Text-to-Speech Synthesis

HUANG-WEI CHEN

DEPT. OF COMPUTER SCIENCE AND INFORMATION ENGINEERING

SPOKEN LANGUAGE PROCESSING LAB.

Reference:

1. Paul Taylor, Text-to-Speech Synthesis
2. Heiga Zen et. al. Statistical parametric speech synthesis
3. Andrew J. Hunt et. al. Unit selection in a concatenative speech synthesis system using a large speech database
4. Speech Synthesis Wikipedia, http://en.wikipedia.org/wiki/Speech_synthesis
5. HTS official slide, http://hts.sp.nitech.ac.jp/archives/2.2/HTS_Slides.zip

History of Speech Synthesis(1/2)

1779

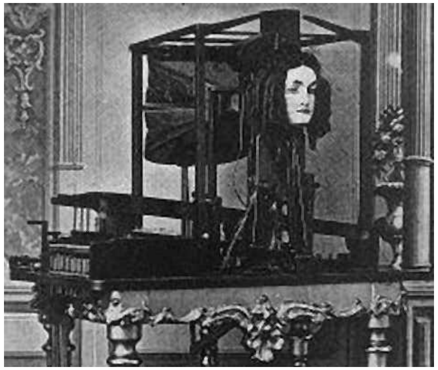
- The Danish scientist built models of the human vocal tract that could produce 5 long vowel sound.

1791

- Acoustic-mechanical speech machine
- Add models of lips and tongue, so consonants(子音) could be produced.

1837+

- Speaking machine
- Euphonia



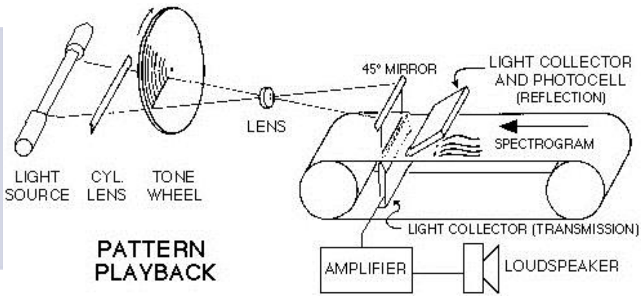
History of Speech Synthesis(2/2)

1930s

- Bells Lab develop a vocoder can analyzed speech into f0 and resourances(共鳴).
- Homer Dudley developed a keyboard-operated voice synthesizer.

1940s-1950s

- Pattern playback



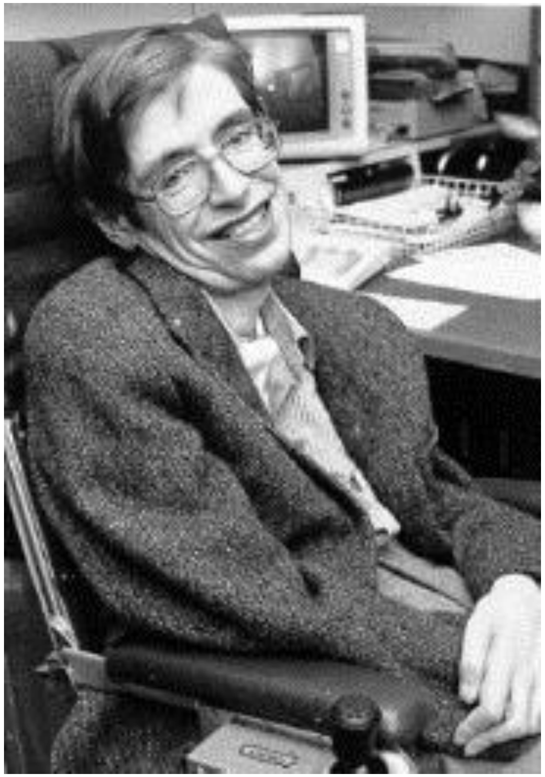
1960s

- In 1961, physicist John Larry Kelly Jr. used an IBM 704 computer to synthesize speech.
- In 1968, the first general English TTS system developed by Noriko Umeda.

http://en.wikipedia.org/wiki/Pattern_playback

The most famous TTS user in the world

[Stephen Hawking's Talk: Questioning the universe]



Stephen Hawking and he's device.

The most famous TTS software in the world(...Might be)

[Song]

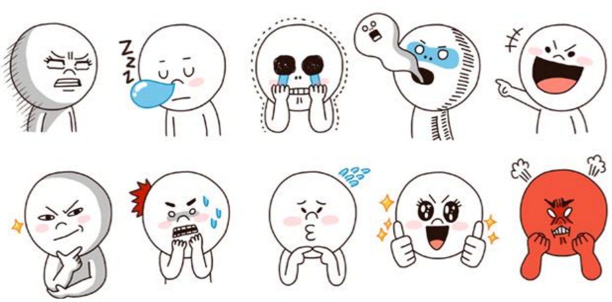
A song synthesis software, Vocaloid2: Hatsune Miku

A thing always happens in our daily life

How to communicate?

By talking…

By texting…



Or even just by a figure

But what is behind the communication?

