INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Automated Screening of Speech Development Issues in Children By Identifying Phonological Error Patterns

Lauren Ward[1,2], *Alessandro Stefani*[1,3], *Daniel Smith*[1], *Andreas Duenser*[1], *Jill Freyne*[4], *Barbara Dodd*[5], *Angela Morgan*[6]

[1] Data61, CSIRO, Hobart, Australia
[2] Acoustics Research Centre, University of Salford, Manchester, UK
[3] RMIT, Melbourne, Australia
[4] Health & Biosecurity, CSIRO, Sydney, Australia
[5] University of Melbourne, Melbourne, Australia
[6] Murdoch Children's Research Institute, Melbourne, Australia
andreas.duenser@csiro.au

Abstract

A proof of concept system is developed to provide a broad assessment of speech development issues in children. It has been designed to enable non-experts to complete an initial screening of children's speech with the aim of reducing the workload on Speech Language Pathology services. The system was composed of an acoustic model trained by neural networks with split temporal context features and a constrained HMM encoded with the knowledge of Speech Language Pathologists. Results demonstrated the system was able to improve PER by 33% compared with standard HMM decoders, with a minimum PER of 19.03% achieved. Identification of Phonological Error Patterns with up to 94% accuracy was achieved despite utilizing only a small corpus of disordered speech from Australian children. These results indicate the proposed system is viable and the direction of further development are outlined in the paper.
Index Terms: Automated Speech Recognition, Phonological Error Patterns, Speech Therapy, Speech Assessment Tools

1. Introduction

In 2011 the ratio of Speech and Language Pathologists (SLPs) to the general population in Australia was just over 1:5000 [1, 2]. When compared to the one in twenty Australian children presenting to school with a speech disorder [3, 4], this ratio highlights a vast disparity between supply and demand. This issue is further exacerbated in the public sector, where it has been reported that 25% of children wait more than 6 months and 18% wait more than 1 year to access an initial assessment [5], with wealthier families more likely to access these services [6]. These factors significantly impact a familys ability to optimize their childs speech and language development. Better service allocation to the most needy could be achieved if clinical technologies were developed, using Automatic Speech Recognition (ASR), to automate initial assessment. Furthermore, such technologies would enable more efficient assessment in countries without SLP shortages and provide significant benefit to countries, such as China, with more severe skill shortages [7].

Previous work in clinical applications of ASR have focused on particular disorders, such as childhood apraxia of speech (CAS) [8]. The aim of this system is to facilitate remote delivery and the monitoring of therapy by skilled SLPs, achieved through modules which detect voice activity and verify lexical stress and pronunciation. Automated systems have also been used to detect dysarthria [9], a motor speech disorder, by identifying the acoustic landmarks where dysarthria affected speech differs from normal speech. Additionally, systems have been developed to detect vocal fold disorders that are symptomatic of particular diseases i.e. throat cancer [10, 11, 12]. Features representing pitch variation and pitch amplitude are used to detect the severity of voice pathologies using supervised classifiers.

ASR models have also been used to provide feedback to patients with speech development issues in therapy applications [13, 14]. The pronunciation quality of test phonemes are measured relative to the posterior probabilities produced by the acoustic models. Whilst suitable for therapy, these approaches are not informative enough for a screening applications which not only need to identify when a phoneme is mispronounced, but what phonological error pattern (PEP) has occurred (whether phonemes are substituted, inserted or deleted from the target word). Studies show that the error type, not just the presence of an error, is an important diagnostic criteria for prognosticating about later speech and literacy outcomes [15, 16].

Existing work, with its focus on SLP monitoring and intervention for specific conditions, has not addressed the need for a broad initial assessment tool. To this end, we have developed a proof of concept system to identify and evaluate the type of PEPs in children's speech. Our system differs from existing work in that it aims at broad assessment not diagnosis of a specific condition. Through integrating expert SLP knowledge, our system accommodates usage by non-experts to alleviate the burden on SLPs. The proposed system aims to triage children for professional SLP assessment based on whether their PEP are low-risk (not present or age-appropriate), moderate risk (typical but delayed more than 6 months) or high risk (atypical). Furthermore, through making available the more detailed outputs of the system to trained SLPs, pre-appointment assessment could take place to streamline the service delivery.

