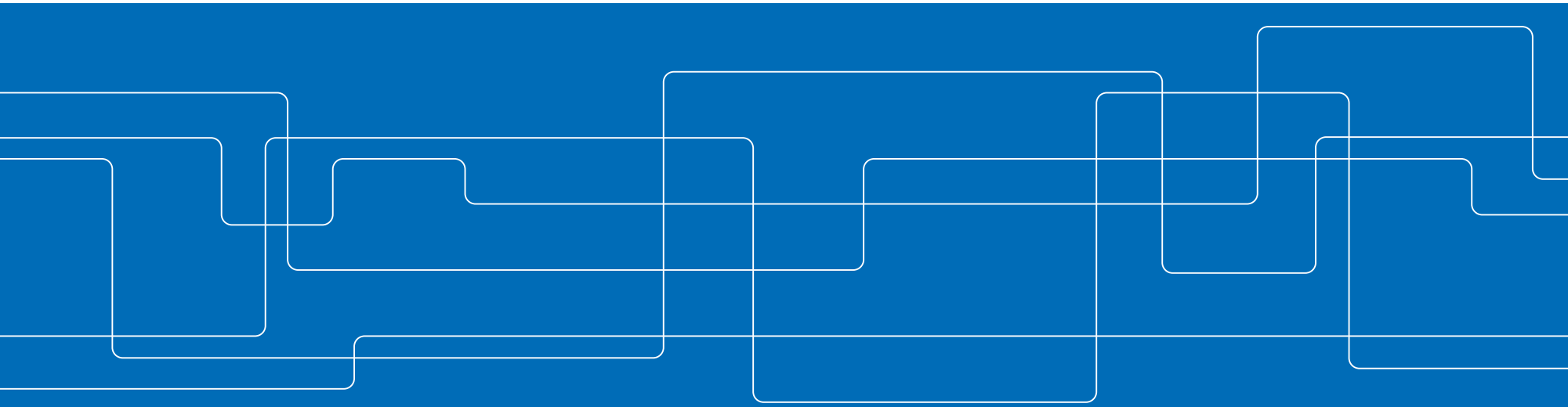
Modern speech synthesis and its implications for speech sciences

Zofia Malisz[1], Gustav Eje Henter[1], Cassia Valentini-Botinhao[2], Oliver Watts[2], Jonas Beskow[1], Joakim Gustafson[1]

[1] Division of Speech, Music and Hearing (TMH),
KTH Royal Institute of Technology, Stockholm, Sweden

[2] The Centre for Speech Technology Research (CSTR),
The University of Edinburgh, UK



Take-home message

- Once upon a time, speech technology and speech sciences were engaged in a dialogue that benefitted both fields
- Differences in priorities have caused the fields to grow apart
- Recent speech-synthesis developments have eliminated old hurdles for speech scientists
- The interests of the two fields are now converging
- This an opportunity for both speech technologists and speech scientists

Speech synthesis contributions to phonetics

- 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 contributions to synthesis

- Speech science was instrumental for speech processing and engineering in the data-sparse formant-synthesis era (King, 2015)
- Phones and phone sets
- Perception-based modelling, e.g., the mel scale (Stevens et al., 1937)
- Sophisticated speech-synthesis evaluation methods derived from, e.g., psycholinguistics (Winters and Pisoni, 2004; Govender and King, 2018)

Why do technologists need speech sciences?

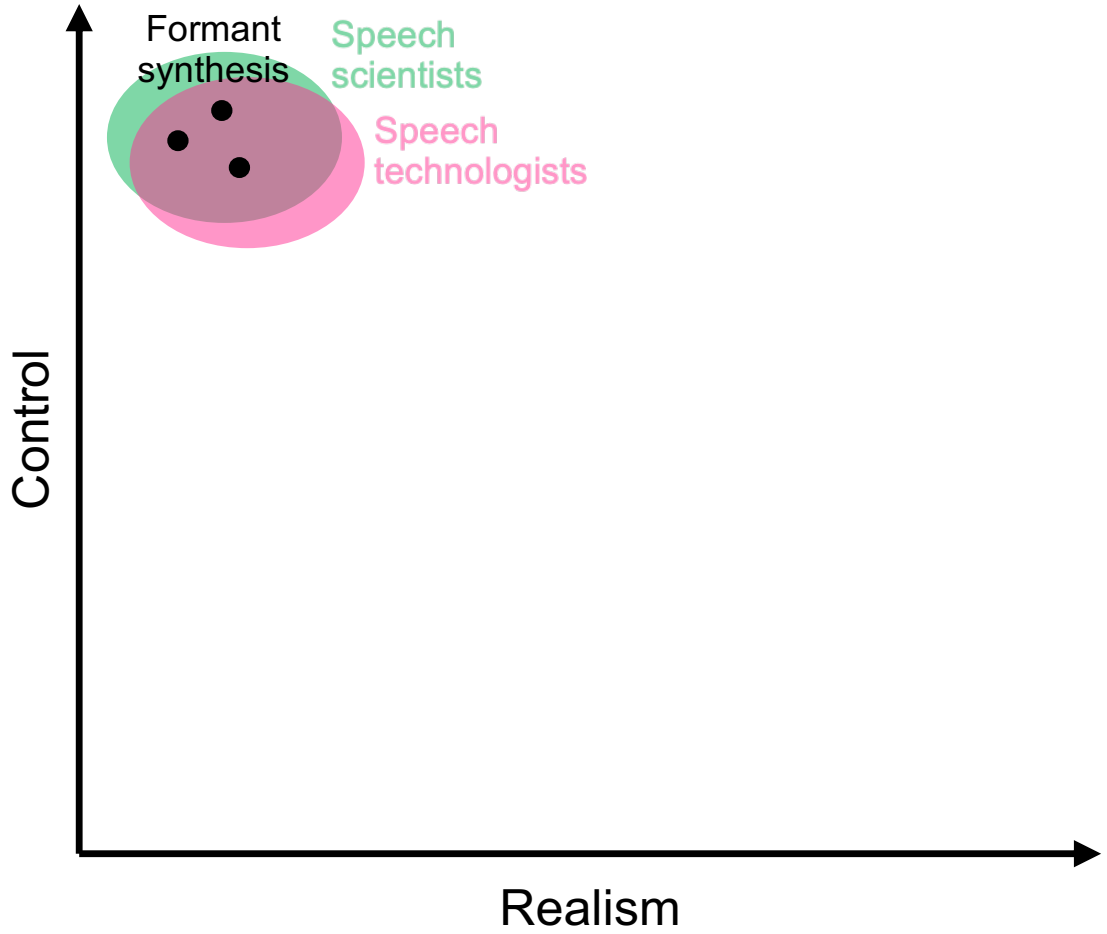
- Synthesis and analysis go hand in hand
- To understand data and results (beyond merely describing them)
- For a rigorous approach to evaluation and analysis

Why do phoneticians need speech synthesis?

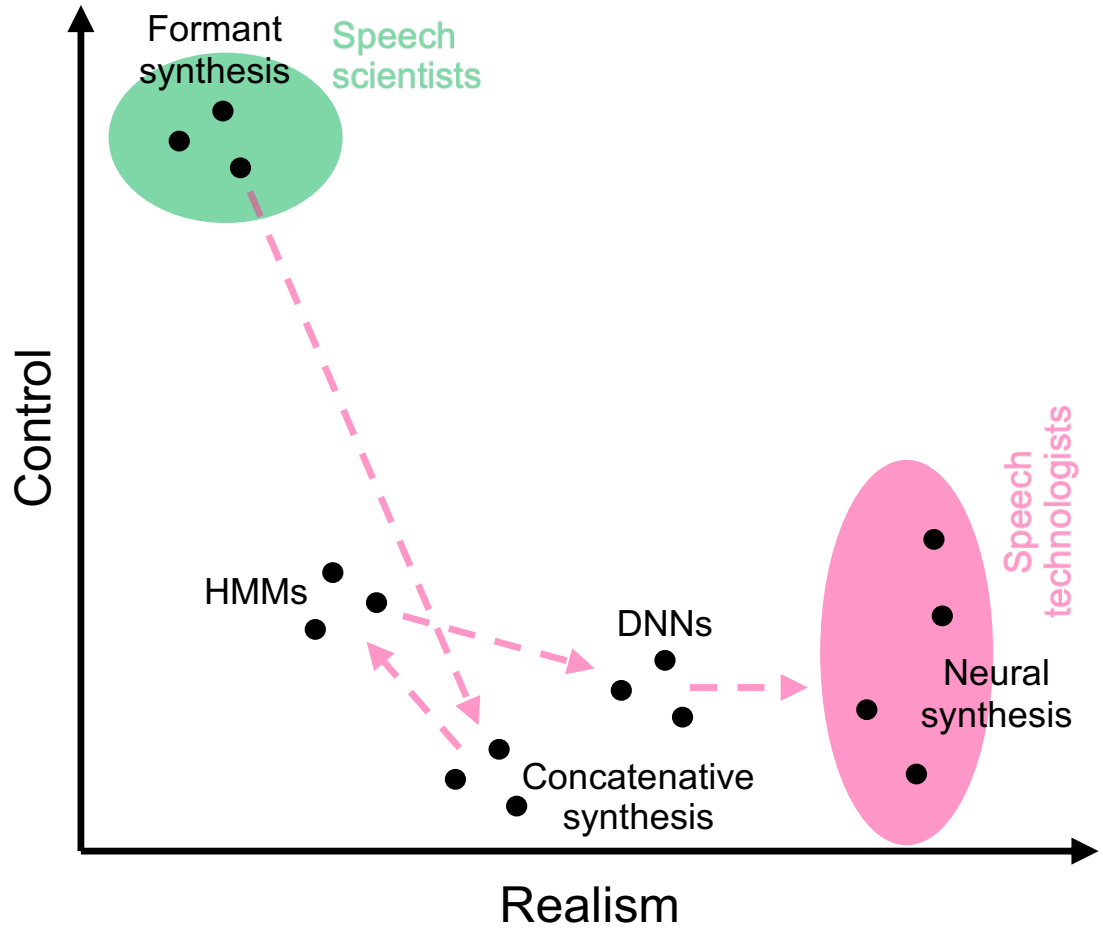
- Stimulus creation: Assess listeners' sensitivity to particular acoustic cues in isolation
  - Manipulation of, e.g., formant transitions while excluding redundant and residual cues to place of articulation
  - Control over single-cue variability, limiting confounds
  - PSOLA, MBROLA, STRAIGHT for creating and manipulating speech (Moulines and Charpentier, 1990; Dutoit et al., 1996; Kawahara, 2006)
  - Speech distortion and delexicalisation; noise-vocoding (White et al., 2015; Kolly and Dellwo, 2014)

