1-1-2018

Reactions of adult listeners to infant speech-like vocalizations and cry

Hyunjoo Yoo

REACTIONS OF ADULT LISTENERS

TO INFANT SPEECH-LIKE VOCALIZATIONS AND CRY

by

Hyunjoo Yoo

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

Major: Communication Sciences and Disorders

The University of Memphis

August 2018

Dedication

This dissertation is dedicated to my family, friends, advisors and colleagues.

Acknowledgements

This work was supported by grants from the National Institutes of Health R01 DC011027 and R01 DC006099 awarded to D. Kimbrough Oller.

Portions of this work were supported by grants from the American Hearing Research Foundation (AHRF) and American Academy of Audiology Foundation (AAF) awarded to Gavin M. Bidelman.

PREFACE

Chapter 2 has been resubmitted as a manuscript to Frontiers in Psychology. Its authors are Hyunjoo Yoo, Dale A. Bowman, and D. Kimbrough Oller.

Chapter 3 is in the process of resubmission as a manuscript to Infant Behavior and Development. Its authors are Hyunjoo Yoo, D. Kimbrough Oller, and Gavin M. Bidelman.

Chapter 4 is to be submitted as a manuscript to PLoS ONE.

Abstract

Yoo, Hyunjoo. PhD. The University of Memphis. August 2018. Reactions of adult listeners to infant speech-like vocalizations and cry. Major Professor: D. Kimbrough Oller, Ph.D.

Caregiver-infant interaction is critical for cognitive, social, emotional, and language development. This dissertation investigated adult responses to infant speech-like (i.e., protophones) and distress vocalizations in three individual projects. Study1 investigated different timing of caregiver responses to protophones and cries. In order for caregivers to respond differently to protophones and cries, they need to be able to differentiate these sounds. Study 2 and Study 3 projects addressed this issue.

Infant recordings from a longitudinal study were used for the dissertation. For Study 1 and 3, all-day LENA home recordings were used, and for Study 2, both LENA and laboratory recordings were used. Adult listeners for Study 2 and 3 were students and/or staff in the School of Communication Sciences and Disorders. Pupillometry and reaction time were used in Study 2 to measure listeners' cognitive load when judging infant vocalizations.

Study 1 found that caregivers tended to take turns with protophones, suggesting they viewed protophones as conversational material, while they tended to overlap with cries from the first months of life. This result is important because it suggests parents know that protophones are precursors to speech even in the first months of life, whereas cries express distress, and caregivers intuitively treat them as not being conversational material.

Study 2 found that nonparent adult listeners were reliably able to identify high-distress wail cry and mid-distress whine. Listeners judged cry faster in a speech-babble noise condition than in a no-noise or a music-masking condition, a pattern consistent with the fast-guessing principle. Greater pupil dilation was found when listeners identified whine than when they identified cry in the noise condition, suggesting there was greater cognitive load in the noise condition.

Study 3 documented that 39 listeners agreed with each other highly in rating levels of distress in infant vocalizations ranging from cries to protophones. The study also showed that moments of the long-term average spectrum in vibratory regimes within utterance, utterance duration, number of acoustic regimes, and maximum f_0 were strong predictors of the ratings of levels of distress. In addition, regardless of experience in infant vocalization coding, listeners were not significantly different in perceiving the level of distress.

Table of Contents

List of Tables

List of Figures

Figure Page

Chapter 1: General Introduction

Caregiver-infant vocal interaction has been examined for several decades, on the assumption that it is critical to build a bond of attachment between parent and infant to support cognitive, social, and language development in later life (Bowlby, 1958, 1969). Timing patterns between parent and infant utterances and correlations between adult perception and acoustic features of infant utterances were addressed in this dissertation; these are important issues in infant speech and language development.

Study 1 investigated timing of parent responses to infant speech-like vocalization (i.e., protophones) and cries. If parents respond by alternating their vocalizations with protophones and overlapping their vocalizations with cry, the results would support the idea that parents intuitively treat infant protophones as communicative vehicles, providing a frame for protoconversation, whereas they attempt to soothe the infant who cries.

Study 2 utilized new methods not previously involved in cry research, reaction time and pupillometry to measure listener reactions to cries and whines. This work provides grounding for new types of systematic investigation of reactions of adult non-parents to infant distress sounds. Different reactions of adults to cries and whines as seen through these new measures may help to more completely characterize perception of vocal distress.

Study 3 addresses definitional questions more directly by investigating acoustic features that contribute to the perception of distress or lack of it in high-distress wail cries, no-distress vocants and intermediate-distress whines. The results of this study may help 1) provide clearer definitions of infant vocalizations based on acoustic parameters, 2) lead to deeper understanding of the process of perceiving distress levels in infant vocalization, and 3) lay groundwork for future work on automatic algorithms to identify various types of infant vocalizations.

