Increase Apparent Public Speaking Fluency by Speech Augmentation

Sagnik **D**as, **N**isha **G**andhi, **T**ejas **N**aik, **R**oy **S**hilkrot
Human Interaction Lab, Stony Brook University, NY, USA

ICASSP 2019 · IEEE Signal Processing Society · Stony Brook University

Introduction & Motivation:

Speech disfluency generally comes in the form of long pauses, discourse markers, repeated words, phrases or sentences and fillers or filled pauses like uh and um. Approximately 6% of speech appears to be non-pause disfluency [1]. **Filled pauses or filler-words are the most common disfluency in any unrehearsed, impromptu speech** [2].

Contributions:

- Filler-word detection on acoustic features.
- Silence classification conditioned on previous speech segment.
- Disfluency repair scheme to aid speakers.

Disfluency Detection:

Filler-word segmentation works using a Convolutional Recurrent Neural Network (CRNN) [3].

- Frame level acoustic features (log mel & MFCCs) are fed into Conv-MaxPool-ReLU blocks.
- Output features are stacked and fed into multiple Gated Recurrent Units (GRUs).
- FC-Softmax layer gives frame level probability.

Silences are classified into fluent and disfluent class.

- Training: Each silence is padded with previous and next word utterances and MFCC features are extracted to train a binary classifier.
- Testing: Around each silence a fixed length time window is used. 0.8-1.0 secs. works pretty well.

Disfluency Repair:

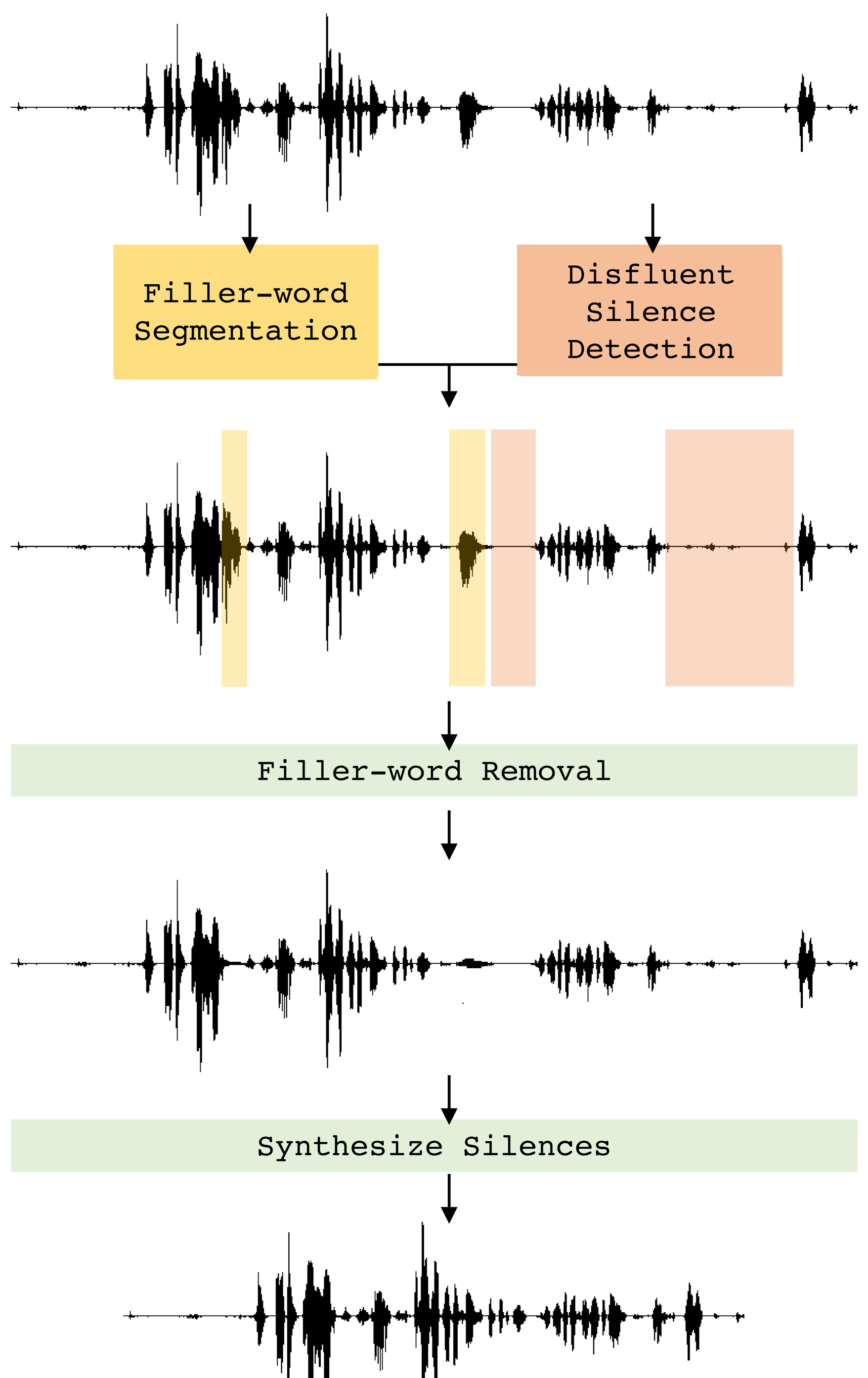
- All fluent silence durations are used to obtain a **histogram of fluent silences**.
- **Median of histogram bins** works well as optimal silence duration.
- Fillers and disfluent silences are **replaced with a silence of optimal duration**.

