Articulatory Speech Synthesis from Static Context-Aware Articulatory Targets

Anastasiia Tsukanova, Benjamin Elie, Yves Laprie

Articulatory Speech Synthesis from Static Context-Aware Articulatory Targets

Anastasiia Tsukanova, Benjamin Elie, Yves Laprie

Université de Lorraine, LORIA, UMR 7503, Vandoeuvre-lès-Nancy, F-54506, France
Inria, Villers-lès-Nancy, F-54600, France
CNRS, LORIA, UMR 7503, Vandoeuvre-llès-Nancy, F-54506, France
anastasiia.tsukanova@inria.fr, benjamin.elie@inria.fr, yves.laprie@loria.fr

Abstract

The aim of this work is to develop an algorithm for controlling the articulators (the jaw, the tongue, the lips, the velum, the larynx and the epiglottis) to produce given speech sounds, syllables and phrases. This control has to take into account coarticulation and be flexible enough to be able to vary strategies for speech production. The data for the algorithm are 97 static MRI images capturing the articulation of French vowels and blocked consonant-vowel syllables. The results of this synthesis are evaluated visually, acoustically and perceptually, and the problems encountered are broken down by their origin: the dataset, its modeling, the algorithm for managing the vocal tract shapes, their translation to the area functions, and the acoustic simulation.

Keywords: articulatory synthesis, coarticulation, articulatory gestures

1. Introduction

Articulatory speech synthesis is a method of synthesizing speech by managing the vocal tract shape on the level of the speech organs, which is an advantage over the state-of-the-art methods that do not usually incorporate any articulatory information. The vocal tract can be modeled with geometric (Öhman 1966; Birkholz, Jackèl, and Kröger 2006; Story 2013), biomechanical (Lloyd, Stavness, and Fels 2012; Anderson et al. 2015) and statistical (Maeda 1990; Howard and Messum 2011) models. The advantage of statistical models is that they use very few parameters, speeding up the computation time. Their disadvantage is that they follow the data a priori without any guidance and do not have access to the knowledge of what is realistic or physically possible. Because of this, to produce correct configurations, they need to be finely tuned.

Previous work performed by Tsukanova (2016) explored the potential in using quite little, and yet sufficient, static magnetic resonance imaging (MRI) data and implemented one of the few existing attempts at creating a full-fledged speech synthesizer that would be capable of reproducing the vast diversity of speech phenomena. This work is its follow-up, with improvements in the statistical articulatory model (better fitting of the articulators), articulation strategies (more fine-grained timing control over separate articulators rather than their ensemble) and acoustic simulation (operating the vocal folds by their relative opening rather than by the pressure and voicing controls, and more work on the production of fricatives and trills).

2. Building an articulatory speech synthesis system

The system is basically made up of three components: the database with the “building blocks" for articulating utterances, the joint control algorithm for the vocal tract and the glottal source, and acoustic simulation. The primary concern of this work are the first two components.

2.1. Dataset

The dataset construction and manipulation were inspired by the work of Birkholz (2013). We used static MRI data, 97 images of the midsagittal section, capturing articulation without phonation: the speaker was instructed to show the position that he would have to attain to produce a particular sound. For vowels, that is the position when the vowel would be at its clearest if the subject were phonating. For consonant-vowel (CV) syllables, that is the blocked configuration of the vocal tract, as if the subject were about to start pronouncing it. The assumption is that such articulation shows the anticipatory coarticulation effects of the vowel V on the consonant C preceding it. There were 13 vowels, 72 CV syllables and 2 semi-vowels in the final dataset. This covers all main phonemes of the French language, but not in all contexts.

