Testing the Consistency Assumption

Pronunciation Variant Forced Alignment in Read and Spontaneous Speech Synthesis

Rasmus Dall, Centre for Speech Technology Research, University of Edinburgh

ICASSP 24/3-2016

Collaborators

Thanks to all collaborators:

Sandrine Brognaux (Universite de Mons/Universite Catholique de Louvain, Belgium)
Korin Richmond (CSTR)
Cassia Valentini Botinhao (CSTR)
Gustav Eje Henter (CSTR)
Julia Hirschberg (Columbia University, USA)
Junichi Yamagishi (CSTR/National Institute of Informatics Tokyo, Japan)
Simon King (CSTR)

Motivation

- Earlier research [1] has found that using manually aligned data for both training and synthesis improves quality.
- This may be due to:
    - Better phonemisation/alignment at training time
    - Better phonemisation at synthesis time
    - Both
- This work focuses on producing a better phonemisation/alignment at training time.
- Tests the “Consistency Assumption”

Consistency Assumption

"Phoneme identity errors made by the forced aligner are compensated for by making the same errors at synthesis time."

- It is often debated whether this is true.
  - Some prefer pronunciation variation in alignment (inconsistent)
  - Others not (consistent)
- So does this assumption hold?
  - Does it for (more difficult) spontaneous speech?

Consistency Assumption

We have the dog here

Standard Training:

$$sil \rightarrow w\ i \rightarrow sp \rightarrow h\ a\ v \rightarrow sp \rightarrow D\ i \rightarrow sp \rightarrow d\ Q\ g \rightarrow sp \rightarrow h\ I@\ r \rightarrow sil$$

Synthesis:

$$sil \rightarrow w\ i \rightarrow h\ a\ v \rightarrow sil \rightarrow D\ i \rightarrow d\ Q\ g \rightarrow h\ I@\ r \rightarrow sil$$

Consistency Assumption

We have the dog here

Variant Training:

$$sil \rightarrow w\ i \rightarrow sp \rightarrow h\ a\ v \rightarrow sp \rightarrow D\ i \rightarrow sp \rightarrow d\ Q\ g \rightarrow sp \rightarrow h\ I@\ r \rightarrow sil$$

$$w\ I \qquad h\ @\ v \qquad D\ @$$

$$@\ v$$

Synthesis:

$$sil \rightarrow w\ i \rightarrow h\ a\ v \rightarrow sil \rightarrow D\ i \rightarrow d\ Q\ g \rightarrow h\ I@\ r \rightarrow sil$$

Consistency Assumption

We have the dog here

Variant Training:

$$sil \rightarrow w\ i \rightarrow sp \rightarrow h\ a\ v \rightarrow sp \rightarrow D\ i \rightarrow sp \rightarrow d\ Q\ g \rightarrow sp \rightarrow h\ I@\ r \rightarrow sil$$

$w\ I$ $\quad$ $h\ @\ v$ $\quad$ $D\ @$

$@\ v$

Synthesis:

$$sil \rightarrow w\ i \rightarrow h\ a\ v \rightarrow sil \rightarrow D\ i \rightarrow d\ Q\ g \rightarrow h\ I@\ r \rightarrow sil$$

Never changes!

Corpora

Training Corpora:

- Two Corpora of approximately 1h/1100 sentences at 48khz, 16 bit.
- “Read” speech
  - Arctic prompts
- “Spontaneous” speech
  - Recorded in the same studio as the read prompts
  - Free conversation with voice talent with webcam view to facilitate natural conversation
  - Orthographically transcribed
- Both corpora from same British English female speaker.

Corpora

Development Corpus:

- Small corpus of 50 read and 50 spontaneous sentences with same content.
    - Only differing in realisation, either spontaneously uttered or recorded as prompt
    - Same set as in [2]
- Transcribed at phoneme level by two annotators
    - Corrected output of standard multisyn forced alignment
    - Corrected for phoneme identity not boundary!
    - Met and agreed on Gold standard

Transcription Accuracy

Phoneme accuracy when compared to Gold standard:

	Del	Add	Sub	Total	PER
Read					
Automatic	149	10	151	310	19.1%
Annotator 1	33	30	69	132	8.1%
Annotator 2	3	9	36	48	3.0%
Spontaneous					
Automatic	202	17	180	399	25.2%
Annotator 1	11	15	42	68	4.3%
Annotator 2	4	15	18	37	2.3%

Pronunciation Variant Alignment

Implemented method for pronunciation variant forced alignment.