Why is synthetic speech so rare in contemporary speech sciences?

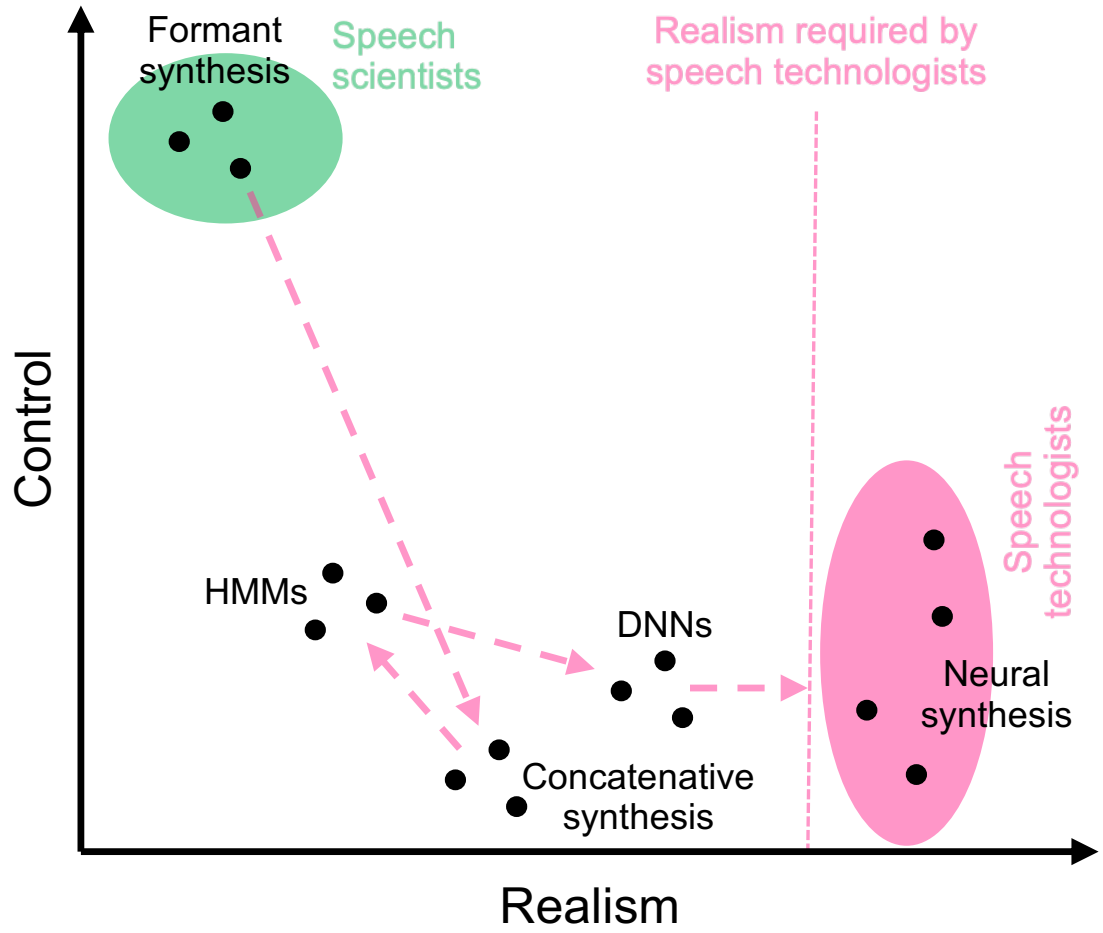
Then and now in synthetic speech



Then and now in synthetic speech



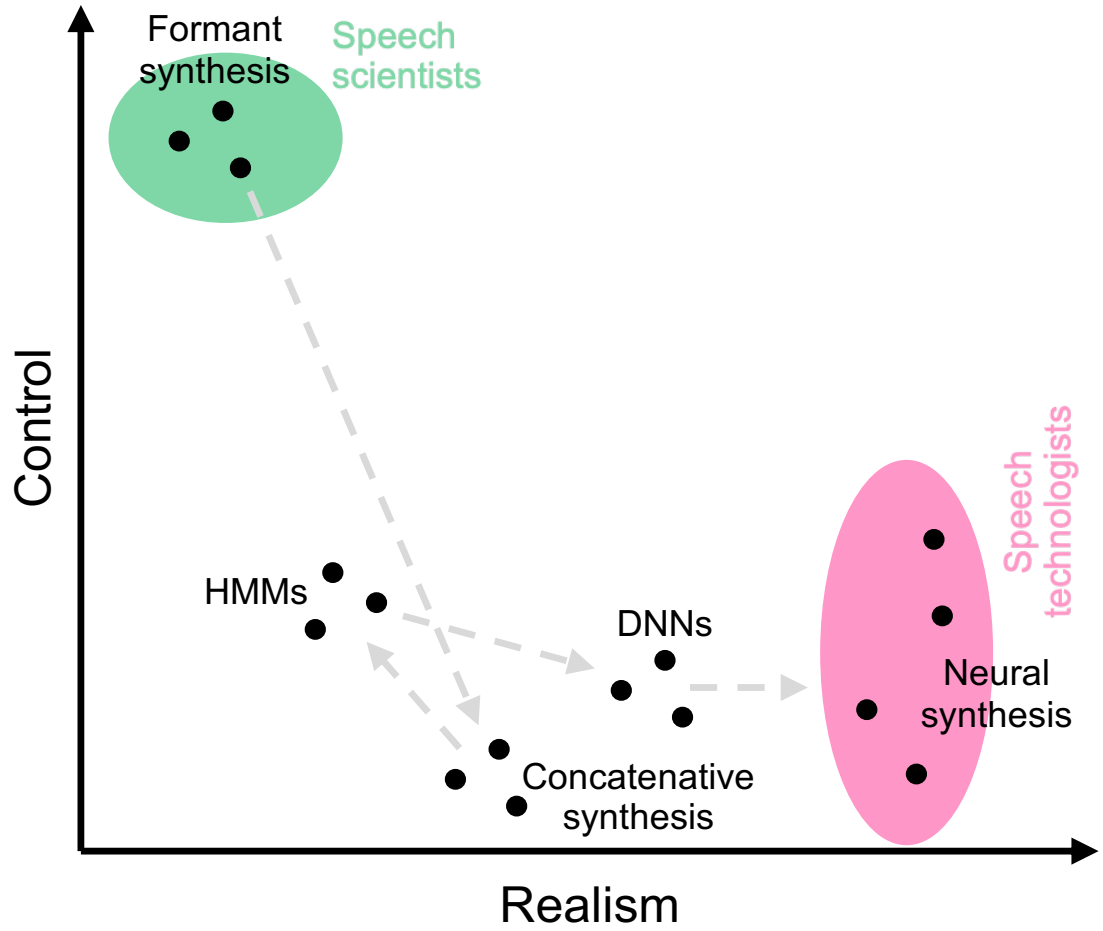
Then and now in synthetic speech



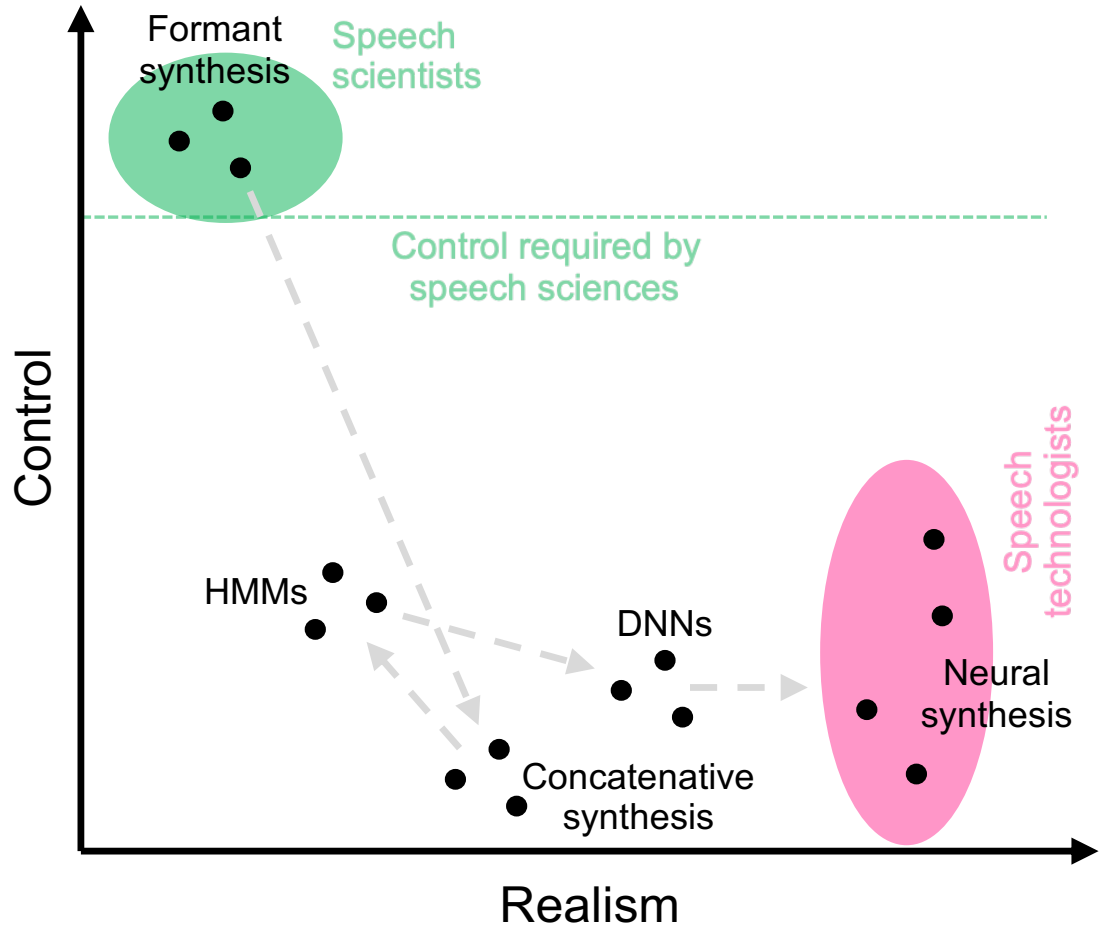
Recent synthesis naturalness achievements

- Highly natural speech-signal generation with neural vocoders such as WaveNet (van den Oord et al., 2016)
