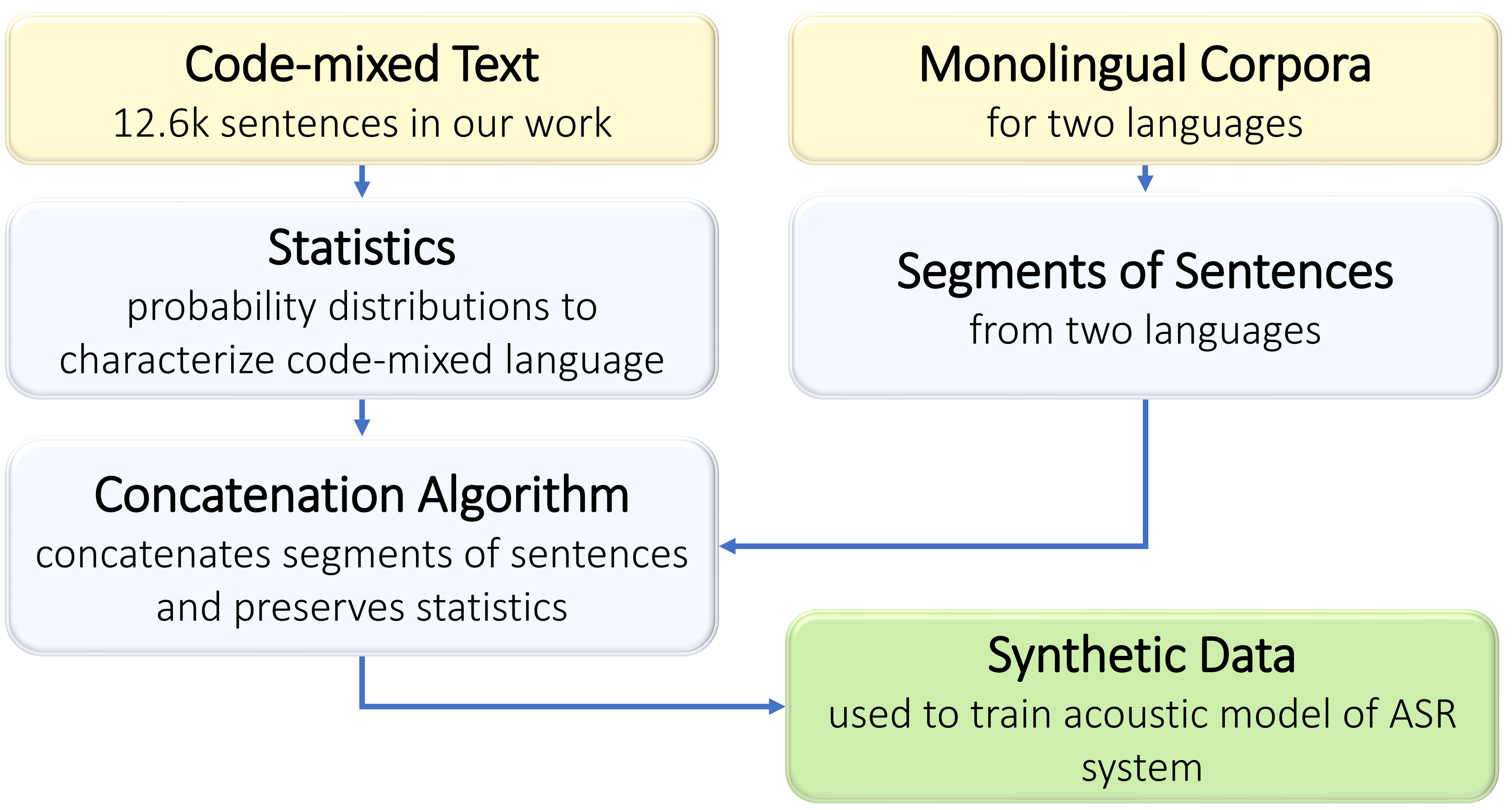
Exploiting Monolingual Speech Corpora for Code-mixed Speech Recognition

