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

An Automatically Aligned Corpus of Child-directed Speech

Micha Elsner, Kiwako Ito

Department of Linguistics, The Ohio State University

melsner@ling.osu.edu, ito.19@edu

Abstract

Forced alignment would enable phonetic analyses of child directed speech (CDS) corpora which have existing transcriptions. But existing alignment systems are inaccurate due to the atypical phonetics of CDS. We adapt a Kaldi forced alignment system to CDS by extending the dictionary and providing it with heuristically-derived hints for vowel locations. Using this system, we present a new time-aligned CDS corpus with a million aligned segments. We manually correct a subset of the corpus and demonstrate that our system is 70% accurate. Both our automatic and manually corrected alignments are publically available at osf.io/ke44q.

Index Terms: 1.18 Special session: Data collection, transcription and annotation issues in child language acquisition settings; 1.11 L1 acquisition and bilingual acquisition; 8.8 Acoustic model adaptation

1. Introduction

Studies of the phonetics of child-directed speech (CDS) can now take advantage of a variety of large transcribed corpora from different speakers, languages and situations [1]. In some cases, data may be collected on an extremely large scale with daily recordings (LENA) [2]. But while many researchers want to focus on specific segments (such as vowels or sibilants), the lack of time alignments between the transcript and the audio recording makes it impossible to tell where the segment of interest occurs. Traditionally, these alignments are supplied by the phonetician, a task which is expensive and time-consuming. For instance, the Buckeye corpus of adult-directed speech [3, 4] was aligned at the segment level by annotators with the aid of a first-pass computer segmentation system; the overall transcription effort took nearly five years. Comparable general-purpose resources for CDS are rare and hard to access; we know of only one, the RIKEN corpus of spoken Japanese [5], and even this is not publically distributed. Thus, many researchers are forced to create time alignments by hand as part of their data analysis. For instance, in her study of sibilant acquisition, Cristiá [6] annotates sibilants in 8167 sentences of CDS.

In adult phonetics, time alignment with a transcript can often be automated using speech recognition technology (“forced alignment”). Any speech recognizer can be used to perform the alignment, although the Penn Forced Aligner [7] has been particularly popular due to its ease of use. In any case, the forced aligner employs the same trained acoustic model as a speech recognizer, and has similar demands for training data. This creates problems for alignment in novel domains such as CDS; training a top-quality recognizer may require on the order of thousands of hours of transcribed audio [8].

We present a new dataset containing automatic alignments for a popular corpus of CDS [10], as well as a small set of manual alignments used for evaluation; size statistics are shown in table 1. To perform the alignments, we use some heuristic methods to improve a Kaldi [13] aligner. Our alignments are about 70% accurate, substantially more so than a previous attempt to align this corpus (Pate 2014, personal communication), using the HTK recognizer [9]. We use our alignments to replicate a previous result on the phonetics of CDS, demonstrating their potential usefulness for research purposes.

Table 1: *Statistics of the alignment dataset.*

	Automatically aligned
Sessions	167
Total duration	152;51
Total words	386307
Total phones	1190420
	Manually corrected
Sessions	4
Total duration	3;41
Total words	16770
Total phones	44290

1.1. Dataset

Our data is drawn from the Providence corpus [10]. This corpus was collected during a two-year longitudinal study of six children from Providence, Rhode Island (0;11–2;1; 3 male, 3 female). Parents (usually mothers) interacted with the child for about one hour every two weeks. The interactions are naturalistic; the researcher was not present and did not prompt any particular kind of interaction. The parent’s utterances were recorded with a wireless lavalier microphone pinned to their collar. However, audio quality is not universally high. Some recordings have long sections of feedback and include noises caused by objects tapping the microphone.

Providence is orthographically transcribed (using CHAT conventions [1]) and time-aligned at the utterance level, but not at the word or phone level. Our corpus covers the data from parents of four of the six children (Alex, Lily, Violet and William; 2 male, 2 female). Two more (Ethan and Naima) were skipped because of difficulty extracting the audio. We filter the Providence transcripts to select only adult utterances (not utterances by the child). Utterances where the transcriber indicated a phonetic pronunciation for a partial or unintelligible word are discarded.

