Improving Children’s Speech Recognition through Out-of-Domain Data Augmentation

Joachim Fainberg¹, Peter Bell², Mike Lincoln², Steve Renals¹

(1) Centre for Speech Technology Research, University of Edinburgh, Edinburgh, EH8 9AB, UK
(2) Quorate Technology Ltd, Edinburgh, EH8 9AG, UK
{j.fainberg, peter.bell, s.renals}@ed.ac.uk, mike.lincoln@quoratetechnology.com

Abstract

Children’s speech poses challenges to speech recognition due to strong age-dependent anatomical variations and a lack of large, publicly-available corpora. In this paper we explore data augmentation for children’s speech recognition using stochastic feature mapping (SFM) to transform out-of-domain adult data for both GMM-based and DNN-based acoustic models. We performed experiments on the English PF-STAR corpus, augmenting using WSJCAMO and ABI. Our experimental results indicate that a DNN acoustic model for children’s speech can make use of adult data, and that out-of-domain SFM is more accurate than in-domain SFM.

Index Terms: speech recognition, data augmentation, children’s speech

1. Introduction

Recognition of children’s speech poses challenges to Automatic Speech Recognition (ASR) due to the small size of easily available corpora and the large acoustical variations. A commonly used British English children’s speech corpus, PF-STAR [1], contains approximately 7.5 hours of training data – about a tenth of the size of WSJCAMO [2], and about 2.5% of the size of the Switchboard training set [3]. On large amounts of data, state-of-the-art results on children’s ASR are impressive [4]. However, large corpora of children’s speech are proprietary. Furthermore, the large age and gender dependent variations in anatomy before adulthood effectively dilutes the data, yielding poorer results on children’s speech compared to adults’ speech with similar amounts of data [5, 6]. From newborn to adulthood the vocal tract length is approximately doubled [7]. With concurrent changes in vocal tract shape, the formants consequently shift with age. Vocal tract length normalisation (VTLN) attempts to alleviate such variations by adjusting the filterbank in the front-end with a suitable frequency warping function. Good results have been observed on children’s data using a piecewise linear warping function [8, 9, 10]. However, the search for parameters is inefficient, even with gradient search [11] in place of a typically exhaustive grid search. Maximum likelihood linear regression (MLLR) adaptation approaches can improve results [8] but not sufficiently to approach corresponding adult models. Using VTLN as a prior for the MLLR family of adaptation transformations [12] has proven to be effective for adapting child speech in HMM speech synthesis.

The effect of age and gender dependent variations in children’s data is demonstrated in Figure 1, showing mean Euclidean distance between single multivariate Gaussians (MVN) within age- and gender-dependent monophone models of PF-STAR. Spikes may be attributed to the small number of speakers (80), yet the figure suggests increasing phone discrimination with age; a corollary of similar results shown by [13].

To alleviate a lack of data, many authors have attempted simple data augmentation setups in which un-modified adult speech data is added to a children’s training set. This has generally not proven fruitful [5, 9, 14, 15]. A range of in-domain data augmentation techniques exist, typically applied in a low-resource speech recognition setting, such as Vocal Tract Length Perturbation [16] and Stochastic Feature Mapping (SFM) [17]. The latter technique is of particular interest as it does not rely on any hyperparameters, aside from the amount of augmentation data to be generated. SFM generates more data by learning label-preserving feature transformations between speakers within a corpus. For speech recognition of the low-resource languages Bengali and Assamese, word error rate (WER) improvements of up to 1.8% and 2.9% have been observed [17].

The best age-independent WER reported on the British English version of PF-STAR is 44.5% using a Hidden Markov Model (HMM)-Gaussian Mixture Model (GMM) triphone system and an equal probability grammar of 1782 words [18]. In this work we first aim to build a strong baseline (Section 3.1), reporting up to 15.5% absolute improvements using Deep Neural Network (DNN) acoustic models and a stronger language model. We then investigate un-modified augmentation (Section 3.2), showing that contrary to the literature, DNN models

This work was partially supported by a PhD studentship funded by Bloomberg.
may be able to use of out-of-domain data effectively. Finally we compare standard SFM with a novel use of SFM on out-of-domain data (Section 3.3). Our results suggest that out-of-domain SFM is, in this case, more applicable than in-domain SFM. Compared to un-modified augmentation, out-of-domain SFM yields further improvements and a WER of 27.2%, improving on the DNN baseline by 6.2% relative.

