**Making Sense of Children's Speech: Man, Machine, and Moms**

Madeleine Yu, Amrita Bagga & Elizabeth K. Johnson
University of Toronto Mississauga
madeleine.yu@mail.utoronto.ca

Toddler's speech is difficult to understand — a word like *spaghetti* can be produced as *getti, shark* as *sock*, and *baby* as *baba*. These pronunciation variants represent known patterns in how child productions systematically deviate from adult-like forms; clusters are simplified, easier-to-produce sounds are substituted for more difficult sounds, and stressed syllables are reduplicated (e.g., Vihman, 1993). How do these transformations impact human listeners' ability to comprehend children? Does experience with young children affect comprehension? And how well do automatic speech recognition (ASR) systems, primarily designed to process adult speech, handle children's speech? Here, we examine who best understands children's speech and the types of errors that are made, in hopes of generating ideas for improving ASR.

Stimuli for the current study were drawn from a database containing longitudinally recorded utterances of 32 words produced by a set of typically developing children at three age points: 2.5, 3.5, and 5.5 years — and by the children's parents. We tested different 'listeners' on their ability identify these productions through transcriptions (Experiment 1) and looking behavior (Experiment 2). 'Listeners' included a commonly used ASR system (i.e., Apple's Siri) and human listeners who vary in exposure to children (i.e., toddlers, young adults, and mothers). We predicted that human listeners would generally outperform Siri, and that their experience would modulate performance particularly with 2.5-year-olds' speech.

In Experiment 1, young adults (N=48), mothers (current N=7, target N=48), and Siri transcribed 32 single word productions-in-noise (0 SNR) by 12 adults and 12 children. Listeners heard productions by the same 12 children at all three ages. Siri dictation and history were cleared after each of the 48 testing sessions. As predicted, all listeners performed worst with 2.5-year-olds' productions, and humans outperformed Siri with all ages ($p$<0.001; Figure 1). With 2.5-year-old productions, performance differences were the most apparent — mothers demonstrated the highest accuracy (86%). Siri also made distinctive types of errors with children's speech. For example, Siri failed to accurately transcribe all productions of the word 'bunny', whereas human listeners were accurate 86% of the time. Siri transcribed 'bunny' as 'funny' 90% of the time; however, humans transcribed 'bunny' as 'funny' only 2% of the time.

In Experiment 2, we used an eye-tracking procedure that used the same 2.5-year-olds' recordings as in Experiment 1. Listeners included 2.5-year-olds (N=48), young adults (current N=16, target N=48), and mothers of young children (current N=16, target N=48). We predicted that mothers and 2.5-year-olds would outperform young adults, given their routine exposure to young children. Across two 500-ms windows starting from target word onset, all listeners fixated the target above chance (50%; $p$<0.05), but contrary to our predictions, adults looked to the target more than toddlers did ($p$<0.001; Figure 2). On average, mothers' target fixations (78%) were numerically greater than young adults' (74%), but with the current sample size, this difference was not statistically significant ($p$=0.17).

ASR technology is becoming increasingly common in homes and classrooms (e.g., Alwan et al., 2007). Yet little is known about how these systems cope with the speech of children under 6. Our findings support previous literature that humans are better at understanding children than ASR systems (e.g., Gerosa et al., 2009), and in Experiment 1 we show dramatic differences in performance accuracy by man versus machine, especially at the younger ages. Interestingly, we found some evidence that experience with children may help adults understand children (i.e., mothers outperformed young adults in Experiment 1), but experience is not the only factor involved (i.e., children performed poorly in Experiment 2). In follow-up experiments, we hope to build on this work by testing more ASR systems on more child voices, and on stimuli containing multi-word utterances. We will also further investigate how varying exposure to children (e.g., teachers, mothers, toddlers in/not in daycare) impacts comprehension of children. Lastly, we will further characterize the types of transcription errors made by ASR systems and humans to help guide future attempts to improve how ASR systems handle child speech.

References

Alwan, A., Bai, Y., Black, M., Casey, L., Gerosa, M., Heritage, M., Iseli, M., Jones, B., Kazemzadeh, A., Lee, S., Narayanan, S., Price, P., Tepperman, J., & Wang., S. (2007). A System for Technology Based Assessment of Language and Literacy in Young Children: the Role of Multiple Information Sources. *In Proc. of International Workshop on Multimedia Signal Processing*, Chania, Creete, GREECE.