Karan Taneja[1] , Satarupa Guha[2] , Preethi Jyothi[1] , Basil Abraham[2]
[1] Indian Institute of Technology Bombay, Mumbai, India [2] Microsoft India Development Center, Hyderabad, India

Introduction

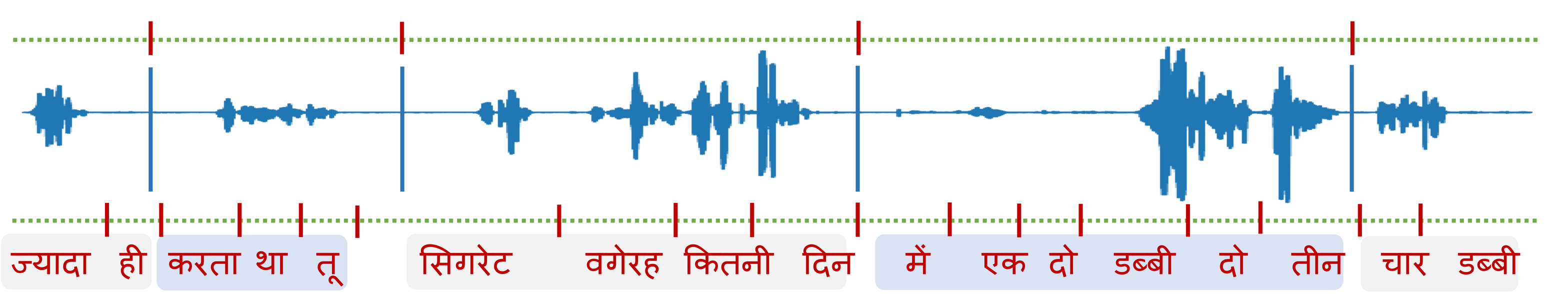
- Code-mixing (CM): speakers alternate between languages within a sentence or a discourse.
- Widespread use of CM among multilingual speakers motivates code-mixed automatic speech recognition (CM-ASR).
- But code-mixed data is not readily available though monolingual resources are quite abundant.
- How can we use monolingual resources for CM-ASR?

Overview



Segmenting Monolingual Utterances

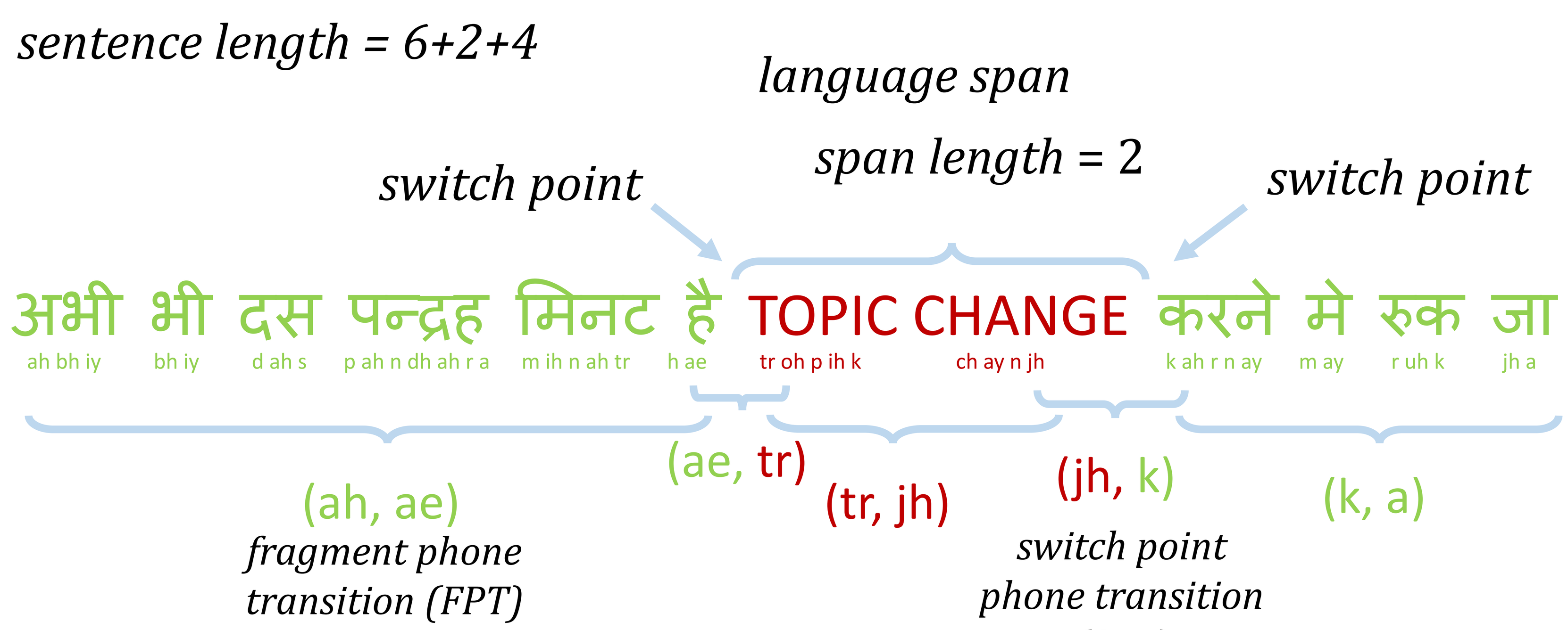
Silence detector: suggests silence markers to split utterances



