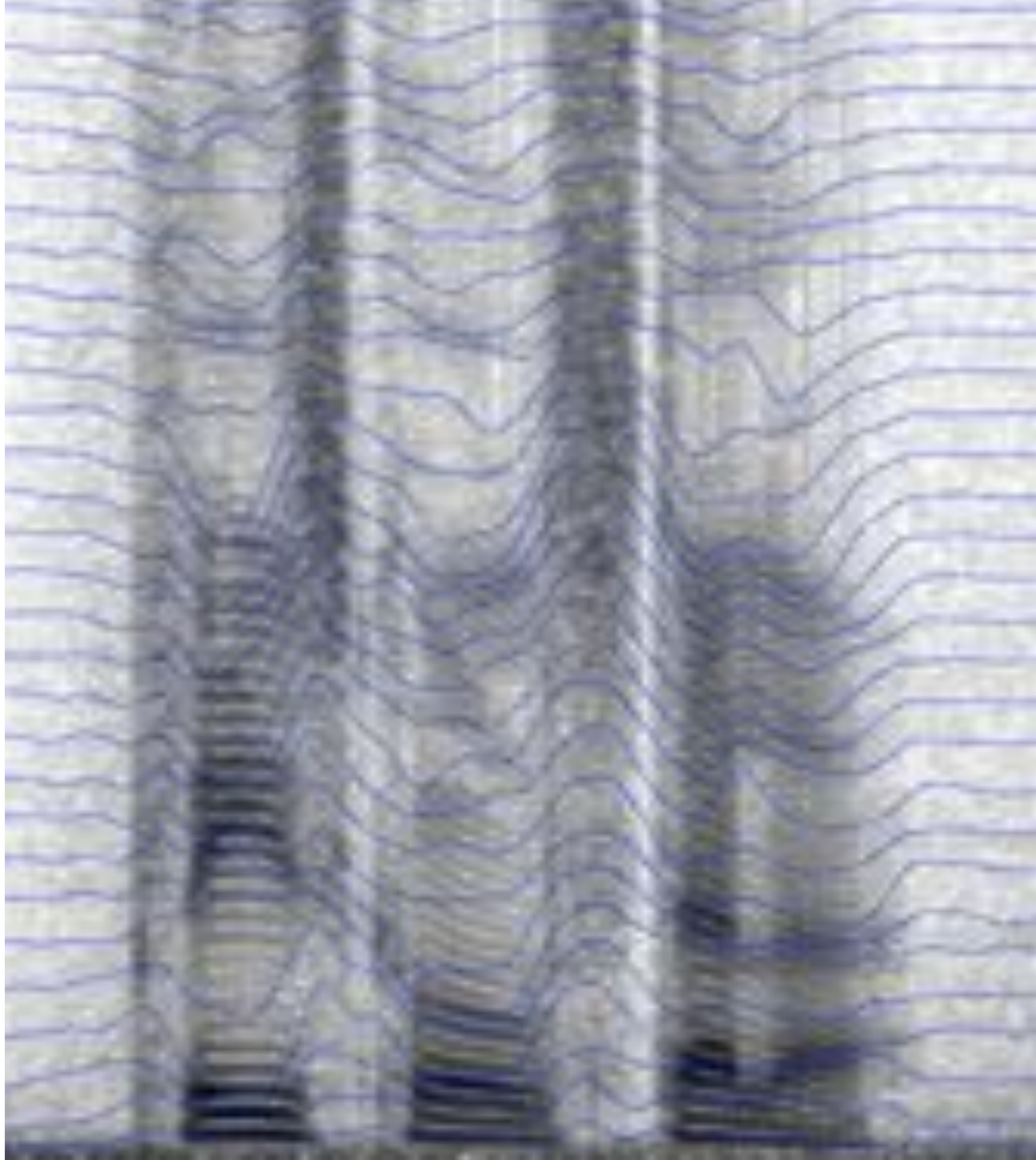
Speech Synthesis

Simon King
University of Edinburgh



Hybrid speech synthesis

- Partial synthesis
- Case study: Trajectory Tiling

Orientation

- SPSS (with HMMs or DNNs)
  - flexible, **robust** to labelling errors

→ but naturalness is limited by vocoder (amongst other things)

- Unit selection
  - potentially excellent **naturalness**
  - target cost and join cost

