Comprehension of synthetic speech with and without natural prosody

Mirjam Wester, Oliver Watts, Gustav Eje Henter

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

Motivation

- Develop more ecologically-valid evaluation techniques that go beyond isolated sentences and measure *comprehension* of synthetic speech.
- The effect of prosody on comprehension is not measured effectively by either MOS or SUS.
- Appropriate evaluation of 'found data'.

Prior work

- 80s & 90s research: post-perceptual measures often did *not* show significant differences in comprehension between synthetic and natural speech [1].
- Online methods did show perceptual difficulties in interpreting high-quality synthetic speech which disappear by the time the entire comprehension process has run its course [2].
- But, online methods generally use sentence-level materials which have been carefully constructed.
- Evaluation techniques that are suitable for found data need to be able to evaluate longer stretches of speech, e.g., dialogues or stories.
- We revisit the post-perceptual approaches to measuring comprehension, arguing that we may be successful this time round as our data is:
  - prosodically rich; it comprises interesting and engaging interviews with comedians,
  - 10 minutes long for each interview,
  - tested using multiple choice questions where the participants are required to recall exact wording or detailed information about the speech content, thus not additionally relying on real-world knowledge.

Experimental set-up

- Speech types: natural (N), synthetic (S) and synthetic-modified (M)
- 3 interviews, 3 speech types (6 different orderings for each)
- 36 listeners for fully balanced design
  - 20 multiple choice questions per interview
  - order of questions and response options were randomised
  - questionnaire after listening:
    - how familiar with Desert Island Discs (DID)
    - how familiar with speakers
    - how difficult were questions
    - any other comments/observations

Speech material

a BBC Radio 4 programme.

Kirsty Young (KY) interviews guests about the eight records they would take to a desert island.

Episodes selected for the evaluation:

David Walliams (DW)

Victoria Wood (VW)

Steve Coogan (SC)

Synthesis

Three DNN-based synthesisers were trained:

	KY	DW & SC	VW
Accent	Scottish F	British M	British F
Data (mins)	238	64	96

- Transcripts passed through TTS front-ends and annotation used in two ways:
  - Completely synthetic (S); front end's predictions of sentence-internal pauses were used directly, durations and acoustic features predicted with the two DNNs for each voice.
  - Duration-modified (M); segmental durations and placement of sentence-internal pauses taken from forced aligned on test data.

Results

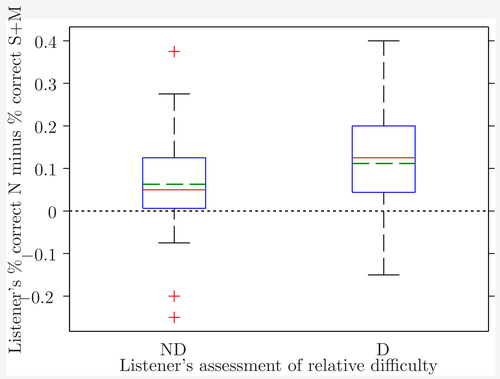
Number and fraction of correct responses across speech speech types (columns) and interviews (rows).

	N	S	M	All types
DW	164/240 68%	181/240 75%	134/240 56%	479/720 67%
SC	176/240 73%	144/240 60%	147/240 61%	467/720 65%
VW	190/240 79%	181/240 75%	157/240 65%	528/720 73%
All int.	530/720 74%	506/720 70%	438/720 61%	1474/2160 68%

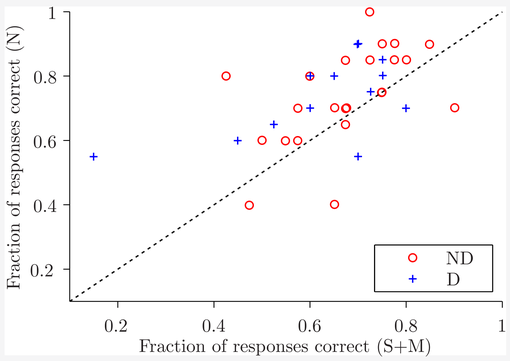
Differences in overall rate of correct response between speech types.

Comparison	N vs. S	N vs. M	S vs. M
Difference	3.3%	13%	9.4%
p-value*	0.18	$< 10^{-6}$	$< 10^{-3}$

*Holm-Bonferroni adjusted



Performance differences between natural and synthetic speech (N easier than S) for listeners in no-difference (ND) and difference (D) groups.



Listeners' rate of correct response on natural and synthetic speech.

Discussion

- In both DW and SC the same synthetic voice was used. Why the differences in scores???

	N	S	M
DW	68%	75%	56%
SC	73%	60%	61%

- Although large in magnitude most of the differences between N and S not significant due to the small sample sizes prior to pooling.
- We expected modified to fall between natural and synthetic due to the natural durations.
- Mismatch training – test data:
  - Acoustic models learned on carefully paced read-speech training data may not produce highly intelligible or comprehensible speech when shoe-horned into the spurt-like duration structure of interview speech.
  - Overlapping speech and laughter tend to have detrimental effect on automatic alignment.

Conclusion

- Overall subjects perform significantly worse on modified synthesis than on natural or synthetic speech.
- Many participants said synthetic speech was more difficult to focus on but the task was do-able.
- However, a couple of listeners pointed out the synthetic speech was nauseating.
- Post-perceptual test not sensitive enough to identify comprehensibility differences, even when using prosodically rich conversational material.
- How to evaluate a voice built on data like DID?
- How should audio books or conversational voices be evaluated?

References

[1] S. Duffy and D. Pisoni (1992) Comprehension of synthetic speech produced by rule: A review and theoretical interpretation *Language and Speech*, vol. 35, no. 4, pp. 351–389.

[2] S. Winters and D. Pisoni (2005) Speech synthesis: Perception and comprehension. In: Brown, K., (ed), *Encyclopedia of Language and Linguistics* volume 12. 31–49.

Acknowledgements

This research was supported by the EPSRC under Programme Grant EP/I031022/1 (Natural Speech Technology).