Methods for eliciting, annotating, and analyzing databases for child speech development

Mary E. Beckman[a], Andrew R. Plummer[b,*], Benjamin Munson[c], Patrick F. Reidy[d]

[a]*Linguistics, Ohio State University*
[b]*Computer Science and Engineering, Ohio State University*
[c]*Speech-Language-Hearing Sciences, University of Minnesota*
[d]*Callier Center for Communication Disorders, University of Texas at Dallas*

Abstract

Methods from automatic speech recognition (ASR), such as segmentation and forced alignment, have facilitated the rapid annotation and analysis of very large adult speech databases and databases of caregiver-infant interaction, enabling advances in speech science that were unimaginable just a few decades ago. This paper centers on two main problems that must be addressed in order to have analogous resources for developing and exploiting databases of young children's speech. The first problem is to understand and appreciate the differences between adult and child speech that cause ASR models developed for adult speech to fail when applied to child speech. These differences include the fact that children's vocal tracts are smaller than those of adult males and also changing rapidly in size and shape over the course of development, leading to between-talker variability across age groups that dwarfs the between-talker differences between adult men and women. Moreover, children do not achieve fully adult-like speech motor control until they are young adults, and their vocabularies and phonological proficiency are developing as well, leading to considerably more within-talker variability as well as more between-talker variability. The second problem then is to determine what annotation schemas and analysis techniques can most usefully capture relevant aspects of this variability. Indeed, standard acoustic characterizations applied to child speech reveal that adult-centered annotation

*Corresponding author
Email address: plummer.321@osu.edu (Andrew R. Plummer)

schemas fail to capture phenomena such as the emergence of covert contrasts in children's developing phonological systems, while also revealing children's nonuniform progression toward community speech norms as they acquire the phonological systems of their native languages. Both problems point to the need for more basic research into the growth and development of the articulatory system (as well as of the lexicon and phonological system) that is oriented explicitly toward the construction of age-appropriate computational models.

Keywords: child speech development, phonetic transcription, spectral kinematics, automatic speech recognition, big data corpora

---

1. Introduction

This paper reviews methods for gathering, annotating, and analyzing collections of recordings to improve our understanding of child speech development. One goal of the review is to evaluate how current methods might be adapted and the toolbox of methods expanded to support the kinds of very large-scale database development efforts that are currently possible for adult speech databases.

In research on adult speech, the prerequisites for progress in science and for advances in technology have long played complementary, mutually beneficial roles in prompting the creation and exploitation of annotated speech databases. For example, the source-filter model of segmental production that was formalized in the acoustic theory of speech production (Chiba and Kajiyama, 1941; Fant, 1960) prompted a phonological theory of acoustically-grounded distinctive features for consonants and vowels (Jakobson et al., 1951) and much seminal work on acoustic correlates which served as the foundation for early formant-based text-to-speech synthesis (TTS) rule systems (see review in Klatt, 1987). The challenge then of developing rules to also specify suprasegmental parameter values and context-related variation in segment-level spectral parameter values in connected speech was an important force driving the creation and annotation of several early adult speech databases (e.g., Umeda, 1975, 1976), as well as the development of other synthesis methods (e.g., Coker, 1976; Olive, 1977). Conversely, the adoption of methods from automatic speech recognition technology to develop forced alignment tools to improve duration models in concatenative TTS systems (e.g., Wightman and Talkin, 1997) aided also in the annota-

tion of collections of recordings that were gathered for other purposes, such as exploring the likelihood and extent of cross-word place assimilation processes (Dilley and Pitt, 2007) and differentiating among models of the relationship between word frequency and phoneme duration in spontaneous speech (Gahl et al., 2012). More recent examples of this interplay between science and technology include experiments to apply analysis-by-synthesis methods to annotate collections of audio recordings with imputed gestural score representations from Articulatory Phonology so as to improve word recognition in noise (Mitra et al., 2012, 2014) and a study combining methods from ASR with information-theoretic measures of functional load to evaluate a distinction between robust and "marginal" phonological contrast (Renwick et al., 2016).

To an increasing extent, research on child speech development also is benefiting from this interplay, most recently from the application of ASR technology to help quantify adult speech directed at infants in hours-long recordings made in their homes (see, e.g. Weisleder and Fernald, 2013). Related methods are also being developed to automatically identify different types of prespeech vocalizations produced by the infants themselves (see, e.g., Oller et al., 2010), offering a clear potential to resolve the long-standing problem of how to gather and efficiently tag representative samples of infant productions at the earliest stages of learning to communicate by voice. In Sections 2 and 3, we briefly review these methods, and then catalog the added challenges of sampling the words and longer referential vocalizations that older infants begin to produce toward the end of these initial stages of learning to talk, in order to evaluate whether current speech technology can be exploited to build comparably large databases of child speech. Section 4 then summarizes the lessons that can be gleaned from the ways in which ASR methods developed for adult speech databases have failed when applied to recordings of children's speech. One of these lessons is that subword annotation schemes and associated speech production models that might be adequate for developing large-scale databases of adult, adolescent, and older children's speech are considerably less useful when applied to younger children's speech. Section 5 then addresses the challenges of developing more useful annotations of recordings of speech produced by preschool children, and Section 6 describes some of the measures that have been developed for interpreting the relationships between annotation categories and speech production models that might be more appropriate for very young speakers. Section 7 will close by speculating on how methods described in Sections 2 through 6 might be

scaled up to begin to exploit the capacity to make hours-long recordings of speech produced spontaneously by young children in ordinary interactions with their families and with others in their widening social circles.

2. Databases of infant-directed speech and infant "speech"

As noted in the introduction, one way in which methods for adult speech databases have been usefully adapted to the study of child speech development is in transcribing hours-long recordings of speech that infants hear in their homes. In Hart and Risley's (1995) landmark study of the effects of variable amounts of such input on early vocabulary development, it took an average of 8 hours to orthographically transcribe each hour-long recording, so that there were resources to make only one recording a month over 2.5 years for each of 60 target infants, yielding a total database size of about 1300 hours. By contrast, the ASR-based rough diarization and transcription of child-directed speech that is provided as part of the Language ENvironment Analysis (LENA) system[1] is reported to be able to produce a first-pass time-aligned annotation of a day-long (16-hour) recording in 2 hours of processing time (VanDam et al., 2016, p. 130). While the adult word counts from these annotations are not error free (cf. Oetting et al., 2009; Canault et al., 2016; VanDam and Silbert, 2016), correlation with human transcribers seems to be high enough to make it feasible to replicate and extend Hart and Risley's results using databases that are an order of magnitude larger (see, e.g., Weisleder and Fernald, 2013; Pae et al., 2016; Suskind et al., 2016).

Moreover, the rough diarization is reliable enough that Oller et al. (2010) were able to use the more than 3.1 million infant vocalizations identified in a database of nearly 1500 day-long recordings to train an algorithm for estimating developmental age from acoustic features that differentiate among the progressively more speech-like vocalization types that have been described in annotation taxonomies developed independently by Stark (1980), Oller (1980), and Koopmans-van Beinum and van der Stelt (1986). They found strong correlations between estimated developmental age and chronological age for both the 106 children with typical language development and the 49 children with language delay not related to autism, with the expected lag in the relationship in the latter group. They also found a much less typical

[1] https://www.lena.org/products

developmental progression for the 77 children with autism on several of the acoustically-defined categories, particularly those related to voice quality and pitch control, which differentiated these 77 children from both other groups, in keeping with the results of a much smaller-scale earlier study by Sheinkopf et al. (2000).

The discovery of the typical developmental progression from earliest reflexive cry to the variegated babbling that is concurrent with first words is itself another example of the rich interplay between speech science and technology. The broad outline of the progression is already evident in the tallies of different vocalization types in Shirley's (1931) longitudinal study. That is, decades before Oller et al. (1976) and Vihman et al. (1985) definitely refuted Jakobson's (1941) hypothesized discontinuity between children's prespeech babbling and their early speech, astute observers already recognized differences in an infant's vocalizations at different ages, with later babbling hardly distinguishable from vocalizations that are recognized by the parents as the infant's first words. These early 20th-century researchers also were keenly aware of the technological limitations of the time, pointing both to "the difficulty of stimulating vocal responses" during a testing interval dedicated to focused observation of vocalization type (Shirley, 1931, p. 47) and to the fact that "the cries and various vocalizations of the infant which are preliminary to the complexities of language proper include many sounds for which we have no adequate written symbols" (McCarthy, 1929b, p. 636).

In his pilot investigation of "the use of magnetic devices" to overcome these difficulties, Lynip (1951) recruited a neonate's parents to collaborate in making a longitudinal database of weekly hour-long audio recordings and then made spectrograms of selected vocalizations that he proposed to use in lieu of any kind of symbolic annotation—e.g., by relating frequencies of spectral peaks in the infant's first "non-crying vocalizations" in the recordings at weeks 8, 9, and 10 to formant patterns in specific adult vowels. Later studies of longitudinal databases of audio recordings, such as Murai (1963), Davis and MacNeilage (1995), Koopmans-van Beinum et al. (2001), and Ishizuka et al. (2007), did not treat spectrographic records as substitutes for phonetic transcription, particularly of canonical babble, but instead related both the spectral analyses and the interpretation of the transcriptions more directly to what was being discovered about the effect of (lack of) auditory input and about how an infant's vocal anatomy changes over the first years of life (e.g. Sasaki et al., 1977; Crelin, 1987). In developing their analyses of the large database of day-long recordings for 232 children, then, Oller et al. (2010)

could build on a wealth of qualitative descriptions and acoustic analyses, accumulated in several decades of intense study of smaller databases of shorter recordings for fewer children (see reviews in Warlaumont et al., 2010; Buder et al., 2013).

Many of these smaller databases also took advantage of the development of video recording technology and of methods for more controlled elicitation of infant vocalizations or caregiver-infant interactions in the laboratory. For example, Hsu et al. (2001) tallied vocalizations produced by young infants in face-to-face interactions with their mothers in weekly video recording sessions between 4 and 24 weeks and found systematically higher rates of fully-resonant vowel-like vocalizations when an infant was looking at the mother's face and the mother was smiling or the infant was about to smile. Relatedly, Franklin et al. (2014) found that 6-month-old infants' volubility and vocalization type could be manipulated experimentally by asking mothers to produce an interval of unresponsive, unsmiling "still face" in the middle of such an interaction session. Gros-Louis et al. (2006) looked at vocalizations produced by even older infants and also categorized their mothers' vocal response types in half-hour video recordings of unstructured play sessions, and found that mothers were more likely to respond to simple vowel-like vocalizations with play vocalizations (e.g., *vroom-vroom* for a toy car) and more likely to respond to canonical babble with imitations and elaborations (e.g. *Ba-ba.* or *Mama, yes, and Dada is working.*). In a subsequent study, Goldstein and Schwade (2008) found that the proportion of canonical babble vocalizations produced by 9-month-old infants could be manipulated experimentally by having the mothers produce imitative responses on a contingent or non-contingent response schedule.

The results reported in Oller et al. (2010) suggest that we are now well-positioned to extend such research to much larger databases of more representative samples of caregiver-infant interactions recorded in their homes. That is, although there is more work yet to be done before we can definitively evaluate the reliability of the algorithm as a tool for automatically tagging the vocalization type, it does seem reliable enough to already be useful for studying differences in interaction patterns across well-differentiated groups, such as children with normal hearing versus children with hearing impairment (e.g. VanDam et al., 2015) and children with and without autism (e.g. Warlaumont and Ramsdell-Hudock, 2016).

There is also a recently-funded project to gather such day-long recordings

into a research archive[2] (VanDam et al., 2016) as well as a large interdisciplinary working group[3] committed to improving the annotation of these recordings by doing such things as making the diarization more precise (i.e., identifying the specific talker and not just the talker type) and differentiating child-directed speech from adult-directed speech. It seems likely, therefore, that we will soon be able to exploit such databases also for other purposes, such as measuring the formants in vowel-like portions of the infants' vocalizations to replicate and extend several landmark studies of cross-language differences in vowel development (de Boysson-Bardies et al., 1989; Rvachew et al., 2006), or basing a more elaborated annotation method on the automatic tagging of the vocalization type. For example, imagine a set of modeling studies that extends Serkhane et al. (2007) versus Nam et al. (2013) to phonetically transcribed canonical babble produced by infants acquiring languages other than American English. The data might be got from day-long recordings gathered for other purposes, such as Canault et al. (2016), by picking out just the vocalizations that have been identified as canonical babble in the recordings made between 6 and 12 months, using these as stimuli in a web-based experiment where several adults imitate each vocalization, and then applying standard ASR to the adult imitations to make a consensus phonetic transcription.

