Kid-Talk: Kid- to- kid Speech Comprehension

Alex Farrow, Sarah Creel (PI, advisor), Anne Beatty-Martínez (secondary advisor), Alexandria Evans (graduate advisor), LASR lab team

Abstract

Understanding how children recognize and process speech is a fundamental aspect of language development. Historically, research has shown that children primarily use adult speech patterns as models for their linguistic development. However, less is known about how children recognize and understand speech from their peers.

In this study, we investigate the ability of children aged 3-5 to recognize self-produced versus peer-produced speech. We used a combination of pointing accuracy and eye-tracking methods to assess how children identify words spoken by themselves and other children. In Particular, we sought to find out if children understand better articulators better, or themselves. These findings would provide valuable insights into the nature of stored/held linguistic representations. Our results so far indicate that better speakers may be better understood by children within their peer groups, and the data does not reveal a strong preference for self-produced speech in language comprehension, highlighting the importance of peer interactions in early language acquisition as well. This contributes with earlier research that emphasized adult speech as a primary model for children's linguistic representations.

These findings underscore the need to consider the social and interactive contexts of children's language environments in understanding their speech recognition processes. By integrating peer interactions into language development research, we can gain a more comprehensive view of how children acquire and process language.

This study broadens our perspective on child speech recognition, emphasizing the dual influences of adult and peer interactions. Future research should explore diverse socio-linguistic environments to further understand the complexities of early language development.

Introduction

There is a problem in the realm of linguistics, some may say a very central problem (Branigan & Pickering, 2016), of the nature of stored linguistic representations and their production in speech. As auditory inputs are processed and mapped (Cooper et al. 2018) onto representative linguistic forms, they must be reproduced in speech, however this mapping is confounded by a variety of linguistic limitations that arise as a result of normal physiological ("vocal tract size, speaking rate, and accent") and linguistic development ("smaller vocabularies, less robust phonemic categories, and [children] are still learning what variation is phonologically relevant for distinguishing between words and what variation can be ignored"). Variability of linguistic outputs also poses a problem when considering exactly how auditory inputs are represented and produced, children are constantly exposed to a wide variety of inputs when developing their linguistic repertoires, and this variety of exposure can also be said to influence their productions. Proximity and frequency would be thought to confer a great impact to child speech development, as in, the people who are with the child the most, speaking the most with the child will affect their development more than anybody else, and may facilitate higher rates of word recognition than unfamiliar inputs. A finding of Cooper et al.'s study found that this was not dramatically true, and that children will always recognize adult speech productions, regardless of relation, more than their own speech productions, and their representations of words are based on general 'adult speech' targets. Dodd (1975) conducted a series of experiments primarily testing the theory "that when a child begins to speak he uses the adult surface phonemic system as an input to his phonological system" and came to a similar conclusion of

Cooper et al.. Experimentally it has been demonstrated so far that children overall are using adult speech productions when forming their own linguistic representations, and when that their own speech productions will attempt to come as close as possible to adult speech productions rather than idiosyncratic child speech productions. Within these experimental procedures the main focus lies on *child to adult*, and *child to self* speech recognition production- reception, but it remains to be seen whether or not kids understand themselves better than they understand other kids, and the extent to which their environment among other children impacts their recognition of speech forms.

There are two hypotheses that are expected to bear out in these experiments, that of the *articulation error hypothesis* and the *multiple representations hypothesis*. The articulation error hypothesis would show that as articulation errors decrease, comprehension would increase, and not reveal a preference for speaker type(self, other), but for level of articulation (as determined by GFTA scores). Multiple representations refer to the fact that multiple representations are held by individuals when utilizing and deploying speech, the most common and familiar representation being one's *own speech*, which would reveal a preference for *self* produced speech.

Methods

Approximately 72 healthy, normally developing, participants aged 36-60 months (3-5 years old) will be recruited for the study. Participants are screened via a language background survey on completion of their signed consent form, as filled out by a parent/guardian, and distributed via the preschool office from researcher to parents of participants and back to the researchers. Part of the screening will gather the linguistic background of the home of the participants, which is considered in the experimental design.