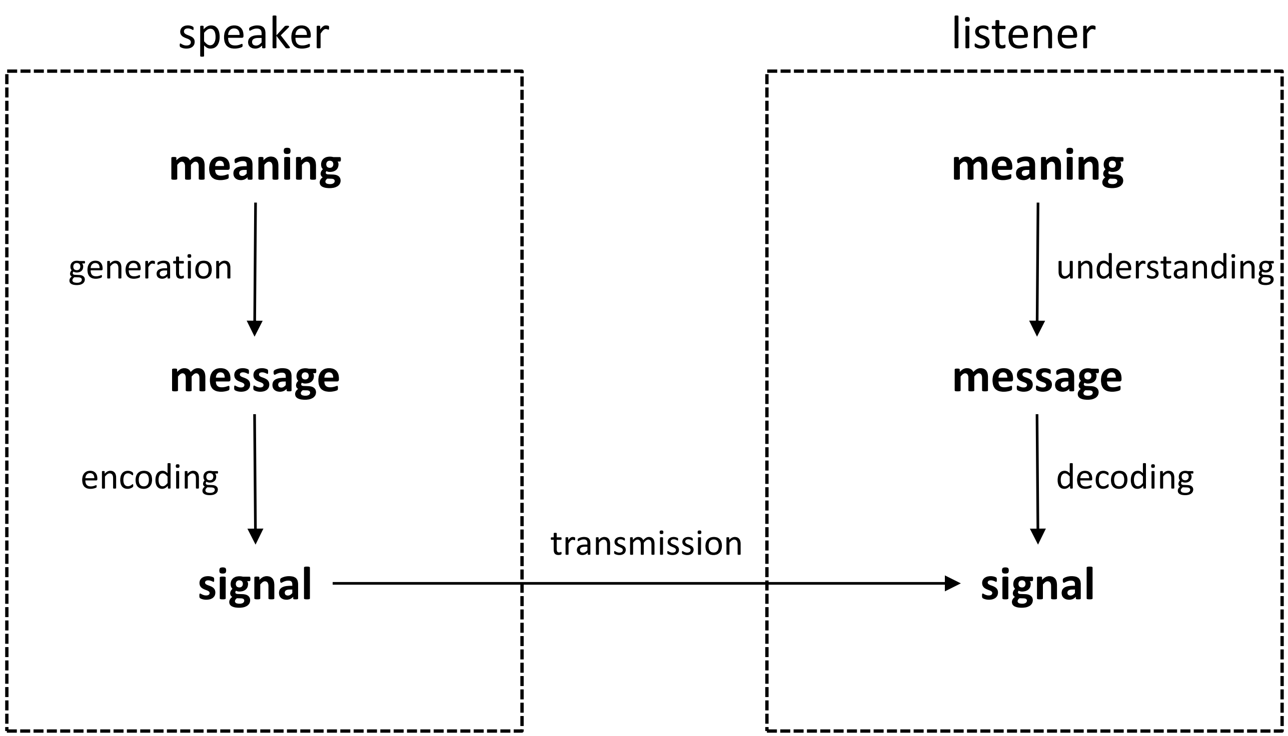


Figure 2.3 Processes involved in communication between two speakers

Thinking a TTS system as a speaker

A TTS system should contain following elements:

1. Text understanding (including text segmentation, organization)
2. Text decoding (part-of-speech, NLP technology)
3. Prosody prediction
4. Phonetics, phonology and pronunciation
5. Speech generation technique

In the Following Slides:

We focus on two part:

1. Text-analysis
2. Speech generation technique

Text Analysis: the Problem

Overview of the problem:

- The job of the text-analysis system is to take arbitrary text as input and convert this into a form more suitable to subsequent linguistic processing.

Example:

Write a cheque from acc 3949293 (code 84-15-56), for $114.34, sign it and take to down to 1134 St Andrews Dr, or else!!!

Text Analysis: the Overview Process(1/2)

Pre-processing:
identification of the text genre

Sentence splitting:
segmentation of the document

Tokenization:
segmentation of each sentence

Text Analysis: the Overview Process(2/2)

- Semiotic classification:
  classification of each token as one of the semiotic classes
- Verbalization:
  Conversion of non-natural-language semiotic classes into words
- Homograph resolution:
  determination of the correct underlying word
- Parsing and prosody prediction

Text Analysis: in Mandarin

	Versus	
Sentence	=?	Utterance
Character	=?	Word
『長』度	=?	『長』官
『不』要	=?	對『不』起

Features in Mandarin

1. Tonal language
2. A character may have difference pronunciation
3. Tone Sandhi problem
4. A Initial and a final composes a character

Technology of TTS

- Concatenative synthesis
- Unit selection synthesis
- Diphone synthesis
- Domain-specific synthesis
- Formant synthesis
- Articulatory synthesis
- HMM-based synthesis

http://en.wikipedia.org/wiki/Speech_synthesis

Speech Synthesis Methods(1/2)

Corpus-based, concatenative synthesis

- Concatenate speech units (waveform) from a database
- Need large data and automatic learning
  - ➢High-quality speech can be synthesized
- Well-known methods
  1. Diphone synthesis
  2. Unit selection synthesis
- Disadvantages
  1. Large database is needed
  2. If the phone combination does not exist in database, the synthetic voice will sound unnatural

Speech Synthesis Methods(2/2)

Corpus-based, statistical parametric synthesis

- Source-filter model and statistical acoustic model
  - ➢Flexible to change its voice characteristic
- HMM as its statistical acoustic model
  - ➢Natural sounding voice can be synthesized with small corpus
  - ➢Familiar with HMM in ASR (Automatic Speech Recognition)
- Well-know methods
  1. HTS speech synthesis
- Disadvantages
  1. Synthetic voice is less high-quality
  2. Corpus should be balanced, nor some phone can not get good sounding after synthesized

Concatenative Speech Synthesis

It is the most subjective way of speech synthesis.

One problem comes up with this method:
Which “unit” is better for speech synthesis? A sentence? A word? Or a phoneme?

Definition of an unit

Considering a large unit like a sentence

- Although it will get the best quality when the target sentence appears in the corpus…but if it doesn't appear in the corpus?

