Training forced aligners on (mis)matched data: the effect of dialect and age

Tünde Szalay[1,2], Mostafa Shahin[2], Kirrie Ballard[1], Beena Ahmed[2]

[1]The University of Sydney, Sydney, Australia
[2]The University of New South Wales, Sydney, Australia
tuende.szalay@sydney.edu.au

Abstract

Training forced phonemic aligners for novel language varieties is non-trivial, as it requires aligned corpora. However, aligning novel corpora requires accurate forced aligners. To align AusKidTalk, an audio corpus of Australian English (AusE) speaking children, we trained three custom aligners on different datasets: age-matched American English (AmE) children, dialect-matched AusE-speaking adults, and their combination. Forced aligner performance using the three custom aligners and the Munich Automatic Segmentation System (MAUS) was evaluated against manual segmentation. The dialect-matched and combined custom aligners outperform MAUS, but the age-matched aligner does not. Our aligners' improved forced segmentation will reduce the time-need of manual correction.

Index Terms: forced phonemic alignment, accent differences, developmental differences, custom aligner for AusE-speaking children, audio corpus

1. Introduction

Segmenting acoustic data into phonemes is necessary for phoneme-level acoustic analysis in linguistics [1, 2]. While manual segmentation is considered the "gold standard" in terms of accuracy [2, 3], the use of forced aligners is recommended as manual phonemic alignment may take 800-times more than the length of the audio [4, 5]. Therefore, aligning large datasets is only possible using automatic forced aligners due to the time associated with manual alignment [4, 5].

During forced alignment, the orthographic transcription of the data is converted to phonemic transcription using a grapheme-to-phoneme pronunciation dictionary, and the phonemic transcription is mapped onto the acoustic data using acoustic models [2]. The acoustic model is created by pre-training the aligner using existing time-aligned speech corpora and a pronunciation dictionary, providing grapheme-to-phoneme mappings [2, 6]. Pre-trained models are not affected by speaker-specific idiosyncrasies, as they are trained on a large number of speakers, enabling them to generalise across them by learning speaker-independent characteristics of a language [1, 6, 7].

The performance of forced aligners is negatively impacted by domain- or population level linguistic differences between training- and novel data, such as differences between read and spontaneous speech, or between dialects [6, 7, 8, 9, 10]. Forced aligners trained on American English (AmE), while generally accurate on other dialects of English, produce larger errors in vowel boundaries as differences between training and novel data increase [8, 9]. For instance, an aligner trained on AmE, places 90% of automatic boundaries within 20 ms of manual boundaries for Received Pronunciation (RP), but only 75% for the Westray variety of Scots, a variety that shows larger differences from AmE than RP [8]. In Trinidad English, automatic vowel boundaries are overall accurate with 9–24 ms discrepancies from human alignment; however, Trinidad English-specific vowels show larger discrepancies [9]. Even small differences between automatic and human boundaries can have a roll-on effect, as vowel duration measured using forced and manual alignment may differ by up to 17 to 67 ms in Trinidad English, with Trinidad English-specific vowels showing the largest measurement differences [9]. Due to the adverse effect of accent differences, accent-specific aligners were developed for American, British, Australian, and New Zealand English [11].

Developmental differences between adults' and children's speech also have a negative effect on the accuracy of forced aligners [10]. As most aligners are pre-trained on adult speech, they show low accuracy on children's speech, with 69% to 79% agreement with human annotation [10]. Forced-aligner accuracy improves for older children compared to younger children, as children become more adult-like, and thus more accurately aligned [10]. Despite the adverse impact of age, no age-specific forced aligner has been pre-trained due to the lack of sufficiently sized and segmented children's corpora [12].

AusKidTalk, a large-scale corpus of Australian English (AusE) speaking children, is currently being developed to provide a corpus large enough for developing automated speech analysis tools for the novel variety of AusE-speaking children [12]. Developing a novel forced aligner for AusE-speaking children requires annotated training data, therefore AusKidTalk must, at least partially, be annotated using existing tools. However, AusKidTalk differs from the training data used in most available forced aligners in its accent (AusE vs. AmE) and age (children vs. adults). Therefore, we developed and tested three custom aligners with different pronunciation dictionaries and acoustic models, each trained on partially matching datasets for AusE-speaking children. The first acoustic model was trained on AmE-speaking children, thus training data matched target age, but not dialect. The second acoustic model was trained on AusE-speaking adults, thus training data matched target dialect, but not age. The third acoustic model was trained on both datasets, thus training data partially matched age and dialect. We evaluated the performance of our custom aligners by comparing it to human ground truth annotation as well as to the Munich Automatic Segmentation System (MAUS) [11].

