Proceedings of Meetings on Acoustics

Volume 24 http://acousticalsociety.org/

170th Meeting of the Acoustical Society of America

Jacksonville, Florida
1 - 6 Nov 2015

Speech Communication: Paper 4pSCb2

Automatic forced alignment on child speech: Directions for improvement

Knowles, Thea
Health and Rehabilitation Sciences, Western University, London, Ontario, Canada
School of Communication Sciences and Disorders, Western University, London, Ontario, Canada; tknowle3@uwo.ca

Clayards, Meghan
School of Communication Sciences and Disorders, McGill University, Montreal, Quebec, Canada;
Department of Linguistics, McGill University, Montreal, Quebec, Canada; meghan.clayards@mcgill.ca

Sonderegger, Morgan & Wagner, Michael
Department of Linguistics, McGill University, Montreal, Quebec, Canada; morgan.sonderegger@mcgill.ca, chael@mcgill.ca

Nadig, Aparna
School of Communication Sciences and Disorders, McGill University, Montreal, Quebec, Canada; aparna.nadig@mcgill.ca

Onishi, Kristine H.
Department of Psychology, McGill University, Montreal, Quebec, Canada; kris.onishi@mcgill.ca

Phonetic analysis is labor intensive, limiting the amount of data that can be considered. Recently, automated techniques (e.g., forced alignment based on Automatic Speech Recognition - ASR) have emerged allowing for much larger-scale analyses. For adult speech, forced alignment can be accurate even when the phonetic transcription is automatically generated, allowing for large-scale phonetic studies. However, such analyses remain difficult for children's speech, where ASR methods perform more poorly. The present study used a trainable forced aligner that performs well on adult speech to examine the effect of four factors on alignment accuracy of child speech: (1) Corpus - elicited speech (multiple children) versus spontaneous speech (single child); (2) Pronunciation dictionary – standard adult versus customized; (3) Training data – adult lab speech, corpus-specific child speech, all child speech, or a combination of child and adult speech; (4) Segment type – voiceless stops, voiceless sibilants, and vowels. Automatic and manual segmentations were compared. Greater accuracy was observed with (1) elicited speech, (2) customized pronunciations, (3) training on child speech, and (4) stops. These factors increase the utility of analyzing children's speech production using forced alignment, potentially allowing researchers to ask questions that otherwise would require weeks or months of manual-segmentation.

1. INTRODUCTION

Acoustic analysis of speech data has traditionally required labour intensive hand annotation. Recently, automatic speech recognition (ASR) techniques and in particular forced alignment have been used to automate some of this process (Gorman et al., 2011; Milne, 2014; Renwick et al., 2013; Schiel, 2004; Yuan & Liberman, 2008; 2011). Forced alignment takes as input an orthographic transcription of the speech signal, the speech signal itself, a pronunciation dictionary, and acoustic models trained to recognize the phones of the pronunciation dictionary and, as output, maps the acoustic signal onto the phones, producing an automatic segmentation. It has been successful for automating acoustic analysis of adult productions (e.g., sibilant spectral centre of gravity, Clayards & Doty, 2011; word-final consonant-cluster variation, Milne, 2014). This technique therefore allows for much larger scale and faster analysis of phonetic data than has been traditionally possible, accelerating the progress of scientific research.

Extending these techniques to other populations such as children would also be of benefit. However, ASR technology is known to perform more poorly with child than adult speech (see Benzeghiba et al., 2007 for review) with error rate generally inversely correlated with age. Some of the problems that ASR systems face is that children's speech is more variable, slower and systematically different in spectral dimensions than adult speech (Lee et al., 1999). In fact, human listeners also have more difficulty in recognizing children's speech, especially young children (D'Arcy & Russell, 2005). The differences between adult and child speech makes recognition of children's speech using acoustic models trained on adult speech problematic (Wilpon & Jacobsen, 1996). Training with child speech improves performance (Wilpon & Jacobsen, 1996), as does warping the speech signal using vocal tract normalization to more closely match adult acoustics (Gerosa et al., 2007; Potamianos et al., 1997) but this later technique may not be as successful as training with child data (Elenius & Blomberg, 2005). Another source of difficulty for automatic systems is that children do not always pronounce words with the same phones as would be found in an adult pronunciation dictionary (Benzeghiba et al., 2007).