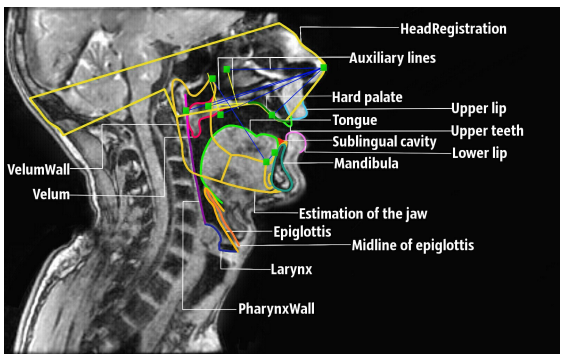

Figure 1: *An example of dataset image annotation (/a/).*

After manually annotating the captures as shown in Fig. 1 we applied a principal-component-analysis (PCA)-based model on the articulator contours (Laprie and Busset 2011; Laprie, Vaxelaire, and Cadot 2014; Laprie, Elie, and Tsukanova 2015). Within that model, the jaw is represented by 3 parameters, the tongue by 12, the lips by 3, the velum by 5, the larynx by 3 and the epiglottis by 3, in total forming a vector from $\mathbb{R}^{29}$ (see Fig. 2 for major parameter contributions to the articulator shape). Since the model uses PCA, the zero configuration

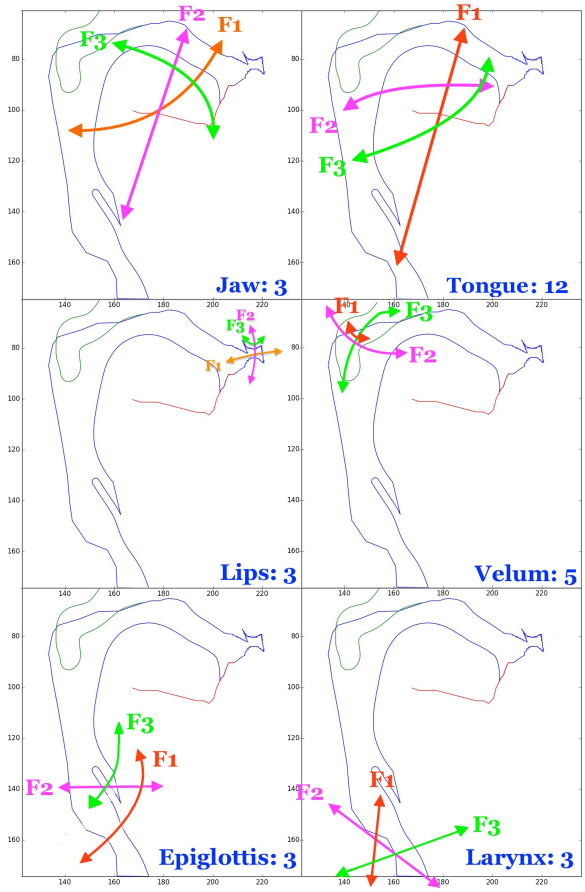

Figure 2: *The PCA-based articulatory model: curve change directions encoded in the first three factors of each articulator (the jaw, the tongue, the lips, the epiglottis, the larynx).*

should correspond to the central position as identified in the dataset, and small changes in the parameter space within a certain neighborhood of zero should correspond to small changes (in terms of distance and shape, not in terms of the resulting acoustics) in the curves. However, there is inter-articulator correlation taken into account as well as various boundary conditions imposed, so the model's behavior is not entirely linear. This non-linear behavior is especially evident in the lips because their shape defines where the vocal tract should end, and while a change may be minor in the articulatory parameter space, it can result in cutting off the vocal tract at quite different positions.

2.1.1. Expanding the dataset

Since the collected French phonemic dataset was limited, we needed to expand it to cover other contexts as well. We used the notion of the cardinal vowels—/a/, /i/ and /u/,—assuming that /a/, /i/ and /u/ represent the most extreme places of vowel articulation, and since then any other vowel articulation can be expanded as a combination of its /a/, /i/ and /u/ "components". Having captured the C+/a/, C+/i/ and C+/u/ context for all consonants C and all non-cardinal vowels V on their own, we were able to estimate the missing C+V samples:

- We projected the vowel V articulatory vector (from $\mathbb{R}^{29}$) onto the convex hull of the /a/, /i/ and /u/ vectors.
- Assuming that the linear relationship between the C+V vector and the C+/a/, C+/i/ and C+/u/ vectors is the same as the one between V and /a/, /i/ and /u/, we estimated C+V from C+/a/, C+/i/ and C+/u/ using the coefficients from the projection of V onto the convex hull of /a/, /i/ and /u/.