As standard ASR approaches utilize HMM decoders trained on large corpora, one of the challenges in developing such a system has been the scarcity of speech data that is available to represent the target population of Australian children. Consequently we have developed a constrained HMM decoder, based on expert SLP knowledge, to exploit a small, but representative, speech corpus of Australian children with Disordered Speech.

http://dx.doi.org/10.21437/Interspeech.2016-850

This paper outlines the architecture of the developed proof of concept system in Section 2. Section 3 presents the results of experiments performed on this system with the conclusions drawn from which are summarized in Section 4.

2. Architecture and Approach

Figure 1 presents the proposed three stage architecture. Input speech is elicited from children utilizing a picture naming task for a set of specific, diagnostically relevant target words. These are based on a commonly used SLP screening protocol from the Diagnostic Evaluation of Articulation and Phonology (DEAP) test [17]. Using this validated protocol as a basis ensures not only clinical relevance but by significantly restricting the dictionary of target words, facilitates the use of knowledge-driven methods.

2.1. Hierarchical Neural Network (HNN)

The acoustic models used in our system are trained by a HNN with split temporal context features. These models were developed for the phoneme recognition system outlined in [18, 19] and were used given its superior performance to Gaussian Mixture Models for small data sets and for data acquired in low signal to noise conditions [19]. The speech signal is split into 25ms frames with 15ms overlap and 13 Mel Frequency Cepstral Coefficients (MFCC) are extracted from each frame. Longer temporal patterns of the MFCC vectors are then modeled for 310ms (31 feature vectors) and split into left (the feature vectors from 0 - 15) and right contexts (feature vectors 15 - 30). The right and left contexts are then processed independently; a half hamming window is applied to emphasize the middle vectors then the Discrete Cosine Transform is used to reduce the dimensionality of each context down to 11 coefficients. Both right and left contexts become the input vectors to separate three layer perceptron classifiers. The output vectors of the two classifiers (posterior probabilities) are concatenated and the log of these probabilities are fed as an input to a third "merged" three layer perceptron classifier. The "merged" classifier maps the posteriors of the left and right temporal contexts to the phoneme classes producing a new set of posterior probabilities that are used to decode the phoneme sequence of each word.

The three neural network classifiers used in the HNN model were trained with classical back propagation. There were 500 neurons in their hidden layer and softmax non-linearity was used to produce the posterior probabilities in the output layer.

2.2. Constrained Hidden Markov Model (HMM) Decoder

The second stage consists of a speaker-independent constrained Hidden Markov Model (HMM) Decoder which utilizes the HNN output as its emission probabilities. This decoder builds on the lattice approach in [20] but not only uses generalized SLP knowledge of common PEPs, but integrates the most likely error patterns for each of the given target words. Phoneme dictionaries were created in consultation with expert SLPS for each target word. These dictionaries consist of the target phonemes followed by expected substitutions for each phoneme position. For example, for the target word 'teeth' the phoneme dictionary would be defined as:

```
teeth.dictionary  =
{0:  ['T'],
 1:  ['IY'],
 2:  ['TH', 'F',  'T']}
```

For the final phoneme, there are two expected PEPs; fronting (TH $\rightarrow$ F) and stopping (TH $\rightarrow$ T), both typical substitution type PEPs present in normal populations of children aged 3yrs-3yrs;11mths and 3yrs-3yrs;5mths respectively. This same error pattern in older children may be indicative of language delay [21].

As opposed to training the transition probabilities between HMM states, the phoneme dictionary is utilized to weight the connections. The traditional Viterbi algorithm is then used to infer the most likely sequence of phonemes [22]. Figure 2 demonstrates the three different transition weights used; W_s, transition back to the current phoneme state, W_e, transition to a phoneme state in the phoneme dictionary (an expected phoneme) and W_u, transition to an unexpected phoneme (whilst $W_u \neq 0$, transitions to phonemes other than those in the phoneme dictionary are possible). Through manual definition and tuning of these weights, it is hypothesized that the system can achieve a good level of accuracy despite utilizing small data-sets for which the statistical occurrence of particular errors is not representative of the general population.

