MAGEFACE: Performative Conversion of Facial Characteristics into Speech Synthesis Parameters

Nicolas d'Alessandro, Maria Astrinaki, and Thierry Dutoit

Institute for New Media Art Technology, University of Mons
Signal Processing Laboratory, 31 Boulevard Dolez, B-7000 Mons
nda@numediart,{maria.astrinaki,thierry.dutoit}@umons.ac.be

Abstract. In this paper, we illustrate the use of the MAGE performative speech synthesizer through its application to the conversion of realtime-measured facial features with FaceOSC into speech synthesis features such as vocal tract shape or intonation. MAGE is a new software library for using HMM-based speech synthesis in reactive programming environments. MAGE uses a rewritten version of the HTS engine enabling the computation of speech audio samples on a two-label window instead of the whole sentence. Only this feature enables the realtime mapping of facial attributes to synthesis parameters.

Keywords: speech synthesis, software library, performative media, streaming architecture, HTS, MAGE, realtime audio software, face tracking, mapping.

1 Introduction

Speech is the richest and most ubiquitous modality of communication used by human beings. Through vocal expression and conversation, we realize a complex process, highly interactive and social. For decades, algorithms for the production of synthetic speech have focused on converting a static text into an intelligible and natural waveform, lately with great success. Ten years ago, the trend of creating expressive or emotional speech brought researchers to realize that such properties were not only a matter of sound quality. Expressivity in speech is contextual, interactive, social, coming in response to other ongoing processes, reaching across most of other human being modalities, and therefore other disciplines. Through this large interdisciplinary redefinition of expressive speech, one property seems to emerge and reach some consensus: speech is a performance; it starts with a gesture and ends up as a message, conveying both informative and affective contents.

As these new trends in understanding expressivity in speech are being explored, one might notice that a real solid platform is missing. Indeed Text-To-Speech (TTS), as a platform, has been tackling and greatly solving other problems: similarity between original and synthetic waveforms, segmental and supra-segmental qualities, intonation modeling, etc. However most of existing TTS systems require a significant amount of text in advance (typically a sentence) and process it into sound as a whole target. Most of the time, the ability to influence the synthesis process has been

M. Mancas et al. (Eds.): INTETAIN 2013, LNICST 124, pp. 179–182, 2013.

limited, disabled or discouraged, as the resulting sound quality quickly degrades. If we consider that expressivity is related to the ability to interact with the artificial speech production process at various production levels and time scales, as it would happen with real speech, then the requirements for such a platform are different: we need a so-called reactive programming architecture, applied to speech synthesis.

2 Performative HMM-Based Speech Synthesis

For the last decade, HMM-based speech synthesis has been constantly improving and became a serious alternative to non-uniform unit selection (NUU), especially when a more lightweight and flexible synthesis engine is required. Particularly, the HTS system [6] is now reaching a reasonably high synthesis quality. Moreover, the model-based approach used in HTS to generate the speech production parameters enables a whole new category of techniques, such as speaker adaptation, speaker interpolation, voice cloning, voice reconstruction, etc.

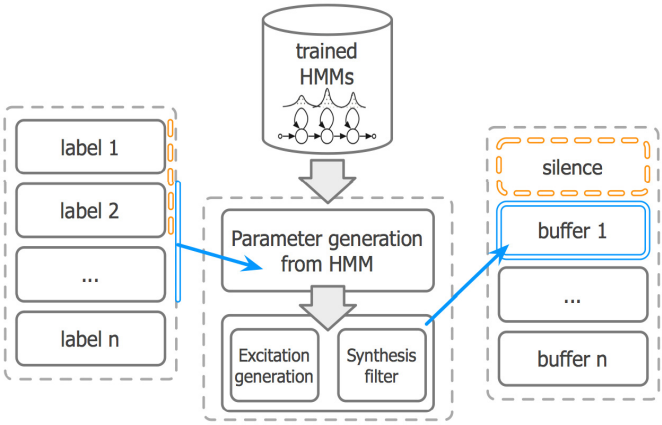

Fig. 1. Performative implementation of HTS: speech signal is computed on a sliding two-label window instead of the whole sentence. This enables the parameters to be changed on the fly.

One fundamental aspect of HTS is the use of context-dependent HMMs. Before the statistical models are trained, each phoneme is characterized by its linguistic context, e.g. previous phoneme, next phoneme, current syllable, previous syllable, next syllable, etc. The trend in synthesizing natural speech trajectories from text has led to add as much phonetic context as possible, to capture variations that are encountered in real speech. As a consequence, the need for a large look-ahead in the future (next phoneme, next syllable, next word) has brought the accessible time scale at run- time in HTS to the current sentence. Therefore, scenarios such as starting to synthesize a sentence with one speaking style and terminating that ongoing sentence with another speaking style based on an unpredictable user control command is impossible.