Providence contains a lot of data from a small set of speakers, making it a valuable resource for child-directed phonetics. While the original study [10] focuses on child productions, subsequent researchers have used Providence as a rich source of CDS for modeling studies [11, 12]. But these studies relied on automatic conversion of the orthography into a phonemic transcript. While phonetic detail could have been useful, the lack of phone-level alignments prevented this kind of analysis from being done.

http://dx.doi.org/10.21437/Interspeech.2017-379

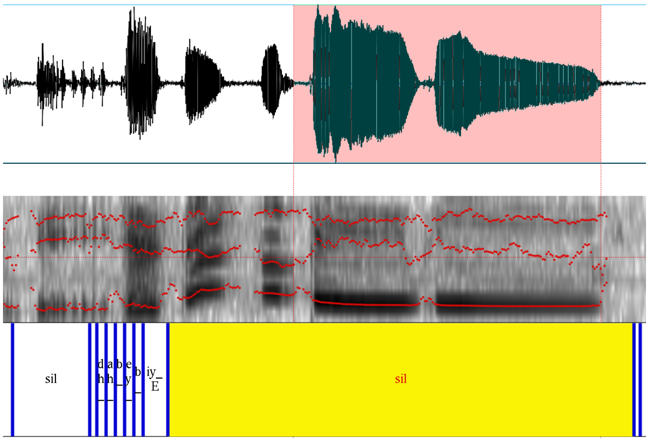

Figure 1: *"Where's the baby?": initial Kaldi alignment. The automatic alignment incorrectly detects all the speech sounds in a small region on the left, with a long following silence. The pink region represents the actual word "baby".*

1.2. Baseline Recognizer

The baseline alignment system is implemented using Kaldi [13], a general-purpose speech recognition toolkit. It is a Gaussian triphone recognizer trained on the Switchboard conversational speech corpus [14]. Switchboard contains conversational speech by American adults, and is therefore a reasonable starting point for modeling the adult speakers in Providence. The baseline system is trained using the Kaldi distribution's premade example script for Switchboard. The alignments produced by Pate (p.c. 2014) use a similar setup, but with HTK [9] as a base recognizer and the Wall Street Journal as a training corpus [15].

The baseline system has relatively low accuracy on the child-directed corpus, due to a variety of phonetic and lexical differences from Switchboard. For instance, Figure 1 shows a severe alignment error in which some long, falsetto segments are misrecognized. Extreme vowel elongation in content words [16, 17] and pitch excursions [18] are typical features of CDS. In other cases, words (such as "elmo", "pumbaa" and "ankylosaurus") are missing from the dictionary; Kaldi simply triggers warnings and then treats these as silence.

1.3. Filtering Low-quality Audio

We detect and exclude segments containing too much microphone feedback. Feedback amplifies a narrow band of frequencies (determined by the configuration of the microphone and the resonant frequencies of the room) [19]. Thus, intervals of feedback can be detected with a formant tracker (although this is an unconventional use of the tracker, which is designed to detect resonances of the vocal tract). We apply the formant detector in Praat [20] to detect a single formant centered at 1300 Hz. (This frequency was selected observationally by looking at the spectrogram for a minute or two of feedback in the corpus.) If the detected power of this formant exceeds .03 of the total power in the signal over a 100ms window, the acoustic frame is marked as feedback. Sequences of feedback frames over 3 seconds long are marked for exclusion, and utterances which overlap them are discarded. This procedure finds 10000 seconds of feedback, about 1.7% of the total recording time in the corpus.

2. Forced Alignment Pipeline

We treat the problem of aligning CDS as a domain adaptation problem [21, 22]. Beginning with the Kaldi baseline system, we modify the dictionary to include some CDS-only words. Next, we use a heuristic method to find likely vowel segments (hopefully marking many of the long, falsetto segments that the baseline misses). We run the system using the heuristic detections to reweight the acoustic model, creating a new set of improved alignments, then run a single iteration of self-training on those alignments.

2.1. Extending the Dictionary

We supply new pronunciation entries for all words in Providence which are not already in the dictionary. These entries are automatically predicted from their spelling using Phonetisaurus, a grapheme-to-phoneme prediction system [23]. For instance, "elmo" is assigned the pronunciation *eh l m ow*. While Phonetisaurus is reasonably accurate, it does make some errors, especially on orthographically atypical words: "pumbaa" (*p uw m b ah*) is assigned the pronunciation *p ah m b ey ey*.

We also add a few dictionary entries by hand: single-sibilant pronunciations of "is" and "does" (*s, z*) and a monosyllabic "little" (*l ih l*). The system continues to make errors by positing phantom vowels in longer multiword phrases with extensive reduction ("whaddya", "dyawanna", etc.). Such phrases could be added to the lexicon as alternate paths through a finite-state transducer [24], but it is not clear how many there are or whether they can be reliably learned from this amount of data. We leave this possibility for future work.

