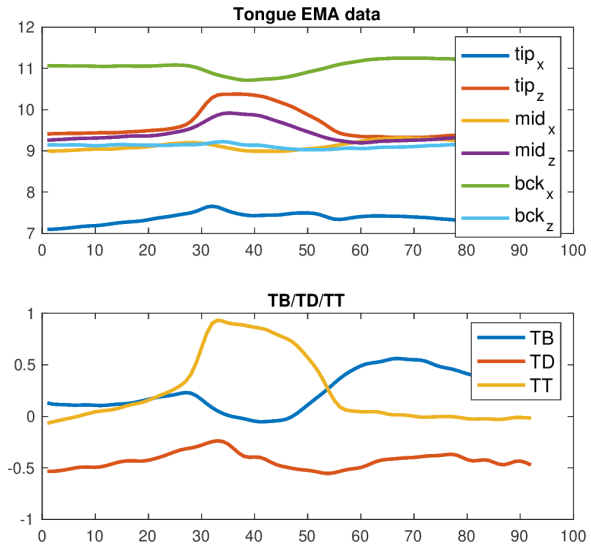

Figure 2: *Trajectories of the 3 EMA coils attached to the tongue (tip, mid, bck, top) while producing the sequence /ala/ and the corresponding articulatory features: Tongue Body (TB), Tongue Dorsum (TD) and Tongue Tip (TT), bottom.*

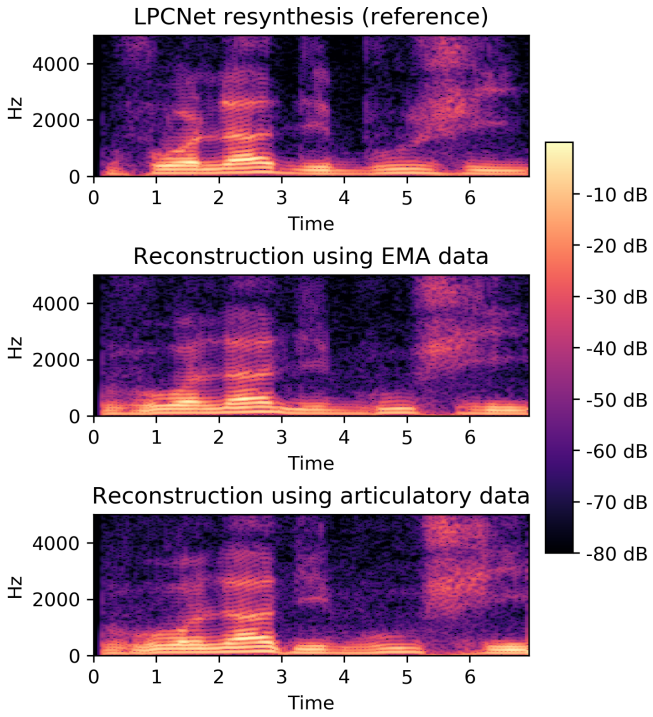

Figure 3: *Spectrograms of the reference and reconstructed signals for the sentence "Voilà des bougies" ("Here are some candles").*

Articulatory-to-acoustic mapping In the proposed approach, a DNN is used to map a vector of articulatory features (i.e., either raw 2D coordinates of EMA coils or articulatory parameters derived from them) to a vector of spectral features (18 Bark-scale cepstral coefficients here). The model used in the present study is composed of 4 fully-connected layers of 512 neurons each (this architecture was selected after preliminary tests on a subset of the database). The hyperbolic tangent was used as activation function for the neurons of the hidden layers. This model was trained using back-propagation with Adam optimizer, on mini-batches of 32 observations. The mean squared error (*MSE*) was used as loss function. In each experiment, 80% of the data (randomly partitioned) were used for training, the remaining 20% were used for testing. 20% of the training data were used for validation (early-stopping). Batch normalization and dropout were also used. All experiments were implemented using the *Keras* toolkit (*https://keras.io*).

Waveform generation Spectral features estimated from articulatory parameters, combined with source parameters (f0 and periodicity) directly extracted from the original signal, are finally fed into a neural vocoder to generate the speech waveform. In the present study, we used the LPCNet neural vocoder [6]. The explicit dissociation of source (f0 and periodicity) and filter (cepstral coefficients) acoustic features makes it well suited to be interfaced with an articulatory model. Starting from an existing version trained on a large acoustic database, we adapted the model parameters to the voice of the reference speaker.

3. Results

We report here preliminary experiments conducted to assess the performance of the proposed articulatory neural speech synthesizer. Here, the reference signal is defined as the analysis-and-resynthesis of an original speech signal (extracted from the test dataset) by the LPCNet vocoder. For each reference signal (REF), we synthesized two other signals: one by using the 12 EMA features as input parameters (EMA-SYN), the other by using the corresponding 6 articulatory features (ART-SYN). Examples of reference and synthesized signals are displayed in Fig 3. Other sound examples are available at (*https://georges.ma/publications/issp2020-abstract/*). For all items in the test dataset, we calculated the PEMO-Q score [8] between REF and EMA-SYN on one hand, and between REF and ART-SYN on the other hand. The average scores on the test dataset were respectively 0.84 ± 0.06 and 0.81 ± 0.07. Initial results are promising and tend to show that the proposed system is able to generate intelligible speech from a few number of articulatory and glottal control parameters.

4. Acknowledgement

This work has been partially supported by MIAI @ Grenoble Alpes (ANR-19-P3IA-0003).

5. References

[1] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Herff, and J. S. Brumberg, "Biosignal-Based Spoken Communication: A Survey," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2257–2271, 2017.

[2] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a Gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215–227, 2008.

[3] T. Hueber and G. Bailly, "Statistical conversion of silent articulation into audible speech using full-covariance HMM," *Computer Speech and Language*, vol. 36, pp. 274–293, 2016.

[4] H. Zen, Y. Nankaku, and K. Tokuda, "Continuous stochastic feature mapping based on trajectory HMMs," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 417–430, 2011.

[5] F. Bocquelet, T. Hueber, L. Girin, C. Savariaux, and B. Yvert, "Real-Time Control of an Articulatory-Based Speech Synthesizer for Brain Computer Interfaces," *PLoS Computational Biology*, vol. 12, no. 11, p. e1005119, 2016.

[6] J.-M. Valin and J. Skoglund, "LPCNet: Improving neural speech synthesis through linear prediction," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5891–5895.

[7] A. Serrurier, P. Badin, A. Barney, L.-J. Boë, and C. Savariaux, "The tongue in speech and feeding: Comparative articulatory modelling," *Journal of Phonetics*, vol. 40, no. 6, pp. 745–763, nov 2012.

[8] R. Huber and B. Kollmeier, "Pemo-q—a new method for objective audio quality assessment using a model of auditory perception," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1902–1911, 2006.