3. Building databases of children's speech

Could such methods be used to exploit day-long recordings also as a source of databases of infants' and young children's speech productions? Obviously, having methods to pick out infant vocalizations and identify the general vocalization type is an important step toward that goal. However, it is only the first step. Insofar as infant speech is speech, the intended target utterance (i.e., the word or phrase that the addressee understands the infant or toddler to be saying) also needs to be identified so that the differences or similarities to community production norms for the target form can be gauged. The difficulty of identifying the intended utterance in very young children's speech is illustrated by McCarthy's (1929a) experience in asking open-ended questions about toys and picture-books in order to elicit

[2]http://homebank.talkbank.org/
[3]http://darcle.org/

50 utterances for the experimenter to write down and analyze to gauge morphosyntactic development in each of 20 children tested at regular intervals starting at age 18 months. On average, only 25% of the fifty utterances elicited from each child at 18 months was "comprehensible" and hence useable in the analyses of vocabulary growth and utterance length. A later study by Weist and Kruppe (1977) showed that only half of words produced even by a 3-year-old child can be identified correctly by listeners who are not the child's parent or sibling, and Flipsen (1995) found similar results for 4- through 7-year-old children with speech sound disorders (SSD, i.e, extremely inaccurate speech production in the absence of a clear medical or psychosocial etiology). It seems likely, then, that if we aim to develop methods for automatically annotating the target child's speech in day-long recordings in the home or day-care center, these methods will need to build on studies of smaller databases of child speech that have been collected in other ways that allow for a more reliable identification of the target utterances.

One potential source of such databases is language samples that have already been transcribed orthographically and/or alphabetically. In the clinical literature, "language sample" is a general term for a recording of connected speech that the target child produces spontaneously in naturalistic interactions either with the caregiver or with an adult observer such as an experimenter or clinician. The methods used to elicit such samples can be more or less structured to elicit target utterance types.

An example of a more structured elicitation is the longitudinal database of play sessions with an experimenter/observer that were recorded by Macken and Barton (1980). For these sessions, the experimenters consulted with each child's parents to be able to use toys (e.g., a *Piglet* doll) or games (e.g., pretending to drive a car into a garage) as child-specific prompts to elicit as many instances as possible of familiar words that began with a voiced or voiceless stop for the express purpose of looking at the distribution of voice onset time (VOT) values at different stages of developing a robust voicing contrast.

An example of a completely unstructured elicitation method is the longitudinal database of mother-child interactions and solitary play recorded for the Stanford Corpus that provided the alphabetic transcriptions for the study reported in Vihman et al. (1985) and for many subsequent studies of early phonological development in English, French, Japanese, and Swedish (e.g., de Boysson-Bardies and Vihman, 1991; Vihman, 1993; McCune and Vihman, 2001). In these weekly recording sessions, the experimenter/observer

set up the audio and video recorders and then took notes as unobtrusively as possible. Because there were fewer than 10 children per language and because the transcription protocol involved frequent inter-transcriber calibration and consensus-building exercises, all of the transcribers became very familiar with every child. This familiarity, augmented by the video record of what the mother and child were doing when the child produced a vocalization, made it possible to develop consistent criteria for differentiating referential words from babbling, and for determining the target utterance for the first category. Many of the longitudinal recordings that have been contributed to the PhonBank archive[4], such as the Davis Corpus for English (Davis and MacNeilage, 1995), the CLPF Corpus for Dutch (Levelt, 1994), the Ota Corpus for Japanese (Ota, 2003), the Lyon Corpus for French (Demuth and Tremblay, 2008), and the "babbling and first words" portion of the Stuttgart Corpus for German (Lintfert, 2009), are similar to the Stanford Corpus in their collection and annotation methods.

Intermediate between these endpoints are many of the conversations that a speech language pathologist (SLP) might record as part of a language assessment. Sometimes the SLP will use toys and picture books as prompts to try to elicit a reasonably phonetically balanced set of words or phrases that can be phonetically transcribed to derive the "Percentage Consonants Correct" (PCC, a measure of segmental production accuracy), as described in Shriberg and Kwiatkowski (1982, 1985), but often the sample is less structured because it is intended instead to let the SLP calculate some other measure such as "Mean Length of Utterance" (MLU, a measure of morphosyntactic development). A specific example of this type of less structured elicitation method is the longitudinal database of conversations recorded by Nittrouer and colleagues as part of a larger battery of tests of language development. Although specific sets of consonant sounds were later extracted from these conversations for the acoustic analyses described in McGowan et al. (2004) and Lowenstein and Nittrouer (2008), the soft toys that were provided to the children to play with during the recording session "were not chosen to elicit specific responses from the children but were rather meant to stimulate communication in general" (Lowenstein and Nittrouer, 2008, p. 1184).

A primary advantage of using language samples to study speech sound development is ecological validity. The contexts in which language samples

[4]http://phonbank.talkbank.org/

are taken resemble natural communicative situations more closely than does the task of naming pictures (a much more commonly used elicitation method in tests of segmental production accuracy—see below). Hence, the results of language samples are thought to be more predictive of real-world communication ability than are the results of single-word naming tests. Two potential disadvantages are that sampling of some target utterance types is not guaranteed unless the elicitation was structured specifically to elicit them and that confounding ambiguity and even substantial data loss can accrue if careful records were not taken at the time of recording so as to be able to clearly identify the intended word or phrase for all potentially relevant utterances. For example, Lowenstein and Nittrouer (2008) found that the target voicing of 24% of word-initial stop consonants could not be determined, confounding the interpretation of the changing distributions of VOT values over the ages sampled.

Another potential source of databases of children's speech is recordings made in the course of administering clinical tests of articulation accuracy or of intelligibility. Examples of tests of intelligibility are the Children's Speech Intelligibility Measure (CSIM, Wilcox and Morris, 1999) and the Beginner's Intelligibility Test (BIT, Osberger et al., 1994). For the CSIM, the tester chooses one word at random from each of 50 lists of 12 similar-sounding words (e.g., one word from the set *tall, stall, wall, shawl, call, all, fall, ball, hall, crawl, mall, Paul*), and says the word for the child to repeat. For the BIT, similarly, the tester pronounces sentences such as *The boy walked to the table.*, optionally with pictures depicting the sentence or with toys that can be used to enact the sentence to help the child understand and remember the sentence better to repeat in its entirety. The child's repetitions of the target words or sentences are then played to at least two independent judges to transcribe. Since the child's utterances are necessarily recorded in order for them to be transcribed, they can then be used for other later analysis. For example, because the CSIM was among the tests administered at 36, 42, and 48 months to the children in the longitudinal study in Lowenstein and Nittrouer (2008), their productions of the CSIM words could be included along with the identifiable words from the language samples in McGowan et al.'s (2014) longitudinal study of vowel formant frequencies.

An example of a very typical test of articulation accuracy is the Goldman-Fristoe Test of Articulation-3 (GFTA, Goldman and Fristoe, 2015, see also earlier versions in 1986, 2001), which is widely used to assess whether an American-English-acquiring child's speech meets age expectations, or is be-

hind. In the GFTA, children are presented with a series of simple pictures that they name. In cases where a child does not produce the intended word (e.g., not knowing the word *shovel* or saying *phone* rather than *telephone*), there is a standard series of prompts to elicit productions that are repetitions of the tester's model utterance. Thus, as in the repetition task that provides the productions for intelligibility, there is a much lower risk of data loss, because the target is always known. Moreover, words to be named can be selected so that they contain a representative sample of the language's phonemes. For example, pictures on the GFTA are designed to elicit the full set of English consonants in the initial, medial, and final positions of words. While the GFTA is not designed to elicit the full set of vowels, other standardized picture-naming tests are (e.g., the Arizona Articulation Proficiency Scale-3, Faluda, 2000).

Indeed, picture naming has been used for many decades to develop the normative statistics that figure in the diagnosis of SSD (see, e.g. Wellman et al., 1931; Templin, 1957; Smit et al., 1990), and tests like the GFTA have been developed for other dialects of English and for many other languages (see reviews in McLeod, 2007; McLeod and Goldstein, 2012), so they are a potentially very rich source of data if clinicians can be encouraged to record the elicited productions. That is, although researchers who include the GFTA as part of a battery of tests accompanying an experimental task often do record the children's responses in order to be able to report inter-transcriber reliability (see, e.g., Rvachew and Grawburg, 2006), the GFTA guidelines are merely to annotate correctness on the fly using a publisher-provided check sheet, which leaves a space for optionally using phonetic transcription to make a record of the perceived error if the child's pronunciation is judged to be incorrect. However, SLPs probably could be recruited to make audio recordings as well if contributing such recordings to a research archive such as PhonBank were to lead to ASR-based tools to aid in the administration of the test and the interpretation of the test results. This should be appealing especially to SLPs who work in the many school districts in the U.S. where they might be tasked with using an instrument such as the GFTA to assess the speech therapy needs of every child at school entry.

Another elicitation method, described in Edwards and Beckman (2008a), combines the elicitation method of the typical intelligibility test with that of the typical test of articulation accuracy. In this picture-prompted word-repetition task, the child sees a picture on a computer screen, hears a recorded audio prompt that names the picture, and is asked to repeat "what the com-

puter said." Since the picture is always named for the child, the constraints on pictureability are relaxed. Indeed, the task can be used even to elicit repetitions of nonsense words. Moreover, the protocols for modeling the picture's name for a child who does not know the intended word are eliminated, so that elicitation is uniform across age groups and can proceed very quickly even when eliciting productions from very young children who have small vocabularies. This also means that many more productions can be elicited. For example, Edwards and Beckman used the task to elicit three words for each of the lingual obstruents of Cantonese, English, Greek and Japanese in each of five broadly defined vowel contexts from 20 two- and three-year-old children per language, to be able to compare transcribed consonant accuracy in productions of similar phoneme sequences with different language-specific type frequencies (Edwards and Beckman, 2008b; Beckman and Edwards, 2010). A subsequent collection of recordings of a somewhat more limited set of target consonants from eighty 2- through 5-year-old children for each of these four languages is now part of the PhonBank archive, and has been used in acoustic studies by other researchers (e.g. Reidy, 2015; Bang et al., 2015) as well as by researchers involved in the original data collection (e.g., Li, 2012; Kong et al., 2012; Holliday et al., 2015).

When there are fewer target utterance types, a different variant of picture-naming can be used in which the child participants are first taught the names of all of the pictures which then are presented in random order for the child to name. This method is particularly appropriate when the "words" to be named are nonsense forms which are taught to the child as proper nouns—e.g., as the names of stuffed toy animals, or puppet figures, or computer game characters (see, e.g. White, 2001; Bunnell et al., 2000). In experiments where articulatory records also will be made using a device such as an Optotrak camera system, electromagnetometer, or ultrasound machine, the children might also be trained to say the target name in a frame that is selected to provide a consistent coarticulatory context (see, e.g., Noiray et al., 2004; Zharkova et al., 2012). Picture-naming and repetition can also be combined to elicit baseline productions and imitations of synthetically manipulated audio prompts (see, e.g. Nielsen, 2014; Kleber and Peters, 2014). Other combinations of the tasks can be used to test the effects of lexicality (i.e., imputed referential status) on phonological accuracy in novel word learning (see, e.g., Goffman et al., 2007; Heisler and Goffman, 2016)

Picture naming can be used with older children, too, although with school-aged children, elicitation will be quicker if written prompts are used

instead, with an adult at hand to produce the prompt for the child to repeat just in case the child misreads the prompt. This is an efficient way to record phonetically balanced materials from a large number of school-aged children. Examples of databases that use this method are the TIDIGITS corpus (Leonard, 1984), the CMU KIDS corpus (Eskenazi, 1996; Eskenazi et al., 1997), the CID corpus (Miller et al., 1996; Lee et al., 1999), and the OGI / CSLU Kid's Speech corpus (Shobaki et al., 2000, 2007) for American English; the ChildIt corpus for Italian (Giuliani et al., 2006; Gerosa et al., 2007); the Swedish portion of the PF_STAR Children's Speech Corpus (Batliner et al., 2005); and the SingaKids–Mandarin corpus for Singapore Mandarin Chinese (Chen et al., 2016).

Other methods that can be used to elicit more spontaneous speech include interactive computer games, robot-guiding tasks, and the like. Examples of databases that use these methods are the Takemaru-kun recordings for Japanese (Nisimura et al., 2003; Cincarek et al., 2007), the AIBO robot interaction portions of the PF_STAR Children's Speech Corpus for German and UK English (Batliner et al., 2005), and the CNG corpus for European Portuguese (Hämäläinen et al., 2013, 2014).