Considering a medium unit like a word

- The discontinuous between words might be serious.
- New words born everyday.
- In total, a large unit might get trouble in synthesis.

Considering a small unit like a phoneme

- Weird sound might be synthesized.
- A phoneme has many candidates to choose.
- In total, a small unit might cause unnatural sound.

Solution

Due to the fact of we can not collect new word once by once, phoneme will be the best choose in concatenative TTS.

The reason of weird sound is because of choosing wrong phoneme candidates.

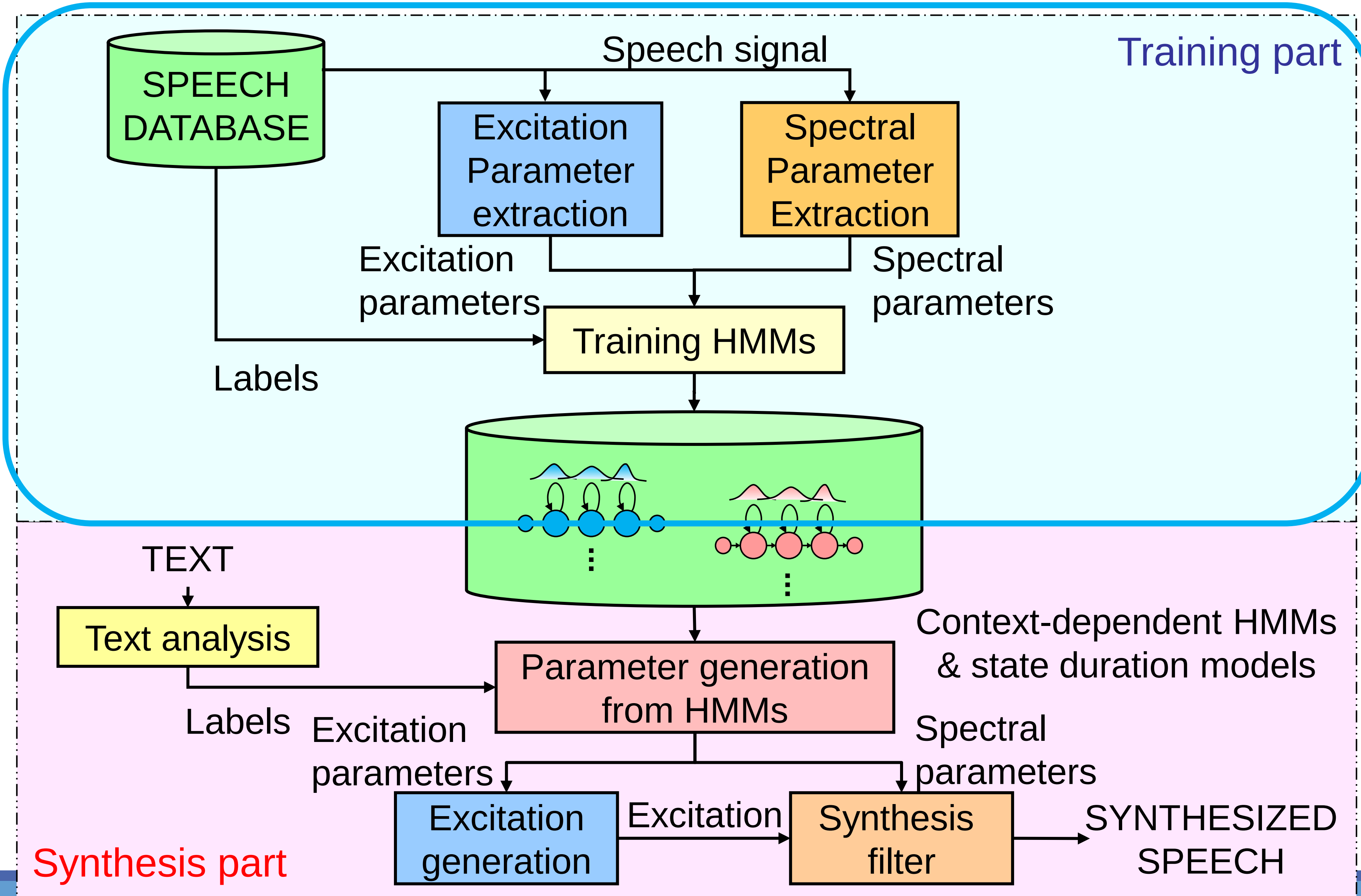
Some solution is provided:

1. A cost function is used.
2. HMM-based unit-selection
3. Diphone speech synthesis

Weakness of concatenative TTS

1. A large database is needed.
2. Each unit should be cut clearly.
3. We loss the flexibility of controlling the characteristic.
4. Due to so many candidates can be chosen in a unit, the cost is large.

HMM-based speech synthesis system



This figure references from HTS official slides, p.21

Context Labeling

Many contextual factors can affect the production of human speech, some important contextual factors like phone identity, stress, accent, position, part of speech, etc.

In HMM-based speech synthesis, the labels of the HMMs is composed of a combination of these contextual factors.

Spectral Parameter Extraction

Mel-generalized Cepstral coefficient is used in our system

Assume that $c_{\alpha,\gamma}(m)$ is mel-generalized cepstral coefficient

$$H(z) = \begin{cases} \left(1 + \gamma \sum_{m=0}^{M} c_{\alpha,\gamma}(m)\, z_{\alpha}^{-m}\right)^{1/\gamma}, -1 \leq \gamma < 0 \\ exp \sum_{m=0}^{M} c_{\alpha,\gamma}(m)\, z_{\alpha}^{-m}, \gamma = 0 \end{cases}$$

$z_{\alpha}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}$ $(|\alpha| < 1)$ First-order all-pass function

If $\gamma = 0$, $c_{\alpha,\gamma}(m)$ is called mel-cepstral coefficients

Excitation Parameter Extraction

Robust Algorithm for Pitch Tracking (RAPT) is used in our system.

