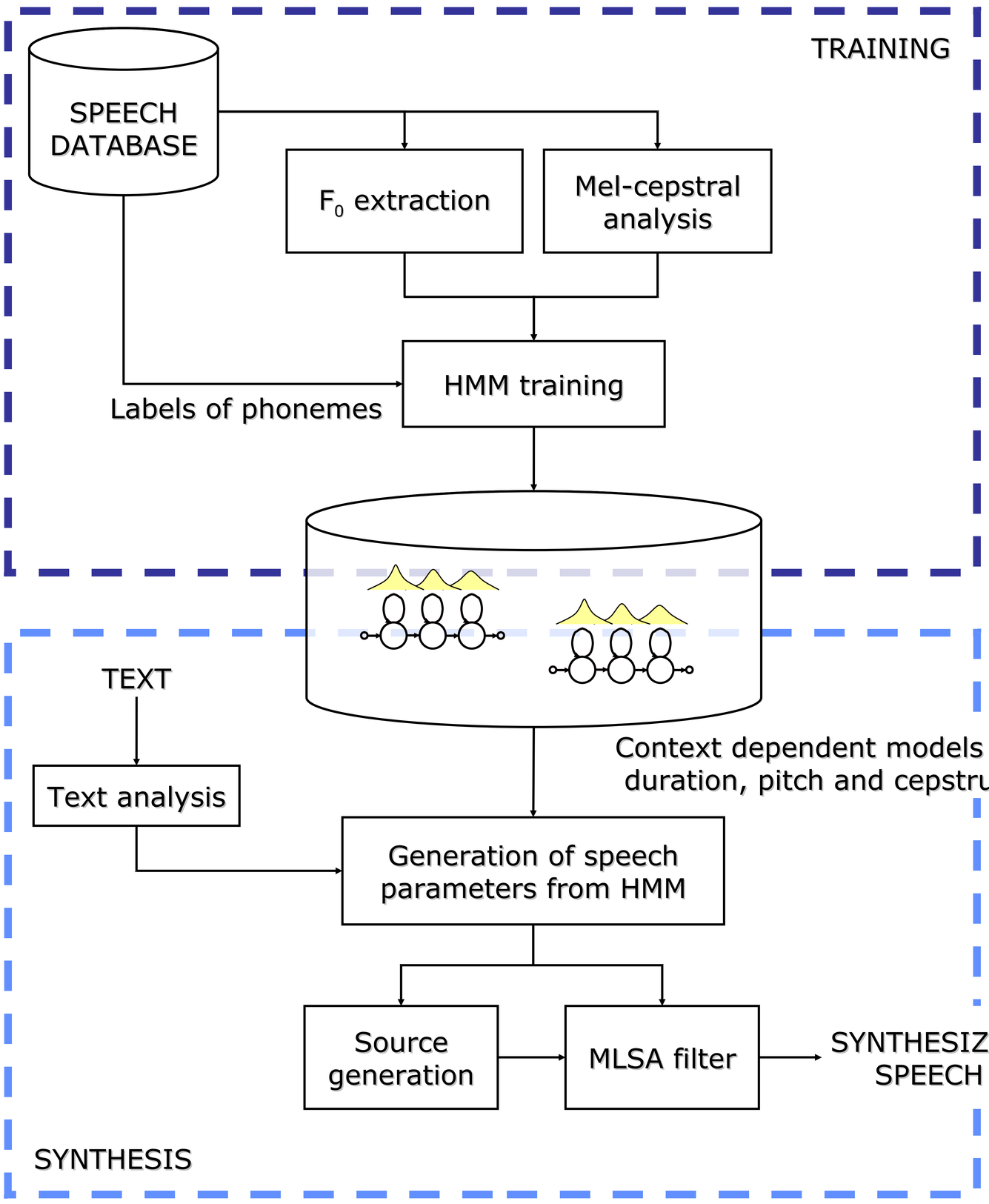
HTS – a HMM-based speech synthesis system

