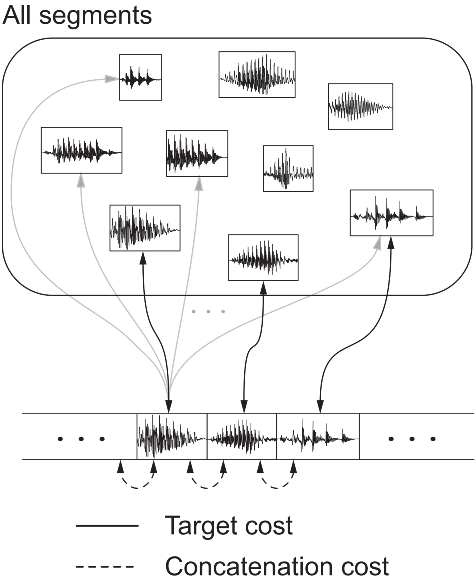
Text-to-Speech Synthesis – Part II

Dr. Mathew Magimai Doss

Outline

Outline

Last Week: Concatenative Synthesis



Zen, Tokuda & Black, 2009

Last Week: Concatenative Synthesis

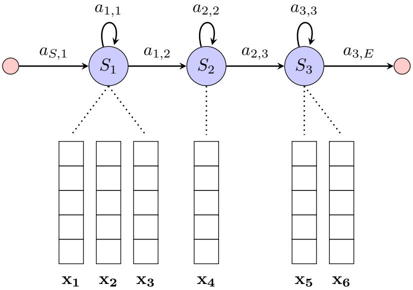
Target cost: Find best match to the target unit, in terms of

- Phonetic context
- F0, stress, phrase position, duration
- Acoustic distance

Join cost: Find a unit that can combine well with neighboring units and has

- Matching formants, energy, F0

These can be seen as emission (*target cost*) and transition (*join cost*) probabilities of HMMs.

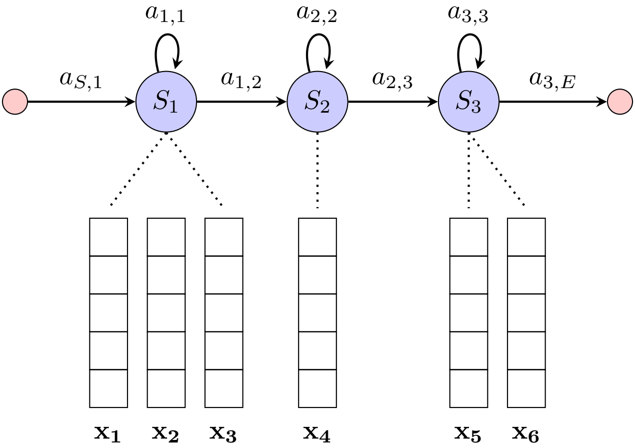


Outline

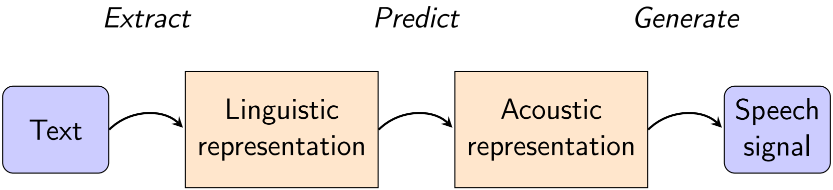
Statistical Parametric TTS

- Uses HMMs, like ASR, but to generate speech.
- Needs less training data, no need to store the unit database.
- Easy to adapt the speech.
- No artefacts from unit joints.
- Buzzy speech quality.
- Interactive online demo

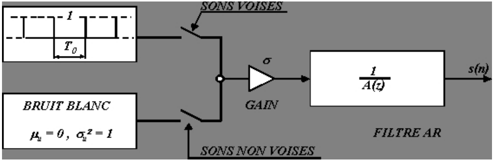
HMM Synthesis



TTS: Basic Steps



Linguistic representation: Context-dependent states (with a lot of context!), representing phonemes
Acoustic representation: Spectral (*system*) and excitation (*source*) features



Linguistic representation

Context for modelling HMM states includes:

- Current, preceding, following phonemes
- Position of current phoneme in syllable
- Numbers of phonemes in current, preceding, following syllables
- Stress and accent of current, preceding, following syllables
- Number of syllables to previous, next stressed syllable
- Position of current word in phrase
- Number of words to next content word
- . . .

