USING MAGE FOR REAL TIME SPEECH-LAUGH SYNTHESIS

Kevin El Haddad, Alexis Moinet, Hüseyin Çakmak, Stéphane Dupont, Thierry Dutoit

TCTS lab-University of Mons (UMONS/Belgium)
{kevin.elhaddad},{alexis.moinet},{huseyin.cakmak},{stephane.dupont},{thierry.dutoit}@umons.ac.be

ABSTRACT

In this paper, we present an ongoing work which aims at synthesizing speech-laugh sentences in real-time. To do so, the Hidden Markov Model (HMM)-based speech-laugh synthesis system will be used along with the MAGE software library. First results are available online on tcts.fpms.ac.be/~laughter/laughterWorkshop15.

Keywords: Real-time speech-synthesis, speech-laughs, laugh

1. INTRODUCTION

Recent research has been focusing on improvements to the expressivity and naturalness of synthetic speech. One way to do so is to add emotions to the synthesis. This paper presents work in the framework of affective speech synthesis, focused more particularly on amused speech. We attempt to present perspectives on real-time speech-laugh synthesis. Speech-laugh designates laughs happening at the same time as speech and intermingled with it. Several phonetic and acoustic studies concerning speech-laugh can be found in [11], [7] and [5]. To the best of our knowledge, very few were made addressing speech-laugh synthesis. Lasarcyk [6] reproduced natural speech-laughs using an articulatory synthepsizer. Also, this work relies on a previously developed Hidden Markov Model (HMM)-based synthesis system for speech-laugh [4]. We will attempt to use this system alongside with the MAGE software [1] to synthesize and control speech-laugh in real-time. This system will be integrated in the Chist-ERA Joker project. This project is currently designing JOKER, a generic intelligent user interface providing a multimodal dialogue system with social communication skills including humor, empathy, compassion, charm, and other informal socially-oriented behaviors (more information about this project can be found in http://www.chistera.eu/projects/joker).
This paper presents the HMM-based speech-laugh synthesis system in section II. We then present the MAGE software library in section III. In section IV, we explain how the two systems will be used together. Finally, section V will conclude.

2. HMM-BASED SPEECH-LAUGH SYNTHESIS SYSTEM

This section only summarizes the system presented in [4]. Please refer to that earlier work for a more detailed presentation. This work attempts to model speech-laughs. Speech-laughs are variable and depend on social and/or situational context. A realistic model would be very complex to create. So further studies, like the one in chapter 9 of [6], in order to develop such a model. Our first approach was therefore, to simplify the model by regarding speech-laughs as laughter bursts replacing vowels in speech-smiles. Fig. 1 shows the HMM-based speech-laugh synthesis system workflow.

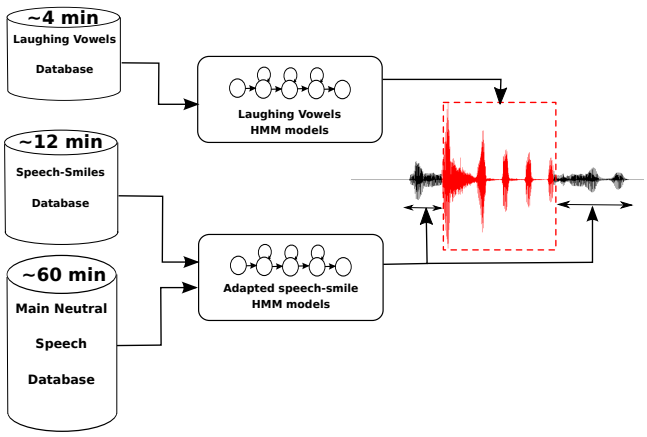

Figure 1: HMM-based speech-laugh synthesis system

HMM models of a neutral voice (reading voice) are first trained using a corpus of approximately 1 hour of sentences in French read by a Belgian native French-speaking actor. Those created parametric probabilistic models are then converted to a speech-smile voice. By speech-smile we mean Duchenne smiles (smiles containing real enjoyment) occurring at the same time as speech and therefore altering it [3]. This conversion is made by adaptation using the Constraint Maximum

Likelihood Linear Regression (CMLLR) algorithm [2]. Moreover speech-smile HMM models were also trained from a smaller corpus (approximately 12 minutes long) of speech-smiles recorded from another French-speaking person. This person was trained to pronounce the desired smiled voice before the recordings. The synthesis results in [4] and in [3] showed the efficiency of those data when used for training HMM models of speech-smiles.
Vowels are then replaced by laughter bursts inside those speech-smile sentences. In order to do that, "laughing vowels" which are laughter bursts occurring inside vowels, were recorded from the same person the speech-smile sentences were recorded from. He was asked to produce sustained French vowels while watching funny videos. HMM models are then also created for the laughing vowels. These models are used to replace the vowels by laughing vowels in the synthesized speech-smile sentences. The sentences synthesized using this system proved to be efficiently perceived as amused as the evaluations in [4] show.
Practically, for synthesis, the input of the system is a list of phonemes. The system then chooses the best suited trajectories for the features (that model the voices) using a maximum-likelihood parameter generation algorithm making use of the HMM model previously made during training [10]. From those trajectories, a waveform is generated from a synthesizer. Fig. 2 shows the laughing vowels labeling.

