Cantonese Automatic Speech Recognition Using Transfer Learning

Bryan Li[1], Xinyue Cindy Wang[1], Homayoon Beigi[1,2]
Columbia University[1], NY, Recognition Technologies, Inc.[2], NY | [bl2557, xw2368]@columbia.edu, beigi@recotechnologies.com

1. Introduction

We propose a system for automatic speech recognition (ASR) of Cantonese through transfer learning from Mandarin. We take a time-delayed neural network trained on Mandarin, and perform weight transfer of several layers to a newly initialized model for Cantonese. Key findings are that this approach allows for quicker training time with less data. We find that for every epoch, log-probability is higher for our best transfer learning model compared to a Cantonese-only model. The Cantonese ASR results for transfer-learned models show slight improvement in CER. We also discuss our ongoing work in further improving results.

2. Background and Motivation

- **Motivation**: features are shared (MFCCs, pitch), leverage high-quality, larger volume data in a high-resource language
- **Transfer learning**: machine learning method to generalize models trained on one task to another
- **Model adaptation**: train a model on one language, retrain all or parts of it on a different one
- **Key idea**: features learned by neural networks are more language-independent in earlier layers than in later layers [1].
- **Prior Work** [2]: transfer learning effective for low-resource languages, especially between related pairs (e.g German to English).

3. Datasets

Dataset	Language	Length	Environments	Tones
BABEL [3]	Cantonese	215 hrs	home, street, car, etc.	6
AISHELL-2 [4]	Mandarin	1000 hrs	studio, living room	4

Dataset	Speakers	Ages	Topics	MF Ratio
BABEL	952	16-67	conversational (used), scripted (unused)	48:52%
AISHELL-2	1991	11-40	voice control, news, sports, etc.	40:60%

Table 1: Dataset statistics

We downsampled AISHELL-2 16 kHz $\rightarrow$ 8 kHz to match BABEL.

4. Model Architecture

Implemented in **Kaldi** following fairly standard pipeline
Language model: 4-gram statistical model, trained on transcripts
Acoustic model: two-stage model, GMM-HMM $\rightarrow$ TDNN

- Objective function: lattice-free maximum mutual information

Selected References

[1] D. Wang and T. F. Zheng, "Transfer learning for speech and language processing," in *APSIPA 2015*, Dec. 2015.

[2] J. Kunze *et al.*, "Transfer learning for speech recognition on a budget," in *Rep4NLP Workshop at ACL 2017*.

[3] T. Andrus *et al.* (2016) IARPA babel cantonese language pack. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2016S02

[4] J. Du, X. Na, X. Liu, and H. Bu. (2018) AISHELL-2: Transforming mandarin ASR research into industrial scale. [Online]. Available: https://arxiv.org/abs/1808.10583

5. Experiments

- **Baseline**: Cantonese model, weights randomly initialized
- **Transferred layers**: Cantonese model, for layers `affine1` to `TDNN{K}`, initialize weights to those of trained Mandarin model
- **Learning rate**: lr={0.0, 0.25, 1.0} for transferred, lr=1.0 for rest
- Tried pretrained AISHELL-2 i-vectors for TDNN – does not help