Participant recruitment is still ongoing, as of June 2024 a preliminary analysis was done with subjects n= 46(/72).

Lists: Language background and explicit age will determine the sequence and list for each participant. For instance, a child three years 11 months and 29 days old will be considered within the three-year-old list, their language background would also determine which version of the three-year-old list they participate in, the division being between monolingual and multilingual background. There are six lists total from this, three-year old monolinguals, three-year-old bilinguals, four-year-old monolinguals, four-year-old bilinguals, five-year-old monolinguals, and five-year-old bilinguals. Again, bilingual status does not explicitly refer to the child's linguistic status, as in if they explicitly *speak* the language, but rather if they have any kind of moderate exposure to another language at home via their parents, relatives, media, etcetc.

Each participant is assigned a sequential number in the list, the sequence is able to continue to the next participant after the successful completion of the experimental trail of the previous participant. Because the experimental design relies on the successful completion of the previous experiment, this does mean that the first participant of each list will not be able to participate in the experiment, as they need to provide the initial recordings (detailed below) to get the experiment going. So each list's first participant is n=0, without an experimental trial. This participant will go through the familiarization and naming phases (detailed below) and will also have the standard GFTA administered by the experimenter, and participant n=1 in each list

will then go through the familiarization, naming, GFTA phases and then will be able to participate in the experiment.

N= 72 participants refers to the number of successful experiential trails run, six children will not receive the experimental trail as they will start out our lists, and there will be numerous children who may not successfully complete the study through the experimental trail, and will need to be dropped. The true number of participants recruited for this reason may be well beyond N=72.

Procedure: Participants engaged in three phases leading up to the experimental trial: Picture familiarization (conceptual pact), naming (recording), and an articulation test(GFTA) administered while another researcher edited the produced recordings for the experimental trial. During the conceptual pact the researcher only uses the noun form of the stimulus, without any leading words as the participant is shown two images, one image is of the stimulus to be used in the experiment, and the other is simply a random object. The researcher will ask the participant, using the noun form only, to point to the stimulus: “Can you point to [stimulus]” and to continue the participant must point to the requested image.

During the naming phase the participant will use a headset and microphone to repeat the stimulus for the experiment, as well as hear the previous participant’s responses for the stimuli. The order in which this happens is counterbalanced according to their position in the list, so in the naming phase the participant will either first be requested to name the stimuli and then hear the previous participant’s productions, or they will first hear the previous productions, and then name the stimuli.

After the naming phase is concluded, a researcher will administer the test of articulation, and these responses are recorded using another device. Simultaneously, another researcher will edit the sound files of the just-produced recordings using an in-house PRAAT script. Recordings will be edited to cut out excess empty noise, or otherwise non-stimuli related sounds from the recording, only the full pronunciation of the content stimuli will be preserved in the recording, i.e., only the name of the stimuli itself.

After the articulation test and recordings are prepared, the experimental trial will begin. Eye tracking software and equipment is readied and deployed as well during this phase. There is again a counterbalance to determine which set of stimuli the participant will hear first, themselves or the other previous participant in the list, which alternates with each participant in the list. There will be 48 total trials during this phase for successful completion, the participant will hear each stimulus twice in this phase. The counterbalance determines which *set* is heard first, so that when the participant hears themselves, they hear every recording of themselves, and then in the next phase of the experimental trial they will hear all of the previous participant’s recordings (or vice versa). Each of the 48 experimental trials will present four of the familiarized stimuli, as seen in Fig. 1. During each trial the images will appear, and simultaneously the audio recording will play either themselves or the other child’s recording. Participants are asked at the beginning of the experimental phase to simply point to whichever word they have heard in the headphones, and based on which image they point to, a researcher will click that image to continue to the next trial. Because all the stimuli were recorded and organized via the name of

the stimuli, there is only one ‘correct’ answer, and this will determine accuracy results.

