

# PHTS FOR MAX/MSP: A STREAMING ARCHITECTURE FOR STATISTICAL PARAMETRIC SPEECH SYNTHESIS

Thierry Dutoit<sup>1</sup>, Maria Astrinaki<sup>1</sup>, Onur Babacan<sup>1</sup>, Nicolas d’Alessandro<sup>2</sup>, Benjamin Picart<sup>1</sup>.

<sup>1</sup> Laboratoire de Théorie des Circuits et Traitement du Signal (TCTS), Faculté Polytechnique de Mons (FPMs), Belgique

<sup>2</sup> MAGIC - Media and Graphics Interdisciplinary Centre, University of British Columbia, Canada

## ABSTRACT

In this report, we present a Max/MSP external for real-time speech synthesis. Statistical parametric speech synthesis, based on Hidden Markov Models has been demonstrated to be very effective in synthesizing high-quality, natural and expressive speech. This technique is also able to provide high flexibility as a speech production model and a small database footprint. In this work, we modify the existing HTS engine in order to establish a streaming architecture, called performative-HTS or pHTS. pHTS is implemented as a Max/MSP external which provides a basis for further research in gesturally-controlled speech synthesis. Quantitative evaluations of the system show that the degradation of speech quality in pHTS is small with reference to HTS. These results are supported by a subjective evaluation, which confirms that HTS and pHTS resulting speech waveforms can hardly be distinguished.

## KEYWORDS

HMM, speech synthesis, statistical parametric speech synthesis, real-time, performative, streaming, HTS, sHTS, pHTS, Max/MSP

## 1. INTRODUCTION

Performative speech synthesis is an emerging research direction which aims at producing speech signals directly through gestural control rather than through the Text-To-Speech (TTS) paradigm. This approach leads to consider the design of performative speech synthesis systems as a digital instrument making task. In our research, we have been investigating *digital lutherie* with a focus on speech synthesis for several years through projects like MaxMBROLA [4], RAMCESS [5] and HandSketch [3].

The recent outgrowth of statistical parametric synthesis has incited the next steps in our research. In contrast to concurrent concatenative synthesis approaches [10], statistical parametric speech synthesis systems train statistical models with various features using natural speech databases, and generate speech from the trained statistical models. A prominent method in this approach employs hidden Markov Models (HMMs). HTS is an implementation of HMM-based speech synthesis and is publicly available and widely used [21]. HTS features small database size, easily changeable voice characteristics and models a large amount of contextual factors. It produces highly intelligible speech.

To the best of our knowledge, a performative HMM-based speech synthesis system has not been attempted yet. Our aim in this work is thus to restructure the HTS architecture into a real-time back-end system, evaluate the segmental quality of the new system and develop an pHTS Max/MSP external.

In this report we give an overview of HTS as a comprehensive implementation of HMM-based speech synthesis. We continue with the description of our streaming architecture for HTS

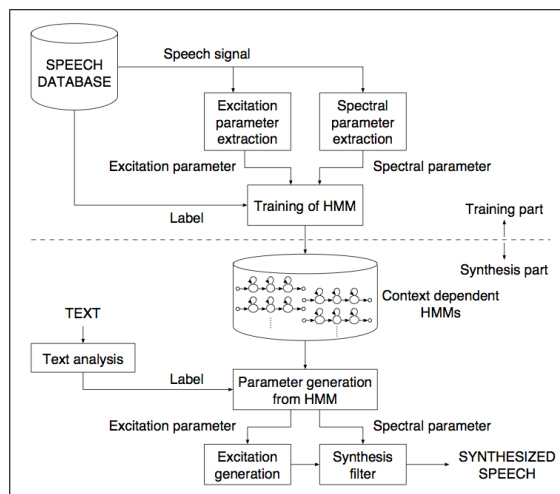


Figure 1: Block diagram of the HMM-based speech synthesis system: the training and synthesis parts [16].

and its implementation as a Max/MSP external. Finally results from objective and subjective evaluations are presented.

## 2. HMM-BASED SPEECH SYNTHESIS WITH HTS

The need for smaller synthesizer footprints and more flexible control of synthesis has brought researchers to envision ways of using the knowledge contained in the database itself rather than manually crafting the available phonetic units. This approach is known as statistical parametric speech synthesis. Instead of using real speech samples at runtime, the pre-recorded database is analyzed, various production parameters are extracted – spectral envelopes, fundamental frequency and duration of the phonemes – and used to train statistical models. Later these models will generate the speech parameters, regarding a given targeted text input. Speech waveform are produced from parameters with typical speech synthesis techniques: subtractive synthesis or harmonic plus noise.

