INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Improving child speech disorder assessment by incorporating out-of-domain adult speech

Daniel Smith[1], *Alex Sneddon*[2], *Lauren Ward*[3], *Andreas Duenser*[1], *Jill Freyne*[4], *David Silvera-Tawil*[4], *Angela Morgan*[5]

[1]Data61, CSIRO, Hobart, Australia
[2]University of Sydney, Sydney, Australia
[3]Acoustics Research Centre, University of Salford, Manchester, U.K.
[4]Health and Biosecurity, CSIRO, Sydney, Australia
[5]Murdoch Childrens Research Institute, Melbourne, Australia
andreas.duenser@data61.csiro.au

Abstract

This paper describes the continued development of a system to provide early assessment of speech development issues in children and better triaging to professional services. Whilst corpora of children's speech are increasingly available, recognition of disordered children's speech is still a data-scarce task. Transfer learning methods have been shown to be effective at leveraging out-of-domain data to improve ASR performance in similar data-scarce applications. This paper combines transfer learning, with previously developed methods for constrained decoding based on expert speech pathology knowledge and knowledge of the target text. Results of this study show that transfer learning with out-of-domain adult speech can improve phoneme recognition for disordered children's speech. Specifically, a Deep Neural Network (DNN) trained on adult speech and fine-tuned on a corpus of disordered children's speech reduced the phoneme error rate (PER) of a DNN trained on a children's corpus from 16.3% to 14.2%. Furthermore, this fine-tuned DNN also improved the performance of a Hierarchal Neural Network based acoustic model previously used by the system with a PER of 19.3%. We close with a discussion of our planned future developments of the system.

Index Terms: Automated Speech Recognition, Speech Therapy, Speech Assessment Tools

1. Introduction

In Australia as many as one in every twenty preschool-aged children have a speech disorder [1]. Childhood speech disorders are similarly prevalent in other countries [2, 3]. Untreated, childhood speech and language difficulties may have a long term, negative effect on a person educationally, vocationally, and socially [4]. It has been shown that the best predictor for children's later speech outcomes is the type of phonological errors made when they are young [5, 6]. To ensure the best outcomes for children, identification of error patterns and treatment is required at an early age [7]. The process of identification and triaging children to receive therapy represents a significant practical challenge, given not only the shortage of Speech and Language Pathologists (SLP) in many countries but their uneven distribution [8]. The application of automatic speech recognition (ASR), to assess and identify children with high-risk error patterns, has the potential to not only reduce burden on SLP but help ensure that the children most in need receive the appropriate clinical treatment.

There has been an increasing body of research investigating the application of ASR to speech and language pathology. Much of this research investigates very specific applications, such as stuttering [9], dysarthria [10, 11] and Parkinson's Disease [12]. Research has also investigated the application of ASR to children's speech development, assessing autism spectrum disorders [13], childhood apraxia of speech [14], cleft lip and palate [15] and general language development [16]. Despite the increasing abundance of such research, the focus remains primarily on therapy rather than screening, with very little work addressing the need for a broadly applicable screening tool [16, 17].

A key challenge which has constrained clinical ASR is the scarcity of data available for the target population. A number of approaches have been leveraged to deal with this scarcity. In adult speech articulation entropy has been used, which is a proxy measure for the number of distinct phonemes a person produces [12]. For children's speech acoustic features like pause events, which are more robust to the variability, have been successfully used with connected speech [16]. Whilst these approaches have been shown to be effective for segregating between individuals with broadly normal and abnormal language production, they cannot identify specific articulation errors and are not suitable for young children. Other approaches have exploited the fact that many assessments of speech and articulation are made with known texts. The improvement in recognition this can yield for clinical ASR has been demonstrated previously by this group [17], where an improvement of 33.1% in phoneme recognition was achieved when the target text was exploited by the constrained decoder. Other groups have had similar success with this approach, improving detection of pathological voice by 20% [18].