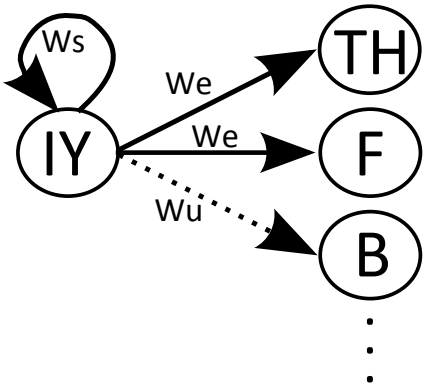

Figure 2: *A state transition in the constrained HMM decoder*

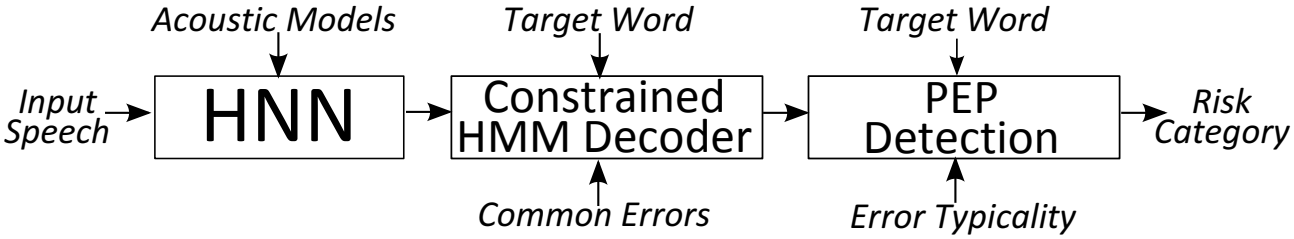

Figure 1: *Three stage architecture of the proposed system*

2.3. Phonological Error Patterns (PEP) Detection

The final stage takes the recognized phoneme string from Stage 2 and performs global alignment with the target word using the Needleman-Wunsch Algorithm [23]. This allows regions of phoneme substitution, insertion and deletion to be detected and passed through a decision tree to determine the PEP. Typical PEPs, like fronting and stopping, are recognized with reference to the phoneme dictionary in the second stage, whilst those not present in the dictionary are deemed to be atypical (e.g. the backing error TH $\rightarrow$ S in the word 'teeth'). This, coupled with information about the child such as age and gender, builds the basis for automating the system to evaluate whether typical PEPs are age appropriate or indicative of delayed speech.

2.4. Speech Corpus

The speech corpus used in this work consisted of 114 unique child speakers, aged between 3 and 14 years, with a range of underlying disorders (57 with Cerebral Palsy, 28 with with idiopathic or development speech disorder, 25 children born preterm $=<$ 30 weeks and 4 normal children). The data was collected by expert SLPs from the Murdoch Children's Research Institute and contains 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 8 words exhibiting typical PEPs from the DEAP Phonology, Inconsistency and Articulation sub-tests were selected to test the proof of concept system. In total 1081 utterances were used, 39.50% which were misarticulated. These utterances represent 21 of the 39 phonemes used in Australian English. There were 92 speakers used for training the HNN models and a hold out set of 12 speakers for validation (i.e. selection of model parameters). There were then 10 individual speakers used for testing the model (unless otherwise indicated).

3. Experiments

In order to demonstrate the feasibility of the proposed system, experiments were performed to test the efficacy of phoneme recognition and PEP detection. We will be comparing our constrained HMM decoder with the results obtained from a speaker-independent HMM decoder (HVite in the HTK toolkit). The HVite HMM decoder was trained using the same phoneme acoustic models as the constrained HMM decoder but utilized a standard lattice, trained using the forward-backwards algorithm. The performance of the two decoders were compared using the speech corpus and testing process outlined in section 2.4. Both decoders utilize the same set of emission outputs (i.e. posterior probabilities) from the split temporal context HNN.

3.1. Stage 2: Constrained HMM Decoder

3.1.1. Optimization

Before evaluating the constrained HMM decoder, the weights in Figure 2 required optimization. Two parameters, magnitude of W_u, and the relative magnitude of W_s:W_e, were optimized individually whilst the other was held constant after, an initial coarse grid search of possible values was performed. W_u was varied from 1×10^4 to 1×10^{-1} and W_s:W_e was varied from 100:1 to 1:100. The phoneme error rate (PER) was measured for these values and weights which balanced the minimum PER for misarticulated and correct phonemes, as well as overall PER, were selected.

