Mismatched Training Data Enhancement for Automatic Recognition of Children's Speech using DNN-HMM

Mengjie Qian[1], Ian McLoughlin[1,2], Wu Guo[1], Lirong Dai[1]
[1]National Engineering Laboratory for Speech and Language Information Processing
University of Science and Technology of China, Hefei, Anhui, China.
[2]School of Computing, University of Kent, Medway, Kent, UK.

Abstract

- Children's speech has greater time and frequency domain variability than typical adult speech, lacks good large scale training data, and presents difficulties relating to capture quality.
- Explore the incorporation of mismatched training data to achieve a better acoustic model and improve performance in the face of limited training data, as well as training data augmentation using noise.
- Explore two arrangements for vocal tract length normalisation and a gender-based data selection technique for training a children's speech recogniser.

Corpus	Data	Utterances	Speakers	Duration
Kids	Train	3545	52	6.18h
	Dev	778	13	1.47h
	Test	713	9	1.12h
Kids	Total	5036	74	8.77h
TIMIT	Train(m)	2352	325	1.98h
	Train(f)	1344	137	1.14h
	Test	1088	168	0.94h
TIMIT	Total	4784	630	4.06h

Table1: Data used in the experiments.

Dataset

1. CMU Kids
   - 24 boys and 52 girls
   - aged from 6 to 9 (one 11)
   - exclude unknown age child and age 11 child
   - Read
2. TIMIT
   - read

Training data adaptation

1. Vocal tract length normalization (VTLN)

- VTL increases from infancy to adulthood both according to body size and differently according to gender (shown in Fig.1).
- The VT shape can be modelled by a uniform lossless acoustic tube with the closed end represented by the glottis and the open end represented by the lips.

$$F_k = \frac{c}{4L}(2k-1) \quad where\ k = 1,2,3,\ldots$$

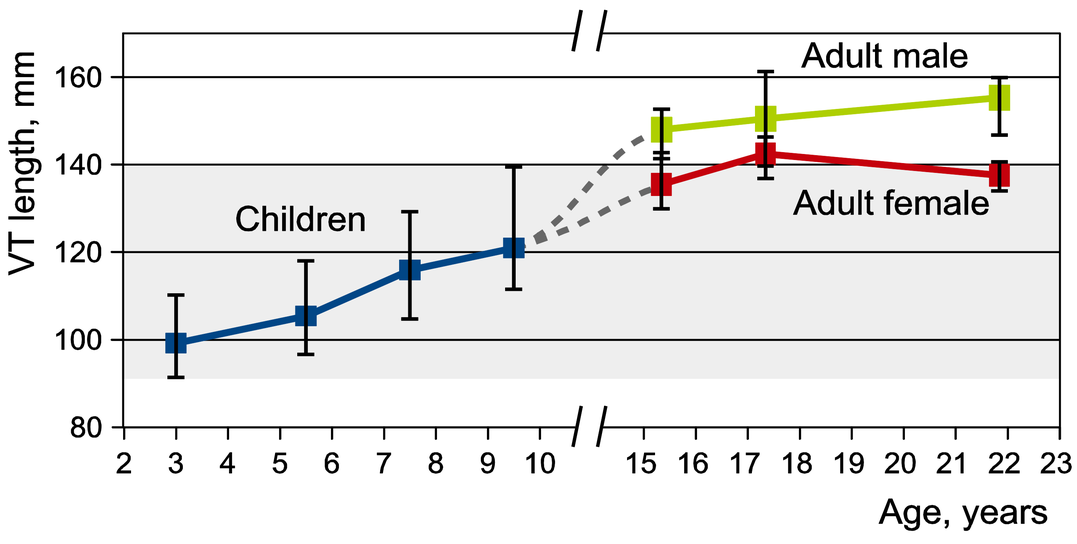

Figure 1: Plot of VT length for various age children and adults using data from [12, 13], with the approximate band of child VT lengths shown shaded.

2. Noise augmentation

- Add acoustic noise to clean utterances to form a larger corpus.
- 115 noise types are used, including 100 noises recorded by G. Hu [16] plus 15 other common noise types recorded locally [17].
- Separate the whole clean dataset into 115 small parts and add one type of noise to each part using the Filtering and Noise Adding Tool (FaNT) [18].
- Each type of noise is added at a SNR of 20dB.

3. Use of mismatched training data

- Adding adult speech from the TIMIT database to the children's speech data for training.
- Inspired by Fig. 1, female adult speech resembles child speech much more closely than adult male speech does. Thus, selecting only adult female data from TIMIT to augment the CMU Kids training corpus.

