Speech Synthesis in Phonetics Teaching

Dafydd Gibbon (University of Bielefeld) and Jolanta Bachan (University of Bielefeld)

Common techniques in phonetics teaching have a history dating back to before the inception of the IPA in the late 19th century: mainly conditioned by foreign language teaching requirements, textbooks provide the core material, teachers embody operational models of speech production and perception. Needs in applied phonetics training have changed drastically, however: software and media development for semi-automatic tutoring, dictation, public announcements, require different kinds of knowledge and skills, and consequently we advocate practical training in modern speech technology techniques and media even in the early stages of general and foreign-language oriented phonetics teaching, particularly using time-aligned signal annotation and speech synthesis.

Speech synthesis is a technology with applications in public domains (automatic announcements) and personal domains (reading devices for the blind, talking navigators); as such it has proved its worth and is undergoing continual development. Less known are uses of rule-based speech synthesis for operational linguistic and phonetic theory testing (e.g. Dirksen & al. 1997; Hertz & al. 1999): correlates of representations in generative or autosegmental phonologies are provided with a literal implementation in a programming environment with a well-defined operational semantics.

We adopt an applications oriented version of the theory testing philosophy and present Close Copy Speech (CCS) synthesis techniques as a component of a methodology for phonetics teaching. Our main requirement specification for an appropriate speech synthesiser is that the major prosodic parameters of duration and pitch which need to be taught in phonetics courses should be easily parametrisable by teachers and students. This is partly the case for manual or scripted re-synthesis using Praat (Boersma et al. 2001). Modern, and highly realistic unit selection synthesisers such as FESTIVAL or BOSS have different goals, and are unsuitable for this purpose as prosodic information is corpus-derived and an interface for additional prosodic parametrisation is not easily accessible or not defined. The older technique of diphone synthesis represented by MBROLA (Dutoit 1997), on the other hand, has a clear interface which matches our specification.

The preliminary step in the teaching process is manual time aligned annotation of speech signal recordings. The next step is manual creation or modification of interface files, based on measurements made using the annotated files, i.e. manual CCS. However, we are more interested in relating large quantities of “real-life” utterances to synthesised utterances more precisely, and therefore introduce an automatic approach to CCS synthesis for the purpose. At a conceptual level, the close copy function is an alternative to NLP input into the DSP synthesis engine in the standard MBROLA synthesis model. We derive this input automatically from the annotated speech files, and the re-synthesised output permits rapid and direct auditory verification of the quality of the annotations; this process is of course much faster than manual CCS. We demonstrate the prototype of an automatic CCS system for Polish with this functionality, and provide a quantitative evaluation of its naturalness and intelligibility in comparison to original recordings and selected parametrisations.

References

Boersma, Paul & Weenink, David 2001. PRAAT, a system for doing phonetics by computer. *Glot International* 5(9/10): 341-345.

Dirksen, Arthur & John S. Coleman. 1997. All-Prosodic Synthesis Architecture. In J. P. H. van Santen, R. W. Sproat, J. P. Olive and J. Hirschberg, eds. *Progress in Speech Synthesis*. New York: Springer-Verlag. 91-108.

Dutoit, Thierry 1997. *An Introduction To Text-To-Speech Synthesis*. Dordrecht: KAP.

Hertz, Susan R., Rebecca J. Younes & Nina Zinovieva 1999. Language universal and language specific components in the multi-language ETI Eloquence Text-To-Speech System. *Proc. 14th ICPhS, San Francisco,* pp. 2283-2286.