Deep Neural Networks (DNN) currently offer state of the art performance for ASR. Due to the high complexity of the models, however, significant amounts of annotated speech data are often required for training (i.e. thousands of hours of speech). For data-scarce tasks, this can often lead to the overfitting of acoustic models. Transfer learning methods address this issue by training the DNN on a larger, out-of-domain corpora and then re-purposing the learnt features to train a network on a smaller, in-domain, dataset. Such methods have been shown to be effective for low-resource languages [11, 19] and disordered adult speech [20]. For low-resource languages, an improvement of 10-30% has been shown when the underlying DNN is trained on the higher-resource language and then the language specific parameters are estimated from the low-resource language [19]. Applications in dysarthric speech have

http://dx.doi.org/10.21437/Interspeech.2017-455

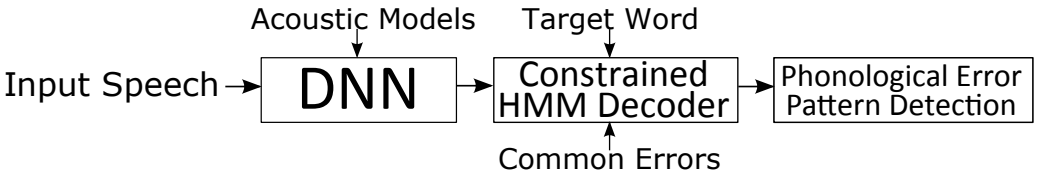

Figure 1: *Architecture of the system*

yielded performance improvements of a similar magnitude, utilising out-of-domain varieties of Dutch to train a DNN which is then adapted to pathological Flemish speech [11]. Stochastic feature mapping has also shown success for recognition of children's speech at the word level, utilising out-of-domain adult speech [21].

In this paper, we investigate whether an approach utilising transfer learning combined with previously developed methods for constrained decoding based on knowledge of the target text [17], can be applied to improve the detection of phonological errors in children's speech. A large and readily available dataset of adult Australian speech is used to train a base DNN, which is then transferred to train our target DNN of disordered Australian children's speech.

2. Automated Assessment System

Figure 1 presents the three stage assessment system (the *'proof of concept'* which was developed in [17] and [22]). This paper presents improvements to the initial stage, whilst maintaining the novel constrained HMM decoder and phonological error pattern (PEP) detection stages from previous work.

The clinical screening protocol upon which this prototype system is based is the Diagnostic Evaluation of Articulation and Phonology (DEAP) test [23]. Utilising a validated protocol as the basis for this system ensures the diagnostic value of the target words. Furthermore, this knowledge of the target words forms the basis of the constrained HMM decoder.

2.1. Acoustic Model

The first stage consists of an acoustic model of children's phonemes. The input to these models is speech made up of target words in isolation, elicited through a picture naming task. The acoustic model used in the previous system [17] used a hierarchical neural network (HNN) with long temporal context features over a 310ms window. This HNN cascaded a pair of neural networks (NN) into a third NN, with the third NN trained on the concatenated posterior probabilities of the first two NN. This model is used as a baseline reference in the current work. The acoustic model proposed in this work replaces the HNN with a DNN comprised of N fully connected layers with an input layer, N-2 hidden layers and an output layer composed of a softmax classifier. A number of different base and target DNN designs were considered in the study and selected by tuning with the validation data set. This design process will be described in greater detail in section 3.

2.2. Constrained HMM Decoder

The second stage is the constrained Hidden Markov Model (HMM) decoder, which was manually constructed from prior knowledge of the target words and expert SLP knowledge about likely phonological error patterns. The Viterbi decoding algorithm is used to infer the most likely phoneme sequence for each test word. Development of this decoder is presented in [17].

