# Inserting Filled Pauses and Discourse Markers for Disfluent Speech Synthesis

**M. Tomalin, R. Dall, M. Wester, X. Liu, B. Byrne, & S. King**

**Cambridge University Engineering Department, CSTR The University of Edinburgh**
**{mt126, xl207, wjb31}@eng.cam.ac.uk, {r.dall@sms, mwester@inf, Simon.King@}.ed.ac.uk**

## AIMS

- Create disfluent input text for a Speech Synthesis (SS) system
- Insert multiple Filled Pauses (FPs) and Discourse Markers (DMs) into fluent text using lattice-based rescoring framework
- Compare performance of Ngrams and Full-Output Recurrent Neural Network Language Models (f-RNNLMs)
- Assess performance of interpolated Ngram + f-RNNLM

## DISFLUENT SPEECH SYNTHESIS

Speech disfluencies (DISs) serve important purposes in human speech:
- indicate psychological/emotional states
- structure spoken discourse
- facilitate word recall
- improve object recognition

Automatically insert FPs and DMs into fluent text:
- text input for SS systems usually fluent
- DISs must be introduced for disfluent SS
- 2 FPs and 2 DMs modelled overtly
- FPs = **UH**, **UM**
- DMs = **I MEAN**, **YOU KNOW**
- c.20M words (1M sentences) of training data:
  - Switchboard, Fisher, and AMI
- these 4 DISs occurred most frequently in the training data

| | #occs [%] |
|---|---|
| **YOU KNOW** | 278,423 [1.4%] |
| **UH** | 213,924 [1.1%] |
| **UM** | 200,500 [1.0%] |
| **I MEAN** | 73,719 [0.4%] |

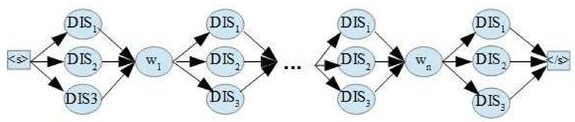
- **UH**, **UM**, and **YOU KNOW** occur comparably frequently (c.1% of all tokens in training data)
- **I MEAN** less frequent, so harder to model well
- DIS-insertion system built using training data
- it automatically inserts these 4 DISs into otherwise fluent texts

An example:
- **I NEVER LIKED GAMES → I MEAN I NEVER LIKED UH GAMES**

## THE DIS-INSERTION SYSTEM

The models used:
- Ngrams and f-RNNLMs built using training data
- f-RNNLM LMs:
  - have strong generalization performance
  - facilitate efficient training parallisation in Graphics Processing Unit (GPU) framework
  - improve WERs in speech recognition systems

The lattice-based framework:
- robust lattice-rescoring framework used
- initial lattices created
- each DIS accessible from each word token




- initial lattices expanded using Ngram (6g)
- lattices rescored using f-RNNLM, and Ngram + f-RNNLM
- $n$-best output for each sentence $S$ generated (where $n = 10000$)
- for fluent $S$ with $q$ tokens, $n$-best varies from $q$ to $(q \times 2) + 1$ tokens
- 1-best for each $p$-token $S$ selected using Ngram

Controlling the degree of disfluency:
- Disfluency Parameter (DP) set:
  - $0 \leq \text{DP} \leq 1$
  - $[0, 1]$ divided between 1-best outputs for all $p$
  - determines degree of disfluency in output

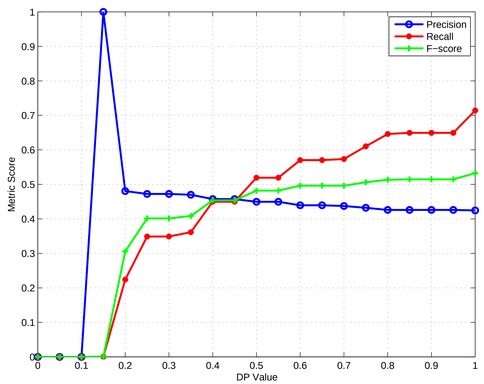
| DP | Output Sentence |
|---|---|
| 0.00 | WELL I GUESS THEY WERE SAYING |
| 0.25 | WELL I GUESS THEY WERE SAYING UM |
| 0.50 | UM WELL I MEAN I GUESS THEY WERE SAYING UM |
| 0.75 | UM WELL I MEAN I GUESS YOU KNOW THEY WERE SAYING UM |
| 1.00 | UM WELL I MEAN I GUESS YOU KNOW THEY UH WERE YOU KNOW SAYING UM |

