Measuring the Perceptual Effects of Speech Synthesis Modelling Assumptions

Natural Speech Technology

Edinburgh – Cambridge – Sheffield

Gustav Eje Henter,
Thomas Merritt,
Matt Shannon,
Catherine Mayo,
Simon King

Summary

"Hear the perceptual effects of modelling assumptions in statistical speech synthesis"

Summary

"Hear the perceptual effects of modelling assumptions in statistical speech synthesis"

1. Through manipulating repeated natural speech

Summary

"Hear the perceptual effects of modelling assumptions in statistical speech synthesis"

1. Through manipulating repeated natural speech
2. Identify which assumptions that limit synthesiser naturalness

Overview

1. Background
2. Methodology
3. Experiments
4. Conclusions and outlook

Naturalness in speech synthesis

Natural Speech Technology

Output naturalness depends on many factors:

- Text processing
- Speech parameter representation (vocoder etc.)
- Probabilistic models
- Parameter generation method

Naturalness in speech synthesis

Natural Speech Technology

Output naturalness depends on many factors:

- Text processing
- Speech parameter representation (vocoder etc.)
- Probabilistic models
- Parameter generation method

Modelling assumptions

Acoustic models make many assumptions:

- High-level assumptions
  - Different parameter streams are conditionally independent
  - Filter parameter trajectories are conditionally independent

Modelling assumptions

Acoustic models make many assumptions:

- High-level assumptions
  - Different parameter streams are conditionally independent
  - Filter parameter trajectories are conditionally independent
- Low-level assumptions
  - A particular decision tree partitioning of linguistic contexts
  - Leaf node distributions are Gaussian

Modelling assumptions

Natural Speech Technology

Acoustic models make many assumptions:

- High-level assumptions
  - Different parameter streams are conditionally independent
  - Filter parameter trajectories are conditionally independent
- Low-level assumptions
  - A particular decision tree partitioning of linguistic contexts
  - Leaf node distributions are Gaussian

Assumption adequacy affects output naturalness

Questions

1. Which high-level assumptions hurt naturalness?
2. How much may we gain if we could remove these assumptions?

Questions

1. Which high-level assumptions hurt naturalness?
2. How much may we gain if we could remove these assumptions?

$\rightarrow$ Where should we direct our improvement efforts?

Traditional fault-finding

Investigate naturalness through trial-and-error:

1. Select an assumption and modify it
2. Compare output naturalness before and after

Traditional fault-finding

Natural Speech Technology

Investigate naturalness through trial-and-error:

1. Select an assumption and modify it
2. Compare output naturalness before and after

Problems:

- Impressions are coloured by other imperfections
  - Low-level assumptions
  - Estimation errors

Traditional fault-finding

Investigate naturalness through trial-and-error:

1. Select an assumption and modify it
2. Compare output naturalness before and after

Problems:

- Impressions are coloured by other imperfections
  - Low-level assumptions
  - Estimation errors
- Does not compare the relative severity of different assumptions

Our insight

- Natural speech is a sample from the true acoustic model

Our insight

- Natural speech is a sample from the true acoustic model
- By manipulating repeated natural speech we can simulate output from
  - highly accurate models
    - only incorporating certain high-level modelling assumptions
    - no low-level assumptions at all

Our insight

- Natural speech is a sample from the true acoustic model
- By manipulating repeated natural speech we can simulate output from
    - highly accurate models
        - only incorporating certain high-level modelling assumptions
        - no low-level assumptions at all
    - with a particular parameter representation
    - and a particular output generation method

Why is this cool?

Nobody knows what these "nearly perfect" models are, yet we can listen to their output!

Why is this cool?

Nobody knows what these "nearly perfect" models are, yet we can listen to their output!