We also estimated the pure C configuration as the average of C+/a/, C+/i/ and C+/u/.

Finally, we assumed that the voiced and unvoiced consonants did not have any differences in the articulation.

2.2. Strategies for transitioning between the articulatory targets

The dataset provided static images capturing idealistic, possibly over-articulated, targets for consonants anticipating particular vowels, whereas the goal was to be also able to deal with consonant clusters and consonants that would not anticipate any vowel at all—for example, due to their ultimate position in a rhythmic phrase. This is why we imposed several restrictions on the anticipatory effect:

- Temporal: no coarticulatory effect if the anticipated phoneme is more than 200 ms ahead;
- Spatial: if there is any movement scheduled between the anticipated vowel, the phoneme in question negates the effect. For example, consider such sequence as /lka/: after /l/, the tongue needs to move backward to produce /k/ before coming back forward for /a/. In this situation, our algorithm does not allow the /l/ to anticipate the coming /a/;
- Categorical: it is not possible to anticipate a vowel more than 5 phonemes ahead, and this restriction becomes stricter if it applies across syllable boundaries.

For vowels, there is also a model of target undershoot.

Having established the articulatory targets, the question is how to transition between them. We have tested out three strategies for interpolation between the target vectors:

- Linear: the interpolation between the target vectors is linear;
- Cosine;
- Complex: transitions vary by the articulators. The critical ones reach for their target position faster than the others, while those articulators whose contribution to the resulting sound intelligibility is not as large move slower. Besides, the articulators composed of heavier tissues (such as the tongue back) move slower than the light and highly mobile ones (such as the lips).

2.3. Obtaining the sound

Each vocal tract position was encoded in an area function. They were obtained by the algorithm of Heinz and Stevens (1965) with coefficients adapted by Shinji Maeda and Yves Laprie. Considering the presence of central and lateral phonemes in the French language and conflicting evidence in the literature (Soquet et al. (2002) and McGowan, Jackson, and Berger (2012)), though, it is quite probable that these coefficients need to be changed dynamically.

We used an acoustic simulation system implemented by Elie and Laprie (2016b) to obtain sound from the area functions and supplementary control files: glottal opening and pitch control.

Glottal opening was modeled by using external lighting and sensing photo-glottography (ePGG) measurements (Honda and Maeda 2008). Within the model, glottis is at its most closed

position when producing vowels, nasals and the liquid sound /l/, and momentarily reaches its most open one when producing voiceless fricatives and stops. Voiced fricatives and stops also create peaks in glottal opening, but not as high.

There was no need to model voicing (high-frequency oscillations of low amplitude superimposed onto the glottal opening waves) since the vocal folds operated by the glottal chink model (Elie and Laprie 2016b; Elie and Laprie 2016a) are self-oscillating.

3. Evaluation

Each step in the system was evaluated on its own, and afterwards the synthesis results were evaluated visually, acoustically and perceptually.

3.1. The articulatory model and the trajectories

One peculiarity of the dataset and therefore of the model was the fact that it used only the sagittal section of the speaker's vocal tract. While full three-dimensional models can capture the full geometry of the vocal tract with such phenomena as lateral phonemes (e.g. /l/), two-dimensional models get the benefit of faster computation time and overall simplicity, but irreversibly lose the spatial information.

In general, the articulatory model captured vocal tract positions correctly or with no critical errors, and some adjustments could be necessary only at the points of constriction, since on its own the model did not impose much control over them. This is a disadvantage brought by the nature of the articulatory model: choosing to operate at the level of articulators rather than the resulting vocal tract geometry.

As for the movements, for now we can say that they were reasonable and the coarticulation-affected targets guided the articulators to the positions necessary to produce a particular utterance. One key point here is the timing strategy. Rule-based timing strategy seems to be very rigid for the dynamic nature of speech; it would be more natural to follow speech production processes in humans and to guide the synthesis with the elicited sound or the speaker's expectation—based on their experience—on what this sound will be. We plan to evaluate the transitions with new dynamic MRI data.