Chapter 2: The origin of protoconversation: An examination of caregiver responses to cry and speech-like vocalizations

Introduction

Overview of the Present Effort

The importance of early caregiver-infant interaction in cognitive, social and language development has been well documented for decades (Ainsworth and Bell, 1974; Beckwith et al., 1976; Bornstein and Bruner, 2014; Feldman, 2007a, 2007b; Jaffe et al., 2001; Murray et al., 1996). The research has emphasized the sense in which early turn taking vocal interactions provide a basis for emotional bonding (Ainsworth, 1979; Bell and Ainsworth, 1972; Blehar, Lieberman, and Ainsworth, 1977; Keller et al., 1996; Völker et al., 1999), a protoconversational frame, and a foundation sociality and for speech communication (Bateson 1975; Goldstein, King, and West 2003; Papoušek, 1995; Trevarthen 1977; Tronick, Als, and Brazelton 1980;). However, there has been a remarkable gap in this literature in that it has ignored the timing of caregiver responses to infant cries, focusing instead on timing of contingent patterns of response to speech-like vocalizations (i.e., protophones, Oller, 2000). The gap is especially notable considering the fact that infants produce *both* protophones and cries from birth (Dominguez et al., 2016; Jhang and Oller, 2017; Keller and Schölmerich, 1987; Nathani-Iyer et al., 2006). Stern et al. (1975) speculated that caregiver responses to cry would tend to overlap rather than alternate, as they had been shown to do with speech-like sounds. Empirical research on this previously unstudied speculation is important because it could illustrate that caregivers express an intuitive awareness of protophones as potential speech material by taking turns with them, while at the same time treating cries differently, speaking over them. The present study aims to systematically investigate timing of caregiver utterances in response to both protophones and cries.

Both very early precanonical protophones and later canonical syllables are foundations for speech (Koopmans-van Beinum and van der Stelt, 1986; Oller, 1980, 2000; Oller et al., 2013). However, compared to canonical syllables, precanonical protophones show far less obvious speech-like characteristics. The present research targets caregiver responses to the earliest precanonical protophones at 0, 1 and 3 months of age, affording the opportunity to evaluate the possibility that caregivers intuitively know protophones are precursors to speech even from the first months of life and treat them as such in the earliest interactions.

The research provides a new perspective on caregiver-infant interaction, because the data are derived from all-day recordings in infant homes. Prior research has almost entirely been conducted in structured settings where caregivers and infants have been expected to interact for the recordings. In these settings, with caregivers and infants always in the same room, caregivers have usually responded to infant vocalizations at very high, and presumably unrepresentative rates (see review in Fagan and Doveikis, 2017). Our approach should provide a more representative portrayal of both rates and timing of interactions.

Vocal Turn-taking in Conversation

In conversation, human adults contingently interact with each other and overwhelmingly take turns (Abney, 2016; Clayman, 2013; Hayashi, 2013; Sacks et al., 1974; Sidnell and Stivers, 2012). Levinson (2016) has suggested several reasons why investigating the turn-taking system in conversation is important both in adults and in parent-infant interaction, and thus why the turn-taking system has drawn increasing attention in the field of psycholinguistics and conversation analysis. The turn-taking system has universal characteristics that allow researchers to evaluate human predispositions and capabilities that are fundamental to language acquisition and language processing (Levinson, 2016; Levinson & Torreira, 2015). However, it has been frequently reported, particularly in the field of anthropology, that there are culture-specific features in human communication (Brown, 1998; Stross, 1972; Tanaka, 1999). For example, although systematic quantification has not been provided, speakers in the Nordic countries have been reported to be silent and to tend to interpose long silences between turn transitions. Long silences between turns may require “tolerance of silence” in American speakers (Lehtonen and Sajavaara, 1985, p. 279). Gender-specific features have also been investigated (Maltz and Borker, 1982; Coates, 1994, 1997). The research indicated that female friends were more likely to overlap or take turns without a gap than male friends. In other words, the collaborative floor (termed the “all-in-together mode”) was found to be more common in conversations between female friends.

Not only adult communication, but also caregiver-infant communication has been investigated to examine cross-cultural variations. Indeed research has suggested that features of parenting or caregiver-infant interaction vary cross culturally (Fogel et al., 1988; Kärtner et al., 2010; Keller et al. 2005; Rabain-Jamin and Sabeau-Jouannet, 1997; Richman et al., 1992). For example, Rabain-Jamin & Sabeau-Jouannet (1997) reported that French mothers tended to interact with their infants in dyads whereas Senegalese mothers (Wolof speaking) frequently included additional conversational partners.

However, a growing body of research has reported relatively universal characteristics of human interaction, particularly focusing on rapid turn-taking (Stivers et al., 2009; Heldner and Edlund, 2010). For example, Stivers et al. (2009) have provided empirical evidence reporting that speakers in 10 different languages (including the Nordic countries) showed similar latencies (around 250 ms) in response to questions, although there were subtle differences across languages. Wilson and Wilson (2005) also claimed that turn-taking patterns are similar regardless of cultures or social classes.

