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

Computational simulations of temporal vocalization behavior in adult-child interaction

Ellen Marklund, David Pagmar, Tove Gerholm and Lisa Gustavsson

Stockholm Babylab, Phonetics Laboratory, Dept. of Linguistics, Stockholm University, Sweden
[ellen.marklund|david.pagmar|tove.gerholm|lisa.gustavsson]@ling.su.se

Abstract

The purpose of the present study was to introduce a computational simulation of timing in child-adult interaction. The simulation uses temporal information from real adult-child interactions as default temporal behavior of two simulated agents. Dependencies between the agents' behavior are added, and how the simulated interactions compare to real interaction data as a result is investigated. In the present study, the real data consisted of transcriptions of a mother interacting with her 12-month-old child, and the data simulated was vocalizations. The first experiment shows that although the two agents generate vocalizations according to the temporal characteristics of the interlocutors in the real data, simulated interaction with no contingencies between the two agents' behavior differs from real interaction data. In the second experiment, a contingency was introduced to the simulation: the likelihood that the adult agent initiated a vocalization if the child agent was already vocalizing. Overall, the simulated data is more similar to the real interaction data when the adult agent is less likely to start speaking while the child agent vocalizes. The results are in line with previous studies on turn-taking in parent-child interaction at comparable ages. This illustrates that computational simulations are useful tools when investigating parent-child interactions.

Index Terms: first language acquisition, parent-child interaction, computational modeling, vocal turn-taking

1. Introduction

As interactants communicate, a number of distinct but coordinated systems are set to work, such as vocalizations [1, 2], breathing [3, 4], gestures [5, 6], touch behavior [7], and gaze patterns [8]. Researchers have for a long time captured these behaviors in transcribed form, and descriptions of how they might be coordinated and used in communication abound [9, 10, 11].

Interactants tend to modify their behaviour depending on the behaviour of the interactional partner, a phenomonenon called entrainment. In terms of vocal behavior, speakers have been shown to adjust for example their speaking rate [12] and their utterance duration [13] to be more similar to that of their conversational partner. Entrainment has also been found for non-verbal behaviors such as bodily postures during conversation [14]. Interaction can thus be described as a highly adaptive process.

When the one interlocutor is an adult and the other a child or infant, adaptations on part of the adult have often been the focus. Numerous studies have shown various ways in which adults modify their behavior when interacting with children. For example, adults' adaptation when interacting with children is not limited to vocalizations; adults interacting with children differ in their handling of interaction patterns and breaches of the same, compared to when they interact with other adults [15]. Speech directed to infants or children differs from speech directed to adults (for a review, see [16]), for example in terms of exaggerated prosodic variation [17, 18] and different weighting of phonetic cues [19, 20]. Another aspect of speech known to be modified when speaking to children is temporal characteristics of vocalization behavior. Utterances are typically shorter [21], and pauses longer [22, 18], although both utterance duration and pause duration vary depending on the age of the child [23]. Further, adults show different temporal vocalization behavior depending on the linguistic proficiency of the child [24].

All in all, adult-child interaction is typically characterized by a high degree of adaptation on the part of the adult. Although adaptive behavior has in some cases been observed also on the part of the child [25], it has been relatively less studied.

1.1. The present study

The purpose of this study is to introduce a computational model of interaction between an adult and a child. The model simulates parent-child interactions based on temporal characteristics of real interaction data, and explores the impact of dependencies between the two interactants' behavior. In essence, the model generates a simulated interaction between two agents, and different ways in which the agents can adapt their behavior to each other can be explored.

In the present paper, two experiments will be reported, both simulations of vocal behavior of a 12-month-old infant and his mother. The purpose of the first experiment is to confirm that there are dependencies between the interactants' vocal behaviors by simulating a situation where there are no dependencies. It is expected that the simulated interactions in this experiment differ from the real interaction used as model. The purpose of the second experiment is to illustrate the impact of adding a contingency to the simulation. Specifically, the likelihood of the adult agent starting to vocalize while the child agent is already vocalizing will be manipulated. Previous research show that parents tend to co-vocalize with their infants when the infants are very young [26], whereas when the infants are somewhat older, nine months have been reported, parents and child take turns to vocalize [27]. Based on this, and the fact that the child in the modeled interaction is older than nine months, it is expected that a lower probability of adult agent co-vocalizing with the child agent will result in the simulated interactions being more similar to the real data, and vice versa.