2. Methods

2.1. Custom aligners

We developed three custom aligners, each with different acoustic models and pronunciation dictionaries (Fig. 1). The acoustic models were implemented using a Factored Time-Delay Neural Network in the Kaldi toolkit [13, 14], and trained on three, partially domain-matched datasets: AmE-speaking children (AmE

Child), AusE-speaking adults (AusE Adult), and on the combination of the two sets (Combined) (Fig. 1). The AmE Child model was trained on four children corpora yielding a total of 400 hours including single words and continuous speech – the Oregon Graduate Institute kids' speech corpus [15], the Carnegie Mellon University kids' speech corpus [16], the Colorado University Kids' corpus [17, 18], and the My Science Tutor Children's speech corpus [19]. Grapheme-to-phoneme conversion was provided by The Carnegie Mellon University (CMU) Pronouncing Dictionary [20]. The AusE adult model was trained on 800 hours of speech using the scripted, single word and continuous speech production tasks from the AusTalk corpus [21]. Grapheme-to-phoneme transcription was provided by orthographic- and phonemic transcriptions of the tasks used to elicit speech in AusTalk. The Combined model was trained on 1200 hours of speech using the AmE-speaking children's and the AusE-speaking adults' corpora. Grapheme-to-phoneme transcription was provided by the CMU Pronouncing Dictionary for the AmE-speaking children and by AusTalk transcriptions for the AusE-speaking adults. The Combined model used Multi-Task Learning to share consonant output layer between the dialects and had dialect-specific vowel output layers [22].

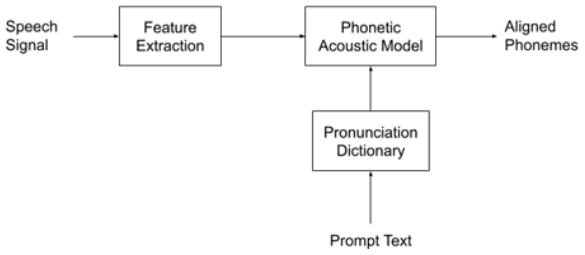

Figure 1: *Schematic diagram of forced aligner architecture.*

During forced alignment, the speech signal is divided into 25 msec frames with 15 msec overlap. Each frame is multiplied by a Hamming windowing function and 40 Mel-Frequency Cepstral Coefficients (MFCC) are extracted from each windowed frame (Fig. 1). The extracted features are fed into the acoustic model along with the expected phoneme sequence created through phonemic transcription of the prompts' orthographic transcription using the appropriate pronunciation dictionary – CMU Pronouncing Dictionary for the AmE Child model and corpus transcriptions for the AusE Adult model. The acoustic model assigns each frame to the most likely phoneme based on the pre-trained phoneme models and is constrained by the given phoneme sequence.

2.2. Test data

Forced-aligner performance was evaluated using data from the AusKidTalk corpus [12]. Speech recordings of eleven (M = 7, F = 4, aged from 4;10 to 11;11, mean = 7;7) children were extracted from the database. Children were native speakers of AusE without any speech disorders. Children produced 18 target words in a picture naming task (range = 11-15 words per child, mean = 13.9 words), giving a total of 153 words. Target words were extracted and saved as single-word wav files.

2.3. Forced and manual aligning

Words were force-aligned with our AmE Child, AusE Adult, and Combined custom aligners, using the sound files and their orthographic transcriptions. Words were force-aligned with the MAUS webtool, using the grapheme-to-phoneme (G2P) → MAUS pipeline without automatic speech recognition, and with the AusE pronunciation dictionary [11, 23, 24, 25]. Expert options were set to default; no custom rules were added. Sound files paired with matching text files containing orthographic transcription of the word with standard English spelling were uploaded to MAUS. MAUS was accessed on 27 August 2021, as well as on 06 June 2022. Results from 2022 are reported.

Manual segmentation was carried out by a trained phonetician in Praat [26], to provide ground truth segmentation prior to observing automatic alignment. Manual segment boundary placement was informed by periodicity, amplitude, and formant structure as presented in the waveform and the spectrogram.