4. How and why standard ASR methods fail for child speech

All of the databases listed in the last two paragraphs of Section 3 were developed for the purpose of training ASR models, and studies using them for that purpose have contributed to our understanding of the differences between adult speech and child speech that make it challenging to adapt ASR-based methods from adult speech databases to expedite the annotation and analysis of databases of child speech. A much-cited study by Wilpon and Jacobsen (1996) is a useful starting point for reviewing these differences. In this study, Wilpon and Jacobsen trained HMM models on various subsets of a corpus of recordings of digits produced by more than 1000 speakers from the Danish portion of the Rafael.0 database (Rosenbeck et al., 1994), including about 100 speakers in each of the categories "children" (8-to-12 year olds) and "elderly adults" (speakers 60 years and older) as well as 200 or 300 speakers in each of three age bands in between 13 and 59 years. A baseline model trained on all of the available data yielded a word error rate for the children's productions that was more than twice the error rates for the middle three bands (4.6%, as compared to 2%, 1.8%, and 1.4% for the teenagers, young adults, and middle-aged adults, respectively). Results of

various experiments in normalizing the total size of the training data set to be equal to that of the smallest age-band-specific dataset showed that the higher error rate for children in the original model was not merely an artifact of data sparsity. Analyzing results separately by talker gender in order to explore "physiological explanations" also suggested more fundamental differences between adult's and children's speech patterns that would need to be better understood before we try to apply the kinds of adaptation methods that have been used to address data sparsity issues when building gender-neutral speaker-independent ASR models for adults. (See Moore, 2003, for another line of argument for this point.)

In the two decades since this seminal study with the Danish portion of the Rafael.0 corpus, many more databases have been created that include speech of school-aged children, recorded for the purpose of exploring the differences between child and adult speech and developing ASR methods that can accommodate to those differences. (See the partial list at the end of Section 3 and also see Claus et al., 2013, for a comprehensive recent survey.) A common theme in studies that use these corpora echoes and expands on the conclusions from Wilpon and Jacobsen (1996). ASR models yield higher error rates when applied to speech produced by children as compared to speech produced by adults, and this is true whether the models are trained on just children, on just adults, or on age-diverse training sets. Moreover, the increase in error rates relative to ASR with adult speech tends to be greater the younger the children (see, e.g., the studies reviewed or reported in Potamianos and Narayanan, 1998; Stemmer et al., 2003; Elenius and Blomberg, 2005; Cincarek et al., 2007; Gerosa et al., 2009b; Shivakumar et al., 2014). Prominent among the explanations that have been offered for these higher error rates is the fact that children's speech is more variable, in ways that strongly compound the perennial data sparsity problem (see, e.g., Potamianos and Narayanan, 2003; Benzeghiba et al., 2007).

Some of this greater variability is related to the fact that children's vocal tracts are changing dramatically across the course of development, with overall vocal tract length increasing on average from about 10 cm at age 5 years to about 15 cm at 12 years of age, according to the model of longitudinal data by Barbier et al. (2015), which conforms with earlier models of cross-sectional data such as Fitch and Giedd (1999) and Vorperian et al. (2009). This means that ASR methods for school-aged children's speech must accommodate to a range of differences in vocal tract length that can be as large as 150% from youngest to oldest speaker (as compared to a difference of about

110% between adult females and males).

Also, while there is no good evidence of gender-related physiological differences in vocal tract characteristics before puberty, school-age children are practicing phonetic markers of different social identities and relationships, including markers of gender identity as well as of inclusion/affiliation versus exclusion/dissociation in relationship to social groups defined by ethnicity, family socioeconomic status, age cohort, and so on (see, e.g. Drager, 2011; Kohn and Farrington, 2012; Alam and Stuart-Smith, 2014; Li et al., 2016, among many others). This leads to potentially large pools of more or less fluent socially-motivated variation, as boys and girls negotiate gender-specific, group-specific, cohort-specific, or even clique-specific pronunciation habits that will continue to evolve even as they grow into adolescents who must further adjust their speech motor habits to the physiological changes to the vocal tract and glottis that are associated with hormonal changes at puberty.

Moreover, when pre-school children are included, models must adjust for differences in vocal tract shape that are at least as consequential as the differences between females and males post-puberty. That is, the rapid 10% growth in vocal tract length when a boy goes through puberty involves no change in oral cavity length, but a small increase in the lip tube length (due to the thickening of the lips) and a 1-to-2 cm increase in pharyngeal cavity length (due to the descent of the larynx). The analogous reshaping at the beginning of life involves an initial descent of the larynx to disengage from the velopharynx and create a short pharynx along with a rapid increase in hard palate length during the first year or so, after which a child's oral cavity length grows hardly at all compared to the further increase in pharyngeal cavity length as the height of the mandible doubles, until the oral and pharyngeal cavities reach the 1:1 adult female proportion at about 5 or 6 years of age (see, e.g., Fig. 3 in Barbier et al., 2015). These differences in growth patterns before and after school entry might help explain some of the differences across studies in the effectiveness of adapting methods developed for adult speaker-independent ASR such as frequency warping and vocal tract length normalization techniques (see, e.g., Potamianos et al., 1997; Claes et al., 1998; Cui and Alwan, 2006; Gerosa et al., 2009a).

Another source of greater variability in children's speech is that speech is an extremely complex motor activity. While speech motor control develops very rapidly across the first three or four years of life, many aspects are not fully adult-like until well into adolescence (see, e.g., literature reviewed in Smith, 2010; Green and Nip, 2010). For example, kinematic analyses of the

lips and jaw in children younger than six years show nothing like the consistent coordination between fluidly independent movements that characterize adult productions of labial stops and labiodental fricatives in variable vocalic and prosodic contexts (Green et al., 2000), and adult-like coordination between labial gestures and lingual gestures for rounded vowels is also slow to develop (see, e.g. Ménard et al., 2016). Moreover, because of the relationship between experience (or practice) and fluency, this protracted development of motor control leads to variability within individual children as well as between children. For example, there are systematic within-talker differences in speech articulator movement variability that are related to differences in sentence complexity, in word length, and even in phonotactic probability and lexicality (see, e.g. Sadagopan and Smith, 2008; Heisler and Goffman, 2016).

This means that there will be differences in the amount of variability across development as children's vocabularies continue to grow and as their phonologies continue to change to accommodate to new phoneme sequences and new prosodic structures. These differences in "fluency" should lead to differences in accuracy of ASR models across children of different ages and also across children of the same age who have different levels of experience or competence with the speaking style that is involved in the elicitation task. For example, if the speaking style is that of read speech, we might expect to see differences in recognition rates that are systematically related to differences in reading proficiency, as Li and Russell (2002) found when they related failures of an ASR model to teacher's ratings of pronunciation "goodness" in the Children's Primary School Reading (PSR) corpus.

More generally, we might expect to see variation between and within children in the degree to which pronunciations of target words or sounds deviate from community pronunciation norms, as noted, for example, by Benzeghiba et al. (2007) and Cincarek et al. (2007), and researchers such as Fringi et al. (2015) and Hämäläinen et al. (2014) have begun to explore the idea that recognition might be improved (i.e., be made more robustly human-like) if ASR systems for younger children are designed to recognize and accommodate to the most commonly noted patterns of deviation in normal speech development. Indeed, these patterns of deviation have become the object of recognition in their own right, in the application of ASR methods in developing tools for assessing and treating atypical speech development (see, e.g., Kewley-Port et al., 1991; Bunnell et al., 2000; Hamidi and Baljko, 2013; Sadeghian and Zahorian, 2015).

This application of ASR technology raises the question of how recordings

of children's speech should be annotated in order to gauge the accuracy of such tools and to generalize them to be useful more generally in research on normal speech development as well. Commonly noted patterns of deviation from adult community norms are typically described using the symbolic categories of the International Phonetic Alphabet. For example, the deviations that Bunnell et al. (2000) examined are typically described as substitutions of [w] for target /r/. Is phonetic transcription the most useful annotation tool for this purpose? The next section address that question.

5. Annotating speech from preschool children

Phonetic transcription has been used to record observations of children's vocalizations for more than a century, from the very earliest diary studies that were the basis for Jespersen's characterization of children's speech development (Jespersen, 1922). Indeed, in the days before there were devices such as the wearable voice-activated audio recorder that is used in the LENA system, on-the-fly phonetic transcription was the only efficient and reasonably unambiguous method available for recording continuous longitudinal day-by-day observations of child speech, and it has continued to be used for this purpose in diary studies to this day (see, e.g., Jaeger, 2005; Inkelas and Rose, 2007). On-the-fly phonetic transcription was also the only viable tool for recording observations of more systematically elicited word productions from sufficiently large samples of children in the earliest studies to establish norms for children's phonological development (e.g., Wellman et al., 1931). Moreover, it continues to be the primary analytic tool in developing and using instruments for clinical evaluation of phonological development such as the GFTA, so that even when audio recordings are made of children's responses in the course of developing new clinical assessment instruments, such as the one described in Dodd et al. (2003), the purpose is typically only to provide a way of doing multiple post-hoc phonetic transcriptions of the same utterances, as a way of assessing the consistency of observations within and between testers. Given this long history of relying on phonetic transcription as a way of recording observations of children's speech, it is perhaps not surprising that post-hoc phonetic transcription also continued to be used as an analytic tool even in many studies where the primary data were audio recordings, such as the studies in support of the Frame-Content model of canonical babble and early words (Davis and MacNeilage, 1995; MacNeilage et al., 1997), and that these phonetic transcriptions remain the

primary annotation schema when the audio recordings are then contributed to the PhonBank archive (Rose and MacWhinney, 2014).

Phonetic transcription is a very useful analytic tool in that it can be used to note deviations from canonical adult forms without any specialized instrumentation beyond the human ear and brain. However, it has limitations. The annotator must impute a particular segmental phonetic model and then parse continuous phonetic variation for each imputed segment into one of a finite set of consonant and vowel categories. In cases where temporal coordination among articulators yields an acoustic pattern that is consistent with the imputed segmental model and the phonetic variation within each category is low relative to the differences between categories, this task is straightforward. In cases where temporal coordination is inconsistent or the distribution of phonetic variation within the parametric articulatory and auditory spaces is not yet controlled in an adult-like way, however, phonetic transcription is considerably less informative.

There is substantial evidence that children's productions of sounds are highly variable, and that they often do not resemble canonical adult forms of the same target sounds. Some of this evidence is presented earlier: children's productions of sounds show greater trial-to-trial variability in temporal coordination and spectral parameter measures than do adults. This is true even for older school-aged children whose speech is transcribed to be correct. This section considers variation in very young children's productions of sounds, and how these might affect the quality of annotations. One frequently cited study of phonetic variation in children's speech is given by Macken and Barton (1980). Macken and Barton examined the development of stop consonant voicing, using instrumental measures of voice onset time (VOT). VOT is the primary acoustic cue to the distinction between target voiced and voiceless stop consonants in English and many other languages. In perception experiments, the influence of VOT on the identification of stop consonant voicing has been shown to be strongly categorical. Stops with VOT values that are intermediate between voiced and voiceless are typically labeled as either voiced or voiceless, and discrimination of sounds that are labeled identically is poor (Liberman et al., 1961). Given this fact, it was generally thought at the time that Macken and Barton conducted their study that listeners should be poor at tracking within-category variation in VOT.

Macken and Barton examined four children longitudinally, eliciting target words in play sessions recorded at intervals from the middle of the second year of life to early in the third year. There were considerable differences

across the recording sessions. Importantly, there were sessions where the child produced statistically significant differences in VOT between target voiced and voiceless consonants, but where the mean values were all in the short-lag range that adults would label with the voiced stop category. For example, participant "Jay" produced a significant difference between target voiced (M=7.5 ms) and voiceless (M=16.6 ms) stops in the recording made at 34 months. These values are likely to be labeled as voiced by adult listeners. A common term for this phenomenon is "covert contrast" (see, e.g., Scobbie et al., 2000). The productions showed a contrast in the sense that the voiced and voiceless targets were produced significantly differently. The contrast was covert in the sense that productions of both targets were in the range that previous research suggested were part of a single voicing category in adults; hence, it was predicted by researchers like Liberman et al. (1961) that adults would find difficulty encoding and denoting VOT differences within categories.

Subsequent research has documented covert contrast for other emerging sounds. For example, Li et al. (2009) found evidence of covert contrast between anterior and posterior lingual sibilant fricatives in 2-3 year old children acquiring English (/s/ vs. /ʃ/) and Japanese (/s/ vs. /ɕ/). There is a particularly robust literature showing that covert contrasts exist in the speech of children with SSD. Baum and McNutt (1990) found acoustic differences between target /θ/ and frontal variants of /s/ (which are often transcribed as [θ]) in children with residual misarticulations. Edwards et al. (1997) found that some children who were transcribed as producing [t] for /k/ nonetheless produced target /t/ and /k/ with systematically different-shaped stop burst spectra (as measured by the spectral mean and skew).

