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

Citation for published version:
Fainberg, J, Bell, P, Lincoln, M & Renals, S 2016, Improving Children's Speech Recognition through Out-of-Domain Data Augmentation. in *Interspeech 2016.* Interspeech, pp. 1598-1602, Interspeech 2016, San Francisco, United States, 8/09/16. https://doi.org/10.21437/Interspeech.2016-1348

Digital Object Identifier (DOI):
10.21437/Interspeech.2016-1348

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Interspeech 2016

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

Joachim Fainberg[1], *Peter Bell*[1], *Mike Lincoln*[2], *Steve Renals*[1]

(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 WSJCAM0 and ABI. Our experimental results indicate that a DNN acoustic model for childrens 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 WSJCAM0 [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 typical 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)

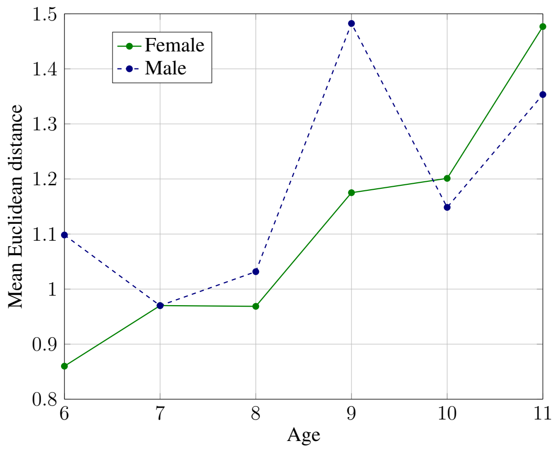

Figure 1: Mean Euclidean distance between MVNs of age- and gender-dependent monophone models trained on PF-STAR.

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 VTLP [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:

$$\mathbf{O}^{(t)} = \mathbf{A}(t,s)\mathbf{O}^{(s)} + \mathbf{b}(t,s), \tag{1}$$

where $\mathbf{O}^{(k)}$ denotes the features for speaker k, and $\mathbf{A}(t,s)$ and $\mathbf{b}(t,s)$ 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, λ_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, λ_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 source speaker given the corresponding speaker-dependent model of the target speaker: $\mathbf{A}(t,s)|\lambda_t$. 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[1].

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[2]. 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 fMLLR. 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 "nnet1" 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

[1] Part 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.

[2] `http://www.mgb-challenge.org`

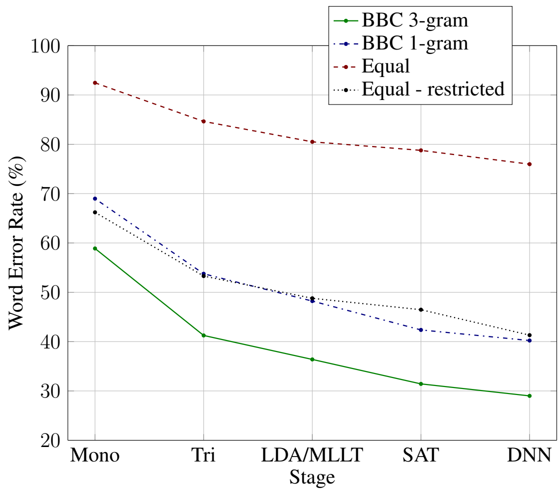

Figure 2: WER of baseline models for various training stages and language models.

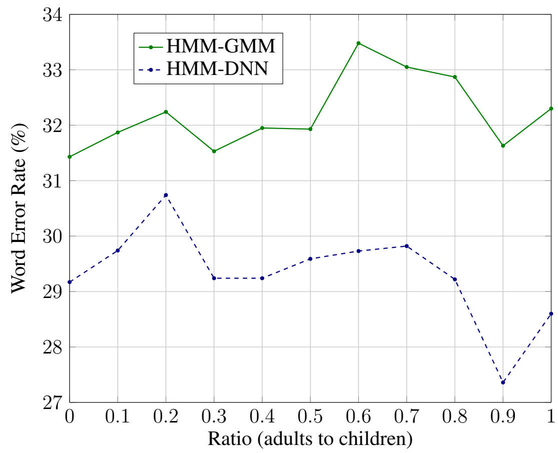

Figure 3: Un-modified augmentation of speakers from WSJ-CAM0 to PF-STAR, decoded on PF-STAR. The DNN system is more robust to out-of-domain data.

