REACTIVE AND CONTINUOUS CONTROL OF HMM-BASED SPEECH SYNTHESIS

Maria Astrinaki, Nicolas d'Alessandro, Benjamin Picart, Thomas Drugman, Thierry Dutoit

TCTS Lab, University of Mons, Belgium

ABSTRACT

In this paper, we present a modified version of HTS, called performative HTS or pHTS. The objective of pHTS is to enhance the control ability and reactivity of HTS. pHTS reduces the phonetic context used for training the models and generates the speech parameters within a 2-label window. Speech waveforms are generated on-the-fly and the models can be reactively modified, impacting the synthesized speech with a delay of only one phoneme. It is shown that HTS and pHTS have comparable output quality. We use this new system to achieve reactive model interpolation and conduct a new test where articulation degree is modified within the sentence.

Index Terms— speech synthesis, HTS, reactive control

1. INTRODUCTION

For more than a decade, speech synthesis based on statistical parametric modeling [1] has constantly improved. Particularly, the use of context-dependent Hidden Markov Models (HMM), such as in HTS [2], has led to a reasonably high-quality output; offering the more preferred (through MOS tests) and more understandable (through WER scores) synthesis [3]. HTS is a publicly available toolkit [4], that nowadays, leads to a growing interest in the speech synthesis community. We think that one important aspect for these new applications to emerge is to work beyond the Text-To-Speech (TTS) paradigm. Indeed, the common conversion from text to speech sound at the sentence level leads to a narrow set of text-reading applications. We think that new applications in speech synthesis can benefit from major improvements in what kind of data speech is generated from and how it is generated. We envision two main application categories where this approach can be relevant:

- *Context-reactive speech synthesis*: as in a real dialogue situation, the spoken content considerably adapts to surrounding conditions [5]. Therefore the speech synthesizer needs to be *reactive*, i.e. adapting the speech production parameters to the continuously evolving context, as the speech sound is generated.
- *Performative speech synthesis*: speech sound primarily results from a continuous articulatory gesture. Therefore the ability to create synthetic speech from a more gestural and *performance-based* input than text – like hand postures in [6] – can lead to using speech synthesis more expressively in some situations.

Current speech synthesis algorithms are unable to provide the level of controllability and reactivity that is required to support these new categories of speech synthesis applications. Indeed, either in non-uniform unit selection or statistical parametric modeling, the whole sentence is always chosen as the target on which the optimization is achieved, respectively by minimizing a cost function [7], [8] or maximizing the likelihood [9]. Moreover the increasing amount of linguistic information being used in speech database clustering techniques [7], [10] increases the dependency between the current speech segment to be synthesized and its phonetic context, and particularly with the future. This non-causal approach to speech synthesis is a drawback when it comes to envision reactive solutions. However we think that HMM-based speech synthesis is among the most promising techniques in order to explore such reactivity and controllability. Indeed it has been shown in many related studies [11], [12] that the generation of speech parameters from HMMs is a flexible process, leading to intelligible speech with a limited computational cost.

The issue of continuous control in HMM-based speech synthesis has already been addressed in various ways. We can highlight the use of articulatory data in HTS to provide a more continuous phonetic space [13]. Model interpolation has also been tested for changing the speaking style [14], [15]. In opposition, the aspect of reactivity in HMM-based speech synthesis has barely been considered. We can only highlight the consideration of computing HMM-based trajectories in real-time in [2], i.e. as fast as the duration of the corresponding sound.

In this paper, we introduce a modified HTS engine called performative HTS (pHTS). pHTS computes the synthetic speech reactively, i.e. enabling the HMM-based generation of speech parameters on the fly, independently for each label and impacting on the ongoing generated sound with only one label of delay. In Section 2, we describe pHTS. Then we present in Section 3 a new effect enabled by pHTS: the reactive and continuous interpolation of articulation degree. In Section 4, we compare HTS and 2 versions of pHTS. In Section 5, we evaluate our new interpolation and in Section 6 we present our conclusions.

2. HMM-BASED SPEECH SYNTHESIS WITH REDUCED PHONETIC CONTEXT

In HMM-based speech synthesis the required phonetic context is related to the whole sentence, which prevents to adapt to any external solicitation within the sentence. We assume that reducing this phonetic context is necessary to reactively control the speech synthesis. Thus, we decided to build the *performative* version of HTS, called *pHTS*. Such a system requires to alter two important aspects of HTS. First, reducing the context considered during the training phase (Section 2.1). Secondly, developing a new approach for the generation of speech parameters with a smaller look-ahead window (Section 2.2).

