Joint Modeling of Accents and Acoustics for Multi-Accent Speech Recognition

Xuesong Yang*, Kartik Audhkhasi†, Andrew Rosenberg†, Samuel Thomas†, Bhuvana Ramabhadran†*, Mark Hasegawa-Johnson*

*University of Illinois at Urbana-Champaign, †IBM T.J. Watson Research Center

Contribution

Dealing with speaker accent mismatch by exploring an alternate model where we **jointly learn an accent classifier and a multi-task acoustic model**.

- Experiments on two accents: Wall Street Journal American and British English
- Our **Joint** model **outperforms** the strong multi-accent acoustic model (**MTLP**) by relative WER improvements:
  - **5.94%** on British English.
  - **9.47%** on American English.

Introduction

ASR systems have achieved human parity on Switchboard [1, 2], but still perform **much worse than human speech recognition when meeting accents**.

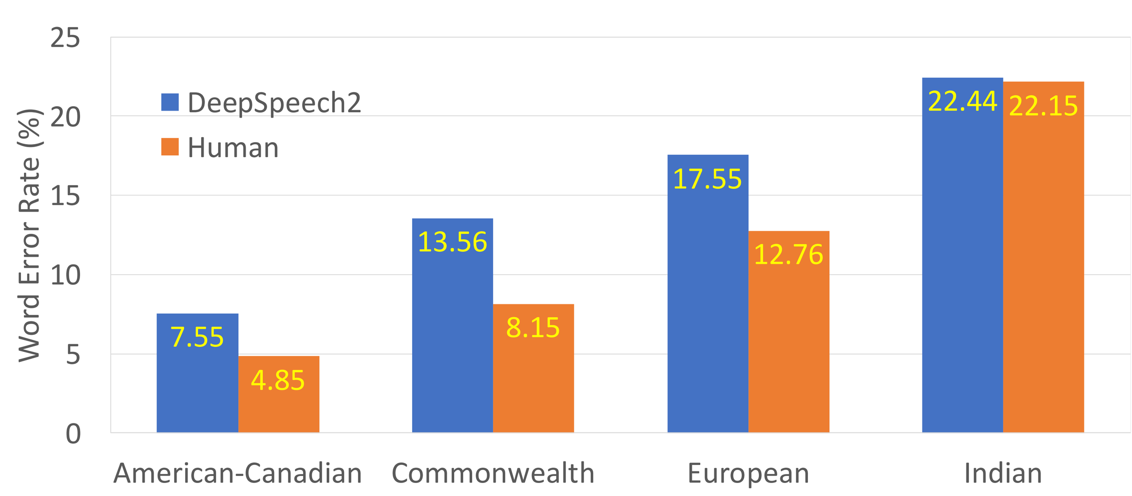

Figure 1: Comparison of WERs between DeepSpeech2 and crowd-sourced human recognition [3] on VoxForge. DeepSpeech2 is trained on 11,940 hours English speech.

Accent Variations in a Language:

- Associated with the residence, ethnicity, social class, and native language of speakers.
- Distinguished by traits of phonology, grammar, and vocabulary.

Pronunciation	British	American
***SCH**EDULE*	[ˈʃɛdjuːl]	[ˈskɛdʒʊl]
*DR**E**SS*	[ɛ] (England) [e] (Wales)	[e]

*The author is currently in Google.

Related Work

1. Hierarchical grapheme and phoneme based acoustic modeling [4]: outperformed accent specific models but achieved competitive WER with multi-accent phoneme models.
2. Adaptive multi-accent phoneme based acoustic modeling [5]: trained a multi-accent phoneme model and adapted it with a target accent.

Methods

Human Accented Speech Perception:

Humans memorize the phonological and phonetic forms of accented speech: "mental representations of phonological forms are extremely detailed," and include "traces of individual voices or types of voices" [6].

