Evolution of Text-to-Speech Systems and Methods of Their Assessment

(Vlado Delić, Milan Sečujski, Piroška Stanić Molcer)

The paper gives a retrospective of the development of speech synthesis systems, from mechanical synthesisers to computer systems for text-to-speech conversion (TTS) and analyses the perspectives of biomechanical and multimodal TTS systems within dialogue systems addressing higher cognitive levels as well. Special attention is given to the methods for assessment of the quality of synthesised speech and the applicability of TTS-based solutions.

The idea of machines speaking has been captivating the imagination of researchers and creative thinkers for centuries. Written accounts of "talking heads" date back to 10th and 12th century, while the first mechanical synthesisers are related to the work of von Kempelen, in the late 18th century. It was at that time that the physiological differences between vowels were explained and the first acoustical-mechanical machines modelling the vocal system (including the vocal folds) were made.

The first electronic TTS systems were articulatory or formant-based, and were developed in the mid-1960s and 1970s, while systems based on concatenation of pre-recorded speech segments were developed later, and they are predominantly used today for conversion of text into speech.

The architecture of modern TTS systems is described in more detail. The first component is **text processing,** charged with conversion of an arbitrary input text (string of characters) into orthographic words. The **phonetic analysis**, basically known as "grapheme-to-phoneme conversion", converts the orthographical symbols into phonological ones using a phonetic alphabet. Speech synthesis systems use two basic approaches to determine the pronunciation of a word based on its spelling: dictionary-based and rule-based approach. The phonemes are followed by corresponding visemes if a TTS system includes visual output in the form of an animated speaking face (avatar). The **prosodic modeling** aims at determining the rhythm of speech, stress pattern and intonation from text, as well as incorporating them into synthesised speech. Finally, the **speech signal synthesis** produces the speech signal, with voice characteristics usually close to those of a person whose voice was used for speech database recording. There are three types of speech signal synthesis: concatenative synthesis, formant synthesis, and articulatory synthesis.

The development and application of each TTS system are followed by a constant need for assessment of quality of synthesised speech and applicability of TTS-based solutions. The paper explains why TTS quality assessment is fundamentally different from measurement of quality of speech coding. It also presents basic methods for objective and subjective TTS quality assessment, and explains the importance of diversity in the choice of testers, including developers themselves and end users, both professional and lay. A possibility of inclusion of particularly interesting categories of users, such as linguists and the visually impaired, is analysed, through experiments in laboratory conditions as well as listening tests available via the Internet. Two versions of the TTS system are going to be tested on-line on the acoustic level. Prosody will be subjectively evaluated by determining the MOSLQS (mean opinion score listening quality). Intelligibility will be evaluated by open response identification test and psychoacoustic test of application specific sentence verification.

Human-machine spoken dialogue systems are based on speech technologies, TTS and ASR, as well as other modules like spoken language generation, spoken language understanding, and dialogue management, which are closer to the cognitive level of speech communications, where the speech production begins and ends. An intelligent TTS system should know not only what to say, but how to say it as well. One of the open questions is how to incorporate emotions and manner of pronunciation into synthesised speech based on the dialogue context and the given text.

[1] *William of Malmesbury's Chronicle of the Kings of England,* ed. J. A. Giles (London: 1847), p. 181; John Gower, *Confessio Amantis,* in *The Complete Works of John Gower,* ed. G.c. Macaulay (Oxford, 1899-1902, repr. 1968), II, 307-8 (lines 234-43).

[2] http://en.wikipedia.org/wiki/Articulatory_synthesis (accessed 13.9.2010)

[3] This information about talking heads, and that which follows, is from George Lyman Kittredge, *A Study of Gawain and the Green Knight* (Cambridge, MA, 1916), pp. 147-94; and Arthur Dickson, *Valentine and Orson: A Study in Late Medieval Romance* (New York, 1929), 190-216.

[4] B. Gold, N. Morgan, "Speech and Audio Signal Processing: Processing and Perception of Speech and Music", JW&S, 2000.

[5] http://knol.google.com/k/text-to-speech-synthesis# (accessed 13.9.2010)

[6] R. Pratt: "Quantifying the Performance of Text-to-Speech Synthesizers", Speech Technology, 1987. Vol. 3, pp. 54-64.

[7] R. van Bezooijen, V. van Heuven: "Quality of Synthesized Speech", Speech Coding and Synthesis, ed. W. B. Kleijn, K. K. Paliwal, 1995.

[8] A. Mahdi, D. Picovici: "New single-ended objective measure for non-intrusive speech quality evaluation", Journal of Signal, Image and Video Processing Springer, 2010. DOI: 10.1007-s11760-008-0092-1.

[9] Z. Becvar, J. Zelenka, M. Brada, T. Valenta: "Comparison of Subjective and Objective Speech Quality Testing Methods in the VoIP Networks" 13th International Conference on Systems, Signals and Image Processing, Budapest, 2006. Accessed 3th September 2010 cyberspace.mht.bme.hu/iwssip06/cikk/1098.pdf

[10] J. Tian, J. Nurminen, I. Kiss: "Modular Text-to-Speech Synthesis Evaluation for Mandarin Chinese", International Symposium on Chinese Spoken Language Processing, Singapore, 2006.

[11]R. Van Bezooijen (1986). Lay ratings of long-term voice-and-speech characteristics. F. Beukema & A. Hulk, Linguistics in the Netherlands 1986, 1-7. Foris, Dordrecht

[12] R. Van Bezooijen & W. Jongenburger (1993). Evaluation of an electronic newspaper for the blind in the Netherlands - intelligibility, acceptability, adequacy, and users'attitudes. Proceedings of the ESCA Workshop on Speech and Language Technology for Disabled Persons, 195-198, Stockholm

[13] E. Hjelmquist, B. Jansson & G. Torell (1987). Psychological aspects on blind people's reading of radio-distributed daily newspapers. B. Knave & P. Widebäck, , Work with display units 86, 187-201. North-Holland, Elsevier Science Publishers, Amsterdam.

[14] C. Delogu, A. Paoloni, P. Ridolfi & K. Vagges (1993b). Intelligibility of Italian text-to-speech synthesizers over ortophonic and telephonic channel. Proceedings of the Eurospeech '93, 3, 1893-1896, Berlin

[15] C. Delogu, A. Paoloni & C. Sementina (1992a). Comprehension of natural and synthetic speech: Preliminary studies. ESPRIT Project 2589 (SAM), , Multilingual speech input/output assessment, methodology and standardisation. University College London, London. Final report, Year three, 1.III.91-28.II.1992. SAM Internal Report II.c.

[16] A. Lampert "Evaluation of the MU-TALK Speech Synthesis System", 2004, http://www.ict.csiro.au/staff/andrew.lampert/writing/SynthesisEvaluation.pdf (accessed 6.9.2010)

[17] Yang A., Lohscheller, J., Berry, D.A., Becker, S., Eysholdt, U., Voigt, D., Döllinger, M., Biomechanical modeling of the three-dimensional aspects of human vocal fold dynamics, Journal of Acoustical Society of America, 127 (2), 2010, 1014-1031.