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

Mengjie Qian[1], *Ian McLoughlin*[2], *Wu Guo*[1], *Lirong Dai*[1]

[1] National Engineering Laboratory of Speech & Language Information Processing
The University of Science and Technology of China, Hefei, Anhui, China.
[2] School of Computing, The University of Kent, Medway, Kent, UK
qmj@mail.ustc.edu.cn, ivm@kent.ac.uk, {guowu, lrdai}@ustc.edu.cn

Abstract

The increasing profusion of commercial automatic speech recognition technology applications has been driven by big-data techniques, using high quality labelled speech datasets. 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. Each of these factors reduces the performance of systems that automatically recognise children's speech. In this paper, children's speech recognition is investigated using a hybrid acoustic modelling approach based on deep neural networks and Gaussian mixture models with hidden Markov model back ends. We 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. We also explore two arrangements for vocal tract length normalisation and a gender-based data selection technique suitable for training a children's speech recogniser.
Index Terms: child speech recognition, children's ASR, vocal tract length normalisation

1. Introduction

Automatic speech recognition (ASR) for children has attracted much attention as it is of great significance for many application domains including entertainment, education and information accessibility tools [**?**]. Despite the good potential, children's speech ASR has been widely noted to perform poorly compared to adult ASR. Performance is particularly poor when models are trained using adult speech data [**?**], which is typically attempted because much better adult speech training data is available than children's speech training data.

Children's ASR performance is much worse than adult systems due to (a) acoustic differences, (b) phonetic differences, between adult and children's speech, and (c) limitations of training data quality and quantity. Children have shorter vocal tracts and vocal fold lengths and lower mass vocal cords compared to adults, resulting in higher positions of formants and fundamental frequency [**?**]. Meanwhile younger children in particular may not have yet learned how to pronounce specific phonemes [**?**], hence exhibit greater speaking variability than adults. Children may not speak clearly, concisely or in accordance with grammatical norms, causing difficulties with language modelling. Although their spoken vocabulary may be smaller (which is beneficial to recognition [**?**]) they tend to use words that do not occur in adult speech, which are used inappropriately or are incorrectly pronounced. Further exacerbating the difficulty, children's speech databases are much smaller than adult speech databases, so that insufficient training data significantly limits the performance of acoustic models [**?**].

Several techniques have been explored to improve children's ASR: (i) Defining better acoustic features for children's speech. The most common features are MFCCs, filter bank and PLP coefficients [**?**], with MFCCs achieving best performance in GMM-based ASR systems [**?**, **?**] and mel filterbank coefficients most commonly used in DNN-based systems. (ii) Pronunciation modelling, since children use different pronunciation, there may be substantial age-dependent differences, hence research to better model the phonemes that children tend to mispronounce [**?**]. (iii) Vocal tract length normalisation (VTLN), to account for formant shifts induced by difference in VT length between speakers. (iv) Model adaptation techniques which are also used in adult ASR such as maximum a posterior (MAP) and maximum likelihood linear regression (MLLR).

1.1. Novelty

This paper explores children's ASR using baseline DNN-HMM and GMM-HMM systems based on the CMU Kids Corpus. Using this baseline, (i) we evaluate training data augmentation for children by adding noise (which is known to work for adult ASR training [**?**] but has not yet been evaluated for children), (ii) we introduce a novel training selection approach based on gender and (iii) trial the use of VTLN in opposing directions, since the two directions are currently unexplored for children's ASR. Section 2 describes the standard speech corpuses used in our experiments. Training data augmentation, VTLN and gender selection are discussed in Section 3, Section 4 presents and discusses details of the various experiments and Section 5 will conclude the paper.

2. Database

CMU Kids Corpus contains about 9 hours of recordings of material read by children. In total, 24 male and 52 female speakers, ranging in age from 6 to 11 years in the first to third US school grades (the 11-year-old was in 6th grade). Based on preliminary experiments, the age 11 child and those of unknown age were excluded due to the limited number of recordings and speakers. Instead we focus on the larger amount of data from children aged 6 to 9 for which utterances have been accompanied by a transcription. We divide into training data, development data and test data at a ratio of approximately 4:1:1, leading to the distribution of data shown at the top of Table 1.

We also test using adult speech from the TIMIT corpus for training. This consists of 16kHz recordings from 630 male (m)

Table 1: *Data distribution in CMU Kids and TIMIT corpuses.*

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

and female (f) speakers of eight American English dialects reading phonetically rich sentences, shown at the bottom of Table 1.