All phoneme-level segmentation was carried out on wav files containing single words to prevent errors caused by confusion between words, such as mistaking the last segment of *snake* with the first segment of *key*, and to minimise the size of alignment errors. All aligners returned the results in Praat textgrids [26]. Boundary locations for all aligners (forced and manual) were extracted from the textgrids using Praat [26].

2.4. Analysis

A total of 816 (phoneme boundaries) × 4 (forced aligners) = 3,264 automatic boundaries were compared to manually placed boundaries. Boundary displacement between automatic and manual boundaries was calculated as the absolute value difference of manual minus automatic boundary [2]. Accuracy of automatically placed boundaries was calculated based on displacement: automatic boundaries were rated as correct when the boundary displacement was 20 ms or below, and as incorrect when displacement exceeded 20 ms [3]. Overlap rate between automatically segmented phonemes was calculated relative to the human annotation, using the time shared between human annotator and forced aligner (*Dur Shared*), the duration of the human aligner (*Dur Hum*), and the duration of the forced aligner (*Dur Forced*) using Equation 1 [2, 27].

$$Overlap = \frac{DurShared}{DurHum + DurForced - DurShared} \quad (1)$$

Equation 1 gives a score from 0 (representing no overlap) to 1 (representing complete overlap) for every phoneme. The distribution of Overlap rate was left-skewed, and bound from 0 to 1, making it conditionally beta-distributed [29]. As 0 and 1, despite being genuine outcomes of Overlap rate, cannot be included in the beta distribution (bound between 0 and 1, non-inclusive), Overlap rate was transformed to beta distribution using the weighted average (N Boundary = 3,264) and a constant 0.5 (Equation 2) [28, 29].

$$OverlapBeta = \frac{Overlap \times (NBoundary - 1) + 0.5}{NBoundary} \quad (2)$$

We constructed a generalised linear mixed effect model (GLM) with the dependent variable Accuracy (binomial family). As the 20ms threshold for accurate boundaries can indicate a quite large discrepancy, especially at fast speech rates, we constructed two more GLMs, one with Displacement (Gaussian family), and one with Overlap (Beta-transformed, Beta family) as dependent variables. The independent variable was Aligner (contrast coded, MAUS as baseline); Speaker was random intercept [30, 31]. p-values were calculated using Satterthwaite's

degrees of freedom method [32]. Planned comparisons with Bonferroni correction were used to compare the AmE Child, AusE Adult, and Combined custom aligners to each other [33]. All data analysis was done in R [34].

3. Results

The Combined custom aligner ($\beta = 0.325, z_{0.103} = 3.165, p = 0.0016$) and the AusE Adult custom aligner ($\beta = 0.254, z_{0.102} = 2.492, p = 0.0127$) produced significantly more accurate boundaries compared to MAUS. Accuracy decreased significantly when using the GenAm Child custom aligner compared to using MAUS ($\beta = -0.489, z_{0.1} = -4.896, p < 0.0001$) (Fig. 2).

Planned comparison confirmed that MAUS is significantly less accurate than the Combined ($p = 0.0093$), and more accurate than the GenAm Child custom aligner ($p < 0.0001$). Contrary to our GLM model, planned comparison did not show a significant difference between MAUS and the AusE Adult custom aligner ($p = 0.0761$). Planned comparison revealed that the GenAm Child custom aligner performs significantly less accurately than the Combined ($p < 0.0001$) and the AusE Adult ($p < 0.0001$) custom aligners. No difference was found between the Combined and the AusE Adult aligners ($p = 1$).