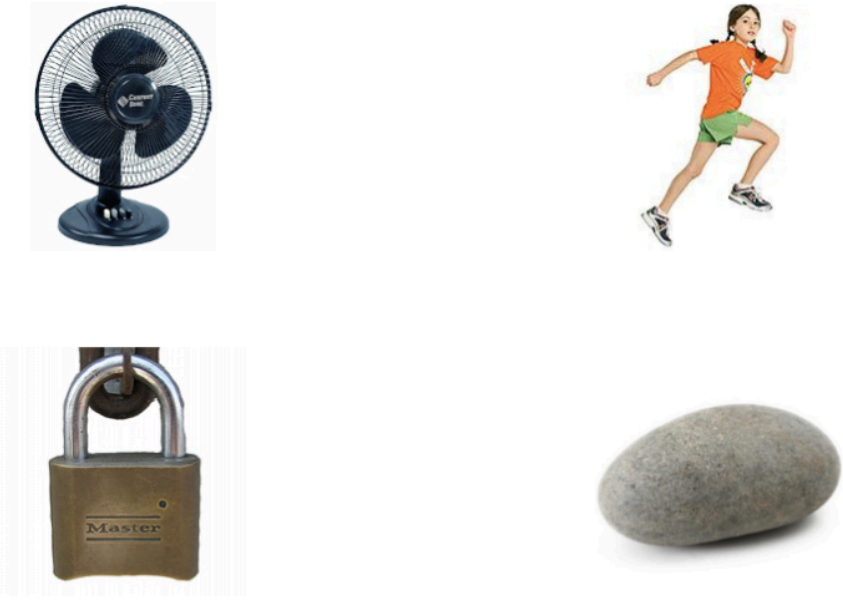

Figure 1: An example of one trial in the experimental phase. In this trial, for example, the correct answer(the audio played) would be ‘lock’ which differs by a minimal pair with ‘rock’. This is the true experiment, the choice here would reflect either better self understanding or better articulator understanding across the two responses to ‘lock’ for each participant.

Stimuli: Stimuli consist of a set of 24 images designed to elicit a set of common phonetic sounds in English, and normed to be conventionally recognizably clear images of common everyday objects, such as “shoe, fox, window” etc. Many of the stimuli are designed to produce minimal pairs, the target words being “ring-wing, lock-rock, chip-ship” etc for the images provided. In this study each child will produce a recording of what they determine the stimuli is called, and will hear both their own and another child’s production in the identification phase (this is dependent on both children using the same production after exposure in familiarization phase). Participants will receive the same set of stimuli. The stimuli in the identification section include a visual image of a common object displayed on the screen, and the participant will be prompted to identify the name of the stimulus which will be recorded during this section (identification+naming).
Equipment/materials: A linux machine, 1920x1080p monitor, periphery mouse, keyboard, a headset with microphone, portable Eyelink eye tracking camera, laptop housing eye tracking software, and separate recording device(for GFTA) are brought to each location for this study. In house matlab scripts are used for each phase of experiment. The Goldman-Fristoe Test of Articulation(GFTA) is administered on site. Transcription for study materials(stimuli productions, GFTA scores) include trained RA within the LASR lab at UCSD.
Data collection(and transcription info): Data to be collected include: audio waveform data collected then processed in PRAAT. Transcriptions will be done after sessions of each recording, and IPA transcriptions will be used in non-normal child speech productions, indicating the variation and recording it via a Google Sheets shared by RAs in the lab. GFTA data will be

analyzed via a researcher and recorded by hand to record a score for production proficiency, the transcription scores judge level of articulation. Eye tracking data (stimuli response time) will be collected using an Eyelink portable camera and software. Identified stimuli responses will be recorded using in house scripts.

Human subjects information:

All participants gave their informed consent in accordance with the protocols approved by the University of California, San Diego Human Research Protections Program. IRB approval was obtained for data collection as well.
Parents of each participant were to fill out the form: 'Using eye tracking to understand speech perception-production relationships in young children (190038)'. Funded by a grant from the National Institutes of Health.
Because data collection was done on-site at local private preschools, the directors of participating preschools were also required to fill out a Director's Letter, giving consent for the lab to operate on the premises.