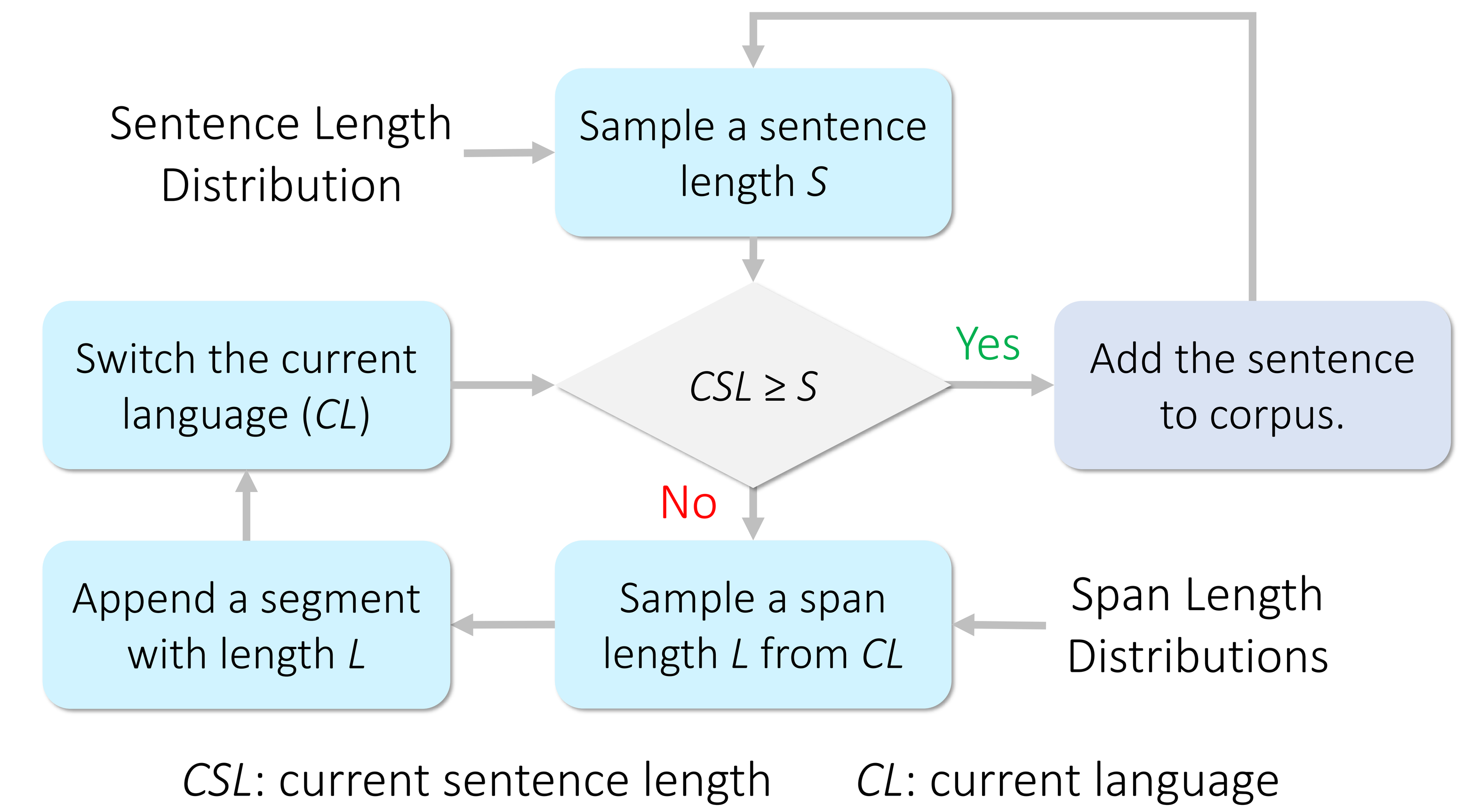
Monolingual ASR: used to get forced alignment between text and utterance