→ but strongly affected by labelling errors

→ hard work to optimise on new data

- Hybrid synthesis
  - robust statistical model
  - waveform concatenation

→ potential to **combine** the best properties of SPSS and unit selection

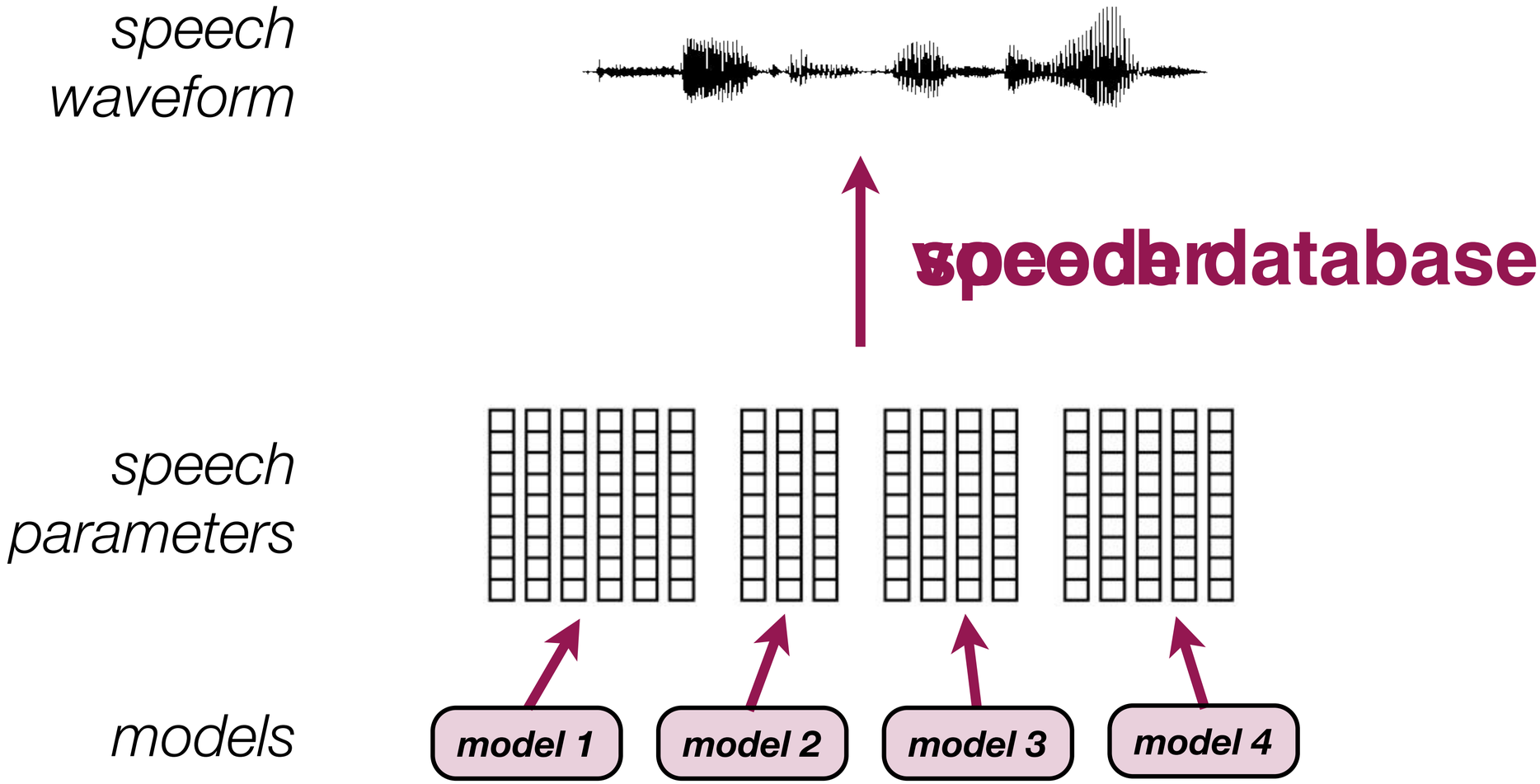
What you should already know

- Signal processing
  - ways to parameterise speech signals
    - for classification (e.g., MFCCs)
    - for vocoding
- Unit selection
  - sparsity in linguistic and/or acoustic space
  - understanding of IFF, ASF target cost
- SPSS
  - sequence-to-sequence regression
  - HMMs & DNNs

Hybrid speech synthesis

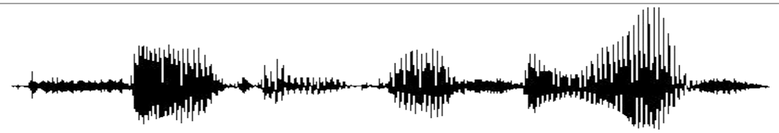
- Partial synthesis
- Case study: Trajectory Tiling

Hybrid speech synthesis, as SPSS with a replacement for the vocoder

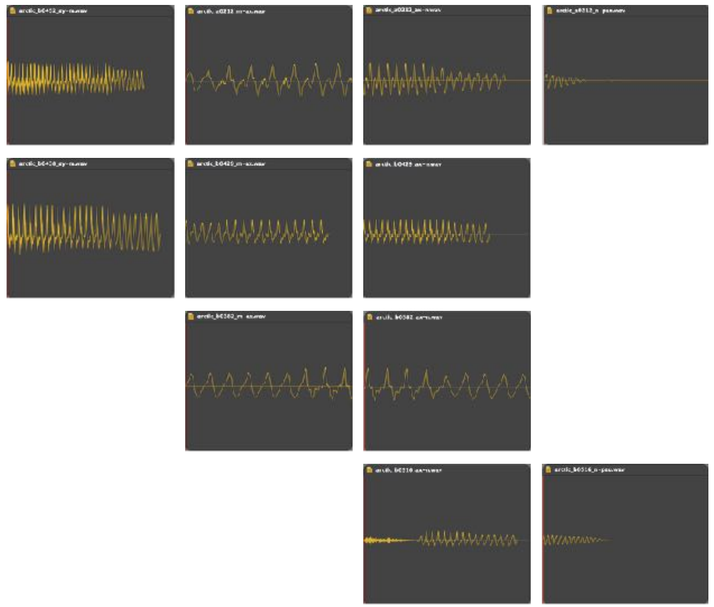


Hybrid speech synthesis, as unit selection with an ASF target cost function

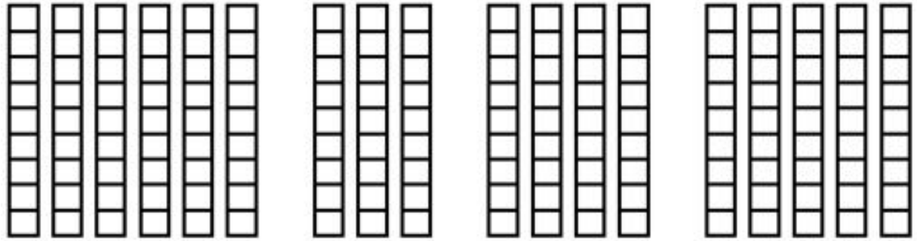
speech waveform



speech database



spectral parameters



Analogy: computer generated images

credit for the following 4 images: Speech Graphics

raw measurement data
from human subject

parametric model

model + shading

model + rendering

Hybrid speech synthesis

- Partial synthesis
- Case study: Trajectory Tiling

IEEE Trans. Audio, Speech, and Language Proc. 21 (2), pp. 280-290, 2013. DOI:10.1109/TASL.2012.2221460

A Unified Trajectory Tiling Approach to High Quality Speech Rendering

Yao Qian, *Senior Member, IEEE*, Frank K. Soong, *Fellow, IEEE*, and Zhi-Jie Yan, *Member, IEEE*

***Abstract*—It is technically challenging to make a machine talk as naturally as a human so as to facilitate "frictionless" interactions between machine and human. We propose a trajectory tiling-based approach to high-quality speech rendering, where speech parameter trajectories, extracted from natural, processed, or synthesized speech, are used to guide the search for the best sequence of waveform "tiles" stored in a pre-recorded speech database. We test the proposed unified algorithm in both Text-To-Speech (TTS) syn-**

smooth and highly intelligible synthesized speech, it has still been perceived as a voice with some traditional vocoder flavor [10]. On the other hand, the waveform concatenation-based unit selection TTS can yield fairly natural sounding speech but occasionally it may still produce some undesirable concatenation glitches. The hybrid approaches, which use HMM to guide the unit selection process to minimize the spectral, pitch

Trajectory tiling

- Core idea
  - **generate** speech parameters using a statistical model
    - spectral envelope
    - F0
    - energy (gain)
  - find a sequence of waveform fragments that **matches** these parameters
  - **concatenate** that sequence