Data analysis

This study measured pointing accuracy, eye movement speeds upon stimulus onset, and level of articulation. The preregistration for the study details that for the purposes of a student's honor's thesis, there would be a partial preliminary analysis of the data for a thesis presentation. Data was analyzed using in house Python scripts, Google Sheets, and RStudio.

Results

Pointing Accuracy by speaker:

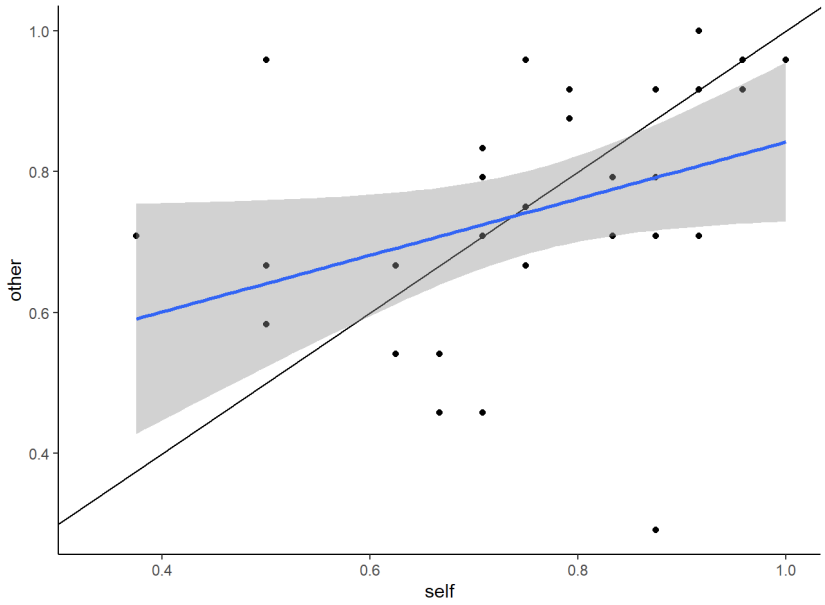

Figure 2.1: Plot of accuracy by speaker quantifying the strength and significance of the relations between "self" and "other" accuracy