In full automatic speech recognition, the system must determine what the words were as well as where the segments are. In forced alignment, however, the transcription is already available making the task more constrained. Thus forced alignment is a potentially viable tool for analyzing children's speech. For example, Lee et al. (1999) used it to examine acoustic properties of speech of 5- to 11-year-olds. However relatively little work has examined factors affecting the accuracy of forced alignment for children's speech. Here we examine the viability of using forced alignment for child data by examining the effect of four factors on alignment performance: corpus; standard vs. phonetic-transcription-based pronunciation dictionary; training set; and type of segment to be aligned.

2. METHOD

2.1 Data

We analyzed two speech corpora with audio files from the Child Language Data Exchange System (CHILDES) (MacWhinney, 2000). The *Julia* corpus included approximately two hours of speech from one female Canadian-English speaking child. Data were collected longitudinally from ages 1;5 to 3;6 in a naturalistic setting (Goad, 2010). The *Paidologos* corpus included approximately five hours of speech from 81 children (40 females) from Columbus, OH, ages 2;0 - 5;11 (Edwards & Beckman, 2008). Speech consisted of single word productions elicited during a picture-prompted word repetition task. Both corpora included orthographic and partial or full phonetic transcriptions. The speech audio files were segmented at the utterance level to prepare for alignment.

2.2 Manual Segmentation

Manual segmentations were collected for both corpora for comparison to the automatic segmentations. For *Julia*, manual segmentation of voiceless stops, voiceless sibilants, and vowels was completed by research assistants using Praat (Boersma & Weenink, 2011). Phoneme boundaries that were too difficult to determine due to background noise or ambiguity in the signal (for example two stops with no release between them, making segmentation of individual stops impossible) were discarded. For *Paidologos*, manual segmentations of word-initial consonants and the following vowels were provided with the corpus.

2.3 Forced alignment

Automatic segmentation was then performed for all data using the Prosodylab-Aligner (Gorman et al., 2011), which uses the Hidden Markov Model Toolkit (HTK) to align text to audio. Forced-alignment was manipulated in eight conditions by transcription type and training data for *Julia*, and in four conditions by training data for *Paidologos*.

2.4 Pronunciation dictionaries

Forced alignment requires a phonetic transcription of the audio speech data to-be-aligned. We included two transcription conditions. A *standard* North American English transcription, the CMU Pronunciation Dictionary, was used for both corpora. The CMU Pronunciation Dictionary is a machine-readable pronunciation dictionary for North American English that provides over 134,000 words and their phonetic transcription in ARPAbet. *Julia* was also aligned using a *customized speaker-specific* pronunciation dictionary developed from the phonetic transcription of her utterances. Each utterance was given a unique entry in the pronunciation dictionary (i.e., no homonyms). The supplied narrow phonetic transcription was collapsed into broader ARPAbet characters to provide more exemplars for each ARPAbet category.

2.5 Training acoustic models

The default acoustic models for the Prosodylab-Aligner are monophone Gaussian mixtures consisting of 39 Mel frequency cepstral coefficients that were pre-trained on approximately ten hours of North American English adult laboratory speech (Gorman et al., 2011). One advantage of this aligner is that it also supports training of new acoustic models based on arbitrary datasets. Training of new acoustic models involves three rounds of model estimation consisting of four iterations each. In the first round, the models are initialized with flat-start monophones. Next, a tied-state “small pause” model is inserted before the second round of estimation. The data are aligned using the most likely pronunciation of all homonyms in the provided pronunciation dictionary, followed by a final round of estimation. The final alignment represents the optimal model estimation given the training input.

Four training conditions were included in the present study. Training set (a) included the default acoustic models described above. Training set (b) included acoustic models trained only on the specific corpus to be aligned (*Julia* or *Paidologos*). Training set (c) included all child data (approximately seven hours of audio) and no adult data. Training set (d) included a combination of both adult laboratory data (default training input) as well as the child data (approximately 6 hours of audio).

2.6 Comparisons

Manual and automatic segmentations were compared for voiceless stops, voiceless sibilants, and vowels by computing the percentage of auto-aligned phones that overlapped with the midpoint of the corresponding manually-aligned phone (%-Match).

