What goals for articulatory speech synthesis?

Pascal Perrier

Université Grenoble Alpes, CNRS, GIPSA-Lab UMR 5216, F-38000 Grenoble, France

Introduction

In their often cited paper published in 1987 in IEEE Trans. ASSP, Sondhi & Schroeter wrote : « *Articulatory speech synthesis has several advantages as follows.*

> *a) Articulatory speech synthesis has the potential for very natural speech output at bit rates below 4800 bits/s, provided that "good" articulatory parameters are available to control the synthesizer.*
>
> *b) The control signals of articulatory speech synthesizers have a direct interpretation in terms of physiological and physical data. In the human voice production system, they vary slowly enough to be potential candidates for efficient coding.*
>
> *c) The model parameters are easier to interpolate than those of more abstract waveform or spectrum synthesizers. This is because interpolated values for the control signals of an articulatory synthesizer are physically realizable. (This is not true in general. An LPC vector interpolated between two realizable vectors might correspond to an unstable filter; interpolation of a set of formants between two reasonable sets of formants might yield a set that corresponds to an unreasonable, if not impossible, vocal tract shape, etc.) For the same reason, slightly erroneous control signals usually do not result in "unnatural'' speech.* » (Sondhi & Schroeter, 1987, p.

Thirty years after we know that not much of this happened to be true. After the advent of the TD-PSOLA technique in the 90s (Dutoit et al., 1993), and, more recently, of statistical methods of speech synthesis exploiting the power of machine learning algorithms to deal with a large body of extremely varied data (among many others Ling et al., 2013, or Ze et al., 2013), we know that the introduction of more explicit knowledge in articulatory synthesis models is not likely to compete with the massive introduction of implicit knowledge in the most recent speech synthesizers. In this context articulatory speech synthesis systems cannot be any longer considered to be useful tools toward future technological developments, except perhaps in the context of second language learning and phonetic correction (Bälter et al., 2005; Badin et al., 2010). However, articulatory speech synthesizers can be powerful tools to investigate in depth the mechanisms underlying speech production, from the neural control level to the aerodynamics of sound production, provided the models used are adapted to the purpose of the study.

In this talk I will present a summary and a perspective of the works that we have carried on at Gipsa-lab, in order to investigate the nature of the motor goals in speech production, using biomechanical models.

Method and results

Our focus was on the temporal nature of the motor goals, trying to evaluate to what extent the hypothesis telling that a series of discrete motor goals related to the phonemic structure underlies the production of a speech sequence, resists the comparison between synthetic articulatory and acoustic speech signals and similar data collected from human subjects.

Using a 2-D biomechanical model of the tongue (Payan & Perrier, 1977) we have shown that using a sequence of discrete goals specified as a set of mechanical equilibrium positions (Feldman, 1986) it is possible to generate complex realistic velocity profiles and complex articulatory patterns such as the articulatory loops observed in Vowel-VelarConsonant-Vowel sequences in different languages (Perrier et al., 2003). We have also shown that the relation between speed and trajectory curvature experimentally observed in tongue movements (Tasko & Wetsbury, 2004) could also naturally emerge in synthetic tongue movements generated from a discrete sequence of motor goals (Perrier & Fuchs, 2008). We have also shown that such a discrete specification of the motor goals is compatible with the

variability of the speech articulators' trajectories experimentally observed in vowel reduction phenomena (Patri et al., 2016).

Using a more complex 3-D biomechanical tongue model we have also shown that, provided a proper account of short-delay feedback in muscle force generation mechanisms, these discrete goals are robust enough to deal with change in the orientation of the gravity field (Buchaillard et al., 2009).

Conclusion

In conclusion, I will discuss the strengths and weaknesses of these findings for a better understanding of speech motor control in relation to the phonological specification of a speech sequence.

References

Badin, P., Tarabalka, Y., Elisei, F., & Bailly, G. (2010). Can you 'read'tongue movements? Evaluation of the contribution of tongue display to speech understanding. *Speech Communication, 52(6)*, 493-503.

Bälter, O., Engwall, O., Öster, A.-M., & Kjellström, H. (2005). Wizard-of-Oz test of ARTUR – a computer-based speech training system with articulation correction. *Proceedings of the Seventh International ACM SIGACCESS Conference on Computers and Accessibility* (pp. 36–43). Baltimore,

Buchaillard, S., Perrier, P., & Payan, Y. (2009). A biomechanical model of cardinal vowel production: Muscle activations and the impact of gravity on tongue positioning. *The Journal of the Acoustical Society of America, 126(4)*, 2033-2051.

Dutoit, T., & Leich, H. (1993). MBR-PSOLA: Text-to-speech synthesis based on an MBE re-synthesis of the segments database. *Speech Communication, 13(3-4)*, 435-440.

Feldman, A. G. (1986). Once more on the equilibrium-point hypothesis (λ model) for motor control. *Journal of Motor Behavior, 18(1)*, 17-54.

Ling, Z. H., Deng, L., & Yu, D. (2013). Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing, 21(10)*, 2129-2139.

Patri, J. F., Diard, J., & Perrier, P. (2016). Bayesian modeling in speech motor control: a principled structure for the integration of various constraints. *Proceedings of Interspeech 2016* (pp. 3588-3592). ISCA

Payan, Y., & Perrier, P. (1997). Synthesis of VV sequences with a 2D biomechanical tongue model controlled by the Equilibrium Point Hypothesis. *Speech communication, 22(2-3)*, 185-205.

Perrier, P., Payan, Y., Zandipour, M., & Perkell, J. (2003). Influences of tongue biomechanics on speech movements during the production of velar stop consonants: A modeling study. *The Journal of the Acoustical Society of America, 114(3)*, 1582-1599.

Perrier, P., & Fuchs, S. (2008). Speed–curvature relations in speech production challenge the 1/3 power law. *Journal of neurophysiology, 100(3)*, 1171-1183.

Sondhi, M., & Schroeter, J. (1987). A hybrid time-frequency domain articulatory speech synthesizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 35(7)*, 955-967.

Tasko, S. M., & Westbury, J. R. (2004). Speed–curvature relations for speech-related articulatory movement. *Journal of Phonetics, 32(1),* 65-80.

Ze, H., Senior, A., & Schuster, M. (2013). Statistical parametric speech synthesis using deep neural networks. *Proceedings of the IEEE International Conference on the Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7962-7966). IEEE.