Hybrid Speech Synthesis

Simon King

Centre for Speech Technology Research

University of Edinburgh
What are you going to learn?

• Another recap of unit selection
  • let’s properly understand the “Acoustic Space Formulation” of the target cost
• Comparing IFF and ASF target cost functions
  • the case of prosody prediction
• Core idea of hybrid speech synthesis
• Case study: Microsoft’s ‘trajectory tiling’ method
Hybrid Speech Synthesis

Recap of unit selection (yes, again!)
“Simon”
“Simon”
Possible formulations of the target cost

• The ‘distance’ between a candidate unit and the ideal (i.e., target) unit is measured by the target function

• Taylor describes two possible formulations of the target function
  • independent feature formulation (IFF) - this is what Festival’s Multisyn engine uses (well, mostly)
  • acoustic-space formulation (ASF) - this is hybrid speech synthesis
The acoustic-space target-function formulation (ASF)

- To use an ASF target cost, we need to do “partial synthesis”
  - i.e., we need to predict some acoustic properties
    - which properties?
    - how do we predict them?
    - how exactly do we then use them in an ASF target cost?

- Predicting acoustic properties
  - classification and regression trees, as we saw in Speech Processing
  - or any other predictive model you care to use
What acoustic properties to predict?

- We have choices:
  - a few simple acoustic properties such as F0 and duration
    - would probably combine with aspects of an IFF target cost function
  - a more detailed specification such as spectral shape (e.g., represented as cepstral coefficients)
    - possibly a full set of vocoder features (as per HMM or DNN synthesis)
The acoustic-space target-function formulation (ASF)

- Visualising the acoustic space (Taylor, figure 16.6)

![Diagram of acoustic space](http://dx.doi.org/10.1017/CBO9780511816338.018)

© Copyright Simon King, University of Edinburgh, 2016. Personal use only. Not for re-use or redistribution.
Hybrid IFF + ASF target cost

- Real systems often actually use a hybrid IFF + ASF target cost function
  - It’s easy enough in principle to combine them: some sub-costs use linguistic features, others use acoustic features

- Why?
  - Partial synthesis is a way to escape some of the sparsity problems of linguistic features: many different feature combinations lead to the same acoustic property value (e.g., F0)
  - But our small set of acoustic properties (F0, duration, ..?) doesn’t capture all possible acoustic variation
    - E.g., voice quality, such as phrase-final creaky voice
Hybrid Speech Synthesis

Understanding the difference between IFF and ASF
- the case of prosody prediction
Prosody generation in unit selection: IFF approach

• the key question is: **what linguistic features** should the target cost compare?

• well - they can be anything we can reliably predict from the text

• should that include **ToBI accents & boundary** tones, for example?
  • how would we predict these?
    • choose your classifier: .................................................................
    • list available predictors: ..............................................................
    • obtain training data: .................................................................
  • how accurate would those predictions be?
Prosody generation in unit selection: ASF approach

- **how to predict** the acoustic features for the target?
  - assume we will use ToBI as the symbolic representation of prosody
  - step 1: predict ToBI symbols from text
    - a classification task, as in the IFF approach
  - step 2: render ToBI symbols as an F0 contour
    - a regression task - will need training on data

- **how to compare** the acoustic features between target and candidate?
  - Euclidean distance between F0 contours?
  - is that perceptually relevant?
Hybrid Speech Synthesis

The core idea
Hybrid approaches

- **HMM or DNN synthesis**
  - flexible, somewhat robust to labelling errors
  - but limited in naturalness by the vocoder (amongst other things)
- **Unit selection**
  - potentially excellent naturalness (due to *waveform* concatenation)
  - but IFF target cost is hand-crafted; join cost rather naive
    - fragile - e.g., easily affected by labelling errors
    - hard to optimise for each new speech database
- **Hybrid synthesis**
  - robustness and learning-from-data
  - waveform concatenation
Hybrid speech synthesis
Hybrid speech synthesis

speech waveform

unit inventory

speech parameters
Various forms of hybrid synthesis

- **Trajectory tiling** (Microsoft Research)
  - generate speech parameters from HMM
  - select closest matching waveform units
    - can formulate this probabilistically
  - effectively, HMMs are the target cost
  - perform unit selection search procedure
- **concatenate waveforms**
- **Multiform synthesis** (Nuance, used in main product)
  - concatenate an **alternating** sequence of
    - waveform units
    - speech generated from HMMs + vocoder
  - perceptual considerations: use HMMs when listener will not hear the difference
Hybrid Speech Synthesis

Hybrid speech synthesis: the “trajectory tiling” approach

This content is based on the paper:


and the following slides contain some figures taken from that paper.
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
© Copyright Simon King, University of Edinburgh, 2016. Personal use only. Not for re-use or redistribution.
Measuring the distance between waveform fragments and the trajectories from the HMM

• How might we do this?
  • extract from the waveforms
    • spectral envelope
    • energy
    • F0
  • **target cost** = Euclidean distance (between the above features, summed over all frames of a unit)
  • **join cost** = Euclidean distance between the above features across a concatenation point

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
© Copyright Simon King, University of Edinburgh, 2016. Personal use only. Not for re-use or redistribution.
Measuring the distance between waveform fragments and the trajectories from the HMM

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

© Copyright Simon King, University of Edinburgh, 2016. Personal use only. Not for re-use or redistribution.
Using linear prediction features (source-filter model)

- extract from the waveforms
  - line spectral pairs (**LSPs**)
  - gain (of the LPC filter)
  - F0
- **target cost** = Euclidean distance (between the above features, summed over all frames of a unit)
Mismatch between natural parameter trajectories and those generated by HMMs

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

© Copyright Simon King, University of Edinburgh, 2016. Personal use only. Not for re-use or redistribution.
LSPs: extracted from waveform vs. generated by HMM
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

- **generate them using HMMs**
  - train models on the full database of waveforms (training data)
  - synthesise parameter trajectories for this training data from these 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

© Copyright Simon King, University of Edinburgh, 2016. Personal use only. Not for re-use or redistribution.
Figure 1 from Y. Qian, F. K. Soong and Z. J. Yan “A Unified Trajectory Tiling Approach to High Quality Speech Rendering”  

© Copyright Simon King, University of Edinburgh, 2016. Personal use only. Not for re-use or redistribution.
What is NCC (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

© Copyright Simon King, University of Edinburgh, 2016. Personal use only. Not for re-use or redistribution.
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

© Copyright Simon King, University of Edinburgh, 2016. Personal use only. Not for re-use or redistribution.
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
  • for the purposes of distance calculation, replace waveform fragments with parameters generated by HMMS (trained on that same data)
  • use a join cost that both
    • measures mismatch
    • finds good concatenation points
Figure 7 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

© Copyright Simon King, University of Edinburgh, 2016. Personal use only. Not for re-use or redistribution.