- Vastly improved text-to-speech prosody (in English) with end-to-end approaches such as Tacotron (Wang et al., 2017)
- TTS naturalness rated close to recorded speech in mean opinion score (Shen et al., 2018)

Speech science point of view



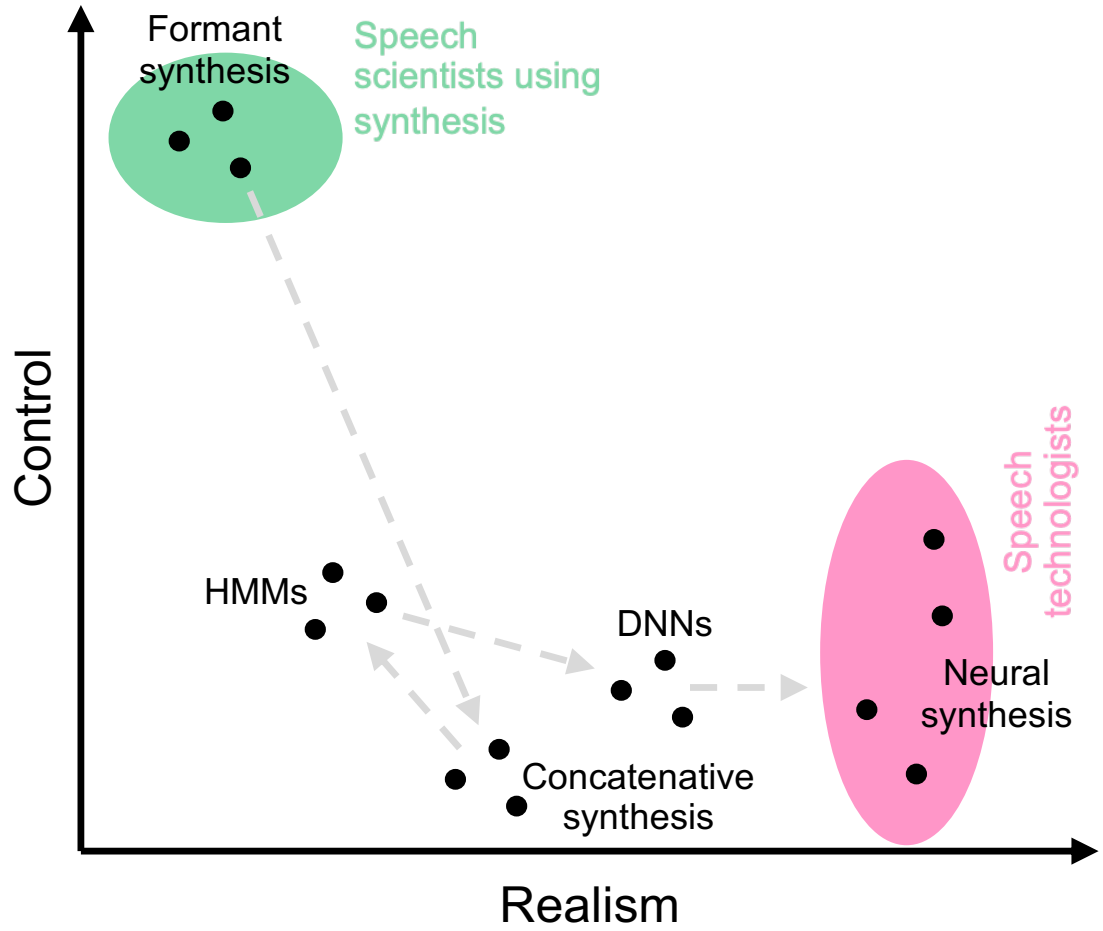
Speech science point of view



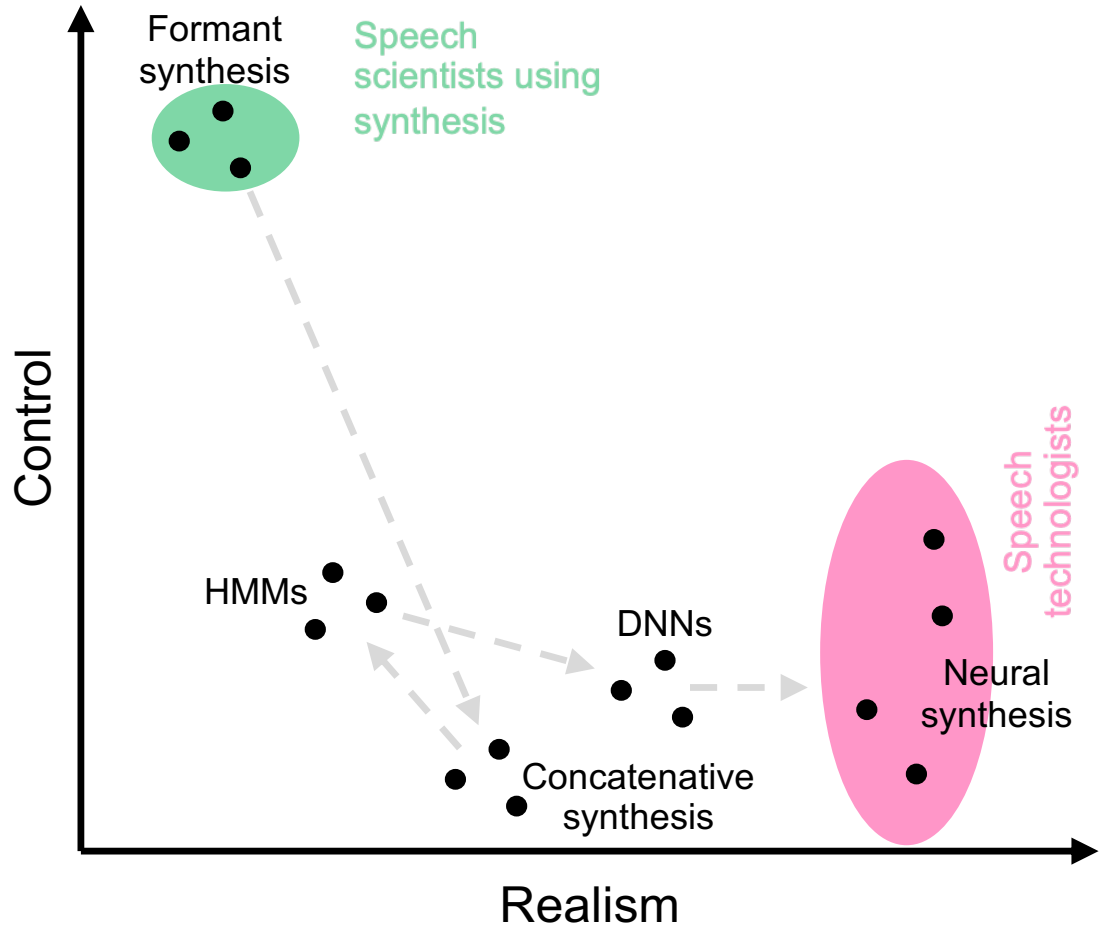
Why so little synthesis in speech sciences?