2. Method

2.1. Real data

Real interaction data was in the present study based on a recording of a mother interacting with her 12-month-old son. The recording was one of several collected within a longitudinal project on early parent-child interaction (MINT, Marcus and Amalia Wallenberg Foundation, 2011.007). A total of 75 fam-

http://dx.doi.org/10.21437/Interspeech.2017-1289

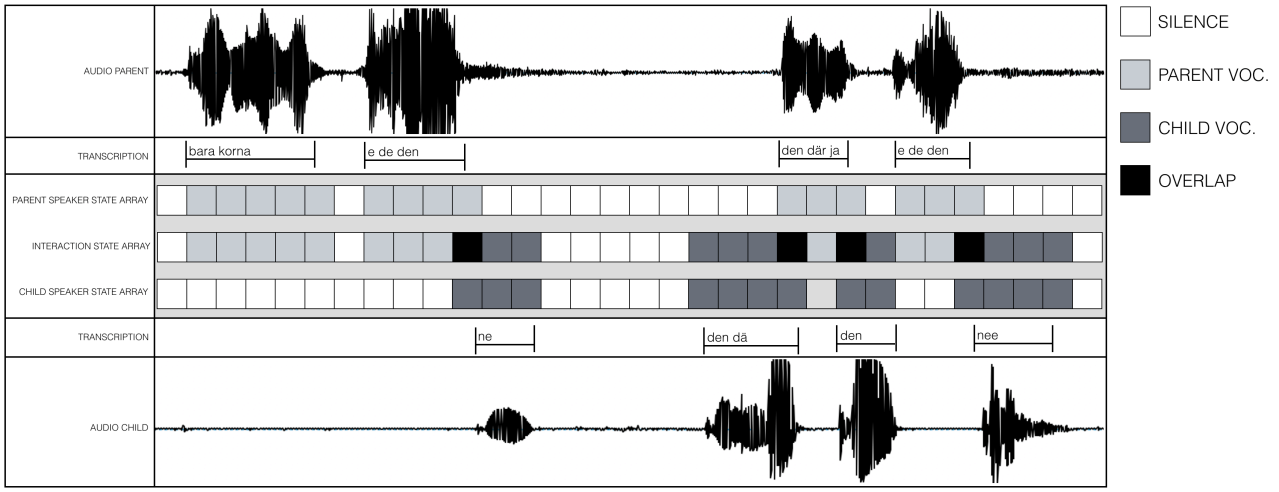

Figure 1: *The audio recordings were transcribed, and converted into down-sampled speaker state arrays. The state consisted of a series of ones and zeros (here illustrated by gray and white boxes respectively). The speaker state arrays of the parent and the child were converted into an interaction state array, denoting the combined state of the two speakers: silence (white boxes), only adult is vocalizing (light gray boxes), only child is vocalizing (dark gray boxes), or overlap (black boxes).*

ilies contributed to the project, parents visiting the laboratory with their child at three-month intervals from the age of three months up to their fourth birthday (every sixth months during the last year). During each lab session, parents and children were recorded while engaged in free play and a number of guided activities and tests. The recordings were annotated according to an extensive mark-up scheme, including both non-vocal behavior such as gaze and touch, as well as transcriptions of parent, child and experimenter vocalizations (for a more in-depth description, see [28]).

One child was randomly selected as model for the simulations in the present study. No hearing issues were reported for the child, nor any speech or language delays for any member of the family. Both parents were native speakers of Swedish. The rationale for using the selected session, with a 12-month-old child, was to ensure there would be vocal behavior to model on the part of both adult and child.

Time-stamped transcriptions of vocalizations from the recorded interaction session were down-sampled to 500 ms time frames, and converted into speaker state arrays. A speaker state array essentially consists of a series of zeros and ones, in which each digit represent a time frame and encodes the state of the speaker (1 = speaker is vocalizing, 0 = speaker is silent). The real data was further converted into an interactional state array, in order to be readily comparable to the simulated data (see Figure 1). The interactional state array is similar to a speaker state array, except that instead of encoding a single speaker's state of vocalization, it encodes the current state of the vocal interaction between the two speakers. If both speakers are vocalizing, the state for that time segment is overlap, whereas if none of the speakers are vocalizing, the state is silence. If only the adult is vocalizing, the interaction state is adult vocalization, and if only the child is vocalizing the interaction state is child vocalization.

For comparison with simulated data, short sections were sampled at random time points from the interaction state array of the entire recording. The number of samples drawn from the real data, as well as their duration, were always matched to the number of simulated samples and their duration. Original transcriptions were performed in ELAN [29], and all following operations were done in Mathematica 9 (Wolfram Research Inc., Champaign, Illinois, USA).

