MAGE – A Platform for Tangible Speech Synthesis

Maria Astrinaki
numediart – Institute for New Media Art Technology of the University of Mons
B-7000, Mons, Belgium
maria.astrinaki@umons.ac.be

Nicolas d'Alessandro
numediart – Institute for New Media Art Technology of the University of Mons
B-7000, Mons, Belgium
nicolas@dalessandro.be

Thierry Dutoit
numediart – Institute for New Media Art Technology of the University of Mons
B-7000, Mons, Belgium
thierry.dutoit@umons.ac.be

ABSTRACT

In this paper, we describe our pioneering work in developing speech synthesis beyond the Text-To-Speech paradigm. We introduce tangible speech synthesis as an alternate way of envisioning how artificial speech content can be produced. Tangible speech synthesis refers to the ability, for a given system, to provide some physicality and interactivity to important speech production parameters. We present MAGE, our new software platform for high-quality reactive speech synthesis, based on statistical parametric modeling and more particularly hidden Markov models. We also introduce a new HANDSKETCH-based musical instrument. This instrument brings pen and posture based interaction on the top of MAGE, and demonstrates a first proof of concept.

Keywords

speech synthesis, Hidden Markov Models, tangible interaction, software library, MAGE, HTS, performative

1. INTRODUCTION

Speech is one of the richest and most ubiquitous modalities of communication used by human beings. Vocal expression involves complex production and perception mechanisms. Conversation is a highly interactive process, with complex timings and wide-ranging variations of quality. It is known that speech production properties have a deep impact on perceived identity and social cues [9]. This critical role of speech production in our life makes anybody an expert listener. The synthesis of artificial speech has been explored for decades to use in many applications, from the purely functional level to artistic exploration. However, human's natural expertise in listening to spoken content makes speech synthesis a really complex problem. Recent synthesizers have made great progress in terms of intelligibility and naturalness but they are still not providing a completely convincing vocal experience to users, neither an expressive tool for artists. In this Section, we describe what led to this situation, as an introduction to our concept of *tangible speech synthesis* and our new speech synthesis system.

The speech research community has been making outstanding progresses over the last decades, but it seems that some aspects of speech production remain misunderstood. For example, we do not have an exhaustive model for vocal folds vibration, because observations *in vivo* are nearly impossible. There is also a big debate in how speech production is influenced by the context, such as speaker's emotional state, listener's reaction or other surrounding stimuli, because real-life measurements are intrusive. These issues result in an elusive mapping between parameters of existing production models and real social impacts of speech.

At the time TTS[1] became the main trend, it was not evident that computing would go mobile so massively. Retrospectively, we understand how the design choices underlying TTS – text input and black-boxed generation of resulting speech – have anchored its use to reading text on desktop computers. Indeed, mobile computing relies on ubiquitous sensing of user's context, and user interaction tends to become more natural. Therefore, high-quality speech synthesis seems to have major issues in being used "in the wild". Nowadays there are two main application types that are prevented to expand because of these limitations:

1. *Context-reactive speech synthesis*: Ubiquitous computing brings current devices to gather a lot of information about our context. However, TTS makes few sense of this context, because the speech production properties can barely be altered, even less on-the-fly.

2. *Performative speech synthesis*: artificial speech can be generated from gestural performance, rather than pretyped text. This requires a major breakthrough in speech synthesis techniques in order to create speech sounds from non-textual fast-changing inputs.

In this paper, we present a software platform for tangible speech synthesis, as a way to address the applicative case 2 – performative speech synthesis – by giving a better grasp on some critical speech production parameters. More motivations about tangible speech synthesis are given in Section 2. The software platform, called MAGE, and its reactive properties are described in Section 3. In Section 4, we show our first musical instrument prototype, that brings the tangible control aspect. Finally we discuss the impact of tangible control on speech intelligibility, in Section 5.

2. MAKING SPEECH TANGIBLE

In this Section, we propose to frame tangible speech synthesis as a way of designing alternatives to TTS architectures. Artificial speech production is now envisioned as the realization of a musical instrument, with tangible properties. Further in this part, we extend the definition of tangible speech synthesis. Then we gather several motivations to explore this new direction. Finally we mention some research fields in which the development of tangible speech synthesis brings interesting contributions.

[1]TTS: Text-To-Speech, synthesize speech from text.