HTS linguistic feature specification

Acoustic representation

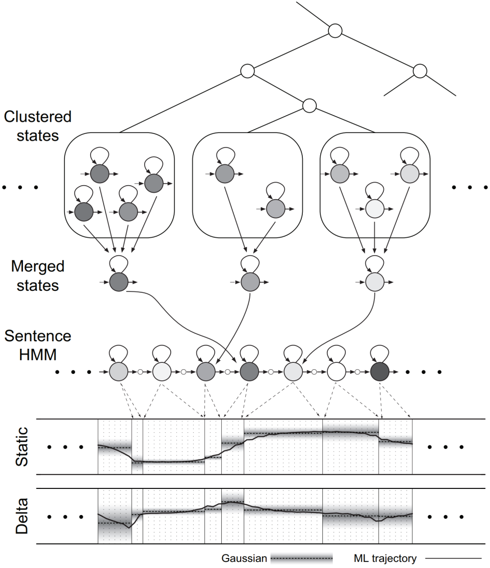
Predicted parameters:

- Spectrum: MFCCs + Δ + $\Delta\Delta$
- Excitation: $\log F_0 + \Delta + \Delta\Delta$
- Possibly further vocoder parameters
- HMM state durations

Prediction models:

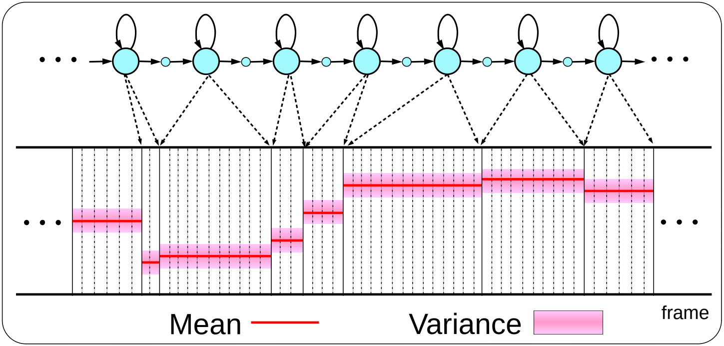
- Regression trees with Gaussian probability distributions
- Neural networks

HMM Synthesis



Zen, Tokuda & Black, 2009

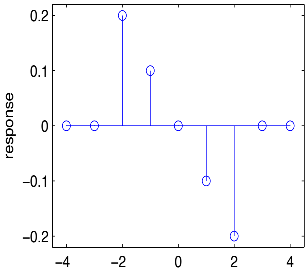
Static Features



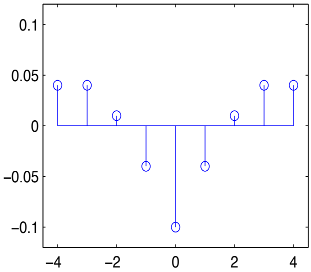
HTS slides

Recall: Temporal derivatives

$$\Delta_{c_m} = \frac{\sum_{k=1}^{K} k \cdot (c_{m+k} - c_{m-k})}{2 \cdot \sum_{k=1}^{K} k^2} \quad (1)$$



Delta (first order derivative)