Rapid turn-taking between conversational partners is a remarkable feature given that one must comprehend, plan to produce and predict when to begin talking, while listening to the other's speech (Levinson, 2016). Obviously, rapid turn taking between speakers requires quick cognitive processing, considering that it takes at least 600 ms to prepare a single word production (Indefrey and Levelt, 2004; Indefrey, 2011). Sacks et al. (1974) systematically characterized turn-taking as a primary pattern in conversation. Other researchers have reported timing (or lags) of turn-taking, indicating that short latencies within hundreds of milliseconds are overwhelmingly common in conversation (Heldner and Edlund, 2010; Levinson and Torreira, 2015). Recent studies have attempted to examine the complex cognitive processing (e.g., prediction of the end of the utterance) that occurs in preparation for rapid turn transitions. Bögels and Levinson (2017) reviewed neurocognitive studies (e.g., brain imaging and electroencephalography) showing that listeners immediately recognized speech acts (such as statements or questions) and planned to produce speech for the next turn while listening.

To demonstrate that the turn-taking system is fundamental to human communication, it is important to investigate whether caregivers and infants show similar turn-taking patterns in vocal interaction (Levinson, 2016). If turn-taking occurs in the earliest interactions, does it show timing similar to that of more mature interactions? Addressing this question will help clarify how conversation emerges in development. And by considering possible differences in timing of parent responses to cries and protophones, we may illuminate the nature of parent awareness of the protophones as potential conversational material very early in life.

It is noteworthy, of course, that turn-taking is not the only way that speakers interact. Sometimes speaking in unison occurs both in adult conversation and in parent-infant interaction (Stern et al, 1975). The function of speaking in unison has been speculated to be associated with various circumstances, including high arousal expressions of coordinated action or thinking or of discord. In the present work, the analysis focuses only on the extent to which unison (or overlapping vocalization) and alternation between parents and infants reflects differences in how parents react to cries and protophones in the first three months of infant life. Ultimately of course it will be desirable to address the functions of overlapping and alternating talk as well as nonverbal behaviors under a single umbrella of theory that differentiates a wide variety of possible functions of coordinated rhythms in interaction.

Development of the Turn-taking System: Focus on the Protophones

Early caregiver-infant vocal interaction has been reported to surprisingly resemble conversation in mature languages (Bateson, 1975; Jasnow and Feldstein, 1986; Papoušek, 1995). Caregiver-infant interaction has been investigated for decades because it has been suggested to influence infant cognitive, emotional, and language development (Bloom et al., 1987; Goldstein et al., 2003, 2009; Jaffe et al., 2001). Researchers have provided evidence that even before speech, caregivers and infants show turn-taking patterns, and this vocal interaction in early infancy has been called "protoconversation" (Bateson, 1975; Trevarthen and Aitken, 2001). For example, Bateson (1975) showed early mother interaction with infants as young as the second month of life in various modalities including gaze and vocalization. After Stern et al. (1975) suggested two different modes of communication in mother-infant dyads, representing coaction (simultaneous or overlapping talk) and alternation (turn taking), researchers attempted to find a transition between the two. It was seemingly assumed by some that there might be a developmental trajectory of the two modes in dyads, with coaction preceding alternation. Similarly it seemed to be assumed that the mother might be primarily responsible for the appearance of vocal interaction at the youngest infant ages, while the infant might need to learn to be an active turn-taker (Miura et al., 2007; Ishihara et al., 2009). To explain how the mother could create the appearance of bilateral interaction at very young ages, consider the possibility that she can anticipate the *offset* of infant utterances (that are produced endogenously) and respond to them, and further that she can anticipate the *onset* of infant utterances and speak before them. In one study, vocal turn-taking was reported to be increased between 12 and 18 weeks of age after overlapping between 7 and 13 weeks (Ginsburg and Kilbourne, 1988). This study has been cited many times in an attempt to argue that infants are more likely to overlap with caregivers in early months and gradually to develop turn-taking capability. The study has sometimes been interpreted to suggest that the mother drives (with limited success) most of the apparent interaction at the youngest ages, and that the baby learns to interact actively with experience, resulting in more consistent alternation of mother and infant voices at older ages. Interpretation of the study is, however, hampered by its small number of dyads (3) and high variability among them, as well as the small number of interactive samples and range of circumstances of interaction that were observed.