6. Diagram

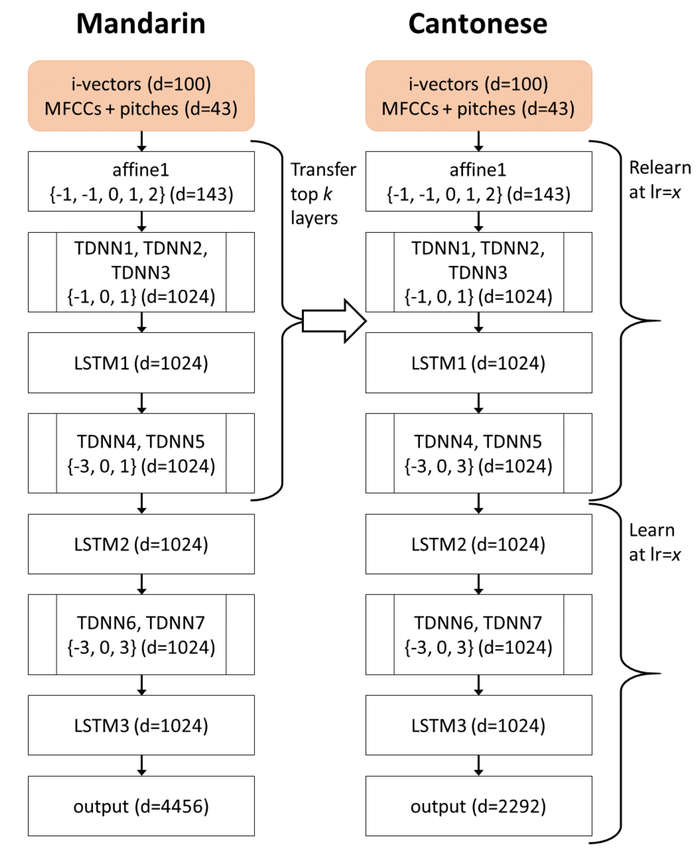

Figure 1: TDNN-LSTM acoustic model + transfer learning

Where *d* is the output dimension and *[...]* are the spliced frames i-vectors are trained with features from the GMM-HMM model

Note: Our model achieves 9.44% CER on 8 kHz AISHELL-2 (vs 8.81% on 16 kHz [4])

7. Results

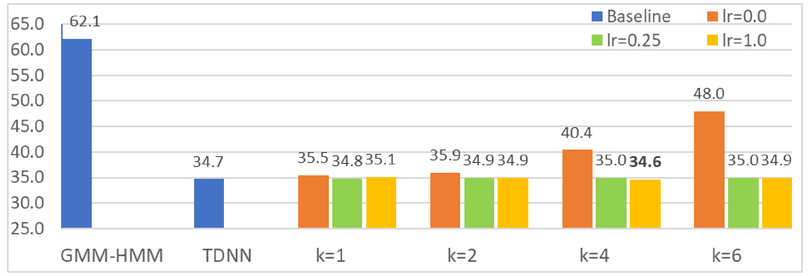

Figure 2: CER of transfer-learned models vs baseline

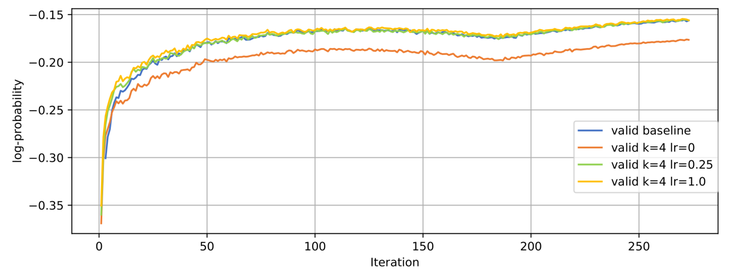

Figure 3: log-prob vs iterations (output)

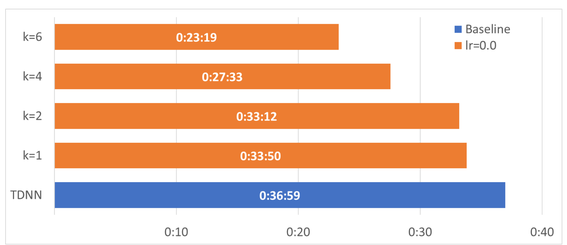

Figure 4: time per epoch

Note: for all models lr>0, train time == baseline train time

7. Conclusion and Future Work

- New state-of-the-art on BABEL Cantonese (CER=34.6%) by transferring a Mandarin model up to TDNN4, lr=1.0
- Experiments show speed up only for lr=0.0
- **Future work**: more informed transfer learning methods
  - BABEL poor audio quality – also add noise to AISHELL2
  - Improve LM: 1) more data, 2) map words + phonemes
  - Fine-tune AISHELL-2 i-vectors, then transfer