Pipeline Model with AID:

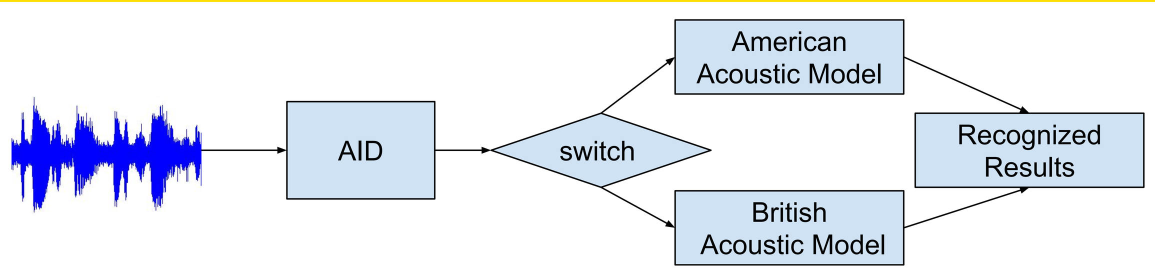

Figure 2: **Pipelines:** acoustic model (AM) is selected based on the hard-switch between accent specific acoustic models.

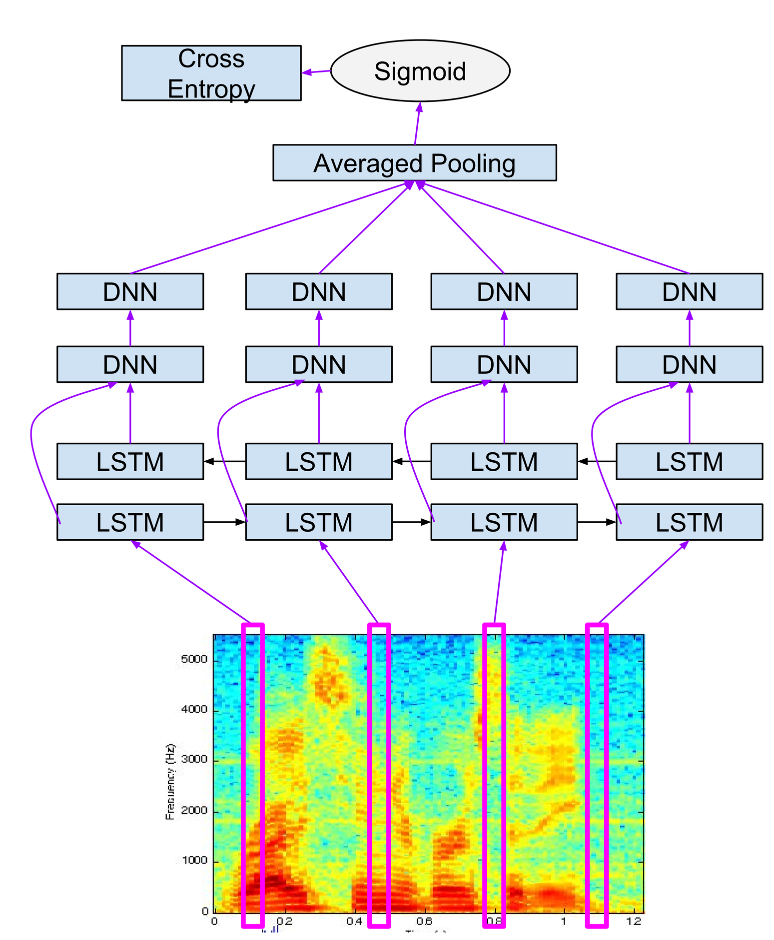

Figure 3: **AID:** accent identification with average pooling.

