Mage - Reactive articulatory feature control of HMM-based parametric speech synthesis

Maria Astrinaki, Alexis Moinet, Junichi Yamagishi, Korin Richmond, Zhen-Hua Ling, Simon King, and Thierry Dutoit. Mage - reactive articulatory feature control of HMM-based parametric speech synthesis. In 8th ISCA Workshop on Speech Synthesis, pages 227-231, Barcelona, Spain, August 2013.

Mage - Reactive articulatory feature control of HMM-based parametric speech synthesis

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 integration of articulatory control into MAGE, a framework for realtime and interactive (reactive) parametric speech synthesis using hidden Markov models (HMMs). MAGE is based on the speech synthesis engine from HTS and uses acoustic features (spectrum and f_0) to model and synthesize speech. In this work, we replace the standard acoustic models with models combining acoustic and articulatory features, such as tongue, lips and jaw positions. We then use feature-space-switched articulatory-to-acoustic regression matrices to enable us to control the spectral acoustic features by manipulating the articulatory features. Combining this synthesis model with MAGE allows us to interactively and intuitively modify phones synthesized in real time, for example transforming one phone into another, by controlling the configuration of the articulators in a visual display.

Index Terms: speech synthesis, reactive, articulators

1. Introduction

For human beings, speech is one of the richest and most sophisticated modalities used for communication. It involves complex production and perception mechanisms and it varies in quality. It is a highly reactive and interactive process, with complex timing, involving the all the articulators (tongue, lips, jaw, lungs, etc.), even the hands. Artificially synthesized speech has been explored for decades and several methods have been developed, such as formant synthesis [1], diphone synthesis [2], articulatory speech synthesis [3], unit selection synthesis [4] and statistical parametric synthesis [5] resulting in various Text-To-Speech (TTS) systems. Nowadays, TTS systems are very intelligible and natural, and can be expressive, but they are also static, they do not support user interaction and they are not sensitive to environmental conditions. Although there has been great progress in terms of intelligibility and naturalness, there is still place for improvement when considering interaction. This "deaf" design of conventional TTS systems is limiting the involvement of the user and the vocal expression influenced by its environment, leaving no room for expression and creativity.

In recent years, there has been an emerging interest in applications that need reactivity and expressivity in speech production. There are different ways of approaching the idea of working beyond TTS. One approach to move further than the standard TTS paradigm comes with MAGE [6], one of the first methods proposed for reactive HMM-based speech and singing synthesis. It is a modified version of the HMM-based parametric speech synthesis approach that has become a mainstream speech synthesis method [7], [8]. MAGE allows reactive control of prosody, context, speaking style and speech quality. It is able to synthesize highly intelligible and smooth speech, it is flexible, and supports adaptation and interpolation methods, combined with a very small footprint and computational weight; advantages inherited from the original system, HTS [9]. However, the quality of the output is constrained by the quality of the training data and the user controls. It is still difficult, even for trained users to produce meaningful expressivity due to the complexity of the speech itself and to the abstract representation of speech through statistical models.

The controls that have been available, have been over the acoustic features, as used in the conventional HMM-based speech synthesis, which are parameters required by a vocoder. Such parameters do not necessarily have a "physical" or "intuitive" meaning to the user. However, the physical nature of human speech production means that an articulatory parameterization of speech has interesting properties. The articulatory features describe the quantitive positions and the continuous movements of a group of human articulators, such as tongue, jaws, lips, velum. Such features have relatively slow and consistent evolution through time, they are not influenced in the same way by acoustic noise and other environmental conditions. They can provide a straightforward and simple explanation for speech characteristics and they provide meaningful interpretation of the speech production to the user.

A method for integrating articulatory features in HMM-based speech synthesis has already been proposed [10], [11], where the articulatory features were recorded using electromagnetic articulography (EMA). In this method, a unified acoustic-articulatory model is trained and a piecewise linear transform is adopted to model the dependency of the acoustic features on the articulatory features. During synthesis, the articulatory features are generated from the previously trained models. Then, these generated articulatory features can be manipulated in arbitrary ways which in turn affect the generation of acoustic features. In this way, the characteristics of the synthetic speech can be controlled via an articulatory representation. The motivation of the work described here is to see wether reactive articulatory control is possible, to evaluate the results and then to explore the potential and the possibilities of different user applications

(see Section 2.2) that take advantage of the physical nature and stability of the articulators.

