KTH ROYAL INSTITUTE
OF TECHNOLOGY

The speech synthesis phoneticians need is both realistic and controllable.

Zofia Malisz[1], Gustav Eje Henter[1], Cassia Valentini-Botinhao[2], Oliver Watts[2], Jonas Beskow[1], Joakim Gustafson[1]
[1]Department of Speech, Music and Hearing, KTH
[2]The University of Edinburgh, UK

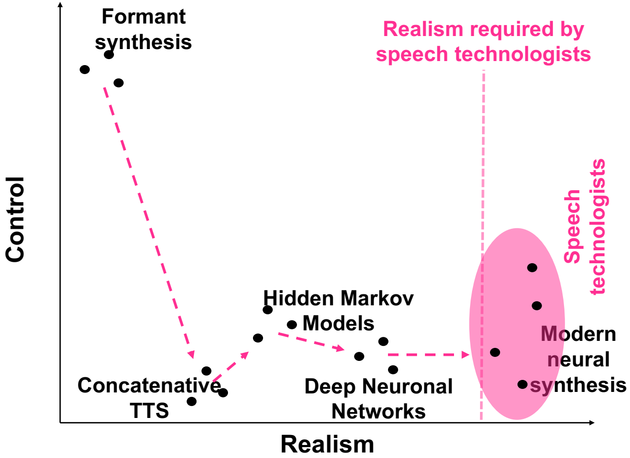
Why do speech engineers need speech sciences?

- There is no synthesis without analysis (mostly)
- More data, better algorithms, better performance - yes, but what about:
  - ... understanding your data?
  - ... modeling your data so that you can manipulate or predict particular aspects of it?
- Methodology: prevent your non-ML statistics muscle atrophy

Why does speech synthesis need speech sciences?

- Instrumental in speech processing and engineering in the formant synthesis age: sparse data, wetware modelling (King, 2015)
- Today: perception-based modelling (e.g. mel scale)
- Benchmarking TTS: advanced evaluation methods crossed over from e.g. psycholinguistics

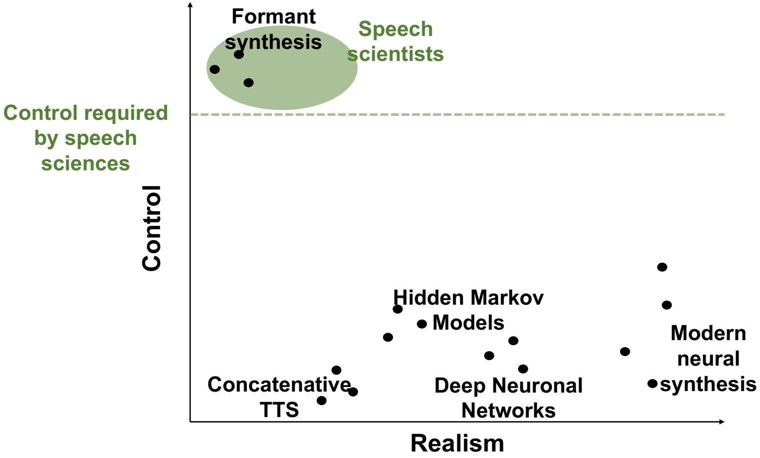
Speech technology point of view



Why do phoneticians need speech synthesis?

- Categorical speech perception: use of synthetic sound continua (Lisker and Abramson, 1970)
- Motor theory of speech perception (Liberman and Mattingly, 1985), acoustic cue analysis
- Analysis by synthesis: modelling frameworks used for testing phonological models (Xu and Prom-On, 2014; Cerňak et al., 2017)