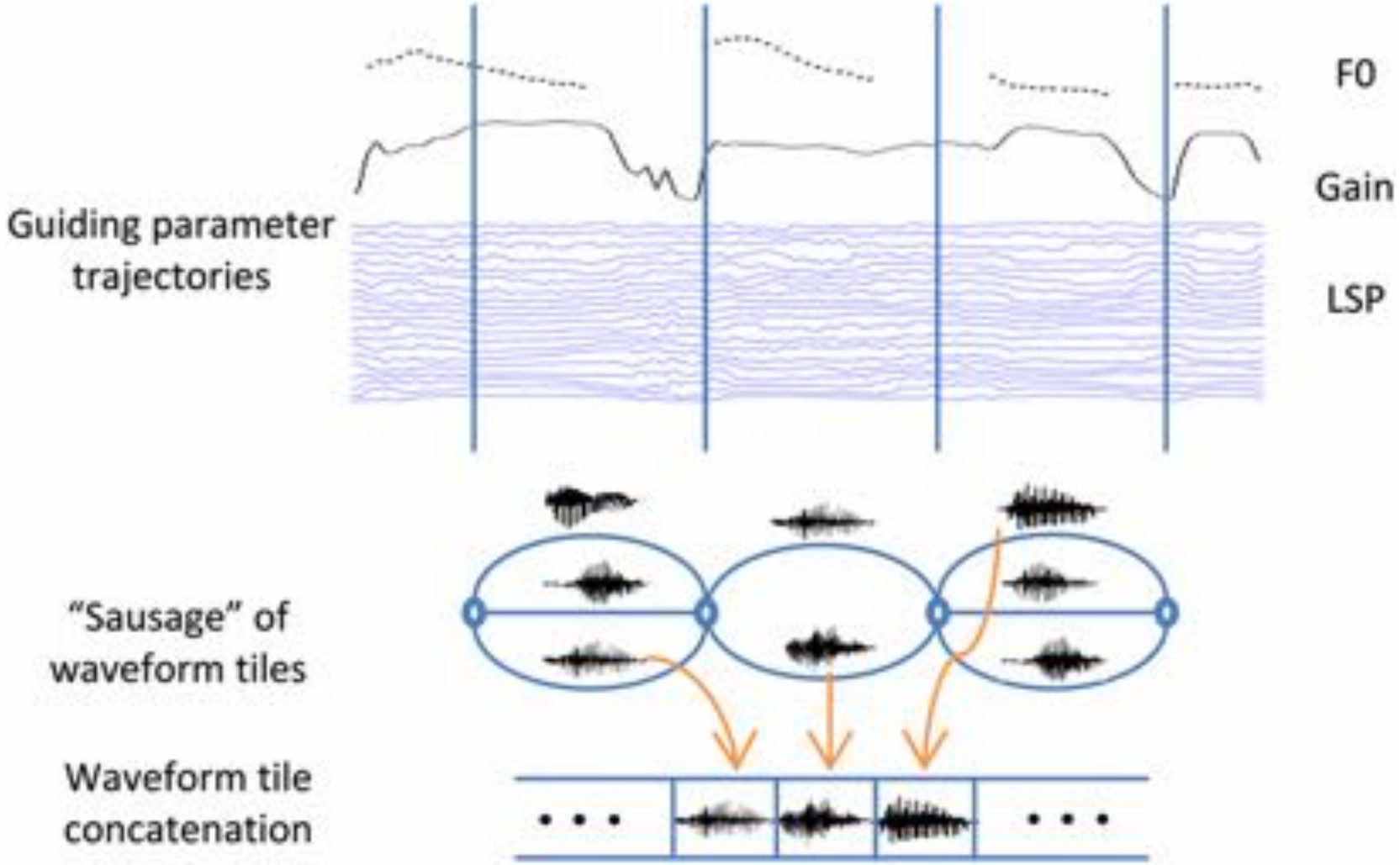


Figure 1 from Y. Qian, F. K. Soong and Z. J. Yan "A Unified Trajectory Tiling Approach to High Quality Speech Rendering" *IEEE Trans. Audio, Speech, and Language Proc.* 21 (2), pp. 280-290, 2013. DOI:10.1109/TASL.2012.2221460

Measuring the distance between waveform fragments and the trajectories from the HMM

- Extract from the waveforms
  - spectral envelope
  - energy
  - F0
- **target cost** = distance between the above features, summed over all frames of a unit
- **join cost** = ?

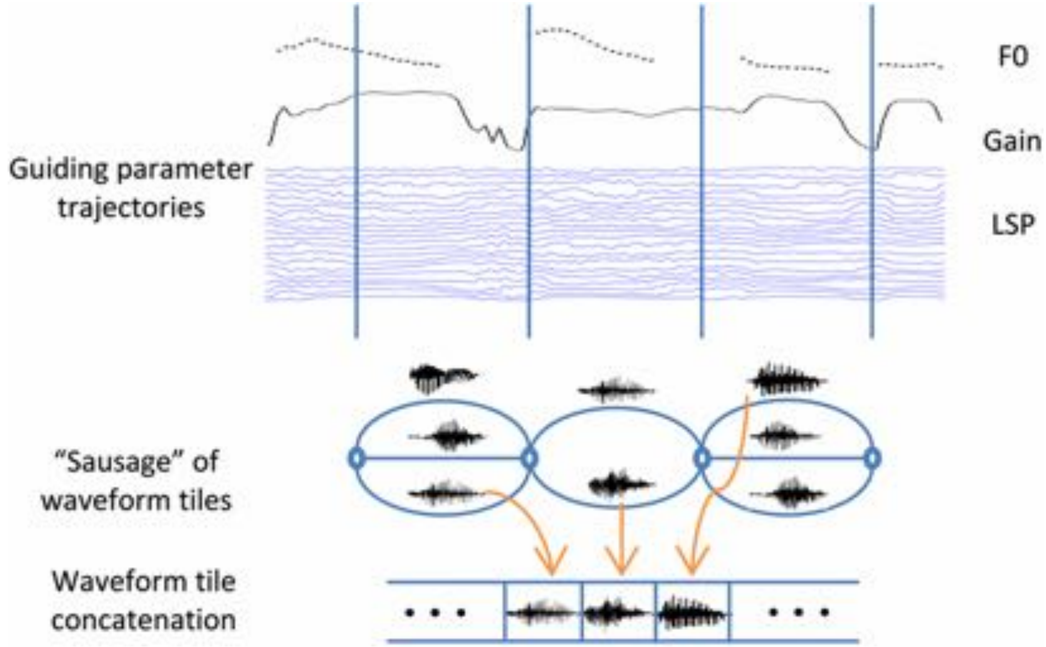
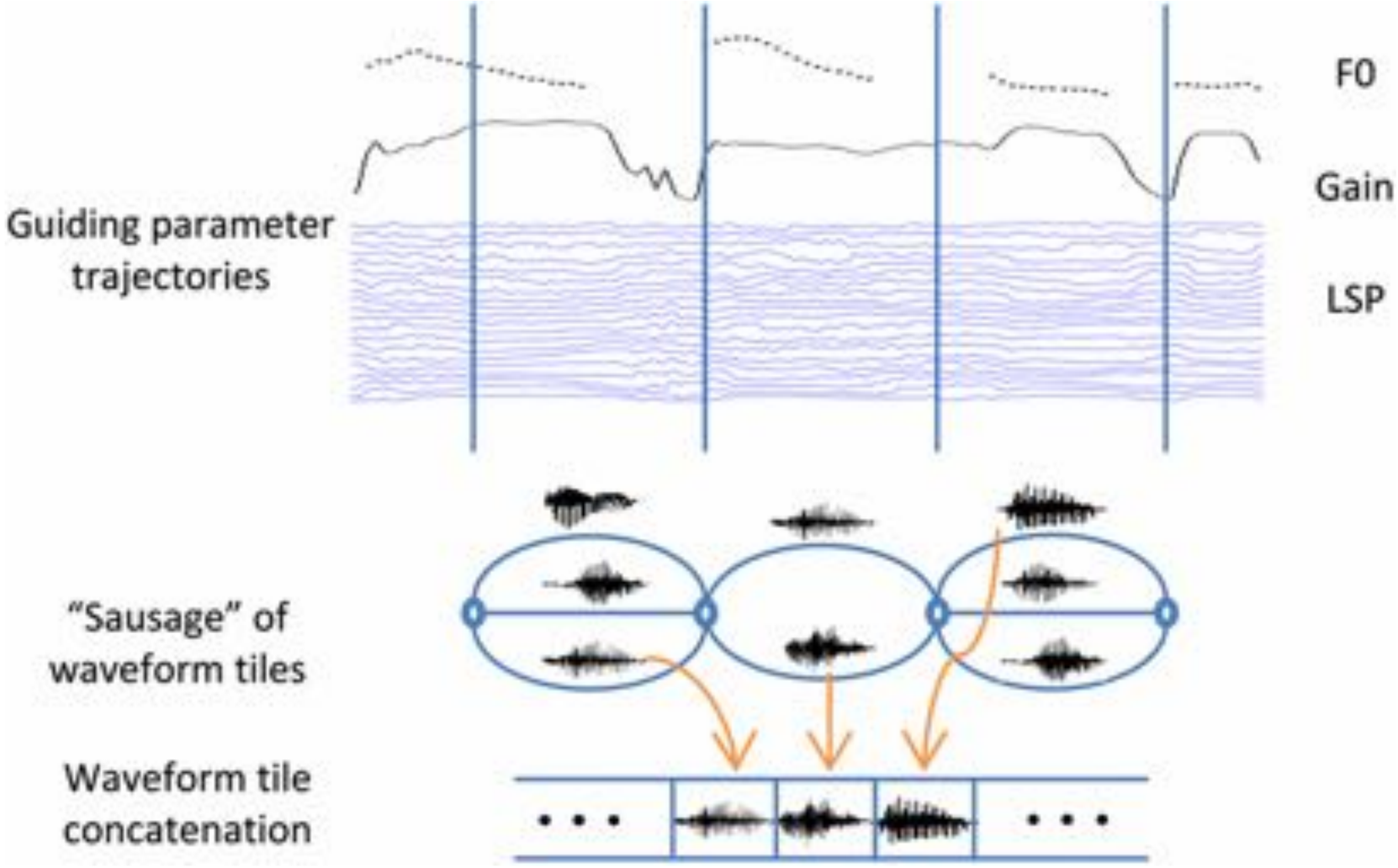