- Newer speech synthesis does not provide the precise control required for phonetic research
- Little overlap between communities means that few phoneticians have the technical knowledge to adapt synthesis developments for their needs

Troubling developments



Troubling developments



The perception problem

- A body of research, as reviewed by Winters and Pisoni (2004), shows that classic formant synthesis:
  - Is less intelligible than recorded speech
  - Overburdens attention and cognitive mechanisms resulting in slower processing times (Duffy and Pisoni, 1992)
- . . . in addition to receiving low naturalness ratings

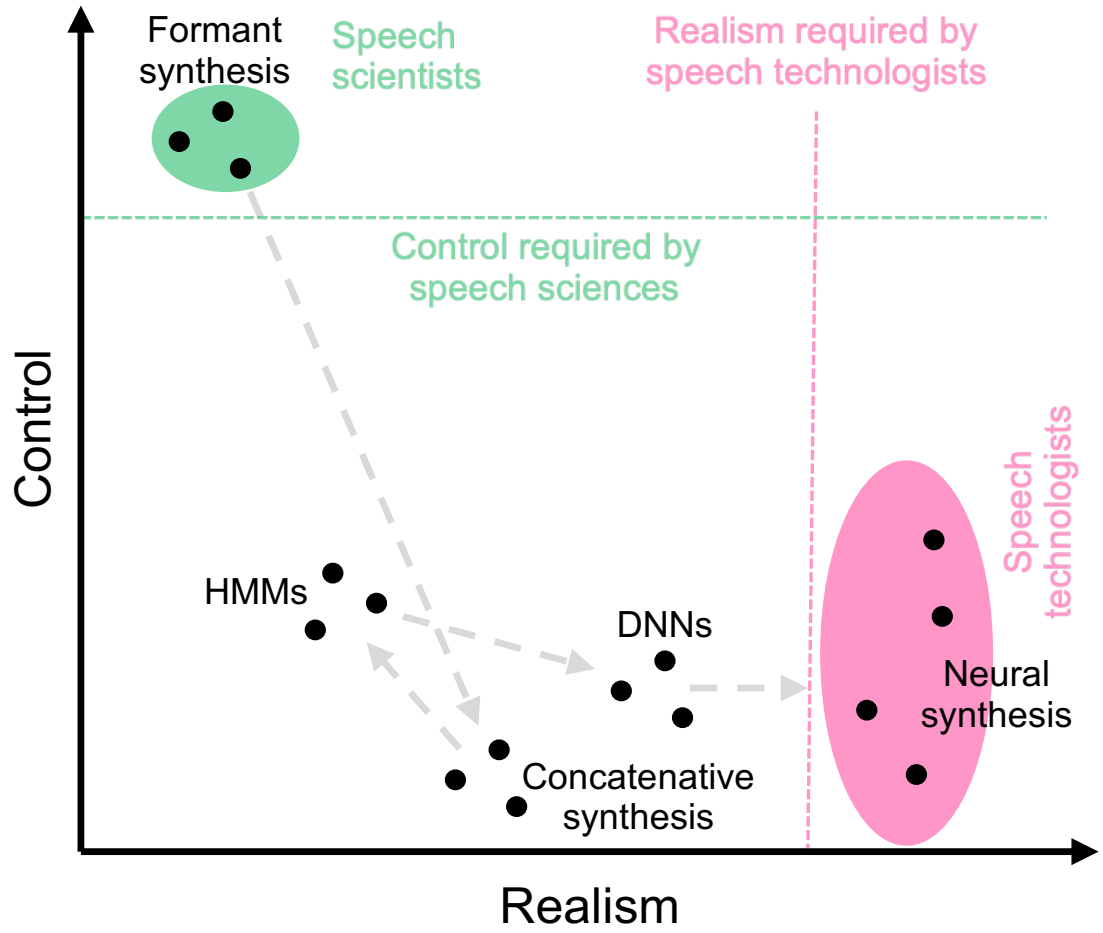
Why so little synthesis in speech sciences?

- Newer speech synthesis does not provide the precise control required for phonetic research
- Little overlap between communities means that few phoneticians have the technical knowledge to adapt synthesis developments for their needs
- Differences in perception between natural and classical synthesised speech cast doubt on the universality of research findings (Iverson, 2003)

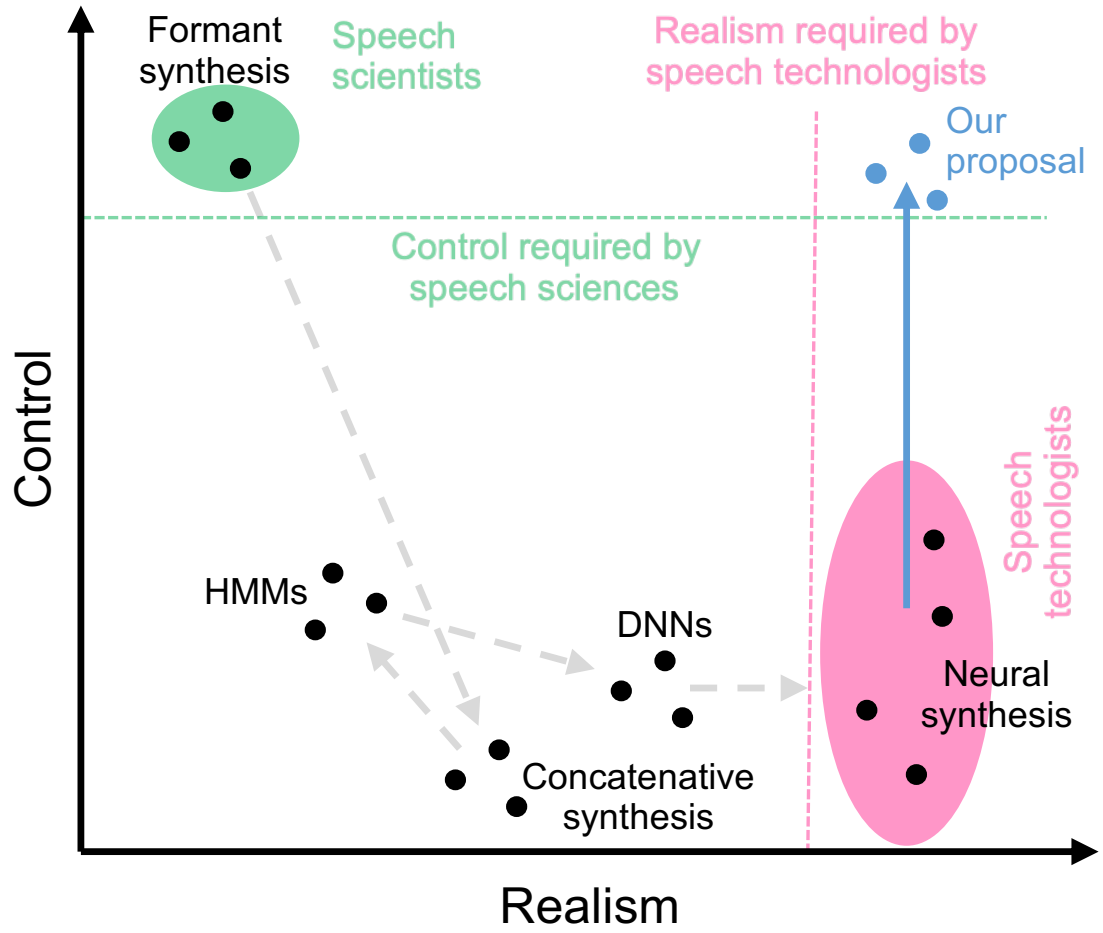
Our beliefs

1. Speech technologists should pursue accurate output-control for modern speech synthesis paradigms
2. Speech scientists should pay attention and contribute to these developments
3. Issues of perceptual inadequacy have largely been overcome