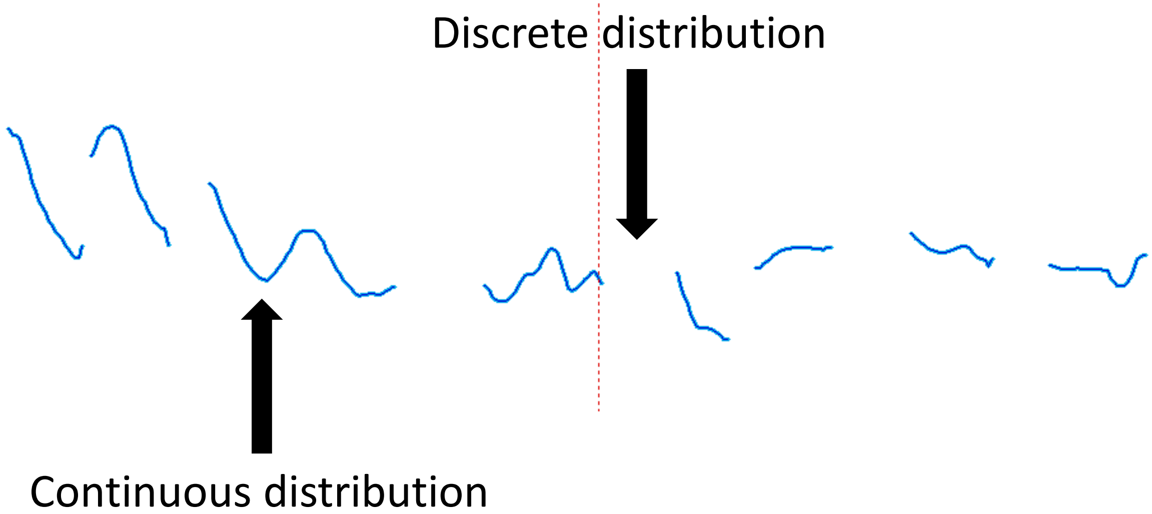
RAPT is the extension of Normalized Cross-correlation Function (NCCF)

$$NCCF(\eta) = \frac{\sum_{j=1}^{n} s(j)s(j+\eta)}{\sqrt{e_0 e_\eta}}, e_j = \sum_{k=j+1}^{j+n} s(k)^2$$

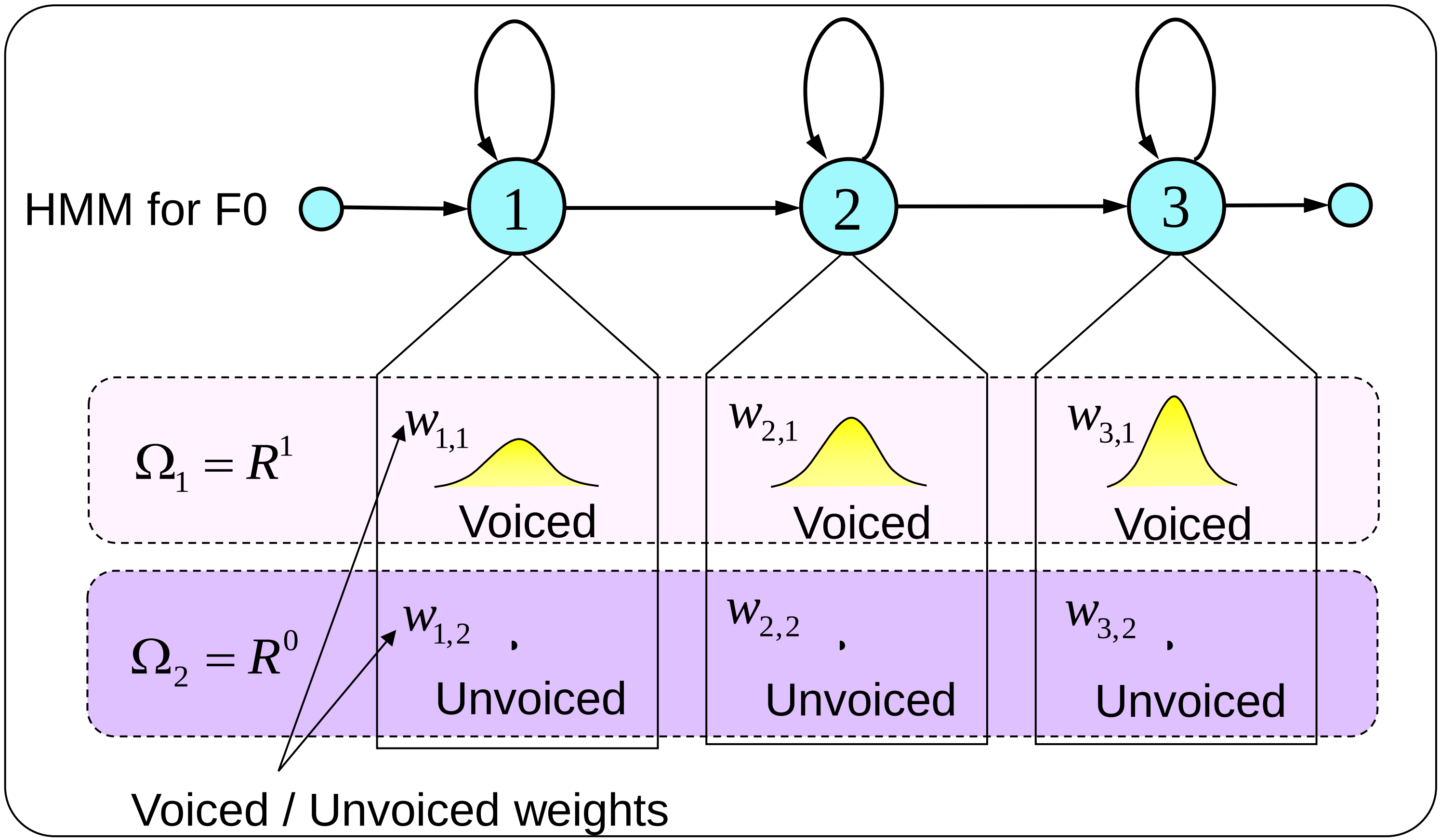
1. We use NCCF to calculate low sampling rate for all lags in the F0 range of interest. And the local maximum of each frame is saved.
2. Compute the NCCF of the high sample rate signal only in the vicinity of promising peaks found in the first pass. Search again for the local maximum in this refined NCCF to obtain improved peak location and amplitude estimates.
3. Each peak retained from the high-resolution NCCF generates a candidate F0 for that frame. At each frame the hypothesis that the frame is unvoiced is also advanced.
4. DP is used to select the set of NCCF peaks or unvoiced hypotheses across all frames that best match the characteristics mentioned above.