A recent study attempted again to investigate developmental trajectories of turn-taking in caregiver-infant interaction. Hilbrink et al. (2015) investigated developmental trajectories of mother-infant interaction with infants ranging from 3 to 18 months of age. The authors reported that infants between 3 to 5 months produced more than 40% of their turns in overlap with caregivers, while this proportion of overlap decreased after 5 months and dropped to around 20% at 18 months. Turn-taking patterns were present from 3 months through 18 months, and only gap durations were different depending on ages. Gratier et al. (2015) also attempted to investigate developmental courses and showed that around 30% of infant vocalizations involved in turn-taking were overlapped with maternal vocalizations both at 8-13 weeks and at 17-21 weeks. In the Gratier et al. work, turn-taking patterns did *not* increase in older infants. Lavelli and Fogel (2002) conducted a longitudinal study on communication through gaze and facial expression between 1 and 14 weeks and found significant developmental changes around 2 months. The authors emphasized that critical neurodevelopmental changes occur at 2 months of age, and that most studies on turn-taking have investigated infants after this critical period. We note the important exception of Dominguez et al. (2016) who recently focused on infants at 2 to 4 *days* of age. These authors reported that 32% of infant vocalizations were overlapped with mothers' vocalizations. Surprisingly, when infants produced vocalizations that followed maternal vocalizations, about 70% were produced within 1 sec, the same time frame typical of older ages.

Taken together, researchers have reported consistent results in terms of presence (or early emergence) of turn-taking in protoconversation, even though many infant vocalizations are overlapped with maternal vocalizations (Bateson, 1975; Beebe et al. 1988; Elias et al., 1986; Gratier, 2003; Hsu and Fogel, 2003). However, the evidence is not conclusive about whether turn-taking increases and overlap decreases as a function of age. In addition, Stern et al. (1975) suggested that both coaction and alternation exist throughout life for different communicative functions, and thus coaction does not necessarily reflect an immature pattern of interaction. Their suggestion creates possibilities that interaction patterns may be different depending on functions of vocalizations. However, surprisingly, almost all prior research on early turn-taking has focused only on protophones and has ignored responses to cries.

Limitations in Prior Research: The Failure to Compare Responses to Protophones and Cries

Since language is primarily vocal, a key question in how vocal interaction develops concerns the nature of infant vocalizations themselves. We emphasize the distinction between early cries and vocalizations deemed to be precursors to speech, the protophones. One might imagine that these sounds would have been systematically differentiated in the study of early vocal interaction. In fact as far back as Stern (1975), it has been speculated, but not quantified, that caregivers may tend to speak simultaneously with cry as opposed to non-cry. Yet, despite decades of research in early caregiver-infant interaction, as far as we know, no prior research has explicitly provided a clear definition of distress vocalizations (e.g., fusses and cries) as opposed to protophones, and consequently no research has differentiated caregiver responses to these importantly different kinds of sounds. Instead, it has been simply mentioned in some research that infant distress/negative sounds (e.g., fusses, whimpers and cries) were excluded (e.g., Gratier et al., 2015; Hsu and Fogel, 2003). In other cases distress and non-distress sounds appear to have been grouped together without clear information about what the definitions were and how groupings were established (e.g., Bell and Ainsworth, 1972). Therefore, it has not been possible to determine what sounds have been included in most caregiver-infant interaction analyses.

In addition, although infants produce both cries and protophones from birth (Nathani-Iyer et al., 2006), most research appears so far to have attempted to investigate caregiver-infant interaction exclusively with speech-like sounds, which they have generally termed “non-distress” sounds (e.g., Hsu et al., 2001). Kaye and Fogel (1980) treated distress sounds somewhat differently from other studies, mentioning that “less extreme fussiness was considered a normal part of the interaction" (p. 455). Still, the authors' criteria for identifying fussiness were vague. In the absence of clear definitions (differentiating non-distress vocalizations as opposed to distress vocalizations), it is not clear exactly what sounds have been included under the heading "non-distress".

We propose that a clear distinction between protophones and distress sounds is critical for the study of caregiver-infant vocal interaction because it makes sense (in accord with the opinion of Stern) to imagine that caregivers will interact differently with the different sounds, since protophones are presumable precursors to speech (and are thus amenable to conversation), while distress sounds may be antithetical to conversation. It is nonetheless important to recognize that infant cries can play a role in establishing attachment with caregivers, which is fundamental to infant social, cognitive, and language development (Ainsworth and Bell, 1974; Bell and Ainsworth, 1972; Sroufe and Waters, 1977). Thus it makes sense to explore caregiver-infant interaction with *both* protophones and cries.

Another key limitation in prior studies on caregiver-infant interaction is that they have been overwhelmingly conducted in artificial structured settings (either in a laboratory or home). Mothers have been asked to interact with her infants with (or without) staff observing only during a brief artificially designed period, usually less than 10 minutes (review in Fagan and Doveikis, 2017). In such structured settings (with staff observing during brief periods), mothers, and infants may not interact naturally, and thus it may not be possible for researchers to obtain representative data. While interaction in well-defined laboratory circumstances is a legitimate target for research, it is also important to evaluate vocal interaction in the totally natural environment of the home. In that environment there are many differences from laboratory sampling. For example, parents are often not in the same room with infants at home, whereas in laboratory research they are usually in the same room with the infant and are expected to interact face-to-face. There is presumably a much reduced such expectation in the context of all-day home recordings. The purpose here is not to *compare* parent-infant interaction between structured and naturalistic settings but merely to present data from all-day home recordings, which we presume to provide a maximally naturalistic characterization that may reflect more representative and valid interactions.