3. Heuristic vowel extraction

Errors involving atypical vowel segments (as in Figure 1) can be partly countered by giving the system "hints" about the vowel position. We obtain a rough estimate of the vowel positions by looking for loud intervals in the 500-1000Hz range of the spectrogram. In particular, we use a Hann band (rectangular) filter to select this spectral region, then use the built-in "silences" detector in Praat [20] to smooth the detections. The smoothing function divides the audio into loud (vowel) and quiet regions, but ignores brief events in which the amplitude varies. We set the silence threshold to -25 dB below the maximum intensity of the audiofile and the minimum length of a region to 0.02 seconds. (These parameters were tuned manually on a small subset of the data.)

The vowel position hints are provided as an extra parameter to a modified version of Kaldi's decoder. The modification acts as a wrapper around an arbitrary acoustic model (in this case, a Gaussian mixture). When within a detected vowel region, the wrapper increases the posterior for vowels by a log-odds factor of 3 and decreases the posterior for everything else by 3. Outside such a region, vowels are downweighted by a factor of 3.

Aligning the corpus using the vowel hints improves the results, but can cause problems, for instance with segments such as nasals, intervocalic fricatives, and liquids (which are often loud enough in the low frequencies to be labeled as vowels). Thus, we perform a step of self-training by retraining the recognizer on the output of the vowel hints system. This adapts the acoustic model towards picking up CDS-specific phenomena, but without forcing it to obey the hints in cases where they are too acoustically implausible.

Our final aligner is this retrained system (which does not use the hints directly). While it is not a strong enough system for general-purpose CDS speech recognition, it is capable of reasonable performance as an aligner.

4. Evaluation

We create a small gold standard alignment for evaluation. Rather than start from scratch, we provide our human annotators with the system output and ask them to correct it. Although this might lead to some biases in favor of our system, it greatly reduces annotation time, and is a standard procedure in projects like Buckeye [4].

Hand alignment was carried out as a class project by undergraduate students taking an advanced phonetics class. Four recordings (one per child) were selected for correction. A total of 32 students were divided into 8 teams to hand-correct the system alignments. Each team was assigned about a half of a recording of a particular parent's speech (i.e., approx. 30 min). Although each student was in charge of a different part of the recording (approx. 7.5 min each), they worked as a team to learn the Arpabet transcription system and to discuss uncertain transcriptions. They were provided with the original recording wav file, an associated Praat TextGrid file with Kaldi-based phonetic and word transcriptions, and the CHA discourse transcription file. They were instructed to insert two extra tiers for copying and editing the existing annotations for the assigned part. They were instructed to listen to the wav file while viewing the CHA transcription to check its accuracy, and adjust/insert boundaries or transcriptions whenever a discrepancy between the sound and its existing transcription was detected. The data presented here are from 27 students who completed the task within the course assignment period (3-4 weeks).

The hand-corrected data contains 44290 segments of which 16820 are vowels. An additional tuning set was hand-corrected by the first author; this set contains 2309 speech segments of which 913 are vowels. The annotators were not required to supply alignments from scratch in cases where the system completely discarded an utterance due to missing words in the original transcript or to feedback. In some cases, they added alignments for these utterances anyway; in others, they did not. Thus, recall scores reflect some losses of phones due to filtering of the utterance set, but mainly losses due to misalignments.

We compute recall measurements by checking each gold phone in the hand-corrected data to see if it was acceptably detected by the system. Non-phone segments for laughter, noise, etc. are not considered when scoring. We consider a segment "acceptably" aligned if it overlaps in 0.75 of its duration with a segment with the same label in the automatic alignment. For vowel segments, the hand-correction often indicates an alternate vowel quality (e.g. *k ih n* rather than *k aa n* for "can"). We allow vowels to overlap with all other vowels, regardless of label. A segment is "catastrophically" aligned if it overlaps in less than 0.05 of its duration. Precision measurements are computed in the same way, except that we check the alignment for each proposed segment in the system output against the gold alignment.