Covert contrasts provide a dilemma for how to best annotate child speech databases. Ideally, the annotation should be as rich as is possible, while maintaining a high level of reliability in the annotations. If there is a way to note the "outlier" productions that are the prime contributors to the statistical differentiation of sounds that are transcribed with the same phonetic symbol, then it should be used, as the studies cited above suggest that many children do go through such stages of covert contrast in the course of early speech development.

The very term covert contrast implies that outlier productions cannot be detected by ears alone. It invokes a strong interpretation of early Structuralist theories of the "psychological reality of phonemes" (Sapir, 1933) as supported by experiments showing the "categorical perception" of speech sounds

(Liberman et al., 1957)—i.e., results suggesting that speech perception is so constrained by native-language phoneme contrasts that adults cannot detect sub-phonemic phonetic differences. However, the result that an adult listener cannot perceive VOT differences between stops that fall within the same phoneme category for the native language is based on research using synthesized CV stimuli in a combination of an identification task with closed-set responses and a discrimination task with similarly categorical choices between just two (ABX) response types (e.g., Liberman et al., 1961) or among just three ("odd man out") response types (e.g., Abramson and Lisker, 1970), and it is possible that other types of stimuli and/or other perception tasks might allow listeners to detect covert contrast.

Indeed, there is evidence that listeners can note small phonetic differences between sounds in the same phoneme category when given finer-grained response alternatives. For example, Miller (1997) reviews experiments using continuous "category goodness" ratings instead of closed-set identification and discrimination, with results that show that adult listeners are sensitive to sub-phonemic variation in VOT values and that they parse this variation in terms of their experience with systematic variation, such as differences in VOT values that are associated with different overall speaking rates. Relatedly, Massaro and Cohen (1983) examined listeners' perception of three speech continua (/bæ/-/dæ/, /bæ/-/pæ/, and /i/-/ɪ/) using an 11-point rating scale (e.g., responses from 0 for the best /bæ/ to 10 for the best /dæ/). They found that listeners' mean ratings changed gradually and monotonically across the steps of the continuum. Moreover, the overall distribution of ratings for most of the listeners followed the predictions of a mathematical model in which categorization was continuous, rather than one in which it was categorical. This is consistent with more recent findings by Toscano et al. (2010) that show continuous perception is reflected in the strength of neurophysiological responses to speech continua.

What annotation options exist for noting sub-phonemic variation, such as the difference between a production of /d/ that might be transcribed as [d] because it has a canonically short-lag VOT value and a production of /t/ that might be transcribed as [d] because it has a VOT that is too short to be assimilated into the aspirated American English /t/ category? One suggestion is simply to give listeners the option of coding sounds as "fuzzy" exemplars (see, e.g., Stoel-Gammon, 2001)—e.g., as productions that are somehow intermediate between two contrasting phoneme categories. This suggestion was implemented in the annotations of consonant voicing and

consonant place of articulation in the Paidologos database (Edwards and Beckman, 2008a). For example, children's productions of English target /s/ could be transcribed as [s], [θ], [ʃ], or as a sound intermediate between [s] and one of the other sounds. These intermediate productions were coded as either intermediate but closer to [s] (noted as [s:θ] or [s:ʃ]) or as intermediate but closer to the other sound (noted as [θ:s] or [ʃ:s]). The annotators found the intermediate categories to be very useful in coding productions that were impossible to assimilate fully to one adult phoneme category.

A second possibility is that we can aggregate closed-set responses over a group of listeners, and use the degree of disagreement among listeners as a measure of the proximity of a production to a prototypical, adult-like example of that sound. The notion underlying this method is that the closer a token is to the best exemplar, the larger a proportion of listeners will rate it as a member of that category. This possibility was examined by McAllister Byun et al. (2016b), who asked untrained listeners to judge whether children's productions of /r/ were correct or incorrect. The proportion of listeners who judged a stimulus as correct was related to an acoustic measure of rhoticity, the difference between the second and third formant frequencies at the point of F3 minimum.

A similar finding is reported by Munson and Urberg Carlson (2016), who conducted a variety of experiments in which listeners provided complex judgements such as first labeling a sound as either the phoneme /s/ or the phoneme /ʃ/ and then rating it for how good an example of /s/ or /ʃ/ it was. Their results showed that the proportion of listeners who labeled a sound as /s/ was correlated with an acoustic measure of the distribution of energy in the interval of frication, the centroid frequency of the middle 40 ms of the frication interval, as illustrated in Fig. 1 for the 21 subjects in one of these experiments who were native speakers of American English. Schellinger et al. (2017) found a similar relationship in a reanalysis of responses in another of the experiments reported in Munson and Urberg Carlson (2016), where the /s/ phoneme response was pitted against /θ/ rather than against /ʃ/. Schellinger and colleagues also showed that this aggregated closed-set response measure validated the use of intermediate categories in transcription. Sounds that phonetically trained annotators had transcribed using the intermediate transcriptions were associated with less agreement across naïve listeners in whether the sound was /s/ or /θ/ than sounds that the experts had transcribed as [s] or [θ].

A final possibility is that listeners can provide continuous ratings of cat-

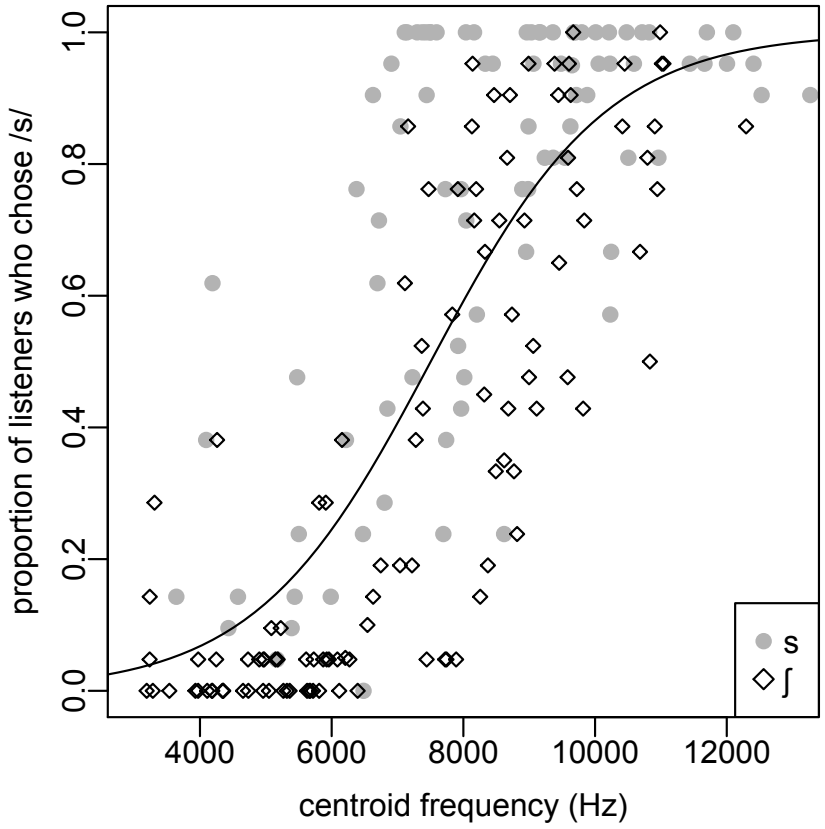

Figure 1: Proportion of 21 subjects (all American English native speakers) who listened to CV syllables extracted from children's productions of /s/-initial words such as *soccer* (plotted with filled circles) and /ʃ/-initial words such as *shower* (plotted with unfilled diamonds) and then labeled the target word-initial sibilant as /s/ rather than as /ʃ/, plotted as a function of the centroid frequency of a spectrum calculated over the middle 40 ms of the frication interval. The overlaid curve is from a mixed effects logistic regression with random (subject-level) intercepts.

egory goodness, when given the right listening task. The specific measure that we and others have examined is inspired by Massaro and Cohen's (1983) results, described above. Munson et al. (2010) examined listeners' continuous ratings of children's productions of the fricatives /s/ and /θ/ in word-initial position. Listeners were presented with consonant-vowel sequences excised from real words and nonwords elicited with a picture-prompted word-repetition task. Listeners were presented with a double-headed arrow (meant to evoke the continuous number line) anchored by the text "the 's' sound" at the left end and "the 'th' sound" at the right end. Listeners were instructed to click wherever on the line they perceived the sound to be relative to the ideal endpoints. These ratings were so fine-grained that they discriminated between correct and incorrect productions of both /s/ and /θ/. That is, the tokens of [θ] for target /θ/ were rated as more [θ]-like than were tokens of [θ] for target /s/, even though they were both transcribed with [θ]. The ratings also discriminated both of the intermediate categories [s:θ] and [θ:s] from one another, and from the other four transcription categories. This promising finding was replicated by Munson et al. (2012), and was extended to children's productions of the /t/-/k/ and /d/-/g/ contrasts. Strömbergsson et al. (2015) similarly found that ratings were useful in discriminating Swedish-speaking children's productions of correct, intermediate, and incorrect tokens of /t/ and /k/. Strömbergsson and colleagues found evidence of an additional utility of VAS measures: tokens of /t/ produced by children with SSD were rated as less prototypically /t/-like than were productions by children without SSD.

Continuous ratings of children's speech elicited with VAS are a potentially very useful tool for annotating child speech databases. For example, Munson et al. (2010) suggest that they might be especially suited to detecting a covert contrast when the specific acoustic parameter that is involved is not the primary acoustic cue to the contrast in the adult language. One example of this is given by Scobbie et al. (2000), who documented covert contrasts involving /st/, /t/, and /d/ in a single child with SSD whose errors demonstrated the common developmental patterns of "cluster reduction" (i.e. producing target clusters as singleton stops) and of voicing neutralization. Scobbie and colleagues found evidence of covert contrast in the voice quality of the vowel following the target stop or cluster, suggesting that the wide glottal-opening gesture associated with target /s/ was extending into the vowel to create an interval of breathy voice, and that the extended aspiration that is typical of singleton /t/ in English also was being realized as breathy voice. This was

found only after examining other seemingly more obvious parameters, such as VOT. Presumably, the covert contrasts between /st/ and /d/ and between /t/ and /d/ in this child's speech would be salient to listeners in a perception task.

Finer-grained perceptual measures also have the potential to be useful target annotations in the application of ASR methods to the evaluation of normal versus atypical speech development. For example, Bunnell et al. (2000) showed that a Likert-scale measure of the category goodness of a misarticulated token of /r/ was correlated with the log-likelihood of the ASR model classifying the token as an /r/ versus a /w/. Relatedly, finer-grained perceptual measures of children's productions have the potential to be useful as ways to evaluate different candidate acoustic measures for sounds that might have different developmental paths in different languages because of cross-language differences in community pronunciation norms for phonemes that might be transcribed using the same symbol (see, e.g., the literature reviewed in Edwards et al., 2015). For example, VAS has been used to measure children's production of lingual stop place contrasts in a number of languages with different community pronunciation norms for phonemes that English-speaking listeners might assimilate to English /t/ versus /k/, including Swedish (Strömbergsson et al., 2015), Greek (Arbisi-Kelm et al., 2010), and Japanese (Beckman et al., 2014).

Indeed, continuous perception measures may be especially useful for characterizing children's progressive mastery of this class of contrasts, which depend on a hidden articulatory target that is only audible earlier and/or later than the time when the target articulation is achieved. That is, for example, while the articulatory difference between /t/ and /k/ in prevocalic position might be characterized in terms of the location in the oral cavity of a coronal versus dorsal occlusion at the time of maximal contact with the teeth or palate, the acoustic evidence of this articulatory gesture is necessarily displaced in time, so that a child (or an adult second language learner) must infer the stop place from the rapidly changing spectrum from the release of the stop closure into the lingual posture for the following vowel. Languages can differ in exactly where the constriction is made and also in the part of the tongue tip or dorsum that makes contact. For example, whereas the /t/ of English is alveolar and typically apical, the /t/ of Swedish is dental and typically laminal. And the only ways to gauge from the audio signal whether the child's production is typical of the ambient language might be to analyze spectral characteristics after the lingual contact has been released and

to play this portion to listeners to judge, as in Stoel-Gammon et al. (1994).

Moreover languages differ in their patterns of temporal coordination among the consonant's occlusion gesture, the glottal opening gesture that makes it a /t/ or /k/ rather than a /d/ or /g/, and the vowel's lingual posturing gesture. And the child must also learn to produce the appropriate coordination patterns in order to sound like a fluent native speaker. While the development of ultrasound methods for observing the articulators themselves has provided another window into the development of this class of sounds (see, e.g., McAllister Byun et al., 2016a), there will always be many more hours of speech recordings available only as acoustic records. Therefore, the exploitation of child speech databases for studying speech development requires methods that can relate perceptual measures to acoustic measures of the ways in which children's articulations can deviate from community pronunciation norms. These measures should be child-centric and not presume the phonologies or the physiologies of adults. The next section describes a very small sample of some of the acoustic measures that we and others have been developing for two classes of lingual obstruent contrast that often are implicated in speech sound disorders.