2. Stochastic Feature Mapping

SFM was proposed by Cui et al [17] as a label-preserving augmentation algorithm. The algorithm augments the data by transformations of the in-domain data itself. In this sense it is similar to VTLN [16], which iteratively applies VTLN to copies of the data, and to speed perturbation [19], which speeds up and slows down copies of the data by resampling. However, these techniques rely on a hyperparameter (e.g. warping factor or resampling rate), while SFM only requires the amount of copies to be made.

In SFM, the label space of a speaker \( t \) is augmented by transforming the features of another speaker \( s \) by an affine transform across labels:

\[
O^{(s)} = A(t, s)O^{(k)} + b(t, s),
\]

where \( O^{(k)} \) denotes the features for speaker \( k \), and \( A(t, s) \) and \( b(t, s) \) are the transform matrix and bias relating source speaker \( s \) and target speaker \( t \).

In practice this is achieved by estimating a feature-space MLLR (fMLLR) transform for \( s \) given a speaker-dependent model of \( t \). One or more target speakers are randomly selected for each speaker in the corpus. The number of targets are called “replicas” in [17]. An fMLLR transform is estimated given a target speaker. The original fMLLR transforms of the target speakers are then applied to the transformed features. The features are subsequently combined with the original dataset and DNN training proceeds normally on the augmented set.

The source speakers \( s \) and target speakers \( t \) are sourced within the same corpus in the original work. We propose to also employ source speakers from external corpora. This allows for larger variety and greater flexibility – it may be that some corpora are more suited to augmenting a corpus than others. It should be noted that applying SFM in this manner inherently assumes that the acoustic model will benefit from more similar data.

In our experiments we estimate MLLR and fMLLR transforms for each speaker in the corpus given the baseline speaker-independent model. A set of speaker dependent models, \( \lambda_t \), are generated from the speaker-independent models and the MLLR transforms. For each source speaker, \( s \), from the source corpus (e.g. in-domain, PF-STAR; or out-of-domain, WSJCAM0 or ABI) we assign at random one or more target speakers, \( t \), from PF-STAR. An fMLLR transform is estimated for the speaker given the corresponding speaker-dependent model of the target speaker: \( A(t, s)\lambda_t \). This moves the source features into the target feature space. Finally, the fMLLR transform of the original target speaker is applied to the transformed features. Training proceeds on the combined dataset. As the label of the source speaker is preserved, the original alignments could be reused. Empirically we found improved performance by instead running one more pass of alignment on the combined data prior to training.

3. Experiments

We use two out of domain British English corpora for the experiments: WSJCAM0 and ABI (Accents of the British Isles), WSJCAM0 [2] is a British English version of the American English WSJ0 corpus [20]. It consists of 140 speakers speaking roughly 110 utterances each from the Wall Street Journal. The majority of the corpus, about 83%, are speakers aged between 18 and 28. The training set amounts to approximately 81 hours of speech.

The ABI corpus [21] consists of 280 speakers, evenly distributed across gender and 14 British English accent groups, recorded in a variety of uncontrolled environments: background noise varies with each accent group. The corpus is not a priori split into training and test sets. We extracted random subsets of speakers for training and test sets with an 80/20 split, yielding about 16 hours of training data.

We use a trigram language model trained on roughly 600 hours of subtitles from the British Broadcasting Corporation (BBC) [22] with corresponding lexicon taken from the 2015 MGB Challenge. The vocabulary consists of 238580 words. There are 50 out-of-vocabulary words within the PF-STAR dataset, the majority of which are mispronunciations labelled as such (e.g. ***TING***). These make up roughly 4.3% of the corpus.