Rationale for the Present Study

In the present study, we pursued the question of the origin of vocal interactivity by investigating the timing of caregiver vocalizations in the hope of illuminating whether (or how) caregivers play a role in controlling or scaffolding vocal interaction. Infants produce both protophones and cries from birth and those vocalizations operate as vehicles for possible interaction with caregivers. Protophones are known to be precursors to speech while cries express distress. Our study evaluates, for the first time, the relative timing of caregiver vocal responses to protophones and distress sounds (e.g., cries and whimpers)[1]. If caregivers tend to take turns with protophones, while speaking simultaneously with cries and whimpers, we can argue that caregivers intuitively treat protophones in a way that allows infants to begin to learn about conversation. Research has so far failed to show caregivers' systematic responses to protophones as opposed to cries because prior research has largely ignored caregivers' interaction with cries. Moreover, no prior interaction research has provided systematic and clear criteria for identifying distress as opposed to protophone sounds.

[1] Cries can be subcategorized into high distress wail cries and lower distress whimpers (sometimes called "fuss" in the literature). In the present work, we did not in the original coding differentiate these cry types, but coded both types as cries.

We investigated timing of caregiver vocalizations in response to infant protophones as opposed to cries specifying acoustic/auditory criteria to differentiate protophones from cries. In addition, to evaluate the origins of the human tendency and learning pattern for interactivity, we sought representative data from the natural interactive setting. We made all-day recordings in the home and selected periods with naturally-occurring high volubility and interactivity. Our approach allowed sampling from entire days of home recording. By using this approach, we hoped to provide maximally representative data on vocal interaction, and to illuminate the beginnings of human conversation.

Methods

Participants

12 infants contributed data for the present study: 9 infants at 0 months and 10 infants at both 1 and 3 months. Among the 12 infants, 7 were fully longitudinally with data available at all three ages (see Appendix A). All infants were Caucasian from English-speaking environments, mid to low-mid SES, and typically developing with no known risk factors.

All the infants were part of a longitudinal study of vocal development on typically developing infants. Parents of the infants were recruited through child-birth education classes and word of mouth for the longitudinal study. Interested individuals were given a consent form and questionnaire. Families returning the questionnaire and meeting inclusion criteria were contacted for an interview. All procedures were approved by The University of Memphis Institutional Review Board for the Protection of Human subjects.

Recordings and Recording Procedure

The battery-powered, palm-sized LENA recorder was placed in the chest pocket of special infant clothing, with the microphone 7-12 cm from the infants' mouth. The recorder allowed us to investigate the naturalistic language environment conveniently with recordings up to 16 hours/day at high sound quality, 16 kHz sampling rate (Xu et al., 2008). Parents were instructed by laboratory staff about how to place and activate the LENA recorder in the pocket of infant clothing at home. The parents brought the recorder to the laboratory after completing recordings according to a prescribed schedule, and laboratory staff uploaded the recordings through the LENA software. Once recordings were uploaded, automated analysis through the LENA software provided an estimated rate of infants' speech-like vocalizations (i.e., protophones) during each 5-minutes.

As a part of the longitudinal study, there were LENA all-day home recordings available for most of the 12 infants at each of the ages of 0, 1 and 3 months, that is during the first, second and fourth months of life—29 recordings in all (see Appendix A where the table indicates the 7 missing recordings). In a prior effort, 34 five-minute segments from each infant had been selected for human coding for each of the 29 recordings (Oller et al., 2014; Yoo et al., 2014). In order to obtain representative segments across each day, 24 of the 34 segments had been selected at equal intervals across each recording day. The researchers had also chosen the 10 segments with highest volubility (infant vocalization count) for each recording based on the automated estimates of the LENA software. That is, we rank-ordered all the five-minute segments for the recording in terms of the counts of infant vocalizations estimated by LENA and selected the 10 segments with the highest counts.

All the selected segments (34 per infant per age) had been coded in real time by trained human coders. Given that there were 29 recordings, there were 986 coded segments available. Each infant utterance was categorized as a protophone (squeal, growl, vocant (i.e., vowel-like sound)), cry, or laugh. Coders also indicated in response to a questionnaire after coding each 5-

minute segment, how much of the time on a five-point scale, caregivers were talking to their infants.