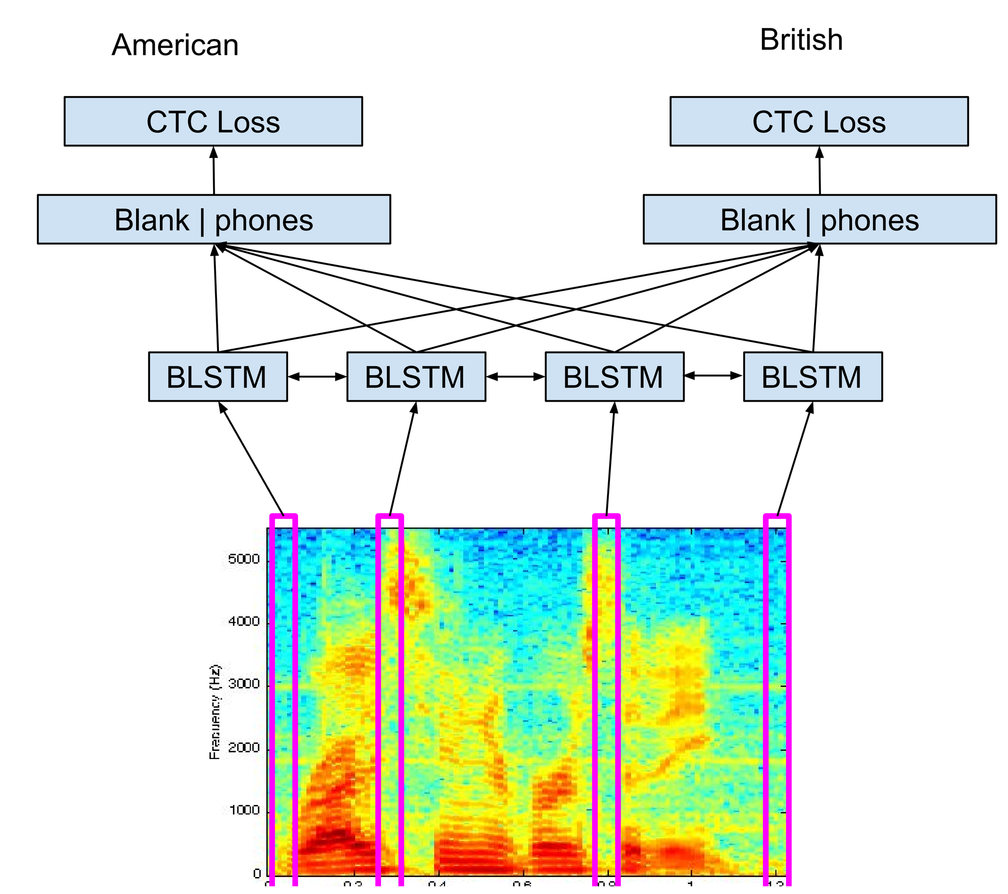

Figure 4: **MTLP:** multi-accent phoneme based acoustic model using connectionist temporal classification (CTC) loss.

$$\min_{\Theta} \mathcal{L}_{\mathsf{AM}}(\Theta) = 0.5 * \mathcal{L}_{\mathsf{UK}}(\Theta) + 0.5 * \mathcal{L}_{\mathsf{US}}(\Theta)$$

Proposed Joint Model:

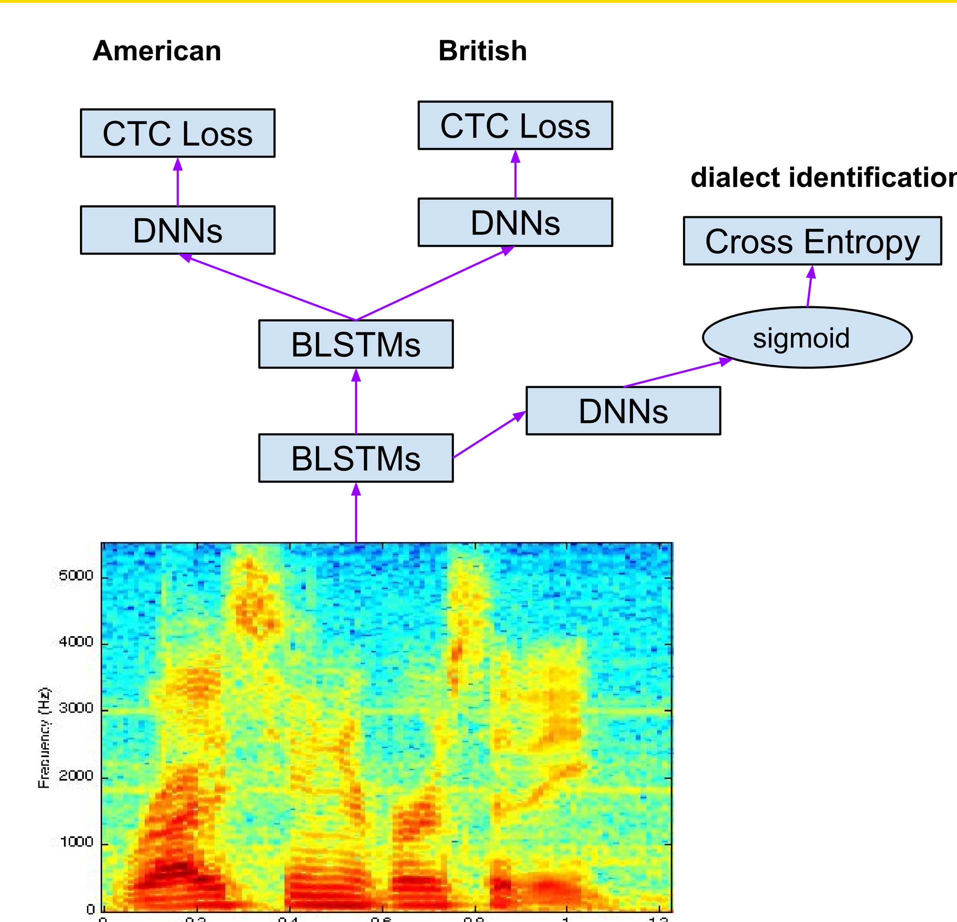

Figure 5: **Joint:** we proposed to link the training of **acoustic models** and **accent identification models** in a manner similar to the linking of these two learning processes in human speech perception.

$$\min_{\Theta} \mathcal{L}_{\mathsf{Joint}}(\Theta) = (1-\alpha) * \mathcal{L}_{\mathsf{AM}}(\Theta) + \alpha * \mathcal{L}_{\mathsf{AID}}(\Theta)$$

Experiments

Speech Corpora:

- **Train:** WSJ American English (42 phones) and Cambridge British English (45 phones), 15 hours speech recordings for each accent.
- **Test:** American English (`eval93`) and British English (`si_dt5b`)

Results:

- **ASpec:** accent specific AMs that are trained separately on corresponding mono-accent data.
- **MTLP:** multi-accent AMs that are jointly trained in a way of multitask learning.
- **Joint:** our proposed acoustic model that explicitly includes accents information.

Table 1: Oracle performance in WER (rel. imp.) that assumes the true accent ID is known in advance. The relative improvement is calculated over **ASpec**.

Corpus	ASpec	MTLP	Joint
British	11.5	10.1 (-12.17)	**9.5** (-17.39)
American	10.2	9.0 (-11.76)	**8.3** (-18.63)
average	10.85	9.55 (-11.98)	**8.9** (-17.97)

Table 2: Real task performance in WER (rel. imp.) that assumes the true accent ID is not known in advance. The relative improvement is calculated over **ASpec**. **Pipelines** model applies AID trained separately while **Joint** model applies AID jointly trained with **MTLP**.

Corpus	Pipelines with AID		Joint	Joint v.s. MTLP
	ASpec	MTLP		
British	11.5	10.1 (-12.17)	**9.5** (-17.39)	9.5 **(-5.94)**
American	11.1	9.5 (-14.41)	**8.6** (-22.52)	8.6 **(-9.47)**

[1] W. Xiong *et al.*, "Toward human parity in conversational speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2410–2423, 2017.

[2] G. Saon *et al.*, "English conversational telephone speech recognition by humans and machines," *Proc. of Interspeech 2017*, pp. 132–136, 2017.

[3] D. Amodei *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International Conference on Machine Learning*, pp. 173–182, 2016.

[4] K. Rao *et al.*, "Multi-accent speech recognition with hierarchical grapheme based models," in *Proc. of ICASSP*, pp. 4815–4819, IEEE, 2017.

[5] J. Yi *et al.*, "Ctc regularized model adaptation for improving lstm rnn based multi-accent mandarin speech recognition," in *Proc. of ISCSLP*, pp. 1–5, IEEE, 2016.

[6] J. Pierrehumbert, "Phonological representation: Beyond abstract versus episodic," *Annu. Rev. Linguist.*, vol. 2, pp. 33–52, 2016.