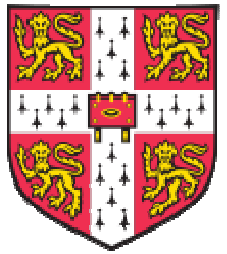


Automatic and Human Filled Pause Insertion for Speech Synthesis

Marcus Tomalin, Rasmus Dall,
Mirjam Wester, Bill Byrne, & Simon King

NST Meeting, May 20th 2014

Filled Pauses and Speech Synthesis

Filled Pauses (FPs) pervasive in conversational speech

- *I'm getting a bit uh specific here*

They serve structural and psychological purposes:

- Planning utterances
- Discouraging interruptions
- Highlighting discourse-structure

FPs rarely modelled overtly in Speech Synthesis (SS) systems:

- fluent sentence:
- disfluent sentence:

Therefore FPs usually sound 'unnatural' in SS outputs

Task: to develop convincingly disfluent SS systems

Experiment 1: Human FP Insertion

Do humans agree on where to insert FPs?

Insert FPs at possible Insertion Points (IPs):

- IP1 *I'm* IP2 *getting* IP3 *a* IP4 *bit* IP5 *specific* IP6 *here* IP7
- if IP5 ➔ *uh, then I'm getting a bit uh specific here*

Can human FPs provide 'gold standard' for automatic FPs?

Experimental framework:

- Sentences from 2 corpora (at least 1 FP, and at most 5):
  - **WSJ**: 30 sentences; fluent text; 50% isolated, 50% in paragraph
  - **AMI**: 30 sentences; disfluent speech; 50% isolated, 50% in paragraph
- 72 native speakers inserted uh and um
- 60 sentences divided into 2 sets (equal text/sentence type)
- Each subject processed 1 set (sentences in random order)
- Each subject inserted at least 1 FP, but more if natural

Experiment 1: Human FP Insertion

Cond	Pos	Used	Ins	Top	Top 3
WSJ	12.77	80.68%	1.40	28.07%	59.14%
AMI	16.48	75.73%	1.67	24.86%	54.19%
Isolated	14.59	77.54%	1.51	26.78%	58.17%
Paragraph	14.60	78.31%	1.55	25.98%	54.91%
All	14.59	77.93%	1.53	26.35%	56.49%
Chance	14.59	97.23%	1.53	9.54%	28.61%

- FP patterns similar for news/speech, isolated/paragraph
- Top IP is 26.35% of all insertions (chance = 9.54%)
- Top 3 IPs are 56.49% of all insertions (chance = 28.61%)
- 96.6% of original AMI sentences had an FP in one of Top 3 IPs

This suggests:

- good consistency between subjects' predictions and actual usage
- good consistency between different subjects' predictions

Experiment 2: FP Perception

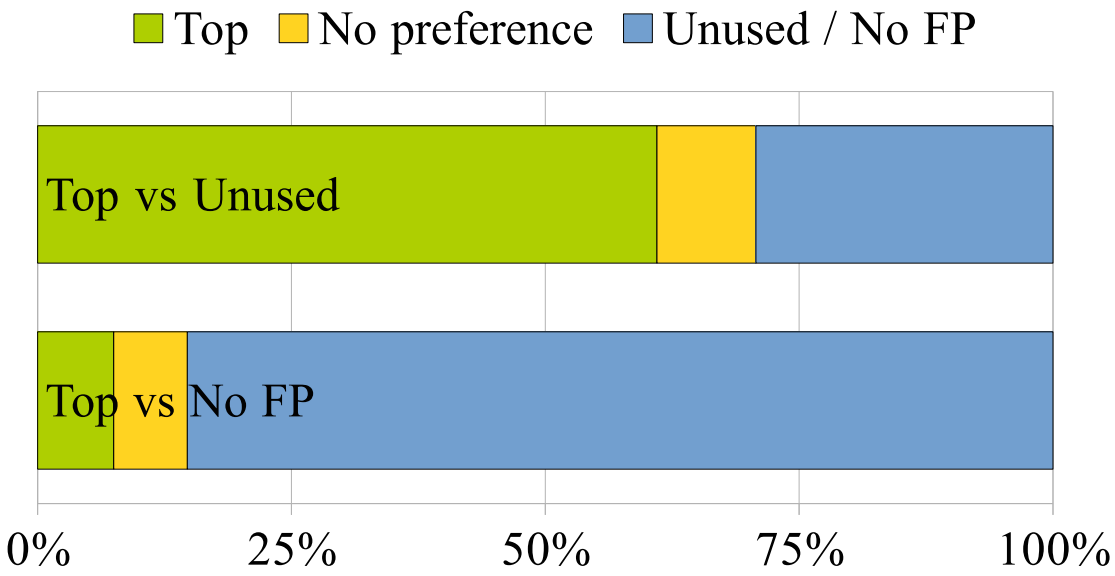
Does chosen IP make a perceptual difference?

- are there IPs where FPs should not be inserted?
- are FPs in well-placed IPs preferred over no FPs?

Experimental framework:

- 20 of the 30 **AMI** sentences used
- uh inserted either
  - at most likely place (using data from Experiment 1)
  - or randomly in an unused IP
- a synthetic female voice used (HTS 2)
- HMM-synthesis; c.8 hours of read speech.
- good synthetic uh used for high-quality SS
- 20 Amazon Mechanical Turk native speakers (‘master worker’ status)
- 20 sentence pairs (random order); had to indicate preference (if any)
- FP in Top position vs either (i) random position or (ii) no FP

Experiment 2: FP Perception



- Listeners preferred FPs inserted in Top IP (61%)
  - there are IPs where FPs are not preferred by listeners
- Listeners favoured no FP rather than an FP
  - most likely they conflated ‘natural’ with ‘fluent’
  - previous work has shown that FPs facilitate comprehension (Dall et al 2014)

Automatic FP Insertion

Training and test data:

- 1M sentences (19M words) training data from **AMI**, **Fisher**, **SWB**
- dev and test sets
  - 2k words each; 50% with FPs, 50% without FPs; from same corpora
- uh and um mapped to a single type (uh)

The systems built to insert FPs automatically:

- *Random*: randomly inserts uh into IP
- *Ngram*: a 4gram
  - SRILM; 68K wordlist; KN-discounting
- *RNN LM*: Recurrent Neural Network
  - 500 neurons; 250 classes
- *Interpolated RNN LM and Ngram*:
  - interpolated on a word-by-word basis

Automatic FP Insertion

System	Precision	Recall	All F	UH F
Ngram				
dev	0.49	0.15	0.23	0.26
test	0.48	0.16	0.24	0.27
RNN LM				
dev	0.31	**0.51**	0.39	0.51
test	0.32	**0.52**	0.40	0.53
RNN LM+Ngram				
dev	**0.53**	**0.51**	**0.52**	**0.57**
test	**0.50**	0.47	**0.48**	**0.54**
Random				
dev	0.13	0.16	0.14	0.16
test	0.14	0.17	0.15	0.18

- All results for Top 3 output
- Best system is the RNN LM + Ngram (highest F-score)
- RNN LM has better Recall; Ngram better Precision
- RNN LM + Ngram better than chance for all metrics

Conclusions and Future Work

- Humans largely agree about FP insertion, regardless of data type (i.e., news, spontaneous speech)
- FP insertion not random, and therefore can be modelled automatically
- RNN LM + Ngram performed best across a range of metrics
- Need to improve modelling of FPs in SS systems
  - e.g., intonation, duration
- Need to model other types of disfluency
  - e.g., repetitions, restarts, revisions