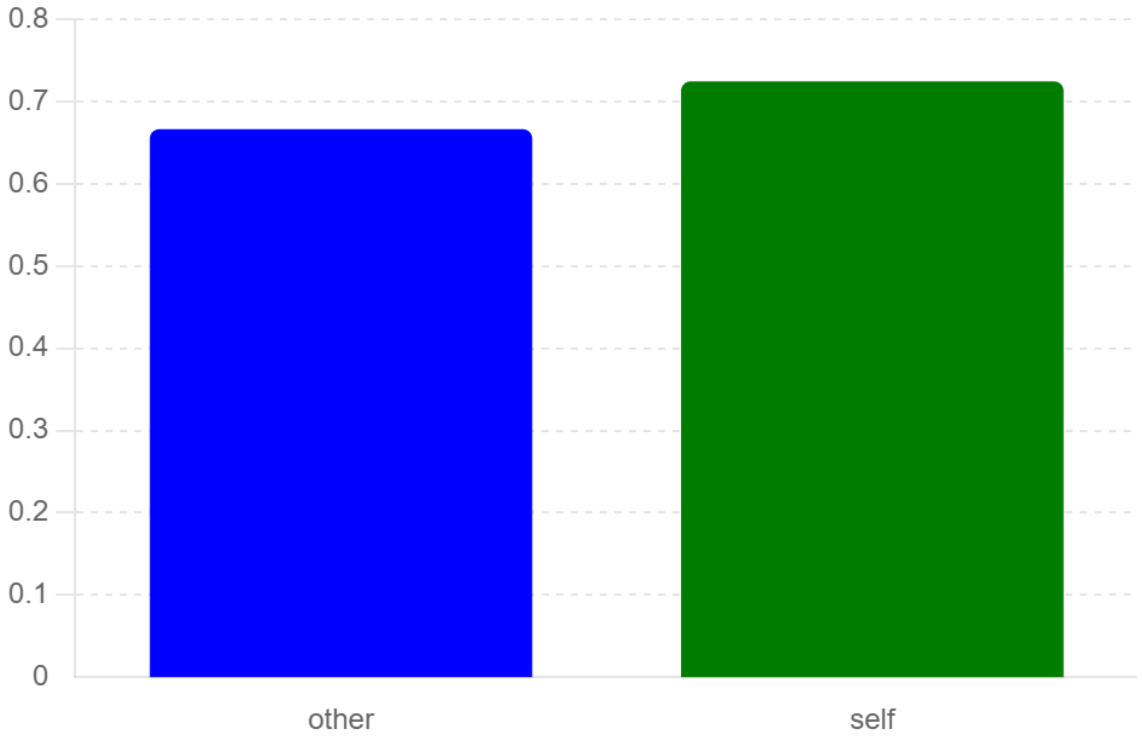

Figure 2.2: Bar graph of pointing accuracy by speaker

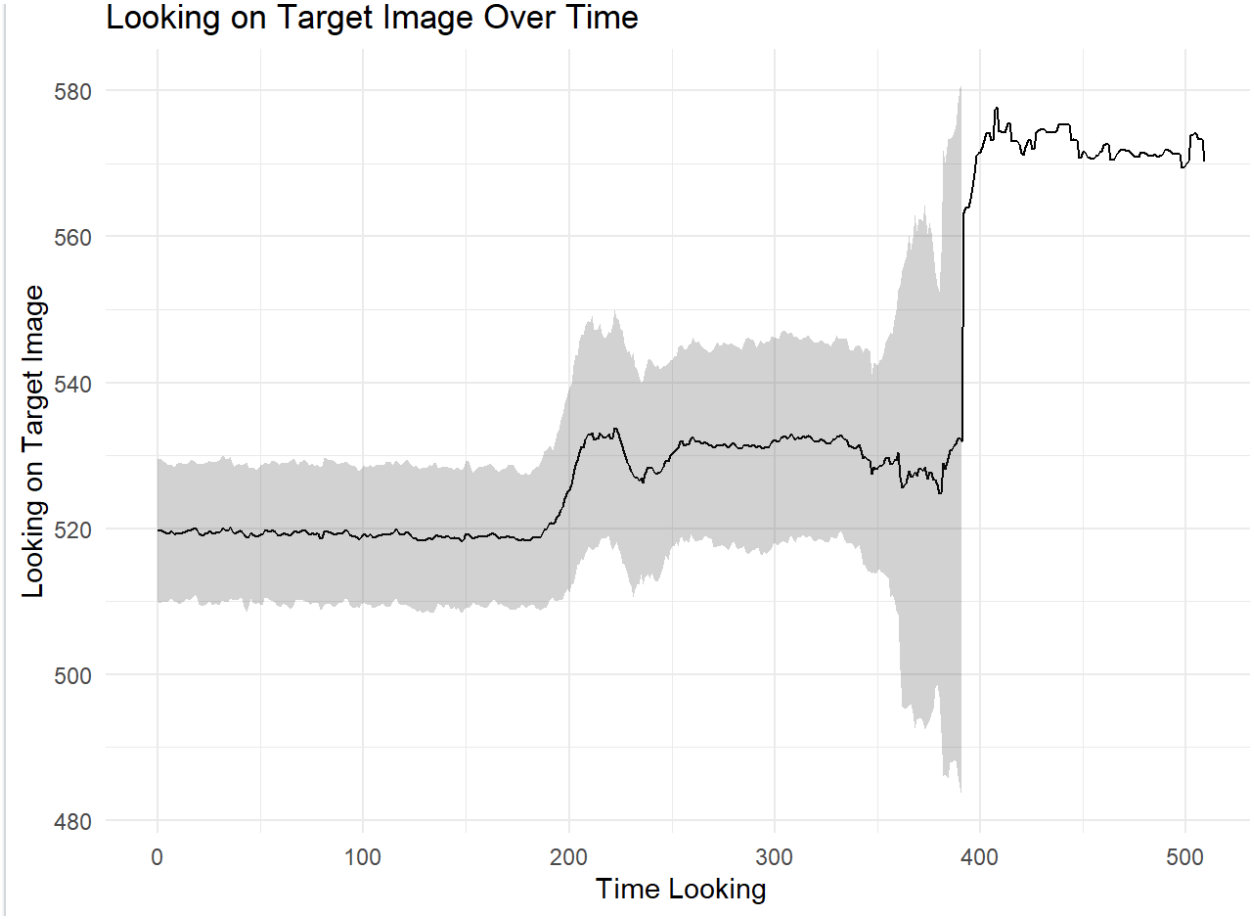

Fig 3: Eye-tracking plot showing the dynamics of how participants' attention shifts over time when looking at a target image. Key phases include stable initial looking behavior, a transition phase with increased attention (due to stimulus), and subsequent stabilization at a higher level of looking. Variability in the shaded area suggests differences in individual responses to the stimuli.

Discussion

Hypotheses: From the data we see pointing accuracy very narrowly showing preferences for self speech when it comes to comprehension in the bar graph for pointing accuracy by speaker, but this preference may not be significant. The plot indicates that pointing accuracy is more even divided by speaker type. This would imply that the *articulation error hypothesis* has been borne out by the data, confirming that better articulators are better understood by children in kid to kid communication. This may indicate that the nature of their held representations are more closely aligned with what is perceived to be ‘good speech’, regardless of the articulation level of one's own speech. Again, data collection is still ongoing, and upon completion it will be confirmed via reference with GFTA scores whether this is the case or not.

Historical Context: Understanding child speech recognition has been a key area of linguistic research for decades, with foundational work exploring how children process and produce language. The concept that children predominantly use adult speech patterns as their linguistic model has been well-documented since the mid-20th century. Early studies, such as those by

Dodd (1975) and Zlatin and Koenigsknecht (1975), established that children's phonological systems are heavily influenced by adult speech forms, which they use as primary targets in speech production. More recent work by Cooper et al. (2018) further supported these findings, emphasizing that children's speech recognition capabilities are more attuned to adult speech rather than their own or their peers'.

Our study contributes to this historical context by examining an area less explored: the comparative recognition of self-produced speech versus peer-produced speech among children. This investigation helps to fill a gap in understanding the developmental nuances of child speech recognition and extends the existing body of knowledge by providing empirical data on how children perceive and process speech from their peers versus their own.