To investigate caregiver *responses* to infant vocalizations in the present study, we selected 290 segments out of the 986 that had been previously coded: the selected segments were required to have 1) some infant-directed-speech (IDS), according to the questionnaire answered by coders at the end of each coding session, and 2) a high rate of protophone or cry as determined by the prior coding. We selected the 5 segments for each recording that had the highest protophone rates along with the 5 segments for each recording that had the highest cry rates (see Appendix A). This procedure constitutes a compromise between selecting completely random samples across the day (for maximal representativeness) and selecting for samples with sufficient numbers of infant vocalizations and parent responses to power our proposed analyses.

On the five-point scale of the questionnaire, "1" indicated that no one was talking to the infant during the 5-minutes and "5" indicated that someone was talking to the infant close to the whole 5-minutes. Segments that were marked "2" (less than half the time) or higher on the questionnaire were designated as candidates for selection. To avoid too many empty cells in the design, additional human listening was conducted to seek indications of IDS even in cases where the questionnaire responses had indicated 1 (no one talking to the infants). The original coding had been done in real time, and so the coders may have failed to notice some IDS. The new coding was conducted in repeat-listening (coders were allowed to listen to the same periods several times). 12% of the 290 selected segments were included in the study based on this additional human listening, which determined that there were indeed some IDS utterances in those segments where the questionnaire data had not indicated that IDS was present. Still, 18 out of the 290 segments (6.2%) had no cases of IDS responses to infant utterances (see Appendix B). See below for definition of IDS responses.

Coding and Measurement

The coding team consisted of 4 Masters students and 1 PhD student in Communication Sciences and Disorders. In several intensive training sessions (with the last author, who has trained coders in infant vocal development for more than 40 years) of about an hour and a half each, all coders were introduced to how to locate boundaries for infant protophones, infant cries and caregiver utterances in AACT (Action Analysis, Coding, and Training, Delgado, 1996) software according to coding criteria listed below.

After training, the 5-min segments were coded by the five coders in repeat-listening mode to locate onset and offset of each vocalization. This coding procedure allowed us to measure lag times between infant and caregiver vocalizations. To locate utterances, we applied the breath-group criterion suggested by Lynch et al. (1995). According to the criterion, one utterance consists of a vocalization occurring on one egress (one expiration) and a new utterance can begin after each inspiration. We used the breath-group criterion because speech is organized in groups of expiration accompanied by phonation and supraglottal articulation, and because this criterion has proven to yield better intercoder agreement than methods based on fixed time intervals of silences (Lynch et al. 1995).

In order to quantify temporal structure of caregiver vocal responses, we first needed to identify cry as opposed to protophones. Protophones are defined as flexibly produced vocalizations including vowel-like sounds, squeals, growls, and so on (Oller, 2000). Cry conveys distress and always expresses negative affect whereas protophones are considered to be precursors to speech, not being bound to specific affect (Oller et al., 2013; Scheiner et al., 2002).

Thus cries are bound to a fixed affective state (i.e., negative) whereas protophones are not bound in this way. Protophones can be produced with different affect (i.e., positive, negative, and neutral) on different occasions. For example, infants can produce squeal (high pitch) sounds with positive affect in a joyful state and the same sounds with negative affect in a distressed state. This variability in usage of protophones (but not cries) is called functional flexibility (Oller et al., 2013). The distinction in functional flexibility between cry and protophones is important because we hypothesized that caregivers would respond differently to cry and protophones. We reasoned that cry is a signal for eliciting caregiver attention and aid, whereas protophones may be more likely to elicit pure social interaction.

Coders were trained to recognize markers for cry in terms of intense nuclei, dysphonation, glottal bursts and catch breaths (Stark et al., 1975; Truby and Lind, 1965). Appendix C provides a few example spectrographic displays and accompanying waveforms. Very intense cries are easy to identify and agree upon. They tend to have very intense, long dysphonated nuclei. They sometimes include glottal bursts or catch breaths at the beginning or end of the utterance. Utterances with glottal bursts or catch breaths are sometimes interpreted as negative even though they have less intense or short nuclei. Coders were trained to recognize one such common negative sound, which we term whimper, as displayed in Appendix C. After this training we found excellent agreement among coders as reported below.

Each caregiver utterance was identified as being infant-directed speech (IDS) or adult-directed speech (ADS). These identifications were quite reliable, because they were based on special phonatory characteristics of IDS, and because the meaning of both IDS and ADS was often clear to the listeners. In fact, the meaningful content usually made it totally unambiguous whether the parent was talking to the baby or not (e.g., “oh, you’re the cutest little thing today” or “let’s change your diaper now”). IDS has often been called “motherese” or “baby talk” because (in addition to special meaningful content) it includes unique phonatory characteristics such as wide pitch range, high pitch, smooth intonation, and long duration per syllable. A recent study by Farran et al. (2016) reported that IDS utterances are identifiable with intercoder agreement > 0.9 as measured by Intraclass Correlation, and our data (see below) confirm very high agreement levels among coders. We identified each utterance of adults as IDS from parents, IDS from other adults, or ADS. For the purposes of the present study, however, only IDS from *parents* was used in determining timing relations with infant utterances.

