

---

# On the use of Machine Learning in Statistical Parametric Speech Synthesis

---

Thomas Drugman  
Alexis Moinet  
Thierry Dutoit

FIRSTNAME.NAME@FPMS.AC.BE

Faculté Polytechnique de Mons, TCTS Lab, 31 Boulevard Dolez, 7000 Mons, Belgium

## Abstract

Statistical parametric speech synthesis has recently shown its ability to produce natural sounding speech while keeping a certain flexibility for voice transformation without requiring a huge amount of data. This abstract presents how machine learning techniques such as Hidden Markov Models in generation mode or context oriented clustering with decision trees are applied in speech synthesis. Fields that are investigated in our laboratory to improve this method are also discussed.

## 1. HMM-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 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. Next subsections describe the two main steps in the bloc diagram of such a synthesizer (see Figure 1).

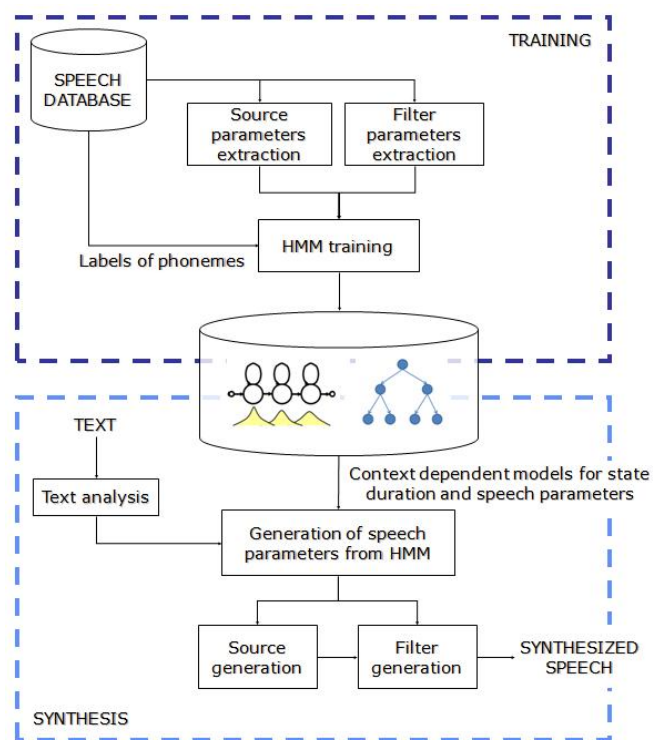


Figure 1. Bloc diagram of a HMM-based speech synthesizer

### 1.1. 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.

## 1.2. 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.

## 2. Our ongoing research activities

Our main goal is to develop an efficient HMM-based speech synthesizer for French. For this, the ACAPELA group kindly provided us with their natural language processor. Since English and French have both their own phonological particularities, an adaptation of the questions used for the context oriented clustering was necessary.

Basically our (ongoing) research activities focus on three main issues:

- **Speech analysis:** A major disadvantage of such a synthesizer is the "buzziness" of the produced speech. This is typically due to the parametrical representation of speech. To overcome this hindrance a particular interest is devoted to speech analysis. Our approach particularly investigates a method of source-filter deconvolution based on the zeros of the Z-transform (B. Bozkurt & Dutoit, 2007). By this way an estimation of the glot-

tal signal and the vocal tract impulse response is achieved. Different models for the source (LF and CALM models) as well as for the spectrum (MLSA, LSP or MFCC coefficients) are tested and their perceptual quality is assessed.

- **Intelligibility enhancement:** In some applications speech has to be synthesized in adverse conditions (in cars, at the station,...). Intelligibility consequently becomes of a paramount importance (Langner & Black, 2005). If we can model the modifications occurring when speech is produced in noise (possibly implying a training), a synthesizer with (adaptive) intelligibility enhancement could be carried out.
- **Voice conversion :** In voice conversion (Y. Stylianou & Moulines, 1995) it is aimed at modifying the source speaker's voice towards a particular target speaker given a limited dataset of his utterances. This approach implies the study of the statistical learning transforming representation spaces of both speakers. This could allow us to easily generate new voices, including the production of more emotions and expressivity in speech.

## Acknowledgments

Thomas Drugman is supported by the "Fonds National de la Recherche Scientifique". The authors also would like to thank the ACAPELA group and the Walloon Region (grant #415911) for their support.

## References

- A. W. Black, H. Z., & Tokuda, K. (2007). Statistical parametric speech synthesis. *Proc. of ICASSP* (pp. 1229–1232).
- B. Bozkurt, L. C., & Dutoit, T. (2007). Chirp group delay analysis of speech signals. *Speech Comm.*, 49, issue 3, 159–176.
- K. Tokuda, H. Zen, A. W. B. (2002). An hmm-based speech synthesis system applied to english. *Proc. of IEEE SSW*.
- Langner, B., & Black, A. W. (2005). Improving the understandability of speech synthesis by modeling speech in noise. *Proc. ICASSP*.
- Y. Stylianou, O. C., & Moulines, E. (1995). Statistical methods for voice quality transformation. *Proc. EUROSPEECH*.