Mage - HMM-based speech synthesis reactively controlled by the articulators

Maria Astrinaki[1], *Alexis Moinet*[1], *Junichi Yamagishi*[2,3], *Korin Richmond* [2], *Zhen-Hua Ling*[4], *Simon King*[2], *Thierry Dutoit*[1]

[1]Circuit Theory and Signal Processing Lab, Numediart Institute, University of Mons, Belgium
[2]The Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK
[3]National Institute of Informatics, Tokyo, Japan
[4]University of Science and Technology of China (USTC), China

maria.astrinaki@umons.ac.be, alexis.moinet@umons.ac.be, jyamagis@inf.ed.ac.uk
korin@cstr.ed.ac.uk, zhling@ustc.edu, simon.king@ed.ac.uk, thierry.dutoit@umons.ac.be

Abstract

In this paper, we present the recent progress in the MAGE project. MAGE is a library for realtime and interactive (reactive) parametric speech synthesis using hidden Markov models (HMMs). Here, it is broadened in order to support not only the standard acoustic features (spectrum and f_0) to model and synthesize speech but also to combine acoustic and articulatory features, such as tongue, lips and jaw positions. Such an integration enables the user to have a straight forward and meaningful control space to intuitively modify the synthesized phones in real time only by configuring the position of the articulators.
Index Terms: speech synthesis, reactive, articulators

1. Reactive HMM-based speech synthesis

MAGE is based on HTS [1], which it extends in order to support realtime architecture and multithreaded control. MAGE uses multiple threads, and each thread can be affected by the user. This allows accurate and precise control over the different production levels of the artificial speech. MAGE integrates three main threads: the *label thread* responsible for the contextual control, the *parameter generation thread* responsible for the reactive parameter generation by means of short-term parameter trajectories and local maximization and the *audio generation thread* responsible for the vocoding. Three queues are shared between threads for sharing and exchanging data: the *label queue*, the *parameter queue* and the *sample queue*. Further details can be found in [2].

2. Reactive articulatory feature control

In this work, MAGE is modified in order to generate and alter articulatory features. Given the unified acoustic-articulatory model [3], [4] and a set of phonetic labels, it is possible to reactively generate the target speech samples. Simultaneously, it is possible to influence the generated acoustic features by replacing the generated articulatory features with the user input. In this way, we can achieve the goal of altering the generated speech samples at the articulatory level rather than directly at the acoustic level. Note that the intention is to reactively alter a given context and its acoustic features by using only modifications over the articulatory features provided by the user. Here we present an application that combines the MAGE synthesizer with a graphical user interface (GUI)[1]. The GUI is dependent on the database we use for the synthesis [3]. It depicts a two dimensional midsagittal view of the vocal tract drawn using 124 points.The position of these EMA points can be reactively controlled by the user using a mouse or touch screen. There are no limits to the possible position of the EMA points providing to the user 12 degrees of freedom. This means that the user is free to place these points in coordinates that are "unnatural" either from a physical point of view or as sequence of movements. The user is also able to load predefined configurations of vocal tract shapes and EMAs and continue to apply his own controls. The shape of the vocal tract can be reactively altered so that the user will have a reference point to the initial configuration chosen. The final part of the application, generating the speech waveform, is implemented by MAGE. The user modifications over the EMA points are taken into account to generate the corresponding articulatory features. These features are used to estimate the acoustic features, then will give the final speech samples.

3. Conclusions

We presented a method that enables reactive articulatory control over HMM-based parametric speech synthesis using MAGE combined with an application that enables the user to reactively control the position of the articulators through a GUI. We see that reactive articulatory control is feasible, and combined with an interface allows us to explore different aspects of the speech production. However the most important aspect of this work is that we actually prove that MAGE can be used also as a reactive mapping tool between different feature streams. In this case MAGE is able to reactively map the user controls over the articulatory feature stream over the acoustic feature stream.

4. References

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, pp. 1039–1064, 2009.

[2] M. Astrinaki, N. d´Alessandro, L. Reboursière, A. Moinet, and T. Dutoit, "MAGE 2.00: New features and its application in the development of a talking guitar," in *Proc. of NIME´13*, 2013.

[3] Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang, "Integrating articulatory features into HMM-based parametric speech synthesis," *IEEE TASLP*, vol. 17, no. 6, pp. 1171–1185, 2009.

[4] Z. Ling, K. Richmond, and J. Yamagishi, "Articulatory control of HMM-based parametric speech synthesis using feature-space-switched multiple regression," *IEEE TASLP*, vol. 21, no. 1, pp. 207–219, 2013.

[1]A video demonstration of the presented system can be found in https://vimeo.com/67404386.