3.2. Glottal opening control

The algorithm for the glottis opening successfully allowed to distinguish between vowels and consonants. Distinguishing between voiced and voiceless consonants, though, stays a point for improvement, as well as well-coordinated control over the glottis and the vocal tract.

3.3. The synthesized sound

Vowels and stops were the most identifiable and correct, although sometimes some minor adjustments in the original data were necessary to obtain formants close to the reference values. When compared to human speech, the formant transitions within the suggested strategies sometimes occurred too fast and sometimes too slowly; again, this highlights the utmost importance of realistic timing strategies. Fig. 3 shows an example of the synthesis when it is not guided by real timing: /aʃa/ as produced by the system and as uttered by a human. The high-frequency contributions in vowels, not appearing in the human sample, are due to the acoustic simulation. The noise of /ʃ/ is at the correct frequencies, but with a bit different energy distribution, probably because of differences in articulation or in the area functions. There are also differences in phoneme duration: the synthesized /ʃ/ is shorter, and both instances of /a/ are longer. It does not mean that the synthesized version is wrong; one human sample does not have to be the only correct way to say this. Nevertheless it could lead a synthesis result to sound unnatural.

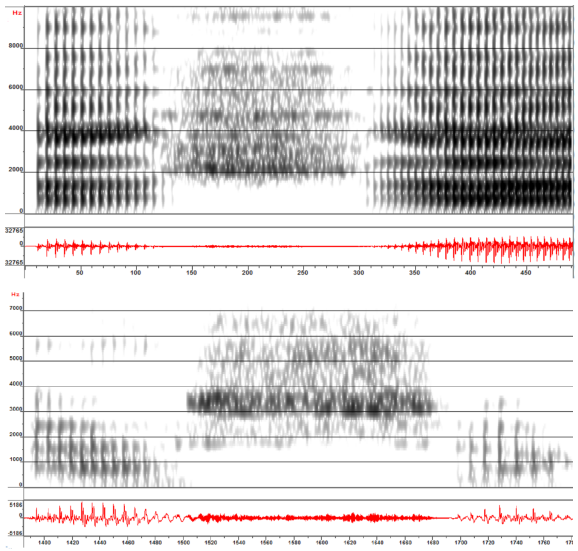

Figure 3: *An example of synthesis of /aʃa/ (above) and its utterance by a human (below).*

4. Conclusion

Regarding speech as a process of transitioning between context-aware targets is an interesting approach that can be connected with the mental processes of speech production: to allow the others to perceive the necessary acoustic cue, the speaker needs to come close enough to the associated articulatory goal. The important difference between a real speaker and the algorithm is the fact that the algorithm solves a static problem, laid out in full; it needs to hit particular targets in a given order. As for humans, we solve a dynamic problem, and coarticulation is not something we put in its definition; rather, coarticulation is our means to make the problem of reaching too many targets in a too short period of time solvable.

The statistically derived articulatory model encodes complicated shapes of the articulators in only 29 parameters, sometimes struggling at the constrictions because of the inherent—and intentional—lack of control over the resulting geometry of the vocal tract.

Those shapes of the articulators change in time according to the produced trajectories of the vocal tract, and those are phonetically sound. Whether there are any important differences between the produced transitions and the ones in real speech, needs to be verified with actual dynamic data.

After the aspect of *how* the articulators move we need to consider *when*. The timing strategies, currently rule-based, apparently need to be extracted from dynamic data, and we can use the approaches by Elie et al. (2016) for that.

A closer, intertwined interaction with the acoustic simulation unit—such as guidance on how to navigate between the area functions at the level of separate acoustic tubes and im-

proved control over the glottal opening—could improve the results for consonants.

5. Acknowledgments

The data collection for this work benefited from the support of the project ArtSpeech of ANR (Agence Nationale de la Recherche), France.