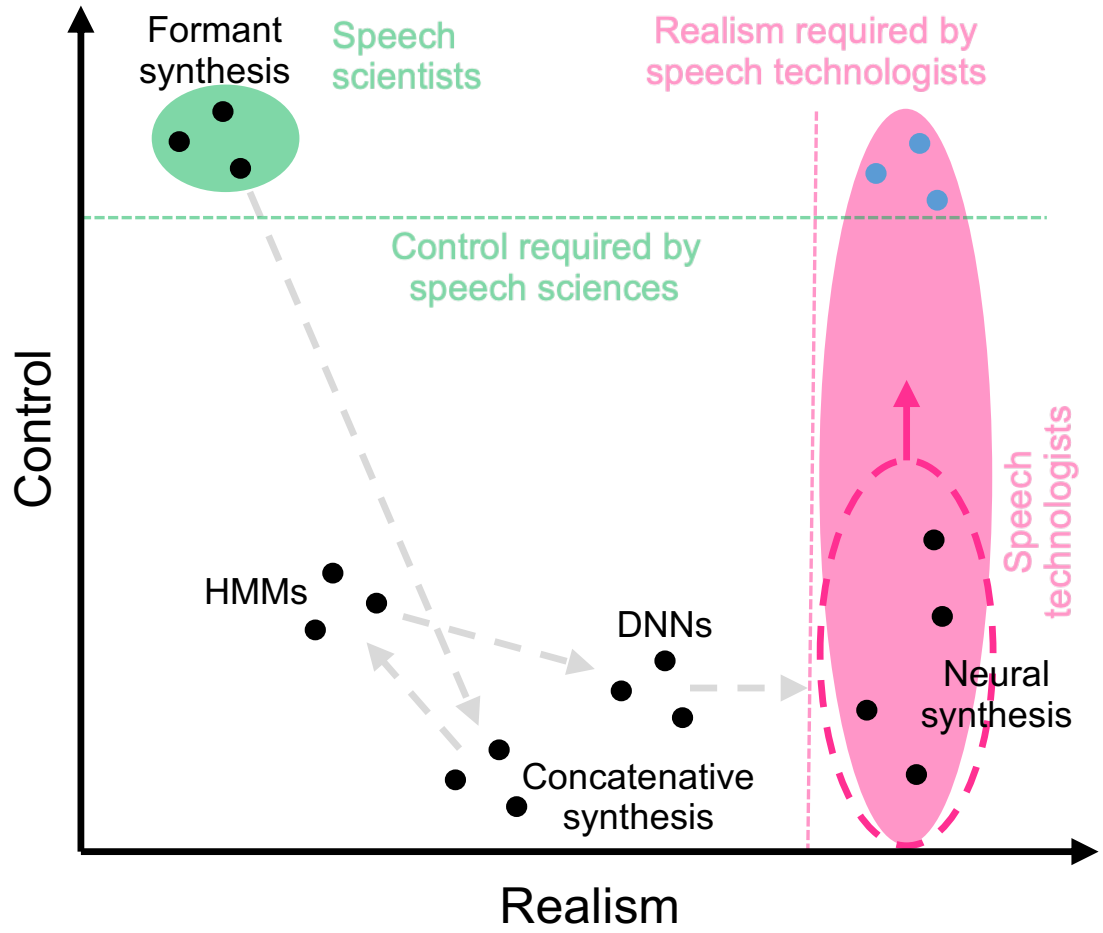
Technological agenda



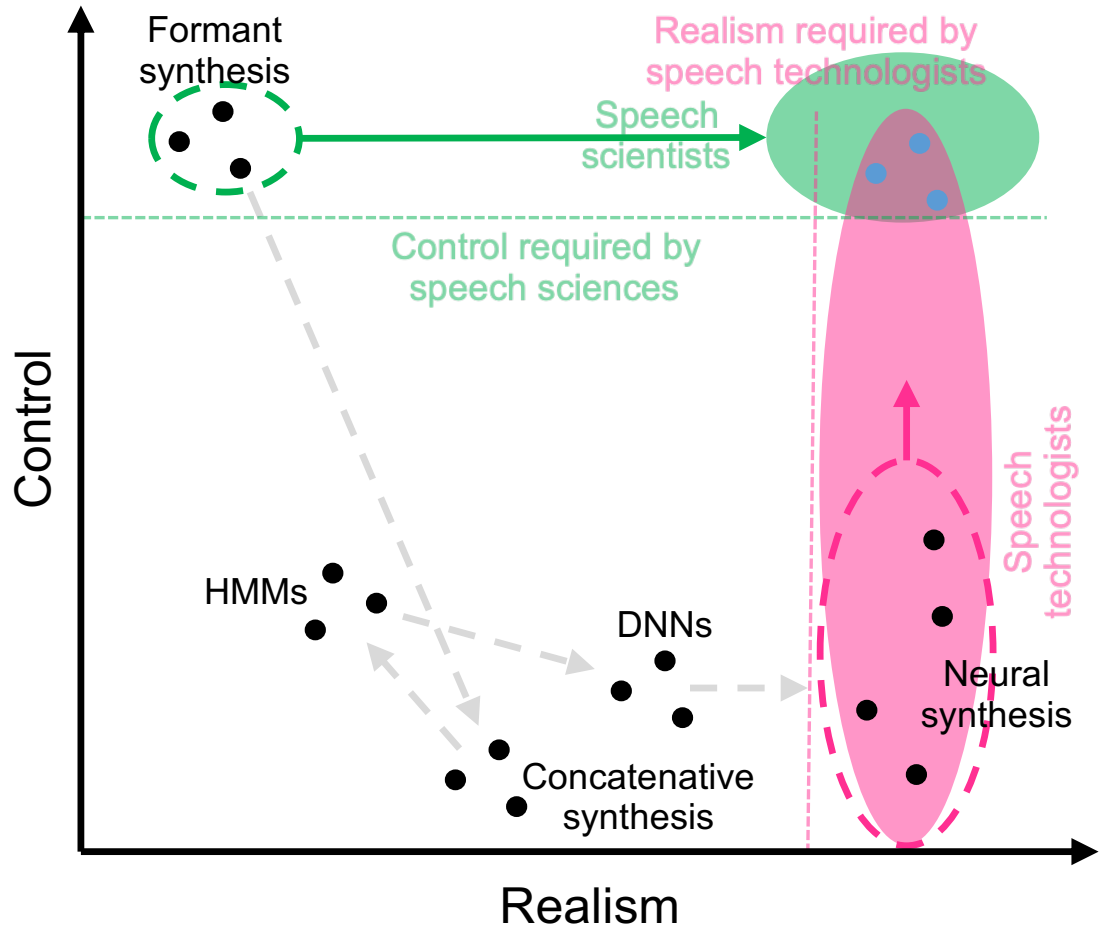
Technological agenda



Technological agenda



Technological agenda



Examples of new technological research

- Controllable neural vocoder for phonetics: MFCC control interface (Juvela et al., 2018) replaced with more phonetically-meaningful speech parameters
  - These speech parameters can alternatively be predicted from text, e.g., using Tacotron
- Control of high-level speech features, e.g., prominence (Malisz et al., 2017)

Examples of new phonetic research areas

- Improved and controllable synthesis not only offers better stimuli for established research directions, but also opens new areas such as. . .
  - Generating conversational phenomena “on demand” (Székely et al., 2019)
  - Generating optional or non-intentional phenomena that are difficult to elicit from human speakers in empirical designs (e.g., conversational clicks)
  - “Artificial speech” vs. realistic speaker babble, e.g., from unconditional WaveNet

Examples of new joint research

- New robust and meaningful evaluation methods for today's highly-capable speech synthesisers
- Result: Rekindling the productive dialogue between speech sciences and speech technology

What about the perceptual issues?

- We know from before that classic speech synthesis:
  - Is rated as less natural than recorded speech
  - Is less intelligible than recorded speech
  - Yields slower cognitive processing times than recorded speech
- To what extent is this still true?

What about the perceptual issues?

- We know from before that classic speech synthesis:
  - Is rated as less natural than recorded speech
  - Is less intelligible than recorded speech
  - Yields slower cognitive processing times than recorded speech
- To what extent is this still true?
- Empirical study: Compare natural speech, classic synthesis, and modern deep-learning synthesis on:
  - Subjective listener ratings
  - Intelligibility
  - Speed of processing
- . . . using open code and databases and modest computational resources

Systems compared

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

- Corpus taken from Cooke et al. (2013), including approximately 2k utterances for voice building
- SISO = Speech in, speech out
- TISO = Text in, speech out

Systems compared

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)

Systems compared

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

Systems compared

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 and Lim, 1984) for phase reconstruction

