Natural Speech Technology
Edinburgh – Cambridge – Sheffield

Measuring the perceptual effects of speech synthesis modelling assumptions

Gustav Eje Henter Thomas Merritt Matt Shannon Catherine Mayo Simon King
University of Edinburgh, U.K. and University of Cambridge, U.K.

Outline

- recorded a corpus where the same text is read aloud multiple times
- created various forms of chimeric speech where different aspects of the generated speech come from different repetitions
  - e.g. cepstra from one repetition, F_0 from another repetition
- evaluated naturalness
- provides a way to investigate the perceptual effect of various high-level modelling assumptions which are common in SPSS

Assumptions in speech synthesis

- naturalness of an SPSS system depends on
  - speech parameter representation (vocoder, etc.)
  - probabilistic model
  - speech parameter generation method
- probabilistic model makes many assumptions:
  - high-level assumptions
    - e.g. source and filter parameters are conditionally independent
    - e.g. different cepstral trajectories are conditionally independent
  - low-level assumptions
    - e.g. for each decision tree leaf a given quantity is Gaussian distributed
- questions in model design:
  - how restrictive are particular high-level assumptions?
  - which ones should we try to remove to improve naturalness?
- hard to investigate without worrying about low-level assumptions

Key insight

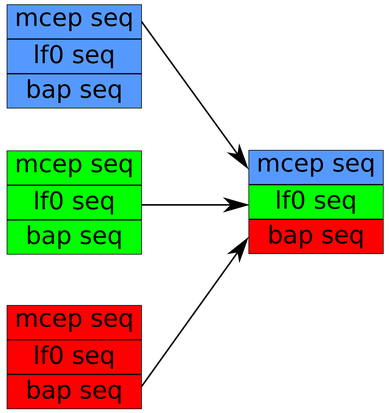
- by manipulating repeated natural speech, it is possible to simulate parameter generation from a model which makes no low-level assumptions
- this allows investigation of the fundamental limit that would be reached by perfecting the low-level part of the probabilistic model
- may help to inform the design of new probabilistic models

REHASP 0.5 corpus

- female British English speaker
- 30 Harvard sentence prompts
- 40 repetitions
- care taken with ordering to prevent list effects
- recorded at 96 kHz, 16 bit
- available under a permissive license

Combining repetitions

- align all repetitions of the same prompt using dynamic time warping
- form a chimeric combination, for example:
  - cepstral sequence from one repetition
  - log F_0 sequence from a different repetition
  - band aperiodicity sequence from a third repetition

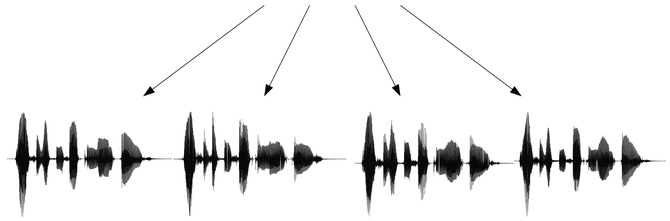


- can also combine different repetitions by taking the mean

Interpretation

- a repetition ≈ a sample from a "perfect" probabilistic model

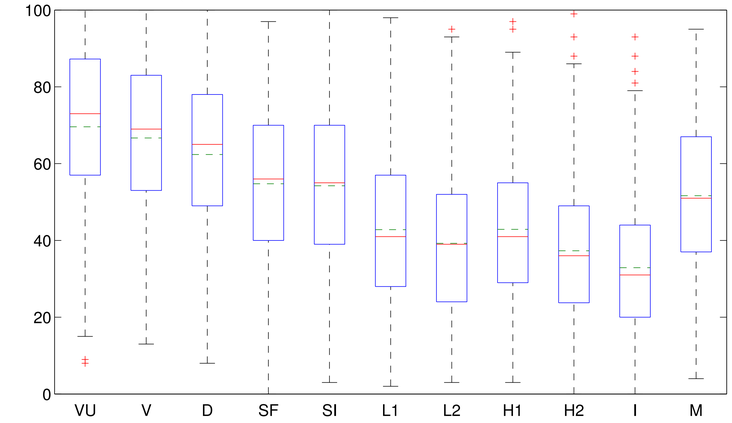
"Rice is often served in round bowls"



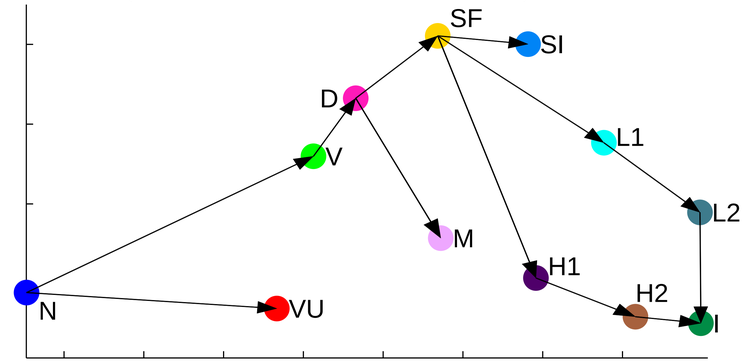
- a chimeric combination ≈ a sample from a probabilistic model that makes given high-level assumptions but no low-level assumptions
- a mean combination ≈ the mean of a probabilistic model
- allows us to hear what speech would sound like in the limit of improving the low-level part of the probabilistic model
  - given a speech parameter representation
  - given a speech parameter generation method
    - mean-based parameter generation
    - sampling parameter generation
  - given particular high-level modelling assumptions

Results

- MUSHRA test (perceived naturalness):



- MDS plot (visualisation of perceptual distance):



- key:
  - baseline conditions:
    - N: natural waveform
    - VU: natural speech parameters (no smoothing)
    - V: natural speech parameters (slightly smoothed)
  - conditions simulating sampling parameter generation:
    - D: no high-level assumptions
    - SF: filter (mcep) and source (lf0, bap) parameter sequences conditionally independent
    - SI: mcep, lf0, bap sequences conditionally independent
    - L1 and L2: lowest mcep trajectories conditionally independent of each other
    - H1 and H2: highest mcep trajectories conditionally independent of each other
    - I: all mcep trajectories conditionally independent
  - conditions simulating mean-based parameter generation:
    - M: any of the above high-level assumptions

Conclusions

- SF is quite restrictive; I is very restrictive
- mean-based generation is better than sampling when using a poor model, but worse than sampling when using a good model