The paper is organized as follows. Section 2 gives a brief overview of the reactive HMM-based speech synthesis approach called MAGE. Section 3 describes our proposed method in detail. Section 4 describes the articulatory control application we developed as a proof of concept as well as the challenges faced for the evaluation and testing. Section 5 presents the future targets and Section 6 gives the conclusions of this work.

2. Reactive HMM-based speech and singing synthesis

MAGE is based on the HMM-based parametric speech synthesis method, which it extends in order to support realtime architecture and multithreaded control. As it is based on HTS, it inherits its features, advantages and drawbacks [5]. The contribution of MAGE is that it opens the enclosed processing loop of the conventional system and allows reactive user control over the available contextual information, the speech prosody and speaking style and quality. Moreover, it provides a simple C++ API, allowing reactive HMM-based speech synthesis to be easily integrated into reactive and realtime frameworks [12], [13]; run in various devices and create different prototypes [14], [15].

2.1. Overview of MAGE

One important feature of MAGE is that it uses multiple threads, and each thread can be affected by the user which allows accurate and precise control over the different production levels of the artificial speech. As illustrated in Figure 1, MAGE integrates three main threads: the *label thread*, the *parameter generation thread* and the *audio generation thread*. Three queues are shared between threads: the *label queue*, the *parameter queue* and the *sample queue*.

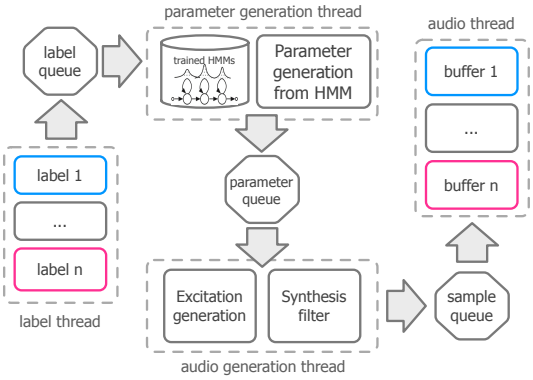

Figure 1: MAGE: reactive parameter generation using multiple threads and shared queues.

Briefly, the *label thread* controls the input sequence of the phonetic labels, by pushing the received phonetic labels onto the *label queue*. Then, the *parameter generation thread* reads from the *label queue* one phonetic label at a time. For that single label the speech parameters are generated (sequences of spectral and excitation parameters including first and second derivatives of the static features), which are locally-maximized using only the current phonetic label and, if available, the two previous labels. In other words, for every single input phonetic label, the feature vectors are estimated by taking into account the HMMs of these specific labels and the input user controls. The generated speech parameters are stored in the *parameter queue*. As shown in [16], the impact of the local maximization of the generated parameters is very small. Then, the *audio generation thread* will generate the actual speech samples and store them in the *sample queue* so that the system's *audio thread* will access them and deliver them to the user. Further details of the MAGE reactive parameter estimation can be found in [17].

Accessing and controlling every thread has a different impact over the synthesized speech. The *label thread* can provide contextual phoneme control, the *parameter generation thread* can reactively modify the way the available models are used for the parameter sequence generation [17], and finally the *audio generation thread* manipulates reactively the vocoding of every sample, resulting in prosody and voice quality controls. The delay in applying any received user control varies between a single speech sample and a phonetic label depending on the thread that is being accessed.

2.2. Potential applications

MAGE aims to combine simple prototyping with meaningful gestural control, through an appropriate mapping and to bring synthetic speech to performative use cases. It can enable a broad set of new types of design and architectures for speech synthesis applications, such as silent speech communication, entertainment and gaming, assistive applications for speech impaired people and performing arts.

Several real life and scientific applications target the creation and use of unique personalized voices, with certain speech characteristics. For example, in the field of new interfaces for musical expression and performing arts, it would be possible to create a voice that has a very specific speaking style that could also gradually change during the performance or adapt to the feedback of the audience. Regarding avatars and gaming applications, users would be able to select, customize and refine the voice used by their avatar. It could also be used in a GPS, where the accent of the used voice changes depending on the position, as the user is driving through a country. The same principle applies to movie dubbing applications or assistive communication devices for speech impaired people, where a voice can be adapted and personalized according to the past and the personality of every person or character.