3. Training data adaptation

The legal and ethical difficulties implicit in collecting training data from minors, coupled with the difficulty of recording children's speech, means that there is a lack of training data. During recordings, children frequently substitute, insert or miss words, and instead of keeping steady when speaking, they tend to fiddle with the microphone, wobble their body, or move their heads. These actions result in poor quality recordings compared to those of adults. Together, there is thus a lack of both quantity and quality in children's speech corpuses suitable for training ASR systems.

3.1. Vocal tract length normalisation

Vocal tract length (VTL), measured along its midline from the glottis to the lips, is an important parameter that accounts for much of the acoustic inter-speaker variability in speech production. This structural characteristic influences many aspects of speech, and varies significantly between speakers.

VTL increases from infancy to adulthood both according to body size and differently according to gender. In infants, the larynx is located much higher up in the throat than in adults. A phenomenon known as larynx descent occurs between the third month and the third year when the larynx moves closer to the adult position. Then, as children grow further, VTL increases steadily with body mass. VTL does not differ significantly between boys and girls until puberty, when a second larynx descent occurs for males. This is argued to be the main reason for gender-based variation in VTL [**?**] in adults. Length varies from an average of approximately 8 cm at birth, to about 16 cm in adulthood (with a range from about 13 to 20 cm). Fig. 1 illustrates VTL against age for both children and young adults.

The influence of VTL on speech acoustics has been explained as part of the filter theory of speech. 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 [**?**]. The relationship is explained as follows:

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

where c is the speed of sound and L is the uniform tube length. Formant positions are thus inversely proportional to VTL. Vocal tract length normalisation (VTLN) is used in ASR to reduce mismatch between speakers, and improve system performance [**?**]. VTLN typically seeks a frequency scale transformation that allows for an optimal comparison between spectral features. Frequency warping functions including linear, nonlinear and power-based, with piece-wise linear warping being most common [**?**].

In general there are 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 features more similar to children's speech and (ii) apply VTLN to children's utterances during training and testing to make them more similar to adults' speech. We will evaluate both directions of normalisation.

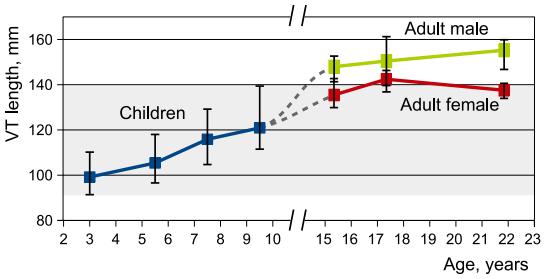

Figure 1: *Plot of VT length for various age children and adults using data from [**?**, **?**], with the approximate band of child VT lengths shown shaded.*

3.2. Noise augmentation

CMU Kids corpus is a clean corpus that has been recorded with the intention of minimising interfering noise. Despite being one of the largest and best corpuses of children's speech, it has wide variability in age, and is too small to train a viable model for either baseline ASR system. One way around this is to add acoustic noise to clean utterances to form a larger corpus of clean recordings augmented with noisy recordings [**?**].

In this study, 115 noise types are used, including some musical noise. These 115 noise types include 100 noises recorded by G. Hu [**?**] plus 15 other common noise types recorded locally [**?**], all of which are readily available for public download. For training, we separate the whole clean data set into 115 small parts and add one type of noise to each part using the Filtering and Noise Adding Tool (FaNT) [**?**]. Each type of noise is added at a SNR of 20dB, resulting in the data distribution shown in Table 2a.

3.3. Use of mismatched training data

Adding adult speech from the TIMIT database to the children's speech data for training, leads to the mismatched training resource shown in Table 2b. The idea being that benefits gained from increasing the quantity of training data may override the known mismatch in age.

A novel method introduced in this paper is inspired by the observation from Fig. 1, that 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 may lead to a better trained children's ASR system. This training combination is shown in Table 2c.

4. Experimental setup

4.1. Baseline system

MFCCs are used as front-end features, with 13 coefficients from a 25 ms frame with 10 ms shift between frames and 16 kHz sampling frequency. Delta and delta-delta coefficients are also used

Table 2: *Training data arrangements for the remaining three ASR models that are compared.*

(a) Training data for the kids+noise model:

Data	Utterances	recording length
kids	3545	6.18h
noisy kids	3545	6.18h
Total	7090	12.36h

(b) Training data for the kids+TIMIT model:

Data	Utterances	recording length
kids	3545	6.18h
TIMIT	3696	3.12h
Total	7241	9.30h

(c) Training data for the kids+TIMIT(female) model:

Data	Utterances	recording length
kids	3545	6.18h
TIMIT	1088	0.94h
Total	4633	7.12h

Table 3: *Performance (WER) of the baseline system.*

Dataset	GMM-HMM	DNN
Dev	32.03%	27.65%
Test	22.32%	19.50%

with a context size of 11 frames (i.e. 5 frames before and 5 after). We use hidden Markov models (HMMs) to represent each phone using a 5-state HMM, while a 3-state HMM is used to model silence, noise and short-pauses. The language model (LM) used by the speech decoder is a trigram model, trained on the reference transcriptions of training data. Before training, the texts were cleaned and normalised by removing punctuation and expanding numbers. Incomplete words were marked.