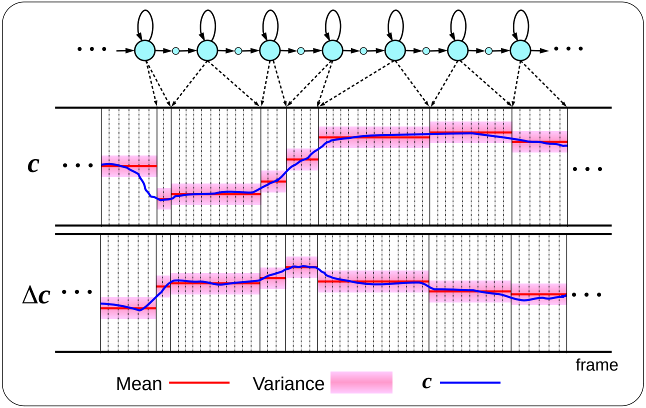


Delta-Delta (second order derivative)

- Savitzky-Golay filtering and temporal derivatives computation

With Dynamic Features

Dynamic features help to generate smooth trajectories.



HTS slides

See formulation of HMM as trajectory model

Duration Modeling

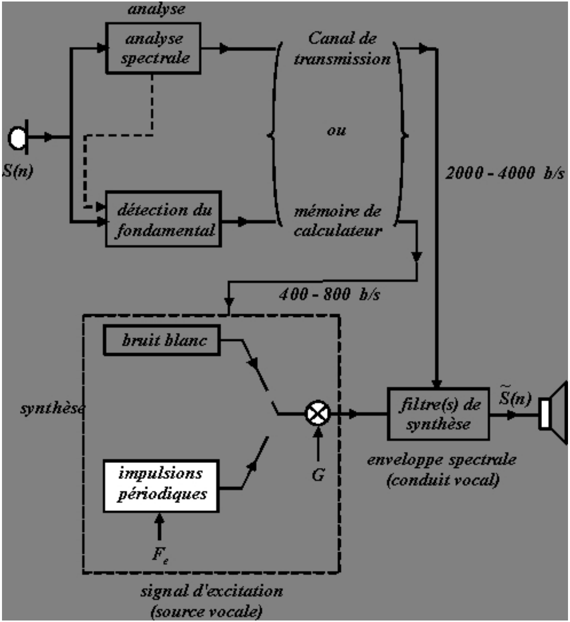
- Normal HMMs model duration through the transition probabilities of self-loops
  - Duration probabilities decay exponentially, which is inaccurate
  - Usually sufficient for ASR, but TTS needs explicit model

Hidden semi-Markov models (HSMM)

- Replace self-transitions with explicit Gaussian duration model

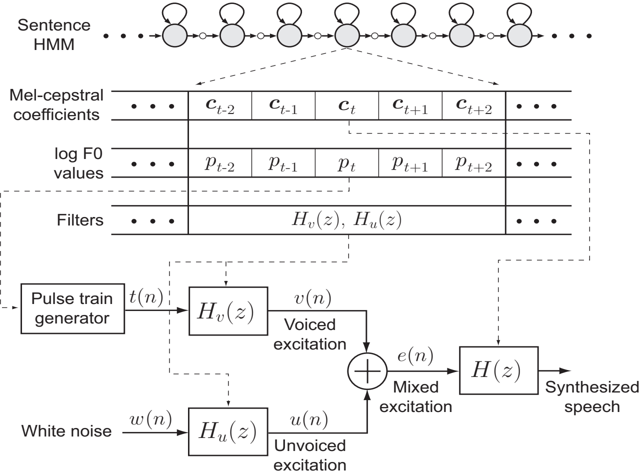
Shun-Zheng Yu,Hidden semi-Markov models, Artificial Intelligence, Vol. 174, 2010, pp 215–243.