We believe that by adding articulatory control to MAGE not only will the range of the potential applications be enriched, but also it will be possible to achieve a deeper understanding of articulatory speech production mechanisms. Applications targeting the fields of speech pedagogy, linguistics and speech therapy in particular will be able to help people come to understand how certain phones are produced at the articulatory level, providing instant acoustic feedback. Our attempt to implement this is discussed in the following sections.

3. Reactive articulatory feature control

In this work, MAGE is modified in order to generate and alter articulatory features. Given the unified acoustic-articulatory model and a set of phonetic labels, it is possible to reactively generate the target speech samples. Simultaneously, it is possible to influence the generated articulatory features by replacing them 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.

3.1. Feature-Space-Switched Multiple Regression HMMs

In HMM-based speech synthesis, a sequence of contextual phonetic labels is used to predict an optimal state sequence and the duration, in frames, of each state. In the case of acoustic features (i.e. spectral parameters), a state j corresponds to a multivariate Gaussian distribution whose parameters are $\boldsymbol{\mu}_j$, its mean vector, and $\boldsymbol{\Sigma}_j$, its covariance matrix. Given these parameters, the sequence of states and their durations are used to generate an optimal sequence of acoustic features $\boldsymbol{X}$ that are then combined with synthetic source parameters to synthesize speech using, for instance, a vocoder. Note that in MAGE, as presented in Section 2, the state sequence and the computation of parameters are performed locally, label by label, as opposed to the one-pass approach that we have just described.

This framework for HMM-based parametric speech synthesis has been expanded in [11] so that the acoustic models become dependent on articulatory features. One of the methods presented is called "feature-space-switched multiple regression HMM" (FSS-MRHMM). MRHMM consists in replacing the mean vector $\boldsymbol{\mu}_j$ of each state by a linear combination of synthetic articulatory features $\boldsymbol{\xi}_t$ and $\boldsymbol{\mu}_j$, before computing the optimal sequence of acoustic features. Therefore, the Gaussian distribution for each frame t is defined as

$$b_j(\boldsymbol{x}_t|\boldsymbol{y}_t) = \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{A}_t\boldsymbol{\xi}_t + \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \quad (1)$$

with $\boldsymbol{x}_t$ a vector of static acoustic features and their first and second derivatives. $\boldsymbol{A}_t$ is the articulatory-to-acoustic projection matrix and $\boldsymbol{\xi}_t$ is an expanded articulatory feature vector, which means it contains $[\boldsymbol{y}_t^\intercal, 1]$, with $\boldsymbol{y}_t$ a vector of static articulatory features and their first and second derivatives. Normally, $\boldsymbol{y}_t$ is generated using standard HMM-based synthesis with its own specific models of articulatory features. However, as explained in Section 3.2, it can also be replaced by other values, thus modifying the identity of the synthetic phones.

In the particular case of FSS-MRHMM, a finite set of M matrices $\{\boldsymbol{A}_1, \ldots, \boldsymbol{A}_M\}$ is trained along with an M-mixture Gaussian mixture model (GMM) of the articulatory space. Then, at synthesis time, instead of a single Gaussian the probability density function of each state is written as

$$b_j(\boldsymbol{x}_t|\boldsymbol{y}_t) = \sum_{k=1}^{M} \zeta_k(t)\mathcal{N}(\boldsymbol{x}_t; \boldsymbol{A}_k\boldsymbol{\xi}_t + \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \quad (2)$$

where $\zeta_k(t)$ is the probability for the mixture component k given $\boldsymbol{y}_t$. However, with such a model, the parameter generation would require to use an EM-based iterative estimation. This cannot be applied in the context of a reactive application such as MAGE and we simplified it by considering only the mixture with maximum $\zeta_k(t)$, as proposed in [11]. Therefore, Equation 2 is rewritten as

$$k_t = \underset{k=[1,\ldots,M]}{\operatorname{argmax}}\ \zeta_k(t) \quad (3)$$

$$b_j(\boldsymbol{x}_t|\boldsymbol{y}_t) = \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{A}_{k_t}\boldsymbol{\xi}_t + \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \quad (4)$$

Further details on the training of the different models (acoustic, articulatory, GMM and $\boldsymbol{A}_k$) can be found in [11].