2.2. Simulated data

In essence, the simulated data consists of speaker state arrays; generated frame-by-frame in time frames of 500 ms, for each of the simulated speakers, the agents. The temporal characteristics of the simulated arrays, that is, how long each state lasts, are modeled on temporal information from the real data (down-sampled and converted to speaker state arrays). The duration of vocalizations and silences were extracted for each of the speakers, and probability functions created for the agents. The probability functions determine the likelihood of a vocalization state change for each of the agents. For example, if the adult agent is currently vocalizing (speaker state = 1), and the immediate state history reveals that this is the fourth segment of vocalization in a row (current vocalization duration = 2000 ms), the probability function dictates a 70% probability of a change of state (the agent stops vocalizing). This probability is calculated based on the adult's vocalization durations in the real data.

When dependencies are added between speakers, an agent's probability of a state change is not only dependent on the agent's own probability function based in its state and state history, but also upon the immediately preceding state of the other agent. In the present setup, any dependency is varied between -1 and 1 in steps of 0.25, and its value is added to the agent's current probability of a state change (based on the probability function as described above), although the combined probability naturally has a floor at 0 and a ceiling at 1.

Each simulated segment was 60 seconds in total. In each run of the simulation, 50 such interaction segments were generated. For each segment, an interaction state was created based on the simulated speaker state arrays, to enable comparison between simulated interaction data and real interaction data. All operations were done in Mathematica 9 (Wolfram Research Inc., Champaign, Illinois, USA).

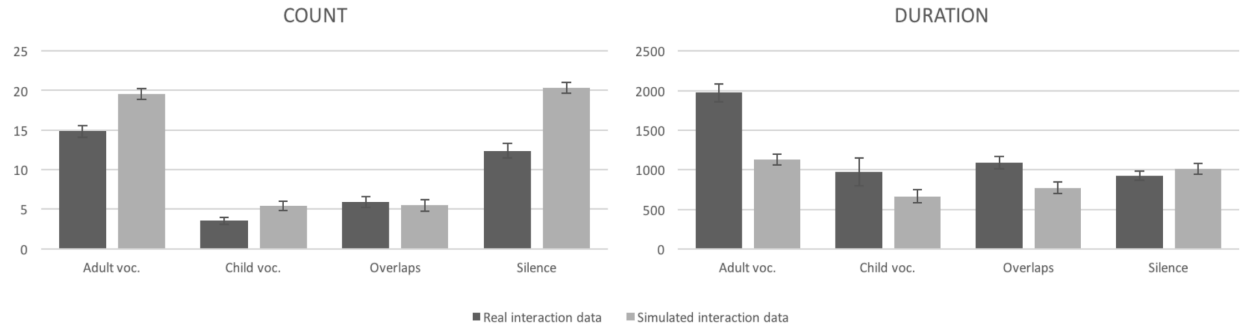

Figure 2: *The result patterns from Experiment 1, in terms of mean interactional state count (left) and mean interactional state duration in ms (right). Dark columns show real interactional data, light columns show simulated interactional data. Error bars denote the 95% confidence interval.*

2.3. Experiments

In both experiments the focus was on vocal behavior. In Experiment 1, no dependencies were added to the simulated interaction. That is, the adult agent and the child agent each modeled the temporal characteristics of the real interactants, but with no relation between them. In Experiment 2, a dependency on the part of the adult was added. The probability of the adult agent starting to vocalize while the child agent was vocalizing was varied from one extreme (the adult agent never starts to vocalize if the child agent is vocalizing), via no dependency (the probability of the adult agent starting to vocalize is unrelated to whether the child agent is vocalizing or not), to the other extreme (the adult agent always starts to vocalize if the child agent is vocalizing).

2.4. Comparison

Each run of the simulation generated 50 simulated interaction state arrays and 50 samples from the real interaction state array. For each of those, the number of occurrences of each of the four states was counted (a segment with the same state repeated for any given duration is counted as one occurrence), and the mean duration of each state's segments was calculated. The mean and 95% confidence interval for each of those measures were then calculated for the entire run, based on all 50 samples.

In order to make sure patterns found in the results were stable, the simulation was run ten times for each condition. In Experiment 1, this entails a total of ten runs. In Experiment 2, the simulation was run ten times per value on the dependency variable, that is, a total of 90 times. A result was deemed stable for a given interaction measure if the confidence intervals did or did not overlap consistently in at least nine out of ten runs.