Systems compared

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

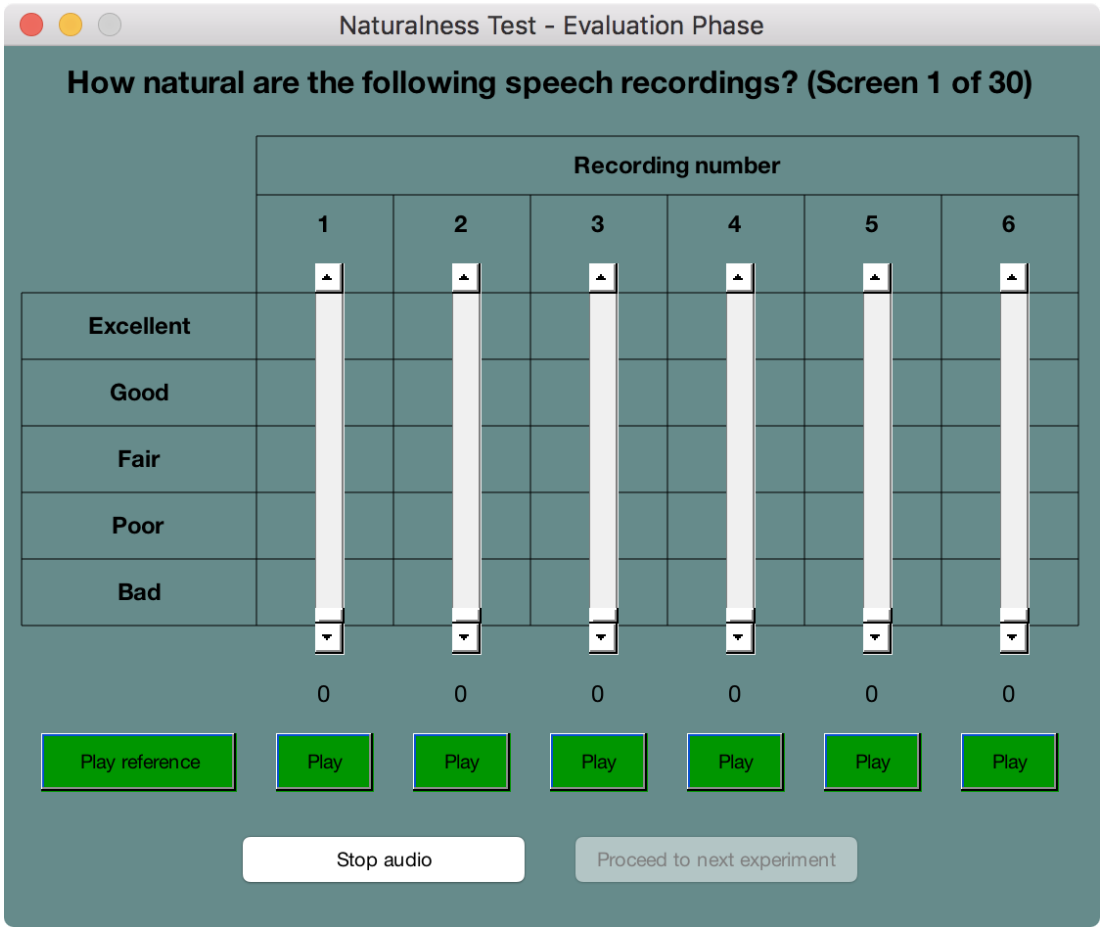
- Tacotron-like TTS using deep convolutional networks as in Tachibana et al. (2018) with Griffin-Lim signal generation
- Pre-trained on 11.6k utterances from another speaker to learn attention and accurate pronunciation

Systems compared

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; Sjölander et al., 1998) configured to use a male RP British English voice
- Research-grade formant-based TTS
- Permits optional prosodic emphasis control

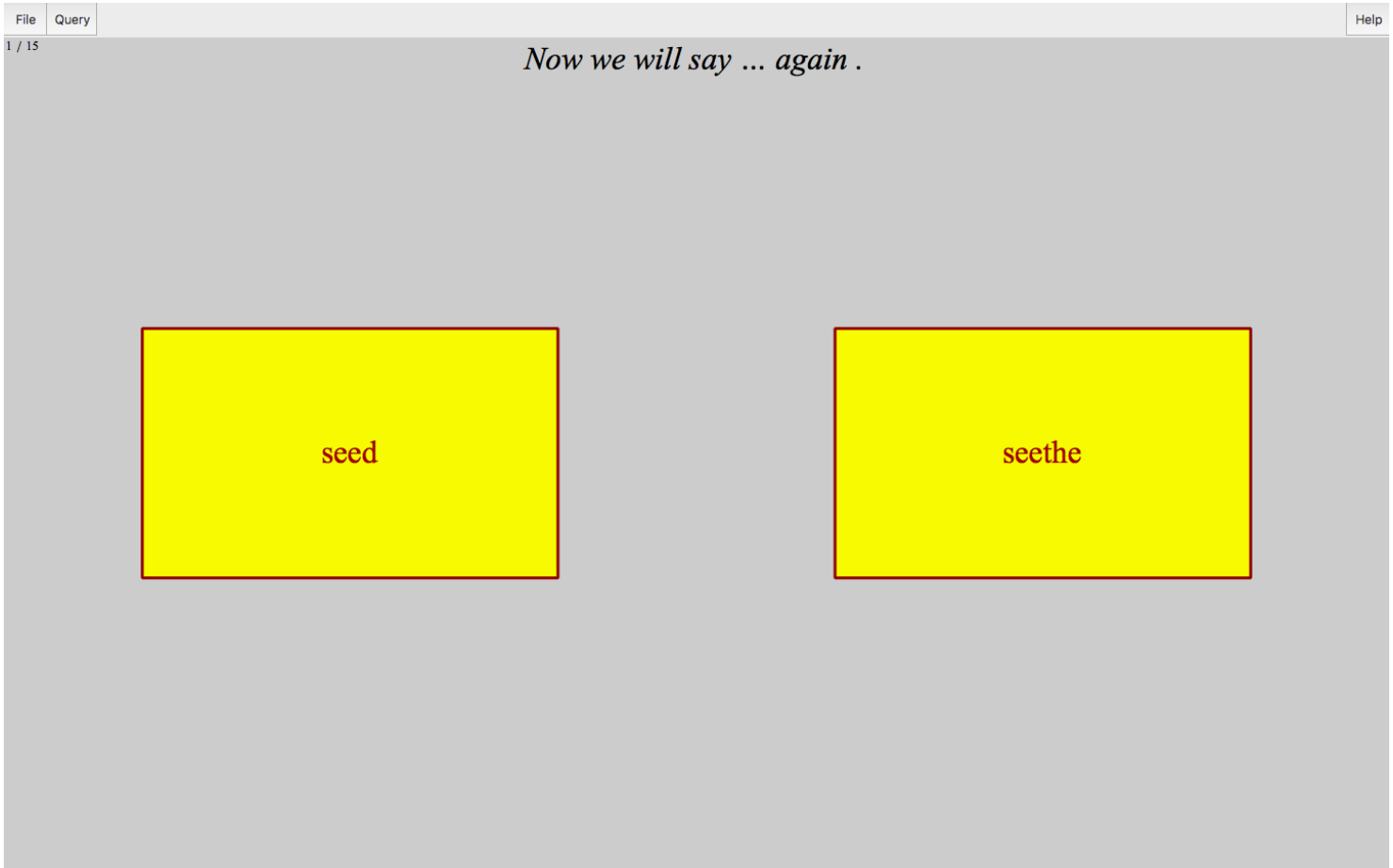
Subjective rating: MUSHRA test



Subjective rating: MUSHRA test

- MUSHRA tests are an ITU standard (ITU, 2015)
- Listeners rated stimuli representing the different systems speaking four sets of ten Harvard sentences (Rothauser and et al., 1969), designed to be approximately phonetically balanced
- 20 native English-speaking listeners provided a total of $N = 799$ ratings per system

Lexical decision: Correct response rate and reaction time test



Lexical decision: Correct response rate and reaction time test

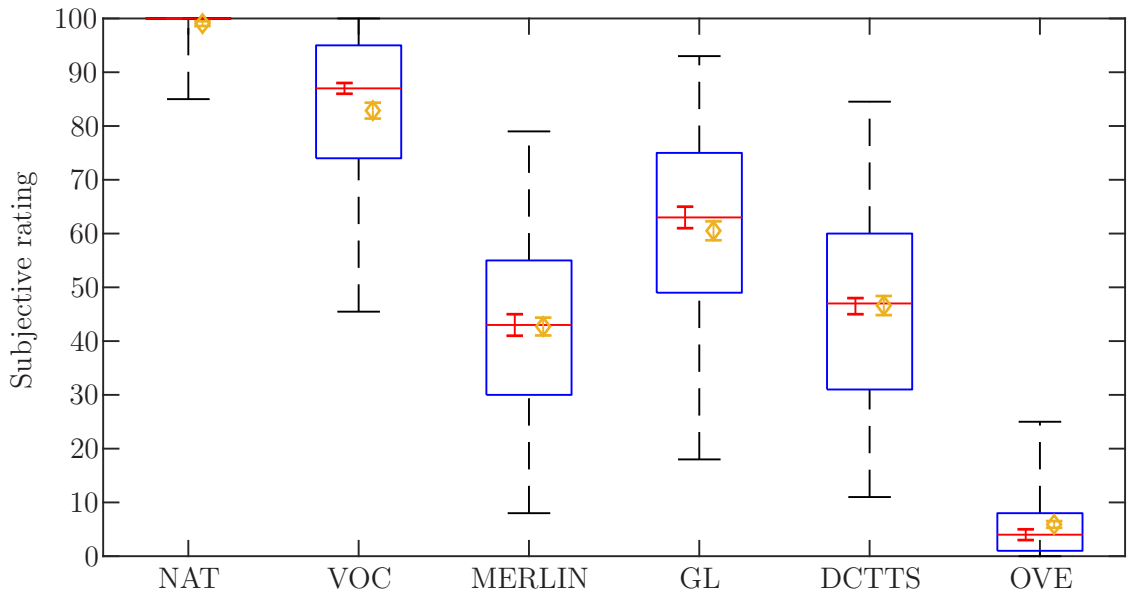
Lexical decision: Correct response rate and reaction time test

- Stimuli were CVC words from 50 minimal pairs selected from the modified rhyme test (House et al., 1963), embedded in a fixed carrier sentence rendered by the six different systems
- We tested 20 listeners, with 600 choices and reaction times per listener

Example stimuli

System	HVD	MRT 1	MRT 2
NAT	Play	Play	Play
VOC	Play	Play	Play
MERLIN	Play	Play	Play
GL	Play	Play	Play
DCTTS	Play	Play	Play
OVE	Play	Play	Play

Results: Subjective naturalness ratings



- Pairwise system differences are all statistically significant ($p < 0.001$),
- VOC was rated above NAT 5.7% of the time
- OVE was rated as the worst system 99% of the time