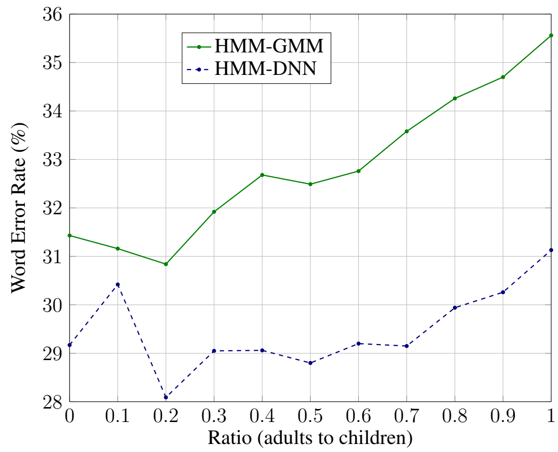

Figure 4: Un-modified augmentation of speakers from ABI to PF-STAR, decoded on PF-STAR.

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.

	GMM	DNN (pre)	DNN (no pre)
PF-STAR	31.4	29.0	30.3

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 constrast 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

	% WER	ins	del	sub
Baseline	29.0	1017	1046	3911
Baseline (no pre)	30.3	901	1191	4147
SFM-1 (pre)	29.4	1303	764	4001
2	29.9	1400	724	4029
3	35.8	2408	536	4435
SFM-1 (no pre)	30.7	1306	861	4158
2	30.1	1263	898	4032
3	32.5	1661	706	4331

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

	Pre	No pre
Baseline	29.0	30.3
AUG 1	*29.6*	30.7
2	30.3	32.5
3	*29.4*	30.4
SFM 1	*28.5*	30.8
2	*28.3*	30.3
3	**27.2**	*28.8*

sults for un-modified augmentation (AUG) with the same speakers and utterances as their SFM counterparts. These models are trained given GMM alignments of the PF-STAR baseline system, in contrast to the above experiments where data is mixed prior to GMM training.

SFM with WSJCAM0 yields at best 27.2% WER, a 6.2% relative improvement upon the baseline. There is some improvements for ABI with a single replica, but with $p = 0.005$, reflecting the lesser performance with ABI in the experiments above. The discrepancy between the two corpora may reflect the different results observed in [17] on different datasets, suggesting that the technique is quite dependent upon the given corpus. Un-modified augmentation yields no improvements. We note the strong speaker dependency shown above, however, out-of-domain SFM generally outperforms un-modified augmentation for the same exact subset of speakers.

It is interesting to note that pretraining is crucial to the observed improvements: either technique without pretraining yields on average worse results than the corresponding baseline. This suggests that pretraining is key to making use of the additional data, and perhaps provides most of the gain.

To gain a better understanding of the difference in results between in-domain and out-of-domain SFM on this data, we investigated the fMLLR matrices generated between speakers

Table 4: Out-of-domain data augmentation with ABI (WER/%). Numbers in italic are not significantly different to the baseline (i.e. $p > 0.005$).

	Pre	No pre
Baseline	29.0	30.3
AUG 1	*29.8*	31.2
2	*29.7*	30.9
3	*29.8*	30.7
SFM 1	**28.0**	*29.5*
2	*28.9*	30.0
3	*30.0*	30.7

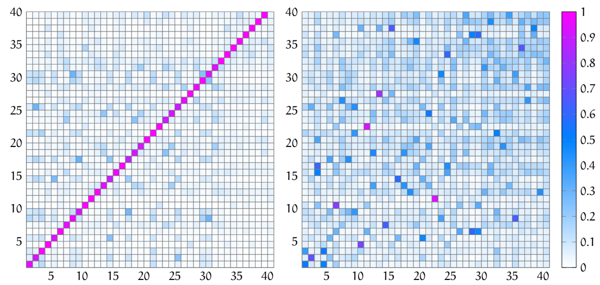

Figure 5: Average fMLLR matrices from performing SFM on in-domain PF-STAR data (left) and out-of-domain WSJCAM0 data (right).

in the SFM algorithm. Figure 5 shows element-wise averages of 80 absolute-value 40x41 fMLLR matrices corresponding to 40-dimensional LDA features for in-domain (left) and out-of-domain (right) transformations, respectively. The transformations with WSJCAM0 are pronounced and varied, while those with PF-STAR are very subtle. This may explain the lack of improvements: the algorithm is practically duplicating existing speakers, albeit slightly modified or distorted.