3. RESULTS

3.1 Alignment parameters

Acoustic models that produced more accurate alignments were identified by a greater %-Match between the automatically and manually aligned audio. Table 1,reports the %-Match for each corpus for each alignment parameter, which ranged from 23% to 89%. Alignment on *Paidologos* resulted in higher accuracy. For both datasets, more accurate alignments were generally obtained when training included the specific child to be aligned. Additionally, for *Julia*, alignment was more accurate with the customized than with the standard dictionary. Specific phone alignment accuracy varied across the two datasets. For *Paidologos*, vowels accounted for the highest percentage of matched segments, whereas voiceless sibilants were aligned more accurately for *Julia*.

Table 1: % Match

Dataset	Transcription	Training	Stop	Sibilant	Vowel	All segments
Paidologos	Standard	A-adult	78	23	81	**61**
		B-specific	89	86	86	**87**
		C-all child	85	64	83	**77**
		D- child + adult	76	53	87	**72**
		All Training	82	57	84	**74**
Julia	Standard	A-adult	27	28	20	**25**
		B-specific	45	52	27	**41**
		C-all child	49	61	27	**46**
		D- child + adult	32	32	23	**29**
		All Training	38	43	24	**35**
	Customized	A-adult	33	36	40	**36**
		B-specific	65	76	63	**68**
		C-all child	53	80	63	**65**
		D- child + adult	39	41	46	**42**
		All Training	48	58	53	**53**
All Datasets	All transcriptions	All Training	54	52	52	**53**

Accuracy of the automatic left and right boundary placements for all matched segments was also examined. For each boundary (left and right) the absolute difference between the manual and hand aligned placement was calculated. Table 2 presents the cumulative distribution of all absolute boundary differences for matched segments. For example, 49.69% of matched automatic and manual boundary placements differed by less than 25 ms for *Paidologos*, compared to 45.79% for *Julia*.

Table 2: Cumulative % of absolute boundary differences between automatic and manual alignments (matched segments)

Absolute boundary difference (ms)	Cumulative %	
	Paidologos	*Julia*
5	2.25	4.03
10	11.05	15.02
25	49.69	45.79
50	87.80	77.28
100	96.61	93.63

3.2 Age of speaker

Alignment accuracy improved with the speech of older children for both datasets (Figure 1). For *Paidologos*, the speech of 4- and 5-year-olds was aligned with greater accuracy than that of 2- and 3-year-olds. In the *Julia* dataset, which was collected longitudinally, speech collected later in time was aligned with greater accuracy.

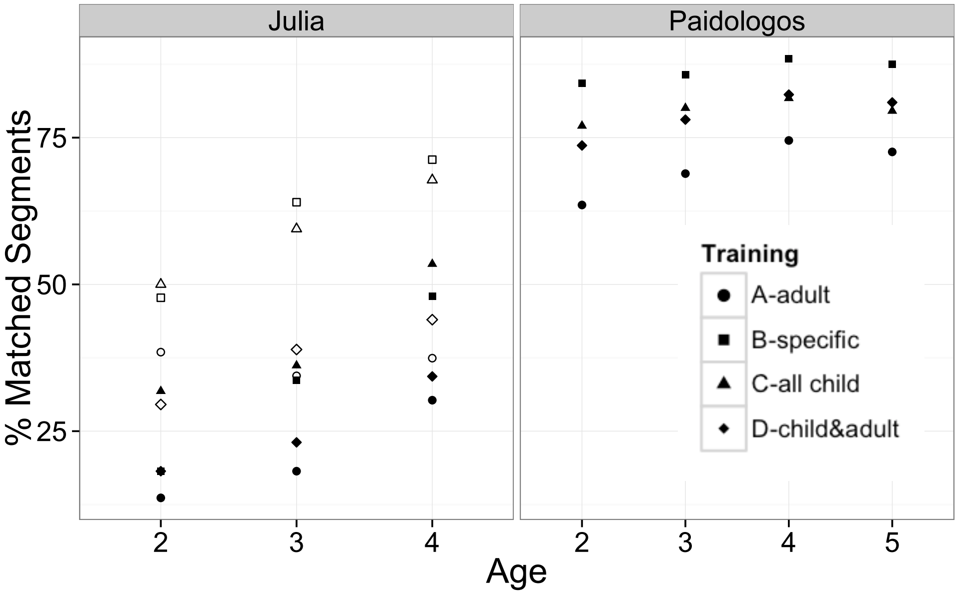

Figure 1: %-Match by speaker age, alignment parameters. Open shapes represent the customized pronunciation dictionary.