2.3. Phonological Error Pattern (PEP) Detection

Finally, the most likely sequence of phonemes is passed through a decision tree to detect PEPs. First the sequence is compared with the target to determine whether an articulation error has occurred (insertion, deletion or substitution of an expected phoneme). Then, utilising other normative data and information about the child, the typicality of the error can be assessed. For example, at age 3 years a fronting error where \TH\ is substituted for \F\ in the word *'teeth'* is considered a typical error, however the backing error \TH\→ \S\ is considered atypical. As atypical errors are indicative of a higher risk of speech delay, knowledge of the typicality can be leveraged to determine whether to recommend that the child to be further assessed by a human SLP.

2.4. Speech Corpora

This work utilises a large out-of-domain corpus of Australian English adult speech, the AusTalk corpus [24] and a much smaller target corpus of disordered children's speech from the Murdoch Children's Research Institute (MCRI).

The AusTalk corpus contains modern Australian English of 861 adult speakers aged between 18 and 83, acquired from both cities and regional areas across Australia. Data from each speaker was collected during three independent one-hour recording sessions consisting of both read and spontaneous speech tasks. Spontaneous and prompted speech each comprise approximately 50% of each recording. We used a subset of data from 109 speakers for which speech has been transcribed and annotated. The data from 98 speakers were used for training the DNN models, whilst the data from 11 speakers was reserved for validation (i.e. optimisation of model parameters).

The MCRI corpus consists of 114 unique child speakers, aged between 3 and 14 years, collected by expert SLPs from the MCRI. The data includes correct and misarticulated word samples as evaluated by an expert SLP. Only recordings deemed to be 'good' (word clearly intelligible, with low background noise and an undistorted recording) were used. A sub-set of 10 isolated words exhibiting typical PEPs from the DEAP Phonology, Inconsistency and Articulation sub-tests were selected to test the proof of concept system. In total 1173 words were used, 39.50% which were misarticulated. These words represent 21 of the 39 phonemes used in Australian English.

3. Experiments

We performed experiments upon different acoustic models to test the efficacy of phoneme recognition and PEP detection in the automated assessment system. In this paper we explore different approaches to train the target network. A common characteristic of each approach involved replacing the output layer of the base network with a new softmax classifier that was then re-trained upon the children's phoneme classes. The input and hidden layers of the base network were then used to:

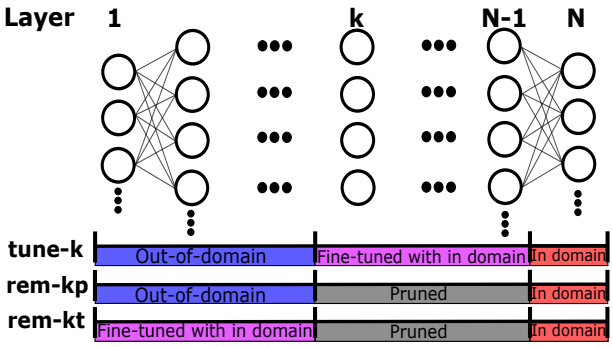

Figure 2: *DNN acoustic models, noting layers using out-of-domain adult and in domain children's speech data*

- **Fine-tune:** The top k layers of the network (excluding the output layer) were fine-tuned with children's speech and the remaining N-1-k hidden layers were preserved. Fine-tuning was performed upon the top k = 0 to N-1 layers of each network and defined as **tune-k**.

- **Prune:** The top k hidden layers of the network were removed, whilst the remaining N-1-k layers were either preserved or fine-tuned with children's speech. Pruning was performed upon the top k = 1 to N-2 layers of each network and defined as **rem-kp** when the remaining layers were preserved or **rem-kt** when the remaining layers were fine-tuned.

Speech signals were split into 25ms frames with 15ms overlap. There were 13 Mel Frequency Cepstral Coefficients and 13 Delta Coefficients extracted from each frame. These were used as the input vectors to the acoustic model. Temporal context was added to the input vectors by concatenating an additional 7 frames on either side of the central frame (15 frame context).