Used multisyn forced alignment tools.

- Standard method
    - Monophoneme mixture models (8 mixes)
    - Power normalisation
    - Silence trimming (>0.5s)
    - Short pause modelling
    - Combilex dictionary
    - Festival as front-end

Pronunciation Variant Alignment

Variant systems introduced lattice decoding at short pause modelling stage

Two sources of information:

- Manual context rules based on observation of speaker pattern
  - e.g. “Any end of word stop can deleted”
- Dictionary encoded variants (from Combilex)
  - ("or" (cc full) (((O r) 1)))
  - ("or" (cc reduced) (((@ r) 0)))
- Also combined the two

Pronunciation Variant Alignment

- These were run on each type of speech.

	Del	Add	Sub	Total	PER
Read					
Standard	10	149	151	310	19.1%
Lattice w. Combilex	6	139	184	329	20.2%
Lattice w. Rules	20	106	120	246	15.2%
Lattice w. Both	22	101	142	265	16.3%
Spontaneous					
Standard	17	202	180	399	25.2%
Lattice w. Combilex	9	178	199	386	24.4%
Lattice w. Rules	37	133	134	304	19.2%
Lattice w. Both	38	130	145	313	19.7%

Pronunciation Variant Alignment

- These were run on each type of speech.

	Del	Add	Sub	Total	PER
Read					
Standard	10	149	151	310	19.1%
Lattice w. Combilex	6	139	184	329	20.2%
Lattice w. Rules	20	106	120	246	15.2%
Lattice w. Both	22	101	142	265	16.3%
Spontaneous					
Standard	17	202	180	399	25.2%
Lattice w. Combilex	9	178	199	386	24.4%
Lattice w. Rules	37	133	134	304	19.2%
Lattice w. Both	38	130	145	313	19.7%

Transcriber Issues

- Starting point influences annotators [3]
- Previous transcribers started from standard system output
  - Skewed toward standard output
- To see this effect we got a third transcriber in
  - Started from Both system output
  - Should be skewed toward Both output

Transcriber Issues

- System accuracy per Annotator:

	A1	A2	A3	Gold
Read				
Standard	17.3%	19.2%	22.6%	19.1%
Lattice w. Combilex	20.1%	20.2%	16.9%	20.2%
Lattice w. Rules	15.7%	15.2%	13.9%	15.2%
Lattice w. Both	16.8%	16.7%	9.1%	16.3%
Spontaneous				
Standard	23.0%	25.7%	32.0%	25.2%
Lattice w. Combilex	23.4%	25.1%	26.1%	24.4%
Lattice w. Rules	18.0%	19.9%	20.8%	19.2%
Lattice w. Both	18.6%	20.8%	16.5%	19.7%

Transcriber Issues

- 3rd transcriber with outset in Both system:

	A1	A2	A3	Gold
Read				
Standard	17.3%	19.2%	22.6%	19.1%
Lattice w. Combilex	20.1%	20.2%	16.9%	20.2%
Lattice w. Rules	15.7%	15.2%	13.9%	15.2%
Lattice w. Both	16.8%	16.7%	9.1%	16.3%
Spontaneous				
Standard	23.0%	25.7%	32.0%	25.2%
Lattice w. Combilex	23.4%	25.1%	26.1%	24.4%
Lattice w. Rules	18.0%	19.9%	20.8%	19.2%
Lattice w. Both	18.6%	20.8%	16.5%	19.7%

Transcriber Issues

- Combilex version IS helpful:

	A1	A2	A3	Gold
Read				
Standard	17.3%	19.2%	22.6%	19.1%
Lattice w. Combilex	20.1%	20.2%	16.9%	20.2%
Lattice w. Rules	15.7%	15.2%	13.9%	15.2%
Lattice w. Both	16.8%	16.7%	9.1%	16.3%
Spontaneous				
Standard	23.0%	25.7%	32.0%	25.2%
Lattice w. Combilex	23.4%	25.1%	26.1%	24.4%
Lattice w. Rules	18.0%	19.9%	20.8%	19.2%
Lattice w. Both	18.6%	20.8%	16.5%	19.7%