6. Acoustic measures of preschool children's speech

In order to acquire adult-like speech motor control, children must meet two articulatory demands: one, differentiate gestures for contrastive phonemes, in order to signal distinctive meanings (e.g., *sip* vs. *ship*); two, coordinate the gestures for phones that compose an utterance, in order to speak fluently. Over the eight decades since the foundational work on acoustic correlates of distinctive features, which enabled the first commercial speech synthesis systems, spectral analysis of child speech has revealed that skills for gestural differentiation and coordination do not necessarily develop uniformly and monotonically in preschool children. Below, we review a narrow slice of the literature where instrumental analysis has been used to describe the acoustic and spectral properties of child speech. Specifically, the review focuses on sibilant fricatives and stop consonants because their production involves a complex coordination among asynchronous supralaryngeal filter shapes and both surpralaryngeal and supralaryngeal sources that makes them more challenging to model than the source-filter combinations of vowels. Unlike the ASR systems reviewed in Section 4, which aimed at learning a speaker-independent model that could be used to predictively annotate speech from

novel talkers, the instrumental analyses have generally sought to describe variation in acoustic and spectral properties of child speech with explicit reference to indexical properties of the talkers, such as their ages.

The spectra of the English sibilant fricatives /s/ and /ʃ/ are characterized by a high-frequency concentration of energy affiliated to the size of the cavity anterior to the constriction (Toda et al., 2010). The frequency-location of this energy concentration can be indexed by measuring either centroid frequency (center-of-gravity of the energy distribution across the frequency scale) or peak frequency (component with greatest amplitude), and many researchers have used such measures to describe how the ability to differentiate these sibilants develops continuously in preschool and school-aged children (see, e.g., Holliday et al., 2015; Nittrouer, 1993; Romeo et al., 2013). Li (2012) described the development of the /s/ vs. /ʃ/ contrast in children between 2 and 5 years of age by regressing centroid frequency values against age. The youngest children's productions were undifferentiated; as age increased, the fitted regression line for /s/ increased only slightly, while the one for /ʃ/ decreased significantly. These results imply, first, that the mapping between annotation categories (discussed in the previous section) and distributions within a given feature-space (e.g., centroid-frequency space) varies with the age of the talker; and, second, that the magnitude of this variation across age is not uniform across consonants.

Despite their theoretical contributions, the above-mentioned spectral analyses of child speech were limited in that spectral properties were measured from a single time point in the frication noise. While sibilant fricatives can be modeled synthetically as being produced by static articulatory postures, recent research indicates that the spectral properties of sibilant fricatives do vary over time (Iskarous et al., 2011; Yu, 2016) and that the spectral-kinematic properties of sibilant fricatives carry language- and consonant-specific acoustic information (Reidy, 2016). Preliminary analyses of cross-sectional age differences in how the /s/ vs. /ʃ/ contrast develops in preschool-aged children acquiring American English indicate development along both static and time-varying spectral properties (Reidy, 2015). Two-year-old children's productions of target /s/ and /ʃ/ did not differ in terms of peak frequency trajectory. Three-year-old children's productions differed in terms of the overall levels (i.e., static aspect) of the consonants' peak frequency trajectories, but not in terms of trajectory shapes (i.e., time-varying aspects). Four- and five-year-old children's productions differed in terms of both the overall level and the shapes of consonants' trajectories. The acoustic im-

plementation of this phonological contrast thus seems to develop initially in terms of spectral statics and then to be refined at later ages in terms of spectral kinematics.

While these preliminary analyses indicate that four- and five-year-old children approach the adult-like community norm of differentiating /s/ vs. /ʃ/ in terms of both spectral statics and spectral kinematics, they further revealed that the children's productions did not yet fully conform to the community norm as they exhibited greater excursions in the peak frequency trajectories of both sibilants. At present the articulatory causes of such differences in the spectral kinematics of children's productions, as compared with adults', remains poorly understood due to the lack of models that fully explain the adult data or that attempt to explain the children's data. Using a computational model of the adult vocal tract, Toda and Maeda (2006) simulated its transfer function in response to turbulence noise sources, while varying the size of the front cavity and the constriction. They then mapped how changes in front cavity size relate to changes in the resonant frequency of the simulated transfer function. Their simulations suggest that changes in front cavity size—which might arise from the upward movement of the jaw—perturb resonant frequency when the front cavity is relatively small. This analysis-by-synthesis suggests that the age-related differences in excursiveness of spectral kinematics may be related to the smaller vocal tract sizes in children; however, the ranges of front cavity sizes studied by Toda and Maeda were too large to be directly applicable to children. The absence of synthesis models applicable to preschool children's speech signals the need for the greater collection of articulatory data (see Katz and Bharadwaj, 2001; Zharkova et al., 2011) that can then be used to build such models.

By contrast to the relatively recent recognition that kinematic spectral measures may be necessary to capture children's gradual acquisition of adult-like control of sibilant fricative contrasts, the history of kinematic spectral measures for stops is quite old (see, e.g., Delattre et al., 1955; Fant, 1973; Kewley-Port, 1983). One very early kinematic measure that has recently been applied to articulatory as well as acoustic records for children's productions is the slope of locus equations.

Locus equations are fitted by regressing F2 frequency at CV boundary (first glottal pulse of the vowel) against F2 frequency at vowel midpoint, with F2 measured from multiple productions where the consonant remains fixed and vowel varies (e.g., /di, dɑ, du/). The slope of the locus-equation regression indexes the extent to which the tongue accommodates to the following

vowel during the production of the consonant, and thus indicates a consonant's extent of lingual coarticulation across vowel contexts, with greater slopes indicating greater extent of coarticulation (see Iskarous et al., 2010). In an early single-child study, Sussman et al. (1999) found that during the early preschool years, extent of lingual coarticulation does not necessarily develop monotonically toward an adult-like level of coarticulation. For [b]-initial syllables, locus-equation slopes increased steadily between 24 and 32 months, dropped sharply between 32 and 36 months, and then rebounded between 36 and 40 months. For [d]-initial syllables, the locus-equation slopes fell between 24 and 26 months, rebounded between 26 and 30 months, and finally decreased between 30 and 40 months. This longitudinal study suggests that coarticulatory ability does not necessarily improve monotonically toward community speech norms for all consonants. Furthermore, locus-equation slopes for [b] and [d] followed idiosyncratic trajectories out of phase with each other; hence, the exact character of the non-monotonic development in coarticulatory ability seems specific to each consonant.

While young preschool children's coarticulatory patterns for stops diverge from adult norms, several studies suggest that they eventually attain such norms during the later preschool years (Sussman et al., 1992). Similarly, Noiray et al. (2013) had adult and 4- and 5-year-old child speakers of Canadian French pronounce /VCV/ sequences, where the consonant was one of /p, t, k/ and the vowel was one of /i, a, u/. Time-aligned audio and ultrasound recordings were made of the productions, and the acoustic signal was used to define the consonant-vowel boundary (first glottal pulse) and the vowel midpoint. Two types of locus equations (acoustic and lingual) were fitted for each participant, for each consonant. Acoustic locus equations were fitted to F2, while lingual locus equations were fitted to the x-coordinate of the tongue body's highest point in the ultrasound image. The acoustic and lingual locus equations revealed similar patterns: slopes were significantly lower for /t/ compared with /p/ and /k/, and there were no differences between adults and children in terms of locus-equation slope for any consonant.

A possible criticism of the locus equation slope as a measure of the degree of consonant-vowel coarticulation is that, because it relies on values taken at just two time points, and because it is a statistical summary measure of the mean relationship between those two values, it can only describe a linear relationship between two acoustic (e.g., F2) or articulatory states (e.g., tongue-body height) that obtains in the aggregate of many syllable productions. Furthermore, because the acoustic-articulatory states are measured at

the consonant-vowel boundary and at the steady-state vowel midpoint, locus equations cannot capture potential developmental differences in how precisely oral and glottal gestures are temporally coordinated at the transition between the stop and the vowel. The most complete response to this criticism of locus equations would be to gather more aligned acoustic-articulatory data (similar to Noiray et al., 2013) and more densely sample both streams of data near the stop-vowel transition. Processing the data in this way would, furthermore, support the development of age-specific (or even child-specific) articulatory synthesis models, to be able to systematically explore the space of possible spatial-temporal coproduction patterns and how they map to the spectral kinematics during the transition from the stop burst into the vowel.

7. Summary and the road ahead

Our review of corpus elicitation methods in Sections 2 and 3 began with a description of methods that are currently being developed to exploit collections of day-long recordings made in homes and day-care centers, to enrich our understanding of infants' vocal development during the first year or two of life. The studies reviewed in VanDam and Silbert (2016) suggest that the rough diarization methods already are sufficient to pick out more than 85% of the infant's own utterances in the context of other talkers in the room (or on a television playing in the background), and Oller et al. (2010) proposed a fairly simple algorithm that can automatically tag these utterances using an annotation schema that is grounded in decades of research on phonetic models and acoustic characterizations of smaller longitudinal databases. The literature reviewed in Section 4 suggests that automatic ASR-based methods adapted from phonetic models of adult speech production might be adequate for annotating comparably large-scale databases of older school-age children's speech.

By contrast, methods for exploiting large-scale databases of preschool children's speech clearly will require that different annotation schemas be developed, in tandem with research to develop age-appropriate production models and associated acoustic measures, to better capture the protracted nature of phonetic development from the mostly not intelligible vocalizations of an 18-month-old to the mostly intelligible speech of an 8-year-old. Section 5 reviewed some ongoing research to develop such annotation schemas, and Section 6 described some research on developing acoustic measures for two classes of consonants that are challenging to model. (Both sibilants

and lingual stops are also often misarticulated in children with speech sound disorders.)

The measures described in Section 6 require that multiple tokens of the target sounds be available from each child in a study. For the databases used in the two studies by Reidy (i.e., Reidy, 2015, 2016), these multiple tokens were elicited using the picture-prompted word-repetition task described in Edwards and Beckman (2008a). This task is an efficient way to elicit a reasonably large sample of productions of a number of target sounds. Other researchers have used even more engaging methods for eliciting multiple tokens of target sounds. For example, Bunnell et al. (2000) used a computer game in which the game character names began with the target sounds /t/, /k/, /r/, and /w/.

It is possible that multiple tokens of particular target sounds can be captured in day-long recordings, too. Indeed, for sounds that occur in many words that young children know, such as /s/, /t/, and /k/ for English, our aim would be to eventually bootstrap from studies with more controlled databases, to be able to develop ASR-based methods to pick out words containing these sounds in day-long recordings to track the changes in an individual child's productions over development. However, the identification of relevant words in such samples cannot be automated yet, given the high word-error rates of ASR models for young children, and it seems likely that even with a human transcriber, there could be even higher rates of data loss than in the more controlled language samples that were analyzed in Lowenstein and Nittrouer (2008).

At the same time, programs such as Providence Talks[5] suggest that elicitation methods could be developed that take advantage of parents' eagerness to assess their children's language development at home. If elicitation methods such as the computer game in the Bunnell et al. (2000) study were combined with such already existing ASR-based resources, new protocols could be developed that could recruit parents to participate as fellow researchers, by playing games with their children to efficiently record multiple tokens of many more consonants than the four that Bunnell and colleagues targeted, and to do so regularly, to get dense longitudinal samples. If some of these parents were willing to bring their children into the lab to get articulatory records as well, then child-specific articulatory synthesis models could be de-

[5]http://www.providencetalks.org/about/

veloped to apply analysis-by-synthesis methods to consonant development that are comparable to those used in studies of vowel development such as Rvachew et al. (2006).

In summary, we probably already have methods that are adequate for eliciting the databases that we need for research in child speech development. What we need now are better schemas for annotations that do not impose adult segmental coordination models and adult phoneme categories onto children's productions. We also need more foundational research on how those annotations map to distributions in the spectral-temporal space, and we need more articulatory data to be able to build synthesis models that can capture the ways in which young children's vocal tracts and speech motor control constrain the relationship between their articulations and their acoustics.

Acknowledgments

Some of the research described in this paper was supported by NSF grant BCS0729277 to Benjamin Munson and by NIH grant DC02932 to Jan Edwards, who also deserves copious thanks for her extensive contributions to the ideas about the development and interpretation of elicitation and annotation protocols for children's speech that are described here.

References

Abramson, A.S., Lisker, L., 1970. Discriminability along the voicing continuum: Cross language tests, in: Proceedings of the 6th International Congress of Phonetic Sciences, pp. 569–573.