2.1. Training HMMs with Reduced Phonetic Context

In HTS, labels are used to describe phonemes with all their contextual dependencies. Each phoneme depends on its preceding/succeeding ones and on the larger segments in which this phoneme is contained, e.g. the syllable [16]. Iteratively, for every larger segment, i.e. syllable, word, phrase and utterance, a similar dependency graph is built. Along with these dependencies, various pieces of information are available: relative position of the current segment in the larger containing one (e.g. relative position of the phoneme in the syllable), and the possible presence of an accent or a stress on the current and/or the surrounding segments [2]. Figure 1 gives a summary of these dependencies in form of a matrix.

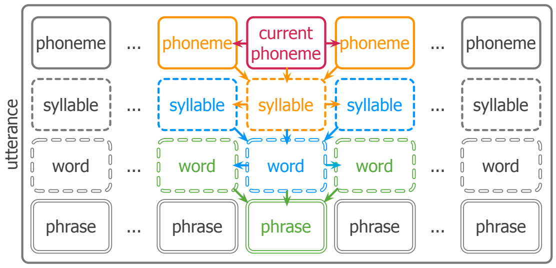

Fig. 1. *Summary graph of the contextual dependencies of the currently produced phoneme inside a given utterance.*

In pHTS, we reduce the phonetic context used at the training stage. All the contextual information related to segments larger than the syllable, i.e. word and phrase, is discarded. The amount of considered linguistic information is now limited only to past, current, future phoneme, previous and current syllable. The whole database of context-dependent HMMs is then retrained in consequence. Note that for the new training we use the standard implementation of the HTS toolkit available in [4], with a reduced set of questions.

2.2. Short-Term Speech Parameter Trajectories

During HTS synthesis, a given sequence of words is converted into a context-dependent label sequence, which is used to concatenate the context-dependent HMMs. The speech parameters – sequences of spectral and excitation parameters – are generated by maximizing the probability of the speech parameter sequence, considering this sequence of concatenated HMMs as the underlying statistical model [2]. Hence, the accessible time scale software-wise is this sequence of targeted words. The speech waveform is then synthesized from the generated mel-cepstral and $F0$ parameter sequences by using Mel Log Spectrum Approximation (MLSA) filter [17], with pulse-train or white-noise excitation.

In pHTS, we realize the parameter generation process on a sliding window of two labels. It means that, for each new label received by the system, only the HMMs corresponding to that label and the previous one are concatenated. Then the speech parameter trajectories corresponding the previous label are generated from these two HMMs, using the maximization in Equation 2, as used in [2]. One consequence of this process is that the resulting parameter trajectories do not correspond to the overall maximum of probability, but only the concatenation of locally-maximized speech parameters.

$$q^\star = \underset{q}{\operatorname{argmax}}\, P(q \mid \boldsymbol{\lambda}^\star, \hat{\boldsymbol{T}}) \tag{1}$$

$$\hat{\boldsymbol{O}} = \underset{O}{\operatorname{argmax}}\, P(O \mid q^\star, \boldsymbol{\lambda}^\star, \hat{\boldsymbol{T}}) \tag{2}$$

where O and q are respectively the sequence of speech parameters and the sequence of states, $q^\star$ and $\lambda^\star$ respectively the estimated sequence of states and the concatenated left-to-right HMMs corresponding to the 2-label window, $\hat{O}$ the sequence of locally-maximized generated speech parameters, and $\hat{T}$ is the time frame corresponding to the first label of the 2-label window on which $\hat{O}$ is computed.

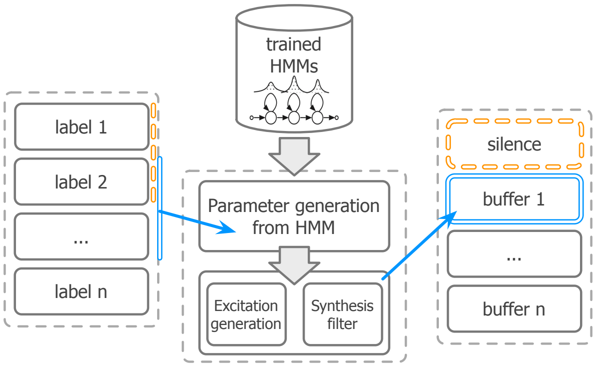

Fig. 2. *pHTS synthesis, using a 2-label sliding window to generate the speech parameter trajectories and audio buffers.*