Multi-Space probability Distribution HMM(1/2)

After we extract the pitch from the data, we find that we can not model pitch contour using HMM because discrete and continuous distribution exist at same data



Multi-Space probability Distribution HMM(2/2)



Training Monophone

Speech data & labels

This figure references from HTS official, p.56

Compute variance floor (HCompV)

Calculate minimum variance of speech parameter

Initialize CI-HMMs by segmental k-means (HInit)

Create monophone HMMs for all phonemes and initialize these parameters

Reestimate CI-HMMs by EM algorithm (HRest & HERest)

Estimate parameters with training data (Context-independent : CI)

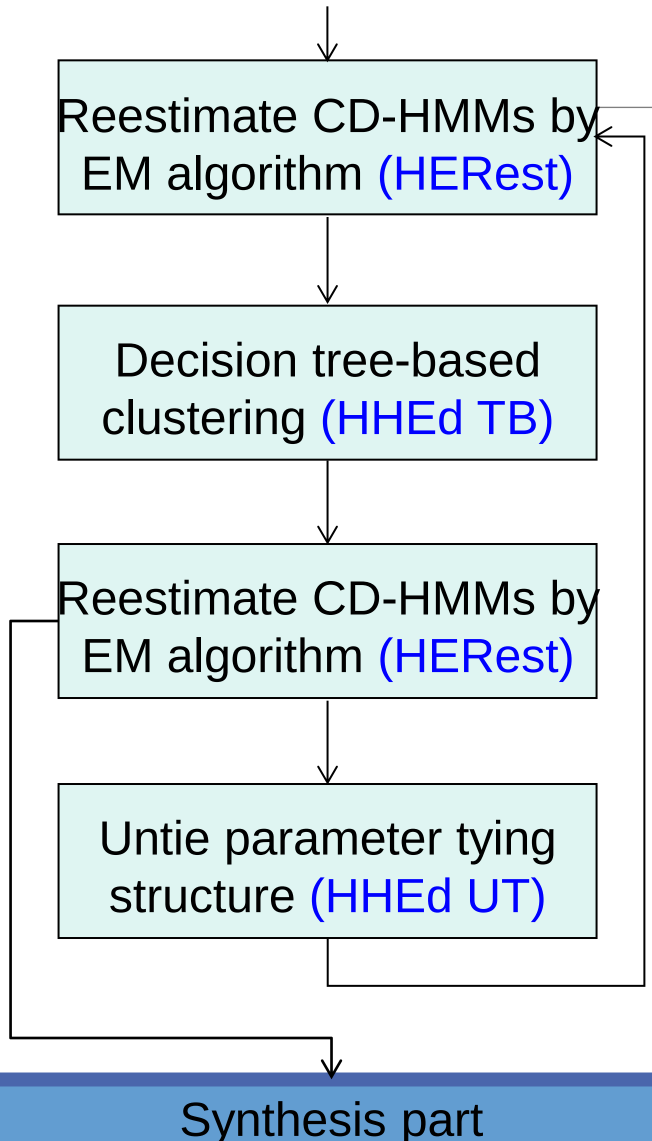
Copy CI-HMMs to CD-HMMs (HHEd CL)

Concatenate monophone HMMs according to fullcontext labels (Context-dependent : CD)