Alam, F., Stuart-Smith, J., 2014. Identity, ethnicity and fine phonetic detail: An acoustic phonetic analysis of syllable-initial /t/ in Glaswegian girls of Pakistani heritage, in: Hundt, M., Sharma, D. (Eds.), English in the Indian Diaspora. John Benjamins, Amsterdam, pp. 29–53.

Arbisi-Kelm, T., Edwards, J., Munson, B., Kong, E.J., 2010. Cross-linguistic perception of velar and alveolar obstruents: A perceptual and psychoacoustic study. Journal of the Acoustical Society of America 1957, 127. Poster presentation at the Acoustical Society of America.

Bang, H.Y., Clayards, M., Goad, H., 2015. A child-specific compensatory mechanism in the acquisition of English /s/, in: Proceedings of the 39th annual Boston University Conference on Language Development (BUCLD 39), Cascadilla Press, Somerville, MA. pp. 75–87.

Barbier, G., Boë, L.J., Captier, G., Laboissière, R., 2015. Human vocal tract growth: A longitudinal study of the development of various anatomical structures, in: Proceedings of INTERSPEECH 2015, pp. 364–368.

Batliner, A., Blomberg, M., D'Arcy, S., Elenius, D., Giuliani, D., Gerosa, M., Hacker, C., Russell, M., Steidl, S., Wong, M., 2005. The PF_STAR children's speech corpus, in: INTERSPEECH-2005, pp. 2761–2764.

Baum, S.R., McNutt, J.C., 1990. An acoustic analysis of frontal misarticulation of /s/ in children. Journal of Phonetics 18, 51–63.

Beckman, M.E., Edwards, J., 2010. Generalizing over lexicons to predict consonant mastery. Laboratory Phonology 1, 319–343.

Beckman, M.E., Munson, B., Edwards, J., 2014. Effects of speaker language and listener language on children's stop place. Poster presentation given at the Conference on Laboratory Phonology, Tachikawa, Japan.

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, 763–786. doi:10.1016/J.SPECOM.2007.02.006.

de Boysson-Bardies, B., Hallé, P., Sagart, L., Durand, C., 1989. A crosslinguistic investigation of vowel formants in babbling. Journal of Child Language 16, 1–17. doi:10.1017/S0305000900013404.

de Boysson-Bardies, B., Vihman, M.M., 1991. Adaptation to language: Evidence from babbling and first words in four languages. Language 67, 297–319. doi:10.1353/lan.1991.0045.

Buder, E.H., Warlaumont, A.S., Oller, D.K., 2013. An acoustic phonetic catalog of prespeech vocalizations from a developmental perspective, in: Peter, B., MacLeod, A.A.N. (Eds.), Comprehensive perspectives on child speech development and disorders: Pathways from linguistic theory to

clinical practice. Nova Science Publishers, Hauppauge, NY. chapter 4, pp. 103–134.

Bunnell, H.T., Yarrington, D.M., Polikoff, J.B., 2000. STAR: Articulation training for young children, in: ICSLP-2000, pp. 85–88.

Canault, M., Le Normand, M.T., Foudil, S., Loundon, N., Thai-Van, H., 2016. Reliability of the Language ENvironment Analysis system (LENA^{TM}) in European French. Behavior Research Methods 48, 1109–1124. doi:10.3758/s13428-015-0634-8.

Chen, N.F., Tong, R., Wee, D., Lee, P., Ma, B., Li, H., 2016. SingaKids-Mandarin: Speech corpus of Singaporean children speaking Mandarin Chinese, in: INTERSPEECH-2016, pp. 1545–1549. doi:10.21437/Interspeech.2016-139.

Chiba, T., Kajiyama, M., 1941. The vowel, its nature and structure. Kaiseikan, Tokyo.

Cincarek, T., Shindo, I., Toda, T., Saruwatari, H., Shikano, K., 2007. Development of preschool children subsystem for ASR and Q&A in a real-environment speech-oriented guidance task, in: INTERSPEECH-2007, pp. 1469–1472.

Claes, T., Dologlou, I., ten Bosch, L., Van Compernolle, D., 1998. A novel feature transformation for vocal tract length normalization in automatic speech recognition. IEEE Transactions on Speech and Audio Processing 6, 549–557. doi:10.1109/89.725321.

Claus, F., Rosales, H.G., Petrick, R., Hain, H., Hoffmann, R., 2013. A survey about databases of children's speech, in: INTERSPEECH-2013, pp. 2410–2414.

Coker, C.H., 1976. A model of articulatory dynamics and control. Proceedings of the IEEE 64, 452–460. doi:10.1109/PROC.1976.10154.

Crelin, E.S., 1987. The human vocal tract: Anatomy, function, development, and evolution. Vantage Press, New York.

Cui, X., Alwan, A., 2006. Adaptation of children's speech with limited data based on formant-like peak alignment. Computer Speech and Language 20, 400–419.

Davis, B.L., MacNeilage, P.F., 1995. The articulatory basis of babbling. Journal of Speech, Language, and Hearing Research 38, 1199–1211.

Delattre, P.C., Liberman, A.M., Cooper, F.S., 1955. Acoustic loci and transitional cues for consonants. Journal of the Acoustical Society of America 27, 769–773. doi:10.1121/1.1908024.

Demuth, K., Tremblay, A., 2008. Prosodically-conditioned variability in children's production of French determiners. Journal of Child Language 35, 99–127.

Dilley, L.C., Pitt, M.A., 2007. A study of regressive place assimilation in spontaneous speech and its implications for spoken word recognition. Journal of the Acoustical Society of America 122, 2340–2353. doi:10.1121/1.2772226.

Dodd, B., Holm, A., Zhu, H., Crosbie, S., 2003. Phonological development: A normative study of British English-speaking children. Clinical Linguistics and Phonetics 17, 617–643.

Drager, K.K., 2011. Sociophonetic variation and the lemma. Journal of Phonetics 39, 694–707.

Edwards, J., Beckman, M.E., 2008a. Methodological questions in studying consonant acquisition. Clinical Linguistics and Phonetics 22, 937–956.

Edwards, J., Beckman, M.E., 2008b. Some cross-linguistic evidence for modulation of implicational universals by language-specific frequency effects in phonological development. Language Learning and Development 4, 122–156.

Edwards, J., Gibbon, F., Fourakis, M., 1997. On discrete changes in the acquisition of the alveolar/velar stop consonant contrast. Language and Speech 40, 203–210.

Edwards, J.R., Beckman, M.E., Munson, B., 2015. Cross-language differences in acquisition, in: Redford, M.A. (Ed.), Handbook of Speech Production. Wiley-Blackwell, Chichester, UK. chapter 23, pp. 530–554.

Elenius, D., Blomberg, M., 2005. Adaptation and normalization experiments in speech recognition for 4 to 8 year old children, in: INTERSPEECH-2005, pp. 2749–2752.

Eskenazi, M., Mostow, J., Graff, D., 1997. The CMU Kids Corpus LDC97S63. Linguistic Data Consortium database.

Eskenazi, M.S., 1996. KIDS: A database of children's speech. Journal of the Acoustical Society of America 100, 2759. doi:10.1121/1.416340.

Faluda, J.B., 2000. Arizona Articulation Proficiency Scale-3. Western Psychological Services, Los Angeles, CA.

Fant, G., 1960. Acoustic theory of speech production, with calculations based on X-ray studies of Russian articulations. Mouton, The Hague.

Fant, G., 1973. Stops in CV-syllables, in: Fant, G. (Ed.), Speech Sounds and Features. MIT Press, Cambridge, MA, pp. 110–139.

Fitch, W.T., Giedd, J., 1999. Morphology and development of the human vocal tract: A study using magnetic resonance imaging. Journal of the Acoustical Society of America 106, 1511–1522. doi:10.1121/1.427148.

Flipsen, Jr., P., 1995. Speaker-listener familiarity: Parents as judges of delayed speech intelligibility. Journal of Communication Disorders 28, 3–19. doi:10.1016/0021-9924(94)00015-R.

Franklin, B., Warlaumont, A.S., Messinger, D., Bene, E., Iyer, S.N., Lee, C.C., Lambert, B., Oller, D.K., 2014. Effects of parental interaction on infant vocalization rate, variability and vocal type. Language Learning and Development 10, 279–296. doi:10.1080/1547544.2013.849176.

Fringi, E., Lehman, J.F., Russell, M., 2015. Evidence of phonological processes in automatic recognition of children's speech, in: INTERSPEECH-2015, pp. 1621–1624.

Gahl, S., Yao, Y., Johnson, K., 2012. Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. Journal of Memory and Language 66, 789–806. doi:10.1016/J.JML.2011.11.006.

Gerosa, M., Giuliani, D., Brugnara, F., 2007. Acoustic variability and automatic recognition of children's speech. Speech Communication 49, 847–860. doi:10.1016/j.specom.2007.01.002.

Gerosa, M., Giuliani, D., Brugnara, F., 2009a. Towards age-independent acoustic modeling. Speech Communication 51, 499–509. doi:10.1016/J.SPECOM.2009.01.006.

Gerosa, M., Giuliani, D., Narayanan, S., Potamianos, A., 2009b. A review of ASR technologies for children's speech, in: Proceedings of the 2nd Workshop on Child, Computer and Interaction (ACM), Association for Computing Machinery. pp. 1–8. doi:10.1145/1640377.1640384.

Giuliani, D., Gerosa, M., Brugnara, F., 2006. Improved automatic speech recognition through speaker normalization. Computer Speech and Language 20, 107–123.

Goffman, L., Gerken, L., Lucchesi, J., 2007. Relations between segmental and motor variability in prosodically complex nonword sequences. Journal of Speech, Language, and Hearing Research 50, 444–458. doi:10.1044/1092-4388(2007/031).

Goldman, R., Fristoe, M., 2015. Goldman-Fristoe Test of Articulation-3. Pearson, San Antonio: TX.

Goldstein, M.H., Schwade, J.A., 2008. Social feedback to infants' babbling facilitates rapid phonological learning. Psychological Science 19, 515–523. doi:10.1111/j.1467-9280.2008.02117.x.

Green, J.R., Moore, C.A., Higashikawa, M., Steeve, R.W., 2000. The physiologic development of speech motor control: Lip and jaw coordination. Journal of Speech, Language, and Hearing Research 43, 239–255.

Green, J.R., Nip, I.S.B., 2010. Some organization principles in early speech development, in: Maassen, B., van Lieshout, P. (Eds.), Speech motor control: New developments in basic and applied research. Oxford University Press. chapter 10, pp. 171–188.

Gros-Louis, J., West, M.J., Goldstein, M.H., King, A.P., 2006. Mothers provide differential feedback to infants' prelinguistic sounds. International Journal of Behavioral Development 30, 509–516. doi:10.1177/0165025406071914.

Hämäläinen, A., Candeias, S., Cho, H., Meinedo, H., Abad, A., Pellegrini, T., Tjalve, M., Trancoso, I., Dias, M.S., 2014. Correlating ASR errors

with developmental changes in speech production: A study of 3-10-year-old European Portuguese children's speech, in: Proceedings of the Fourth Workshop on Child, Computer and Interaction (WOCCI 2014), pp. 7–13.

Hämäläinen, A., Pinto, F.M., Rodrigues, S., Júdice, A., Silva, S.M., Calado, A., Dias, M.S., 2013. A multimodal educational game for 3-10-year-old children: Collecting and automatically recognising European Portuguese children's speech, in: Proceedings of the Workshop on Speech and Language Technology in Education (SLaTE 2013), Grenoble, France. pp. 31–36.

Hamidi, F., Baljko, M., 2013. Automatic speech recognition: A shifted role in early speech intervention?, in: Proceedings of the Fourth Workshop on Speech and Language Processing for Assistive Technologies, Association for Computational Linguistics, Grenoble, France. pp. 55–61. URL: http://www.aclweb.org/anthology/W13-3910.

Hart, B., Risley, T.R., 1995. Meaningful differences in the everyday experience of young American children. Paul H. Brookes Publishing Co., Baltimore, MD.

Heisler, L., Goffman, L., 2016. The influence of phonotactic probability and neighborhood density on children's production of newly learned words. Language Learning and Development 12, 338–356. doi:10.1080/15475441.2015.1117977.

Holliday, J.J., Reidy, P.F., Beckman, M.E., Edwards, J., 2015. Quantifying the robustness of the English sibilant fricative contrast in children. Journal of Speech, Language, and Hearing Research 58, 622–637. doi:10.1044/2015_JSLHR-S-14-0090.

Hsu, H.C., Fogel, A., Messinger, D.S., 2001. Infant non-distress vocalization during mother-infant face-to-face interaction: Factors associated with quantitative and qualitative differences. Infant Behavior and Development 24, 107–128.