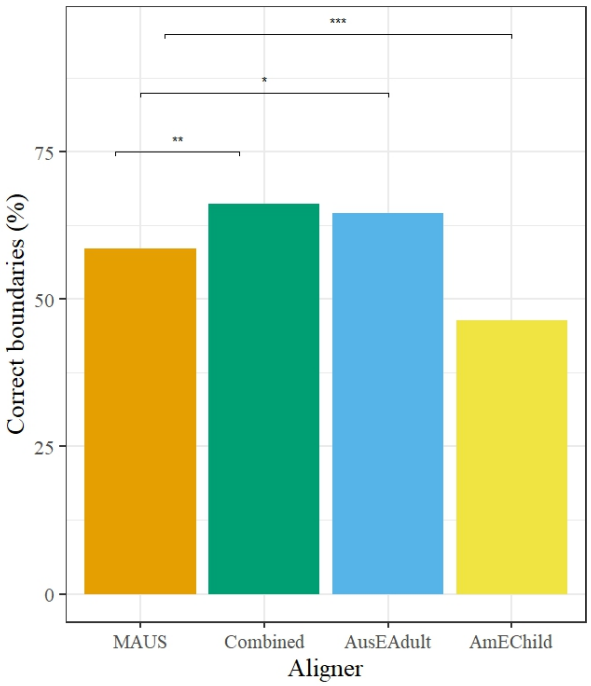

Figure 2: *Boundary accuracy. Significance taken from GLM.*

Choice of forced aligner had no significant effect on boundary displacement in the GLM (Fig. 4) or in the planned comparisons. Boundary displacement was non-significantly smaller (i.e., better) using the AusE Adult ($\beta = -3.452$) and the AmE Child aligners ($\beta = -0.441$) and non-significantly larger (i.e., worse) using the Combined aligner ($\beta = 3.892$) than MAUS.

Overlap rate increased significantly using the custom forced aligners compared to MAUS (Combined: $\beta = 0.196, z_{0.048} = 4.112, p < 0.0001$; AusE Adult: $\beta = 0.268, z_{0.048} = 5.604, p < 0.0001$; AmE Child: $\beta = 0.104, z_{0.048} = 2.155, p < 0.0311$). Planned comparisons confirmed the significantly larger (i.e., better) overlap rate for the Combined ($p = 0.0002$) and the AusE Adult ($p < 0.0001$) aligners compared to MAUS. Contrary to the GLM results, planned comparison showed no difference between MAUS and the AmE Child custom aligner ($p = 0.1873$). Planned comparison revealed that the AmE Child custom aligner shows significantly less overlap with the human annotation compared to the AusE Adult ($p = 0.0043$) aligner but not from the Combined ($p = 0.3396$) aligner. No difference was found between the Combined and the AusE Adult aligners ($p = 0.8045$).

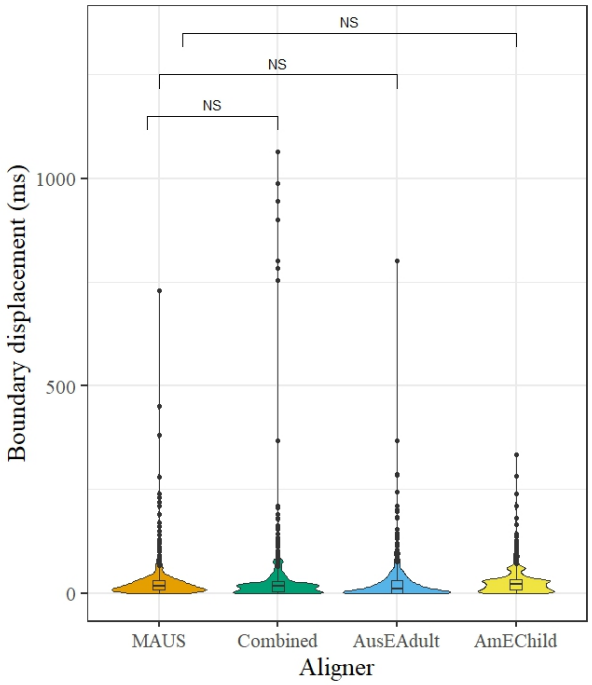

Figure 3: *Boundary displacement. Significance taken from GLM.*

4. Discussion

Our goal was to explore the effect of age- and dialect mismatch on forced-aligning AusE-speaking children's speech. The age-matched, but dialect mismatched custom aligner used an acoustic model trained on AmE-speaking children and a North American pronunciation dictionary; the age-mismatched but dialect-matched aligner used an acoustic model trained on AusE-speaking adults and an AusE pronunciation dictionary; the Combined acoustic model used both datasets with both pronunciation dictionaries. Our AusE Adult and Combined custom aligners outperformed MAUS, both of which used accent-matched training data and pronunciation dictionary. However, performance decreased when our custom aligner used an acoustic model trained on age-matched data and an accent-mismatched pronunciation dictionary. Overall quality of all forced aligners remained low (Table 1), therefore, manual correction of automatically placed boundaries is required for phoneme-level linguistic analysis of AusE-speaking children.