4. DISCUSSION

We found differences in the % Match across corpora. This may have been due to the type of speech - the elicited single-word productions (*Paidologos*) were aligned with greater accuracy than the spontaneous speech in a conversational setting (*Julia*) - however it could also have been due to other differences between the corpora such as recording quality. We found that training acoustic models on child speech yielded better alignment than using models that were trained on only adult data – consistent with the literature on factors that improve ASR performance. In particular, the most accurate segmentations for each corpus used acoustic models trained exclusively on the data of the specific child (or for *Paidologos*, children) to be aligned. This is likely due to differences between the two corpora (e.g., style of speech and recording quality) leading to somewhat different acoustic models. In the case of Julia's data, providing a customized pronunciation dictionary based on a phonetic transcription of the speech also improved performance. This may be especially useful for spontaneous speech where deviations from canonical pronunciation are more likely. We also found differences across segment types; the relative success of each segment type depended on the corpus (for Julia, sibilants were the most accurate, but for Paidologos sibilants were the least well aligned, even if we consider only alignments with the standard dictionary). This suggests that categories of phones may be treated differently by automatic methods, depending on the number of examples in the training data, the amount of segmental variability in the data, and/or the nature of the speech task. Finally, alignment performance increased with speaker age, presumably due to the decreased variability in older children's speech. These results suggest that, despite limitations, the parameters identified here may improve the semi-automatic analysis of speech from children, contributing to our ability to conduct larger-scale analyses of child speech.

REFERENCES

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., & Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication*, *49*(10), 763-786.

Boersma, P., & Weenink, D. (2011). Praat: Doing phonetics by computer [Computer program]. Version 5.3, retrieved from http://www.praat.org/.

Clayards, M., & Doty, E. (2011). Automatic analysis of sibilant assimilation in English. *Canadian Acoustics*, *39*(3), 194-195.

D'Arcy, S., & Russell, M. J. (2005). A comparison of human and computer recognition accuracy for children's speech. In *Proceedings of Interspeech*, 2197-2200.

Edwards, J., & Beckman, M. E. (2008). Methodology questions in studying consonant acquisition. *Clinical Linguistics & Phonetics, 22*, 939-958.

Elenius, D., & Blomberg, M. (2005). Adaptation and normalization experiments in speech recognition for 4- to 8-year-old children. In *Proceedings of Interspeech*, 2749-2752z

Gerosa, M., Giuliani, D., & Brugnara, F. (2007). Acoustic variability and automatic recognition of children's speech. *Speech Communication*, *49*(10), 847-860.

Goad, H. (2010). English-Goad: Online corpus of phonological development. McGill University. ISBN 1-59642-438-9. Web access: http://childes.talkbank.org/data/

Gorman, K., Howell, J., & Wagner, M. (2011). Prosodylab-Aligner: A tool for forced alignment of laboratory speech. In *Proceedings of Acoustics Week in Canada*, *39*(3), 192-193.

Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *Journal of the Acoustical Society of America*, *105*(3), 1455-1468.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Third Edition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Milne, P. (2014). The variable pronunciations of word-final consonant clusters in a force aligned corpus of spoken French. Doctoral dissertation, University of Ottawa.

Potamianos, A., Narayanan, S., & Lee, S. (1997). Automatic speech recognition for children. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech '97)*, 2371-2374.

Renwick, M. E., Baghai-Ravary, L., Temple, R., & Coleman, J. S. (2013). Assimilation of word-final nasals to following word-initial place of articulation in UK English. In *Proc. Interspeech,* 3047-3051.

Schiel, F. (2004). MAUS goes iterative. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, 1015-1018.

Wilpon, J. G., & Jacobsen, C. N. (1996). A study of speech recognition for children and the elderly. In *Proceedings of Acoustics, Speech, and Signal Processing,* 349-352.

Yuan, J., & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. In *Proceedings of the Acoustical Society of America, 123*(5), 5687-5690.

Yuan, J., & Liberman, M. (2011). /l/ variation in American English: A corpus approach. *Journal of Speech Science*, *1*, 35-46.