POLYTECH.MONS

A. Auria[1], N. D'Alessandro[2], T. Dutoit[2], A. Moinet[2]

1. Universitat Politècnica de Catalunya, Barcelona, Spain; 2. Faculté Polytechnique de Mons, TCTS Lab, Mons, Belgium;



Generation of speech parameters :

$o_t = [c_t , \Delta c_t , \Delta^2 c_t]$

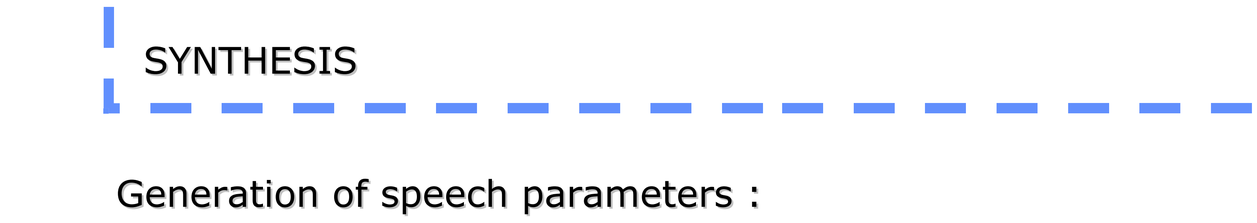
$$\Delta c_t = \sum_{\tau=-L_-^{(1)}}^{L_+^{(1)}} w^{(1)}(\tau) c_{t+\tau} \qquad \Delta^2 c_t = \sum_{\tau=-L_-^{(2)}}^{L_+^{(2)}} w^{(2)}(\tau) c_{t+\tau}$$

→ O = WC → max P(O|Q,λ,T) = max P(WC|Q, λ,T)

$$\frac{\partial}{\partial C} \log P(WC|Q,\lambda,T) = 0$$

→ $W^T\Sigma^{-1}W\,\boldsymbol{C} = W^T\Sigma^{-1}\mu$ (linear equations system)

MLSA Filter :



$$D(z) = \exp \sum_{m=0}^{M} c(m)\tilde{z}^{-m}, \quad \tilde{z}^{-1} = \frac{z^{-1} - a}{1 - az^{-1}}$$

(the exponential function is approximated linearly)

HTS is a speech synthesizer developed and maintained at the Nagoya Institute of Technology (Nitech), in Japan.

Training

The system uses the Hidden Markov Model Toolkit (HTK) from Cambridge Univ. to train a set of HMMs, modeling duration, pitch and cepstral envelopes of the speech database.

Each HMM is a left-to-right sequence of 5 states. It represents one phoneme within its phonemic context (i.e. next and previous phonemes, place of the phoneme in the current syllable, place of the syllable in the current word,…).

For each context, 3 complementary HMMs are created, one for the phoneme duration, one for the evolution of the pitch and one for the evolution of the cepstral envelope.

Then the models are classified using 3 CART (regression trees). Again, one for duration, one for F_0 and one for the cepstrum.

Synthesis

The text is analyzed and transcribed into a sequence of phonemes. Then, using the CARTs, HTS selects three HMMs models* for every phoneme. When this has been done for each phonemes, we have a sequence of states and we can generate a sequence of observations* which most likely fits that state sequence.

Finally, speech can be synthesized using those observations.

A small but powerful system

This system synthesizes speech with no discontinuities at all and very natural prosody while state-of-the-art unit-selection systems often present discontinuities and unnatural prosody.

Moreover, the whole system has a very small footprint (~ 2Mb) while unit-select. synthesizers use DVDs to store their speech databases.

Drawback (our challenge)

The quality of the output sounds quite metallic (vocoded), this is mainly due to the synthesis model (a pulse train through a filter).

We, at TCTS Lab, have a good experience in high quality speech synthesis using more realistic voice production model. We are currently combining such a model with HTS.

This should allow us to improve the acoustic quality of the output as well as easily modify high order voice quality parameters such as spectral tilt, open quotient and asymmetry coefficient.

*duration, pitch and cepstrum

http://hts.sp.nitech.ac.jp/

http://htk.eng.cam.ac.uk/

http://www.cstr.ed.ac.uk/projects/festival/