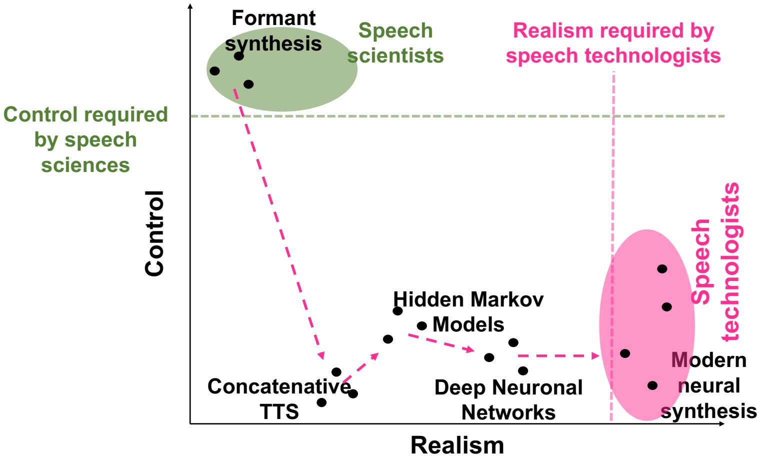
Speech science point of view



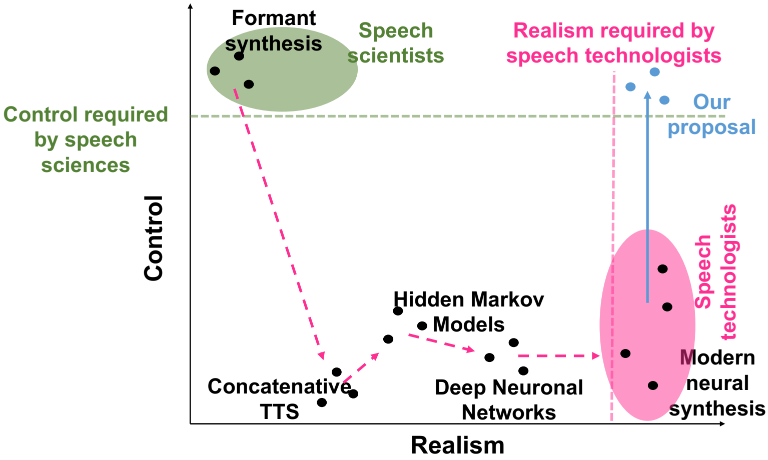
Why do phoneticians need speech synthesis?

- Stimuli creation: assess listeners' sensitivity to a particular acoustic cue in isolation
- Manipulation of e.g. formant transitions: how to exclude redundant and residual cues to place of articulation
- Control over single-cue variability limiting confounds
- MBROLA, PSOLA (Dutoit et al., 1996; Moulines and Charpentier, 1990) (Gao, this conference)
- Speech distortion and delexicalisation, noise-vocoding (White et al., 2015; Kolly and Dellwo, 2014)

Current situation



Proposed development



Simultaneous routes towards the goal

- Resources: What can be achieved by open code and databases with modest computation?
- Evaluation: a case for careful evaluation leading to robust and standardised benchmarking
- We are in new territory in terms of what TTS can do, new evaluation methods necessary
- Renewing dialogue between speech sciences and technology

New areas for research

- Generating conversational phenomena "on demand" (Szekely et al. submitted)
- Phenomena difficult to elicit from human speakers in empirical designs (optional, non-intentional)
- "Artificial speech" vs. realistic speaker babble (WaveNet)

Control

- Controllable neural vocoder: MFCCs re-placed with more phonetically meaningful speech parameters (Juvela et al., 2018)
- Same parameters can be predicted from text (Tacotron, Wang et al. (2017))
- Control of high-level features (Malisz et al. 2017; SSW submitted)

Modern speech synthesis for phonetic sciences: a discussion and an evaluation

Where are we on realism exactly?

- What is the actual perceptual difference between natural speech and modern synthesis?
- Winters and Pisoni (2004) showed that classic synthesis:
  - is less intelligible
  - overburdens attention and cognitive mechanisms resulting in slower processing times
- Compare natural speech, classic synthesis and modern synthesisers on:
  - listener preference
  - intelligibility
  - speed of processing

System	Type	Paradigm	Signal gen.
NAT	-	Natural	Vocal tract
VOC	**SISO**	**Copy synthesis**	**MagPhase**
MERLIN	TISO	Stat. parametric	MagPhase
GL	SISO	Copy synthesis	Griffin-Lim
DCTTS	TISO	End-to-end	Griffin-Lim
OVE	TISO	Rule-based	Formant

- Copy synthesis (acoustic analysis followed by re-synthesis) with the MagPhase vocoder (Espic et al. 2017)

System	Type	Paradigm	Signal gen.
NAT	-	Natural	Vocal tract
VOC	SISO	Copy synthesis	MagPhase
MERLIN	**TISO**	**Stat. parametric**	**MagPhase**
GL	SISO	Copy synthesis	Griffin-Lim
DCTTS	TISO	End-to-end	Griffin-Lim
OVE	TISO	Rule-based	Formant

- Synthetic speech generated by the Merlin TTS system Wu et al. (2016) using the MagPhase vocoder.
- Standard research grade statistical-parametric TTS.

