Hidden Markov Model-based Speech Synthesis

Gustav Sto. Tomas, 2016

Keiichi Tokuda, 1995

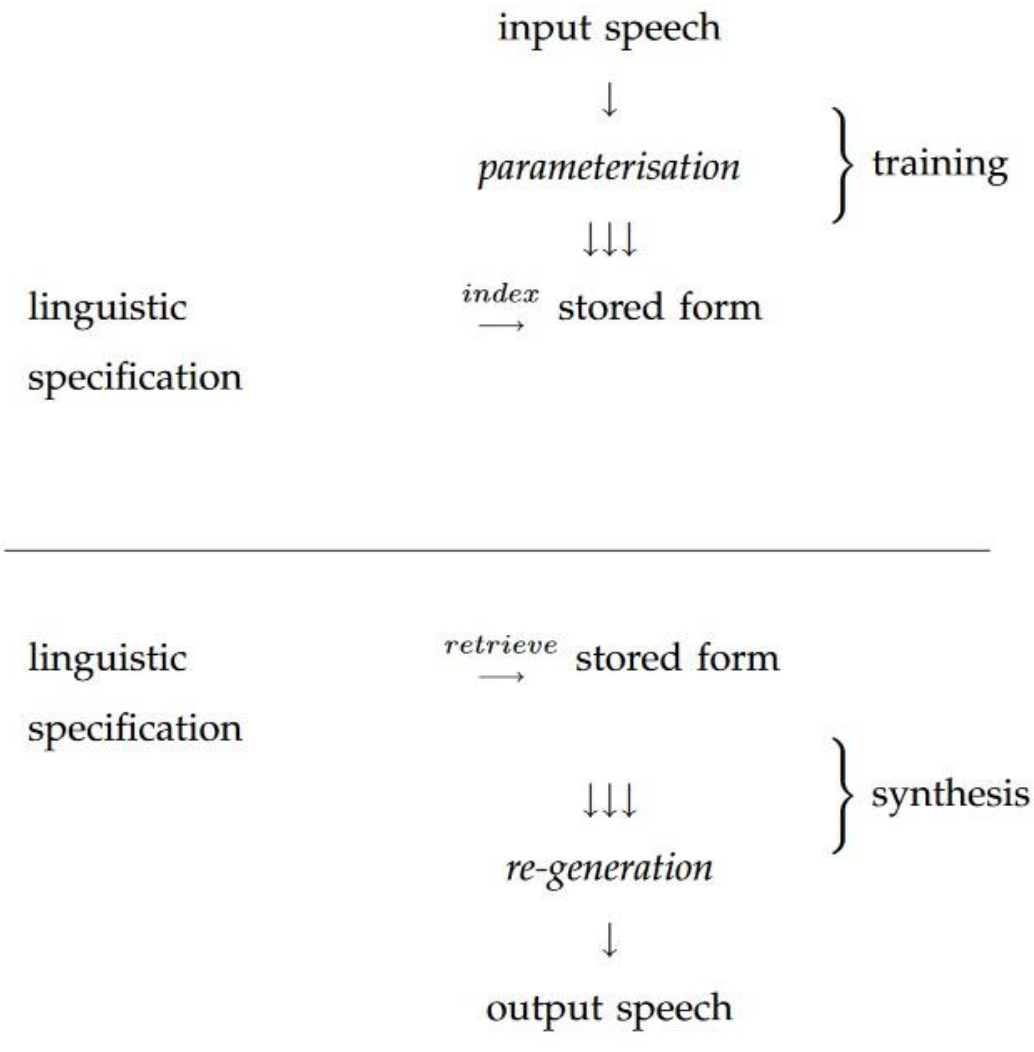
“HMM-based Speech Synthesis System”, HTS