pHTS opens the enclosed synthesis loop of HTS and reduces the accessible time scale to the phoneme level. Indeed, once the speech parameters have been generated for the delayed input label from the sliding window, the corresponding speech sound can be synthesized right away and stored in a buffer. Within an appropriate real-time audio software architecture, it means that the sound can be synthesized on the fly, as illustrated in Figure 2. As a result, any kind of modification which is achieved on the models during the generation of the speech parameters has an impact on the corresponding speech sound with a delay of only one label. It means that either the pitch curve, the phoneme durations or the spectral envelopes can be altered in many different ways, while the synthesis process is running and impact on the speech sound with only one phoneme of latency.

3. REACTIVE AND CONTINUOUS MODEL INTERPOLATION

pHTS enables the reactive and continuous modification of the models at the label level. If this feature leads to interesting results while directly applied on pitch curve and durations, the leap forward comes with using it with more complex model transformations. In this work, we want to evaluate how some existing speaker adaptation techniques can benefit from our new reactive and continuous control.

Speaker adaptation can be performed by adapting an average voice model to a specific target speaker [2]. There are two major techniques in the speaker adaptation domain: Maximum A Posteriori (MAP) [18] estimation and Maximum Likelihood Linear Regression (MLLR) [19]. In [14] and this work, our average voice model corresponds to the standard neutral HMM model. It is then adapted using Constrained Maximum Likelihood Linear Regression (CMLLR) transform with hypo/hyperarticulated speech data [15]. “Hyperarticulated speech” refers to the situation of a teacher/speaker talking in front of a large audience (important articulation efforts have to be made to be understood by everybody). “Hypoarticulated speech” refers to the situation of a person talking in a narrow environment or very close to someone (few articulation efforts have to be made to be understood). Due to the parametric representation of speech production (segmental and suprasegmental) used in HMM-based synthesis, interpolation between these different speaking styles is possible. This technique allows us to synthesize speech sentences on a continuous articulation degree scale [15].

When integrated into pHTS, this model interpolation algorithm can be applied at the label level and impact the resulting speech sound reactively with a delay of only one label. It means that we can vary the articulation degree on the continuous scale as the speech signal is being synthesized and heard by the user. This property enables the application of variable hypo/hypoarticulated speech in quickly-changing use cases, e.g. the listener talking back to the system, sudden background noise, various under interaction gestures, etc. Note, that real-time control of the articulation degree is just an illustration of the real-time interpolation concept. This approach can be applied potentially to any model interpolation scheme.

4. COMPARATIVE EVALUATION OF HTS/PHTS

Objective and subjective experiments were conducted to evaluate the similarity between HTS and two versions of pHTS. We call pHTS-1 the system with only the modifications described in Section 2.2. We call pHTS-2 the system with both modifications described in Section 2.2 and the reduced context at training time, as described in Section 2.1.

4.1. Experimental Protocol

For our tests, we used the speaker-dependent training demo in English that is provided in [4], applying the SLT female speaker and the BDL male speaker from the CMU ARCTIC database. All training sentences included in the demo were used to train our system.

For the objective evaluation, we synthesized 40 sentences based on the phonetic labels provided in the demo, which were not used for training. Speech signals were sampled at 16 kHz and the traditional Mel Generalized Cepstral (MGC) coefficients (with α = 0.42, γ = 0 and order of MGC analysis = 24) were obtained. Speech waveforms were synthesized with MLSA filtering. Note that state durations vary between HTS and pHTS when using full or reduced context HMMs, since durations are modeled differently in every approach. In order to compare our results on a frame-by-frame basis, both synthesizers were forced to use the original phoneme durations, ensuring that all streams are synchronous and comparable.

4.2. Objective Evaluation

In order to evaluate the distortion introduced by pHTS-1 and pHTS-2, we make use of objective quality measurements by applying two different metrics, previously used in related research [14]. The first metric used is the *mel-cepstral distortion* (Mel-CD - expressed in decibel); a distance measure calculated between the target and the estimated mel-cepstrum. Table 1 shows the mel-cepstral distortion averaged over the test sentences, for both female and male speakers using a 95% confidence interval.

Table 1. *Mean mel-cepstral distortion [dB] introduced by pHTS-1 and pHTS-2 using a 95% confidence interval.*

Mel-CD (dB)	**Male**	**Female**
pHTS-1	0.92 ± 0.17	0.96 ± 0.15
pHTS-2	1.12 ± 0.15	1.31 ± 0.22

Results indicate that both pHTS-1 and pHTS-2 lead to a segmental quality that is very close to the HTS output obtained on the same sentences. Indeed, 1 dB is usually accepted as the difference limen for spectral transparency [20].