Training fullcontext HMMs

Training Fullcontext

This figure references from HTS official, p.57



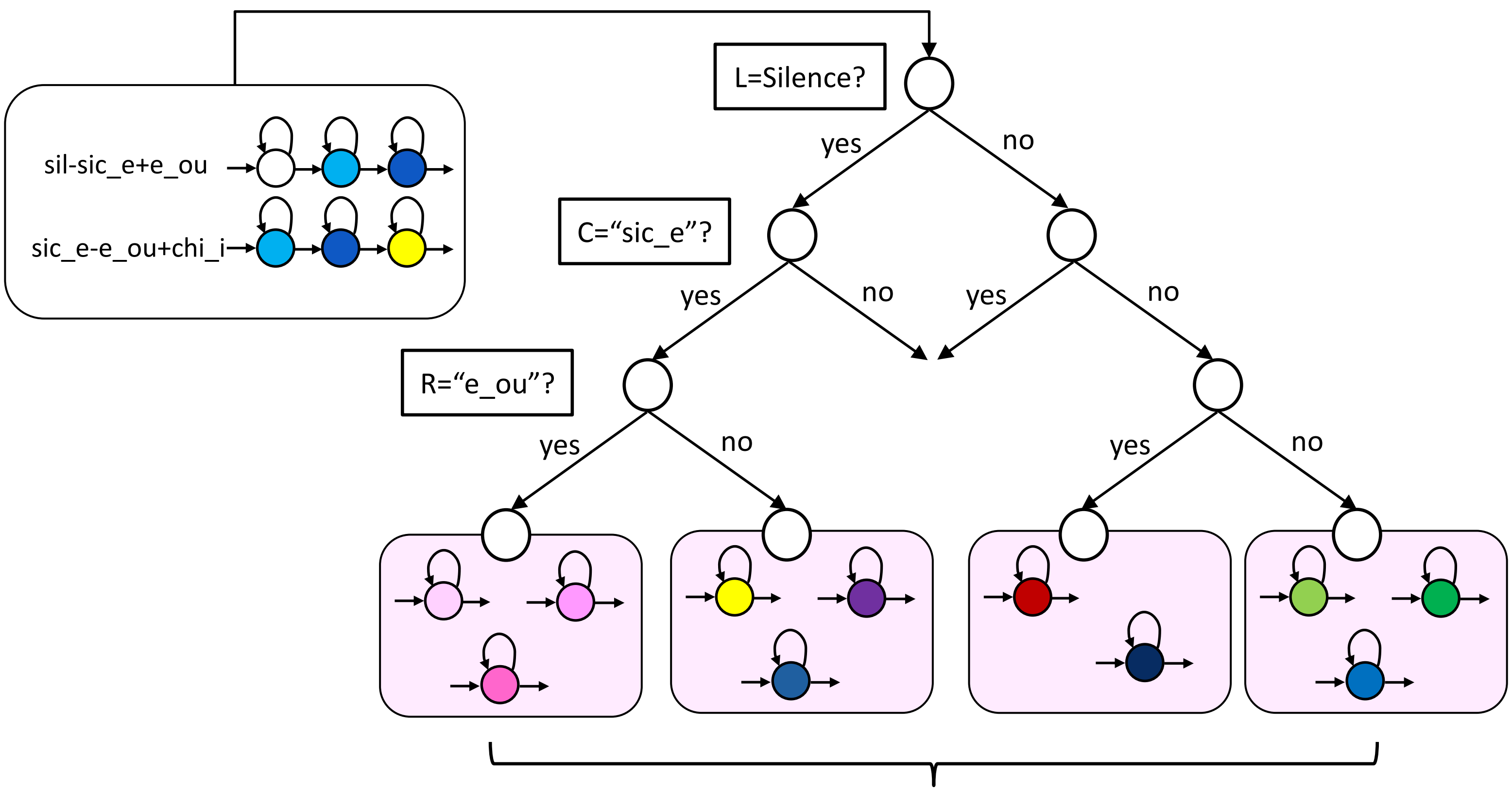
Estimate parameters with training data
(parameters are not shared)

Construct a decision tree by
the context clustering technique
(Use TB command)

Estimate parameters with training data
(parameters are shared)

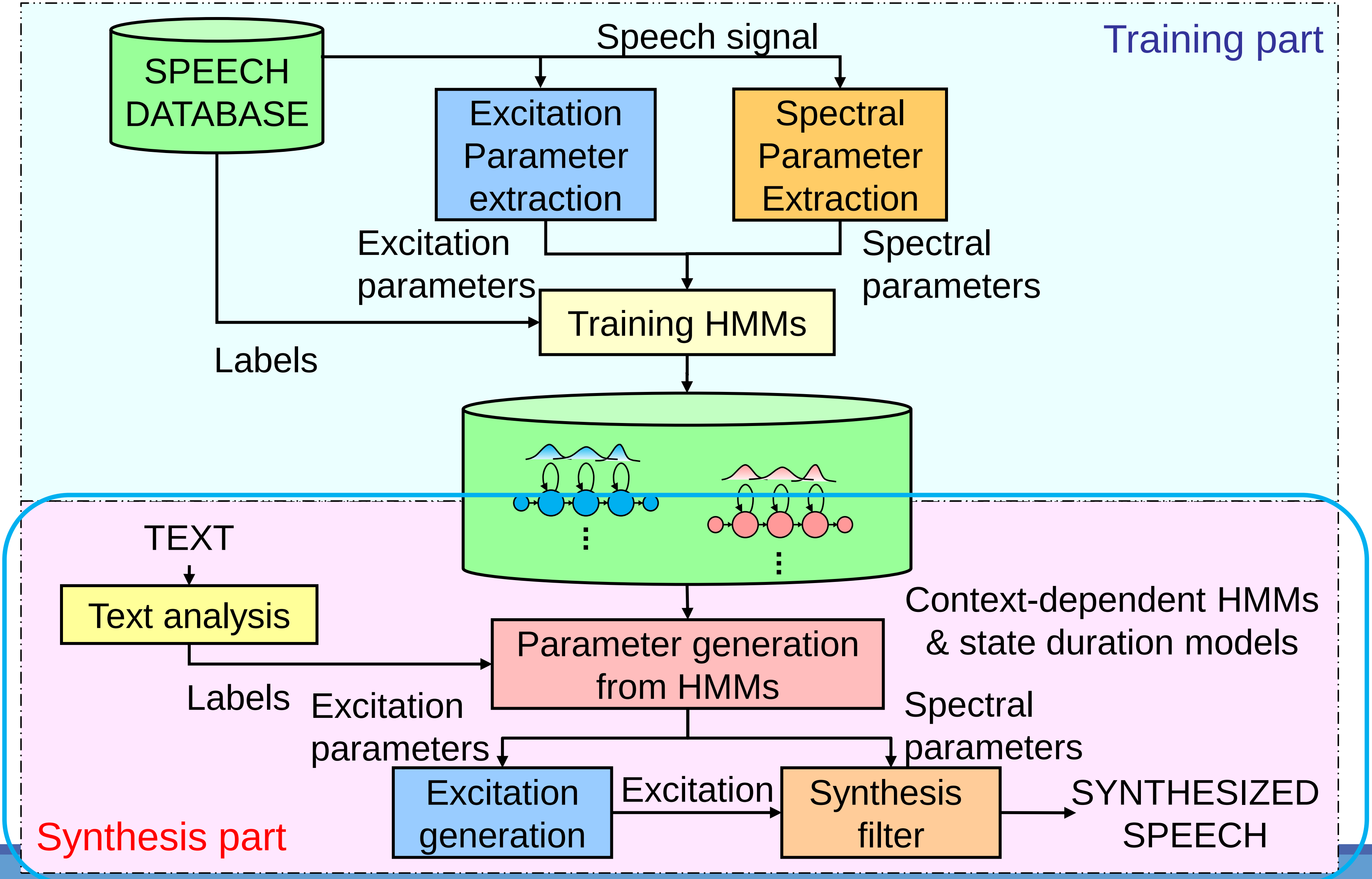
Untying shared structure
(Use UT command)