Calculating Lag Time

To address the hypotheses of the present study, we measured how fast and how often caregivers responded vocally to infant vocalizations. We follow a tradition (based on the floor transfer offset, for review see Holler et al., 2015) where lag is treated as the relation between the offset of one vocalization and the onset of another within a limited frame. In our approach, positive lag occurs when caregiver vocal responses begin after infant vocalization offset (but within 5 sec). Negative lag occurs when responses begin before the infant vocalization is over. A positive lag can be viewed as suggesting turn taking, because there is no overlap.

Positive and negative lag values were measured in TF32, a flexible real-time acoustic analysis program with both waveform and spectrographic displays (Milenkovic, 2015). Cursors were placed at the beginning (onset) and end (offset) of each infant vocalization, and at the onset and offset of each caregiver IDS utterance, using the waveform displays supplemented (especially in cases of overlap) by narrow-band spectrographic displays that facilitated discrimination between the caregiver and infant voices. For the purposes of the present study, we only included the first caregiver responses within 5 seconds of infant vocalization offset. In

Figure 1, we illustrate the principles for determining lags of caregiver vocal responses. We emphasize that each event represented by a green or purple box is an utterance (vocalization), defined by the breath-group criterion (see above).

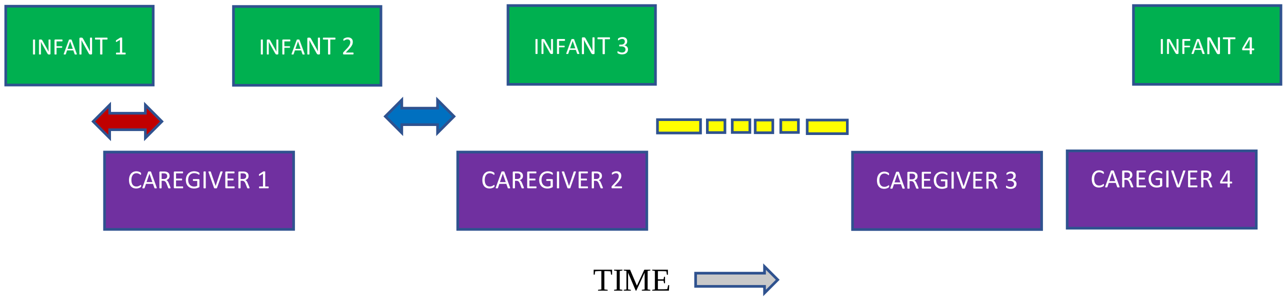


Figure 1. Calculating lags, as the relation between offset of infant utterances/vocalizations to onset of caregiver utterances/vocalizations. Green blocks represent 4 infant vocalizations arranged in time. Purple blocks represent 4 caregiver vocalizations arranged in time. The red arrow indicates that infant vocalization 1 is overlapped with caregiver vocalization 1, showing negative lag. The blue arrow, on the other hand, shows alternating of caregiver vocalization 2 with infant vocalization 2, positive lag. The broken yellow bar represents a time period longer than 5 sec. If a caregiver vocalization occurs >5 sec after the offset of an infant vocalization (as in the relation between infant vocalization 3 and caregiver vocalization 3), the caregiver vocalization is not defined as a response. Also, because caregiver vocalization 4 begins before the onset of infant vocalization 4, no vocal response to the infant vocalization is counted, even though the two vocalizations are overlapped. Similarly, caregiver vocalization 2 is a response to infant vocalization 2 but not to infant vocalization 3.

In accord with our method, a caregiver vocalization can be assigned as a response to one and only one infant vocalization, and an infant vocalization can only be assigned to one caregiver vocalization as a response. Consider caregiver vocalization 2 with respect to infant vocalizations 1 and 2; if the duration from the offset of infant vocalization 1 to the onset of caregiver vocalization 2 is less than 5 sec, then a decision must be made about assignment. First, caregiver vocalization 2 cannot be assigned to infant vocalization 1 because caregiver vocalization 2's onset is closer in time to the onset of infant vocalization 2 than to the onset of infant vocalization 1 and thus must be assigned to infant vocalization 2. In addition infant vocalization 1 must be assigned to caregiver vocalization 1 and thus leaves no option for caregiver vocalization 2 to be assigned to infant vocalization 1.