HTS is an existing software system that specifically uses Hidden Markov Models (HMMs) as the statistical models [21]. Figure 1 shows the block diagram of the HTS system. HTS is mainly composed of two distinct parts: the training and the synthesis part.

In the training part of HTS, both spectrum and excitation parameters and their dynamic features [23], are extracted from a natural speech database and then are modeled by a set of context-dependent HMMs. Linguistic and prosodic context is taken into account. In order to model speech temporally, HMMs model the state duration densities by using multivariate Gaussian distributions [18]. So to handle all the contextual factors, such as phone

identity and stress or accent related factors that affect the targeted synthetic speech output, decision-trees based on context clustering techniques [22] are used. Magnitude spectrum, fundamental frequency and duration are modeled independently, therefore there is a different phonetic decision tree for each of these features [16].

In the synthesis part of HTS, the input is the targeted text to be transformed into synthetic speech. This text is parsed using natural language processing, and a phonetic label sequence containing contextual information is generated. The HMMs of each phoneme in the target sentence are concatenated, and spectral envelopes, pitch, and duration trajectories of the synthetic speech are generated from HMMs themselves based on a maximum likelihood criterion [17, 15]. These trajectories are then used to control the cepstrum-based voice production model in order to synthesize the speech waveform that corresponds to the text input [19].

The main advantages of HTS are its flexibility and its small footprint. It is possible to change the voice characteristics, speaking styles, emotions and prosodic features simply by transforming the parameters of the model [13, 14, 20]. HTS has a small number of tuning parameters. It is based on well-defined statistical principles and it uses source-filter representation of speech, providing the flexibility to control and modify the magnitude spectrum, fundamental frequency and duration of speech output separately. Furthermore, a small amount of training data is enough to create statistical parametric speech synthesis systems. HTS has also a memory-efficient, low-delay speech parameter generation algorithm and a computationally-efficient speech synthesis filter.

### 3. RECORDING AND SYNTHESIS OF ACTORS' VOICES

In the framework of HMM-based expressive speech synthesis, a collaboration with Jean-François Peyret (artistic director) and Thierry Coduys (artist), currently working on “Re : Walden” project, was carried out. This project [8] [9] allows J.F. Peyret to think as Henry David Thoreau, American transcendentalist from the beginning of the 19th century, who decided to spend two years in an isolated cabin in the woods (“Walden; or, Life in the Woods”, book, 1854). The book was a deep reflexion about the use or non-use of technology.

Jean-François Peyret’s play will be based on re-creating the cabin on the scene (one of the places was in the Fresnoy national studio, Tourcoing), using all sorts of technological systems, establishing a paradox between the book and the way it will be acted.

On the stage, the actors will interact not only between them but also with themselves, using their own HMM-based synthetic voice. The latter should be altered in real time to produce special effects (stammering, faltering, etc.) using speech processing techniques. This is in line with the current Numediart project. The synthetic voice of two actors (a male and a female) was provided some months ago. This was achieved by training an HMM-based speech synthesizer for each actor and deliver an easy to use version of these latter synthesizers for the show (collaboration with Acapela Group).

### 4. PHTS : PERFORMATIVE HTS

The concept of real-time can usually be understood in various ways. In the field of gestural control of sound synthesis, we use a performative definition of real-time, the time scales available for control are reduced and at the same time a strict causality between gestures and parameters of sound synthesis processes is preserved.

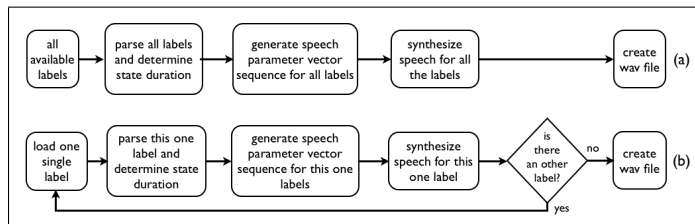


Figure 2: Block diagram of synthesis part of (a) HTS and (b) pHTS.

For the specific case of speech synthesis, coarticulation is known to have forward and backward effects. This implies that a given phoneme cannot be produced correctly if some further information is not known. The main question in this case comes down to: “how much of the future of a linguistic stream must be known to be able to produce natural sounding speech?”. Statistical parametric synthesizers offer a wonderful testbed for exploring this question.

In order to make this possible, we needed to make modifications to the architecture of HTS. As described in the previous section, HTS synthesizes speech on the basis of complete sentences. Thus, in the offline architecture of HTS, all the contextual information from the full input is used. In other words, the minimal available time scale of this architecture is one sentence.