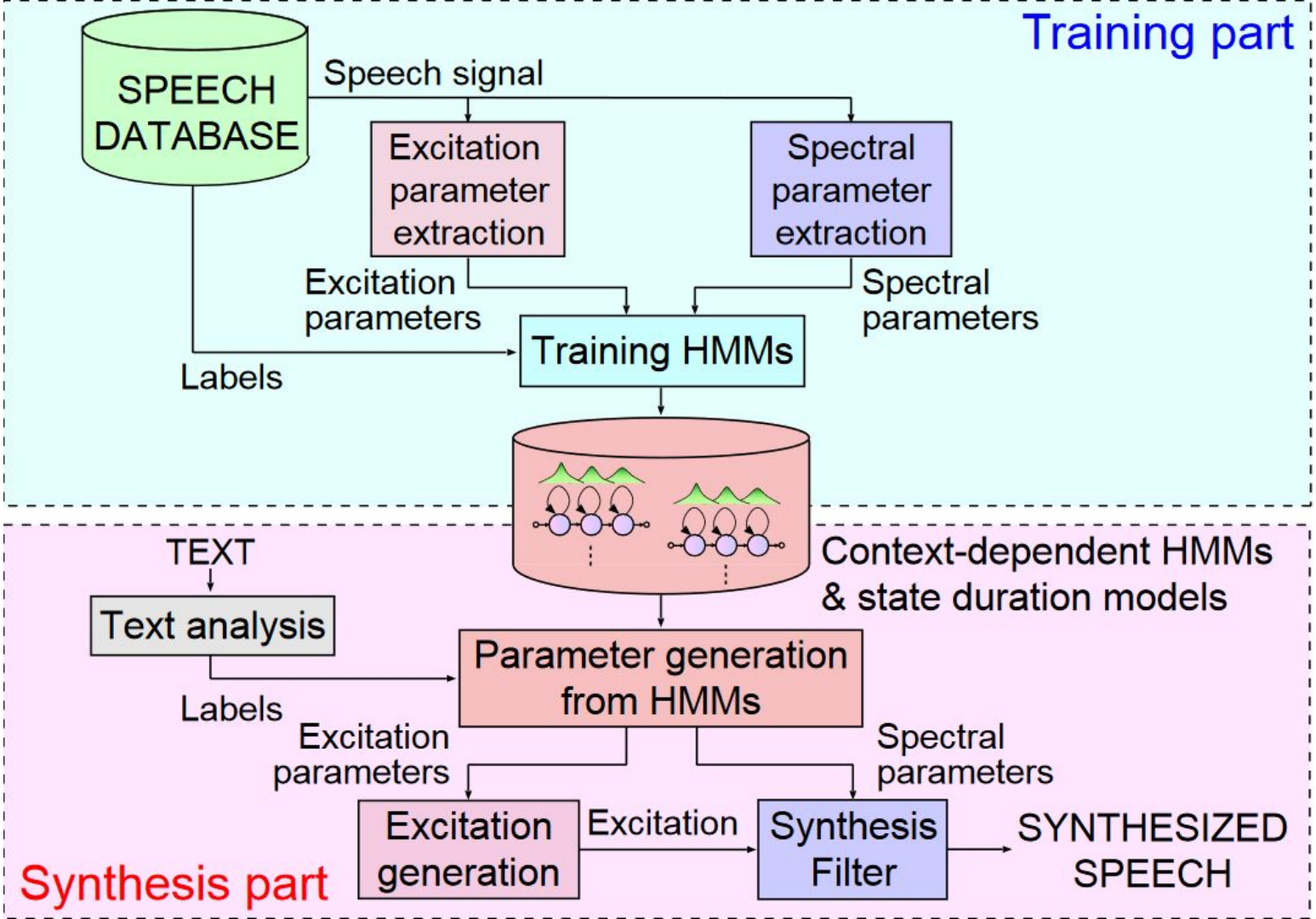
http://hts.sp.nitech.ac.jp/

Statistical Parametric Speech Synthesis

Two components:

- Training Data / Speech Corpus
- Synthesis


input speech
parameterisation
training
linguistic specification
index
stored form
linguistic specification
retrieve
stored form
synthesis
re-generation
output speech