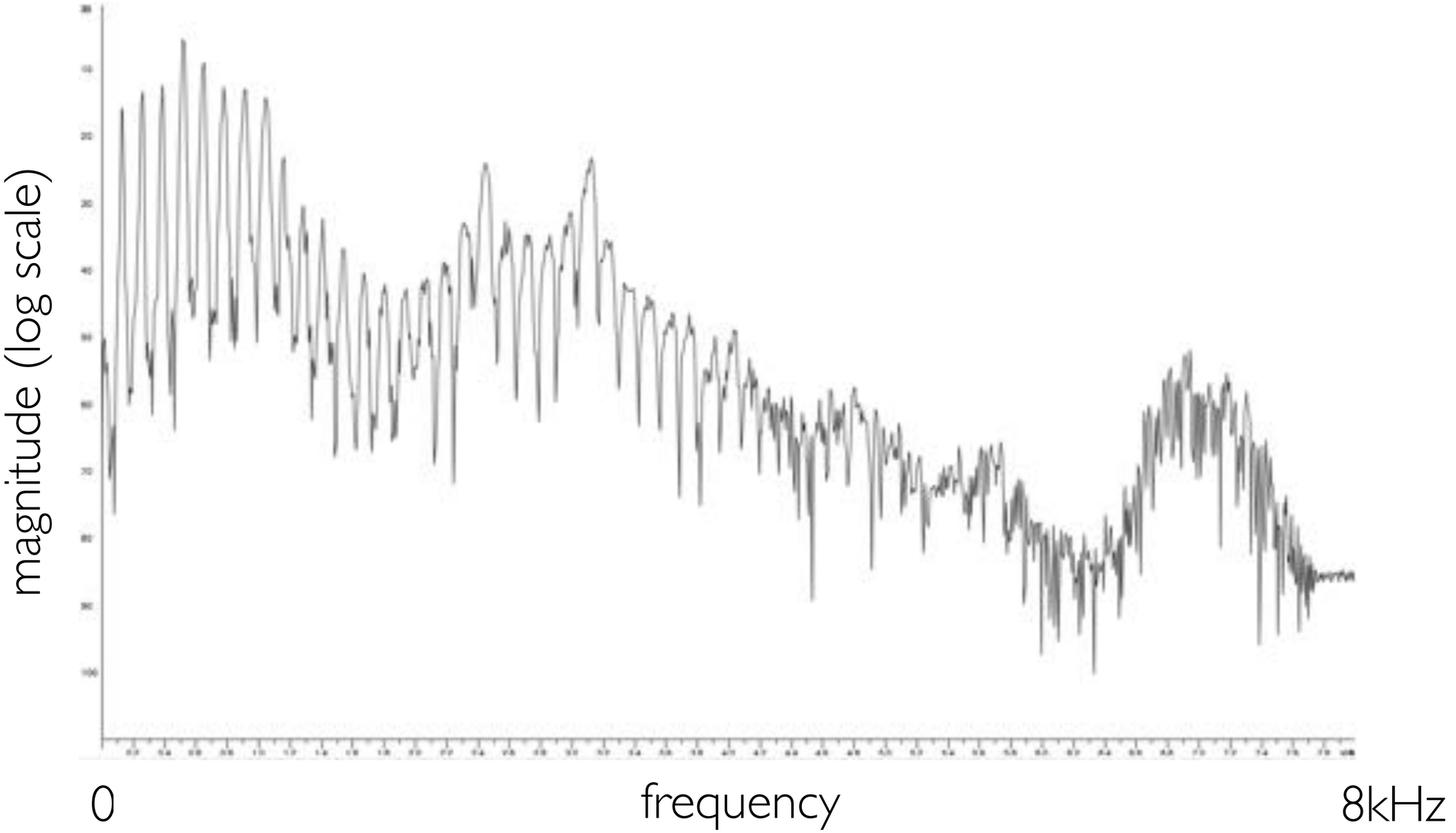
Figure 1 from Y. Qian, F. K. Soong and Z. J. Yan "A Unified Trajectory Tiling Approach to High Quality Speech Rendering" *IEEE Trans. Audio, Speech, and Language Proc.* 21 (2), pp. 280-290, 2013. DOI:10.1109/TASL.2012.2221460

Measuring the distance between waveform fragments and the trajectories from the HMM



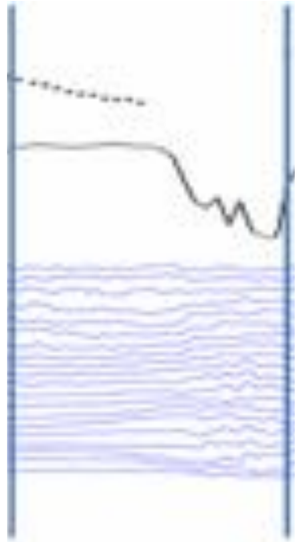
What are Line Spectral Pairs (LSPs) ?