TL; DR

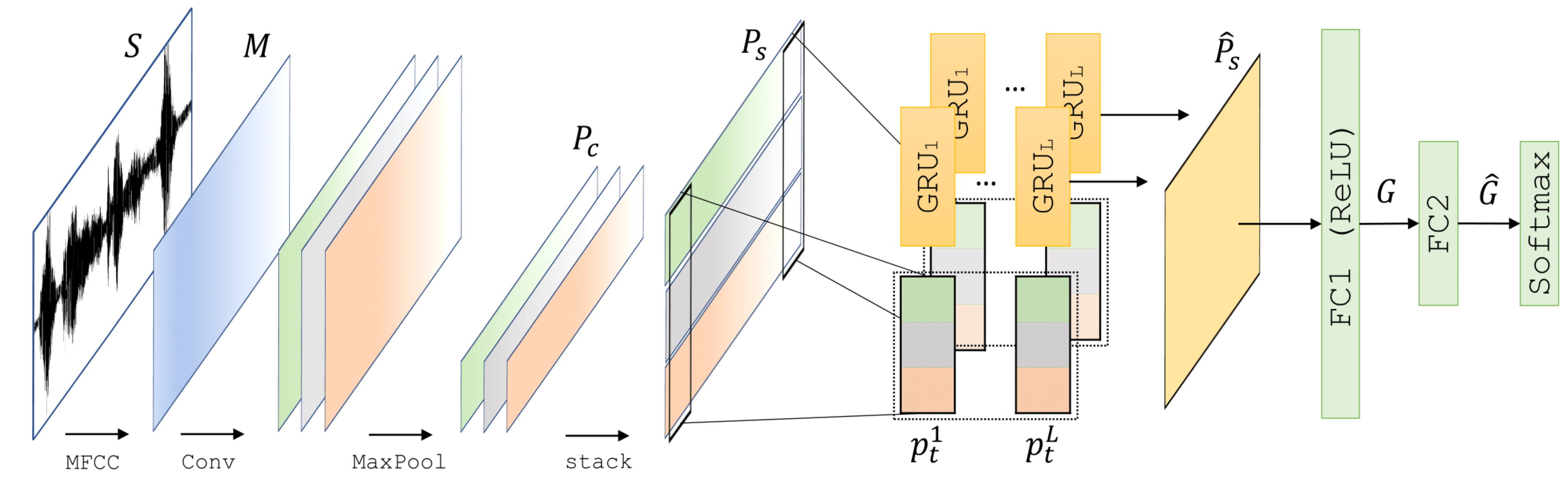
Everyone desires fluent & professional-like speech.