Inkelas, S., Rose, Y., 2007. Positional neutralization: A case study from child language. Language 83, 707–736.

Ishizuka, K., Mugitani, R., Kato, H., Amano, S., 2007. Longitudinal developmental changes in spectral peaks of vowels produced by Japanese

infants. Journal of the Acoustical Society of America 121, 2272–2282. doi:10.1121/1.2535806.

Iskarous, K., Fowler, C.A., Whalen, D.H., 2010. Locus equations are an acoustic expression of articulator synergy. Journal of the Acoustical Society of America 128, 2021–2032. doi:10.1121/1.3479538.

Iskarous, K., Shadle, C.H., Proctor, M.I., 2011. Articulatory–acoustic kinematics: The production of American English /s/. Journal of the Acoustical Society of America 129, 944–954. doi:10.1121/1.3514537.

Jaeger, J.J., 2005. Kids' slips: What young children's slips of the tongue reveal about language development. Lawrence Erlbaum, Mahwah, NJ.

Jakobson, R., 1941. Kindersprache, Aphasie und allgemeine Lautgesetze. Almqvist and Wiksell, Uppsala. Translated by Allan R. Keiler as Child language, aphasia, and phonological universals, The Hague: Mouton, 1968.

Jakobson, R., Gunnar, F., Halle, M., 1951. Preliminaries to speech analysis: The distinctive features and their correlates. MIT Press, Cambridge, MA.

Jespersen, O., 1922. Language: Its nature, development, and origin. Henry Holt and Company, New York, NY.

Katz, W.F., Bharadwaj, S., 2001. Coarticulation in fricative-vowel syllables produced by children and adults: A preliminary report. Clinical Linguistics and Phonetics 15, 139–143.

Kewley-Port, D., 1983. Time-varying features as correlates of place of articulation in stop consonants. Journal of the Acoustical Society of America 73, 322–335. doi:10.1121/1.388813.

Kewley-Port, D., Watson, C.S., Elbert, M., Maki, D., Reed, D., 1991. The Indiana Speech Training Aid (ISTRA) II: Training curriculum and selected case studies. Clinical Linguistics and Phonetics 5, 13–38.

Klatt, D.H., 1987. Review of text-to-speech conversion for English. Journal of the Acoustical Society of America 82, 737–793.

Kleber, F., Peters, S., 2014. Children's imitation of coarticulatory patterns in different prosodic contexts. Paper presented at the 14th Conference on Laboratory Phonology.

Kohn, M.E., Farrington, C., 2012. Evaluating acoustic speaker normalization algorithms: Evidence from longitudinal child data. Journal of the Acoustical Society of America 131, 2237–2248.

Kong, E.J., Beckman, M.E., Edwards, J.R., 2012. Voice onset time is necessary but not always sufficient to describe acquisition of voiced stops: The cases of Greek and Japanese. Journal of Phonetics 40, 725–744.

Koopmans-van Beinum, F.J., Clement, C.J., van den Dikkenberg-Pot, I., 2001. Babbling and the lack of auditory speech perception: A matter of coordination? Developmental Science 4, 61–70.

Koopmans-van Beinum, F.J., van der Stelt, J.M., 1986. Early stages in the development of speech movements, in: Lindblom, B., Zetterström, R. (Eds.), Precursors of early speech. Springer, pp. 37–50.

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, 1455–1468.

Leonard, R.G., 1984. A database for speaker-independent digit recognition, in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-1984), IEEE. pp. 328–331. doi:10.1109/ICASSP.1984.1172716.

Levelt, C., 1994. On the Acquisition of Place. Ph.D. dissertation. Leiden University.

Li, F., 2012. Language-specific developmental differences in speech production: A cross-language acoustic study. Child Development 83, 1303–1315.

Li, F., Edwards, J., Beckman, M.E., 2009. Contrast and covert contrast: The phonetic development of voiceless sibilant fricatives in English and Japanese toddlers. Journal of Phonetics 37, 111–124.

Li, F., Rendall, D., Vasey, P.L., Kinsman, M., Ward-Sutherland, A., Diano, G., 2016. The development of sex/gender-specific /s/ and its relationship to gender identity in children and adolescents. Journal of Phonetics 57, 59–70.

Li, Q., Russell, M.J., 2002. An analysis of the causes of increased error rates in children's speech recognition, in: ICSLP-2002, pp. 2337–2340.

Liberman, A.M., Harris, K.S., Hoffman, H.S., Griffith, B.C., 1957. The discrimination of speech sounds within and across phoneme boundaries. Journal of Experimental Psychology 54, 358–368. doi:10.1037/H0044417.

Liberman, A.M., Harris, K.S., Kinney, J.A., Lane, H., 1961. The discrimination of relative onset-time of the components of certain speech and nonspeech patterns. Journal of Experimental Psychology 61, 379–388.

Lintfert, B., 2009. Phonetic and Phonological Development of Stress in German. Ph.D. dissertation. Universität Stuttgart.

Lowenstein, J.H., Nittrouer, S., 2008. Patterns of acquisition of native voice onset time in English-learning children. Journal of the Acoustical Society of America 124, 1180–1191.

Lynip, A.W., 1951. The use of magnetic devices in the collection and analysis of the preverbal utterances of an infant. Genetic Psychology Monographs 44, 221–262.

Macken, M.A., Barton, D., 1980. The acquisition of the voicing contrast in English: a study of voice onset time in word-initial stop consonants. Journal of Child Language 7, 41–74.

MacNeilage, P.F., Davis, B.L., Matyear, C.L., 1997. Babbling and first words: Phonetic similarities and differences. Speech Communication 22, 269–277. doi:10.1016/S0167-6393(97)00022-8.

Massaro, D.W., Cohen, M.M., 1983. Categorical or continuous speech perception: A new test. Speech Communication 2, 15–35.

McAllister Byun, T., Buchwald, A., Mizoguchi, A., 2016a. Covert contrast in velar fronting: An acoustic and ultrasound study. Clinical Linguistics and Phonetics 30, 249–276.

McAllister Byun, T., Harel, D., Halpin, P.F., Szeredi, D., 2016b. Deriving gradient measures of child speech from crowdsourced ratings. Journal of Communication Disorders 64, 91–102. URL: http://dx.doi.org/10.1016/j.jcomdis.2016.07.001.

McCarthy, D.A., 1929a. The Language Development of the Preschool Child. Institute of Child Welfare, University of Minnesota, Monograph No. 4, University of Minnesota Press.

McCarthy, D.A., 1929b. The vocalizations of infants. Psychological Bulletin 26, 625–651. doi:10.1037/H0072848.

McCune, L., Vihman, M.M., 2001. Early phonetic and lexical development: A productivity approach. Journal of Speech, Language, and Hearing Research 44, 670–684.

McGowan, R.S., Nittrouer, S., Manning, C.J., 2004. Development of [ɹ] in young, Midwestern, American children. Journal of the Acoustical Society of America 115, 871–884. doi:10.1121/1.1642624.

McGowan, R.W., McGowan, R.S., Denny, M., Nittrouer, S., 2014. A longitudinal study of very young children's vowel production. Journal of Speech, Language, and Hearing Research 57, 1–15.

McLeod, S., 2007. The International Guide to Speech Acquisition. Thomson Delmar Learning, Clifton Park, NY.

McLeod, S., Goldstein, B.A. (Eds.), 2012. Multilingual aspects of speech sound disorders in children. Multilingual Matters, Bristol.

Ménard, L., Perrier, P., Aubin, J., 2016. Compensation for a lip-tube perturbation in 4-year-olds: Articulatory, acoustic, and perceptual data analyzed in comparison with adults. Journal of the Acoustical Society of America 139, 2514–2531. doi:10.1121/1.4945718.

Miller, J.D., Lee, S., Uchanski, R.M., Heidbreder, A.F., Richman, B.B., Tadlock, J., 1996. Creation of two children's speech databases, in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-96), pp. 849–852. doi:10.1109/ICASSP.1996.543254.

Miller, J.L., 1997. Internal structure of phonetic categories. Language and Cognitive Processes 12, 865–869.

Mitra, V., Nam, H., Espy-Wilson, C.Y., Saltzman, E., Goldstein, L., 2012. Recognizing articulatory gestures from speech for robust speech recognition. Journal of the Acoustical Society of America 131, 2270–2287. doi:10.1121/1.3682038.

Mitra, V., Sivaraman, G., Nam, H., Espy-Wilson, C.Y., Saltzman, E., 2014. Articulatory features from deep neural networks and their role in speech recognition, in: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3017–3021. doi:10.1109/ICASSP.2014.6854154.

Moore, R.K., 2003. A comparison of the data requirements of automatic speech recognition systems and human listeners, in: EUROSPEECH-2003, pp. 2581–2584.

Munson, B., Edwards, J., Schellinger, S.K., Beckman, M.E., Meyer, M.K., 2010. Deconstructing phonetic transcription: Covert contrast, perceptual bias, and an extraterrestrial view of *Vox Humana*. Clinical Linguistics and Phonetics 24, 245–260.

Munson, B., Johnson, J., Edwards, J., 2012. The role of experience in the perception of phonetic detail in children's speech: A comparison of speech-language pathologists with clinically untrained listeners. American Journal of Speech-Language Pathology 21, 124–139.

Munson, B., Urberg Carlson, K., 2016. An exploration of methods for rating children's production of sibilant fricatives. Speech, Language, and Hearing 19, 36–45.

Murai, J., 1963. The sounds of infants : Their phonemization and symbolization. Studia Phonologica 3, 17–34. URL: `http://hdl.handle.net/2433/52620`.

Nam, H., Goldstein, L.M., Giulivi, S., Levitt, A.G., Whalen, D.H., 2013. Computational simulation of CV combination preferences in babbling. Journal of Phonetics 41, 63–77.

Nielsen, K., 2014. Phonetic imitation by young children and its developmental changes. Journal of Speech, Language, and Hearing Research 57, 2065–2075. doi:10.1044/2014_JSLHR-S-13-0093.

Nisimura, R., Nishihara, Y., Tsurumi, R., Lee, A., Saruwatari, H., Shikano, K., 2003. Takemaru-kun: Speech-oriented information system for real world research platform, in: Proceedings of the First lnternational Workshop on Language Understanding and Agents for Real World lnteraction, 2003, pp. 70–78.

Nittrouer, S., 1993. The emergence of mature gestural patterns is not uniform: Evidence from an acoustic study. Journal of Speech and Hearing Research 36, 959–972. doi:10.1044/jshr.3605.959.

Noiray, A., Ménard, L., Cathiard, M., Abry, C., Savariaux, C., 2004. The development of anticipatory labial coarticulation in French: a pionering study, in: Proceedings of the 8th International Conference on Spoken Language Processing (INTERSPEECH-2004), pp. 53–56.

Noiray, A., Ménard, L., Iskarous, K., 2013. The development of motor synergies in children: Ultrasound and acoustic measurements. Journal of the Acoustical Society of America 133, 444–452. doi:10.1121/1.4763983.

Oetting, J.B., Hartfield, L.R., Pruitt, S.L., 2009. Exploring LENA as a tool for researchers and clinicians. The ASHA Leader 14, 20–22. doi:10.1044/leader.FTR3.14062009.20.

Olive, J.P., 1977. Rule synthesis of speech from dyadic units, in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-77), pp. 568–570. doi:10.1109/ICASSP.1977.1170350.

Oller, D.K., 1980. The emergence of the sounds of speech in infancy, in: Yeni-Komshian, G., Kavanaugh, J.F., Ferguson, C.A. (Eds.), Child phonology, Vol. 1: Production. Academic Press, New York, pp. 93–112.

Oller, D.K., Niyogi, P., Gray, S., Richards, J.A., Gilkerson, J., Xu, D., Yapanel, U., Warren, S.F., 2010. Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. Proceedings of the National Academy of Sciences of the United States of America 107, 13354–13359.

Oller, D.K., Wieman, L.A., Doyle, W.J., Ross, C., 1976. Infant babbling and speech. Journal of Child Language 3, 1–11.

Osberger, M.J., Robbins, A.M., Todd, S.L., Riley, A.I., 1994. Speech intelligibility of children with cochlear implants. Volta Review 96, 169–180.

Ota, M., 2003. The development of prosodic structure in early words: Continuity, divergence and change. John Benjamins, Amsterdam.

Pae, S., Yoon, H., Seol, A., Gilkerson, J., Richards, J.A., Ma, L., Topping, K., 2016. Effects of feedback on parent–child language with infants and toddlers in Korea. First Language 36, 549–569. doi:10.1177/0142723716649273.

Potamianos, A., Narayanan, S., 1998. Spoken dialog systems for children, in: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-1998), IEEE. pp. 197–200. doi:10.1109/ICASSP.1998.674401.