4.1. Custom aligners: the effects of dialect and age

Using the GenAm Child forced-aligner for AusE-speaking children with an age-matched but dialect mismatched acoustic

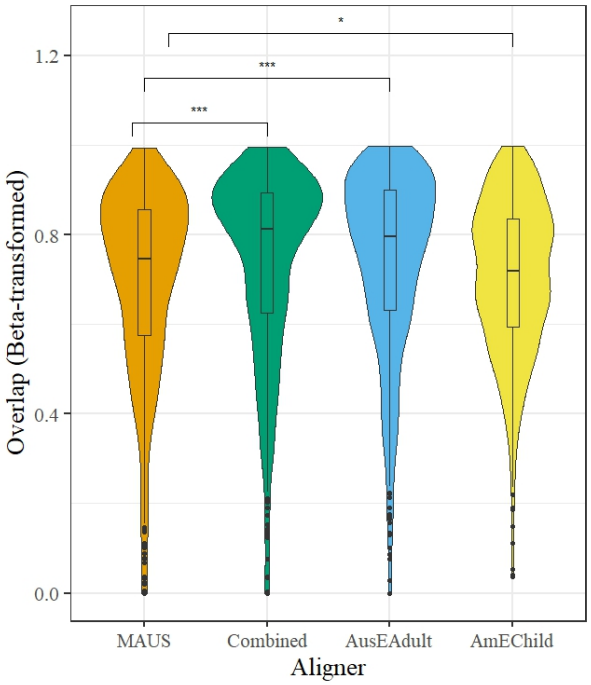

Figure 4: *Overlap rate. Significance taken from GLM.*

Table 1: *Summary of results with accuracy (Acc., %), mean displacement (Disp., ms), and mean overlap rate (Overlap, 0-1, inclusive) for each forced aligner.*

Aligner	Age	Dialect	Acc. (%)	Disp. (ms)	Overlap (0-1)
MAUS	×	✓	59	28	0.69
AusE Adult	×	✓	65	24	0.74
AmE Child	✓	×	46	27	0.71
Combined	✓	✓	66	32	0.73

model yields worse performance than using a similar custom forced aligner with accent-matched acoustic models. The poor performance of the AmE Child aligner indicates that accent differences (AmE vs. AusE) outweigh developmental differences (children vs. adults), as accent differences between AmE and AusE adversely impact the usability of the acoustic model trained on AmE data and of the pronunciation dictionary.

The AmE Child aligner's acoustic model might perform poorly due to dialectal differences. For instance, acoustic vowel differences between AmE training and AusE test data (e.g., *goose* contains back /uː/ in GenAm, but central /ʉː/ in AusE) might lead to incorrect feature mapping between the raw speech and the phonemes. In addition, similarities between the AmE Child training and the AusE child test data might be smaller than expected. For instance, patterns for acquiring /l/, a late acquired sound in both dialects, are similar but not identical [37]. Differences in developmental trajectories between AmE and AusE may cause incorrect feature-to-phoneme mapping and further reduce the suitability of the acoustic model. Thus, using accent-matched AusE training data for AusE-speaking children is required due to the considerable accent differences between GenAm and AusE and the small age-related similarities between children of the two dialects.

The AmE Child aligner's errors in words containing a post-vocalic /ɹ/ in AmE, but not in AusE (e.g., *car, spiderweb*) can be attributed to applying AmE grapheme-to-phoneme mapping onto AusE speech. The pronunciation dictionary in the AmE Child acoustic model maps the letter "r" onto the phoneme /ɹ/ in word-final and pre-consonantal positions, as AmE is a rhotic accent, allowing /ɹ/ in word-final, pre-consonantal, and pre-vocalic positions [36]. In contrast, AusE is a non-rhotic accent, in which /ɹ/ only occurs pre-vocalically (e.g., /ɹ/ occurs in *red*, but not in *car* and *spiderweb*) [35]. As a result, the GenAm Child aligner attempts to map the single AusE vowel into a vowel-/ɹ/sequence, resulting in a vowel offset placed before the acoustic end of the vowel and an unnecessary /ɹ/ interval. Errors caused by /ɹ/-insertion show the detrimental effect of incorrect grapheme-to-phoneme mapping, although previously no such effect of mismatched pronunciation dictionary was found [2]. Therefore, using a dialect-matched, AusE pronunciation dictionary is recommended.