Decision Tree-based Clustering



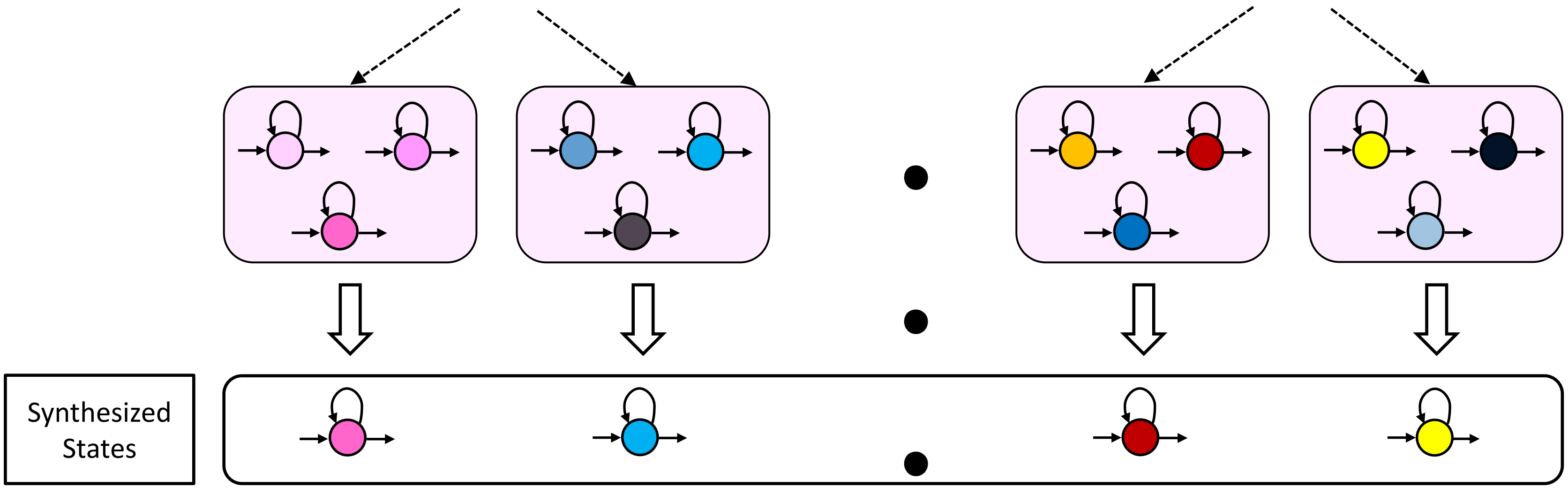
Sharing the same HMMs' parameters in the same leaf node

HMM-based speech synthesis system



This figure references from HTS official, p.21

Synthesis from Context-label



Each state is connected according to the context-label

If the an unknown HMM in develop set is needed to synthesize, its parameter can be generated by the familiar HMM