- Compare naturalness degradations due to different high-level assumptions in an otherwise perfect model
- Identified key naturalness bottlenecks in speech synthesis

Overview

1. Background
2. Methodology
3. Experiments
4. Conclusions and outlook

Repeated speech

Even when controlling for context, the same text can be realised acoustically in many different ways

Repeated speech

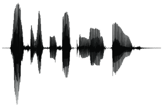
Even when controlling for context, the same text can be realised acoustically in many different ways

"Rice is often served in round bowls"

Repeated speech

Even when controlling for context, the same text can be realised acoustically in many different ways

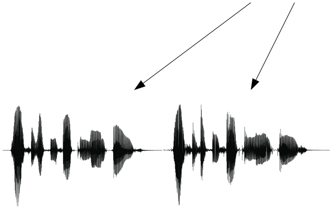
"Rice is often served in round bowls"



Repeated speech

Even when controlling for context, the same text can be realised acoustically in many different ways

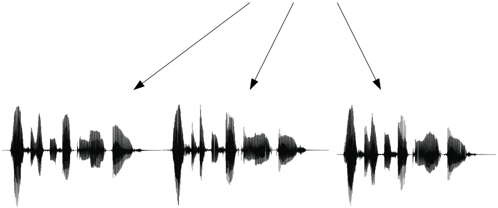
"Rice is often served in round bowls"



Repeated speech

Even when controlling for context, the same text can be realised acoustically in many different ways

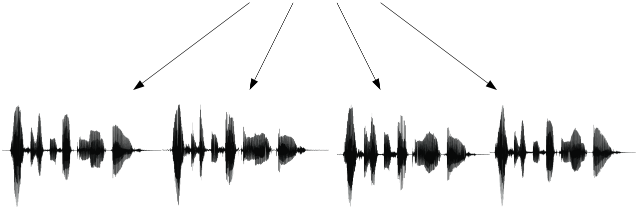
"Rice is often served in round bowls"



Repeated speech

Even when controlling for context, the same text can be realised acoustically in many different ways

"Rice is often served in round bowls"



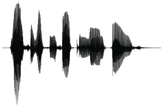
Natural
Speech
Technology

REHASP 0.5 corpus

- "REpeated HArvard Sentence Prompts"
- Female British English talker "Lucy"
- 30 Harvard sentence prompts
- Each read aloud 40 times
  - Presented in random order
- Recorded at 16 bit 96 kHz
- Publicly available under a permissive license
  - datashare.is.ed.ac.uk/handle/10283/561

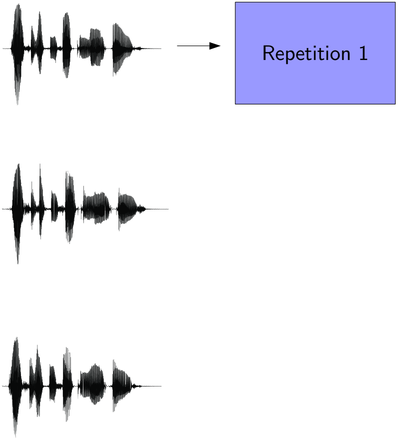
In pictures

0. Start with natural speech repetitions:



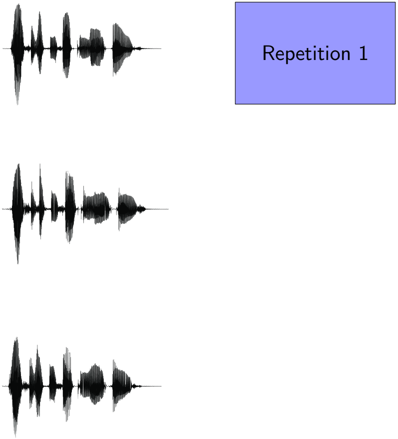
In pictures

1. Extract parameters:



In pictures

1. Extract parameters:



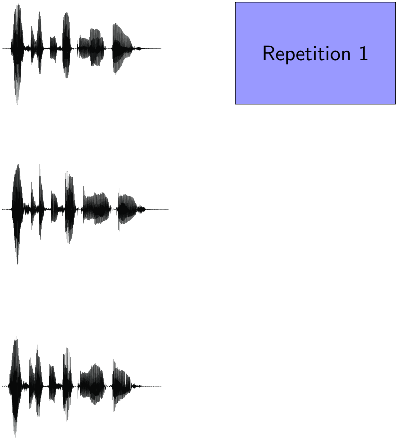
Speech representation

Standard parametric speech representation used for experiments:

- 16 kHz operating point
- Matlab STRAIGHT for parameter extraction
- 46-dimensional parameter vector with three streams:
  - 40 MCEPs (0–39), representing filter coefficients
  - Log-F0
  - 5 band aperiodicities (BAPs)
- 5 ms frame shift

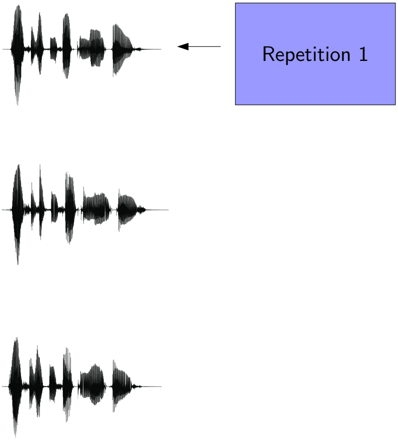
In pictures

1.b. Resynthesise (baseline “V”):



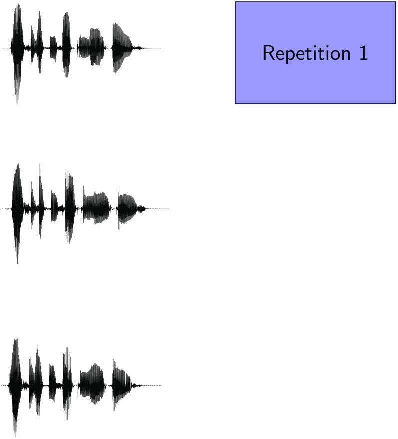
In pictures

1.b. Resynthesise (baseline "V"):



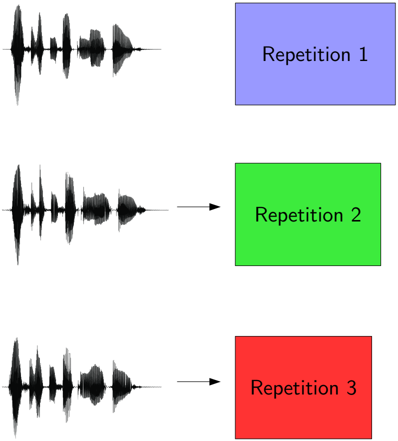
In pictures

1. Extract parameters:



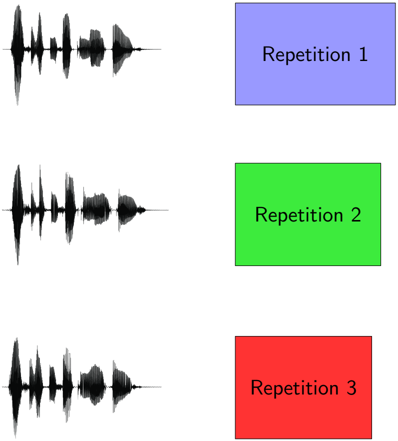
In pictures

1. Extract parameters:



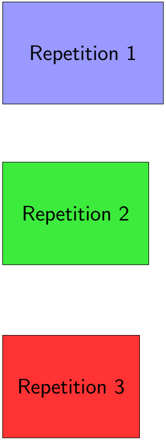
In pictures

1. Extract parameters:



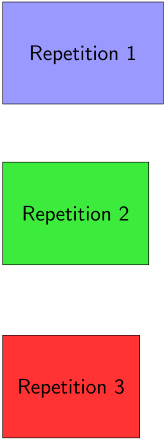
In pictures

1. Extract parameters:



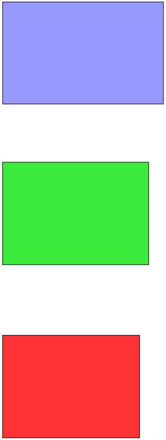
Match timings

2.a. Match frames:



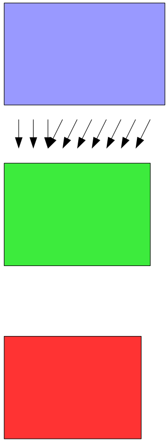
Match timings

2.a. Match frames:



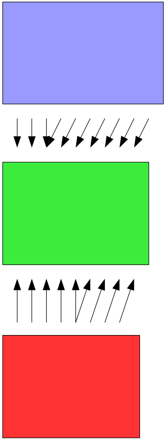
Match timings

2.a. Match frames:



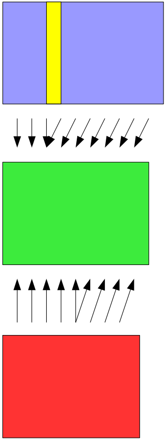
Match timings

2.a. Match frames:



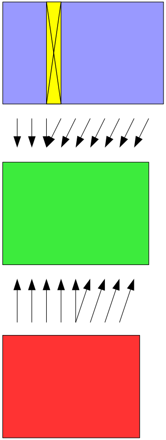
Match timings

2.b. Warp timings:



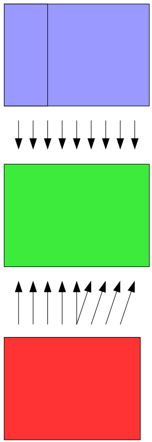
Match timings

2.b. Warp timings:



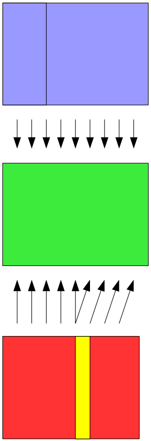
Match timings

2.b. Warp timings:



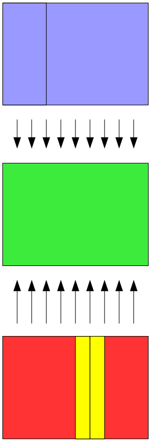
Match timings

2.b. Warp timings:



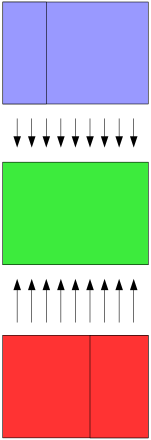
Match timings

2.b. Warp timings:



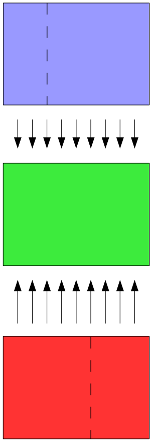
Match timings

2.b. Warp timings:



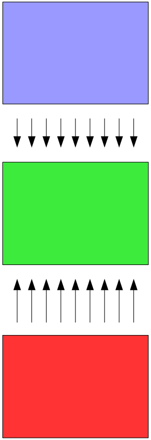
Match timings

2.b. Warp timings:



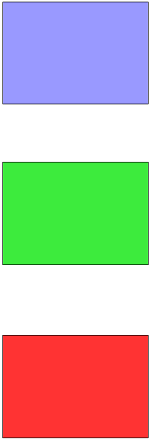
Match timings

2.b. Warp timings:



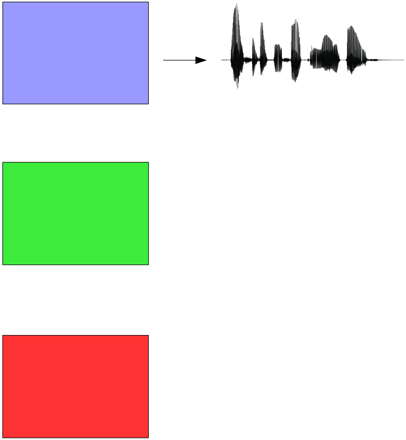
Match timings

2.b. Warp timings:



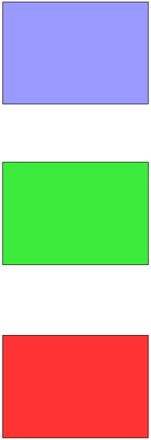
Match timings

2.c. Resynthesise (baseline “D”):



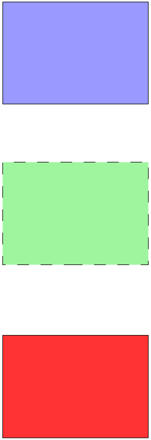
Match timings

2.d. Remove reference:



Match timings

2.d. Remove reference:



Match timings

2.d. Remove reference:

Match timings

We now have “LEGO pieces” of aligned repetitions

Create chimeric speech

3.a. Combine parameters from independent repetitions:

Create chimeric speech

3.a. Combine parameters from independent repetitions:

Create chimeric speech

3.a. Combine parameters from independent repetitions:

Create chimeric speech

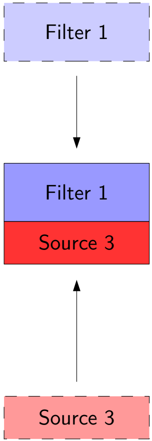
3.a. Combine parameters from independent repetitions:

Create chimeric speech

3.a. Combine parameters from independent repetitions:

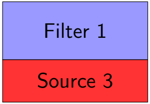
Create chimeric speech

3.a. Combine parameters from independent repetitions:



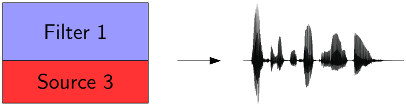
Create chimeric speech

3.a. Combine parameters from independent repetitions:



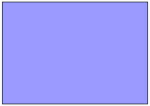
Create chimeric speech

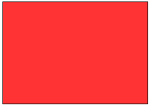
3.a. Resynthesise chimeric speech (here condition “SF”):



Create mean speech

3.b. Take the mean of all repetitions:



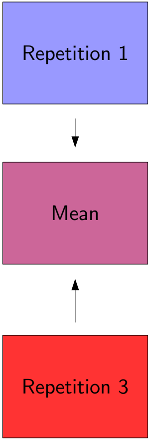


Create mean speech

3.b. Take the mean of all repetitions:

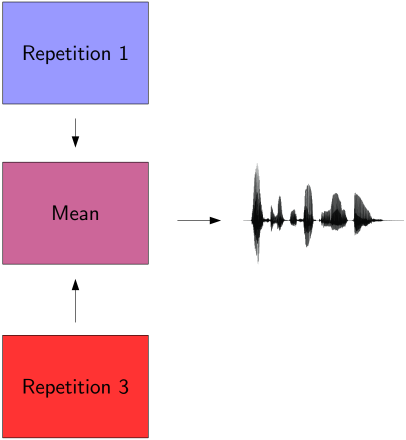
Create mean speech

3.b. Take the mean of all repetitions:



Create mean speech

3.b. Resynthesise mean speech (condition “M”):



Interpretation

- Repeated speech $\approx$ independent samples from a “perfect” acoustic model
- Chimeric speech $\approx$ samples from a model making certain high-level assumptions but no low-level assumptions
- Mean speech $\approx$ the mean of a probabilistic model

Overview

Present investigation

- Two model assumption classes:
  1. Stream independence assumptions
     - 1.1 Source and filter parameters independent
     - 1.2 Filter, pitch, aperiodicities independent
  2. Independence assumptions among filter coefficients

Present investigation

Natural Speech Technology

- Two model assumption classes:
  1. Stream independence assumptions
     1.1 Source and filter parameters independent
     1.2 Filter, pitch, aperiodicities independent
  2. Independence assumptions among filter coefficients
- Two output generation methods:
  1. Random sampling from probability distribution
  2. Mean parameter generation

Natural Speech Technology

Present investigation

- Two model assumption classes:
  1. Stream independence assumptions
     1.1 Source and filter parameters independent
     1.2 Filter, pitch, aperiodicities independent
  2. Independence assumptions among filter coefficients
- Two output generation methods:
  1. Random sampling from probability distribution
  2. Mean parameter generation

