# Dual Language Models for Code Switched Speech Recognition

**Saurabh Garg, Tanmay Parekh, Preethi Jyothi**

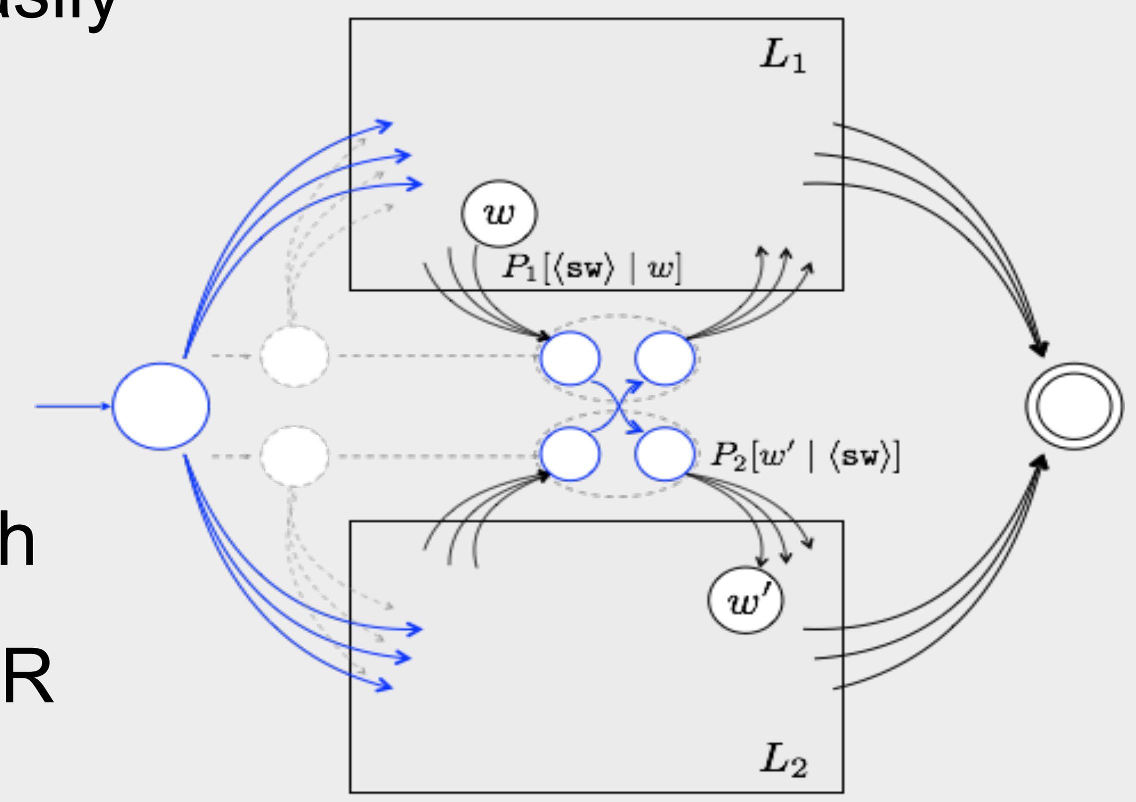
Department of Computer Science and Engineering, Indian Institute of Technology Bombay

## MAIN GOAL

- Code-switching is when speakers switch between multiple languages within a single utterance.
- Common phenomenon in multilingual societies.
- Limited availability of data poses challenges for building computational models for code-switched speech.
- **Objective:** How do we build better language models (LMs) for code-switched speech without the aid of external resources?

## APPROACH

- *Dual language models* (DLM): Combine two monolingual LMs and use a probabilistic model to switch between them
- DLMs are structured as a cooperative game between two players, each in charge of generating tokens in one of two languages:
  - Each player produces at least one token before switching or terminating
  - For simplicity, players do not retain any state on switching
- DLMs can be easily represented as finite-state machines and incorporated with the standard ASR pipeline




## EXPERIMENTS & RESULTS

- SEAME corpus of conversational Mandarin-English code-switched speech

| | Train | Dev | Test |
|---|---|---|---|
| # Speakers | 90 | 37 | 30 |
| Durations (hrs) | 56.6 | 18.5 | 18.7 |
| # Utterance | 54,020 | 19,976 | 19,784 |
| # Tokens | 539,185 | 195,551 | 196,462 |

- Perplexity on the dev/test sets using standard LMs and DLMs with different smoothing techniques

| Smoothing Technique | Dev | | Test | |
|---|---|---|---|---|
| | Mixed LM | DLM | Mixed LM | DLM |
| Good Turing | 338.3 | **329.2** | 384.5 | **371.1** |
| Kneser-Ney | 329.7 | **324.9** | 376.1 | **369.9** |

- Kneser-Ney smoothed dev/test set perplexities using varying amounts of training data

| Training data | Dev | | Test | |
|---|---|---|---|---|
| | Mixed LM | DLM | Mixed LM | DLM |
| Full | 329.7 | **324.9** | 376.1 | **369.9** |
| 1/2 | 362.1 | 350.6 | 400.6 | 389.8 |
| 1/3 | 368.6 | **356.0** | 408.6 | **394.2** |

ASR token error rates using DLMs & standard LMs

| ASR System | Data | Mixed LM | DLM | combined |
|---|---|---|---|---|
| SAT | Dev | 45.59 | 45.59 | **44.93** |
| | Test | 47.43 | 47.48 | **46.96** |
| TDNN + SAT | Dev | 35.20 | 35.26 | **34.91** |
| | Test | 37.42 | 37.35 | **37.17** |
| RNNLM Rescoring | Dev | 34.21 | 34.11 | **33.85** |
| | Test | 36.64 | 36.52 | **36.37** |

Token error rates with ½ training data

| ASR System | Data | Mixed LM | DLM | combined |
|---|---|---|---|---|
| SAT | Dev | 48.48 | 48.17 | **47.67** |
| | Test | 49.07 | 49.04 | **48.52** |
| TDNN + SAT | Dev | 40.59 | 40.48 | **40.12** |
| | Test | 41.34 | 41.32 | **41.13** |
| RNNLM Rescoring | Dev | 40.20 | 40.09 | **39.84** |
| | Test | 40.98 | 40.90 | **40.72** |

## OBSERVATIONS

- **Code-switching boundaries.** Code-switched bigrams with counts of ≤ 10 occupy 87.5% of the total number of code-switched bigrams in the training data (of which 55% are singletons)
- **Illustrative examples.**

| Sentence | Mixed LM perplexity | DLM perplexity |
|---|---|---|
| 我们 的 total 是 五十七 | 920.8 | 720.4 |
| 哦 我 没有 meeting 了 | 92.2 | 75.9 |

## SUBSEQUENT WORK

- Can we retain state when switching between languages? Can we use monolingual data to pretrain each individual monolingual LM?
  - Garg, et al. “Code-switched Language Models Using Dual RNNs and Same-Source Pretraining”, To appear in EMNLP 2018.