Significance testing is performed with the matched pairs test from the NIST Scoring Toolkit [23]. The test affords comparison of two different models given the same test set, given the null hypothesis that the average difference in errors between segments of the two models is zero [24].

3.1. Baseline models

We built baseline models with the Kaldi speech recognition toolkit [25], (http://kaldi-asr.org). PF-STAR is labelled with long and short silences that are superfluous in Kaldi and were removed. We extract 39-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) on audio downsampled from 22.05kHz to 16kHz. Monophone and triphone GMM models were then trained using the MFCC features. Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) were applied to the features before a pass of speaker adaptive training with iMLLR. The number of Gaussians and leaves were optimised on held-out test data. The final GMM model had 2250 leaves and 12500 Gaussians. We then trained neural networks on top of the adapted features and 10 frames of context with a frame cross-entropy error function. The remaining setup is similar to the standard “net1” recipe in Kaldi, with six layers of sigmoid nonlinearities, but with 1024 units in a layer instead of the standard 2048. Our initial experiments indicated that this made no significant differences to the WERs. We performed Restricted Boltzmann Machine (RBM) pretraining and subsequently trained the network with a learning rate of 0.0008 with early stopping. As the corpus is small we also experimented with dropout [26] on all hidden layers, with a retention of 0.8 and twice the amount of epochs as our baseline model, but we were unable to achieve gains on top of RBM pretraining.

The results using the baseline PF-STAR models are shown

---

1Part of the corpus consists of lists of phonetically similar words – these were excluded, as we empirically found them to solely lead to substitution errors; for the mixture experiments it is more important to have well estimated adults’ models.

2http://www.mgb-challenge.org
in Table 1. The GMM model improves upon the previously reported score of 44.53% WER [18] by approximately 29.5% relative. The DNN system further improves the performance by 7.7% relative. Pretraining is reasonably effective resulting in about 1% absolute difference. Further experiments with various language models show that the majority of the reduction in WER compared to the previously published result is due to a stronger language model, as shown in Figure 2. Using a unigram language model or an equal probability language model, restricted to the data vocabulary, yields similar results to [18] for the GMM models, while the DNN models are a few percentage points better.

Table 1: % WER for the baseline models. The DNN model improves WER by approximately 7.7% relative.

<table>
<thead>
<tr>
<th></th>
<th>GMM</th>
<th>DNN (pre)</th>
<th>DNN (no pre)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PF-STAR</td>
<td>31.4</td>
<td>29.0</td>
<td>30.3</td>
</tr>
</tbody>
</table>

3.2. Un-modified mixtures

Figure 3 shows the results for iteratively adding adult speakers from WSJCAM0 to PF-STAR, training as above and decoding on PF-STAR. The results for the GMM systems corroborate the general consensus in literature discussed above: adding unmodified adults speech to a children’s speech corpus does not reduce WER on children’s speech, despite an increase in training data. The DNN models, however, demonstrate somewhat more robust performance with the increased data. WER mainly hovers around the baseline, at best 1.8% absolute better for a ratio of 0.9. The large fluctuations may be attributed to the small amount of data; a speaker selection scheme may ensure more consistent changes to WER.

The results for ABI are not as conclusive and somewhat more erratic than above (Figure 4). This may be due to the large variety of speakers in ABI. Yet, the results show a widening performance gap between the two types of models as the number of adult speakers increases.

These results suggest that the DNN models were in some cases able to discern useful features from the additional data that the GMMs could not, and that performance gains with adult speakers is possible, though strongly speaker dependent.

3.3. Stochastic Feature Mapping

The results for performing SFM on PF-STAR itself are shown in Table 2. There are no significant decreases in WER for any number of replicas; for three replicas, WER increases by a statistically significant 6.8% absolute. This is in stark contrast to the results in [17], where up to 2.9% absolute improvements were observed. This discrepancy may be explained by the nature of the PF-STAR corpus: SFM preserves the labels of transformed features, but there is considerable overlap in the utterances spoken within PF-STAR. Hence, transforming between speakers does not increase the variety - or label-space - in utterances of a given speaker. Instead it produces somewhat distorted duplicates of existing utterances.