3. Results

Considering the high number of data points in each simulation, and the fact that each simulation was run multiple times in order to establish a stable pattern of results (see Method, Comparison), a visual representation of the overall result pattern is reported in lieu of specific statistical tests. The mean value of each measure, as well as its 95% confidence interval, is found in the figures.

The comparison between the simulated interaction data and the real interaction data of Experiment 1 can be seen in Figure 2. Real and simulated data differ in terms of number of adult vocalizations, number of child vocalizations, number of silences, duration of adult vocalizations, duration of child vocalizations, and duration of overlaps. The results were stable across runs for all measures, except for duration of silences. While in the majority of the runs there was no difference between real and simulated data for this measure, in 30% of the runs there was a difference. Still, five out of eight interactional timing measures differ consistently between the simulated and real interactional data when there is no dependency between the two agents in the simulations.

In Experiment 2, the likelihood of the adult agent starting to vocalize while the child agent was already vocalizing was varied. Three interactional measures were influenced by the variation: number of child vocalizations, child vocalization duration, and number of overlaps.

Number of child vocalizations in Experiment 2 can be found in Figure 3, top panel. Low probability of the adult agent starting to vocalize while the child agent is vocalizing resulted in real and simulated interaction data being indistinguishable from each other. When there was no dependency (value 0 in the figure) or the probability of the adult agent to start vocalizing while the child agent was vocalizing was high, the number of infant vocalizations was higher in the simulated data than in the real data.

Child vocalization duration in Experiment 2 is found in Figure 3, middle panel. The measure does not differ between real and simulated data when the adult agent is unlikely to start vocalizing if the child agent is vocalizing. With increased likelihood of the adult agent starting to vocalize while the child agent is already vocalizing, real and simulated data differ, with child vocalizations being shorter in the simulated data than in the real data.

Number of overlaps in Experiment 2 is shown in Figure 3, bottom panel. With low probability of the adult agent starting to vocalize while the child agent is vocalizing, the number of overlaps is smaller than in the real data, and vice versa. When there is no dependency (value 0 in the figure), there is no difference between real and simulated data.

The remaining five interaction measures stayed unchanged despite the variation in dependency. Number of adult vocalizations, number of silences, duration of adult vocalizations and duration of overlaps differed between the simulated interactional data and the real interactional data, regardless of the adult agent's behavior in the simulation. The silence durations

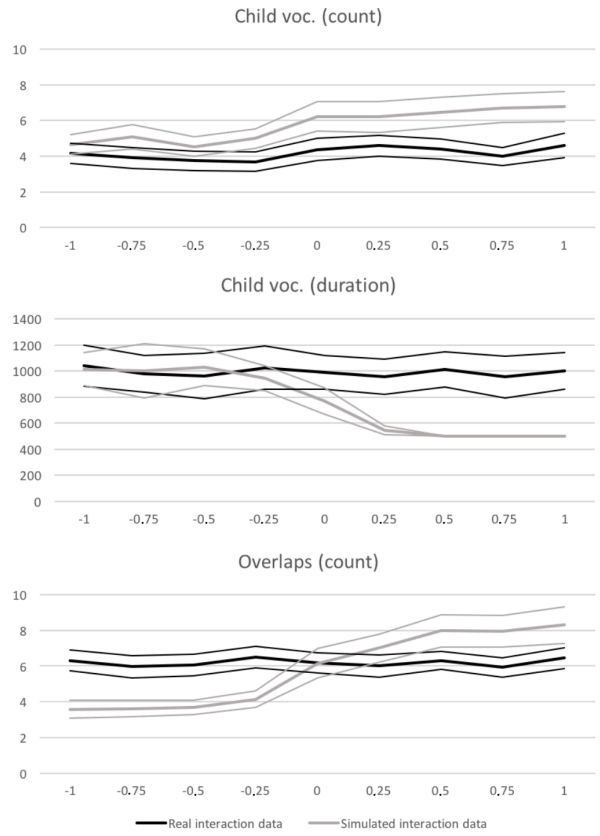

Figure 3: *The result pattern from Experiment 2, in terms of mean number of child vocalizations (top), mean child vocalization duration (middle), and mean number of overlaps (bottom). Dark lines show real interaction data (re-sampled for each run of the simulation), and light lines show simulated interaction data. Thick lines indicate the mean value and thin lines show the 95% confidence interval. The y-axis shows the modification of the probability of the adult agent starting to vocalize while the child agent was vocalizing.*

also remained unstable regardless of the modifications to the adult agents' vocalization behavior.