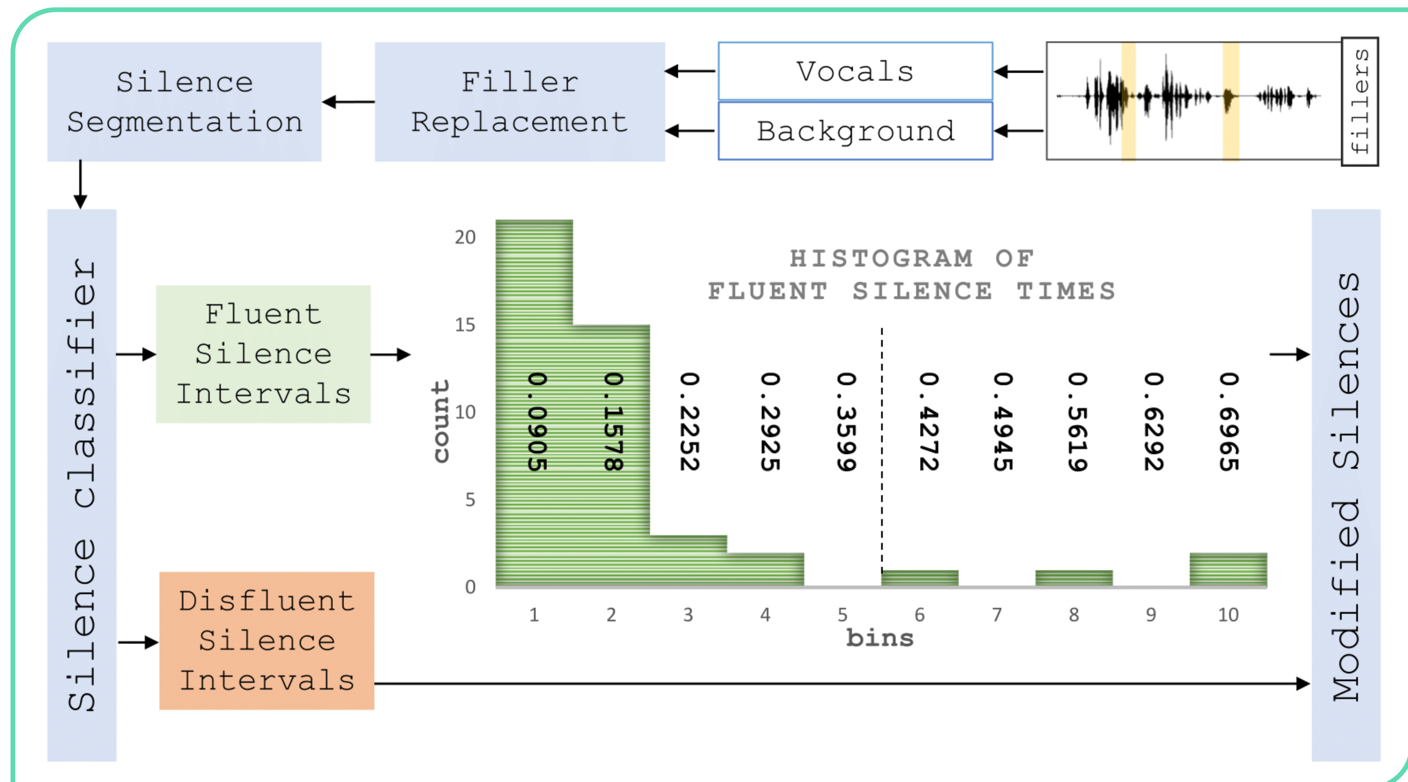
Our paper proposes a method to **detect** & **correct** common speech disfluencies, the **filler-words** (uh, umm), and **long pauses** to make the speaker more fluent.

We achieve this task by a learning based **filler word segmentation** & **silence classification** approach.



Some example outputs are available here!





Convolutional Recurrent Network (CRNN) based **filler-word segmentation**

Weakly-supervised silence classifier trained on fluent speech data. **Histogram based silence modification** allows to maintain speaker's pace.

Data:

- **TIMIT**: Silence classifier training. Silences are weakly-labeled using a probabilistic silence model [4] & disfluency detection model [5].
- **Switchboard I & II, AutoManner** : CRNN training and validation.

Experiments:

- **Validation of filler-word segmentation:**

Features	Precision	Recall	F1
MFCC	0.9482	0.9610	0.9534
Log Mel	**0.9495**	**0.9629**	**0.9550**

Method	Precision	Recall	F1
ASR	**0.9774**	**0.9792**	**0.9775**
CRNN	0.9495	0.9629	0.9550

- **Validation of silence classification:**

Method →	SVM (rbf)	Logistic Reg.	XGBoost
F1	**0.9774**	**0.9792**	**0.9775**

- **Disfluency repair quantitative Metrics** [6]: Speech rate (SR), Articulation rate (AR), Phonation-time ratio (PTR), Mean length of runs (MLR), Mean length of pauses (MLP) and Filled pauses per min. (FPM).

Metrics	SR ↑	AR ↑	PTR ↑	MLR ↑	MLP ↓	FPM ↓
Original	191.456	198.155	66.717	0.420	0.789	4.379
Proposed	206.465	208.437	**77.151**	0.495	**0.422**	1.813
+ASR	**206.710**	**208.770**	76.974	**0.504**	0.438	**1.608**

Future Works:

- Extension for other disfluencies.
- Generate silences instead replacing (GANs).

References:

1. Jean E Fox Tree, "The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech," Journal of memory and language, vol. 34, no. 6, pp. 709–738, 1995.
2. Kathryn Womack, Wilson McCoy, Cecilia Ovesdotter Alm, Cara Calvelli, Jeff B Pelz, Pengcheng Shi, and Anne Haake, "Disfluencies as extra-propositional indicators of cognitive processing," ACL, 2012.
3. Emre Cakır, Giambattista Parascandolo, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," arXiv preprint arXiv:1702.06286, 2017.
4. Guoguo Chen, Hainan Xu, Minhua Wu, Daniel Povey, and Sanjeev Khudanpur, "Pronunciation and silence probability modeling for asr," in Sixteenth Annual Conference of the International Speech Communication Association, 2015.
5. Vicky Zayats, Mari Ostendorf, and Hannaneh Hajishirzi, "Disfluency detection using a bidirectional lstm," arXiv preprint arXiv:1604.03209, 2016.
6. Judit Kormos and Mariann Denes, "Exploring measures and perceptions of fluency in the speech of second language learners," System, vol. 32, no. 2, pp. 145–164, 2004.