Vocoding: recall LP-based speech coding



Vocoding: applied to HMM-based TTS

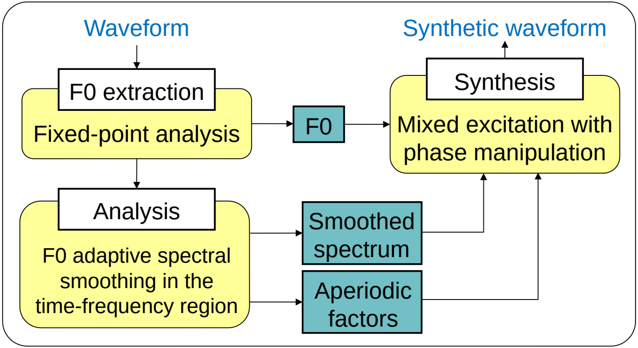
Waveform generation using source-filter model given cepstral feature and F_0 information estimates



source: Zen, Tokuda & Black, 2009

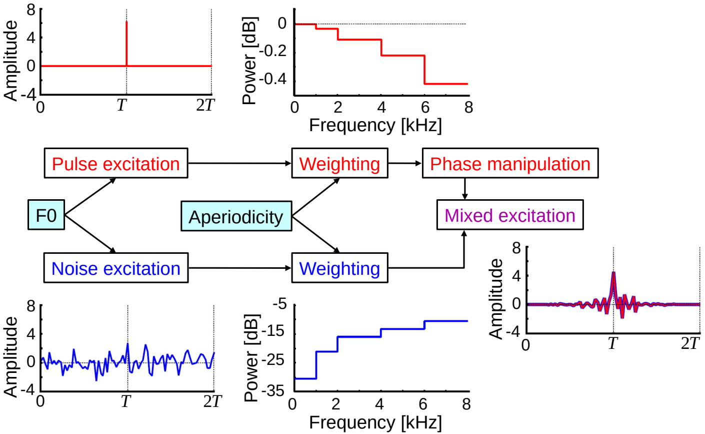
STRAIGHT Vocoder

Speech **T**ransformation and **R**epresentation by **A**daptive **I**nterpolation of wei**GHT**ed spectrogram

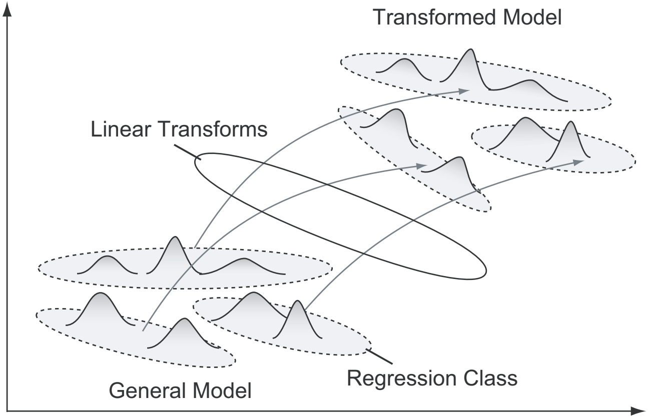


HTS slides

STRAIGHT excitation generation



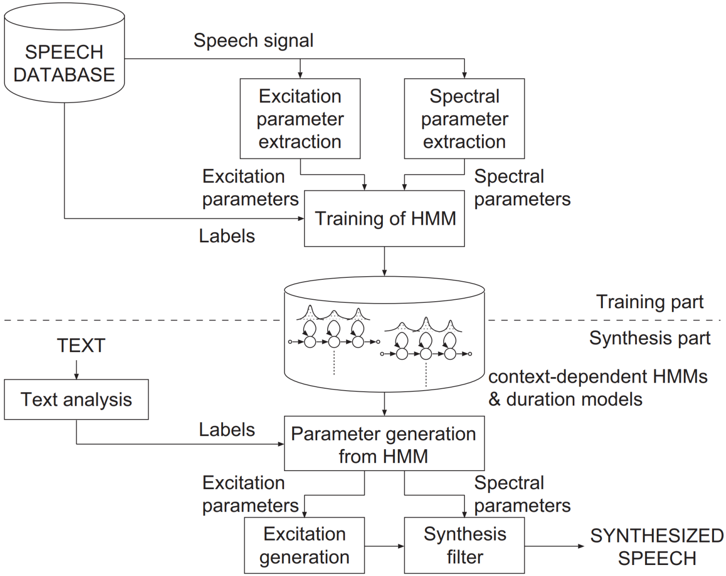
Speaker Adaptation



source: Zen, Tokuda & Black, 2009

Using adaptation techniques such as, maximum likelihood linear regression (MLLR) (applied to model parameters), constrained MLLR (applied to features).