System	Type	Paradigm	Signal gen.
NAT	-	Natural	Vocal tract
VOC	SISO	Copy synthesis	MagPhase
MERLIN	TISO	Stat. parametric	MagPhase
GL	**SISO**	**Copy synthesis**	**Griffin-Lim**
DCTTS	TISO	End-to-end	Griffin-Lim
OVE	TISO	Rule-based	Formant

- Copy synthesis from magnitude mel-spectrograms using the Griffin-Lim algorithm (Griffin 1984) for phase reconstruction.

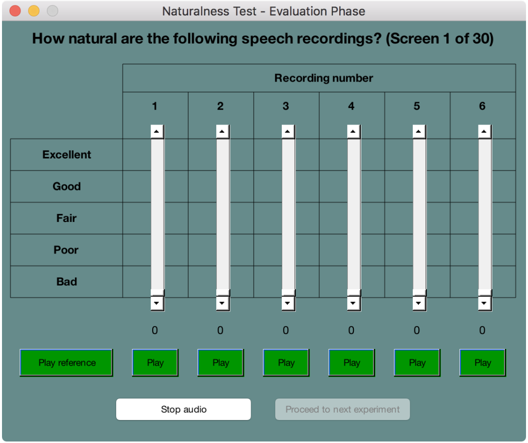
System	Type	Paradigm	Signal gen.
NAT	-	Natural	Vocal tract
VOC	SISO	Copy synthesis	MagPhase
MERLIN	TISO	Stat. parametric	MagPhase
GL	SISO	Copy synthesis	Griffin-Lim
DCTTS	**TISO**	**End-to-end**	**Griffin-Lim**
OVE	TISO	Rule-based	Formant

- Tacotron-like TTS using deep convolutional networks as in (Tachibana et al. 2018) with Griffin-Lim signal generation.

System	Type	Paradigm	Signal gen.
NAT	-	Natural	Vocal tract
VOC	SISO	Copy synthesis	MagPhase
MERLIN	TISO	Stat. parametric	MagPhase
GL	SISO	Copy synthesis	Griffin-Lim
DCTTS	TISO	End-to-end	Griffin-Lim
OVE	**TISO**	**Rule-based**	**Formant**

- Rule-based formant TTS system (Carlson et al. 1982, Sjolander et al. 1998) configured to use a male RP British English voice.
- Research-grade formant-based TTS.
- Permits optional prosodic emphasis control.

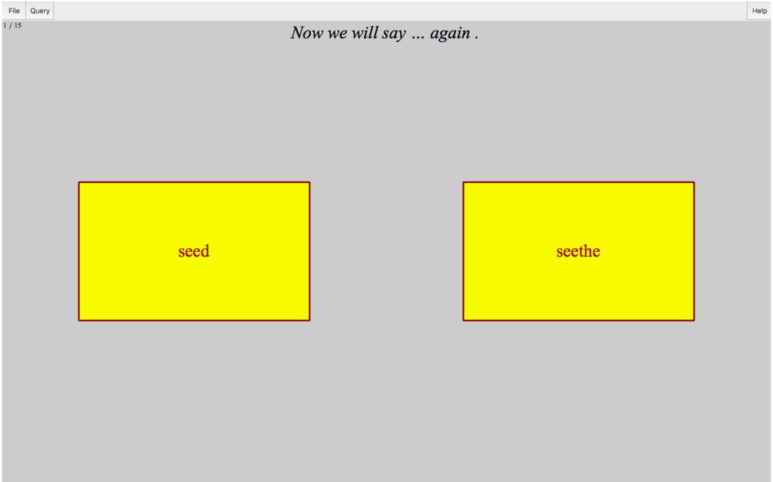
Subjective rating: MUSHRA test



Subjective rating: MUSHRA test

- The test used 20 native English-speaking listeners, N=799 ratings per system
- Listeners rated stimuli representing the different systems speaking four sets of ten Harvard sentences (designed to be approximately phonetically balanced)