The second metric used is the *root-mean-square* (RMS) error of $\log F0$ (expressed in cent). Note that the RMS error is computed for regions where both data models are considered to be voiced, since $\log F0$ is not observed in unvoiced regions. Table 2 shows the obtained results using a 95% confidence interval for both SLT and BDL speakers when comparing HTS to pHTS-1 and pHTS-2.

Table 2. *Root-mean-square error of* $\log F0$ *[cent] with 95% confidence interval between HTS and pHTS-1 / pHTS-2.*

RMSE (cent)	**Male**	**Female**
pHTS-1	79.64 ± 0.10	50.91 ± 0.30
pHTS-2	106.63 ± 0.20	74.16 ± 0.30

The resulting $\log F0$ RMS error is no more than one semitone (100 cent) in average. We know that 25 cent is the smallest noticeable pitch difference for pure tones [21], so we might conjecture that 100 cent is a noticeable interval, although it appears to be still within a limited range. It is known that measuring distortion on the $F0$ trajectory (or any speech parameter trajectory) only gives a partial understanding of how the suprasegmental quality is affected by the performative modifications. Therefore this objective evaluation must be completed with a set of subjective tests.

4.3. Subjective Evaluation

The subjective evaluation of the two proposed systems was conducted via a Comparative Mean Opinion Score (CMOS) test. Listeners grade the perceived quality of a speech signal in relation to a reference speech signal. The CMOS scale is a 7-point scale ranging from -3 for “much worse” to +3 for “much better” with 0 for “about the same”. A group of 33 people participated to the test, among which speech and non-speech experts, native and non-native English speakers. Each test was composed of 30 randomly-chosen sentences equally separated in 3 parts: comparing HTS to pHTS-1, HTS to pHTS-2 and pHTS-1 to pHTS-2. For each sentence, subjects were asked to listen to both versions (randomly shuffled) and to attribute a score according to their overall preference. Table 3 shows the preference scores resulting from this test, computed with a 95% confidence interval for male and female speakers.

Table 3. *Preference scores for the male and female speakers, HTS, pHTS-1 and pHTS-2 compared by pairs.*

	HTS/pHTS-1	**HTS/pHTS-2**	**pHTS-1/pHTS-2**
Male	0.28 ± 0.14	0.51 ± 0.19	0.01 ± 0.16
Female	0.48 ± 0.17	0.49 ± 0.18	0.01 ± 0.16

The first thing that can be highlighted from these results is that the reduction of subjective quality introduced by the performative modifications is relatively minor, for both pHTS-1 and pHTS-2. Indeed the scores obtained with this test show that we stay between 0 (about the same) and 1 (slight preference for HTS). While being noticeable, it means that the distortions introduced by pHTS-1 or pHTS-2 are not stressed by the listeners as a real issue regarding the suprasegmental quality. Besides, when we compare directly the two pHTS approaches, users do not make any distinction. Therefore we can argue that reducing the amount of contextual information at the training stage has an unnoticeable impact on the suprasegmental quality, when used in conjunction with the short-term trajectory generation technique.

5. EVALUATION OF REACTIVE MODIFICATION OF THE ARTICULATION DEGREE

In this section, our goal is to evaluate how real-time interpolation between various models can be perceived by the user. Here, we present the results of a subjective test that we conducted to evaluate the user’s perception of abruptness/fluidity in transitioning between hypo and hyperarticulated speech, although this approach can be applied to any interpolation scheme.

It is important to highlight that this test can only be performed with pHTS, and not with the regular version of HTS. Indeed it takes the short-term generation of speech parameter trajectories described in Section 2.2 to actually be able to perform continuous and reactive modification over the provided models. Even though model interpolation is possible with HTS, this is done over all the generated speech parameter trajectories, resulting in a full sentence with a constant (and hence not a transitional) speaking style. Therefore this listening test on the perception of abruptness/fluidity in transitioning between different articulation degrees only involves the pHTS system.

For this test, we address the idea of speaking style varying along the sentence, as if an external condition would have changed it: sudden background noise, listener reactions or gestures, change of emotion, etc. We are interested in how a time-varying continuous control (in this case of the articulation degree) is perceived by the listener, as a way to evaluate what level of refinement is needed in a context-reactive speech synthesizer. However, such an evaluation is difficult to define. Moreover, we think that pHTS is among the first tools that allows us to explore this new category of listening tests, involving more user interaction during the experiment.