HTS System Overview



Zen, Tokuda & Black, 2009

Summary: HMMs for ASR vs. TTS

	ASR	**TTS**
Acoustic features	About 13 spectral parameters + Δ + $\Delta\Delta$	40–60 spectral parameters + Δ + $\Delta\Delta$ + source features
Frame shift	10 ms	5 ms
Modeling unit	Triphone	Quinphone with full linguistic context
States per model	3	5
State emission distribution	GMM	Single Gaussian
Duration model	HMM self-loops	Explicit model (HSMM)
Parameter estimation	Baum-Welch (EM)	Baum-Welch (EM)
Decoding	Viterbi search	Not usually required
Generation	Not required	Maximum likelihood

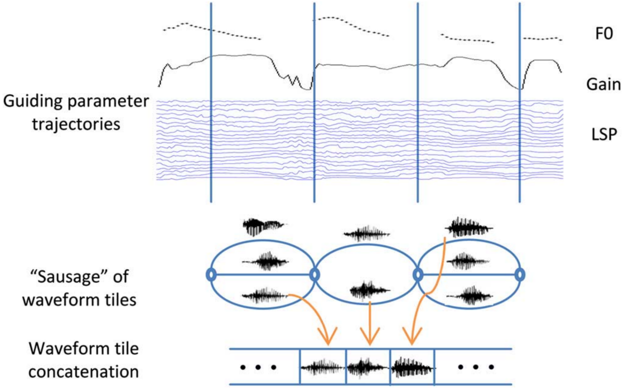
Dines et al., 2010 and *King, 2011*

Outline

Hybrid TTS

Statistically driven unit selection synthesis:

- Like SPSS, but replace the vocoder with unit concatenation
- Like unit selection, but select units based on predicted acoustic parameters



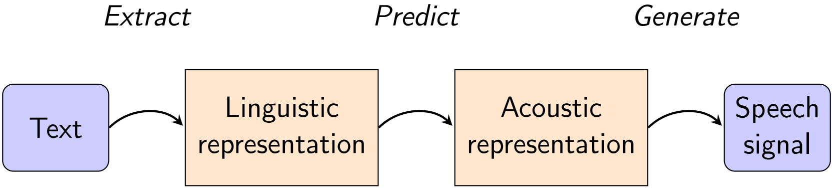
Qian, Soong & Yan, 2012

Outline

End-to-end TTS

Aims to replace hand-crafted TTS components with neural networks, in particular:

- NLP pre-processing pipeline
  - Text normalization (difficult!)
  - Lexicons and grapheme-to-phoneme (G2P) conversion
- Vocoding

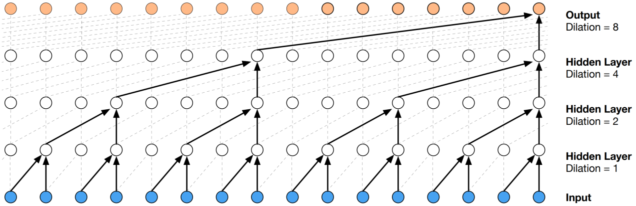


WaveNet (Neural Vocoder)

- Directly predicts speech samples ($x[1] \ldots x[n] \ldots x[N-1]$) given acoustic and linguistic features f