Voice Testing

- We have improvement in alignment accuracy, does it help TTS quality?
- Trained HTS voices on each alignment using each speech type
- 30 sentences split into two groups of 15
    - Subset of the 50 dev sentences
    - Included natural read and spontaneous sentences
- 30 participants
    - Each rated one of the two groups of 15 sentences
- MUSHRA-style listening test
    - Side-by-side comparison on 100-point sliding scale

Voice Testing

Too many systems (8) to play samples here, so:

http://dx.doi.org/10.7488/ds/1314

MUSHRA-style Test

R = Read

S = Spontaneous

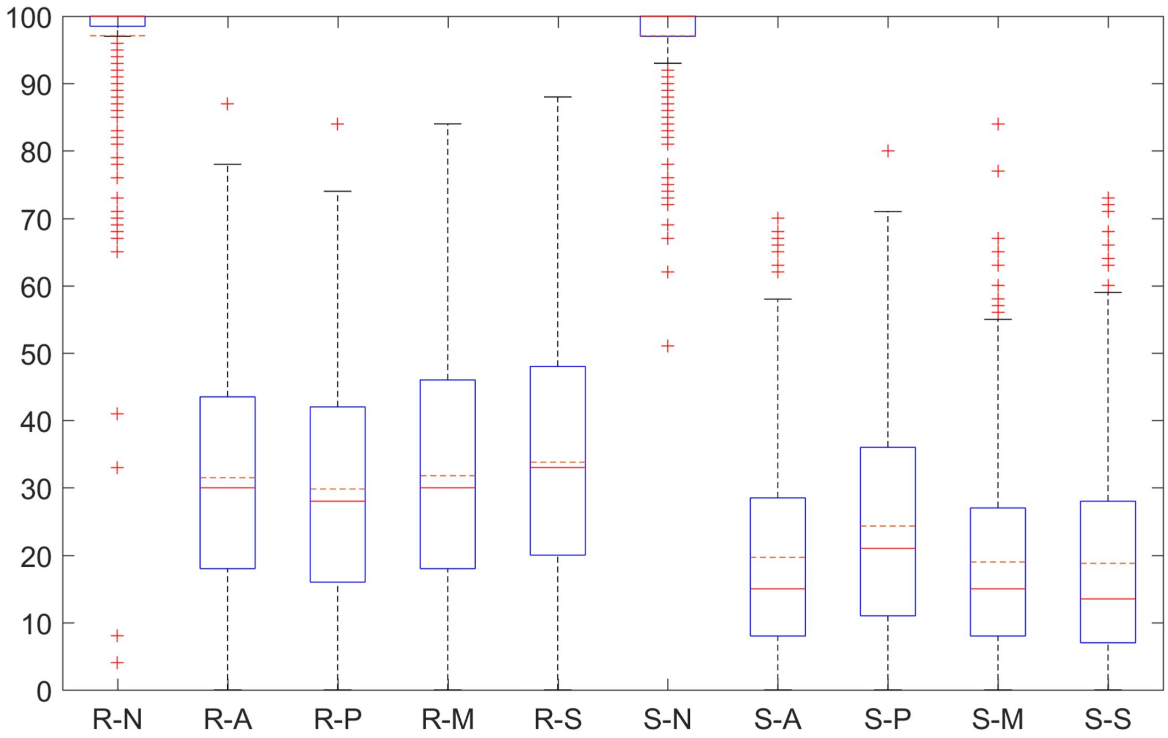
N = Natural

A = Both

P = Combilex

M = Manual

S = Standard



MUSHRA-style Test

R = Read

S = Spontaneous

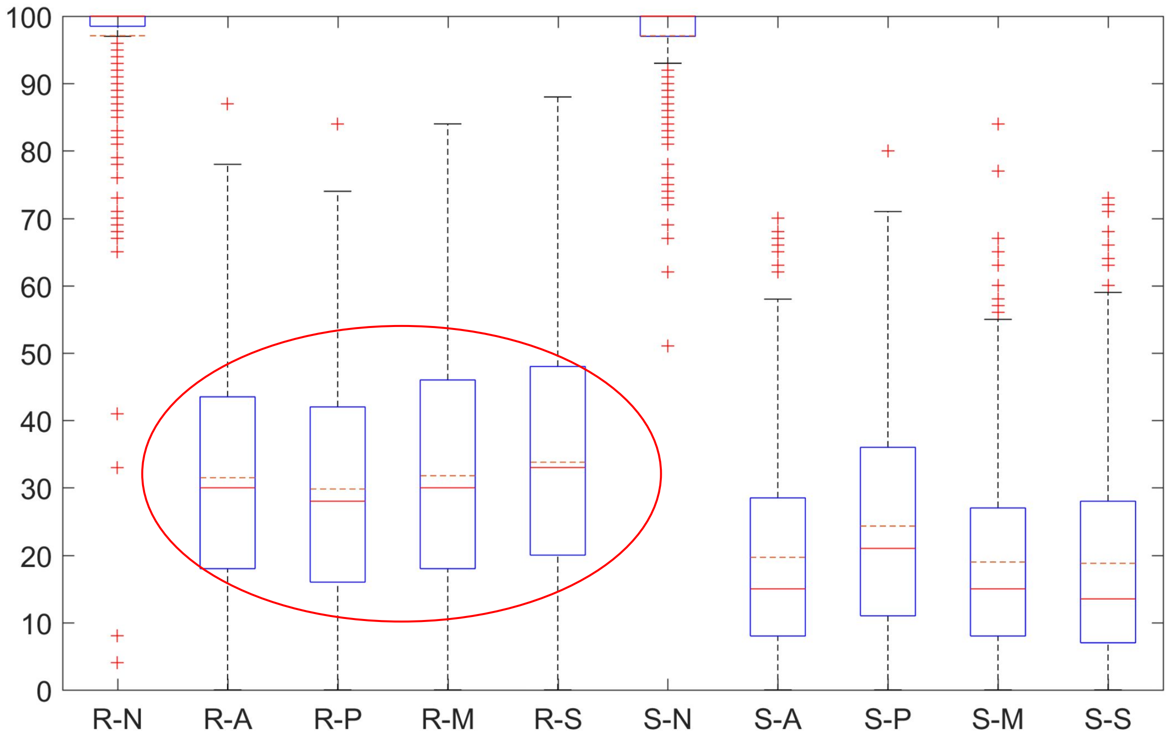
N = Natural

A = Both

P = Combilex

M = Manual

S = Standard



Hyper-articulation?

- The improved alignment did not help Read speech in the test
- But if we listen to some samples of the “worst” system:

Standard Combilex

Standard Combilex

- We can hear that we are producing hyper-articulated sentences
- Arguably what we are asking for at synthesis time