Context-dependent deep neutral networks (DNN) have demonstrated gains in many challenging ASR tasks. In this paper, we explore the performance of DNNs for children's ASR. Specifically, an eleven frame context window of filter bank features (5 frames to each side of the current frame) is used as input to form a 253 dimensional feature vector. Training data is used to layer-wise initialise the weights of a deep, feed-forward network and then back propagation is used to fine-tune the network weights. The DNN has 4 hidden layers each with 1024 neurons, thus the resulting architecture is $253 - 1024 - 1024 - 1024 - 1024 - 1200$.

The baseline children's speech recognition system is trained and tested using kids speech as described in Table 1 and yields the performance is shown in Table 3. From the results we can see that the DNN system exhibits lower WER than the GMM-HMM system both on development and test sets. Specifically, the DNN system improves WER over GMM-HMM by 13.7% relative to the development set and by 12.6% relative to the test set. Results are consistent with those of other authors [**?**].

4.2. Models with mixed data

Children's speech recognition is clearly more challenging than ASR for adults, not only because of children's physical immaturity but also because of the quality and quantity of the speech data set, thus we explore how to overcome the training resource issues using the techniques of Section 3.

Results are plotted in Fig. 2 in terms of word error rate

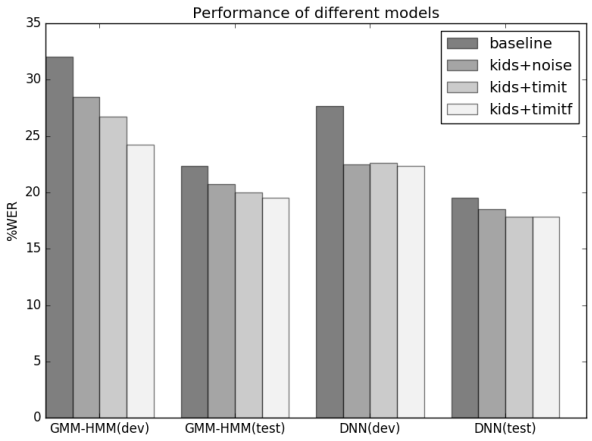

Figure 2: *Word error rate for various types of mixed data models for four ASR systems.*

for the baseline systems, and different combinations of training data. In each case, testing was conducted using only the CMU Kid's corpus test data.

Examination of the results reveals several trends that hold true for each of the tested systems. These are namely:

- Adding noise to the clean data to multiply the amount of training data can result in a better acoustic model, yielding improved performance over the baseline system.
- Adding TIMIT adult data to the kids' dataset improves performance further than the noisy data enhancement does, despite resulting in a smaller training set.
- Selecting only TIMIT female speech to add to the kids' dataset results in even better performance than either of the above, despite resulting in the smallest training data set. This is effectively trading off the amount of data against its relevance, and clearly supports the hypothesis that adult female speech is useful for training a kids' speech model.
- The training approaches evaluated here achieve greater improvements for the GMM-based system than for the DNN-based system.
- The final DNN performance on the test dataset (17.83%) is a substantial improvement on the equivalent GMM baseline performance with kids' data alone (22.32%).

4.3. VTLN normalized features

Considering the VT variability among children and adults, we investigate whether VTLN is effective at mitigating against that variability. Since the kids+TIMIT model has better performance than the baseline model, yet includes a much wider range of VT lengths, we aim to demonstrate improved results by employing VTLN on this data set. In total, we consider a set of 9 warping factors which are evenly distributed in the range of 0.8 to 1.2, with a step size of 0.05. As mentioned in Section 3, we evaluate two alternative methods of applying VTLN: (i) normalise adults' utterances to make the features more similar to children's speech during training and (ii) normalise children's utterances to make them more similar to adults' speech during training and testing. These systems are denoted 'kids+TIMIT(VTLN)' and 'TIMIT+kids(VTLN)' respectively.