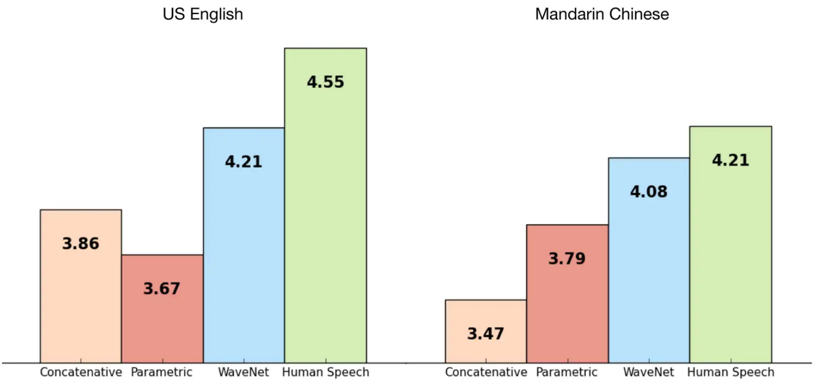
$$\prod_{n=1}^{N-1} p(x[n] | x[n-1], \ldots x[0], f)$$

  - Dilated convolutions allow covering long ranges
- Initially very slow, but now used for real-time, cloud-based TTS. Still requires non-negligible computing resources.
- Very natural speech compared to traditional SPSS (Samples)
- Can train one model for multiple speakers



van den Oord et al., 2016

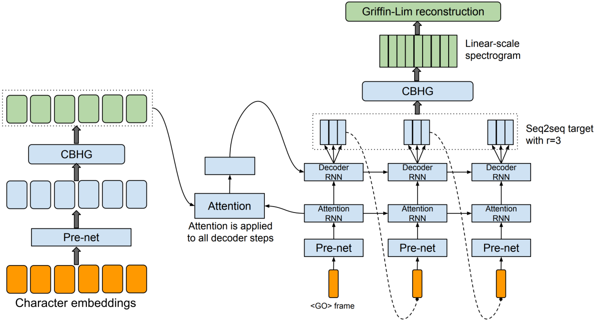
WaveNet MOS Scores



van den Oord et al., 2016

Tacotron

- Generate spectrograms directly from text (character embeddings)
  - No need for HMM alignments to train
  - No need for G2P conversion
  - Assumes text normalized input ("16" is "sixteen")
- Spectrogram inversion with Griffin-Lim algorithm (Tacotron 2 uses WaveNet)



Wang et al., 2017

TTS Training Data Requirements

Typical amounts of data required for training:

Architecture	**Training data**
Diphone synthesis	1 instance per diphone (total of 1000)
Unit selection	5–40 hours (Taylor, 2009), difficult to adapt to new speakers (can use voice conversion)
Statistical parametric synthesis	5+ hours for initial system, but can easily adapt to new speakers
WaveNet	25+ hours, but can combine multiple speakers and can adapt to new speakers with <10 minutes (Chen et al., 2019)

Outline

Evaluating Synthetic Speech

- Subjective vs. objective evaluation
- Naturalness vs. intelligibility
- Other evaluation measures?

Subjective vs. Objective Evaluation

Subjective

- Tests with human listeners
- Expensive and time-consuming, but very flexible
- Not easy to reproduce

Objective

- Automatic evaluation through computers
  - E.g. with ASR systems or by measuring distances to human reference samples
- Cheap and fast, but more difficult to interpret
- Reproducible

Naturalness vs. Intelligibility

Naturalness

- Material: Target domain text or phonetically balanced sentences
- Metrics: Mean opinion scores (MOS) $\{\mathrm{Excellent}, \mathrm{Good}, \mathrm{Fair}, \mathrm{Poor}, \mathrm{Bad}\}$, A/B preference tests

Intelligibility

- Material: Semantically unpredictable sentences
  The dog fights under the red beach.
  The deaf dress sees the bear.
  When does the gold take the beige fear?
  The wheat attempts the time trembling.
  The real glass opens the corner.
  Turn the date or the hand.
- Metric: Word error rate
- Can be made more challenging by adding noise

Blizzard Challenge

- Yearly challenge task to build TTS systems on a shared dataset
- Thorough, centrally organized listening tests
- http://www.festvox.org/blizzard/

Thank you for your attention!

Dr. Mathew Magimai Doss

Idiap Research Institute, Martigny, Switzerland