Results: Correct response rate and log-response time on lexical decision task

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

Results: Correct response rate and log-response time on lexical decision task

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

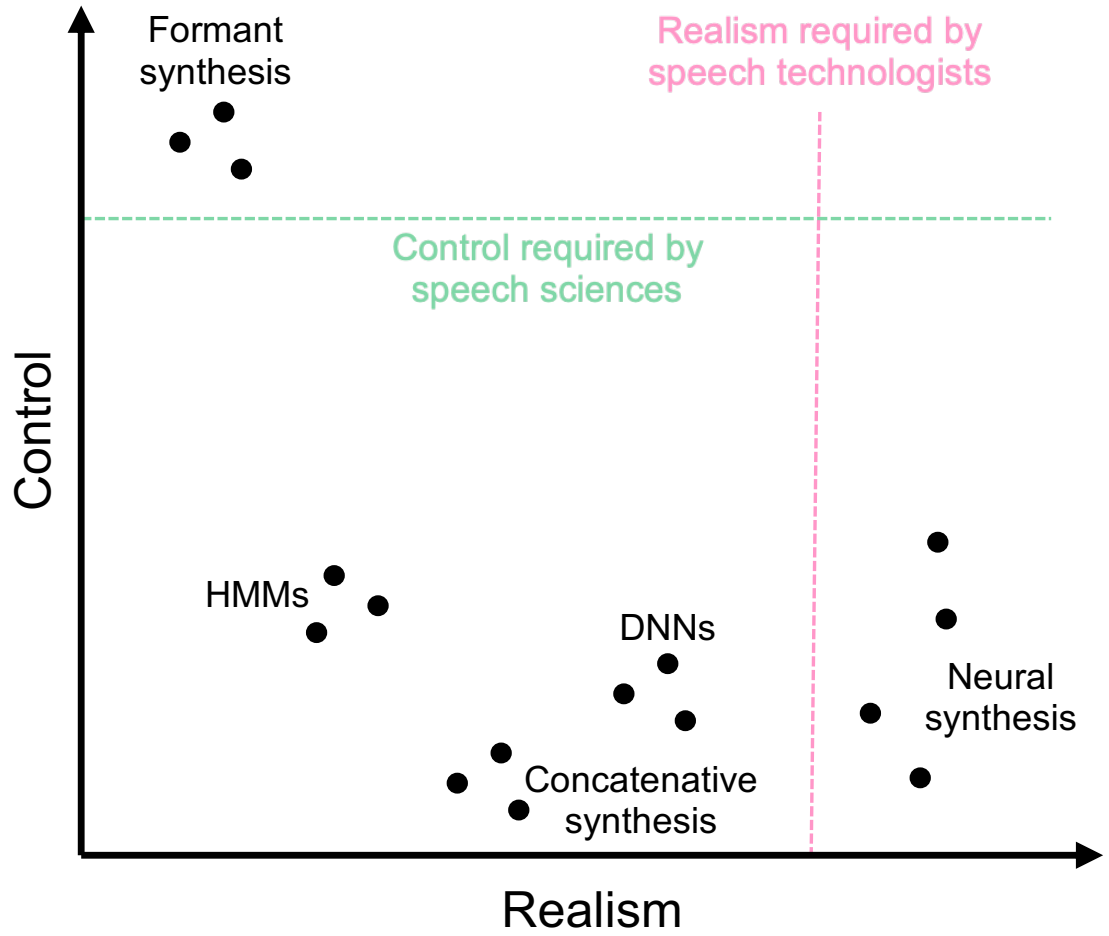
- Modern SISO and TISO systems can be close to natural speech in terms of intelligibility

Results: Correct response rate and log-response time on lexical decision task

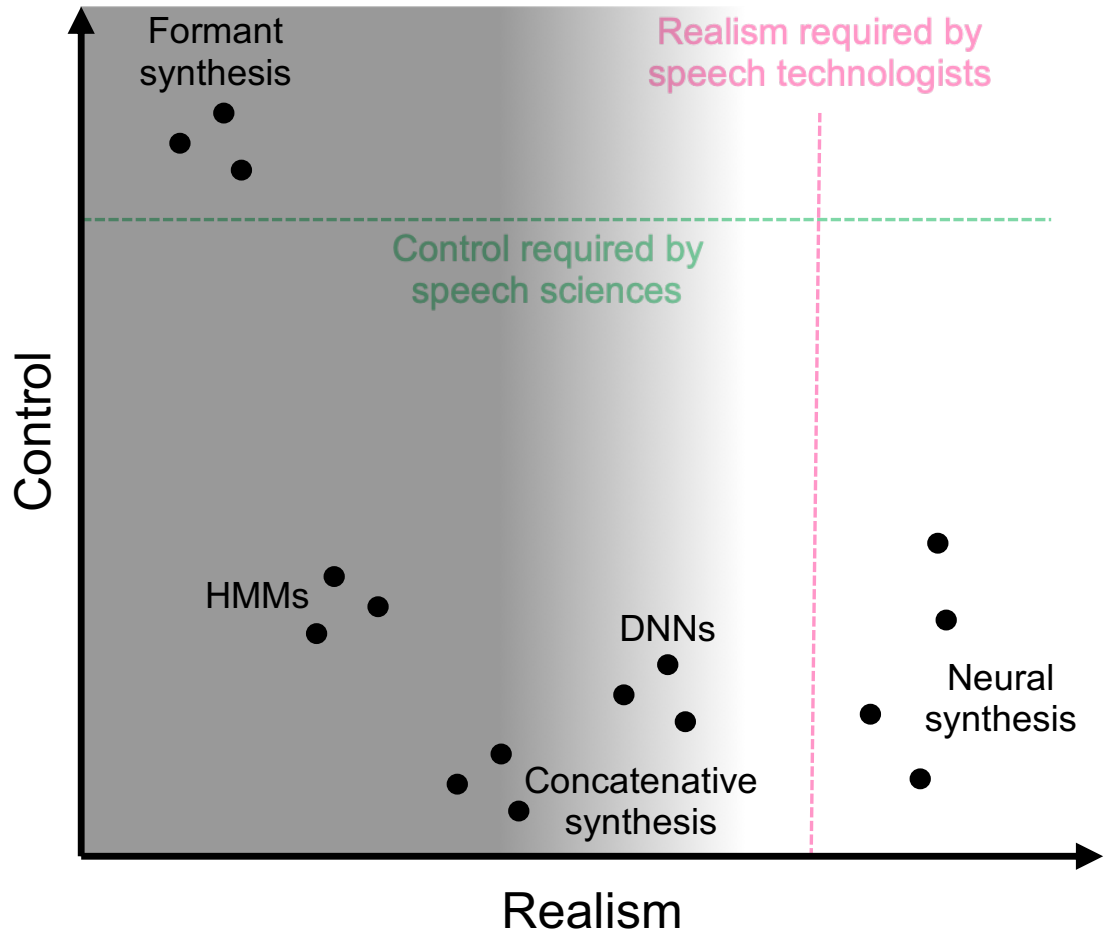
System	Est. effect	*p*-value	Incorrect
NAT (ref.)			2.6%
VOC	0.02	0.33	2.5%
MERLIN	0.02	0.14	3.0%
GL	-0.001	0.94	4.0%
DCTTS	0.04	$<$0.01	5.8%
OVE	0.09	$<$0.001	6.0%

- Modern SISO and TISO systems can be close to natural speech in terms of response time
- Classic formant synthesis shows slower processing times, consistent with prior literature

Graphical interpretation



Graphical interpretation



Summary and future work

- Modern speech synthesis with precise control is of interest to both scientists and technologists
  - This can bring the fields back in touch again
- Modern synthetic speech has largely overcome the perceptual inadequacies of systems commonly used in speech sciences
  - The situation for manipulated speech needs to be studied
  - Neural vocoders and more data or better adaptation should further improve technological capabilities
- Let's work together to make this happen!

Thank you for listening!

For more details see our ICPhS paper

Acknowledgements

This research was funded by:

- **ZM & JB:** Swedish Research Council grant no. 2017-02861
- **GEH:** Swedish Foundation for Strategic Research no. RIT15-0107
- **CVB & OW:** EPSRC Standard Research Grant EP/P011586/1
- **JG:** Swedish Research Council grant no. 2013-4935.

ZM and GEH thank Jens Edlund for helpful discussions.

References I

Carlson, R., Granström, B., and Hunnicutt, S. (1982). A multi-language text-to-speech module. In *Proc. ICASSP*, pages 1604–1607.

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

Cooke, M., Mayo, C., and Valentini-Botinhao, C. (2013). Intelligibility-enhancing speech modifications: The Hurricane Challenge. In *Proc. Interspeech*, pages 3552–3556.

References II

Duffy, S. A. and Pisoni, D. B. (1992). Comprehension of synthetic speech produced by rule: A review and theoretical interpretation. *Lang. Speech*, 35(4):351–389.

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.

Espic, F., Valentini-Botinhao, C., and King, S. (2017). Direct modelling of magnitude and phase spectra for statistical parametric speech synthesis. In *Proc. Interspeech*, pages 1383–1387.

References III

Govender, A. and King, S. (2018). Using pupillometry to measure the cognitive load of synthetic speech. In *Proc. Interspeech*, pages 2838–2842.