Training part
SPEECH DATABASE
Speech signal
Excitation parameter extraction
Spectral parameter extraction
Excitation parameters
Spectral parameters
Training HMMs
Labels
Context-dependent HMMs & state duration models
TEXT
Text analysis
Parameter generation from HMMs
Labels
Excitation parameters
Spectral parameters
Excitation generation
Excitation
Synthesis Filter
SYNTHESIZED SPEECH
Synthesis part

Linguistic Specification in Corpus: Context

- Preceding and following phonemes
- Position of segment in syllable
- Position of syllable in word & phrase
- Position of word in phrase
- Stress/accent/length features of current/preceding/following syllables
- Distance from stressed/accented syllable
- POS of current/preceding/following word
- Length of current/preceding/following phrase
- End tone of phrase
- Length of utterance measured in syllables/words/phrases
- Corpus may consist of either speech or a statistical model

- Every ~5 ms of speech signal in training data is measured
- All parameters describing fundamental frequency, spectral envelope, etc are stored as vectors
- All this data must be labelled!

What is a Markov Chain?

- A Markov chain is a stochastic process where the next step of the sequence is dependent only on the current state of said sequence

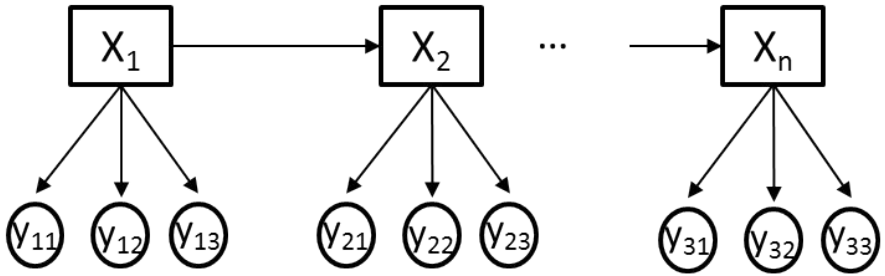
- Markov Assumption: $P(q_i|q_1...q_{i-1}) = P(q_i|q_{i-1})$