6. References

Anderson, Peter, Negar M Harandi, Scott Moisik, Ian Stavness, and Sidney Fels (2015). "A Comprehensive 3D Biomechanically-Driven Vocal Tract Model Including Inverse Dynamics for Speech Research". In: *Sixteenth Annual Conference of the International Speech Communication Association*.

Birkholz, P., D. Jackèl, and B. J. Kröger (2006). "Construction and control of a three-dimensional vocal tract model". In: *Proc. Intl. Conf. Acoust., Spch., and Sig. Proc. (ICASSP 2006)*, pp. 873–876.

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

Elie, Benjamin and Yves Laprie (2016a). "A glottal chink model for the synthesis of voiced fricatives". In: *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, pp. 5240–5244.

— (2016b). "Extension of the single-matrix formulation of the vocal tract: Consideration of bilateral channels and connection of self-oscillating models of the vocal folds with a glottal chink". In: *Speech Communication* 82, pp. 85–96.

Elie, Benjamin, Yves Laprie, Pierre-André Vuissoz, and Freddy Odille (2016). "High spatiotemporal cineMRI films using compressed sensing for acquiring articulatory data". In: *Eusipco, Budapest*, pp. 1353–1357.

Heinz, John M. and Kenneth N. Stevens (1965). "On the relations between lateral cineradiographs, area functions and acoustic spectra of speech". In: *Proceedings of the 5th International Congress on Acoustics*, A44.

Honda, Kiyoshi and Shinji Maeda (2008). "Glottal-opening and airflow pattern during production of voiceless fricatives: a new noninvasive instrumentation". In: *The Journal of the Acoustical Society of America* 123.5, pp. 3738–3738.

Howard, Ian S and Piers Messum (2011). "Modeling the development of pronunciation in infant speech acquisition". In: *Motor Control* 15.1, pp. 85–117.

Laprie, Yves and Julie Busset (2011). "Construction and evaluation of an articulatory model of the vocal tract". In: *19th European Signal Processing Conference - EUSIPCO-2011*. Barcelona, Spain.

Laprie, Yves, Benjamin Elie, and Anastasiia Tsukanova (2015). "2D articulatory velum modeling applied to copy synthesis of sentences containing nasal phonemes". In: *International Congress of Phonetic Sciences*.

Laprie, Yves, Béatrice Vaxelaire, and Martine Cadot (2014). "Geometric articulatory model adapted to the production of consonants". In: *10th International Seminar on Speech Production (ISSP)*. Köln, Allemagne. URL: http://hal.inria.fr/hal-01002125.

Lloyd, John E, Ian Stavness, and Sidney Fels (2012). "ArtiSynth: a fast interactive biomechanical modeling toolkit combining multibody and finite element simulation". In: *Soft tissue biomechanical modeling for computer assisted surgery*. Springer, pp. 355–394.

Maeda, S. (1990). "Compensatory articulation during speech: evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model". In: *Speech production and speech modelling*. Ed. by W.J. Hardcastle and A. Marchal. Amsterdam: Kluwer Academic Publisher, pp. 131–149.

McGowan, Richard S., Michel T.-T. Jackson, and Michael A. Berger (2012). "Analyses of vocal tract cross-distance to area mapping: an investigation of a set of vowel images". In: *Journal of the Acoustical Society of America* 131.1, pp. 424–434.

Öhman, S.E. (1966). "Coarticulation in VCV utterances: spectrographic measurements". In: *Journal of the Acoustical Society of America* 39.1, pp. 151–168.

Soquet, Alain, Véronique Lecuit, Thierry Metens, and Didier Demolin (2002). "Mid-sagittal cut to area function tranformations: Direct measurements of mid-sagittal distance and area with MRI". In: *Speech Communication* 36.3-4, pp. 169–180.

Story, Brad H (2013). "Phrase-level speech simulation with an airway modulation model of speech production". In: *Computer speech & language* 27.4, pp. 989–1010.

Tsukanova, Anastasiia (2016). *A Coarticulation Model for Articulatory Speech Synthesis*. Master thesis. URL: https://lct-master.org/getfile.php?id=1934&n=1&dt=TH&ft=pdf&type=TH.