Lexical decision: correct response rate and reaction time test

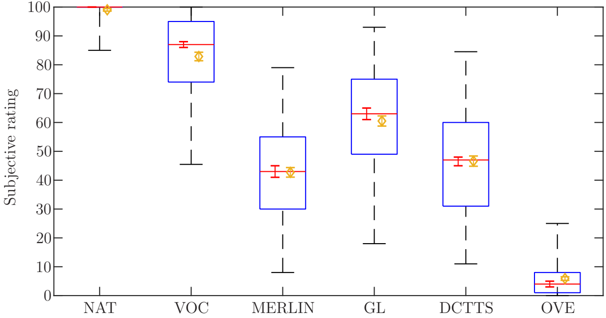


Lexical decision: correct response rate and reaction time test

Lexical decision: correct response rate and reaction time test

- We tested 20 listeners, 600 choices and reaction times per listener
- Stimuli: CVC words from 50 minimal pairs selected from MRT, embedded in a fixed carrier sentence rendered by the six different systems.

Results: subjective rating via MUSHRA



- Pairwise system differences all statistically significant ($p < 0.001$),
- VOC was rated above NAT 5.7% of the time
- MERLIN was rated above NAT 0.38% of the time

Results: correct response rate and reaction time via lexical decision

System	Estimate	*p*-value	Incorrect
NAT (ref.)			2.6%
GL	-0.001	$= 0.94$	4.0%
VOC	0.02	$= 0.33$	2.5%
DCTTS	0.04	< 0.01	5.8%
MERLIN	0.02	$= 0.14$	3.0%
OVE	0.09	< 0.001	6.0%

Results: correct response rate

System	Estimate	p-value	Incorrect
NAT (ref.)			**2.6%**
GL	-0.001	$= 0.94$	4.0%
VOC	0.02	$= 0.33$	**2.5%**
DCTTS	0.04	< 0.01	5.8%
MERLIN	0.02	$= 0.14$	**3.0%**
OVE	0.09	< 0.001	6.0%

Results: reaction times

System	Estimate	*p*-value	Incorrect
NAT (ref.)			2.6%
GL	-0.001	**= 0.94**	4.0%
VOC	0.02	**= 0.33**	2.5%
DCTTS	0.04	< 0.01	5.8%
MERLIN	0.02	**= 0.14**	3.0%
OVE	0.09	< 0.001	6.0%

Conclusions

- Modern methods largely overcome the processing inadequacies of systems commonly used in speech sciences.
- Include speech manipulation and neural vocoders to further improve on the quality of systems for speech sciences
- You can always use OVE for the "artificial speech" quality but realistic synthesis should generalise better to actual speech perception

Thank you!
Tack så mycket!

Acknowledgements

This research was funded by

Cerňak, M., Beňuš, Š., and Lazaridis, A. (2017). Speech vocoding for laboratory phonology. *Comput. Speech Lang.*, 42:100–121.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., and Van der Vrecken, O. (1996). The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. In *Proc. ICSLP*, pages 1393–1396.

Juvela, L., Bollepalli, B., Wang, X., Kameoka, H., Airaksinen, M., Yamagishi, J., and Alku, P. (2018). Speech waveform synthesis from MFCC sequences with generative adversarial networks. In *Proc. ICASSP*, pages 5679–5683.

King, S. (2015). What speech synthesis can do for you (and what you can do for speech synthesis). In *Proc. ICPhS*.

Kolly, M.-J. and Dellwo, V. (2014). Cues to linguistic origin: The contribution of speech temporal information to foreign accent recognition. *Journal of Phonetics*, 42:12–23.

Liberman, A. M. and Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, 21(1):1–36.

Lisker, L. and Abramson, A. S. (1970). The voicing dimension: Some experiments in comparative phonetics. In *Proc. ICPhS*, pages 563–567.

Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Commun.*, 9(5-6):453–467.

Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., and et al. (2017). Tacotron: Towards end-to-end speech synthesis. In *Proc. Interspeech*, pages 4006–4010.