NIME'12, May 21 – 23, 2012, University of Michigan, Ann Arbor.

2.1 Definition

The typical conversion pattern for TTS takes one sentence of raw text, estimates pronunciation features (phonemes and prosody) from the linguistic knowledge, then converts these features into speech waveforms. We define a tangible speech synthesis system as a combined hardware/software system that can afford two main features at synthesis time:

- a consistent and meaningful mapping between some critical speech production parameters and graspable entities or dimensions of a tangible control surface; critical speech production parameters that we consider are: articulation of phonetic cells (e.g. syllable), pitch trajectory, pronunciation timing and voice quality;
- the sufficient reactivity of these speech parameter trajectories to user manipulation, which is a similar guideline as in the design of new musical instruments; it is known that users have different tolerances to latency and time resolution, depending on which speech production feature is considered [9].

2.2 Motivations

There are different ways of approaching the idea of working beyond TTS. In this paper, we investigate tangible speech synthesis. This choice is motivated by several observations in various fields, from human communication and interaction to historical perspectives in artificial speech production. These motivations are described in this part.

Speech and Hands in Communication

According to McNeill, vocal apparatus and hands have competed for a long time for the language function, along the evolution of human beings [8]. If speech motor control and speech perception regions of the brain finally developed further towards the ability to communicate through sound, it is really likely that hands have a similar underlying potential. Their major role in co-verbal gestures, as well as the richness of existing sign languages in deaf communities reinforce this assumption. Therefore, researchers are wondering if hand-based languages actually activate speech regions in the brain. This question is still widely open.

Historical Perspective

Prior to computed-based speech synthesis, artificial speech production devices have been mechanical and then electrical. We can respectively highlight the von Kempelen machine in the XIXth century [12] and the Voder in the late 1930s [4]. Before the ability of computers to programmatically generate parameter trajectories of the speech production model, these trajectories had to be realized by hand, thanks to a well-trained operator. Therefore, these devices were including an ad hoc control interface.

The idea of creating musical instruments that speak and sing is part of a contemporary history in computing, showing that this trend actually did not stop when TTS systems became the mainstream. Fels' GloveTalk [5] to DiVA [6] systems have explored the hand-based control of coarticulated speech since the early 1990s. Cook's SqueezeVox is also one from many attempts of the author to provide tangible control of singing synthesis [2].

3. MAGE SPEECH SYNTHESIZER

MAGE is a new C/C++ software platform for reactive speech synthesis. In this case, "reactive" means that both phonetic content and prosody of synthetic speech can be controlled by the user in realtime. Our main focus is to provide such reactivity while preserving a state-of-the-art intelligibility and naturalness in the speech output. In order to afford this trade-off, we decided to use HMM-based speech synthesis, and more precisely the HTS system [13]. MAGE is a partial redesign of HTS, allowing the user to alter ongoing parameter trajectories of the speech production model. In this Section, we first describe the HTS system. Then we present our modifications respectively achieved on the DSP and NLP[2] modules of HTS.

3.1 HTS: HMM-Based Speech Synthesis

Nowadays, the most intelligible and natural-sounding synthesized speech is obtained with unit selection algorithms. Unit selection uses a very large database of natural speech sounds from the same speaker, which is segmented into so-called *units*. These units correspond to a wide variety of time scales, from phone to sentence. At runtime, these speech segments are optimally selected from the database and concatenated, according to the original requested speech sequence, called the *target*. Although these techniques lead to unprecedented intelligibility and naturalness, they have major drawbacks. Indeed, these algorithms exhibit an important computational footprint. They also work like a black box, with no production model available. The resulting production trajectories are entirely data-driven [7].

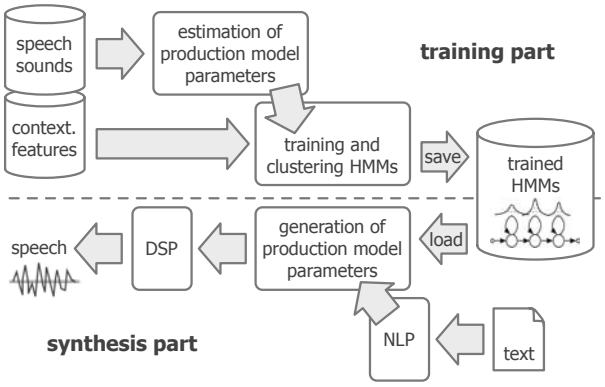

Figure 1: Training and synthesis parts of the HMM-based speech synthesis system (HTS) [13].

