Enriched technology-enabled annotation and analyses of child speech

Mary E. Beckman
Department of Linguistics, Ohio State University, Columbus, OH, USA
mbeckman@ling.osu.edu

Abstract
This paper reviews a range of studies illustrating the ways in which speech technology has enabled richer analyses of corpora of young children’s speech and infants’ speech-like vocalizations. One set of studies illustrates the use of speech synthesis technology, such as VLAM, an articulatory synthesis system that models the transfer functions of child- or infant-proportioned vocal tracts. VLAM has been used to evaluate cross-language differences in the emergence of contrasting vowel categories. Another set of studies illustrates the use of modern corpus development and annotation tools to create and analyze the paidologos corpus, a database of utterances elicited in a picture-prompted word-repetition task from 2- through 5-year-old child speakers of a variety of languages. Flexible, incremental annotation on multiple tiers allows researchers to extract target sounds for spectral analysis as well as for calculating transcribed accuracy rates. Tag sets also can be used to extract stimuli for perception experiments, yielding naive-listener responses that can become another layer of tags.

Index Terms: child speech corpora, articulatory synthesis, perceptual evaluation of child productions

1. Introduction

Corpus studies have long played an important role in advancing our understanding of first language acquisition. For example, in the early part of the 20th century, Otto Jespersen and Roman Jakobson, among others, portrayed children’s phonological development as an orderly sequence of stages, which begins when the rich inventory of sounds randomly produced in babbling abruptly gives way to a very reduced phonetic inventory in first words, followed by a re-acquisition and reorganization of consonant contrasts that proceeds in “a strict and invariable temporal sequence” [1]. This characterization of phonological development as invariable across children and fundamentally different from the richly variable productions of babbling was an extrapolation from the few published diary studies, in which the researcher (typically the parent) recorded typical pronunciations of words and word-like forms observed in the course of daily interaction with a child. It was the dominant view in linguistics until the commercialization of good-quality magnetic analog tape recorders enabled research using more controlled phonetic transcription and even acoustic analysis of babbling and early word productions in corpora that included more children (e.g., [2],[3]).

This paper reviews a range of studies illustrating the ways in which more recent developments in speech technology have enabled even richer analyses of corpora of children’s speech and younger infants’ non-speech vocalizations. One set of studies uses speech synthesis programs such as VLAM [4], an articulatory synthesis program based on Maeda’s Vtcalcs [5] that can model the transfer functions of infant- and child-proportioned vocal tracts. Another set of studies illustrates the application of resources that were developed in building speech databases for training speech synthesis and automatic speech recognition systems in designing and annotating the paidologos cross-language corpus of words and nonwords elicited from 2- through 5-year-old children [6]. Sections 2 and 3 review these two ways of applying speech technology to the analysis of child speech. Each section also cites several studies that use speech synthesis or the tagged corpus to create controlled stimulus sets for perception experiments.

2. Analysis by speech synthesis

The general method of analysis by speech synthesis has been used to investigate phonological representations and processes in adults for many decades. Classic examples include Cohen and ’t Hart’s use of close-copy stylization of the fundamental frequency (F0) contour to develop their model of the Dutch intonation system [7], and Stevens and House’s application of a simple tongue-arc model [8] to explore the features of the American English vowel space as estimated for adult male talkers by the formant values measured in the Peterson and Barney corpus [9].

In the decades since these classic studies, the revolution in computer hardware that gave the world small, fast, powerful computers also led to the eventual incorporation of more recent techniques such as LPC analysis-resynthesis [10] and PSOLA [11] into free signal-analysis programs, such as Praat [12], that would have been unimaginable even as recently as the 1990s. This development in turn has made the analysis-by-synthesis method of close-copy stylization available to anyone with access to a personal computer, so that any researcher studying the acquisition of intonation (e.g., [13]) or of lexical tone (e.g., [14]) can apply the method to estimate pitch values in young children’s productions even for intervals where the interaction of glottal flow and vocal tract resonances is highly nonlinear (cf. [15]), making F0 estimation as well as formant value estimation difficult.