= 12 *conditions* (4 baselines)
  - For each of the 30 Harvard sentences

What it sounds like

Natural Speech Technology

Sampling-based generation:

Database examples:	3	7	26	32	
Baselines:	N	VU	V	D	
Stream independence:	SF	SI			
Filter coefficient independence:	L1	L2	H1	H2	I

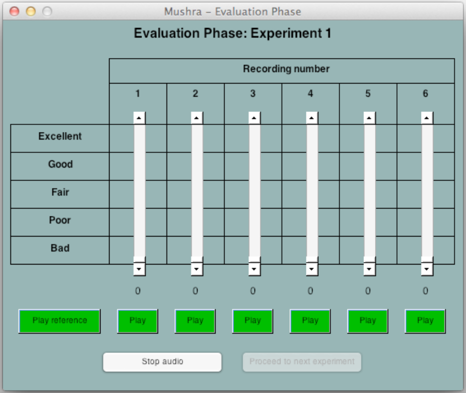
Mean-based generation:

Averaging:	M

(Also available online at homepages.inf.ed.ac.uk/ghenter)

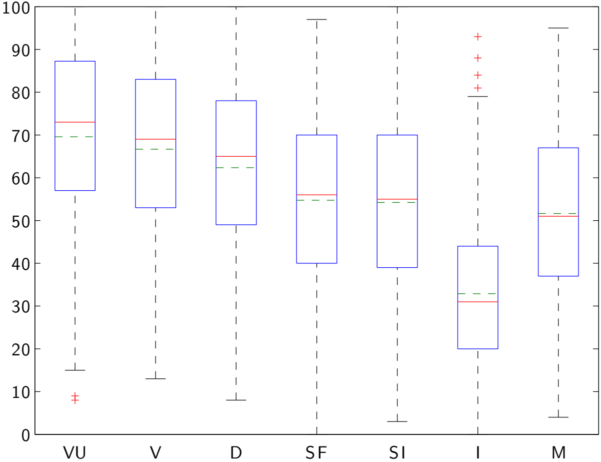
Naturalness test

MUSHRA test for parallel, fine-grained naturalness assessment



Naturalness results

Box plot of 549 comparisons rating natural speech at 100:



Overview

Conclusions

- When sampling from models:

Conclusions

- When sampling from models:
  1. Source-filter independence assumption reduces naturalness

Conclusions

- When sampling from models:
  1. Source-filter independence assumption reduces naturalness
  2. Independence assumptions among filter coefficients further reduces naturalness

Conclusions

- When sampling from models:
  1. Source-filter independence assumption reduces naturalness
  2. Independence assumptions among filter coefficients further reduces naturalness
- Using mean-based parameter generation:

Conclusions

- When sampling from models:
  1. Source-filter independence assumption reduces naturalness
  2. Independence assumptions among filter coefficients further reduces naturalness
- Using mean-based parameter generation:
  1. Better than sampling for poor models

Conclusions

- When sampling from models:
  1. Source-filter independence assumption reduces naturalness
  2. Independence assumptions among filter coefficients further reduces naturalness
- Using mean-based parameter generation:
  1. Better than sampling for poor models
  2. Less natural than sampling for accurate models

Limitations

Natural Speech Technology

Conclusions not applicable to:

- Other speech representations
- Other parameter generation methods
    - E.g., postfiltering, global variance modelling

Future work

- Record REHASP 1.0 corpus

Future work

- Record REHASP 1.0 corpus
- Expanded investigation
  - Consider additional assumptions
    - Cover the entire spectrum from natural speech to TTS system
  - Consider additional parameter generation methods

Future work

- Record REHASP 1.0 corpus
- Expanded investigation
  - Consider additional assumptions
    - Cover the entire spectrum from natural speech to TTS system
  - Consider additional parameter generation methods
- Effect of different parameter representations

The end

The end

Thank you for listening!