#### 4.1. Towards a performative architecture of HTS

In this work we achieve a first step in giving access to smaller time scales. The performative version of HTS that we propose, called performative-HTS or pHTS, works on a phoneme-by-phoneme basis, i.e. it produces the samples for a given phoneme each time a new phoneme is sent to it as input. Additionally, its delay, i.e. the number of future phonemes required for synthesizing the current phoneme, can be constrained to an arbitrary number of phonemes. We use an already trained HTS system, in which we change the way the HMM that models the sentence is constructed from the phonemes, and consequently the way the sequences of spectrum and excitation parameters are generated. We still use pre-computed context-dependent phonetic labels that contain all the contextual information of a full text sentence, but this information is used in a different way. For each given label, the context-dependent HMM is driven and for this single label the sequence of spectrum and excitation parameters are generated. The speech waveform synthesized from these parameters contains the synthetic speech that corresponds only to the given input phonetic label. In this way, the pHTS parameter generation is achieved by computing many locally-optimal feature paths, which, when combined make a sub-optimal total path, instead of computing one total optimal path for the whole input sentence. In the current implementation, even though the complete sentence context information is available, the system only has access to information corresponding to a single label at each iteration. In Figure 2 we present a block diagram of how labels are processed in the synthesis part of HTS and pHTS.

#### 4.2. pHTS Look-ahead

In a first experiment, we synthesized speech by processing the input phonetic labels one-by-one, using only the current label at each iteration, no information related to past or future labels, and obtained the complete sentence waveform by simply concatenating the phoneme-level waveforms. As expected, this approach synthesizes phonemes as if they would be produced in isolation, sur-

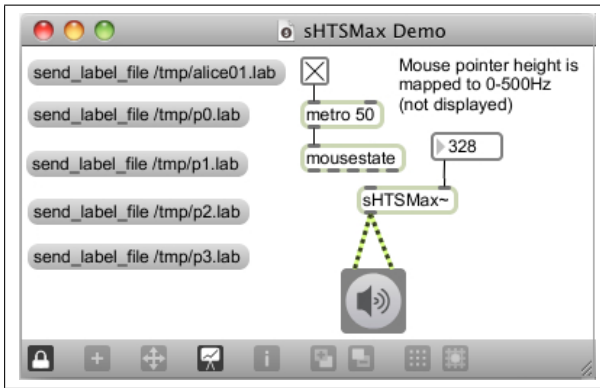


Figure 3: pHTSMax- object in a demo Max/MSP patch using mouse position as pitch input for the object.

rounded by silence, and imposing artificial phoneme on- and off-set. In order to overcome this problem we introduced a small look-ahead buffer, whose length can be set at run-time. This greatly improved the output speech quality, even in the single phonetic label look-ahead case: the resulting synthesized speech sounds almost as natural and as intelligible as the synthesized speech from the offline version of HTS. Details of the tests we made to assess this first impression are in Section 5.

#### 4.3. pHTS as a Max/MSP external

In order to create a real-time implementation, we chose Max/MSP [2] as our platform. Using the pHTS code, we created an MSP external object, named pHTSMax-. At the current state of implementation, the pHTSMax- object has two inputs, one for the path to a label file that contain the context-dependent labels and one for the pitch value in Hz. The phonetic labels, as described above, are pre-computed and contain all the contextual information of a full text sentence. The labels are parsed one-by-one inside the object, using a look-ahead buffer of one phoneme. The speech parameters that correspond to the current phonetic label are produced while the pitch information from the model is ignored for voiced speech segments, and is replaced by the user’s pitch input in real-time. Then the corresponding speech is synthesized and directed to the output. In the example patch (Figure 3), we load labels with Max messages and use the mouse pointer position to control the pitch. A demonstration of this patch can be found online<sup>1</sup>. Of course any other control could be used instead. In the future we will expand this real-time interface of the object, by making it possible to control duration, intensity and voice quality, by using the DSM excitation model for instance [6].

### 5. OBJECTIVE AND SUBJECTIVE MEASUREMENTS

For our tests we used the speaker-dependent training demo in English that is provided in [7]. The contextual factors that we took into account are the same as [16]. At run time, we used full label files, where each line is an alpha-numerical character sequence encoding all the information listed above for one phoneme, which is then input into the system. In HTS, all of the labels were used at once. In pHTS, one phonetic label was input into the system at a time. We synthesized speech produced by both HTS and pHTS,

<sup>1</sup><http://www.youtube.com/watch?v=WaAXiCU49Xw>

using all the pre-computed label files. With this technique, we observed that the state durations were different between the two approaches. This is because durations are modeled more smoothly when all the information from the phonetic labels is provided to the system. Consequently we forced both systems to use the same state durations, those provided by the HTS system, so as to be able to compare our results frame by frame. Using this modification, we evaluated the quality of speech synthesized by pHTS by comparison to the original offline HTS synthetic speech, by using both objective and subjective measurements. We evaluated pHTS results using look-ahead buffers of one, two and three phonetic labels, which we refer to as pHTS1, pHTS2, pHTS3, respectively.