Results for the full pipeline, the initial Kaldi baseline, and Pate's alignment are shown in Table 2. It is clear that the full system is substantially more accurate than either competitor. It more than doubles the number of acceptable segments (70% vs 30%) and halves the number of catastrophic ones (24% vs 60%) versus the baseline. Bearing in mind that the evaluation is potentially biased against Pate due to the influence from the initial machine alignment, it is nonetheless clear that an off-the-shelf HTK aligner is no better (and in fact is probably worse) than a Kaldi one. Results for vowels are shown in Table 3. Most systems perform slightly better for vowels than for other segments.

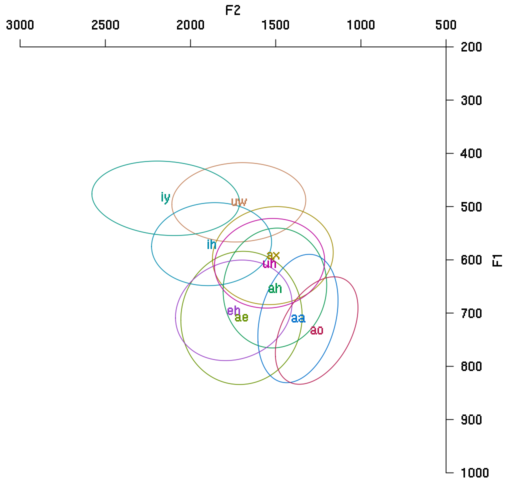

Figure 2: *Vowel midpoint formants (25% confidence intervals) for automatically aligned vowels.*

5. Applications

While 70% precision may not be acceptable for applications which demand that every token be accurately measured, it does enable the automatic extraction of useful information in the aggregate. As an illustration of what is possible with this level of accuracy, we demonstrate two simple phonetic analyses of our automatically aligned dataset.

First, we extract a vowel formant plot (measured at midpoint) from our dataset and graph the 25% confidence intervals (Fig 2). The plot represents all (151852) non-diphthongal vowel tokens with durations between .08 and .3 seconds, excluding 294 tokens with $F_1 + F_2 > 4000$, which are excluded as obvious formant tracking errors. The plot clearly shows the canonical locations of the English vowels (cf. [25]). Vowel formant measurements are widely used as input for computational models of distributional learning [26, 27] as well as studies of dialect perception and processing [28]. The measurements from our corpus appear suitable for these purposes as well.

Second, we analyse the duration of vowels in selected function and content words, replicating the analysis of Swanson et. al. [16]. This study compared the duration of lax vowels in monosyllabic function words versus content words. They determine that content words are lengthened in CDS compared to adult-directed speech (ADS), while function words are not; they also show well-known effects of phrase-final lengthening [29] for content words. The data for the study was experimentally collected, with 15 mothers reading a story to their daughters and to an adult listener. There were 90 critical sentences containing a key word, and the critical segments were manually aligned.

We run the same analysis on our corpus, providing Swanson's measurements as a comparison. We use the same lexical items as Swanson, adding a few extra content words to ensure adequate coverage. Like Swanson's, our additional words are monosyllables with lax vowels and stop-stop contexts ("cat", "bat", "bit", "pick", "tip", "kick", "kid", "duck", "gum", "bet", "dead"). Tokens are labeled phrase-medial if they are directly adjacent in the alignment to the words on either side. Tokens

Table 2: *Comparative results of three forced alignment systems.*

	Recall		Precision	
	Acceptable	Catastrophic	Acceptable	Catastrophic
Full pipeline	70% (30819)	24% (10449)	70% (30490)	24% (10499)
Baseline	30% (10170)	60% (26772)	30% (10170)	46% (15220)
Pate HTK	7% (3017)	72% (31890)	8% (3485)	71% (29546)

Table 3: *Comparative results of three forced alignment systems on vowels only.*

	Recall		Precision	
	Acceptable	Catastrophic	Acceptable	Catastrophic
Full pipeline	77% (12914)	17% (2924)	71% (12302)	21% (3613)
Baseline	16% (2742)	53% (8983)	43% (5716)	36% (4730)
Pate HTK	13% (2255)	46% (7676)	18% (2929)	45% (7502)

Table 4: *Duration statistics for function and content words in CDS and ADS.*

		N	Mean dur	Std. dev
Func.	Swanson ADS	519	38.36	14.23
	Swanson CDS	599	42.26	14.72
	Our CDS	27315	62.31	45.63
Cont. medial	Swanson ADS	355	139.6	37.70
	Swanson CDS	350	154.3	48.77
	Our CDS	1078	108.5	84.23
Cont. final	Swanson ADS	355	178.4	49.75
	Swanson CDS	337	204.9	59.75
	Our CDS	648	204.1	127.96