Identify nearest candidate from silence and alignments to make a cut

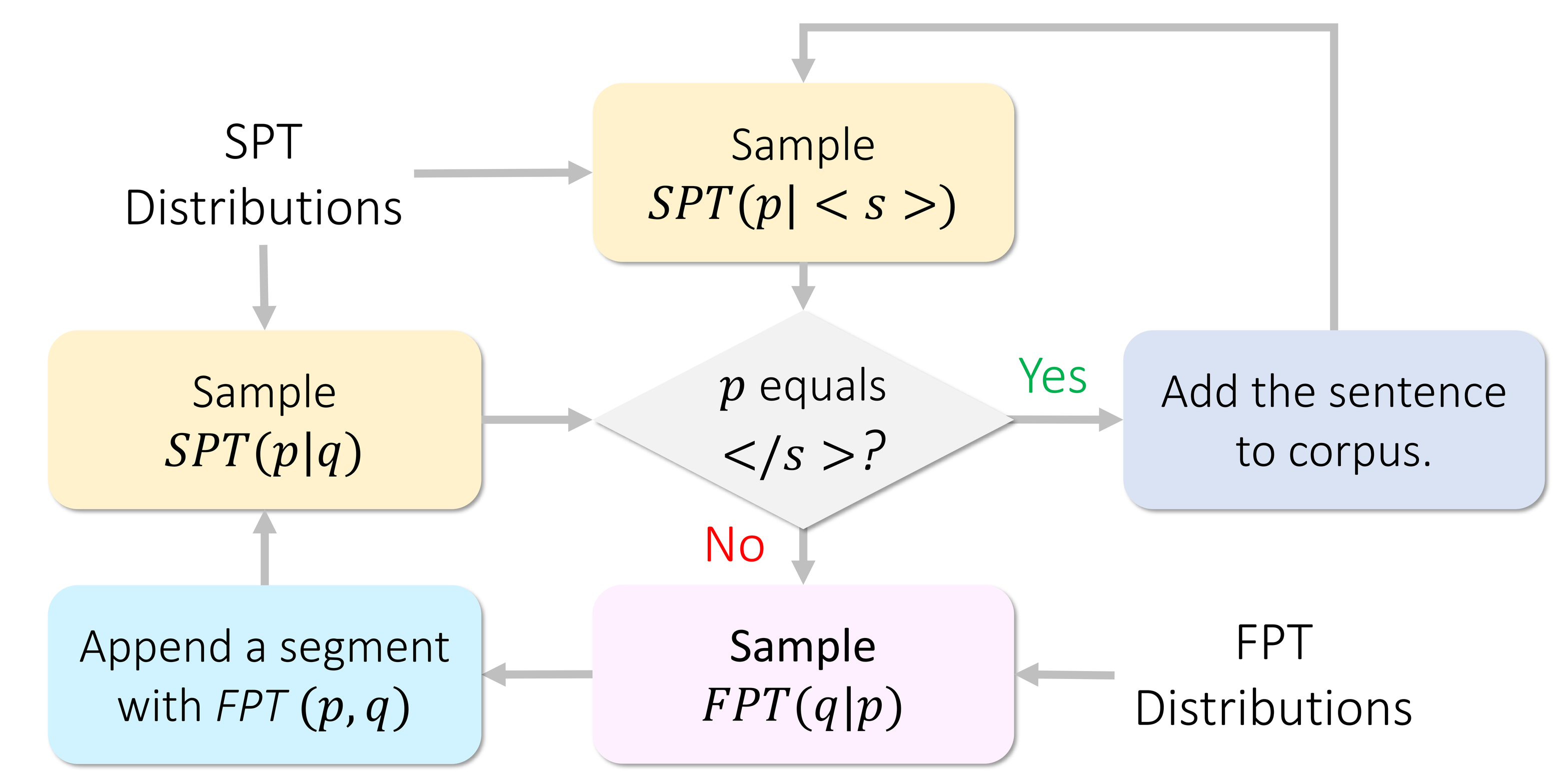
Some Definitions

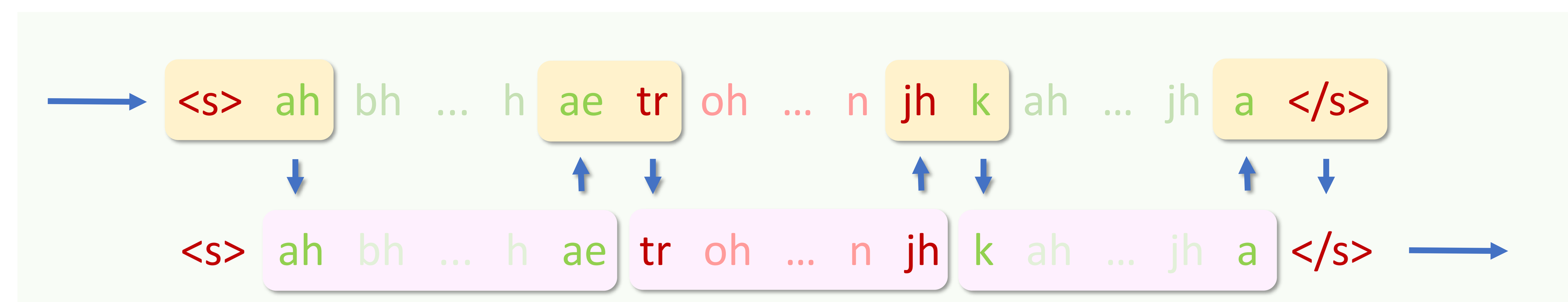


Span Length Based Concatenation

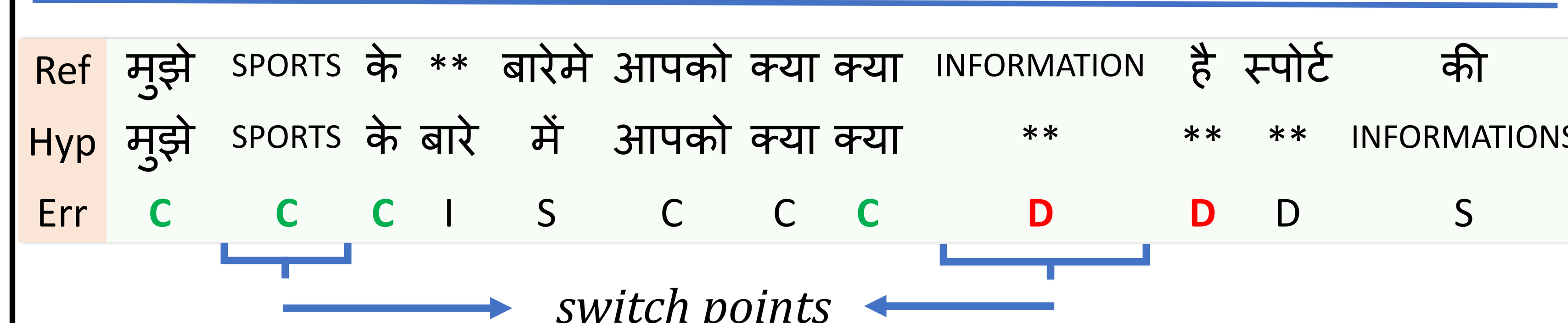


SPT and FPT Based Concatenation





CM-WER Metric and Datasets



If the reference has M words in switch points, and the hypothesis has N edits, then $CM\text{-}WER = N/M$. CM-WER = 3/8 WER = 6/11

Dataset	Description
HI/EN(50)	50 hours of Hindi/English
SynPT / SynSL / SynConcat (100)	100 hours of synthetic data based on PT or SL distributions or naïve concatenation.
CMtext	12.6k code-mixed text utterances

Experiments with Acoustic and Language Models

Training	Dev WER	Test WER
HI(50)	63.01	65.14
HI(50)+SynConcat(100)	60.89	62.44
HI(50)+SynSL(100)	60.22	62.28