Caveats and Limitations: While our study provides valuable insights, it is not without limitations. One significant caveat is the sample size and demographic limitations. Although our participant pool is sizable, it is not fully representative of the broader population, potentially limiting the generalizability of our findings. The participants were predominantly from similar socio-economic and linguistic backgrounds, which may not capture the variability seen in more diverse settings.

Another limitation is the potential variability in the children's familiarity with the stimuli used. Despite efforts to standardize the images and ensure they were commonly recognizable objects, individual differences in prior exposure did influence recognition accuracy. Additionally, the reliance on pointing accuracy and eye-tracking as primary measures, while robust, may not capture the full spectrum of cognitive processes involved in speech recognition.

Technical limitations also exist in the use of eye-tracking equipment and PRAAT scripts for sound editing. Variations in equipment calibration and environment and script precision could introduce minor inconsistencies in data collection and analysis. Furthermore, the sequential nature of the experimental design, where each child's responses depended on the previous participant's recordings, could introduce a compounding error effect, and requires careful vetting.

Data collection is still ongoing, and all results presented here are *highly* preliminary. Further and more advanced cleaning will also be required, as will cross correlation between GFTA scores and pointing accuracy. The author of the paper will see to it that these things are eventually rectified upon total data collection at N=72 participants.

Forward Thinking: This study highlights several important findings and opens new avenues for future research. Firstly, the clear preference for recognizing adult speech over peer speech or self-produced speech suggests that adult models(good articulators) remain central to children's speech processing well into early childhood. This reinforces the importance of adult-child interactions in language development.

