Text-to-Speech Synthesis for Whispered Speech

Liliya I. Tsirulnik[1], Valery A. Petrushin[2] and Veronika Makarova[3]

[1] Computer Audition Lab, UCSD, San Diego, CA, USA
liliya.tsirulnik@gmail.com
[2] Opera Solutions, San Diego, CA, USA
vapetr3@hotmail.com
[3] Department of Languages and Linguistics
University of Saskatchewan, Saskatoon, Canada
v.makarova@usask.ca

Whispering is a modality of speech that is defined as an unvoiced mode of phonation in which the vocal cords do not vibrate, but are adducted sufficiently to create audible turbulence as the speaker exhales during speech [1]. Whispered speech production is found in most world languages and is considered to be one of linguistic universals [2]. The social role of whispering is to communicate information to neaby listerners without being overheard by other people. Whisper is also used for communication by aphonic individuals who may be unable to produce vocal cord vibrations [3]. Over the last century, this speech modality has been examined within frameworks provided by a range of disciplines, such as speech science, phonetics, acoustics, engineering, medicine and health science. Studies of whispered speech find practical applications in evaluation of voice and hearing disorders, speaker and speech recognition for forensic, security, military and other purposes. A recent wave of interest in whispered speech research is explained by a wide spread of speech enabled devices such as mobile phones, smart phones and PDAs.

This research is devoted to analysis and test-to-speech synthesis of whispered speech. The practical purpose of this research is to extend voice cloning techniques [4] to whispered speech modality. The major challenge in whispered speech synthesis is that the absence of fundamental frequency (pitch) in whisper does not allow using traditional techniques for prosodic features modification. The "pitch" of whispered speech is represented by formants and prosodic modeling requires formant transformation. In this research the authors present the analysis of prosodic features that contribute to the expression of sentence intonation in whispered speech. The current work includes analysis of intonation contours for declarative (both complete and incomplete), interrogative and exclamatory types of sentences. It is shown that formants F1 and F2 have different profiles for different intonation types [5]. For modeling prosody of whispered speech, an extension of the Accent Unit Portrait Model [6], which includes formant portraits, is proposed [7]. The algorithms for creating and modification of melodic (using F1 and F2), rhythmic (duration) and energy (amplitude) portraits of accent units have been developed. Currently, the proposed algorithms are being embedded into an experimental concatenative text-to-speech synthesizer.

References:

[1] Laver, J., *Principles of Phonetics.* Cambridge University Press, 1994.
[2] Cirillo, J & Todt, D. Perception and judgment of whispered vocalizations. *Behaviour,* Vol. 142, 2005, pp. 113-128
[3] Morris, R. W., Clements, M. A. Reconstruction of Speech from Whispers. *Medical Engineering & Physics*, Vol. 24, 2002, pp 515-520.
[4] Lobanov, B., Tsirulnik L. "Phonetic-Acoustical Problems of Personal Voice Cloning by TTS", Proc. SPECOM'2004, pp 17-21.
[5] Tsirulnik, L. I., Petrushin, V. A., Makarova, V. "Analysis and TTS-synthesis of Russian Whispered Speech", Proc. SPECOM 2009, pp 180-185.
[6] Lobanov B., Karnevskaya H., "Auditory Estimation of Effectiveness of the AUP-Stylization Model of the Melodic Contour", Proc. SPECOM 2009, pp 155-158.
[7] Petrushin, V.A., Tsirulnik, L.I., Makarova V. "Whispered Speech Prosody Modeling for TTS Synthesis", Proc. "Speech Prosody 2010", p. 1151-1154.

Topic: Speech Technology, Test-to-Speech Synthesis

Preference: poster