For the last ten years, a new approach to corpus-based speech synthesis has emerged: statistical parametric speech synthesis. Particularly, the HTS system [13] brought the necessary leap forward for envisioning this technique as an alternative to unit selection. The architecture of HTS consists in two main parts: the training and the synthesis. Both training and synthesis parts are illustrated in Figure 1.

HTS is the best starting point to create a reactive speech synthesizer – and further a tangible speech synthesis device – since it is model-based from the ground and therefore highly flexible. Indeed, HTS is already used in many voice transformation and morphing techniques. It also produces intelligible and natural speech, with a small footprint.

3.2 Reactive Trajectory Generator

MAGE relies on the same database of trained context-dependent HMMs than HTS, meaning that the training phase is strictly identical. The DSP, which converts speech production model parameters into sound, also uses the same MLSA filtering technique [13] in both systems. Our major modification of the HTS engine is the way of generating the parameter trajectories. In HTS, the whole sequence of labels corresponding to the targeted sentence is used to gener-

[2]Natural Language Processor

ate speech production parameters. Indeed, once the HMMs have been selected in the database, the statistical generation of parameter trajectories is optimized over the whole target, as a way to improve the speech output. As a result, the accessible time scale in HTS is the sentence. MAGE opens this enclosed loop and reduces the time scale by optimizing parameter trajectories locally. MAGE generates the production model parameters and their corresponding audio samples on a sliding window of two labels. Such a window is used to have a look-ahead buffer of one label and be able to properly generate the inter-phoneme coarticulation. Within this window, most of the generated trajectories can be altered. This redesign results in a speech synthesis process in which most of the production properties can be manipulated with a delay of only one label.

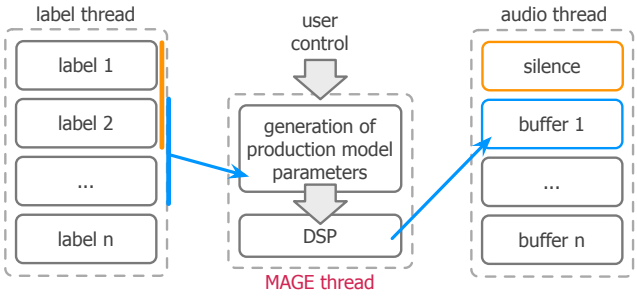

Figure 2: Multithread architecture of MAGE: speech synthesis thread makes the connection between the stack of labels and the audio thread.

MAGE has a multithread architecture and respects the typical software design pattern for realtime audio applications. Labels are added to a First-In-First-Out data structure running in its own thread. The MAGE speech synthesis thread pops labels out of the FIFO, according to the two-label sliding window, then generates the parameter trajectories and the speech samples. Samples are sent to the audio thread as consecutive buffers. Each buffer contains the audio samples of one label. Due to the one-label delay, the first audio buffer always corresponds to silence. This process is summarized in Figure 2. The pitch curve, the speech timing and the vocal tract length can be modified.

3.3 Reactive Natural Language Processor

Although the parameter trajectories are generated within a two-label window, the NLP module used with HTS still assumes that the whole text is provided at once. This assumption has a significant impact on the mechanisms that are encountered in the computation of contextual information, necessary to generate the labels. It relates the current phoneme to its preceding/following ones and to the larger segment that contains this phoneme, i.e. the syllable. Then a similar graph is built iteratively with larger segments, i.e. syllable, word, phrase and utterance. This approach introduces dependencies with future segments that are not compatible with creating labels on-the-fly.

In MAGE, we replace the default NLP component in order to address this issue and enable reactive control of the phonetic content. Firstly, we reduce the amount of considered contextual information to past, current, future phoneme, previous and current syllable. When applied similarly during training and synthesis steps, we think that reducing the context has a rather limited impact on the output quality. Secondly, we replace the offline text interface by an online messaging layer, called reactive NLP or RNLP. With the RNLP, the phonetic content is reactively submitted to MAGE by *chunks*. Each chunk is a group of phonemes with the only condition of containing one vowel. The message contains the phoneme sequence, the included vowel and whether the vowel is stressed or not.