Gerosa, M., Giuliani, D., Narayanan, S., & Potamianos, A. (2009). A review of ASR technologies for children's speech. In Proceedings of the 2nd Workshop on Child, Computer and Interaction (WOCCI '09)., *Association for Computing Machinery* New York, NY, USA, Article 7, 1–8. https://doi.org/10.1145/1640377.1640384

Vihman, M. (1993). Variable Paths to Early Word Production. Journal of Phonetics, 21, 61–82.

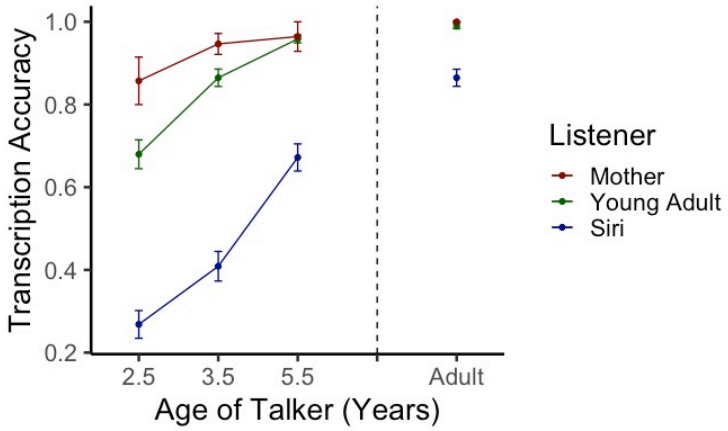


**Figure 1**. Average transcription accuracy by mothers, young adults and Siri on productions by children at three age points and adults (error bars indicate SE). With the youngest age group, mothers demonstrated the highest transcription accuracy (86%), followed by young adults (68%), and then Siri (27%).

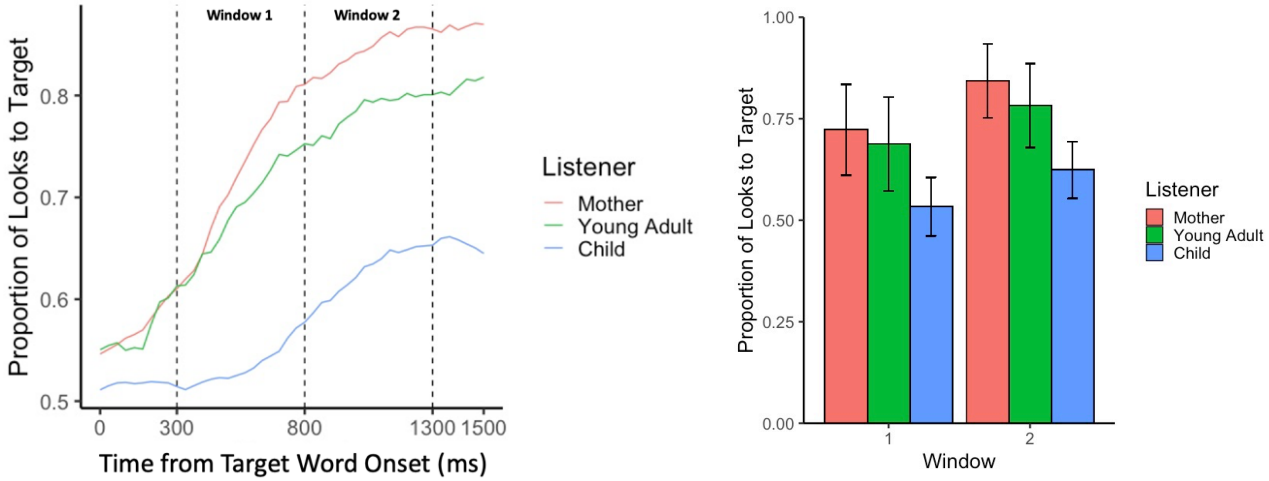


**Figure 2**. Average proportion of looks to target by mothers, young adults, and children across 300-800ms (Window 1) and 800-1300ms (Window 2) from word onset (error bars indicate SE).