Hidden Markov Model

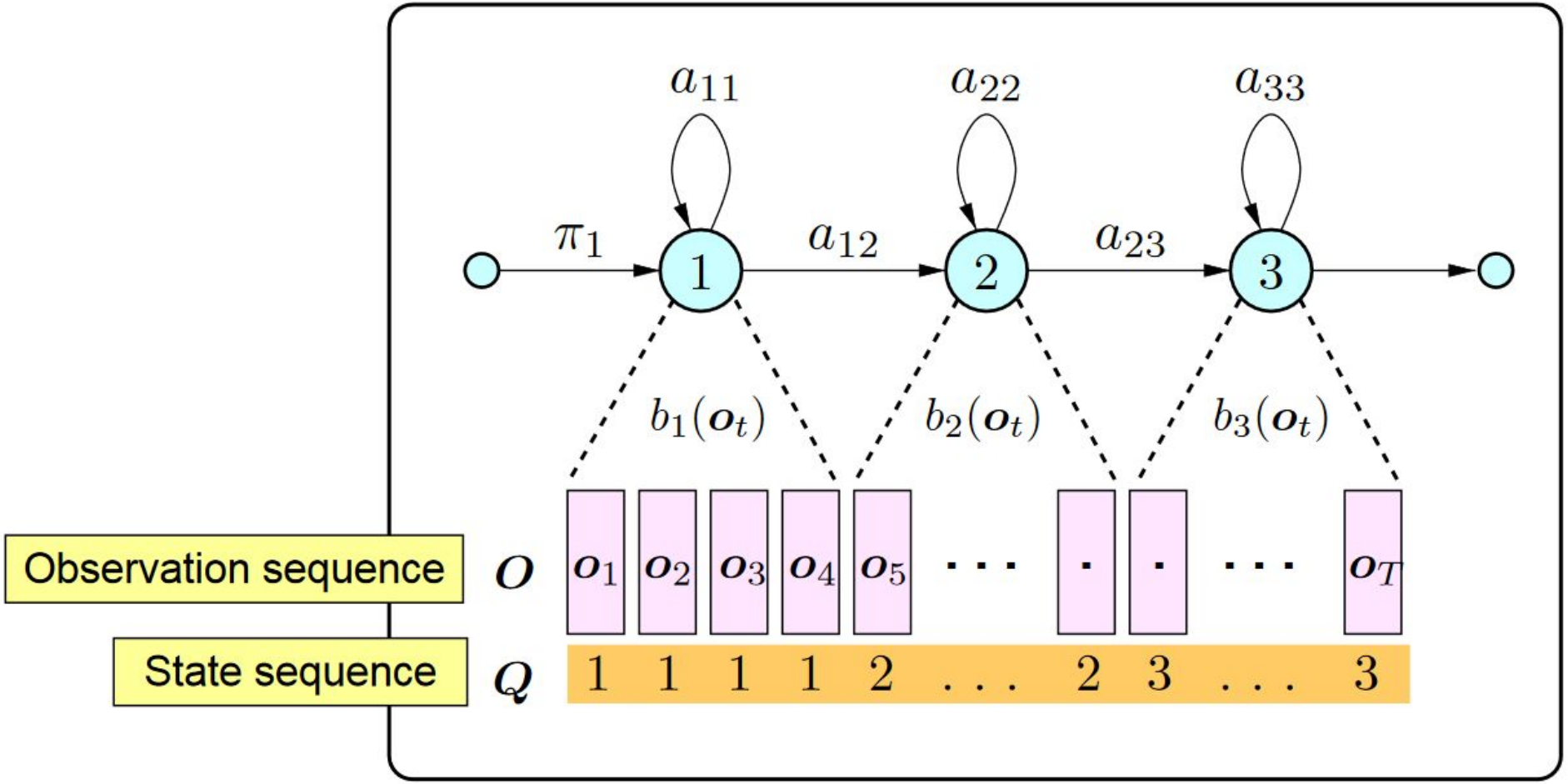
- A HHM is a Markov model where the state is not visible, and the next step in the sequence instead depends on visible output, or emissions