Based on these incoming messages, the contextual information is computed on-the-fly, in order to generate the labels that are used in the MAGE label thread. Since MAGE introduces one-label delay, we can read one phoneme in the future and use this future phoneme to compute the reduced contextual information required by the RNLP.

4. HANDSKETCH-BASED PROTOTYPE

In this work, we aim to integrate MAGE in a new musical instrument prototype, as a way to explore the concept of tangible speech synthesis. In this Section, we present the prototype that we have developed. It is based on the HandSketch interface, a tablet-based bi-manual controller [3], that was originally developed for singing synthesis control. Performing with HandSketch is a combination of pen-based drawing gestures on a vertical surface (for the preferred hand) and pressure-sensitive grips on the side of the device (for the non-preferred hand).

The pen-based gesture achieved on the HandSketch playing diagram is composed of various dimensions: polar coordinates on the fan (angle, radius), pen pressure, pen tilt in x and y directions, and two buttons on the pen. In the mapping between HandSketch and MAGE, we use angle on the fan to control pitch curves. Clicking on the front button of the pen switches between overwriting the pitch curve and deviating it by a given ratio. The radius on the fan controls the speech speed. Low radius values corresponds to slowing down and high values to speeding up. The pen pressure is naturally mapped to speech volume. Finally pen x tilt allows to change the vocal tract length, e.g. interpolate between male and female speakers.

The hand grips performed on the side of the tablet are captured with a network of eight Force Sensing Resistors (FSRs), five on the front side and three on the back side. This configuration allows the performer to grab the side of the instrument with various postures by changing the thumb position and the various combinations of 1–4 fingers on five sensors. In our current mapping with MAGE, these postures are mapped to various phonetic chunks.

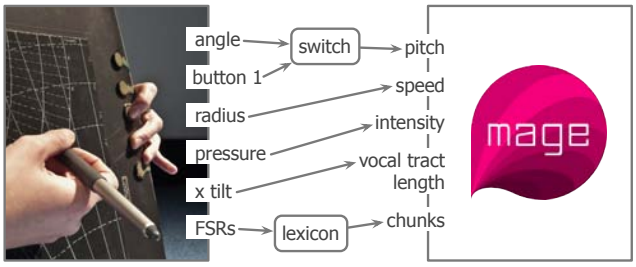

Figure 3: Mapping of the HandSketch control space to the parameters of the MAGE synthesizer.

In this musical instrument prototype, the MAGE speech synthesis functionalities are assembled and integrated as an openFrameworks application. This application continuously listens to incoming OSC messages, for receiving pitch, speed, volume, vocal tract length modifications and chunk messages. The /chunk OSC message follows this syntax :

```
/chunk blah ah 1 ... /chunk blaw aw 0
```

The values received through these OSC messages are then mapped to the appropriate MAGE function calls within the

OSC listener thread. The speech sound is also generated within the openFrameworks application. The mapping from HandSketch control dimensions to MAGE inputs is done in a Max patch. This patch handles the posture-to-chunk look-up table, the pressure thresholding, dimension scaling, and generating the OSC messages. A diagram of our musical instrument is given in Figure 4.

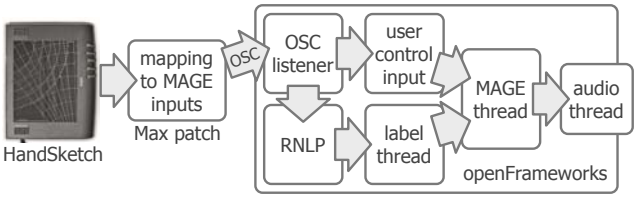

Figure 4: Architecture of the MAGE-based musical instrument prototype, enabling tangible control of speech properties with the HandSketch.

5. RESULTS

In this Section, we present some preliminary results that have been obtained along the prototyping process of the MAGE-based musical instrument. These evaluations were targeting the assessment of two different aspects of the instrument. On the one hand, we estimate the speech quality of MAGE and compare it to the original speech quality of HTS. On the other hand, we provide an informal feedback on the usability of the instrument, for the purpose of controlling artificial speech with bi-manual gestures.