3. MAGE

MAGE [1] is a software library built as a realtime layer around HTS engine [9]. It allows modifications of the parameters and inputs of an HMM-based speech synthesizer on-the-fly, therefore affecting its output while it is speaking.

For instance, a user can give, one-by-one, a sequence of phonemes (complying with the format specified in [8]) as inputs of the system and it will start talking as soon as it receives the first phonemes. Then the user can change the pitch trajectories, or the vocal tract length of the voice as it is produced. When no more phonemes are sent to MAGE, it can either sustain indefinitely the last received one or just stop talking.

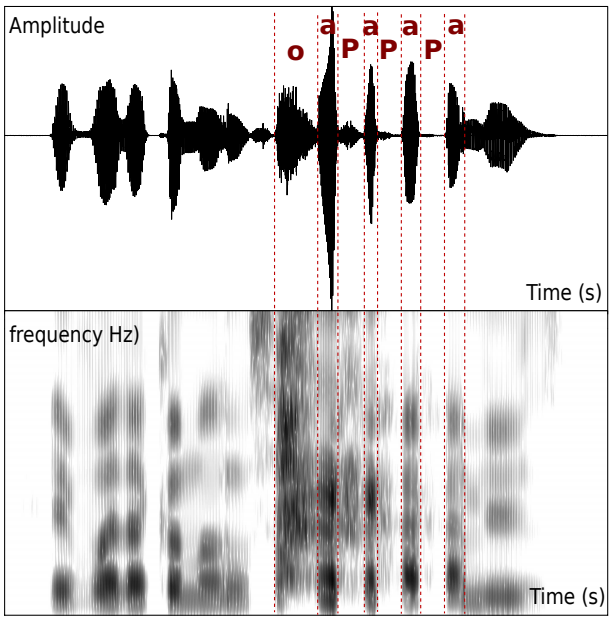

Figure 2: Laughing-vowels: pulse-vowel-pulse-vowel pattern [4].

4. REAL TIME SPEECH-LAUGH SYNTHESIS

Since phonemes can be processed by MAGE progressively, whenever we want to add a laughter in a sentence, we intend to replace a given vowel with a sequence “pulse-vowel-pulse-vowel” just before it is sent as an input. The readers are invited to download some first examples synthesized using MAGE while this paper was being written on tcts.fpms.ac.be/~laughter/laughterWorkshop15

5. CONCLUSION

In this extended abstract, we presented our ongoing work concerning real-time speech laugh synthesis. A concrete example of application will be its integration into the Chist-ERA Joker Project cited above.

6. REFERENCES

[1] Astrinaki, M., d'Alessandro, N., Dutoit, T. 2012. MAGE-a platform for tangible speech synthesis. *Proceedings of the International Conference on New Interfaces for Musical Expression* 353–356.

[2] Digalakis, V. V., Rtischev, D., Neumeyer, L. G. 1995. Speaker adaptation using constrained estimation of Gaussian mixtures. *IEEE Transactions on Speech and Audio Processing* 3(5), 357–366.

[3] El Haddad, K., Dupont, S., d'Alessandro, N., Dutoit, T. in press. 2015. An HMM-based speech-smile synthesis system: An approach for amusement synthesis. *International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE 2015)*.

[4] El Haddad, K., Dupont, S., Urbain, J., Dutoit, T. in press. 2015. Speech-laughs: An HMM-based Approach for Amused Speech Synthesis. *Internation Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*.

[5] Kohler, K. J. 2008. "Speech-smile","speech-laugh","laughter" and their sequencing in dialogic interaction. *Phonetica* 65(1-2), 1–18.

[6] Lasarcyk, E. 2014. *Empirical evaluation of the articulatory synthesizer VocalTractLab as a discovery tool for phonetic research: Articulatory-acoustic investigations of paralinguistic speech phenomena*. PhD thesis Saarland University.

[7] Menezes, C., Igarashi, Y. 2006. The speech laugh spectrum. *Proc. 7th ISSP* Ubatuba, Brazil. 517–524.

[8] Oura, K. 2011. An example of context-dependent label format for HMM-based speech synthesis in English. *The HTS CMUARCTIC demo*.

[9] Oura, K. consulted on August, 2014. HMM-based speech synthesis system (HTS) [computer program webpage]. http://hts.sp.nitech.ac.jp/.

[10] Tokuda, K., Zen, H., Black, A. W. 2002. An HMM-based speech synthesis system applied to english. *Proceedings of IEEE Workshop on Speech Synthesis*. IEEE 227–230.

[11] Trouvain, J. 2001. Phonetic aspects of "speech laughs". *Oralité et Gestualité: Actes du colloque ORAGE, Aix-en-Provence. Paris: L'Harmattan* 634–639.