Spontaneous Speech

R = Read

S = Spontaneous

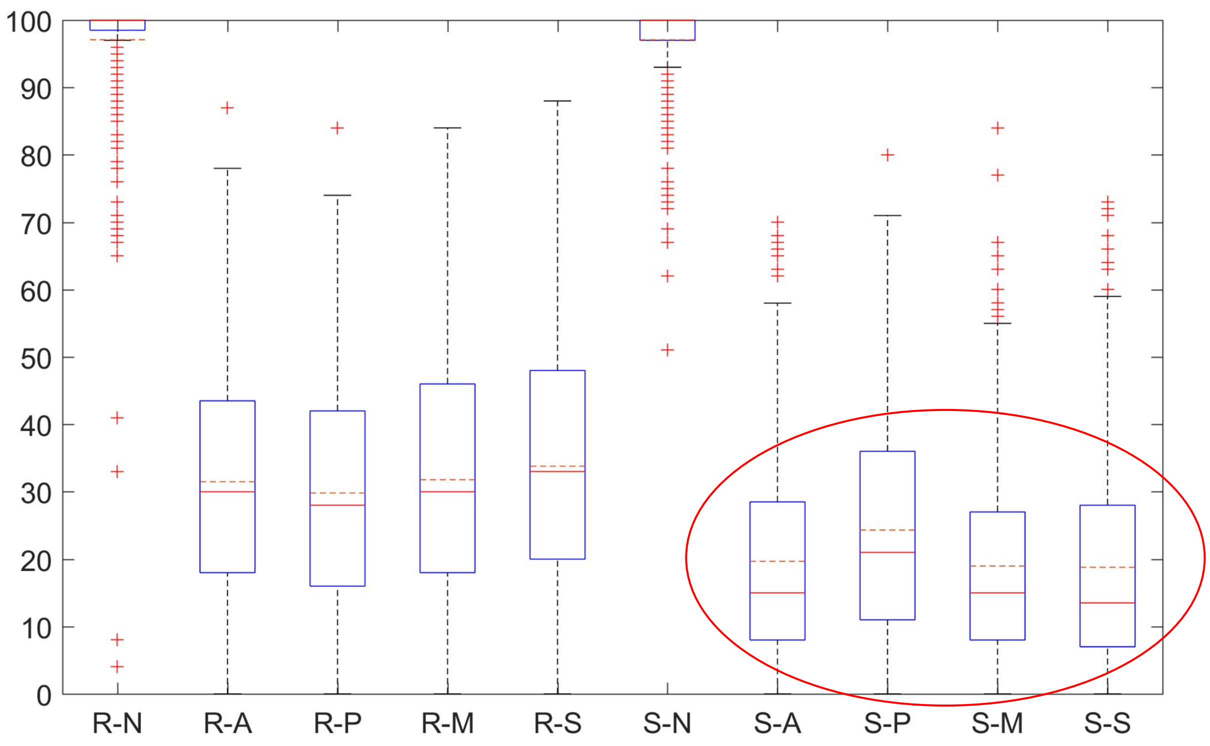
N = Natural

A = Both

P = Combilex

M = Manual

S = Standard



Spontaneous Speech

- Some variation (combilex) in training seems beneficial
  - Neither the most consistent nor the most accuracte
- Too much (manual rules) seems to become too inconsistent with synthesis phonemisation
  - Albeit it helps alignment accuracy
- No variation (standard) too inaccurate
  - Although it retain consistency across training and synthesis

Conclusions

- Pronunciation variant forced alignment improves phoneme accuracy
    - Using both manual rules and combilex derived variants the best
- The consistency assumption seems to hold for Read speech
- But not in Spontaneous speech