The speech output quality of MAGE is compared to the speech output quality of HTS in two different ways. Firstly, we measure the distortion between the two outputs, evaluating Norden's criterion [10] on waveforms and Toda's criterion on production parameters [11]. In average, these measurements give respectively 2.6dB and 1.1dB, which is fairly low. Secondly, we conducted ABX hearing tests [1] on 59 participants. Our participants have on average 37% ($\pm$0.12%) of chance to make a mistake in discriminating between MAGE and HTS. It is an acceptable result, considering that 50% corresponds to total confusion.

Although we have not conducted any formal user study, our new musical instrument has been performed by several people. During these performances, two main qualities have been assessed. Primarily, it has been observed that the one-label delay does not impact on the ability to accurately control prosody. Furthermore, the chunk appears to be the appropriate time scale for triggering spoken content with finger gestures. However, the use of a lexicon is a limiting factor for the user to generate arbitrary sentences.

6. CONCLUSIONS

We think that there is a design space for high-quality performative speech synthesis applications, by using statistical parametric modeling. We proved that it is feasible to overcome the constrain of working at the sentence level, brought from the original system and transformed HTS into a reactive system that provides on-the-fly phonetic and prosodic control. MAGE is a very advanced and flexible speech synthesizer with built-in interactive properties. It can also bring the necessary tools to analyze various situations involving artificial speech production, mapping strategies for interactive prototypes and integrations in various devices. It also envisions the use of synthetic speech for more realistic social experience scenarios. The MAGE/ HandSketch approach is our first realization of tangible speech synthesis, and still among a small amount of speaking musical instruments. But there is a growing interest in using this approach to address remaining issues at various levels of speech production understanding, from acoustical to social. We aim to work on reinforcing graspability of speech production parameters and study the relevance of considered time scales. We also want to assess the interface design process by conducting comprehensive user studies.

7. ACKNOWLEDGEMENTS

Authors would like to thank the various financial and academic supports around the MAGE project: University of Mons, Région Wallonne (grant 716631), and Acapela Group S.A. Also, G. Wilfart and A. Moinet for their contribution.

8. REFERENCES

[1] D. Clark. High-resolution subjective testing using a double-blind comparator. *Journal of the Audio Engineering Society*, 30(5):330–338, 1982.

[2] P. Cook and C. N. Leider. SqueezeVox: A New Controller for Vocal Synthesis Models. In *Proc. of the International Computer Music Conference*, pages 5–8, 2000.

[3] N. d'Alessandro and T. Dutoit. Handsketch Bi-Manual Controller: Investigation on Expressive Control Issues of an Augmented Tablet. In *Proc. of the International Conference on New Interfaces for Musical Expression*, pages 78–81, 2007.

[4] H. Dudley. The Carrier Nature of Speech. *Bell System Technical Journal*, 19:495–515, 1940.

[5] S. Fels and G. E. Hinton. Glove-Talk II - A Neural-Network Interface which Maps Gestures to Parallel Formant Speech Synthesizer Controls. *IEEE Transactions on Neural Networks*, 9(1):205–212, 1998.

[6] S. Fels, R. Pritchard, and A. Lenters. ForTouch: A Wearable Digital Ventriloquized Actor. In *Proc. of the International Conference on New Interfaces for Musical Expression*, pages 274–275, 2009.

[7] A. J. Hunt and A. W. Black. Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database. In *Proc. of the IEEE International Conference on Audio, Speech and Signal Processing*, pages 373–376, 1996.

[8] D. McNeill, editor. *Language and Gesture: Window into Thought and Action*. Cambridge University Press, 2000.

[9] B. C. J. Moore, L. K. Tyler, and W. D. Marslen-Wilson, editors. *The Perception of Speech: From Sound to Meaning*. Oxford University Press, 2009.

[10] F. Norden and T. Eriksson. A Speech Spectrum Distortion Measure with Interframe Memory. In *Proc. of the IEEE International Conference on Audio, Speech and Signal Processing*, 2001.

[11] T. Toda, A. W. Black, and K. Tokuda. Mapping from Articulatory Movements to Vocal Tract Spectrum with Gaussian Mixture Model for Articulatory Speech Synthesis. In *Proc. of ISCA Speech Synthesis Workshop*, pages 31–36, 2004.

[12] W. von Kempelen. *Mechanismus der menschlichen Sprache nebst der Beschreibung seiner sprechenden Maschine*. J. B. Degen, Wien, 1791.

[13] H. Zen, K. Tokuda, and A. W. Black. Statistical Parametric Speech Synthesis. *Speech Communication*, 51:1039–1064, 2009.