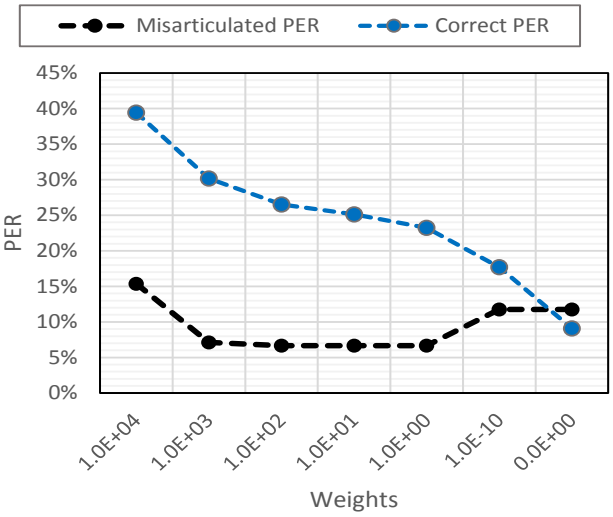

Figure 3: *Phoneme error rate (PER) for different values of* W_u

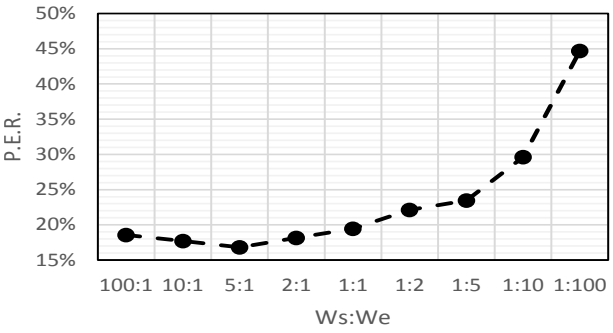

Figure 4: *Phoneme error rate (PER) for different values of the ratio* W_e : W_e

Figure 3 shows the results of varying W_u. As W_u is decreased towards zero, the overall PER also reduces. However the PER for misarticulated phonemes increases slightly when $W_u < 1$, despite the PER for correct phonemes continuing to decrease. To balance these conditions $W_u = 1$ was selected.

Figure 4 indicates that as the ratio is increased in the favor of W_e the PER increases dramatically. It was observed that this was caused by a rapid increase in the number of erroneous insertions (from an average of 0.07 insertions per word at W_s:W_e = 100:1 to 0.27 insertions per word at W_s:W_e = 1:100). The rate of deletions increased slightly at ratios favoring W_s however this effect was minimal. To ensure stable decoding, the ratio W_s:W_e = 5:1 was selected. It was also found that weighting all typical W_e equally increased the rate of false negatives, it also increased the rate of true positives for mispronunciations. Thus equal weighting for all typical W_e was used as in this work prioritized the correct recognition of mispronunciations over false negatives.

The dimensionality of the features vectors used in the HNN acoustic models was investigated for three cases; use of MFCC + Δ + $\Delta\Delta$ coefficients, MFCC + Δ and base MFCC only. Seen in Table 1, the lower dimension vectors improve the PER of the misarticulated and correct phonemes. For the misarticulated phonemes, the PER improvement of the base MFCC features was significant, with a 77.35% and 84.78% improvement over the MFCC + Δ and MFCC + Δ + $\Delta\Delta$. This is an unusual re-

Table 1: *Phoneme error rate (PER) for correct and misarticulated phonemes for decreasing feature vector dimensionality*

	MFCC$\Delta\Delta$	MFCCΔ	MFCC
Correct Phonemes	20.48%	20.1 %	19.90 %
Misarticulated Phonemes	43.75%	29.41 %	6.66%

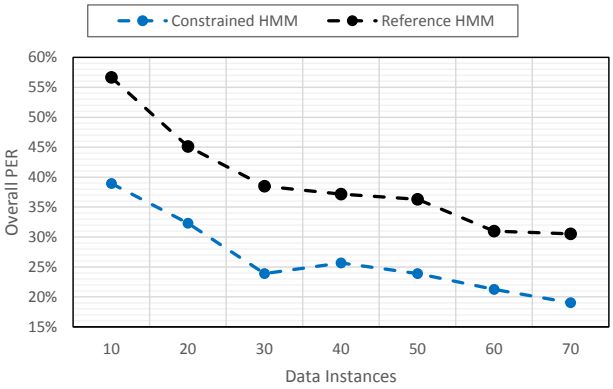

Figure 5: *Phoneme error rate (PER) for different data quantities using the HVite HMM and constrained HMM decoders*

sult for an ASR system and requires further investigation across larger data-sets.