Griffin, D. and Lim, J. (1984). Signal estimation from modified short-time Fourier transform. *IEEE T. Acoust. Speech*, 32(2):236–243.

House, A. S., Williams, C., Hecker, M. H. L., and Kryter, K. D. (1963). Psychoacoustic speech tests: A modified rhyme test. *J. Acoust. Soc. Am.*, 35(11):1899–1899.

ITU (2015). Method for the subjective assessment of intermediate quality levels of coding systems. ITU Recommendation ITU-R BS.1534-3.

References IV

Iverson, P. (2003). Evaluating the function of phonetic perceptual phenomena within speech recognition: An examination of the perception of /d/–/t/ by adult cochlear implant users. *J. Acoust. Soc. Am.*, 113(2):1056–1064.

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.

Kawahara, H. (2006). STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. *Acoust. Sci. Technol.*, 27(6):349–353.

References V

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. *J. 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.

References VI

Malisz, Z., Berthelsen, H., Beskow, J., and Gustafson, J. (2017). Controlling prominence realisation in parametric DNN-based speech synthesis. In *Proc. Interspeech*, pages 1079–1083.

Malisz, Z., Henter, G. E., Valentini-Botinhao, C., Watts, O., Beskow, J., and Gustafson, J. (2019). Modern speech synthesis for phonetic sciences: a discussion and an evaluation. In *Proc. ICPhS*.

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.

References VII

Rothauser, E. H. and et al. (1969). IEEE recommended practice for speech quality measurements. *IEEE T. Acoust. Speech*, 17(3):225–246.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., and et al. (2018). Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In *Proc. ICASSP*, pages 4799–4783.

Sjölander, K., Beskow, J., Gustafson, J., Lewin, E., Carlson, R., and Granström, B. (1998). Web-based educational tools for speech technology. In *Proc. ICSLP*.

References VIII

Stevens, S. S., Volkmann, J., and Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *J. Acoust. Soc. Am.*, 8(3):185–190.

Székely, É., Henter, G. E., Beskow, J., and Gustafson, J. (2019). How to train your fillers: uh and um in spontaneous speech synthesis. Submitted to SSW 2019.

Tachibana, H., Uenoyama, K., and Aihara, S. (2018). Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In *Proc. ICASSP*, pages 4784–4788.

References IX

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and et al. (2016). WaveNet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

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.

White, L., Mattys, S. L., Stefansdottir, L., and Jones, V. (2015). Beating the bounds: Localized timing cues to word segmentation. *J. Acoust. Soc. Am.*, 138(2):1214–1220.

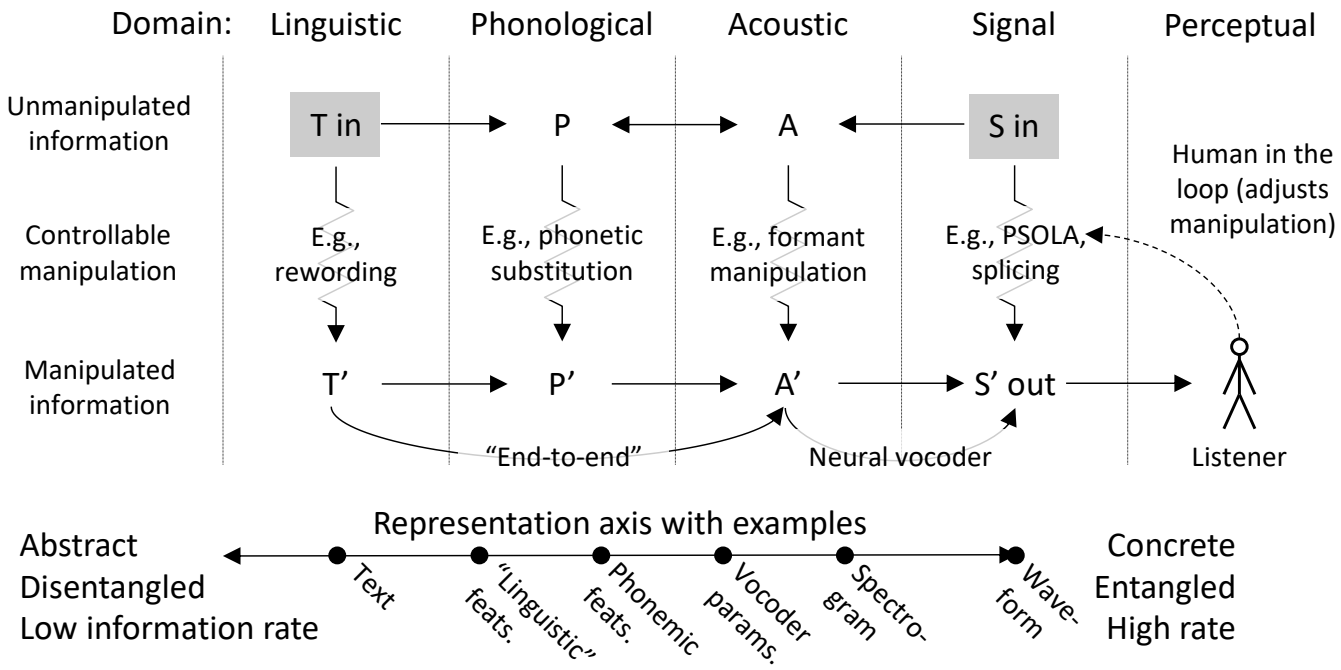
References X

Winters, S. J. and Pisoni, D. B. (2004). Perception and comprehension of synthetic speech. *Research on Spoken Language Processing Progress Report*, (26):95–138.

Wu, Z., Watts, O., and King, S. (2016). Merlin: An open source neural network speech synthesis system. In *Proc. SSW*, volume 9, pages 218–223.

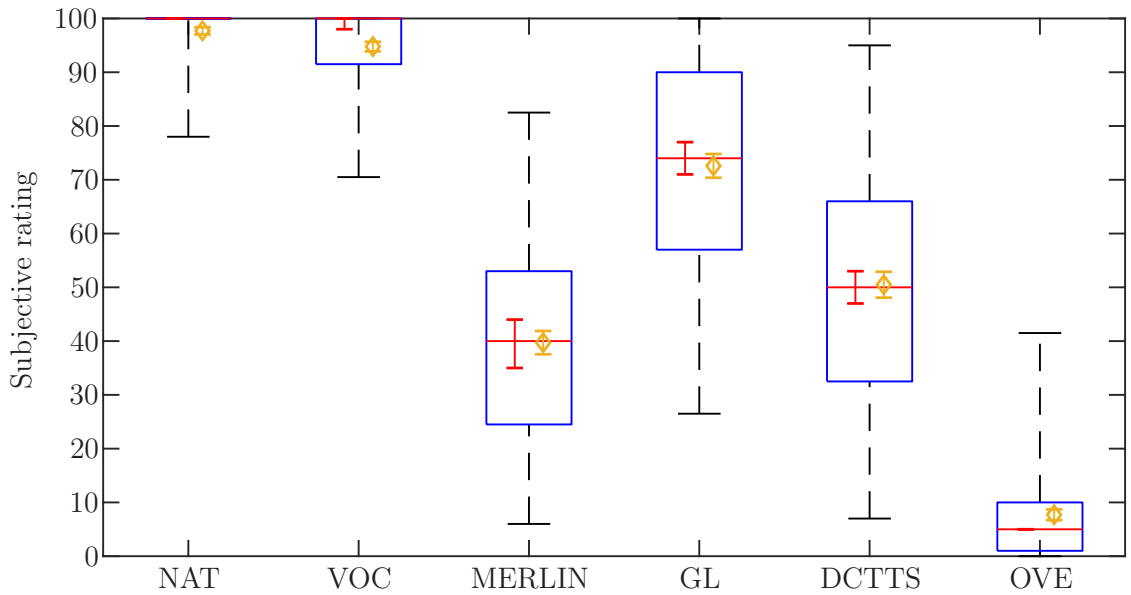
Xu, Y. and Prom-On, S. (2014). Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning. *Speech Commun.*, 57:181–208.

A unifying view



- Capital letters are speech representations
- Horizontal arrows are transformations between them
- Vertical arrows are controllable manipulations

MUSHRA results from pre-study



- The test used 12 listeners and 30 Harvard sentences
- DCTTS used a simpler fine-tuning approach yielding greater acoustic quality but more mispronunciations

Lexical decision task results from pre-study

System	Est. effect	*p*-value	Incorrect
NAT (ref.)			3%
GL	0.02	n.s.	3%
VOC	0.002	n.s.	4%
DCTTS	0.06	$<$0.05	9%
MERLIN	-0.004	n.s.	4%
OVE	0.06	$<$0.005	7%

- 14 listeners with 300 responses and reaction times each
- DCTTS performed significantly worse due to mispronunciations

Example stimuli

System	HVD	MRT 1	MRT 2
NAT	Old, New	Old, New	Old, New
VOC	Old, New	Old, New	Old, New
MERLIN	Old, New	Old, New	Old, New
GL	Old, New	Old, New	Old, New
DCTTS	Old, New	Old, New	Old, New
OVE	Old, New	Old, New	Old, New

- Old = Stimulus from pre-study
- New = Stimulus from main study reported in Malisz et al. (2019)