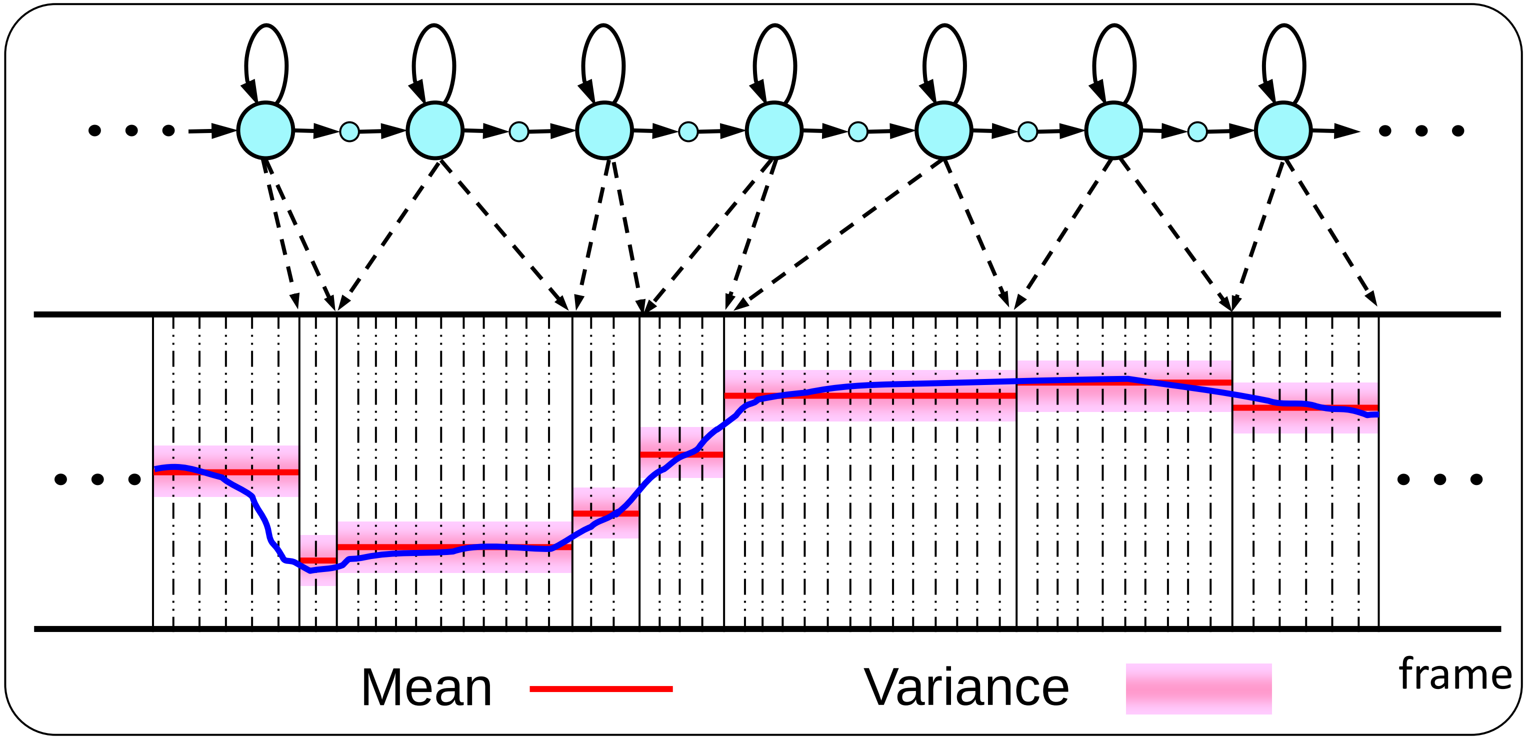
Speech Parameter Generation from HMMs(1/3)

For Given a HMM λ, determine a speech vector sequence $\mathrm{O} = \{o_1, o_2, \cdots, o_T\}$, and o_t is consist of by cepstral coefficient and its delta cepstral coefficient, which maximum

$$P[O|\lambda] = \sum_{all\ q} P[q, O|\lambda] = \sum_{all\ q} P[q|\lambda] \cdot P[O|q, \lambda]$$

$$\approx \max_{q} P[q|\lambda] \cdot P[O|q, \lambda]$$

Speech Parameter Generation from HMMs(2/3)



$\hat{o}$ becomes a sequence of mean vectors
⇒ discontinuous outputs between states

This figure references from HTS official, p.29

Speech Parameter Generation from HMMs(3/3)

To maximum the parameter vector sequence, actually is to maximum cepstral coefficient parameter, so we get

$$\frac{\partial \log P[O|q,\lambda]}{\partial c} = \frac{\partial \log P[Wc|q,\lambda]}{\partial c} = 0_{TM}$$

And we can obtain

$$W^{\top}\Sigma_q^{-1}Wc = W^{\top}\Sigma_q^{-1}\mu_q$$

Where

$$c = [c_1^{\top}, c_2^{\top}, \cdots, c_T^{\top}]$$

$$\mu_q = \left[\mu_{q_1}^{\top}, \mu_{q_2}^{\top}, \cdots, \mu_{q_T}^{\top}\right]$$

$$\Sigma_q = \left[\Sigma_{q_1}^{\top}, \Sigma_{q_2}^{\top}, \cdots, \Sigma_{q_T}^{\top}\right]$$

Synthesis Filter

Mel Log Spectrum Approximation (MLSA) Filer is used for synthesizing speech

$$H(z) = \exp(F_\alpha(z))$$

$$F_\alpha = \sum_{m=0}^{\mathrm{M}} c_{\alpha,\gamma}(z) z_\alpha^{-m}$$

Due to exponential function cannot realize, so by using Pade Approximation

$$H(z) = \exp(F_\alpha(z)) \approx R_L(F_\alpha(z))$$

Scoring on TTS System

Objective Evaluation:

1. Log likelihood evaluation
2. Log likelihood difference

Subjective Evaluation :

1. Mean Opinion Scale (MOS) test
2. Preference test
   - AB test
   - Non-AB test

Research Groups

Nagoya Institute of Technology

- HMM-based speech synthesis
- http://hts.sp.nitech.ac.jp/

訊飛科技

- Hybrid TTS
- http://www.iflytek.com/

Wakayama University

- 音声分析変換合成法STRAIGHT
- http://www.wakayama-u.ac.jp/~kawahara/STRAIGHTadv/index_j.html

Conclusion

In a human interaction system, TTS plays an important part.

In recently research, TTS has two main genres:

- Corpus-based concatenative TTS: High quality sounding can be generated but it relies on a large database and also loss the flexibility of characteristic changing.
- Corpus-based statistical parameter TTS: Natural sounding can be generated from a small database but the quality is less good compare with original sounding.