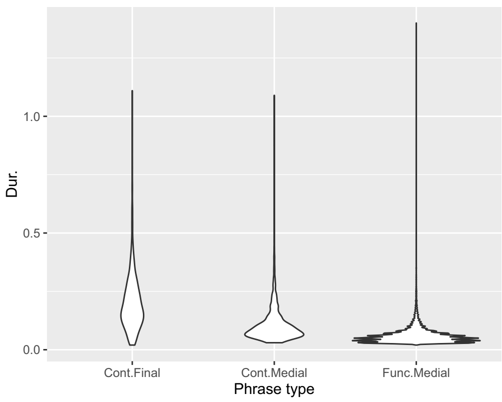

Figure 3: *Distribution plot of vowel durations in selected content and function words in our aligned data. Outliers > 2s have been removed.*

followed by an interval labeled silence are labeled phrase-final. Outliers with duration > 2 seconds were filtered out as obvious errors. (Following Swanson, phrase-final function words are not analyzed.)

Figure 3 shows the distributions from our corpus; Table 4 gives case-by-case comparisons with Swanson's data. Our results show the same effects for CDS (but without the ADS comparison): function words are the shortest (mean duration 62 ms). All content words are longer than function words: Medial content words have a mean duration of 108 ms, and there is a final lengthening effect for phrase-final content words (204 ms).

Methodologically, the corpus approach to such a study has advantages and disadvantages. Since the corpus already exists, the analysis is extremely quick and cheap, but at the same time can use very large numbers of samples. Ecological validity is high— there is no manipulation that subjects could potentially detect. On the other hand, we cannot collect speaker-matched ADS for these speakers; a full replication of Swanson's results using only corpus data would require cross-speaker comparisons with a separate corpus. Moreover, there is less ability to control for prosodic environment; our "medial" and "final" categories are loose, since we cannot control for lengthening due to prosodic focus or hesitations. This lack of control is probably one reason that our variances are higher than Swanson's; another contributing factor is that the alignments are imperfect. But although the variances differ, the means seem relatively reliable.

These analyses suggest that our alignments are accurate enough to perform statistical analyses of CDS phonetics using the Providence corpus. This could significantly reduce the cost of data annotation for a variety of phonetic studies.

6. Future work

Although the current system has reasonably high accuracy, possibilities for improvement still exist. Preliminary tests with Kaldi's neural network systems [30] had poor performance on this dataset, probably because of the very limited data available for retraining the network. But with larger datasets, we anticipate that neural network recognizers will be able to outperform HMM/GMMs. Another possibility is automatic quality control— even if the aligner is not perfect, a system that can detect and filter out mishandled segments or utterances would create cleaner datasets for phonetic analysis. We are in the process of designing such a system.

7. Acknowledgements

We are grateful to John Pate and to Eric Fosler-Lussier, Stephanie Antetomaso and the Ohio State Language Acquisition reading group for their comments. This work was funded by NSF 1422987 to the first author.

8. References

[1] B. MacWhinney, *The CHILDES Project: Tools for Analyzing Talk. Vol 2: The Database*, 3rd ed. Mahwah, NJ: Lawrence Erlbaum Associates, 2000.

[2] J. Gilkerson and J. A. Richards, "The LENA natural language

study," *Boulder, CO: LENA Foundation. Retrieved March*, vol. 3, p. 2009, 2008.

[3] M. A. Pitt, K. Johnson, E. Hume, S. Kiesling, and W. Raymond, "The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability," *Speech Communication*, vol. 45, no. 1, pp. 89–95, 2005. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0167639304000974

[4] M. A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier, "Buckeye corpus of conversational speech (2nd release)," 2007.

[5] R. Mazuka, Y. Igarashi, and K. Nishikawa, "Input for learning Japanese : RIKEN Japanese mother-infant conversation corpus(coe workshop session 2)," *Technical report of IEICE. Thought and language*, vol. 106, no. 165, pp. 11–15, jul 2006. [Online]. Available: http://ci.nii.ac.jp/naid/110004809988/en/

[6] A. Cristiá, "Fine-grained variation in caregivers' /s/predicts their infants' /s/category," *The Journal of the Acoustical Society of America*, vol. 129, p. 3271, 2011.

[7] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus," *Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3878, 2008.

[8] R. K. Moore, "A comparison of the data requirements of automatic speech recognition systems and human listeners." in *INTERSPEECH*, 2003.