The out-of-domain DNN was trained and validated upon the AusTalk speech corpus using 294 hours of speech for training and 11 hours of speech for validation. A greedy search was used to select a set of 12 parameters that minimise the frame error rate (FER) of the AusTalk validation set. The parameters that were tuned during the network construction were classed as being either optimisation (method, weight initialisation, learning rate, learning decay, decay step, drop-out, regularisation, weight decay and batch size) or architecture (number of hidden layers and number of activation units/layer) based.

The target DNN were then trained, validated and tested upon the MCRI corpus using an 80%, 10% and 10% split of its 1173 utterances, respectively. Given that the out-of-domain DNN was used as the base for training the target DNN, it was only possible to tune a small number of parameters. Only five of the optimisation based parameters (learning rate, learning decay, drop-out, regularisation and weight decay) were part of the greedy search and used to minimise the FER of the MCRI validation set. This same process was used to produce the different network architectures for transfer learning.

Once the acoustic models were tuned, the constrained HMM decoder was then optimised. The transition weights of the constrained decoder, which are described in [17], were selected using a three dimensional grid search to minimise the PER of the MCRI validation set. Finally, the optimised assessment system was applied to the MCRI test set to evaluate its performance.

3.1. Results and Discussion

The out-of-domain DNN selected as the base for training the children's phoneme classes produced a minimum FER of 20.5% on the AusTalk validation set. As this base network possessed five layers overall (N=5), fine-tuning was performed on between zero to four layers. Also, one to three hidden layers were pruned from the base network with the remaining layers either being fine-tuned or frozen. Consequently, there were eleven different versions of the target DNN to investigate using transfer learning.

3.1.1. Acoustic models

Figure 3 shows the FER and PER of the eleven different DNNs using transfer learning. The pruned networks where the top one or two hidden layers were removed and the remaining layers were fine-tuned offered the highest performance across all of the networks. The **rem-2t** network had a slightly lower FER than the **rem-1t** network, and hence, achieved the best overall performance with an FER and PER of 32.5% and 14.2%, respectively. This result demonstrated that shallow target networks of 1 to 2 hidden layers offered superior performance for the children's phoneme recognition task. As the the base network was trained with an adult speech corpus that was significantly larger than our target corpus of disordered children's speech, during fine-tuning, there will be a tendency for the target speech to overfit a network of the same depth as the base model [25].

The DNNs that had a majority of their network layers being fine-tuned offered a performance advantage over the DNNs that either had all or a majority of their network layers being preserved. As features became more generic at the lower levels of a network, it is often found these more generic features can be successfully transferred to a closely related task without requiring adaptation [25]. Figure 3 shows this is not the case in this study, given there is a monotonic increasing relationship between the recognition performance and number of layers being fine-tuned. An improvement in the performance continued until all of the hidden layers were fine-tuned (the **tune-3** network with a FER of 34.4% and PER of 14.8%) suggesting that the adult and children's corpus were not closely related in terms of their content.

The networks using transfer learning were benchmarked against two alternative acoustic models. Figure 4 reveals that the **rem-2t** network reduced the PER of the best performing DNN (strictly trained on the MCRI corpus of children's speech) from 16.3% to 14.2%. Furthermore, the **rem-2t** network offered an improvement over the HNN that was adopted in our previous assessment system [17] with a reported PER of 19.3%.

3.1.2. PEP Detection

Test words of eight children speakers were selected from the MCRI corpus to evaluate the extent to which the system can detect specific PEPs. These PEPs have been used by expert SLPs to diagnose four of these speakers as having low risk speech (all words were correctly pronounced) and the other four speakers as having moderate risk speech (fronting error in the word *'teeth'* (\TH \→ \F \) and gliding errors in either or both of the test words, *'rain'* (\R \→ \W \) and *'girl'* (\L \→ \W \)).