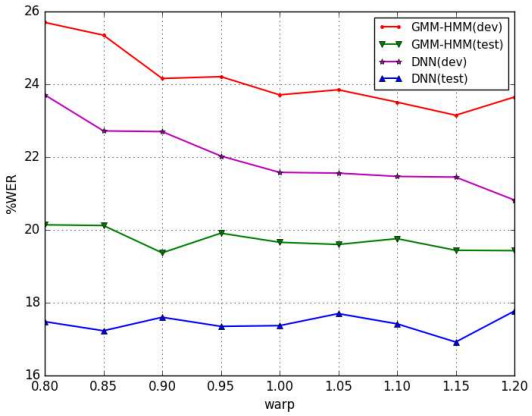

Figure 3: *Error rate for kids+TIMIT(VTLN) model.*

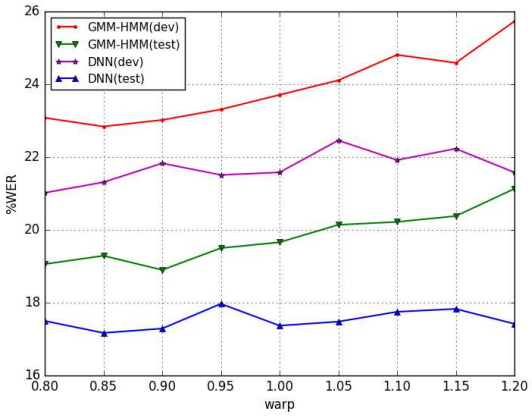

Figure 4: *Error rate for TIMIT+kids(VTLN) model.*

To build the former, a linear feature transform corresponding to each VTLN warp factor is implemented. These warping factors are then used in feature extraction from TIMIT by changing the spacing and width of the filters in the mel filterbank while maintaining the frequency axis of the speech power spectrum unchanged. The process to build the TIMIT+kids(VTLN) system is similar, except that in this case, VTLN is applied to the kids corpus rather than to the TIMIT speech.

The results are shown in Fig. 3, revealing that performance increases slightly for all systems when the warp increases, which is unsurprising since larger warps move the effective TIMIT VTLs closer to those in the kids' corpus. By contrast, in Fig. 4, when the warp is smaller, the performance of TIMIT+kids(VTLN) improves, since this causes the kids VTL to become more similar to that of the adults speech. The best result for each model is shown in Table 4, where we can see that;

- The improvement in moving from GMM to DNN is slightly larger for kids+TIMIT(VTLN) than for TIMIT+kids(VTLN), when evaluated on both development and test sets, by a relative WER reduction of 10.1% on the former and 12.9% on the latter for kids+TIMIT(VTLN) against 8.0% and 9.9% respectively for TIMIT+kids(VTLN).

Table 4: *Best performance (in %WER) for each ASR system when employing VTLN for training with kids' and TIMIT data.*

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

- When considering the GMM systems, both the development set and the test set perform better with TIMIT+kids(VTLN) training data. Interestingly, the opposite is true for the DNN systems. Evidently the DNN system outperforms GMM-HMM for all evaluations.

A 2-pass decoding strategy was also tested based on [**?**], where we use the result from the first-pass decoding as supervision to obtain a maximum likelihood estimate of the warping factor. While this is a logical strategy, the results from this method slightly under performed the GMM-HMM system (23.20% and 22.93% respectively for the development set for systems V1 and V2 respectively), and thus are not explored further. The warping factor which is estimated to be best in the first pass does not perform best in the second pass.

5. Conclusion

Augmenting the children's speech training database with adult speech is an attractive idea given the potentially vast amounts of adult speech data available for training. In this paper it has been shown that the resulting mismatched training set is slightly beneficial when considering CMU Kids corpus and TIMIT. However this paper also introduced the novel idea of gender selection on the basis that adult female speech is more similar to child speech than adult male speech is, due to female speakers not having undergone the second larynx descent. Gender-selected adult training data is demonstrated to be much more beneficial to results, despite the obvious halving of the additional training resources that this entails.

Adding noise to effectively augment the training data has also been shown to provide some benefits, but not to the same extent as adding adult female speech, since this is evidently more relevant in nature to the target speech of children than additional noise would be.

VTLN has been shown 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 (testing and training data) is a question that needs to be considered when using different training arrangements. This paper has evaluated both approaches and shown that which method is best depends to some extent on the nature of the data, with both DNN and GMM-based recognition systems being in agreement concerning which is the optimal choice for each condition. In future work, we would like to further explore how vocal tract length variability affects ASR performance for children's speech, and use this knowledge to better exploit ways to track variability in training data in order to build a better recognition system for children's speech.

6. Acknowledgements

This work was partially funded by National Key Research and Development Program (Grant No.2016YFB100130300)

and Natural Science Foundation of Anhui Province (Grant No. 1408085MKL78).