#### 5.1. Objective Evaluation

In order to evaluate the distortion introduced by pHTS, we created a test database consisting of 40 sentences synthesized by HTS, pHTS1, pHTS2 and pHTS3, for both male and female voices. We chose two metrics previously used in related research to evaluate test data. The first metric we used is *mel-cepstral distortion* [12], a distance measure calculated between mel-cepstral coefficients. We applied this metric to the mel-cepstral coefficients generated by three test groups (pHTS1, pHTS2, pHTS3) and the results are presented in Table 1. The second metric we used is *spectral distortion* (SD) [11]. We applied the discrete-time version of SD to our three test groups (pHTS1, pHTS2, pHTS3) and the results are presented in Table 2. Both methods show that as the buffer size is decreased, the degradation increases as expected, albeit not significantly. This shows that using only one look-ahead phonetic label results in segmental quality that is very close to (if not hardly distinguishable from) situations with more known phonetic future.

Table 1: Mean mel-cepstral distortion (Mel-CD) and 95% confidence intervals between HTS and sHTS with one, two and three future label buffers, in dB.

Mel-CD (dB)	Male Voice	Female Voice
pHTS1	2.76 ± 0.31	2.49 ± 0.33
pHTS2	2.68 ± 0.32	2.40 ± 0.34
pHTS3	2.63 ± 0.32	2.38 ± 0.34

Table 2: Mean spectral distortion (SD) and 95% confidence intervals between HTS and pHTS with one, two and three future label buffers, in dB.

SD (dB)	Male Voice	Female Voice
pHTS1	0.88 ± 0.12	1.21 ± 0.19
pHTS2	0.75 ± 0.10	1.19 ± 0.21
pHTS3	0.73 ± 0.10	1.20 ± 0.22

#### 5.2. Subjective Testing

For the subjective evaluation of the three pHTS approaches compared to the HTS we used the ABX method [1]. Our test database contains 66 different speech samples of a female voice with durations of 2 to 4 seconds for each approach. By using this database we created an ABX test with 30 questions, 10 questions for each one of the three test groups (pHTS1, pHTS2, pHTS3) speech output compared HTS output. For each user a uniquely randomized

test was generated, and for each question of the test, A and B options were randomly selected between the HTS sample and the version of pHTS that was being tested. In total, 59 different tests were conducted, by both speech and non-speech experts. Note that in the ABX method, 50% error rate means perfect confusability, i.e. that the two methods being tested are indistinguishable from each other. The results we obtained show relatively high confusability, enough to confirm that pHTS can be used in place of HTS.

Table 3: Error rate between HTS and pHTS1, pHTS2, pHTS3

pHTS1 vs. HTS	pHTS2 vs. HTS	pHTS3 vs. HTS
37.83%	37.66%	38.83%

## 6. CONCLUSIONS

Converting HTS from the original offline architecture to a streaming architecture takes us one step closer to a full real-time performative system. The results we have obtained so far are very encouraging, as they confirm that HTS can be used with only one look-ahead phoneme, with no significant distortion compared to full sentence look-ahead, given that adequate higher level linguistic information is still provided to the synthesizer, i.e. stress. However, this architecture of pHTS cannot be faster than one phoneme delay, otherwise output quality is degraded significantly. Architectural re-design of pHTS will be needed in order to have intraphonetic control. We want to introduce less contextual information, and to train our system with less possible linguistic information, but in all cases we want to preserve high intelligibility and naturalness. Going further will require interfacing pHTS with controllers for inputting phonetic labels on the fly, and for controlling pitch, duration, and voice quality in real-time. The development of an appropriate interface that will combine continuous gestural inputs and voice synthesis is essential. Finally a major issue we will have to face is that of the ability of a performer to control a large amount of control dimensions.

## 7. ACKNOWLEDGMENTS

M. Astrinaki and O. Babacan's work is supported by a PhD grant funded by UMONS and ACAPELA-GROUP SA.

numediart is a long-term research program centered on Digital Media Arts, funded by Région Wallonne, Belgium (grant N°716631).