Bringing in speakers from a different corpus will increase the label space. Results for SFM using WSJCAM0 and ABI for source speakers are shown in Tables 3 and 4. We also show re-
Table 2: In-domain data augmentation with PF-STAR

<table>
<thead>
<tr>
<th></th>
<th>% WER</th>
<th>ins</th>
<th>del</th>
<th>sub</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>29.0</td>
<td>1017</td>
<td>1046</td>
<td>3911</td>
</tr>
<tr>
<td>Baseline (no pre)</td>
<td>30.3</td>
<td>901</td>
<td>1191</td>
<td>4147</td>
</tr>
<tr>
<td>SFM-1 (pre)</td>
<td>29.4</td>
<td>1303</td>
<td>764</td>
<td>4001</td>
</tr>
<tr>
<td>SFM-1 (no pre)</td>
<td>30.1</td>
<td>1306</td>
<td>801</td>
<td>4158</td>
</tr>
<tr>
<td>SFM-1 (pre)</td>
<td>29.9</td>
<td>1400</td>
<td>724</td>
<td>4029</td>
</tr>
<tr>
<td>SFM-1 (no pre)</td>
<td>30.1</td>
<td>1263</td>
<td>898</td>
<td>4032</td>
</tr>
<tr>
<td>SFM-1 (pre)</td>
<td>35.8</td>
<td>2408</td>
<td>536</td>
<td>4435</td>
</tr>
<tr>
<td>SFM-1 (no pre)</td>
<td>32.5</td>
<td>1661</td>
<td>706</td>
<td>4331</td>
</tr>
</tbody>
</table>

Table 3: Out-of-domain data augmentation with WSJCAM0 (WER/%). Numbers in italic are not significantly different to the baseline (i.e. $p > 0.005$).

<table>
<thead>
<tr>
<th></th>
<th>Pre</th>
<th>No pre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>29.0</td>
<td>30.3</td>
</tr>
<tr>
<td>AUG 1</td>
<td>29.6</td>
<td>30.7</td>
</tr>
<tr>
<td>AUG 2</td>
<td>30.3</td>
<td>32.5</td>
</tr>
<tr>
<td>AUG 3</td>
<td>29.4</td>
<td>30.4</td>
</tr>
<tr>
<td>SFM 1</td>
<td>28.5</td>
<td>30.8</td>
</tr>
<tr>
<td>SFM 2</td>
<td>28.3</td>
<td>30.3</td>
</tr>
<tr>
<td>SFM 3</td>
<td>27.2</td>
<td>28.8</td>
</tr>
</tbody>
</table>

The novel use of out-of-domain SFM was shown to be more effective than in-domain SFM on PF-STAR. At best it produced 6.2% relative improvement with speakers from WSJCAM0 over the baseline of 29.0% WER. Augmenting the DNN models with un-modified features aligned on children’s GMM models proved not useful, resulting in a reduction of WER by 1.4% relative.

In future work, speaker selection schemes, such as the distance measure proposed in [27], may ensure more consistent WER reductions. It may also be used to cluster target speakers to afford more robust estimation of the transformation matrices. Other possibilities include employing multi-lingual children’s data through for example pre-training, tandem features, or domain adaptation of a larger adults’ model in a Multi-Level Adaptive Networks (MLAN) scheme [28].

4. Conclusions

In contrast to published literature, we have shown that DNN acoustic children’s models are able to make use of adults’ speech. Although, due to the small size of the children’s corpus, only a limited number of additional speakers may be selected, resulting in highly speaker dependent performance. This may explain previous findings.

The novel use of out-of-domain SFM was shown to be more effective than in-domain SFM on PF-STAR. At best it produced 6.2% relative improvement with speakers from WSJCAM0 over the baseline of 29.0% WER. Augmenting the DNN models with un-modified features aligned on children’s GMM models proved not useful, resulting in a reduction of WER by 1.4% relative.
5. References