Future research should aim to expand the demographic diversity of participants to examine how socio-economic, cultural, and linguistic backgrounds influence speech recognition. Longitudinal studies could provide deeper insights into how these recognition patterns evolve with age and continued exposure to peer speech.

Technological advancements, such as more sophisticated eye-tracking software and automated transcription tools, could enhance data accuracy and analysis efficiency. Additionally, exploring other modalities of speech recognition, such as incorporating visual lip-reading cues, could

provide a more comprehensive understanding of how children integrate multi-sensory information in speech processing.

Grand Conclusion: In conclusion, our study provides critical insights into the dynamics of child speech recognition, particularly in comparing self-produced and peer-produced speech. Consistent with historical findings, children show a strong preference for adult speech representations, which underscores the enduring influence of adult interactions on language development. Despite the study's limitations, the findings emphasize the nuanced nature of speech recognition in children and pave the way for future research to build on these initial insights. Moving forward, expanding demographic inclusivity and leveraging advanced technologies will be essential in deepening our understanding of child language acquisition and its developmental trajectory.

References

Branigan, H. P., & Pickering, M. J. (2016). An experimental approach to linguistic representation. *Behavioral and Brain Sciences*, *40*(40). https://doi.org/10.1017/s0140525x16002028

Cooper, A., Fecher, N., & Johnson, E. K. (2018). Toddlers' comprehension of adult and child talkers: Adult targets versus vocal tract similarity. *Cognition*, *173*, 16–20. https://doi.org/10.1016/j.cognition.2017.12.013

Dodd, B. (1975). Children's Understanding of their Own Phonological Forms. *Quarterly Journal of Experimental Psychology*, *27*(2), 165–172. https://doi.org/10.1080/14640747508400477

Zlatin, M. A., & Koenigsknecht, R. A. (1975). Development of the Voicing Contrast: Perception of Stop Consonants. *Journal of Speech and Hearing Research*, *18*(3), 541–553. https://doi.org/10.1044/jshr.1803.541

Further literature sources/ reading:

Creel, S. C. (2012). Preschoolers' Use of Talker Information in On-Line Comprehension. *Child Development*, *83*(6), 2042–2056. https://doi.org/10.1111/j.1467-8624.2012.01816.x

Creel, S. C., & Bregman, M. R. (2011). How Talker Identity Relates to Language Processing. *Language and Linguistics Compass*, *5*(5), 190–204. https://doi.org/10.1111/j.1749-818x.2011.00276.x

Creel, S. C., & Jimenez, S. R. (2012). Differences in talker recognition by preschoolers and adults. *Journal of Experimental Child Psychology*, *113*(4), 487–509. https://doi.org/10.1016/j.jecp.2012.07.007

Hazan, V., & Markham, D. (2004). Acoustic-phonetic correlates of talker intelligibility for adults and children. *The Journal of the Acoustical Society of America*, *116*(5), 3108–3118. https://doi.org/10.1121/1.1806826

MacDonald, Ewen N., Johnson, Elizabeth K., Forsythe, J., Plante, P., & Munhall, Kevin G. (2012a). Children's Development of Self-Regulation in Speech Production. *Current Biology*, *22*(2), 113–117. https://doi.org/10.1016/j.cub.2011.11.052

MacDonald, Ewen N., Johnson, Elizabeth K., Forsythe, J., Plante, P., & Munhall, Kevin G. (2012b). Children's Development of Self-Regulation in Speech Production. *Current Biology*, *22*(2), 113–117. https://doi.org/10.1016/j.cub.2011.11.052

Schuerman, W. L., Meyer, A., & McQueen, J. M. (2015). Do We Perceive Others Better than Ourselves? A Perceptual Benefit for Noise-Vocoded Speech Produced by an Average Speaker. *PLOS ONE*, *10*(7), e0129731. https://doi.org/10.1371/journal.pone.0129731

Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, *76*(2), 147–166. https://doi.org/10.1016/s0010-0277(00)00081-0

Yu, M., Cooper, A., & Johnson, E. (2021). Do you speak “kid”? The role of experience in comprehending child speech. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43). https://escholarship.org/uc/item/20p86589