At an even earlier period of this revolution in computer hardware, most phonetics laboratories and speech laboratories acquired enough computational power to be able to apply statistical techniques such as factor analysis to articulatory data such as cross-sectional distances measured at a vector of points along the mid-sagittal tongue surface, making it possible to build more sophisticated articulatory synthesis models than the simple tongue-arc model of Stevens and House – see, e.g., Harshman et al.’s model [16], as well as Maeda’s model [5] mentioned above (5, [17]-[19]). Among these, Maeda’s Vtcalcs model is especially noteworthy, both for the size of the database on which it was based (more than 1000 frames of simultaneous cineradiographic and labiofilm recordings from two adult female speakers of French) and for the way in which the code has been distributed freely [20] to allow its use as a research tool by other researchers (e.g., [21]) and also as a tool for teaching phonetics and speech science (e.g., [22]).

Maeda and his colleagues have incorporated Vtcalcs into several different programs for estimating transfer functions of infant- or child-proportioned vocal tracts, the earliest being the
Boë and Maeda’s Variable Linear Articulatory growth Model (VLAM) mentioned above [4]. VLAM squeezes (or stretches) the physical scale corresponding to the model parameters that specify the adult female mid-sagittal tongue surface and jaw position, so that the tongue can fit into the dimensions of a scaled-down (or scaled-up) model of the various fixed vocal tract structures that determine the lengths of the pharyngeal and oral cavities, with the sizes of the relevant fixed structures estimated using Goldstein’s [23] model of vocal tract growth from birth to 21 years of age, which was based on her survey of extant measurements reported in the literature from the fields of dentistry, anatomy, and physical anthropology. (A more recent model developed by Callan et al. [24] estimates the sizes of the relevant fixed vocal tract structures instead from measurements of several of the first images of infants and children to be included in the Vorperian et al. database of medical MRIs [25].)

Boë, Ménard, and their colleagues have used VLAM to synthesize both maximal vowel spaces (MVS) and a set of “vowel prototype” stimuli for perception experiments ([26]-[32]), in studies designed to explore how vocal tract anatomy constrains the vowel qualities that infants and children can produce at different ages, as the vocal tract is restructured from the shape that enables safe sucking in early infancy to the adult female or male shape. Together these simulations and perceptual studies provide strong support for a picture of vowel acquisition in which vowel-like vocalizations are at first focused in the mid front or the low central region of the vowel space, with gradual expansion along both the first and second formant dimensions until the child has achieved control of the full range of qualities needed to contrast the vowel phonemes of the ambient language.

One of the simulation studies ([28]) also is an explicit cross-language study that uses two different synthesized MVS to interpret formant patterns in a cross-sectional corpus of recordings of 10- to 19-month-old toddlers who were acquiring either Canadian French or Canadian English as their first language. The two groups of toddlers differed both in the typical formant values for the youngest participants and in the dimension that was observed to expand more as a function of the toddler’s age. Specifically, the youngest English-learning toddlers had a higher (more front) mean F2 value than did the youngest French-learning toddlers, in keeping with the values that de Boysson-Bardies and colleagues reported in their earlier cross-linguistic study of vowel formants in babbling of 10-month-old infants [3]. The primary developmental trends were that the MVS expanded in the F2 dimension for the older English-learning toddlers, as they mastered the less frequent high and mid back vowels of their language. By contrast, the MVS expanded (lowered) in the F1 dimension for the older French-learning toddlers, as they gained control of the more densely populated high vowel region of French, and the contrast among [y] and a more peripheral target for both [i] and [u] by comparison to the analogous high vowel of English.

Two of the most recent perception studies ([31]-[32]) also are explicit cross-language comparisons, with results showing that adult speakers of different languages interpret the synthesized vowel qualities differently, in ways that suggest the influence of culture-specific processes of talker-size normalization as well as the influence of language-specific representations for even “prototypical” focal vowel categories.

All together, these studies illustrate how speech synthesis technology can be applied effectively to enable a much richer picture of the role of ambient language input even at the very earliest stages of phonological acquisition.

3. Developing the paidologos corpus

Another way in which speech technology has enabled a richer picture of ambient language input is by providing ancillary resources for developing controlled cross-language databases of children’s speech. This section illustrates this point by describing the development of the paidologos corpus [6].