5.1. Experimental Protocol

For this subjective evaluation, listeners were asked to listen to sentences synthesized with pHTS-1 combined with one of the three different articulation trajectories:

- step trajectory, where the first half of the sentence is over-hypo (over-hyper) articulated and the second half is over-hyper (over-hypo) articulated;
- linear trajectory, where the degree of articulation changes linearly from over-hypo (over-hyper) articulated to over-hyper (over-hypo) articulated over the whole sentence;

- step-linear trajectory, where the first quarter of the sentence is over-hypo (over-hyper) articulated, the last quarter is over-hyper (over-hypo) articulated and the degree of articulation is linearly transitioning between the two extrema in the middle.

For the cases of linear and step-linear trajectory, the articulation degree is actually evolving linearly with the phoneme sequence (whose duration is dependent upon the speech rate), and not with time. The test consisted of 30 sentences randomly chosen among the held-out set of the database (used neither for training nor for adaptation). Participants were asked to score the fluidity of the transition between the two articulation degrees. In order to do that, they were given a continuous scale ranging from 1 (very abrupt) to 5 (very fluid). These scales were extended one point further on both sides in order to prevent border effects. We can also note that these sentences are relatively short, 2 or 3 seconds each. It means that we rather try to assess listener's ability to discriminate short-term variations in the articulation degree.

5.2. Results

The results of the experiment are illustrated in Figure 3. It is observed that participants could correctly discriminate between various types of changes in the articulation degree, happening within a sentence, as expected from the reference. Particularly the step and linear variations of the articulation degree in pHTS led to a clear listener decision, respectively for considering this change very abrupt or very fluid. Furthermore, as expected, the step-linear transition is observed to be perceived in between the step and the linear transitions on the abruptness/fluidity scale. Varying the parameterization of articulation changes – from step to linear – and showing that such parametrization is discriminated by listeners is a way to validate that the *suprasegmental quality of the articulation degree* is a meaningful parameter to be included in a reactive speech synthesizer. Indeed, we can conjecture that quality of variation in the speaking style is a consistent cue in verbal communication and provide useful information on the speaker's state and context. It can then be concluded from this experiment that the proposed pHTS approach allows both a continuous and reactive interpolation of models in the framework of HMM-based speech synthesis, which is here shown to enable a perceptually-motivated control of the articulation degree by the user.

6. CONCLUSIONS AND FUTURE WORKS

In this paper, we have introduced a significant variant of the HTS engine, called performative HTS or pHTS. This new engine relies on two main modifications: the reduction of the surrounding phonetic context for both the training of the HMMs and the generation of speech parameter trajectories from these HMMs. We have shown by objective and subjective measurements that these modifications have a limited impact on the quality of the speech output.

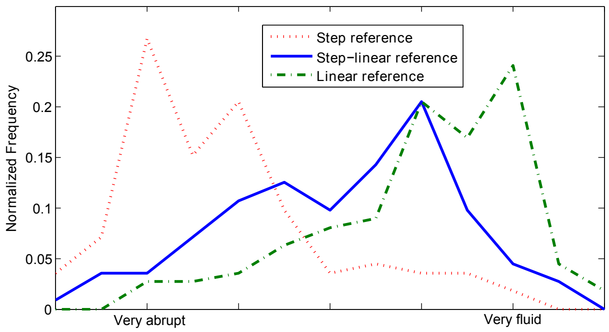

Fig. 3. *Results on the perception of the abruptness/fluidity in transitionning between hypo and hyperarticulated speech.*

We have also presented a new listening test that could only be realized with pHTS. Indeed it involved the continuous and reactive modification of the degree of articulation within the synthesized sentence, which is not possible with the regular HTS engine. We used this new property to assess the listener's ability to distinguish various types of speech sentences, corresponding to various trajectories of the degree of articulation within the sentence. The result of this experiment is that the listeners are able to accurately and continuously track back the quality of the time-varying degree of articulation during synthesis and associate it with a level of abruptness in the change of the speaking style.