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 at best 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].

5. References

[1] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani, M. Gerosa, C. Hacker, M. J. Russell, S. Steidl, and M. Wong, "The PF_STAR children's speech corpus," in *INTERSPEECH*, 2005, pp. 2761–2764.

[2] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAMO: a British English speech corpus for large vocabulary continuous speech recognition," in *, 1995 International Conference on Acoustics, Speech, and Signal Processing, 1995. ICASSP-95*, vol. 1, 1995, pp. 81–84 vol.1.

[3] J. Godfrey and E. Holliman, "Switchboard-1 Release 2 LDC97S62." Philadelphia: Linguistic Data Consortium, 1993.

[4] H. Liao, G. Pundak, O. Siohan, M. K. Carroll, N. Coccaro, Q.-M. Jiang, T. N. Sainath, A. Senior, F. Beaufays, and M. Bacchiani, "Large vocabulary automatic speech recognition for children," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[5] J. Wilpon and C. Jacobsen, "A study of speech recognition for children and the elderly," in *, 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings*, vol. 1, 1996, pp. 349–352 vol. 1.

[6] M. J. Russell and S. D'Arcy, "Challenges for computer recognition of children's speech." in *SLaTE*, 2007, pp. 108–111.

[7] R. Mugitani and S. Hiroya, "Development of vocal tract and acoustic features in children," *Acoustical Science and Technology*, vol. 33, no. 4, pp. 215–220, 2012.

[8] M. Gerosa, D. Giuliani, and F. Brugnara, "Acoustic variability and automatic recognition of childrens speech," *Speech Communication*, vol. 49, no. 10-11, pp. 847–860, 2007.

[9] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 603–616, 2003.

[10] P. G. Shivakumar, A. Potamianos, S. Lee, and S. Narayanan, "Improving speech recognition for children using acoustic adaptation and pronunciation modeling," *Proceedings of Workshop on Child, Computer and Interaction*, 2014.

[11] S. Panchapagesan and A. Alwan, "Multi-parameter frequency warping for VTLN by gradient search," in *2006 IEEE International Conference on Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings*, vol. 1, 2006, pp. I–I.

[12] L. Saheer, J. Yamagishi, P. Garner, and J. Dines, "Combining vocal tract length normalization with hierarchical linear transformations," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 262–272, 2014.

[13] S. Lee, A. Potamianos, and S. Narayanan, "Acoustics of childrens speech: Developmental changes of temporal and spectral parameters," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.

[14] D. Elenius and M. Blomberg, "Adaptation and normalization experiments in speech recognition for 4 to 8 year old children." in *INTERSPEECH*, 2005, pp. 2749–2752.

[15] R. Serizel and D. Giuliani, "Deep neural network adaptation for childrens and adults speech recognition," in *Proc. of the First Italian Computational Linguistics Conference*, 2014.

[16] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, 2013.

[17] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 5582–5586.

[18] S. M. DArcy, L. P. Wong, and M. J. Russell, "Recognition of read and spontaneous childrens speech using two new corpora," in *Proc. of ICSLP*, 2004, pp. 1473–1476.

[19] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proceedings of INTERSPEECH*, 2015.

[20] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[21] S. M. DArcy, M. J. Russell, S. R. Browning, and M. J. Tomlinson, "The accents of the British Isles (ABI) corpus," *Proceedings Modlisations pour lIdentification des Langues*, pp. 115–119, 2004.

[22] P. Bell, M. J. F. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester *et al.*, "The MGB Challenge: evaluating multi-genre broadcast media recognition," *Proc. of ASRU, Arizona, USA*, 2015.

[23] National Institute of Standards and Technology (NIST), "Speech recognition scoring toolkit (SCTK) version 2.4.0." 2010.

[24] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *1989 International Conference on Acoustics, Speech, and Signal Processing, 1989. ICASSP-89*, 1989, pp. 532–535 vol.1.

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," 2011.

[26] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[27] R. C. Vipperla, S. Renals, and J. Frankel, "Augmentation of adaptation data," *Proc. Interspeech*, 2010.

[28] P. Bell, P. Swietojanski, and S. Renals, "Multi-level adaptive networks in tandem and hybrid ASR systems," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6975–6979.