The corpus is a set of audio recordings of words and nonwords that are elicited in a picture-prompted word-repetition task from 2- through 5-year-old child speakers. We designed this task so as to be able to elicit productions both of highly familiar early-acquired words and of less familiar words and even nonsense words, using prompts that are consistent across all types of words and across all ages. We wrote a tcl/tk script that loads a suitably randomized stimulus list and then presents, for each stimulus, a picture prompt (shown on the monitor of a laptop computer) and an associated audio prompt (an audio file recording of a woman’s voice saying the target word in a child-directed voice, that is played over loudspeakers plugged into the laptop’s audio output port).

To the left of the picture prompt, there is a picture of a duck (or frog or koala), which moves up a ladder when the program is advanced to the next stimulus, to give the child feedback about the progress of the experiment.

Our primary original aim in developing this picture-prompted word-repetition method was to be able to compare transcribed production accuracy for word-initial pre-vocalic lingual consonants that occur in several languages, but that differ either in phoneme frequency or phonotactic probability between pairs of target languages (see [33],[34]). For example, the consonant [ts] occurs in both Cantonese and Greek, but where it occurs in many words of Cantonese, it has a much lower phoneme frequency in Greek. The accuracy rates for [ts] in young Cantonese and Greek children’s productions differ in a way that reflects the impact of the cross-language difference in phoneme frequency. Also, the consonant-vowel sequence [ti] occurs in both English and Japanese, but where this sequence begins many words of English, it occurs in only a handful of words of Japanese. And the accuracy rates for [t] before [i] in young English- and Japanese-speaking children’s productions differ in a way that reflects the cross-language difference in phonotactic probability.

The first languages we recorded using this paradigm were Hong Kong Cantonese, U. S. English, a northern variety of Greek, and standard (Tokyo) Japanese (see [6], [33], [34]). Not long after, graduate students working with us expanded the corpus to include recordings of a northeastern variety of Mandarin Chinese [35] and of standard (Seoul) Korean [36]. More recently, other graduate student collaborators have adapted the paidologos paradigm to record bilingual speakers of Drehu and French [37] and of the Taiwan varieties of Southern Min and Mandarin Chinese [38].

For many of the languages that we have recorded we used pronunciation dictionaries developed for speech and language technology in order to determine the frequencies of different target consonants and consonant-vowel sequences. We also used these resources to be able to devise nonwords that controlled for phoneme frequency and phonotactic probability in the “frame” portion – i.e., the portion after the form-initial
consonant-vowel sequence that was the target of our investigation.

To be able to calculate transcribed phoneme accuracy, we developed an annotation protocol and annotation scripts that used the tool and scripting language of Praat to automate as much of the process as possible. In developing the protocol, we benefited from from the by-now long history of annotating corpora for speech synthesis and automatic speech recognition (e.g., [39]), to design the annotation to proceed on as many different annotation tiers as we needed for different analysis purposes, with the annotation scripts linking time points that should be linked. The tags we developed allow researchers to easily extract target sounds or shorter intervals for spectral analysis (e.g., [40]-[42]) as well as for calculating and comparing transcribed accuracy rates.

The technology for tagging child productions and for extracting target sounds for spectral analysis also can be applied to develop more sensitive measures of phonological contrast (e.g., [40],[42]). And the tag sets also can be used to extract stimuli for perception experiments, yielding naïve-listener responses of various types, including continuous category-goodness ratings, which can act as an added layer of tags to augment the much more time-consuming symbolic tags that are provided by phonetically-trained transcribers (e.g., [41], [42], [44], [45]). Moreover, speech language pathologists also can use these more continuous response types, to evaluate the role of clinical experience in training sensitivity to relevant sub-phonemic variation [47] and to develop more sensitive diagnostics of phonological delay or of incremental progress in response to therapy for speech sound disorders [47].

In short, as with the application of speech synthesis technology, the application of auxiliary tools and resources from speech technology has enabled much richer annotations and analyses of child speech corpora. In both cases, too, the applications yield to enhanced tools for educating students of speech, including speech pathologists as well as phoneticians and engineers.

4. Acknowledgements

Work on this paper was supported by NIDCD grant 02932.

5. References


[27] Rvachev, S., Mattock, K., Polka, L., and Ménard, L., “Developmental and cross-linguistic variation in the infant vowel space: The case of Canadian English and Canadian