Our system (with the **rem-2t** network) was applied to eight test words, ensuring these speakers were excluded from training and validation of the acoustic model. Table 1 shows that for each of the four speakers diagnosed with moderate risk speech, at least one fronting or gliding error was identified by the sys-

Table 1: *The system was evaluated by identifying the percentage of test words that were correctly classified as having a gliding error, a fronting error or as being correctly pronounced for 8 children speakers selected from the MCRI corpus. Speakers 1-4 were assessed by a SLP as being at low risk of speech development issues given there were no PEPs associated with its set of 8 test words. Speakers 5-8 were assessed by a SLP as having moderate risk speech due to a fronting error being identified in the test word 'teeth', and for speakers 5, 6 and 8, a gliding error being found in either or both of the test words, 'rain' or 'girl'.*

	Speaker							
	1	2	3	4	5	6	7	8
Correctly Pronounced	75	62.5	100	100	75	100	100	87.5
Fronting	N/A	N/A	N/A	N/A	100	100	100	0
Gliding	N/A	N/A	N/A	N/A	50	100	N/A	50

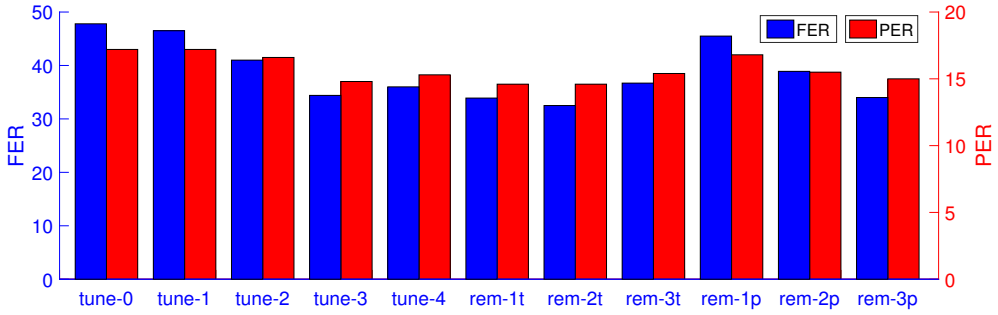

Figure 3: *The DNNs that use different fine tuning and pruning strategies were compared to retrain a base model of adult speech for disordered children's speech. The DNNs were evaluated on the MCRI test set with the FER and PER.*

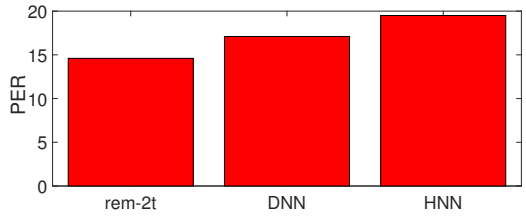

Figure 4: *A comparison of three different types of acoustic models; the HNN previously used in the assessment tool, a DNN strictly trained with children's speech and the* ***rem-2t*** *network that used transfer learning.*

tem. For speakers 6 and 7, every PEP was correctly identified. Of the four speakers with low risk speech, all eight test words were classified as being correctly pronounced for two of the speakers (3 and 4). Speakers one and two were incorrectly classified to have two and three PEPs, respectively.

These results suggest that for a tool based on our current approach, it could be more challenging to correctly assess speakers with low risk speech as opposed to speakers with high risk speech. To assess speakers as having low risk speech, the phoneme recognition model may need to ensure that each of the correctly pronounced test words are classified accurately. Whilst for a speaker to be assessed as high risk, not every atypical PEP in the test words need to be correctly classified. In some cases, only one atypical PEP may need to be classified correctly. Whilst this is not an ideal situation, it is still preferable to the reverse scenario where there is an increased likelihood of children with high risk speech being missed by the system.

