



# Thomas Drugman

*PhD student - FNRS research Fellow*

## Affiliation

TCTS Lab - FPMs

Mons **Belgium**

## Application domain

Human-computer interaction

## Contact

Boulevard Dolez 31

7000 Mons

003265/37.47.49

thomas.drugman@fpms.ac.be

---

## Presentation : Hidden Markov Models-based speech synthesis

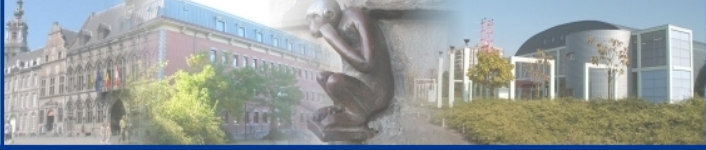
Before the last five years, synthetic speech was typically produced by concatenating frames of natural speech selected from a huge database, possibly applying signal processing to them so as to smooth discontinuities. In 2002, Tokuda et al. (K. Tokuda, 2002) proposed a system relying on the Hidden Markov Models (HMM) generation of speech parameters. Compared to the previous one, this approach has the advantage to allow voice transformation without requiring a large amount of data, merely by adapting its statistics through a short training (A. W. Black & Tokuda, 2007). By voice transformation we here mean voice conversion towards a given target speaker or expressive/emotive speech production from the initial trained system.

The key idea of a HMM-based synthesizer is to generate sequences of speech parameters directly from the trained HMMs. The framework is presented here below and consists of two main steps : the training and the synthesis.

---

## Keywords

Speech Synthesis - Hidden Markov Models - Statistical Parametric Speech Synthesis



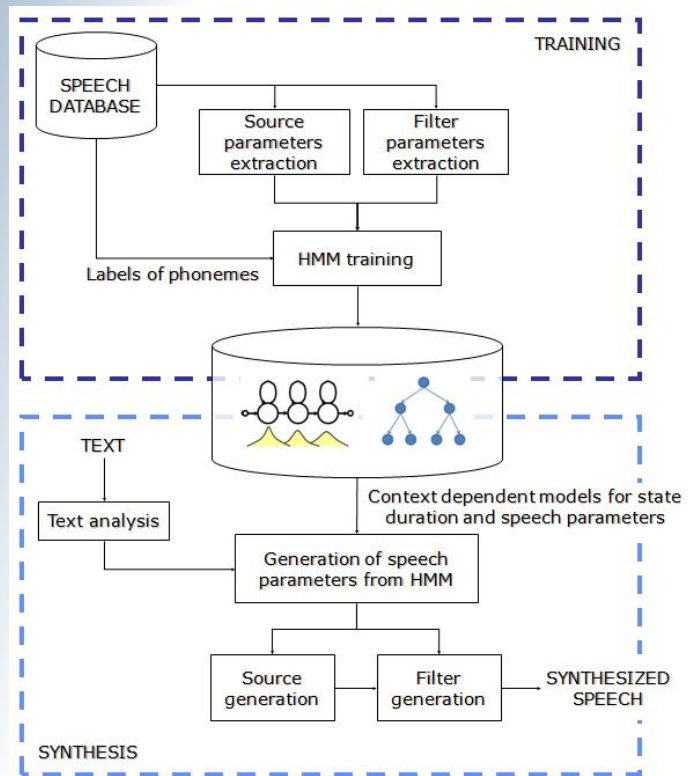
### ***The training part:***

Training our system assumes that a large segmented speech database is available. Labels consist of phonetic environment description. First, speech waveforms are decomposed into their source (glottal) and filter (vocal tract) components. Representative features are then extracted from both contributions. Since source modeling is composed either of continuous values or a discrete symbol (respectively during voiced and unvoiced regions), multi-space probability density HMMs have been proposed. Indeed this approach turns out to be able to model sequences of observations having a variable dimensionality.

Given these latter parameters and the labels, HMMs are trained using the Viterbi and Baum-Welch re-estimation algorithms. Till that point this may seem very close to building a speech recognizer. Nevertheless decision tree-based context clustering is here used to statistically model data appearing in similar contextual situations. Indeed contextual factors such as stress-related, locational, syntactical or phone identity factors affect prosody (duration and source excitation characteristics) as well as spectrum. More precisely an exhaustive list of possible contextual questions is first drawn up. Decision trees are then built for source, spectrum and duration independently (as factors have a different impact on them) using a maximum likelihood criterion. Probability densities for each tree leaf are finally approximated by a Gaussian mixture model.

### ***The synthesis part:***

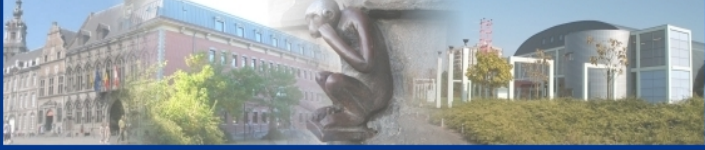
The text typed by the user is first converted into a sequence of contextual labels. From them, a path through context-dependent HMMs is computed using the duration decision tree. Source and spectrum parameters are then generated by maximizing the output probability. The incorporation of dynamic features makes the coefficients evolution more realistic and smooth. Speech is finally synthesized from the generated parameters by an operation of signal processing.



Bloc diagram of a HMM-based speech synthesizer

**Bibliography:**

- A.W. Black, H. Zen & K. Tokuda (2007), Statistical parametric speech synthesis. Proc. of ICASSP (pp.1229-1232).
- K. Tokuda, H. Zen & A.W. Black (2002), An HMM-based speech synthesis system applied to English. Proc. of IEEE Speech Synthesis Workshop.



---

## Academic career

### **From September 2007 : FNRS Research Fellow, PhD Student**

*TCTS Lab, FPMS, Mons*

My PhD thesis aims at incorporating the Hidden Markov Models in voice conversion for speech synthesis

### **January-June 2007 : Master Thesis**

*Swiss Federal Institute of Technologies (EPFL), Lausanne, Switzerland*

My Master thesis dealt with feature selection and multimodal integration for Audio-Visual Speech Recognition

---

## Professional experience

### **July 2006 : Internship**

*Multitel, Mons*

Developing a program in C++ driving the mouse cursor thanks to speech recognition

---

## Research

### *Topics addressed during the thesis:*

Artificial Intelligence - Human-computer interaction - Software Engineering