[9] P. C. Woodland, J. J. Odell, V. Valtchev, and S. J. Young, "Large vocabulary continuous speech recognition using HTK," in *ICASSP-94*, vol. 2, 1994.

[10] K. Demuth, J. Culbertson, and J. Alter, "Word-minimality, epenthesis and coda licensing in the early acquisition of English," *Language and Speech*, vol. 49, no. 2, pp. 137–173, 2006.

[11] J. K. Pate and M. Johnson, "Syllable weight encodes mostly the same information for English word segmentation as dictionary stress," in *Proceedings of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2014.

[12] B. Börschinger and M. Johnson, "Exploring the role of stress in Bayesian word segmentation using adaptor grammars," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 93–104, 2014.

[13] D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glambek, N. Goel, M. Karafiat, A. Rastrow, R. Rose, P. Schwarz, and S. Thomas, "The subspace gaussian mixture model-a structured model for speech recognition," *Computer Speech & Language*, vol. 25, no. 2, pp. 404–439, April 2011.

[14] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proceedings of ACL*, vol. I, San Francisco, 1992, pp. 517–520.

[15] K. Vertanen, "Baseline WSJ acoustic models for HTK and Sphinx: Training recipes and recognition experiments," Technical report). Cambridge, United Kingdom: Cavendish Laboratory, Tech. Rep., 2006.

[16] L. A. Swanson, L. B. Leonard, and J. Gandour, "Vowel duration in mothers' speech to young children," *Journal of Speech, Language, and Hearing Research*, vol. 35, no. 3, pp. 617–625, 1992.

[17] N. B. Ratner, "Patterns of vowel modification in mother–child speech," *Journal of child language*, vol. 11, no. 03, pp. 557–578, 1984.

[18] A. Fernald, T. Taeschner, J. Dunn, M. Papousek, B. de Boysson-Bardies, and I. Fukui, "A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants," *Journal of child language*, vol. 16, no. 03, pp. 477–501, 1989.

[19] C. P. Boner and C. R. Boner, "Minimizing feedback in sound systems and room-ring modes with passive networks," *The Journal of the Acoustical Society of America*, vol. 37, no. 1, pp. 131–135, 1965.

[20] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (computer program)," 2017. [Online]. Available: http://www.praat.org

[21] R. Iyer and M. Ostendorf, "Transforming out-of-domain estimates to improve in-domain language models," in *eurospeech*, vol. 4, 1997, pp. 1975–1978.

[22] J. Foster, J. Wagner, D. Seddah, and J. van Genabith, "Adapting WSJ-trained parsers to the British National Corpus using in-domain self-training," in *Proceedings of the Tenth International Conference on Parsing Technologies*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 33–35. [Online]. Available: http://www.aclweb.org/anthology/W/W07/W07-2204

[23] J. R. Novak, N. Minematsu, and K. Hirose, "Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the wfst framework," *Natural Language Engineering*, pp. 1–32, 2015.

[24] P. Jyothi, E. Fosler-Lussier, and K. Livescu, "Discriminatively learning factorized finite state pronunciation models from dynamic bayesian networks," in *INTERSPEECH*, 2012, pp. 1063–1066.

[25] J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," *The Journal of the Acoustical society of America*, vol. 97, p. 3099, 1995.

[26] G. K. Vallabha, J. L. McClelland, F. Pons, J. F. Werker, and S. Amano, "Unsupervised learning of vowel categories from infant-directed speech," *Proceedings of the National Academy of Sciences*, vol. 104, no. 33, pp. 13 273–13 278, 2007.

[27] N. H. Feldman, E. B. Myers, K. S. White, T. L. Griffiths, and J. L. Morgan, "Word-level information influences phonetic learning in adults and infants," *Cognition*, vol. 127, no. 3, pp. 427–438, 2013.

[28] J. Maye, R. N. Aslin, and M. K. Tanenhaus, "The weckud wetch of the wast: Lexical adaptation to a novel accent," *Cognitive Science*, vol. 32, no. 3, pp. 543–562, 2008.

[29] A. E. Turk and S. Shattuck-Hufnagel, "Multiple targets of phrase-final lengthening in American English words," *Journal of Phonetics*, vol. 35, no. 4, pp. 445–472, 2007.

[30] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, "Sequence discriminative training of deep neural networks," in *Proc. INTERSPEECH*, 2013, pp. 2345–2349.