Potamianos, A., Narayanan, S., 2003. Robust recognition of children's speech. IEEE Transactions on Speech and Audio Processing 11, 603–616. doi:10.1109/TSA.2003.818026.

Potamianos, A., Narayanan, S., Lee, S., 1997. Automatic speech recognition for children, in: EUROSPEECH-1997, pp. 2371–2374.

Reidy, P.F., 2015. The spectral dynamics of voiceless sibilant fricatives in English and Japanese. Ph.D. thesis. The Ohio State University.

Reidy, P.F., 2016. Spectral dynamics of sibilant fricatives are contrastive and language specific. Journal of the Acoustical Society of America 140, 2518–2529. doi:10.1121/1.4964510.

Renwick, M.E., Vasilescu, I., Dutrey, C., Lamel, L., Vieru, B., 2016. Marginal contrast among Romanian vowels: Evidence from ASR and functional load, in: INTERSPEECH-2016, pp. 2433–2437. doi:10.21437/Interspeech.2016-762.

Romeo, R., Hazan, V., Pettinato, M., 2013. Developmental and gender-related trends of intra-talker variability in consonant production. Journal of the Acoustical Society of America 134, 3781–3792. doi:10.1121/1.4824160.

Rose, Y., MacWhinney, B., 2014. The PhonBank Project: Data and software-assisted methods for the study of phonology and phonological development, in: Durand, J., Gut, U., Kristoffersen, G. (Eds.), The Oxford Handbook of Corpus Phonology. Oxford University Press, Oxford, pp. 380–401.

Rosenbeck, P., Baungaard, B., Jacobsen, C., Barry, D., 1994. The design and efficient recording of a 3000 speaker Scandinavian telephone speech database: Rafael.0, in: Proceedings of the Third International Conference on Spoken Language Processing (ICSLP-1994), pp. 1807–1810.

Rvachew, S., Grawburg, M., 2006. Correlates of phonological awareness in preschoolers with speech sound disorders. Journal of Speech, Language, and Hearing Research 49, 74–87. doi:10.1044/1092-4388(2006/006).

Rvachew, S., Mattock, K., Polka, L., Ménard, L., 2006. Developmental and cross-linguistic variation in the infant vowel space: The case of Canadian English and Canadian French. Journal of the Acoustical Society of America 120, 2250–2259.

Sadagopan, N., Smith, A., 2008. Developmental changes in the effects of utterance length and complexity on speech movement variability. Journal of Speech, Language, and Hearing Research 51, 1138–1151. doi:10.1044/1092-4388(2008/06-0222).

Sadeghian, R., Zahorian, S.A., 2015. Towards an automated screening tool for pediatric speech delay, in: INTERSPEECH-2015, pp. 1650–1654.

Sapir, E., 1933. La réalité psychologuique des phonèmes. Journal de Psychologie Normale et Pathologique 30, 247–265.

Sasaki, C.T., Levine, P.A., Laitman, J.T., Crelin, E.S., 1977. Postnasal descent of the epiglottis in man: A preliminary report. Archives of Otolaryngology – Head & Neck Surgery 103, 169–171.

Schellinger, S.K., Munson, B., Edwards, J., 2017. Gradient perception of children's productions of /s/ and /θ/: A comparative study of rating methods. Clinical Linguistics and Phonetics 31, 80–103. doi:10.1080/02699206.2016.1205665.

Scobbie, J., Gibbon, F., Hardcastle, W.J., Fletcher, P., 2000. Covert contrasts as a stage in the acquisition of phonetics and phonology, in: Broe, M., Pierrehumbert, J. (Eds.), Papers in laboratory phonology V: Language acquisition and the lexicon. Cambridge University Press, Cambridge, U. K., pp. 194–207.

Serkhane, J.E., Schwartz, J.L., Boë, L.J., Davis, B.L., Matyear, C.L., 2007. Infants' vocalizations analyzed with an articulatory model: A preliminary report. Journal of Phonetics 35, 321–340. doi:10.1016/J.WOCN.2006.10.002.

Sheinkopf, S.J., Mundy, P., Oller, D.K., Steffens, M., 2000. Vocal atypicalities of preverbal autistic children. Journal of Autism and Developmental Disorders 30, 345–354.

Shirley, M.M., 1931. The first two years: A study of twenty-five babies. volume II, Intellectual Development. The University of Minnesota Press, Minneapolis, MN.

Shivakumar, P.G., Potamianos, A., Lee, S., Narayanan, S., 2014. Improving speech recognition for children using acoustic adaptation and pronunciation modeling, in: Fourth Workshop on Child Computer Interaction (WOCCI 2014), pp. 15–19.

Shobaki, K., Hosom, J.P., Cole, R., 2007. CSLU: Kid's Speech Version 1.1. Linguistic Data Consortium database. URL: `https://catalog.ldc.upenn.edu/LDC2007S18`.

Shobaki, K., Hosom, J.P., Cole, R.A., 2000. The OGI Kids' Speech Corpus and recognizers, in: Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP-2000), pp. 258–261.

Shriberg, L.D., Kwiatkowski, J., 1982. Phonological disorders III: A procedure for assessing severity of involvement. Journal of Speech and Hearing Disorders 47, 256–270. doi:10.1044/jshd.4703.256.

Shriberg, L.D., Kwiatkowski, J., 1985. Continuous speech sampling for phonologic analyses of speech-delayed children. Journal of Speech and Hearing Disorders 50, 323–334. doi:10.1044/jshd.5004.323.

Smit, A.B., Hand, L., Freilinger, J.J., Bernthal, J.E., Bird, A., 1990. The Iowa Articulation Norms project and its Nebraska replication. Journal of Speech and Hearing Disorders 55, 779–798.

Smith, A., 2010. Development of neural control of orofacial movements for speech, in: Hardcastle, W.J., Laver, J., Gibbon, F.E. (Eds.), Handbook of Phonetic Sciences. 2nd ed.. Wiley-Blackwell. chapter 7, pp. 251–296.

Stark, R.E., 1980. Stages of speech development in the first year of life, in: Yeni-Komshian, G., Kavanaugh, J.F., Ferguson, C.A. (Eds.), Child phonology, Vol. 1: Production. Academic Press, New York, pp. 73–92.

Stemmer, G., Hacker, C., Steidl, S., Nöth, E., 2003. Acoustic normalization of children's speech, in: EUROSPEECH-2003, pp. 1313–1316.

Stoel-Gammon, C., 2001. Transcribing the speech of young children. Topics in Language Disorders 21, 12–21.

Stoel-Gammon, C., Williams, K., Buder, E., 1994. Cross-language differences in phonological acquisition: Swedish and American /t/. Phonetica 51, 146–158. doi:10.1159/000261966.

Strömbergsson, S., Salvi, G., House, D., 2015. Acoustic and perceptual evaluation of category goodness of /t/ and /k/ in typical and misarticulated children's speech. Journal of the Acoustical Society of America 137, 3422–3435.

Suskind, D.L., Leffel, K.R., Graf, E., Hernandez, M.W., Gunderson, E.A., Sapolich, S.G., Suskind, E., Leininger, L., Goldin-Meadow, S., Levine, S.C., 2016. A parent-directed language intervention for children of low socioeconomic status: A randomized controlled pilot study. Journal of Child Language 43, 366–406.

Sussman, H.M., Duder, C., Dalston, E., Cacciatore, A., 1999. An acoustic analysis of the development of CV coarticulation: A case study. Journal of Speech, Language, and Hearing Research 42, 1080–1096. doi:10.1044/jslhr.4205.1080.

Sussman, H.M., Hoemeke, K.A., McCaffrey, H.A., 1992. Locus equations as an index of coarticulation for place of articulation distinctions in children. Journal of Speech, Language, and Hearing Research 35, 769–781. doi:10.1044/jshr.3504.769.

Templin, M.C., 1957. Certain language skills in children, their development and interrelationships. University of Minnesota Press, Minneapolis, MN.

Toda, M., Maeda, S., 2006. Quantal aspects of non-anterior sibilant fricatives: A simulation study, in: Proceedings of the 7^{th} International Seminar on Speech Production, pp. 573–580.

Toda, M., Maeda, S., Honda, K., 2010. Formant-cavity affiliation in sibilant fricatives, in: Fuchs, S., Toda, M., Żygis, M. (Eds.), Turbulent Sounds: An Interdisciplinary Guide. De Gruyter Mouton, Berlin, Germany, pp. 343–374.

Toscano, J., McMurray, B., Dennhardt, J., Luck, S., 2010. Continuous perception and graded categorization: Elecrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech. Psychological Science 21, 1532–1540.

Umeda, N., 1975. Vowel duration in American English. Journal of the Acoustical Society of America 58, 434–445.

Umeda, N., 1976. Linguistic rules for text-to-speech synthesis. Proceedings of the IEEE 64, 443–451. doi:10.1109/PROC.1976.10153.

VanDam, M., Oller, D.K., Ambrose, S.E., Gray, S., Richards, J.A., Xu, D., Gilkerson, J., Silbert, N.H., Moeller, M.P., 2015. Automated vocal analysis of children with hearing loss and their typical and atypical peers. Ear and Hearing 36, e146–e152. doi:10.1097/AUD.0000000000000138.

VanDam, M., Silbert, N.H., 2016. Fidelity of automatic speech processing for adult and child talker classifications. PLoS ONE 11, e0160588. doi:10.1371/journal.pone.0160588.

VanDam, M., Warlaumont, A.S., Bergelson, E., Cristia, A., Soderstrom, M., De Palma, P., MacWhinney, B., 2016. Homebank: An online repository of daylong child-centered audio recordings. Seminars in Speech and Language 37, 128–142. doi:10.1055/s-0036-1580745.

Vihman, M.M., 1993. Variable paths to early word production. Journal of Phonetics 21, 61–82.

Vihman, M.M., Macken, M.A., Miller, R., Simmons, H., Miller, J., 1985. From babbling to speech: A re-assessment of the continuity issue. Language 61, 397–445.

Vorperian, H.K., Wang, S., Chung, M.K., Schimek, E.M., Durtschi, R.B., Kent, R.D., Ziegert, A.J., Gentry, L.R., 2009. Anatomic development of the oral and pharyngeal portions of the vocal tract: An imaging

study. Journal of the Acoustical Society of America 125, 1666–1678. doi:10.1121/1.3075589.

Warlaumont, A.S., Oller, D.K., Buder, E.H., Dale, R., Kozma, R., 2010. Data-driven automated acoustic analysis of human infant vocalizations using neural network tools. Journal of the Acoustical Society of America 127, 2563–2577. doi:10.1121/1.3327460.

Warlaumont, A.S., Ramsdell-Hudock, H.L., 2016. Detection of total syllables and canonical syllables in infant vocalizations, in: Proceedings of INTERSPEECH 2016, pp. 2676–2680.

Weisleder, A., Fernald, A., 2013. Talking to children matters: Early language experience strengthens processing and builds vocabulary. Psychological Science 24, 2143–2152. doi:10.1177/0956797613488145.

Weist, R.M., Kruppe, B., 1977. Parent and sibling comprehension of children's speech. Journal of Psycholinguistic Research 6, 49–58. doi:10.1007/BF01069574.

Wellman, B.L., Case, I.M., Mengert, I.G., Bradbury, D.E., 1931. Speech sounds of young children. University of Iowa Studies: Child Welfare 5.

White, S.D., 2001. Covert contrast, merger, and substitution in children's productions of /k/ and /t/. MA thesis. Ohio State University. Columbus, OH.

Wightman, C.W., Talkin, D.T., 1997. The Aligner: Text-to-speech alignment using Markov models, in: van Santen, J.P.H., Sproat, R.W., Olive, J.P., Hirschberg, J. (Eds.), Progress in Speech Synthsis. Springer-Verlag, New York, NY. chapter 25, pp. 313–324.

Wilcox, K., Morris, S., 1999. Children's Speech Intelligibility Measure. The Psychological Corporation, San Antonio: TX.

Wilpon, J.G., Jacobsen, C.N., 1996. A study of speech recognition for children and the elderly, in: ICASSP-96, pp. 349–352. doi:10.1109/ICASSP.1996.541104.

Yu, A.C.L., 2016. Vowel-dependent variation in Cantonese /s/ from an individual-difference perspective. Journal of the Acoustical Society of America 139, 1672–1690. doi:10.1121/1.4944992.

Zharkova, N., Hewlett, N., Hardcastle, W.J., 2011. Coarticulation as an indicator of speech motor control development in children: An ultrasound study. Motor Control 15, 118–140.

Zharkova, N., Hewlett, N., Hardcastle, W.J., 2012. An ultrasound study of lingual coarticulation in /sV/ syllables produced by adults and typically developing children. Journal of the International Phonetic Association 42, 193–208. doi:10.1017/S0025100312000060.