2.1 Performative Implementation of HTS (pHTS)

In pHTS, we have developed a series of modifications, enabling a much more reactive control of speech parameter trajectories. The main modification is the optimization of the generation of speech parameters on a sliding window of 2 labels rather than on the whole sentence, as shown in Fig. 1. When speech parameters have been generated for such a 2-label window, audio samples corresponding to the past label can be synthesized right away. If these samples are used within a realtime audio architecture, it means that modifications achieved on pHTS models will have an impact on the ongoing speech audio output with a delay of only one label.

2.2 MAGE: Flexible API for Speech Synthesis

MAGE is the software umbrella that provides the appropriate real- time audio architecture, in order to plug the pHTS speech synthesis engine. Indeed it schedules the various tasks encountered in the pHTS synthesis, so that the sound is constantly synthesized from an ongoing stream of asynchronous user-provided phonemes:

- on the fly generation of label-formatted streams;
- scheduling of model selection from the database;
- scheduling of speech parameters generation;
- scheduling of MLSA filtering.

The MAGE framework transparently uses concurrent programming techniques in order to guarantee the reactivity and flexibility of the application to unpredictable inputs, such as new labels, modification of F0 models, duration models or MGC models. MAGE is also the opportunity to encapsulate the HTS functionalities into a user-friendly API, so that more developers can integrate our new speech synthesis features in their applications.

3 Prototypes

MAGE as a software library has already been used in various prototypes. These prototypes tend to highlight a common aspect of our research work: exploring how HMM-based speech synthesis can be gesturally controlled. The concept of gesture or performance is here considered in a very large sense. Indeed we work with any user input that can have a meaningful impact on speech production properties. Our primary interest is the manipulation of speech phonemes and prosody with hand gestures, as in [5] and [1], but here we consider more indirect causes, such as facial expression.

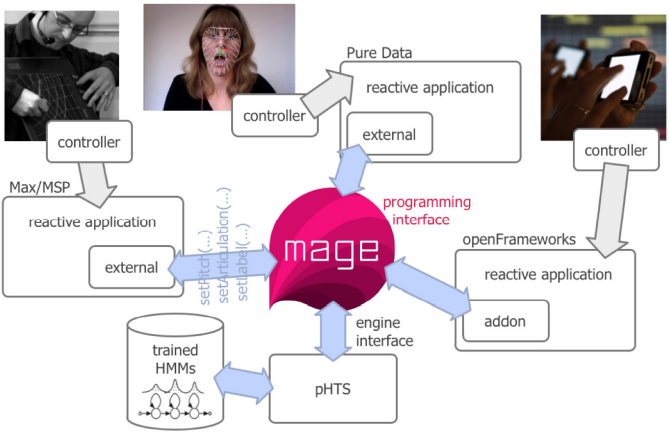

Fig. 2. Various scenarios of MAGE used to create a speech synthesis external object or add-on, as to enable the use of various controllers for impacting on the speech synthesis output

Fig. 2 gives an overview of integrating MAGE into various software environments, with the aim of connecting controllers to the speech synthesis. We can highlight various prototypes that have already been built, following this process. Firstly, the HandSketch musical instrument [1], formerly used for singing synthesis, is now able to change speaking speed and intonation with pen or finger gestures [3]. We can demonstrate a version in Max/MSP with a tablet and an iPhone version. We also built a virtual speaker, triggering syllables directly from mouth motion, using a face tracking software [2] in openFrameworks and synthesizing the speech in PureData. Face features come from FaceOSC [7]. We use various mouth opening sequences and eye broses position to influence the speech intonation and the vocal tract length.

References

1. d'Alessandro, N., Dutoit, T.: HandSketch Bi-Manual Controller: Investigation on Expressive Control Issues of an Augmented Tablet. In: Proc. International Conference on New Interfaces for Musical Expression, pp. 78–81 (2007)
2. MAGE and Face Tracking, `https://vimeo.com/39567236`
3. MAGE and HandSketch, `https://vimeo.com/39558917`
4. MAGE website, `http://mage.numediart.org`
5. Nordstorm, K., et al.: Developing Vowels Mappings for an Interactive Voice Synthesis System Controlled by Hand Motions. Journal of the Acoustical Society of America 127, 2021 (2010)
6. Zen, H., Tokuda, K., Black, A.: Statistical Parametric Speech Synthesis. Speech Communications 51(11), 1039–1064 (2009)
7. FaceOSC, `https://vimeo.com/26098366`