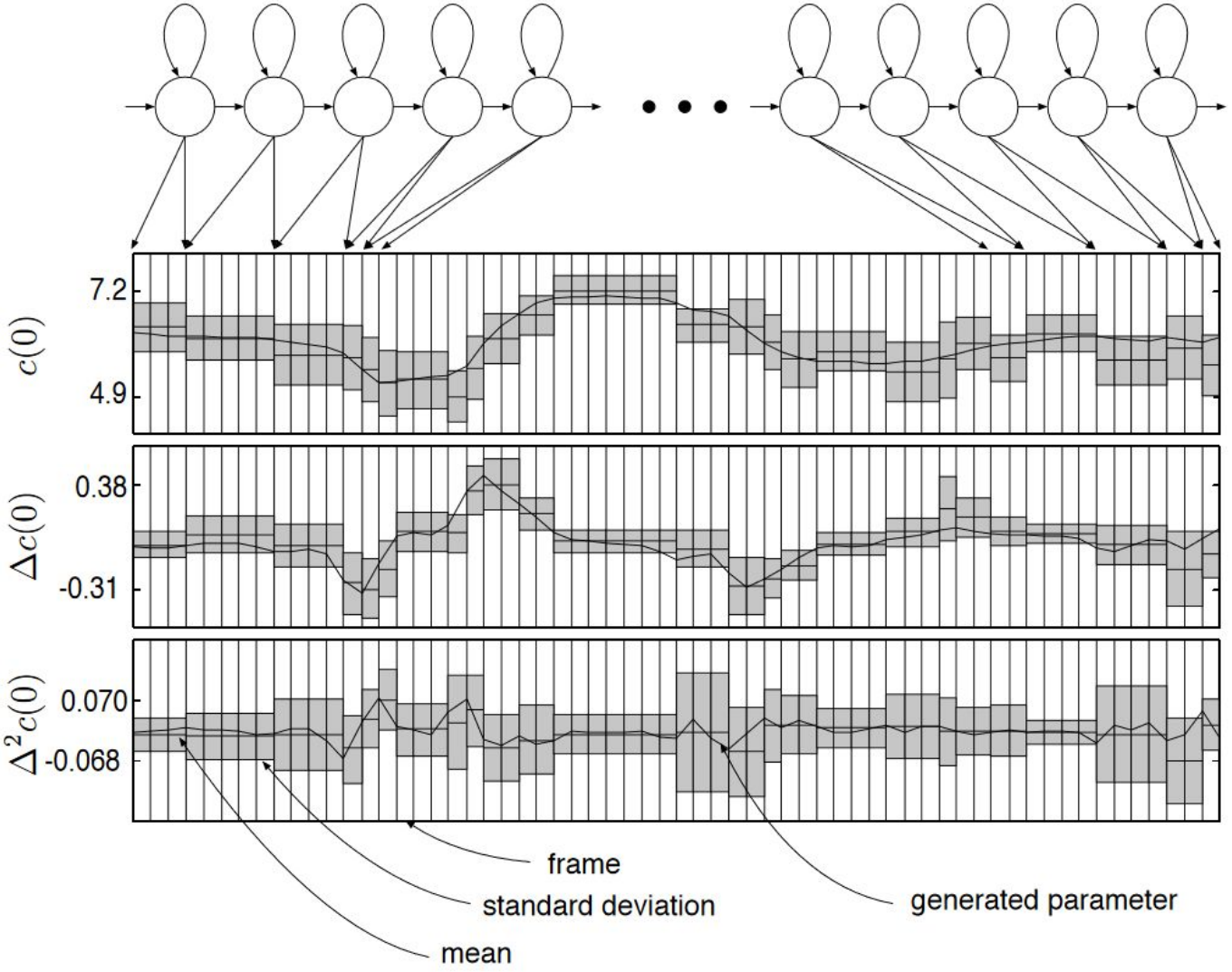
X_t : hidden state variables

y_{ti} : i^{th} observed variable @ t



- Output independence: $P(o_i|q_1...q_T,...,o_1,...,o_i,...,o_T) = P(o_i|q_i)$


a_{11}
a_{22}
a_{33}
π_1
1
a_{12}
2
a_{23}
3
$b_1(\boldsymbol{o}_t)$
$b_2(\boldsymbol{o}_t)$
$b_3(\boldsymbol{o}_t)$
Observation sequence
$\boldsymbol{O}$
$\boldsymbol{o}_1$ $\boldsymbol{o}_2$ $\boldsymbol{o}_3$ $\boldsymbol{o}_4$ $\boldsymbol{o}_5$ . . . . . $\boldsymbol{o}_T$
State sequence
$\boldsymbol{Q}$
1 1 1 1 2 . . . 2 3 . . . 3


7.2
4.9
$c(0)$
0.38
-0.31
$\Delta c(0)$
0.070
-0.068
$\Delta^2 c(0)$
frame
standard deviation
mean
generated parameter

Problems with HMMs

- Because no real-life corpus can ever cover every possible utterance, there will be occurrences of zero-types
- To solve this, the sound is averaged from many sources and smoothed to cover up, which leads to muffled, noisy sound
- The synthesis process consists of vocoding, which makes the output sounds buzzy
- More training data = Better sound: Arrival of DNNs

Literature

Manning, Christopher S & Hinrich Schütze. Foundations of Statistical Natural Language Processing. 1999.

King, Simon. A beginners' guide to statistical parametric speech synthesis. 2010.

Jurafsky, Daniel & James H. Martin. Speech and Language Processing. 2016.

Tokuda, Keiichi & Heiga Zen. Fundamentals and recent advances in HMM-base speech synthesis. 2009.

Zen, Heiga. Deep Learning in Speech Synthesis. 2013.

http://www.festvox.org/voicedemos.html

http://www.cs.cmu.edu/~awb/festival_demos/general.html

http://sp-tk.sourceforge.net/

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