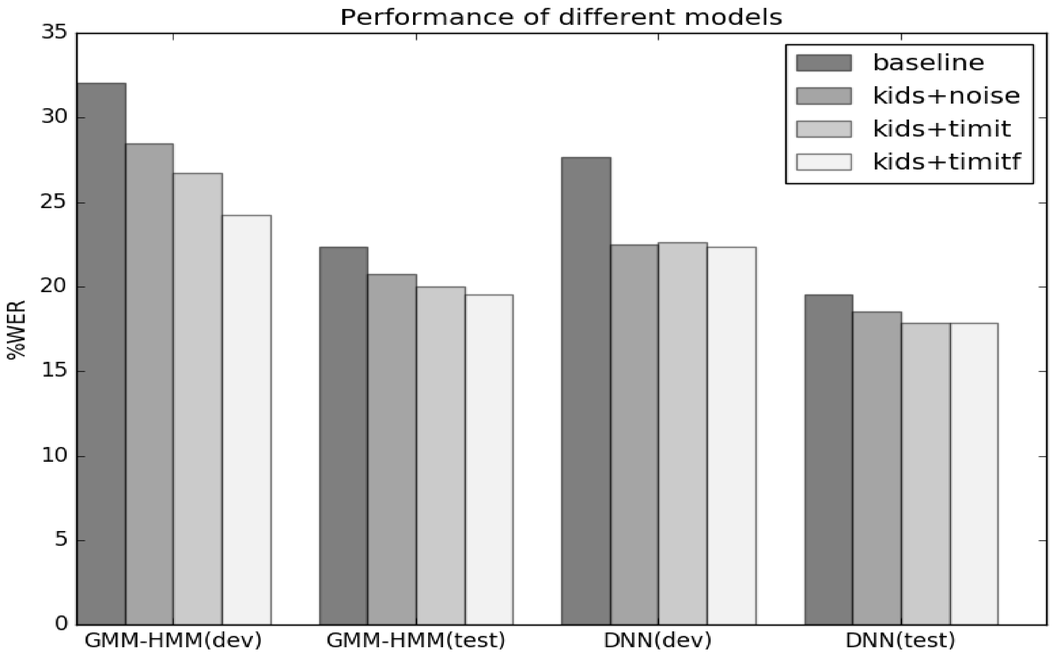

Figure 2: Word error rate for various types of mixed data models for four ASR systems.

Experiments and results

- DNN structure: 253-1024-1024-1024-1024-1200.
- Explore how to overcome the training resource issues using the techniques discussed above, and the results are shown in Fig.2.
- Two ways to account for the large VTL difference between adult and child speech:

(i) apply VTLN to adult utterances during training to make the normalised feature more similar to children's speech;

(ii) apply VTLN to children's utterances during training and test.
Results shown in Fig.3.

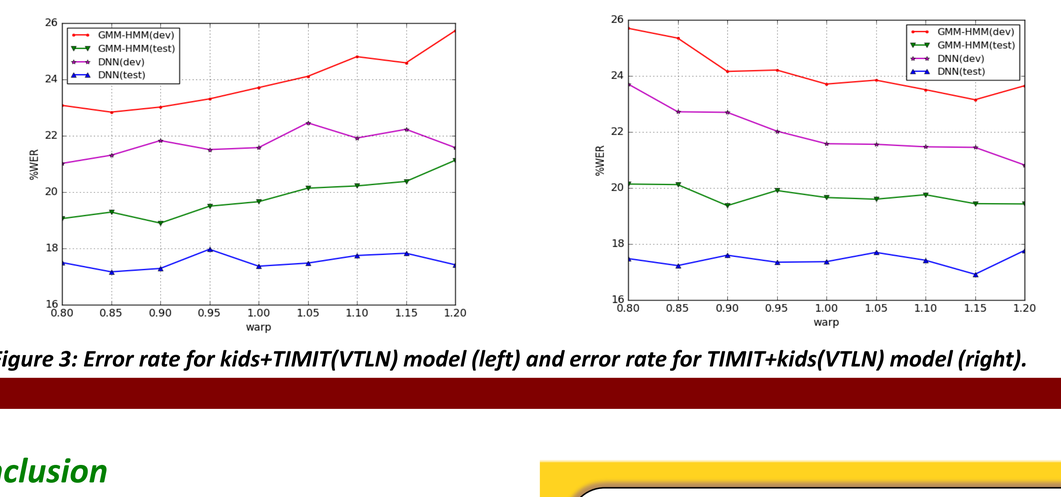

Figure 3: Error rate for kids+TIMIT(VTLN) model (left) and error rate for TIMIT+kids(VTLN) model (right).

Conclusion

- Augmenting children's speech training database with adult speech is an attractive idea given the potentially vast amounts of adult speech data available for training.
- Gender-selected adult training data is much more beneficial to results, despite the obvious halving of the additional training resources.
- Adding noise to augment the training data also provide some benefits, but not to the same extent as adding adult female speech.

System: evaluation	V1	V2
GMM-HMM: dev	23.15	22.84
DNN: dev	20.82	21.02
GMM-HMM: test	19.43	19.06
DNN: test	16.92	17.17

V1: kids+TIMIT(VTLN) training
V2: TIMIT+kids(VTLN) training

Table 2: Best performance (in %WER) for each ASR system when employing VTLN for training with kids' and TIMIT data.

- VTLN is effective at dealing with the variability between children's and adults' speech, however whether to apply VTLN on adults' speech (training data) or children's speech (test and training data) depends to some extent on the nature of the data.