4. Discussion

The real and simulated interaction data differs in Experiment 1, showing that the modeled real parent and child do not vocalize randomly, but that their behavior are somehow related to each other. This in line with both intuitive expectations and previous knowledge about interaction as a process in which interactants typically adapt their behavior to each other [12, 13, 14]. More importantly, this demonstrates that the measures used in the present paper capture an aspect of functioning interaction.

One measure that consistently does not differ between real and simulated interaction data is number of overlaps. This means overlaps happen in the real data just as often as they would if the parent and the child would just be vocalizing randomly next to each other. Interestingly, the duration of the overlaps are actually longer in the real data. Intuitively, one would figure that random vocalizations would generate more and longer overlaps than coordinated interaction between the two speakers. It would be of interest to compare this unexpected finding to real and simulated data from interactions between two adults, to find out whether the long overlaps may be a result of immature turn-taking behavior on the part of the child. Adding information about other modalities to the picture might also shed light upon the long overlaps.

In Experiment 2, three interactional measures were influenced by the dependency added to the simulation. Number of child vocalizations and child vocalization duration were indistinguishable between simulated and real data when the adult agent's propensity to co-vocalize with the child agent was low, but differed from real data when the propensity was high. Number of overlaps differed between simulated and real data both when the propensity of the adult agent to co-vocalize with the child agent was high, and when it was low. Thus, expectations for the results of Experiment 2 were fulfilled, in that simulated data looked more like real data when the probability of the adult agent co-vocalizing with the child agent was low, than when it was high.

At first glance it might seem as though manipulation of "adult" agent behavior result in differences in "child" agent behavior. This was not the case as the "child" agent behavior remained the same throughout the experiments. However, when the propensity of the adult agent to co-vocalize the child agent is high, the duration of child vocalizations will be short, since most of the vocalization will be encoded as overlap. Similarly, as the "adult" agent more frequently co-vocalizes with the "child" agent, the child vocalization count will be higher due to its vocalizations being encoded as alternating sequences of child vocalizations and overlaps.

There are, as always, a number of aspects of the model that could be improved in future studies. More fine-grained temporal resolution may paint a more nuanced picture, as might smaller steps in the dependency variable. Different ways of combining the basic temporally-based probability of a state change with the dependency-based probability could potentially be of interest. Lastly, more detailed ways of comparing real and simulated data should be explored.

All in all, the present study is a successful first introduction of a computational model of temporal interaction behavior in adult-child interactions. The results were in line with expectations and previous knowledge about parent-child interaction, and it has given rise to new areas of interest to study further.

The model is versatile, in the sense that a large number of dependencies can be added and tested, separately or in combination. Further, it is not restricted to vocal behavior, but interactional behavior in any modality can be simulated. Finally, different sets of interaction data can be used as a model, for example parent-child interaction with children of different ages. The model presented here thus has the potential to contribute to a comprehensive picture of the multi-modal behaviors that constitute interaction.

5. Acknowledgements

The present study was funded by the MINT project (Marcus and Amalia Wallenberg Foundation, 2011.007) and the Department of Linguistics, Stockholm University.

Thanks to our colleagues Johan Sjons, Mats Wiren, Lena Renner, Elisabet Cortes and Iris-Corinna Schwarz for helpful comments on an earlier version of the manuscript. We are also grateful to all participating families in the MINT-project.

6. References

[1] J. Jaffe and S. Feldstein, *Rhythms of dialogue*. Academic Press, 1970, vol. 8.

[2] J. Jaffe, B. Beebe, S. Feldstein, C. L. Crown, M. D. Jasnow, P. Rochat, and D. N. Stern, "Rhythms of dialogue in infancy: Coordinated timing in development," *Monographs of the society for research in child development*, pp. i–149, 2001.

[3] M. Wlodarczak, M. Heldner, and J. Edlund, "Breathing in conversation: an unwritten history," in *Proceedings of the 2nd European and the 5th Nordic Symposium on Multimodal Communication, August 6-8, 2014, Tartu, Estonia*, no. 110. Linköping University Electronic Press, 2015, pp. 107–112.

[4] D. H. McFarland, "Respiratory markers of conversational interaction," *Journal of Speech, Language, and Hearing Research*, vol. 44, no. 1, pp. 128–143, 2001.

[5] A. Kendon, *Gesture: Visible action as utterance*. Cambridge University Press, 2004.