    - Likely too different from actual realisation
- Being inconsistent in a “consistent” manner is helpful
    - Perhaps we can come up with ideas to retain consistency while using better alignments?

References

[1] Brogneaux, S., Picart, B., Drugmann, T. & Louvain, D. (2014). Speech synthesis in various communicative situations: Impact of pronunciation variations. In *Proc. Interspeech,* Singapore, Singapore.

[2] Dall, R., Yamagishi, J. & King, S. (2014). Rating Naturalness in Speech Synthesis: The Effect of Style and Expectation. In *Proc. Speech Prosody,* Dublin, Ireland.

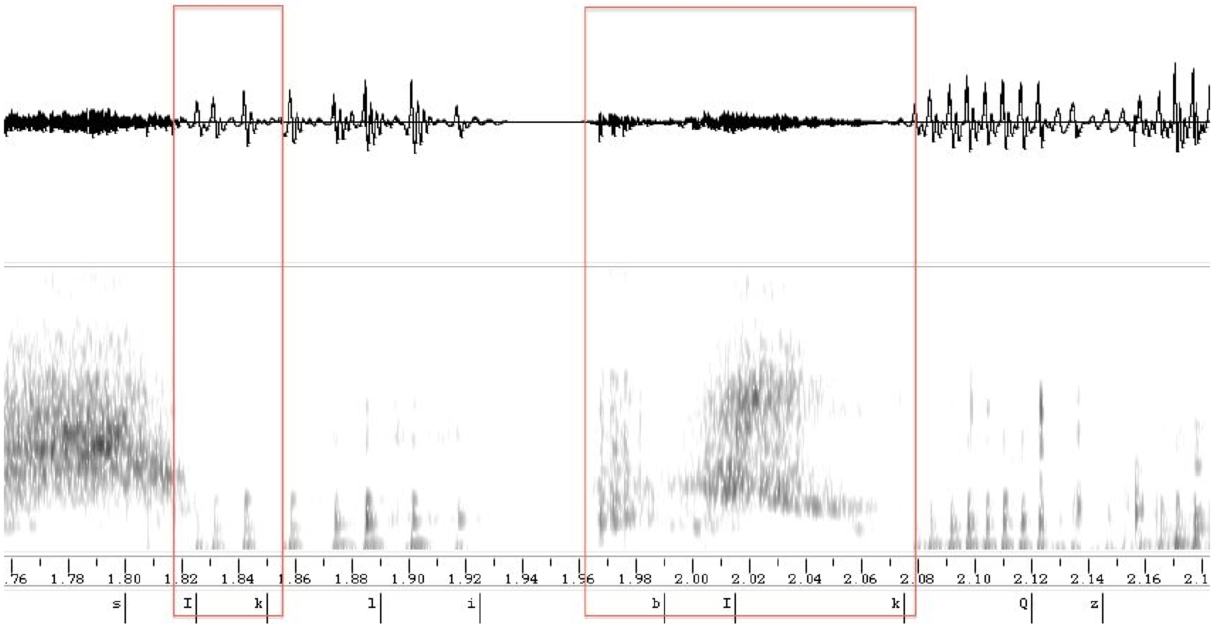
[3] Van Bael, C. (2007). Validation, Automatic Generation and Use of Broad Phonetic Transcriptions. *PhD Thesis,* Radboud University Nijmegen.