Sometimes called Line Spectral Frequencies (LSFs)

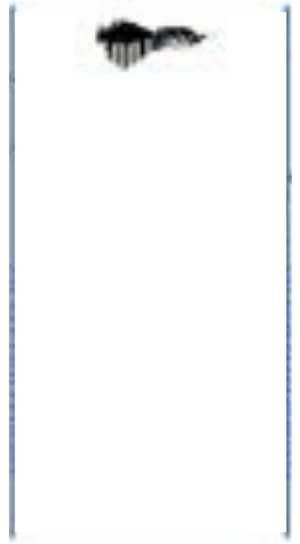


Measuring the distance between waveform fragments and the trajectories from the HMM

guiding parameter trajectories (from HMM)

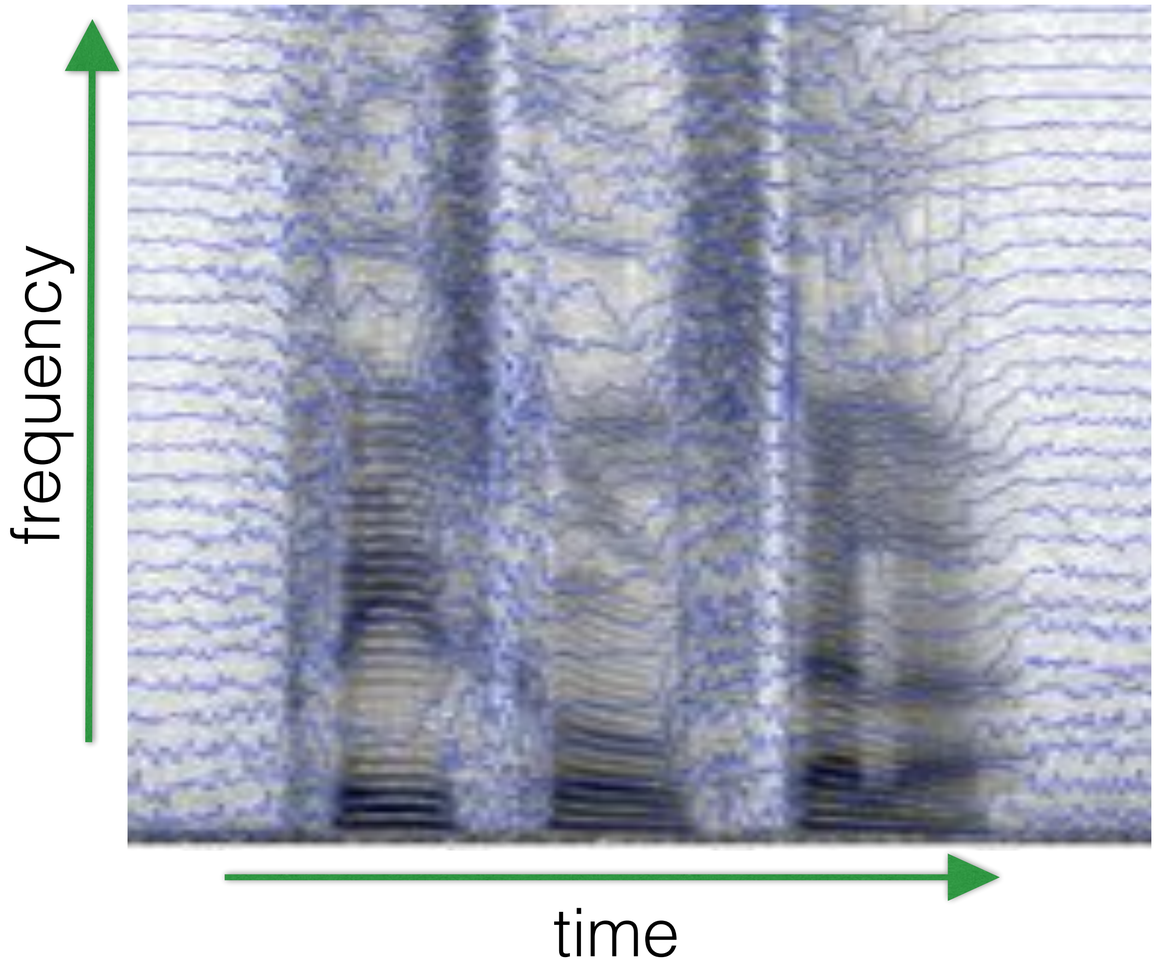


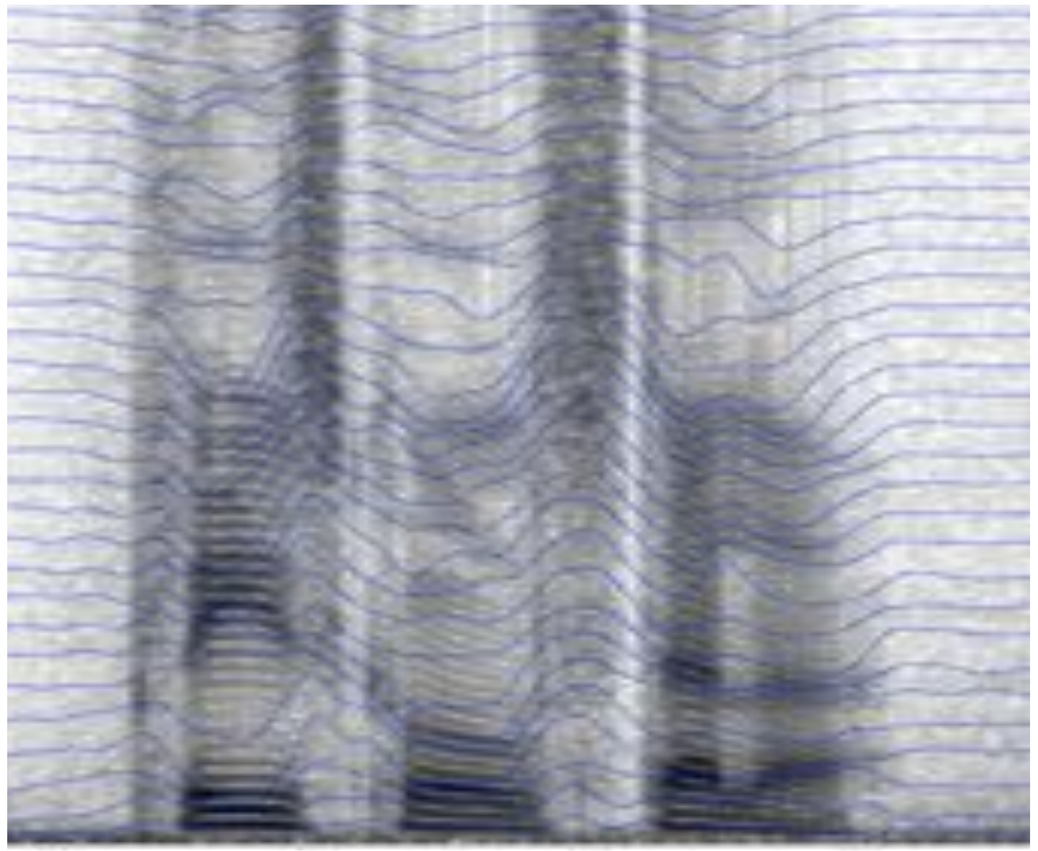
waveform

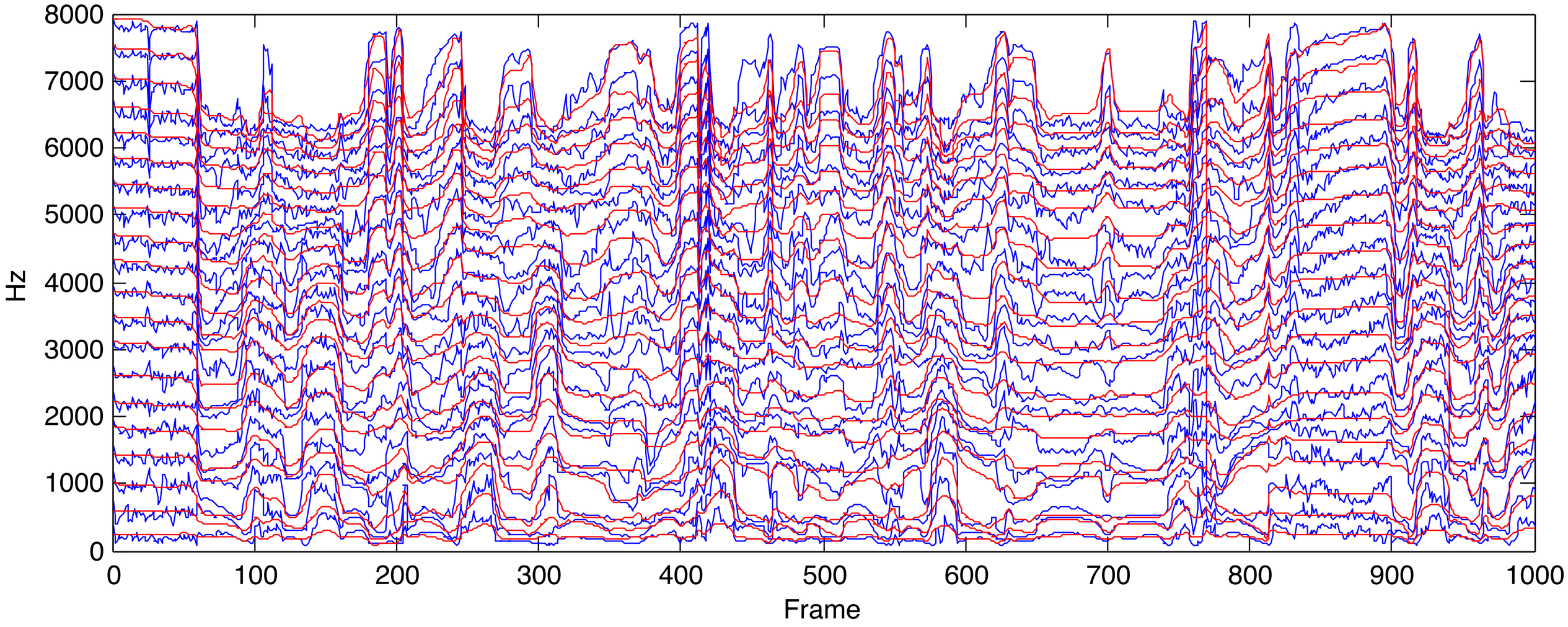


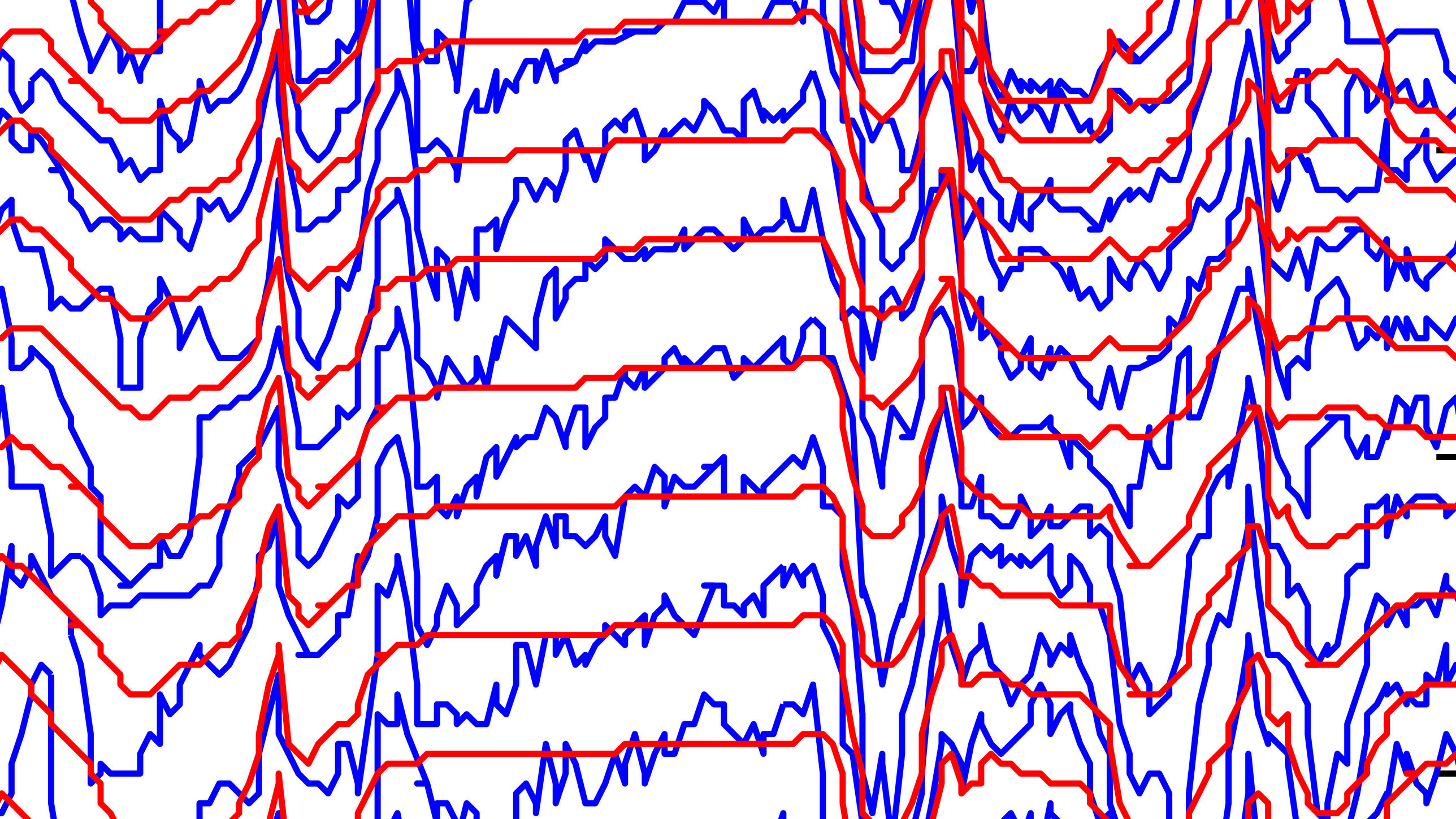