[6] O. Capirci, A. Contaldo, M. C. Caselli, and V. Volterra, "From action to language through gesture: A longitudinal perspective," *Gesture*, vol. 5, no. 1, pp. 155–177, 2005.

[7] I. Mantis, D. M. Stack, L. Ng, L. A. Serbin, and A. E. Schwartzman, "Mutual touch during mother–infant face-to-face still-face interactions: Influences of interaction period and infant birth status," *Infant Behavior and Development*, vol. 37, no. 3, pp. 258–267, 2014.

[8] M. Tomasello, C. Moore, and P. Dunham, "Joint attention as social cognition," *Joint attention: Its origins and role in development*, pp. 103–130, 1995.

[9] M. L. Rowe and S. Goldin-Meadow, "Early gesture selectively predicts later language learning," *Developmental science*, vol. 12, no. 1, pp. 182–187, 2009.

[10] C. Yu, D. H. Ballard, and R. N. Aslin, "The role of embodied intention in early lexical acquisition," *Cognitive science*, vol. 29, no. 6, pp. 961–1005, 2005.

[11] T. Stivers and J. Sidnell, "Introduction: multimodal interaction," *Semiotica*, vol. 2005, no. 156, pp. 1–20, 2005.

[12] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions." in *Interspeech*, 2011, pp. 3081–3084.

[13] J. D. Matarazzo, M. Weitman, G. Saslow, and A. N. Wiens, "Interviewer influence on durations of interviewee speech," *Journal of Verbal Learning and Verbal Behavior*, vol. 1, no. 6, pp. 451–458, 1963.

[14] K. Shockley, M.-V. Santana, and C. A. Fowler, "Mutual interpersonal postural constraints are involved in cooperative conversation." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 29, no. 2, p. 326, 2003.

[15] T. Gerholm, "Children's development of facework practicesan emotional endeavor," *Journal of Pragmatics*, vol. 43, no. 13, pp. 3099–3110, 2011.

[16] M. Soderstrom, "Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants," *Developmental Review*, vol. 27, no. 4, pp. 501–532, 2007.

[17] D. L. Grieser and P. K. Kuhl, "Maternal speech to infants in a tonal language: Support for universal prosodic features in motherese." *Developmental psychology*, vol. 24, no. 1, p. 14, 1988.

[18] A. Fernald and T. Simon, "Expanded intonation contours in mothers' speech to newborns." *Developmental psychology*, vol. 20, no. 1, p. 104, 1984.

[19] P. K. Kuhl, J. E. Andruski, I. A. Chistovich, L. A. Chistovich, E. V. Kozhevnikova, V. L. Ryskina, E. I. Stolyarova, U. Sundberg, and F. Lacerda, "Cross-language analysis of phonetic units in language addressed to infants," *Science*, vol. 277, no. 5326, pp. 684–686, 1997.

[20] U. Sundberg and F. Lacerda, "Voice onset time in speech to infants and adults," *Phonetica*, vol. 56, no. 3-4, pp. 186–199, 1999.

[21] J. Van de Weijer, "Language input to a prelingual infant," in *the GALA'97 Conference on Language Acquisition*. Edinburgh University Press, 1997, pp. 290–293.

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

[23] D. N. Stern, S. Spieker, R. Barnett, and K. MacKain, "The prosody of maternal speech: Infant age and context related changes," *Journal of child language*, vol. 10, no. 01, pp. 1–15, 1983.

[24] U. Marklund, E. Marklund, F. Lacerda, and I.-C. Schwarz, "Pause and utterance duration in child-directed speech in relation to child vocabulary size," *Journal of child language*, vol. 42, no. 05, pp. 1158–1171, 2015.

[25] E.-S. Ko, A. Seidl, A. Cristia, M. Reimchen, and M. Soderstrom, "Entrainment of prosody in the interaction of mothers with their young children," *Journal of child language*, vol. 43, no. 02, pp. 284–309, 2016.

[26] B. J. Anderson, P. Vietze, and P. R. Dokecki, "Reciprocity in vocal interactions of mothers and infants," *Child Development*, pp. 1676–1681, 1977.

[27] M. Jasnow and S. Feldstein, "Adult-like temporal characteristics of mother-infant vocal interactions," *Child development*, pp. 754–761, 1986.

[28] T. Gerholm and L. Gustavsson, "The MINT-project," In preparation.

[29] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "Elan: a professional framework for multimodality research," in *Proceedings of LREC*, vol. 2006, 2006, p. 5th.