2

Towards More Versatile Signal Generation Systems

Gérard Bailly
Institut de la Communication Parlée – UMR-CNRS 5009
INPG and Université Stendhal, 46, avenue Félix Viallet, 38031 Grenoble Cedex 1, France
bailly@icp.grenet.fr

Introduction

Reproducing most of the variability observed in natural speech signals is the main challenge for speech synthesis. This variability is highly contextual and is continuously monitored in speaker/listener interaction (Lindblom, 1987) in order to guarantee optimal communication with minimal articulatory effort for the speaker and cognitive load for the listener. The variability is thus governed by the structure of the language (morphophonology, syntax, etc.), the codes of social interaction (prosodic modalities, attitudes, etc.) as well as individual anatomical, physiological and psychological characteristics. Models of signal variability –and this includes prosodic signals – should thus generate an optimal signal given a set of desired features. Whereas concatenation-based synthesisers use these features directly for selecting appropriate segments, rule-based synthesisers require fuzzier[1] coarticulation models that relate these features to spectro-temporal cues using various data-driven least-square approximations. In either case, these systems have to use signal processing or more explicit signal representation in order to extract the relevant spectro-temporal cues. We thus need accurate signal analysis tools not only to be able to modify the prosody of natural speech signals but also to be able to characterise and label these signals appropriately.

Physical interpretability vs. estimation accuracy

For historical and practical reasons, complex models of the spectro-temporal organisation of speech signals have been developed and used mostly by rule-based

[1] More and more fuzzy as we consider interaction of multiple sources of variability. It is clear, for example, that spectral tilt results from a complex interaction between intonation, voice quality and vocal effort (d'Alessandro and Doval, 1998) and that syllabic structure has an effect on patterns of excitation (Ogden *et al.*, 2000).

synthesisers. The speech quality reached by a pure concatenation of natural speech segments (Black and Taylor, 1994; Campbell, 1997) is so high that complex coding techniques have been mostly used for the compression of segment dictionaries.

Physical interpretability

Complex speech production models such as formant or articulatory synthesis provide all spectro-temporal dimensions necessary and sufficient to characterise and manipulate speech signals. However, most parameters are difficult to estimate from the speech signal (articulatory parameters, formant frequencies and bandwidths, source parameters, etc.). Part of this problem is due to the large number of parameters (typically a few dozen) that have an influence on the entire spectrum: parameters are often estimated independently and consequently the analysis solution is not unique[2] and depends mainly on different estimation methods used.

If physical interpretability was a key issue for the development of early rule-based synthesisers where knowledge was mainly declarative, sub-symbolic processing systems (hidden Markov models, neural networks, regression trees, multilinear regression models, etc.) now succeed in producing a dynamically-varying parametric representation from symbolic input given input/output exemplars. Moreover, early rule-based synthesisers used simplified models to describe the dynamics of the parameters such as targets connected by interpolation functions or fed into passive filters, whereas more complex dynamics and phase relations have to be generated for speech to sound natural.

Characterising speech signals

One of the main strengths of formant or articulatory synthesis lies in providing a *constant* number of *coherent*[3] spectro-temporal parameters suitable for any sub-symbolic processing system that maps parameters to features (for feature extraction or parameter generation) or for spectro-temporal smoothing as required for segment inventory normalisation (Dutoit and Leich, 1993). Obviously traditional coders used in speech synthesis such as TD-PSOLA or RELP are not well suited to these requirements.

An important class of coders – *spectral models*, such as the ones described and evaluated in this section – avoid the oversimplified characterisation of speech signals in the time domain. One advantage of spectral processing is that it tolerates phase distortion, while glottal flow models often used to characterise the voice source (see, for example, Fant *et al.*, 1985) are very sensitive to the temporal shape of the signal waveform. Moreover spectral parameters are more closely related to perceived speech quality than time-domain parameters. The vast majority of these coders have been developed for speech coding as a means to bridge the gap (in

[2] For example, spectral slope can be modelled by source parameters as well as by formant bandwidths.

[3] Coherence here concerns mainly sensitivity to perturbations: small changes in the input parameters should produce small changes in spectro-temporal characteristics and vice versa.

terms of bandwidth) between waveform coders and LPC vocoders. For these coders, the emphasis has been on the perceptual transparency of the analysis-synthesis process, with no particular attention to the interpretability or transparency of the intermediate parametric representation.

Towards more 'ecological' signal generation systems

Contrary to articulatory or terminal-analogue synthesis that guarantees that almost all the synthetic signals could have been produced by a human being (or at least by a vocal tract), the coherence of the input parameters guarantees the naturalness of synthetic speech produced by *phenomenological models* (Dutoit, 1997, p. 193) such as the spectral models mentioned above. The resulting speech quality depends strongly on the *intrinsic* limitations imposed by the model of the speech signal and on the *extrinsic* control model. Evaluation of signal generation systems can thus divided into two main issues: (a) the intrinsic ability of the analysis-synthesis process to preserve subtle (but perceptually relevant) spectro-temporal characteristics of a large range of natural speech signals; and (b) the ability of the analysis scheme to deliver a parametric representation of speech that lends itself to an extrinsic control model. Assuming that most spectral vocoders provide toll-quality output for any speech signal, the evaluation proposed in this part concerns the second point and compares the performance of various signal generation systems on *independent* variation of prosodic parameters without any system-specific model of the interactions between parameters.

Part of this interaction should of course be modelled by an extrinsic control about which we are still largely ignorant. Emerging research fields tackled in Part III will oblige researchers to model the complex interactions at the acoustic level between intonation, voice quality and segmental aspects: these interactions are far beyond the simple superposition of independent contributions.

References

d'Alessandro, C. and Doval, B. (1998). Experiments in voice quality modification of natural speech signals: The spectral approach. *Proceedings of the International Workshop on Speech Synthesis* (pp. 277–282). Jenolan Caves, Australia.

Black, A.W. and Taylor, P. (1994). CHATR: A generic speech synthesis system. *COLING-94*, Vol. II, 983–986.

Campbell, W.N. (1997). Synthesizing spontaneous speech. In Y. Sagisaka, N. Campbell, and N. Higuchi (eds), *Computing Prosody: Computational Models for Processing Spontaneous Speech* (pp. 165–186). Springer Verlag.

Dutoit, T. (1997). *An Introduction to Text-to-speech Synthesis*. Kluwer Academics.

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

Fant, G., Liljencrants, J., and Lin, Q. (1985). A Four Parameter Model of the Glottal Flow. Technical Report 4. Speech Transmission Laboratory, Department of Speech Communication and Music Acoustics, KTH.

Lindblom, B. (1987). Adaptive variability and absolute constancy in speech signals: Two themes in the quest for phonetic invariance. *Proceedings of the XIth International Congress of Phonetic Sciences*, Vol. 3 (pp. 9–18). Tallin, Estonia.

Ogden, R., Hawkins, S., House, J., Huckvale, M., Local, J., Carter, P., Dankovičová, J., and Heid, S. (2000). ProSynth: An integrated prosodic approach to device-independent, natural-sounding speech synthesis. *Computer Speech and Language*, *14*, 177–210.