## 8. REFERENCES

### 8.1. Scientific references

- [1] D. Clark. "High-resolution subjective testing using a double-blind comparator". In: *J. Audio Eng. Soc.* 30 (1982). Pp. 330–338. P.: 9.
- [2] N. D'Alessandro and T. Dutoit. "HandSketch Bi-Manual Controller: Investigation on Expressive Control Issues of an Augmented Tablet". In: *Proceedings of the 7th International Conference on New Instruments for Musical Expression (NIME'07)*. 2007. Pp. 78–81. P.: 7.
- [3] N. D'Alessandro et al. "MaxMBROLA: A Max/MSP MBROLA-Based Tool for Real-Time Voice Synthesis". In: *Proceedings of the EUSIPCO'05 Conference*. Antalya, Turkey 2005. P.: 7.
- [4] N. D'Alessandro et al. "RAMCESS 2.X framework - expressive voice analysis for realtime and accurate synthesis of singing". In: *Journal on Multimodal User Interfaces (JMUI), Springer Berlin/Heidelberg 2.2* (2008). Pp. 133–144. P.: 7.
- [5] T. Drugman, G. Wilfart, and T. Dutoit. "A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis". In: *Proc Interspeech* (2009). P.: 9.
- [6] A. Hunt and A. Black. "Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database". In: *Proc. IEEE International Conference of Acoustics, Speech, and Signal Processing*. 1996. Pp. 373–376. P.: 7.
- [7] F. Norden and T. Eriksson. "A Speech Spectrum Distortion Measure with Interframe Memory". In: *Proc. IEEE International Conference on Audio, Speech and Signal Processing*. 2001. P.: 9.
- [8] B. Picart, T. Drugman, and T. Dutoit. "Analysis and Synthesis of Hypo and Hyperarticulated Speech". In: *Proceedings of the Speech Synthesis Workshop 7 (SSW7)*. NICT/ATR, Kyoto, Japan 2010. Pp. 270–275. P.: 9.
- [9] K. Shichiri et al. "Eigenvoices for HMM-based speech synthesis". In: *Proceedings of International Conference on Spoken Language Processing*. 2002. Pp. 1269–1272. P.: 8.
- [10] M. Tamura et al. "Adaptation of pitch and spectrum for HMM-based speech synthesis using mlr". In: *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*. 2001, Vol. 2. Pp. 805–808. P.: 8.
- [11] T. Toda, A. Black, and K. Tokuda. "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory". In: *IEEE Trans. Audio Speech Lang. Process.* 15.8 (2007). Pp. 2222–2235. P.: 8.
- [12] K. Tokuda, H. Zen, and A.W. Black. "An HMM-based speech synthesis system applied to English". In: *IEEE Speech Synthesis Workshop*. 2002. Pp.: 7–9.
- [13] K. Tokuda et al. "Speech parameter generation algorithms for HMM-based speech synthesis". In: *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*. 2000, Vol. 3. Pp. 1315–1318. P.: 8.
- [14] T. Yoshimura et al. "Duration Modeling in HMM-based Speech Synthesis System". In: *Proc. of ICSLP*. 1998, Vol. 2. Pp. 29–32. P.: 7.
- [15] T. Yoshimura et al. "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis". In: *Proc. Eurospeech*. 1999. Pp. 2347–2350. P.: 8.
- [16] T. Yoshimura et al. "Speaker interpolation in HMM-based speech synthesis system". In: *Proceedings of European Conference on Speech Communication and Technology'97*. 1997, Vol. 5. Pp. 2523–2526. P.: 8.
- [17] H. Zen, K. Tokuda, and A. Black. "Statistical Parametric Speech Synthesis". In: *Speech Communication* 51.11 (2009). Pp. 1039–1064. P.: 7.
- [18] H. Zen, K. Tokuda, and T. Kitamura. "Decision tree based simultaneous clustering of phonetic contexts, dimensions, and state positions for acoustic modeling". In: *Proc. Eurospeech*. 2003b. Pp. 3189–3192. P.: 8.

- [23] H. Zen, K. Tokuda, and T. Kitamura. “Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences”. In: *Comput. Speech Lang* 21.1 (2006c). Pp. 153–173. P.: 7.

### **8.2. Artistic references**

- [8] <http://theatre-chailot.fr/theatre/re-walden>. P.: 8.
- [9] <http://www.liberation.fr/culture/0101642943-une-cabane-a-outils-multiples>. P.: 8.

### **8.3. Software and technologies**

- [2] Cycling 74. “Max/MSP”. Available at: <http://www.cycling74.com>. P.: 9.
- [7] HMM-based Speech Synthesis System (HTS). “Available at: <http://hts.sp.nitech.ac.jp>”, accessed March, 2011. P.: 9.