3.1.2. Performance of Constrained HMM Decoder for small data sets

The constrained HMM decoder was tested against the HVite decoder to evaluate whether the hypothesized improvement in PER for small data sets could be achieved. In this experiment the acoustic models were trained on an utterance set, ranging in size from 20 to 70 utterances per word, randomly selected from the 92 training speakers. For each word with a common mispronunciation, half of the training utterances would have this.

From Figure 5 it can be seen that the constrained HMM approach offers consistently better PER than the HVite decoder yielding, on average, a relative PER improvement of 33.12%. Furthermore, the constrained HMM decoder achieves approximately the same overall PER with 20 utterances per word as is possible with the 70 utterances using the HVite decoder. For less than 30 utterances a sharp increase in PER was noted, so utilizing a minimum of 30 utterances is recommended.

3.2. Stage 3: PEP Detection

Two experiments were performed, the first to determine whether the system can demonstrate clinical accuracy by replicating human SLP diagnoses and the second to test whether the accuracy of PEP detection can be maintained for small data sets. The first experiment selects six test speakers, half of whom have have been identified by an expert human SLP as having as having correct, low-risk speech and half of whom have been diagnosed with typical PEPs indicative of moderate-risk speech (focusing on two specific PEPs; fronting and gliding).

For the speakers with low-risk speech, 77.77%, 90.00% and 96.72% of phonemes were recognized correctly (on average 87.26% across speakers) indicating that a threshold-style approach would be effective for identifying low-risk speakers. For the speakers with typical PEPs, 92.86% of correctly spoken phonemes were identified as such with only 12.00% false positives. Table 2 outlines in detail the recognition accuracy for the misarticulated phonemes, grouped by PEP for each speaker. Table 2 shows that the majority of fronting and gliding errors can be accurately detected not only as errors, but as their specific error pattern. This accuracy indicates that, along with utilization of normative data about PEP production [21], the system could be used to evaluate whether these speaker's PEP are delayed or age-appropriate.

Table 2: *PEP recognition accuracy for fronting and gliding for the test speakers*

	Fronting	Gliding
Speaker 4	100%	N/A
Speaker 5	100%	60%
Speaker 6	80%	100%

Table 3: *Recognition accuracy for gliding and fronting when varying the quantity of training utterances*

	30	50	70
True Positive PEP	63.16 %	82.35%	94.12%

The second experiment utilises the same random test set from Section 3.1 and evaluates the accuracy rate (true positives) for fronting and gliding type PEPs for different size utterance data sets. Table 3 shows larger data sets offer greater error pattern recognition accuracy, however, more than half of the tested PEPs could still be correctly identified when smaller sets of training utterances were utilized. This indicates that even in the presence of scarce data, PEP detection could be utilized in the system's evaluation of children's speech.

4. Conclusions

A proof of concept system has been developed to screen for broad speech development issues in children by identifying PEPs. Results indicate that the proposed system can identify phoneme errors with a relatively high level of accuracy despite using only a small data set of disordered Australian children speakers for training. Building upon the data efficiency of the HNN acoustic model, a constrained HMM decoder encoded with SLP knowledge reduced the PER by an average of 33.12% relative to a HVite HMM decoder trained on the same corpus and achieved a minimum PER of 19.03%. The results also demonstrate that specific error patterns, gliding and fronting, can be detected with a recognition accuracy of up to 94.12%.

Future work involves transitioning from a proof of concept system to a functional screening tool. This will included the addition of a larger set of target words, which will enable more atypical PEPs to be identified. To allow the screening tool to be used by non-experts, further integration of SLP knowledge into the system will be pursued. This will include a knowledge base of developmental norms, which along with the detected PEPs, age and gender of the child, will allow for determinations of typicality of error and inform risk assessments. Finally, we will develop a mobile platform to administer the test. For this delivery component of the work, we would also like to work together with children to present the test content in new and engaging ways, possibly in the form of a game.