Questions?

Thanks for listening - Questions?

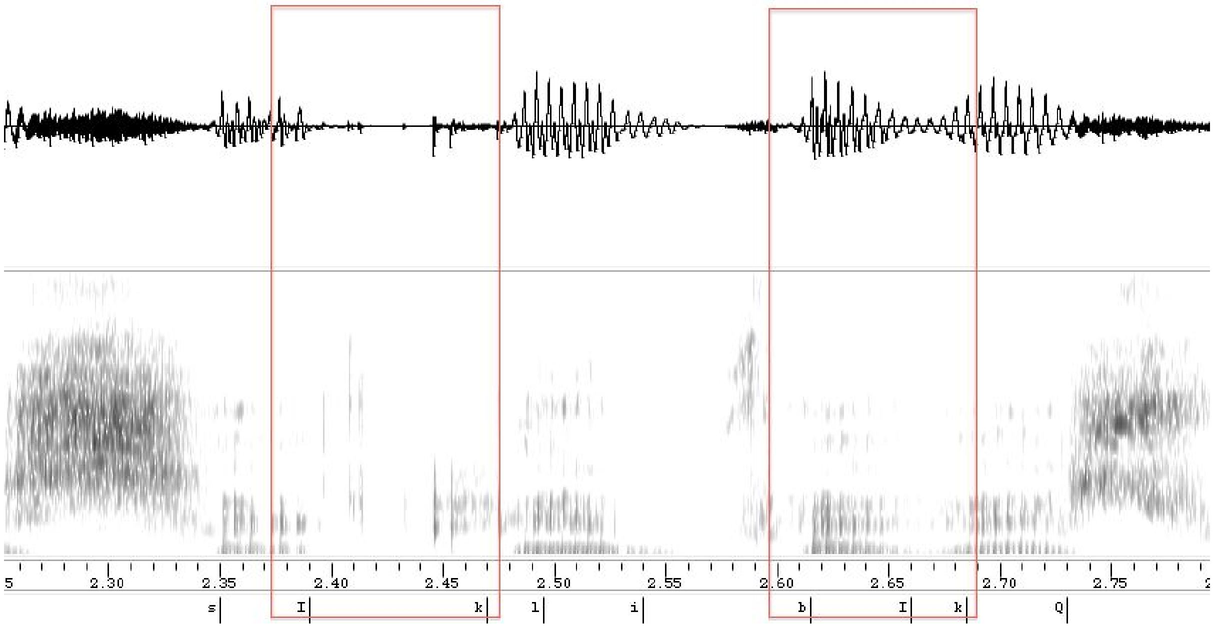
Transcription Accuracy

Spontaneous speech makes cascading errors



Transcription Accuracy

Not present in the Read speech



Predicting Pronunciation Variation

Notice what happens if we improve the alignment AND keep the consistency:

Standard vs Improved Inconsistent vs Improved Consistent

Predicting Pronunciation Variation

Two approaches so far:

- Word based language model to determine word reduction.
  - Based on [15] this should work.
- Phoneme based language model to determine pronunciation variant.
  - Use training data alignment for LM.
  - Retains consistency!
- As this is brand new I can only play you samples of word LM:

From Alignment vs No Reduction vs Half Reduction vs Full Reduction