3.2. Reactive synthesis using articulatory features

During synthesis, a given sequence of context-dependent phonetic labels is used to concatenate the context-dependent HMMs. The articulatory and acoustic features are then predicted from the sentence HMMs by means of a maximum output probability parameter generation algorithm that incorporates dynamic features. It is possible though during synthesis that $\boldsymbol{\xi}_t$, the generated articulatory features, may be modified or replaced either by user input or according to phonetic knowledge (as explained in [11]). Hence, the corresponding acoustic features are regenerated, using Equations 3 and 4, in order to reflect those articulatory changes. The speech waveform is then synthesized from the generated mel-cepstral and f_0 parameter sequences using Mel Log Spectrum Approximation (MLSA) filter [18], with pulse-train (voiced frames) or white-noise excitation (unvoiced frames).

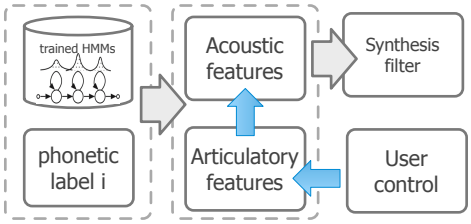

Figure 2: Generation of acoustic features with articulatory control using acoustic-articulatory model and user input controls.

As explained in Section 2, only the current phone label and, if available, the two previous labels are taken into account for the feature generation and therefore the generated parameter trajectories are locally-maximized. Here, for every phone label, the articulatory features are generated, taking into account the control of the user over the articulators (if any). Then, the acoustic features are generated in order to correspond to these articulatory features, as illustrated in Figure 2. Finally, the speech waveform is synthesized. Note that the feature generation is taking place in the *parameter generation thread*, and therefore the application of user control has a delay of one phone label.

The aim of this approach is to use the articulatory features in order to replace to some extent the predefined context and to modify the acoustic features accordingly. In other words, 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.

4. Reactive articulatory control application

To evaluate the proposed method requires an implementation that combines a graphical user interface (GUI) with the MAGE synthesizer. Such an application[1] is essential for multiple reasons. First and foremost, it allows us to assess the quality of the final speech samples. But, moreover we can explore how this output is influenced by fast changing articulatory inputs as well as how proficient users must be at controlling such features.

4.1. Graphical user interface design

The design of the graphical user interface was highly dependent on the database we used for the reactive synthesis. In this work for our experiments and reactive synthesis we have used a multichannel articulatory database containing the acoustic waveform

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

recorded concurrently with EMA data of a male British English speaker. Six EMA receivers were used, and for each receiver three dimensional coordinates were recorded as described in [10]. However, only two dimensions were used in the experiments here (front-to-back and bottom-to-top) resulting in a total of 12 static articulatory features.

Based on these six EMA points with two dimensional movements, we designed the GUI illustrated in Figure 3. The GUI depicts a two dimensional midsagittal view of the vocal tract drawn using 124 points. The six EMA points are represented as white circles placed on the articulators as described in [10] (indicated by the red arrows). 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.

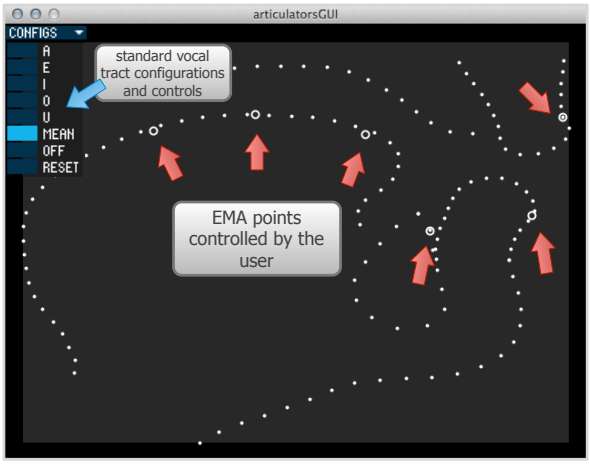

Figure 3: Instance of the graphical user interface showing the available configurations and controls, the six EMA points (red arrows) that can be reactively moved and predefined configurations of vocal tracts that can be applied (blue arrow).

On the left of the interface there is a menu listing predefined vocal tract shapes and EMA sets. These predefined configurations of vocal tract shapes were obtained from speaker-dependent magnetic resonance imaging (MRI) scans and electromagnetic articulography (EMA) data, as described in [19]. The GUI shows vocal tract shapes from models trained on MRI scans combined with EMA points translated in the MRI coordinate space. When the user selects one of these vocal tract and EMA configurations, his previous controls are instantly overwritten and the selected shape is displayed. The user by selecting one EMA point from the GUI is able to move it in the two dimensional MRI coordinate space. The current shape of the vocal tract it is not reactively altered so that the user will have a reference point to the initial configuration chosen. However, it is possible to transform interactively the shape of the vocal tract given the user controls by using specific transformation matrixes. Note here that the controls of the user take place in the MRI space (GUI) and not in the EMA space (synthesis).This means that the controls of the user have to be appropriately transformed in order to be an acceptable input for MAGE. This task is controlled by the interface.