parameters extracted from the waveform

LSPs extracted from waveform vs. generated by HMM *notice the mismatch!*






8000
7000
6000
5000
4000
3000
2000
1000
0
Hz
0
100
200
300
400
500
600
700
800
900
1000
Frame



Reduce mismatch between natural parameter trajectories and those generated by HMMs

- instead of **extracting** these features from the waveforms
  - line spectral pairs (LSPs)
  - gain (of the LPC filter)
  - F0

- **regenerate** them using HMMs
  - train models
  - synthesise speech parameter trajectories **for the training data** from the models

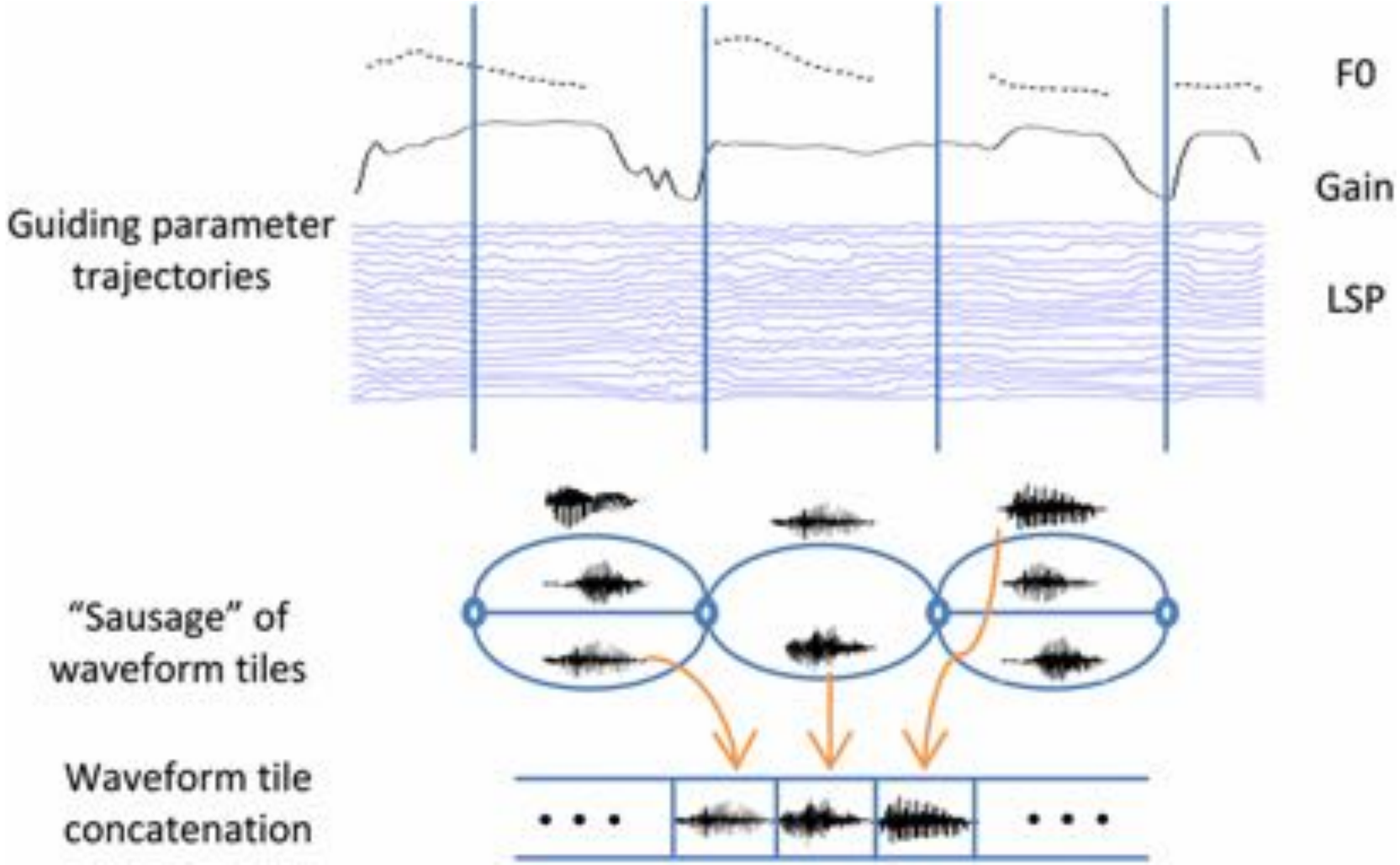


Figure 1 from Y. Qian, F. K. Soong and Z. J. Yan "A Unified Trajectory Tiling Approach to High Quality Speech Rendering" *IEEE Trans. Audio, Speech, and Language Proc.* 21 (2), pp. 280-290, 2013. DOI:10.1109/TASL.2012.2221460