- the higher the DP value the more disfluent the speaker
- final 1-best disfluent output generated for specified DP

## DIS-INSERTION RESULTS AND ANALYSIS

Dev and Test datasets defined:
- Dev and Test sets derived from same corpora as training data:
  - Dev = 5,310 sentences (126k words)
  - Test = 7,632 sentences (180K words)

Experiments and metrics:
- lattices rescored using Ngram, f-RNNLM, and Ngram + f-RNNLM
- component Ngram + f-RNNLM LMs had 50% / 50% weighting
- performance assessed using Precision, Recall, and F-score metrics




- metric scores for Dev data determined DP value used (0.5)
- inverse relationship between Precision and Recall/F-score

Comparison of the DIS-insertion systems:

| | Precision (Dev/Test) | | Recall (Dev/Test) | | F-score (Dev/Test) | |
|---|---|---|---|---|---|---|
| Ngram | 0.43 | 0.42 | 0.53 | 0.56 | 0.47 | 0.48 |
| f-RNNLM | 0.45 | 0.46 | 0.50 | 0.53 | 0.47 | 0.49 |
| Ngram + f-RNNLM | 0.45 | 0.45 | 0.52 | 0.54 | 0.48 | 0.49 |

- Ngram higher Recall than f-RNNLM; f-RNNLM higher Precision
- Ngram + f-RNNLM has complimentary properties of component LMs

Analysis for DIS subtypes:
- DIS-subtypes have different occs in Dev/Test ref/hyp data
- Dev/Test ref/hyp files compared for Ngram + f-RNNLM system

| | Dev occs ref/hyp [%] | | Test occs ref/hyp [%] | |
|---|---|---|---|---|
| UH | 3667 [2.9%] | 2662 [2.1%] | 3658 [2.0%] | 3339 [1.8%] |
| UM | 3338 [2.6%] | 3301 [2.6%] | 3391 [1.9%] | 4078 [2.2%] |
| I MEAN | 436 [0.3%] | 598 [0.5%] | 1239 [0.7%] | 983 [0.5%] |
| YOU KNOW | 1997 [1.6%] | 4141 [3.2%] | 4525 [2.5%] | 7457 [4.1%] |

- Ngram + f-RNNLM inserts **UH**, **UM**, and **I MEAN** proportionately
- Ngram + f-RNNLM overinserts **YOU KNOW** (1.6% abs)

Results for DIS subtypes:

| | Precision (Dev/Test) | | Recall (Dev/Test) | | F-score (Dev/Test) | |
|---|---|---|---|---|---|---|
| **UH** | 0.55 | 0.43 | 0.42 | 0.41 | 0.48 | 0.42 |
| **UM** | 0.58 | 0.43 | 0.53 | 0.51 | 0.55 | 0.47 |
| **I MEAN** | 0.11 | 0.23 | 0.16 | 0.16 | 0.13 | 0.21 |
| **YOU KNOW** | 0.35 | 0.50 | 0.76 | 0.80 | 0.48 | 0.62 |

- considerable DIS-specific variation
- Precision for FPs relatively stable; Recall varies up to 0.10% abs
- Precision for **YOU KNOW** varies by 0.15% abs
- Recall for **YOU KNOW** relatively stable
- metric scores for **I MEAN** lower than for the other DISs

## CONCLUSIONS

- A burgeoning interest in emotional / expressive SS
- Ngram + f-RNNLM most robust of DIS-insertion systems compared
- Explore other DIS modelling techniques
- Develop text processing for other disfluency subtypes (e.g., repetitions, restarts)
- Develop disfluent SS system that uses voices with different emotional states and personality types