Combining the training data for AusE-speaking adults with GenAm-speaking children increased the data-need of our custom aligner, without leading to a significant improvement in performance. A non-significant reduction in the number of errors, coupled with an increase in the size of errors was observed. As the time needed for manual correction of automatic segmentation depends on the number of errors rather than on the size of errors, even a small difference in accuracy is likely to lead to a considerable reduction in the time and resources required for manual correction. As both accent-matched and age-matched acoustic data are readily available through open-source corpora [15, 16, 17, 18, 19, 21], combining accent-matched and age-matched training data is recommended.

4.2. MAUS and the custom aligners

Our accent-matched aligners outperformed MAUS. The acoustic models of MAUS and our custom aligner's AusE Adult and Combined acoustic models were trained using the same AusTalk dataset [38]. However, MAUS uses the Hidden Markov Toolkit, whereas our custom aligner uses the Factored Time-Delay Neural Network in the Kaldi toolkit [14, 39]. The improved performance of our custom aligner is attributed to the more advanced network used by Kaldi. Similarly, the Kaldi-based Montreal Forced Aligner outperformed other aligners with the Hidden Markov Toolkit [2].

MAUS was accessed on 27 August 2021 and on 06 June 2022. Between 2021 and 2022, the AusE dictionary for MAUS was corrected, and the grapheme-to-phoneme, syllable, and word stress models were re-trained. Using MAUS 2021 versus MAUS 2022 with the same settings did not change the results, despite some improvements: from 2021 to 2022, accuracy of MAUS improved from 58% to 59%, displacement from 28 ms to 27 ms, while overlap rate remained the same (0.69). MAUS allows a high-level of customisation in forced aligning, including the addition of custom rules. These custom features were not used in our current study. It is possible that custom settings would improve forced alignments.

5. Conclusion and future directions

To date, the best-performing forced aligner is our custom built forced aligner using a combined acoustic model of AusE-speaking adults and AmE-speaking children. The main drawback of our custom aligner is its lack of accessibility - while MAUS is easily accessible through its web interface, our aligner is located on a private server. Therefore, future work will include sharing our custom built aligner.

6. Acknowledgements

This project was supported by the Australian Research Council LE190100187 grant. We would like to thank the participants who provided their voice for the AusKidTalk project, and without whom this research would not have been possible.

7. References

[1] Fromont, R. and Watson, K., "Factors influencing automatic segmental alignment of sociophonetic corpora", Corpora, 11(3):401—431, 2016.

[2] González, S., Grama, J., and Travis, C., "Comparing the accuracy of forced-aligners for sociolinguistic research", Linguistics Vanguard, 6(1).

[3] Cosi, P., Falavigna, D., and Omologo. M., "A preliminary statistical evaluation of manual and automatic segmentation discrepancies", Proc. Eurospeech, 693–696, 1991.

[4] Gibbon, D., Moore, R., and Winski, R., Handbook of standards and resources for spoken language systems, 1997.

[5] Schiel, F., Draxler, C., Baumann, A., Ellbogen, T., and Steffen, A. The production of speech corpora, 2012.

[6] Brognaux, S., Roekhaut, S., Drugman, T., and Beaufort, R., "Automatic phone alignment", Proc Int Conf on NLP., 300–311.

[7] Chen, L., Liu, Y., Harper, M. P., Maia, E., and McRoy, S,., "Evaluating Factors Impacting the Accuracy of Forced Alignments in a Multimodal Corpus", Proc LREC, 2004.

[8] MacKenzie, L., and Turton, D., "Assessing the accuracy of existing forced alignment software on varieties of British English", Linguistics Vanguard, 6(s1), 2020.

[9] Meer, P., "Automatic alignment for New Englishes: Applying state-of-the-art aligners to Trinidadian English." JASA 147(4):2283–2294, 2020.

[10] Mahr, T. J., Berisha, V., Kawabata, K., Liss, J., and Hustad, K. C., "Performance of forced-alignment algorithms on children's speech", J. Speech Lang. Hear. Res. 64(6S):2213-2222 (2020).

[11] Kisler, T.,Reichel, U. D., and Florian Schiel "Multilingual processing of speech via web services" Computer Speech & Language, 45:326–347, 2017.

[12] Ahmed, B., Ballard, K., Burnham, D., Tharmakulasingam S., Mehmood, H., Estival, D., Baker, E., Cox, F., Arciuli, J., and Benders, T., "AusKidTalk: An Auditory-Visual Corpus of 3-to 12-year-old Australian Children's Speech", ISCA, 2021.