Join cost: Normalised Cross Correlation

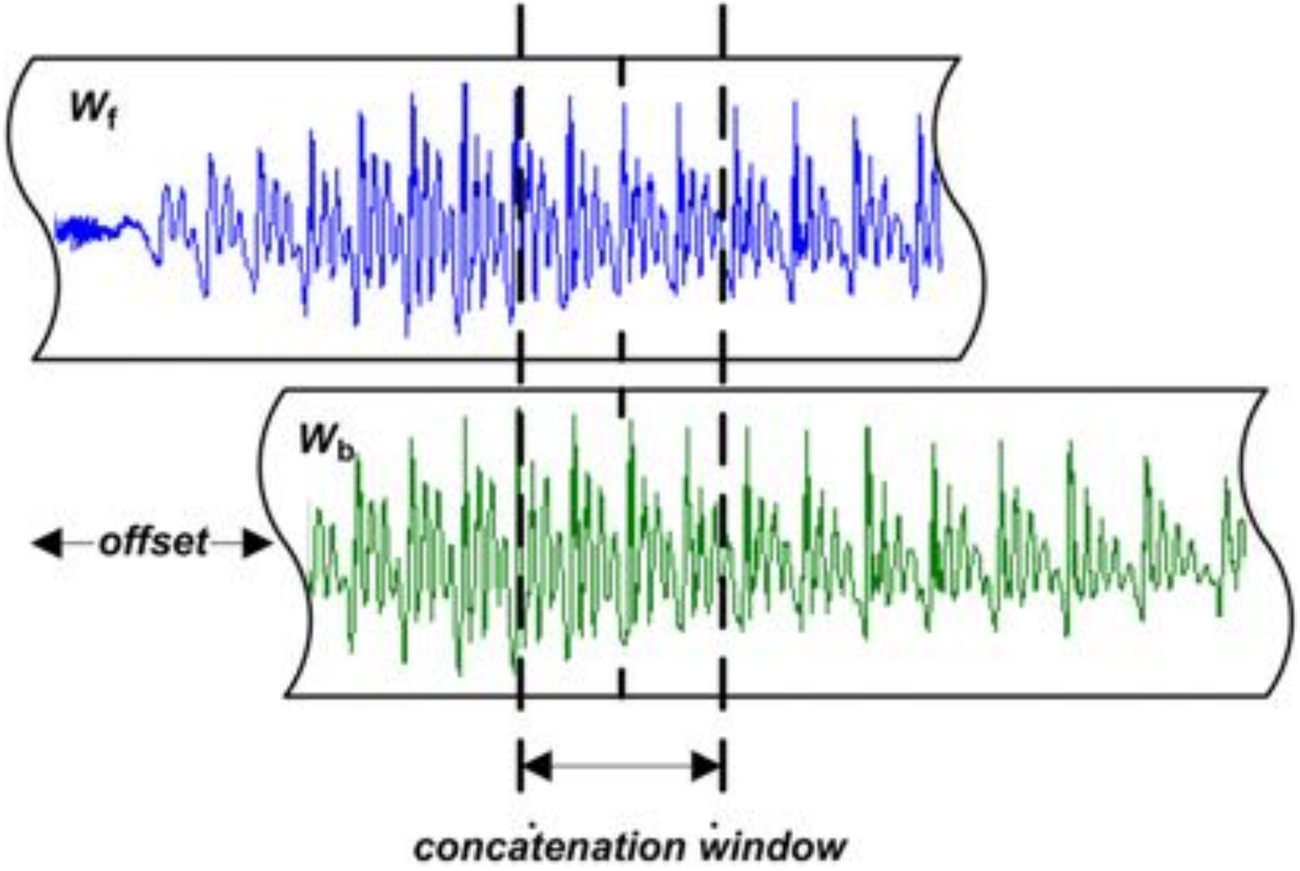


Figure 4 from Y. Qian, F. K. Soong and Z. J. Yan “A Unified Trajectory Tiling Approach to High Quality Speech Rendering” *IEEE Trans. Audio, Speech, and Language Proc.* 21 (2), pp. 280-290, 2013. DOI:10.1109/TASL.2012.2221460

Training the ‘guide’ HMM system

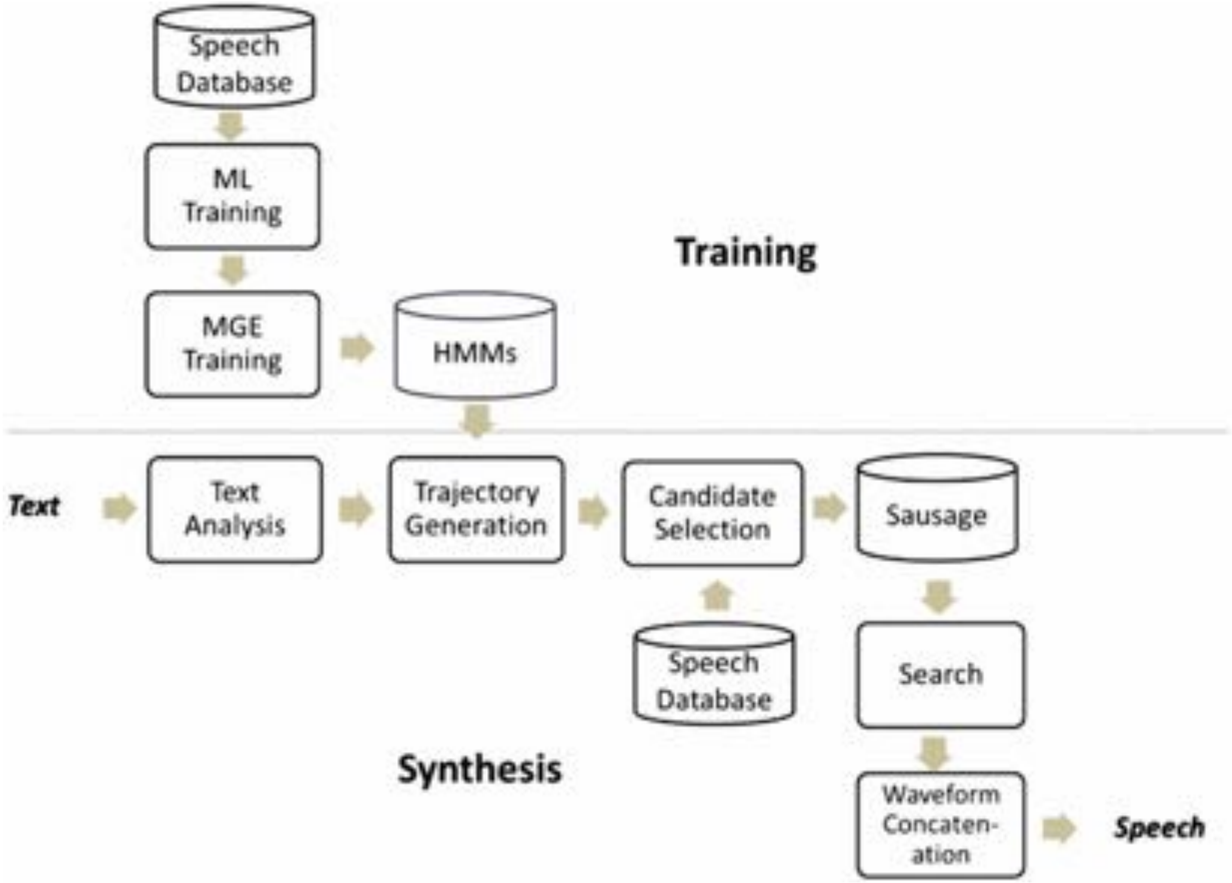


Figure 2 from Y. Qian, F. K. Soong and Z. J. Yan “A Unified Trajectory Tiling Approach to High Quality Speech Rendering” *IEEE Trans. Audio, Speech, and Language Proc.* 21 (2), pp. 280-290, 2013. DOI:10.1109/TASL.2012.2221460

Trajectory tiling

- Core idea
  - **generate** speech parameters using a statistical model
    - spectral envelope
    - F0
    - energy (gain)
  - find a sequence of waveform fragments that **matches** these parameters
  - **concatenate** that sequence

- Additional details
  - use **LSFs** for spectral envelope
  - to calculate the target cost, represent waveform fragments with parameters **generated** by HMMs (trained on the same data)
  - use a **join cost** that both
    - measures mismatch
    - finds good concatenation points