2.5. Methodology

Figure 2.2: Example of applying temporal smoothing to LSF parameterisation using a sliding Hanning window.

The width of the window was varied, to impose varying amounts of smoothing. Figure 2.2 shows an example of this process.

2.5.1.2 Variance scaling

Variance adjustment was implemented as a simple scaling of the standard deviation by a fixed factor. For each parameter (i.e., each LSF) in turn, the mean value over the utterance was found and subtracted before multiplying the parameter by a scalar value, and finally adding the mean back in. By altering the scalar value, the standard deviation is correspondingly adjusted, to simulate both reduced variance (which is commonly observed in HMM synthesis) and increased variance (e.g., as may happen if a Gaussian p.d.f. is poorly estimated during training, or when GV fails to re-instate the appropriate amount of variance). This approach of variance scaling is similar to the postfiltering method investigated by Silén and Helander (2012). Figure 2.3 shows an example of this process.
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

- **speech waveform**
- **speech parameters**
- **models**
  - model 1
  - model 2
  - model 3
  - model 4
Hybrid speech synthesis, as unit selection with an ASF target cost function

speech waveform

speech database

spectral synthesis
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
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) synthesis applications. Experimental results demonstrate that effective voice transformation algorithms must adapt the trajectory of the synthesized speech intelligibly and with a high degree of naturalness. For a given language, the waveshape units are selected based on the statistical distribution of the speech database. The trajectory tiling approach improves the quality of 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, and duration mismatch and concatenation distortion, tend to preserve the advantages of both approaches [7]. A probabilistic approach 

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

2.5. Methodology

Figure 2.2: Example of applying temporal smoothing to LSF parameter generation using a sliding Hanning window.

Variance scaling was implemented as a simple scaling of the standard deviation by a fixed factor. For each parameter (i.e., each LSF) in turn, the mean value over the utterance was found and subtracted before multiplying the parameter by a scalar value, and finally adding the mean back in. By altering the scalar value, the standard deviation is correspondingly adjusted, to simulate both reduced variance (which is commonly observed in HMM synthesis) and increased variance (e.g., as may happen if a Gaussian p.d.f. is poorly estimated during training, or when GV fails to re-instate the appropriate amount of variance). This approach of variance scaling is similar to the postfiltering method investigated by Sillén and Helander (2012). Figure 2.3 shows an example of this process.
Figure 5: MLPG generated LSP by system F (phones, state units, DNN) in red compared with natural LSPs in blue.

References


Figure 5: MLPG generated LSP by system F (phones, state units, DNN) in red compared with natural LSPs in blue.

References


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
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
What next?

• **The state of the art**

• **No videos** on this
  • because it changes too quickly

• You need to read the **primary literature** yourself
  • i.e., journal and conference papers
  • not textbooks