[13] Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohammadi, M., and Khudanpur, S., "Semi-orthogonal low-rank matrix factorization for deep neural networks", Proc Interspeech, 3743-3747, 2018.

[14] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P. and Silovsky, J., "The Kaldi speech recognition toolkit", Proc IEEE workshop on automatic speech recognition and understanding, 2011

[15] Shobaki, K., Hosom, J.-P., and Cole, R. A., "The OGI kids' speech corpus and recognizers", Proc Sixth Int Conf on Spoken Language Processing, 2000.

[16] Eskenazi, M., Mostow, J., and Graff, D., "The CMU Kids Corpus LDC97S63", LDC database, 1997.

[17] Cole, R., Hosom, P., and Pellom, B., "University of Colorado prompted and read children's speech corpus", Technical Report TR-CSLR-2006-02, 2006.

[18] Cole, R. and Pellom, B., "University of Colorado read and summarized story corpus", Technical Report TR-CSLR-2006-03, 2006.

[19] Ward,W., Cole, R., and Pradhan, S., "My Science Tutor and the MyST Corpus," 2019.

[20] Weide, R., The Carnegie Mellon pronouncing dictionary, 1998.

[21] Burnham, D., Estival, D., Fazio, S., Viethen, J., Cox, F., Dale, R., Cassidy, S., Epps, J., Togneri, R., Wagner, M., Kinoshita, Y., Göcke, R., Arciuli, J., Onslow. M., Lewis, T., Butcher, A., and Hajek, J., "Building an audio-visual corpus of Australian English: large corpus collection with an economical portable and replicable Black Box", Proc Interspeech, 841–844, 2011.

[22] Caruana, R., "Multitask learning", Machine learning 28(1):41-75, 1997.

[23] Reichel, U. D. "PermA and Balloon: Tools for string alignment and text processing", Proc Interspeech, 2012.

[24] Schiel, F. "Automatic Phonetic Transcription of Non-Prompted Speech", Proc ICPhS 607-610, 1999.

[25] Schiel, F. "A Statistical Model for Predicting Pronunciation", Proc ICPhS, 2015.

[26] Boersma, P., and Weenink, D., "Praat: doing phonetics by computer". Version 6.1.41, retrieved 25 March 2021 from www.praat.org

[27] Paulo, S., and Oliveira, L. C. "Automatic phonetic alignment and its confidence measures", Proc. Advances in Natural Language Processing, 36–44, 2004.

[28] Macmillan, N.A., and Creelman, D.C., "Detection theory: a user's guide", Lawrence Erlbaum Associates, 2005.

[29] Smithson, M., and Verkuilen, J., "A better lemon squeezer? Maximum likelihood regression with beta-distributed dependent variables", Psychological methods 11(1), 54–71, 2006.

[30] Bates, D., Mächler, M., Bolker, B., and Walker, S. "Fitting Linear Mixed-Effects Models Using lme4", J Stat Softw 67(1):1–48, 2015.

[31] Brooks, M., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., and Bolker B. M., "glmmTB: Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling", The R Journal 9(2):378–400, 2017.

[32] Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B., "lmerTest Package: Tests in Linear Mixed Effects Models", J Stat Softw 82(13):1–26, 2017.

[33] Lenth, R.V., "emmeans: Estimated Marginal Means, aka Least-Squares Means", 2021.

[34] R Core Team, "R: A Language and Environment for Statistical Computing", R Foundation for Statistical Computing, 2021.

[35] Cox, F., and Fletcher, J., Australian English pronunciation and transcription, Cambridge University Press, 2017.

[36] Labov, W., Ash, S., and Boberg, C., The Atlas of North American English, Phonetics, Phonology and Sound Change, Gruyter Mouton, 2008.

[37] Lin, S., and Demuth, C., "Children's acquisition of English onset and coda /l/: Articulatory Evidence", J. Speech Lang. Hear. Res. 58:13–27, 2015.

[38] Cassidy, S., Estival, D., and Cox, F., "Case study: the AusTalk corpus", in N. Ide, and J. Pustejovsky, J. [Ed] Handbook of linguistic annotation, 1287-1301, Springer 2017.

[39] Young, S., Evermann, G., Hain, T., Kershaw, D., Xunying A. L., Odell, J., Ollason, D., Povey, D., Valtchev V., and Woodland, P., "The HTK Book (For Version 3.4)" Cambridge University Engineering Department, 2006.