4. Conclusions

We are developing a new proof of concept screening system to provide assessment of young children's speech development. This system combines a new acoustic model with existing methods for HMM decoding based upon knowledge of the test words. We attempt to address the scarcity of disordered children's speech that is available for training an acoustic model by leveraging an out-of-domain DNN of adult speech to train our target DNN of children's phonemes. Results showed that adopting transfer learning improved the PER compared to a DNN strictly trained on children's speech and a HNN used in our previous system. Furthermore, we performed experiments to assess the system's ability to detect PEPs in the test words of eight children that had been diagnosed by a human SLP. It was shown that typical PEPs were detected for all four children assessed as having moderate risk speech. For two of the four speakers assessed as having low risk speech, the system correctly classifies each of the phonemes as having the right pronunciation.

Future work will involve attempting to improve the acoustic models by collecting additional disordered children's speech for training and identifying other sources of out-of-domain speech (i.e. other child speech corpora) that may be more transferable to our assessment application than the adult speech from AusTalk. Further integration of expert SLP knowledge is required for full implementation of the decision support stage. This will include a knowledge base of developmental norms which along with the detected PEPs, age and gender of the child will allow for effective risk assessments to be made. In addition, we are in the process of developing an end-user interface for the assessment tool and are testing it with the target user group (early childhood teachers and carers).

5. Acknowledgements

Angela Morgan is funded by the National Health and Medical Research Council (NHMRC) Practitioner Fellowship (#1105008), NHMRC Centre of Research Excellence in Speech and Language NeurobioloGy (CRE-SLANG; #1116976) and NHMRC Project grant (#1127144).

6. References

[1] P. Eadie, A. Morgan, O. C. Ukoumunne, K. Ttofari Eecen, M. Wake, and S. Reilly, "Speech sound disorder at 4 years: prevalence, comorbidities, and predictors in a community cohort of children," *Developmental Medicine & Child Neurology*, vol. 57, no. 6, pp. 578–584, 2015.

[2] U.S. Department of Education. (2014) 36th annual report to congress on the implementation of the individuals with disabilities education act, 2014. [Online]. Available: http://www2.ed.gov/about/reports/annual/ osep/2014/parts-b-c/36th-idea-arc.pdf

[3] S. A. Karbasi, R. Fallah, and M. Golestan, "The prevalence of speech disorder in primary school students in yazd-iran," *Acta Medica Iranica*, vol. 49, no. 1, p. 33, 2011.

[4] S. McLeod, L. J. Harrison, L. McAllister, and J. McCormack, "Speech sound disorders in a community study of preschool children," *American Journal of Speech-Language Pathology*, vol. 22, no. 3, pp. 503–522, 2013.

[5] A. Morgan, K. Ttoffari Eecen, A. Pezic, K. Brommeyer, C. Mei, P. Eadie, S. Reilly, and B. Dodd, "Who to refer for speech therapy at 4 years of age versus who to watch and wait?" *Journal of Pediatrics*, 2017.

[6] J. G. Foy and V. A. Mann, "Speech production deficits in early readers: Predictors of risk," *Reading and Writing*, vol. 25, no. 4, pp. 799–830, 2012.

[7] H. M. Sharp and K. Hillenbrand, "Speech and language development and disorders in children," *Pediatric Clinics of North America*, vol. 55, no. 5, pp. 1159–1173, 2008.

[8] P. A. Mashima and C. R. Doarn, "Overview of telehealth activities in speech-language pathology," *Telemedicine and e-Health*, vol. 14, no. 10, pp. 1101–1117, 2008.

[9] P. A. Heeman, R. Lunsford, A. McMillin, and J. S. Yaruss, "Using clinician annotations to improve automatic speech recognition of stuttered speech," in *Proc. Interspeech 2016: 17th Annual Conf. of International Speech Communication Association*. San Francisco, U.S.A.: ISCA, 2016, pp. 2651–2655.

[10] M. Ganzeboom, E. Yılmaz, C. Cucchiarini, and H. Strik, "An asr-based interactive game for speech therapy," 2016.

[11] E. Yılmaz, M. Ganzeboom, C. Cucchiarini, and H. Strik, "Combining non-pathological data of different language varieties to improve dnn-hmm performance on pathological speech," pp. 218–222, 2016.