HI(50)+SynPT(100)	59.05	60.81
HI(350)	58.99	60.47
HI(350)+SynConcat(100)	57.91	59.65
HI(350)+SynSL(100)	57.22	58.29
HI(350)+SynPT(100)	57.31	58.73
HI(350)	58.99	60.47
HI(350)+SynPT(100)	57.31	58.73
HI(350)+SynPT(200)	56.62	58.73
HI(350)+SynPT(350)	56.28	58.29

Training	Dev WER	Test WER
HI	58.65(63.81)	60.83(65.50)
HI + EN	72.36(77.66)	73.81(80.62)
HI + ECT	58.39(62.24)	60.86(65.41)
HI + EN + ECT	57.72(60.57)	60.12(62.92)
HI + EN + SynConcat	57.75(60.60)	60.15(63.04)
HI + EN+ SynSL	57.51(60.03)	60.11(62.75)
HI + EN + SynPT	57.49(60.00)	60.12(62.86)
SynSL + SynPT + ECT (S-All)	57.88(60.05)	60.25(62.79)
CMtext	55.10(52.79)	57.38(55.33)
CMtext + S-All	54.59(52.92)	56.97(55.05)

Numbers in brackets denote CM-WER after transliteration.
ECT: Equivalence Constraint Theory, SynConcat: Naïve concatenation

Summary and Conclusions

- *Span length distribution* and *phone transition distributions* are effective in characterizing code-mixed language.
- Augmenting ASR training with synthetic speech that preserves these distributions lead to an improved ASR performance on code-mixed speech.
- Language models also benefit from using text from the synthetic speech.
- Future work: Explore text-to-speech (TTS) systems to improve the quality of synthetically generated speech.

References

[1] A. Pratapa, G. Bhat, M. Choudhury, S. Sitaram, S. Dandapat, and K. Bali, "*Language modeling for code-mixing: The role of linguistic theory based synthetic data*," in Proceedings of ACL, 2018.
[2] H. Seki, S. Watanabe, T. Hori, J. L. Roux, and J. R. Hershey, "*An end-to-end language-tracking speech recognizer for mixed language speech*," in Proceedings of ICASSP, 2018.