5. References

[1] Health Workforce Australia. (2014, July) Australia's Health Workforce Series - Speech Pathologists in Focus. Australian Government. [Accessed: 11/03/2016]. [Online]. Available: http://industry.gov.au/Office-of-the-Chief-Economist/SkilledOccupationList/Documents/2015Submissions/Speech-Pathology-Australia.pdf

[2] Australian Bureau of Statistics. (2011) 2011 Census QuickStats. [Accessed: 11/03/2016]. [Online]. Available: http://www.censusdata.abs.gov.au/census_services/getproduct/census/2011/quickstat/0

[3] 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.

[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] Community Affairs References Committee. (2014, Sept.) Prevalence of different types of speech, language and communication disorders and speech pathology services in australia. [Accessed: 18-Feb-2016]. [Online]. Available: http://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Community_Affairs/Speech_Pathology/Report

[6] J. Skeat, L. Gold, M. Wake, O. C. Ukoumunne, and S. Reilly, "The costs of preschool communication problems." *The Medical Journal of Australia*, vol. 195, no. 6, pp. 322–323, 2011.

[7] D. Meyer. (2011, November) Speech-Language Pathology in China: Challenges and Opportunities. The ASHA Leader. [Accessed:11/03/2016]. [Online]. Available: http://leader.pubs.asha.org/article.aspx?articleid=2280098

[8] 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.

[9] T. DiCicco and R. Patel, "Automatic landmark analysis of dysarthric speech." *J. Medical Speech-Language Pathology*, vol. 16, pp. 213–219, 2008.

[10] R. Fraile, J. Godino-Llorente, N. Senz-Lechn, V. Osma-Ruiz, and P. Gmez-Vilda, "Automatic detection of laryngeal pathology on sustained vowels using short-term cepstral parameters: analysis of performance and theoretical justification." *Biomedical Engineering Systems and Technologies*, pp. 228–241, 2009.

[11] J. Arias-Londoo, J. Godino-Llorente, V. Osma-Ruiz, and G. Castellanos-Domnguez, "An improved method for voice pathology detection by means of a HMM-based feature space transformation." *Pattern Recognition*, vol. 43, no. 30, pp. 3100–3112, 2010.

[12] 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.

[13] S.-C. Yin, R. Rose, O. Saz, and E. Lleida, "A study of pronunciation verification in a speech therapy application," in *Proc. 2009 IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Taipei, Taiwan, 2009, pp. 4609–4612.

[14] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2, pp. 95–108, 2000.

[15] 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.

[16] A. Morgan, K. Ttofari-Eecen, P. Eadie, S. Reilly, and B. Dodd, "Speech sound error type at 4 predicts speech outcome at 7: findings from the early language in victoria community study," (in preparation).

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

[18] P. Schwarz, P. Matejka, and J. Cernocky, "Hierachial structures of neural networks for phoneme recognition," in *Proc. IEEE 2006 Int. Conf. on Acoustics, Speech and Signal Processing*, Toulouse, France, 2006, pp. 325–328.

[19] P. Schwarz, "Phoneme recognition based on long temporal context," Ph.D. dissertation, Dept. of Computer Graphics and Multimedia, Brno University of Technology, Brno, 2009.

[20] M. Shahin, B. Ahmed, J. McKechnie, K. Ballard, and R. Gutierrez-Osuna, "A comparison of GMM-HMM and DNN-HMM based pronunciation verification techniques for use in the assessment of childhood apraxia of speech," in *Proc. Interspeech 2014: 15th Annual Conf. of the International Speech Communication Association*, Singapore, 2014, pp. 1583–1587.

[21] B. Dodd, A. Holm, Z. Hua, and S. Crosbie, "Phonological development: a normative study of british english-speaking children," *Clinical Linguistics & Phonetics*, vol. 17, no. 8, pp. 617–643, 2003.

[22] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[23] S. Needleman and C. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins." *Journal of Molecular Biology*, vol. 48, no. 33, pp. 443–453, 1970.