The interface also allows to reset the synthesis by clicking on the "reset" button. It is also possible to stop using the reactive articulatory controls and use the generated articulatory data by clicking on the "off" button.

4.2. Synthesis

The final part of the application, generating the speech waveform, is implemented by the MAGE reactive speech synthesizer. The graphical user interface sends to MAGE the modifications applied by the user through open sound control (OSC) messages [20].When received, these 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 with only one phonetic label delay, as explained in Section 3. Let us note here that the articulatory features provided by the user overwrite the estimated articulatory features, $\boldsymbol{\xi}_t$ used in Equation 4. Therefore, the first and second derivatives used are the ones from the contextually estimated articulatory features.

4.3. Challenges

One of the aims of this work is to allow the user to reactively control the speaking style as well as the content by modifying the articulatory features. However, the movement of the articulators is so fast that the user is not able to input the expected movements fast enough through the interface. Therefore, instead of trying to contextually manipulate a full phrase we decided to try and transform only one vowel into another vowel. This simplifies the problem of the fast changing articulatory features but introduces the problem of the duration of the synthesized vowel. If it is synthesized with the standard model duration is too short to be intelligible. Hence, in this case we synthesize long vowels by increasing the generated duration of every state, using a bigger scale factor for the stable state (i.e. 10, 20 or 50), followed by a long pause.

The EMA points can be placed at any position and with any speed. This results in movements of the articulators that do not always respect the "physiology" and "mechanisms" of the human articulators. In other words, these input movements have not been "seen" during the training phase and, consequently, the models cannot accurately estimate them. Hence, when these "new" articulatory features will be used to generate the acoustic features it is highly probable that they will give an "unstable" result. In order to tackle the problem of the extended contextual control and to minimize the possible instabilities caused by the extreme movements of the articulators, MAGE constrains the contextual control only over the voiced frames.

As a test case, the user is asked to listen to a vowel synthesized using phonetic labels and by moving the six EMA points reactively from the interface he transforms it into another target vowel. However, after some exploratory tests we see that this approach requires from the user to move the six EMA points, in the two dimensional MRI space (12 degrees of freedom) accurately enough so that he will achieve the acoustic target. Such a task is very demanding and rather difficult, and in most cases the user does not reach the target. What makes this task more difficult is that we use context-dependent model. Although the FSS-MRHMM approach can determine the regression matrices without using context information, the $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ as shown in Equation 4 are still context-dependent. More specifically, the required modifications over the articulatory features (EMA points) in order to acoustically achieve a target, differ depending on the initial phone synthesized using the provided label.

A solution to this would be to use "tailored" context features, where the vowel identities are removed from the question set for acoustic model clustering and therefore better articulatory controllability can be achieved [11].

5. Future work

Based on the preliminary testing of the application we see some essential modifications regarding the interface. It is important to either decrease the degrees of freedom available to the user or to allow the manipulation of only some EMA points while the system provides the correct coordinates for the remaining ones. For example, a case would be where the user is allowed to manipulate only the three EMA points over the tongue. while the remaining three are automatically adjusted. Providing a "color-coded map" denoting the "accepted" or "suggested" regions of every EMA point could advice or guide the user to choose suitable coordinate sets. This will help the user to have a better understanding of the required modifications of the articulators in order to achieve the acoustically desired target. However, such a simplification of the interface must by all means be combined with using "tailored" context feature during synthesis for better articulatory controllability.

Based on such a framework, it would be meaningful to conduct user studies and listening tests. The user studies will show us how users manipulate the acoustic space by means of articulatory control as well as how skilled a user should be. The listening tests will help us to measure how other listeners perceive the result of these manipulations. Initially, as explained above, the user is asked to transform a given vowel to a target vowel only by controlling the articulatory features. The success of the user will be determined by objectively evaluating the acoustic and articulatory features generated by taking into account or not the user input for the target vowel. The same test can be conducted by using monosyllabic words embedded into a carrier sentence in order to conduct a second vowel identity modification experiment. Then, it would be interesting to see how other users perceive these modifications, and in addition to the objective evaluation, we would like to perform also some listening tests to subjectively evaluate performance on the vowel modification task.