These results are very promising in terms of enriching our current way of modeling the suprasegmental quality in speech synthesis. For instance, by using various databases containing different accents, speaking styles or emotions and, by means of interpolation techniques, it is possible to create both unique personalized voices and a system reactively adjustable to its environmental changes. In the same concept of real-time model interpolation, it would be interesting to explore the idea of not only speech synthesis, but also singing synthesis by using opera recordings and more flexible control of duration models for the synthesis of long vowels, like in [22]. We can also envision the improvement of the vocoder quality. Indeed it has been shown in [23] that a more suited excitation modeling allows to significantly increase the naturalness of the produced speech. We therefore plan to implement a real-time version of the vocoder based on the Deterministic plus Stochastic Model (DSM, [23]) of the residual signal and combine it with the pHTS engine.

We started to use these reactive properties for building a new speech synthesis platform, called MAGE [24], where pHTS is integrated into a real-time audio software architecture and accessed via a reactive query protocol. MAGE encapsulates the HTS functionalities and pHTS properties into a user-friendly API, so that more developers can integrate our new speech synthesis features in their applications.

7. ACKNOWLEDGEMENTS

The authors would like to thank the financial and academic supports of the MAGE project: University of Mons and Acapela Group S.A. Also, we would like to thank G. Wilfart and A. Moinet for their contributions.

8. REFERENCES

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis," *IEICE Transactions On Information And Systems*, vol. 83, no. 11, pp. 2347–2350, 1999.

[2] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[3] C. L. Bennett, *Large Scale Evaluation of Corpus-based Synthesizers: Results and Lessons from the Blizzard Challenge 2005*, pp. 105–108, 2005.

[4] [Online], "Hmm-based speech synthesis system (hts)," http://hts.sp.nitech.ac.jp.

[5] N. Campbell, "Differences in the speaking styles of a japanese male according to interlocutor ; showing the effects of affect in conversational speech," *Computational Linguistics Chinese Language Processing*, vol. 12, no. 1, pp. 1–16, 2007.

[6] K. I. Nordstrom, S. Fels, C. D. Hassall, and B. Pritchard, "Developing vowel mappings for an interactive voice synthesis system controlled by hand motions.," *Journal of the Acoustical Society of America*, vol. 127, no. 3, pp. 2021, 2010.

[7] A. W. Black and P. Taylor, *Automatically clustering similar units for unit selection in speech synthesis*, vol. 97vol2pp, pp. 1–4, International Speech Communication Association, 1997.

[8] A. W. Black, *Unit selection and emotional speech*, pp. 1649–1652, ISCA, 2003.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society Series B Methodological*, vol. 39, no. 1, pp. 1–38, 1977.

[10] J. J. Odell, *The Use of Context in Large Vocabulary Speech Recognition*, Ph.D. thesis, 1995.

[11] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, *Voice characteristics conversion for HMM-based speech synthesis system*, vol. 3, pp. 1611–1614, Institute of electrical engineers (IEE), 1997.

[12] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Speaker interpolation in hmm-based speech synthesis system," pp. 2523–2526, 1997.

[13] Zhen-Hua Ling, K. Richmond, J. Yamagishi, and Ren-Hua 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.

[14] B. Picart, T. Drugman, and T. Dutoit, "Continuous control of the degree of articulation in hmm-based speech synthesis," in *Proceedings of Interspeech*, August 2011, pp. 1797–1800.

[15] B. Picart, T. Drugman, and T. Dutoit, "Analysis and synthesis of hypo and hyperarticulated speech," in *SSW 2010*, 2010.

[16] A. W. Black and K. A. Lenzo, "Building synthetic voices," *Voices*, pp. 1–206, 2003.

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

[18] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using hsmm-based speaker adaptation and adaptive training," *IEICE Transactions On Information And Systems*, vol. E90-D, no. 2, pp. 533–543, 2007.

[19] M. J. F. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[20] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of lpc parameters at 24 bits/frame," *IEEE Transactions On Speech And Audio Processing*, vol. 1, no. 1, pp. 3–14, 1993.

[21] I. Peretz and K. L. Hyde, "What is specific to music processing? insights from congenital amusia," *Trends in Cognitive Sciences*, vol. 7, no. 8, pp. 362–367, 2003.

[22] K. Oura, A. Mase, T. Yamada, S. Muto, Y. Nankaku, and K. Tokuda, "Recent development of the hmm-based singing voice synthesis system - sinsy," in *SSW7 in Kyoto, Japan*, 2010, pp. 211–216.

[23] T. Drugman and T. Dutoit, "The deterministic plus stochastic model of the residual signal and its applications," *IEEE Transactions On Audio Speech And Language Processing*, vol. 20, no. 3, pp. 968–981, 2012.

[24] [Online], "Mage speech synthesis platform (source code)," http://numediart.org/demos/mage_phts.