Development of HMM-based acoustic laughter synthesis

Jérôme Urbain*, Hüseyin Cakmak, and Thierry Dutoit

TCTS Lab, Faculty of Engineering, University of Mons, Belgium

Laughter is a key signal in human communication, conveying information about our emotional state but also providing social feedback to the conversational partners. With the development of more and more natural human-computer interactions (with the help of embodied conversational agents, etc.), the need emerged to enable computers to understand and express emotions. In particular, to enhance human-computer interactions, talking machines should be able to laugh.

Yet, compared to speech synthesis, acoustic laughter synthesis is an almost unexplored domain. Sundaram and Narayanan [5] modeled the laughter intensity rhythmic envelope with the equations governing an oscillating mass-spring and synthesized laughter vowels by Linear Prediction. This approach to laughter synthesis was interesting, but the produced laughs were judged as non-natural by listeners. Lasarcyk and Trouvain [3] compared laughs synthesized by an articulatory system (a 3D modeling of the vocal tract) and diphone concatenation. The articulatory system gave better results, but they were still evaluated as significantly less natural than human laughs.

To improve laughter synthesis naturalness, we propose to use Hidden Markov Models (HMMs), which have proven efficient for speech synthesis. We opted for the HMM-based Speech Synthesis System (HTS) [4], as it is free and widely used in speech synthesis and research. The data used comes from the AVLaughterCycle database (AVLC), which contains around 1000 laughs from 24 subjects and includes phonetic transcriptions of the laughs [6].

HTS provides a demonstration canvas for speech synthesis, which enables to quickly obtain synthesis models with standard speech parameters. Our first works were to use this canvas to build a baseline for HMM-based laughter. Then, we looked at adapting our data and modifying some parts of the HTS demo to improve the quality of the obtained laughs.

The major improvement of the AVLC database to better exploit the potential of HTS is the annotation of laughter "syllables"[1]. This enables to include contextual parameters (e.g. the position of the "phoneme" within its "syllable", the position of the current "syllable" within the current "word", etc.) in the synthesis models.

[*] jerome.urbain@umons.ac.be

[1] We use quotation marks around the terms *syllable, phoneme and word* to distinguish the laughter units from their speech counterparts.

Two important modifications have also been done in the HTS process compared to the demonstration algorithms. First, the standard Dirac pulse train for voiced excitation has been replaced by the DSM model [1], which better fits the human vocal excitation shapes and reduces the buzziness of the synthesized voice. Second, the standard vocal tract and fundamental frequency estimation algorithms provided by HTS have been replaced by the STRAIGHT method [2], which is known in speech processing to provide better estimations.

These modifications largely improved the quality of the synthesized laughs. Some examples of HMM-based laughter synthesis are available on `http://www.ilhaire.eu/blog~Acoustic-Laughter-Synthesis`. It is important to note that we are currently not able to generate new laughter phonetic transcriptions, and in consequence we re-synthesize existing human transcriptions. Future work includes the development of a module to generate (or modify existing) phonetic transcriptions, further optimizations of the synthesis parameters and a perceptive evaluation study to quantify the improvements and provide a benchmark for future developments.

Acknowledgment

This work was supported by the European FP7-ICT-FET project ILHAIRE (grant n°270780).

References

[1] T. Drugman and T. Dutoit. The deterministic plus stochastic model of the residual signal and its applications. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20:968–981, 2012.

[2] H. Kawahara. Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds. *Acoustical science and technology*, 27(6):349–353, 2006.

[3] E. Lasarcyk and J. Trouvain. Imitating conversational laughter with an articulatory speech synthesis. In *Proceedings of the Interdisciplinary Workshop on the Phonetics of Laughter*, pages 43–48, Saarbrücken, Germany, August 2007.

[4] Keiichiro Oura. Hmm-based speech synthesis system (hts) [computer program webpage]. `http://hts.sp.nitech.ac.jp/`, consulted on June 22, 2011.

[5] S. Sundaram and S. Narayanan. Automatic acoustic synthesis of human-like laughter. *Journal of the Acoustical Society of America*, 121(1):527–535, January 2007.

[6] Jérôme Urbain and Thierry Dutoit. A phonetic analysis of natural laughter, for use in automatic laughter processing systems. In *Proceedings of the fourth bi-annual International Conference of the HUMAINE Association on Affective Computing and Intelligent Interaction (ACII2011)*, pages 397–406, Memphis, Tennesse, October 2011.