6. Conclusions

In this paper we have presented a method that enables reactive articulatory control over HMM-based parametric speech synthesis using the MAGE framework. We present also an application that enables the user to reactively control the position of the articulators through a graphical user interface. We see that reactive articulatory control is feasible, and combined with an interface allows us to explore different aspects of the speech production. However, we realize that the manipulation of the articulators by the user, even though it seems rather straightforward, is very demanding and difficult. It is very easy to transform a phone into another random phone while experimenting with the interface, but it becomes rather complicated when a specific target vowel modification is asked. There are aspects of the system that would benefit from improvement. Currently, there are no restrictions over the user manipulation patterns, but probably limiting or "guiding" the possible user controls might lead in more distinguishable vowel modifications. Subjective and objective evaluations are essential in order to assess the final quality of the system. By conducting user studies we want to evaluate the efficiency of the user to achieve certain vowel or monosyllabic word targets. Through listening tests we want also to evaluate how the output of these reactive modifications over the articulatory features is perceived acoustically.

7. References

[1] R. Carlson and B. Granstrom, "A text-to-speech system based entirely on rules," in *Proc. of ICASSP*, 1976.

[2] F. Charpentier and M. Stella, "Diphone synthesis using an overlap-add technique for speech waveforms concatenation," in *Proc. of ICASSP*, vol. 11, 1986, pp. 2015–2018.

[3] B. Brent and W. J. Strong, "Windbag – a vocal-tract analog speech synthesizer," *Acoustical Society of America*, vol. 45, 309(A), 1969.

[4] A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. of ICASSP*, 1996, pp. 373–376.

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

[6] M. Astrinaki, A. Moinet, G. Wilfart, N. d´Alessandro, and T. Dutoit. (2010, September) Mage platform for performative speech synthesis. [Online]. Available: http://mage.numediart.org/

[7] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," *Proc. Eurospeech*, vol. 83, no. 11, pp. 2347–2350, 1999.

[8] K. Tokuda, H. Zen, and A. Black, "HMM-based approach to multilingual speech synthesis," *Text to speech synthesis: New paradigms and advances*, pp. 135–153, 2004.

[9] K. Oura. (2010, Sept) HMM-based speech synthesis system (hts). [Online]. Available: http://hts.sp.nitech.ac.jp/

[10] Z.-H. Ling, K. Richmond, J. Yamagishi, and R.-H. Wang, "Integrating articulatory features into HMM-based parametric speech synthesis," *IEEE Transactions On Audio Speech And Language Processing*, vol. 17, no. 6, pp. 1171–1185, 2009.

[11] 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.

[12] M. Puckette. (2009, September) Pure data. [Online]. Available: http://puredata.info/

[13] Z. Lieberman, T. Watson, A. Castro, and etc. (2009, September) openframeworks. [Online]. Available: http://www.openframeworks.cc

[14] R. A. Clark, M. A. Konkiewicz, M. Astrinak, and J. Yamagishi, "Reactive control of expressive speech synthesis using kinect skeleton tracking," Tech. Rep. 30, December 2012.

[15] M. Astrinaki, A. Moinet, N. d´Alessandro, and T. Dutoit, "Pure data external for reactive HMM-based speech and singing synthesis," *Proc. of the 16th Int. Conference on Digital Audio Effects (DAFx-13)*, 2013.

[16] M. Astrinaki, N. d'Alessandro, B. Picart, T. Drugman, and T. Dutoit, "Reactive and continuous control of HMM-based speech synthesis," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*, 2012, pp. 252–257.

[17] 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 the 13th Conference on New Interfaces for Musical Expression (NIME´13)*, 2013.

[18] K. Sumita and R. Members, "Mel log spectrum approximation (mlsa) filter for speech synthesis," *Electronics and Communications in Japan*, vol. 6, no. 2, pp. 10–18, 1983.

[19] K. Richmond and S. Renals, "Ultrax: An animated midsagittal vocal tract display for speech therapy," in *Proc. Interspeech*, 2012.

[20] A. Schmeder, A. Freed, and D. Wessel. (2009, September) opensoundcontrol.org. [Online]. Available: http://opensoundcontrol.org