[12] Y. Jiao, V. Berisha, J. Liss, S.-C. Hsu, E. Levy, and M. McAuliffe, "Articulation entropy: An unsupervised measure of articulatory precision," *IEEE Signal Processing Letters*, 2016.

[13] J. R. Dykstra, M. G. Sabatos-DeVito, D. W. Irvin, B. A. Boyd, K. A. Hume, and S. L. Odom, "Using the language environment analysis (lena) system in preschool classrooms with children with autism spectrum disorders," *Autism*, vol. 17, no. 5, pp. 582–594, 2013.

[14] M. Shahin, B. Ahmed, A. Parnandi, V. Karappa, J. McKechnie, K. J. Ballard, and R. Gutierrez-Osuna, "Tabby talks: An automated tool for the assessment of childhood apraxia of speech," *Speech Communication*, vol. 70, pp. 49–64, 2015.

[15] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, and E. Nöth, "PEAKS–a system for the automatic evaluation of voice and speech disorders," *Speech Communication*, vol. 51, no. 5, pp. 425–437, 2009.

[16] J. J. Gong, M. Gong, D. Levy-Lambert, J. R. Green, T. P. Hogan, and J. V. Guttag, "Towards an automated screening tool for developmental speech and language impairments," in *Proc. Interspeech 2016: 17th Annual Conf. of International Speech Communication Association*. San Francisco, U.S.A.: ISCA, 2016, pp. 112–116.

[17] L. Ward, A. Stefani, D. Smith, A. Duenser, J. Freyne, B. Dodd, and A. Morgan, "Automated screening of speech development issues in children by identifying phonological error patterns," in *Proc. Interspeech 2016: 17th Annual Conf. of International Speech Communication Association*. San Francisco, U.S.A.: ISCA, 2016.

[18] G. K. Anumanchipalli, H. Meinedo, M. Bugalho, I. Trancoso, L. C. Oliveira, and A. W. Black, "Text-dependent pathological voice detection." in *Proc. Interspeech 2012: 13th Annual Conf. of International Speech Communication Association*. Portland, U.S.A.: ISCA, 2012, pp. 530–533.

[19] B. Abraham, S. Umesh, and N. M. Joy, "Overcoming data sparsity in acoustic modeling of low-resource language by borrowing data and model parameters from high-resource languages," pp. 3037–3041, 2016.

[20] H. Christensen, M. Aniol, P. Bell, P. D. Green, T. Hain, S. King, and P. Swietojanski, "Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech." in *Proc. Interspeech 2013: 14th Annual Conf. of International Speech Communication Association*. Lyon, France: ISCA, 2013, pp. 3642–3645.

[21] J. Fainberg, P. Bell, M. Lincoln, and S. Renals, "Improving childrens speech recognition through out-of-domain data augmentation," in *Proc. Interspeech 2016: 17th Annual Conf. of International Speech Communication Association*. San Francisco, U.S.A.: ISCA, 2016, pp. 1598–1602.

[22] A. Duenser, L. Ward, A. Stefani, D. Smith, J. Freyne, A. Morgan, and B. Dodd, "Feasibility of technology enabled speech disorder screening," in *Digital Health Innovation for Consumers, Clinicians, Connectivity and Community: Selected Papers from the 24th Australian National Health Informatics Conference (HIC 2016)*, vol. 227. IOS Press, 2016, p. 21.

[23] B. Dodd, H. Zhu, S. Crosbie, A. Holm, and A. Ozanne, *Diagnostic evaluation of articulation and phonology (DEAP)*. London: Psychology Corporation, 2002.

[24] D. Estival, S. Cassidy, F. Cox, D. Burnham *et al.*, "Austalk: an audio-visual corpus of australian english." in *LREC*, 2014, pp. 3105–3109.

[25] J. Yosinki, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *Advances in Neural Information Processing Systems*, vol. 27, pp. 3320–3328, 2014.