Coding and Measurement Agreement

For coder agreement tests, 28 out of the 290 segments were randomly selected: 6 segments at 0 months, 15 segments at 1 month and 7 segments at 3 months. Each of the 5 coders coded all the 28 segments in repeat listening mode (just as coders did during primary data collection), locating the onset and offset of each utterance of infants and caregivers. Intraclass Correlation Coefficients (ICC) were calculated to assess inter-rater agreement on the number of each vocal type (i.e., cry, protophone, and IDS). The average measure ICC for cries was .92 with a 95% confidence interval from .85 to .96 ($F_{(27, 108)}$= 85.2, p < .001). In case of protophones, the average ICC was .87 with a 95% confidence interval from .74 to .94 ($F_{(27, 108)}$= 61.2, p < .001). A high degree of inter-rater agreement was also found in identifying IDS. The average measure ICC was .93 with a 95% confidence interval from .88 to .96 ($F_{(27, 108)}$= 71.2, p < .001). Pearson correlations for each vocal type between all the possible pairings of coders were also calculated (M = .94, range: .89 to .98).

The temporal relation between infant and caregiver utterances is the primary research question of the present study, and so we determined the extent to which the coders identified similar patterns of relative timing between infant and caregiver utterances. With the 28 segments, we calculated mean response lags of caregiver utterances to infant cries as well as those to infant protophones (see Results).

Statistical Analysis

Generalized Estimating Equations (GEE) were implemented in R to model lag time as a function of various covariates. GEE models are an extension of Generalized Linear Models (GLM) (McCullagh and Nelder, 1989). GLM are useful to account for dependent variables (DVs) that do not meet the assumptions that DVs are normally distributed and linearly related to predictors. GEE were proposed by Liang and Zeger (1986) to account for correlated, in other words, nested or clustered, DVs. GEE models are also flexible for handling missing data as well as a variety of outcome variable distributions (Zeger et al., 1988).

As explained earlier, 5-min segments were selected based on rate of occurrence of infant protophones and cries that had been determined in the original coding from the prior study. This provision resulted in nesting (or clustering) of the data within each infant. In addition, 6.2% of the segments had no IDS, and thus constituted missing data. Also 5 of the 12 infants had no recording at least one age and thus the data were not equally balanced across the infants (see Appendix B).

Independent and dependent variables used in GEE models for the study are summarized in Table 1. Various combinations of covariates, including interaction terms, were tested to find a good model fit for the data and the variables in the final model, which had the following form: Lag = Age + Vocal Type + Duration (Infant vocalizations). This model was chosen because it was associated with the only significant effects. We initially tested Birth order on the assumption that first-born infants may receive more caregiver responses (Downey 1995), but this variable was dropped in the final model. Similarly we tested for Caregiver vocalization duration, because it seemed possible that infant vocalizations might be influenced by the duration of caregiver vocalizations. But again, this factor showed no notable effects on the dependent variable and was dropped in the final model.

Table 1. Variables used in the GEE Model

Variables		Description
Independent Variables	Age	Infant age in months
	Vocal Type	Vocal type of infant utterance: protophone or cry
	Duration (Infant Vocalizations)	Utterance duration of protophone or cry
	Duration (Caregiver Vocalizations)	Utterance duration of IDS
	Birth order	Birth order of each infant
Dependent Variable	Lag	Time difference between offset of infant utterance and onset of caregiver utterance

Results

Infant and Caregiver Vocalizations in Naturalistic Environments

The average percentage of infant utterances that were responded to with IDS in these segments selected from all-day recordings ranged for the three ages from 10 to 21% for protophones and from 13 to 17% for cries (Table 2, and for more details see Appendix B). In contrast, in laboratory studies with infants as young as 3 months, the percentage of infant utterances with responses has been much higher (generally more than 50% responses), presumably because in the laboratory, caregivers have usually been *instructed* to interact with infants and have stayed always in the same room with the infants (review in Fagan and Doveikis, 2017).

Table 2. Infant and caregiver vocalizations in the segments selected from the all-day recordings

Infant	Sum of IDS and ADS utterances	# of IDS	# of IDS responses		Mean proportions of IDS responses to infant vocalizations	
			To Protophones	To Cries	To Protophones	To Cries
0 months	2191	1697	778	355	.15	.17
1 month	1495	1493	626	259	.10	.13
3 months	2234	2234	1129	111	.21	.15

IDS = Infant-Directed Speech, ADS = Adult-Directed Speech

To see how much IDS was produced within the 5-min segments from all-day recordings, we summed durations of all IDS within each segment. Then, mean, median, min and max of IDS durations at each age were calculated (Table 3). On average about 10% of the time within the 5-min segments was occupied by IDS. In contrast prior results based on short-term recordings where caregivers have been instructed to interact with infants have shown from 40-70% of the time occupied by IDS (e.g., Kärtner et al., 2010; Gros-Louis et al., 2006).

In sum, caregiver responsivity was very different in our naturalistic environments compared to prior results obtained in structured laboratory environments. In our data caregivers tended to produce less IDS and consequently responded less to infant vocalizations than in studies where parents were instructed to interact with their infants in a laboratory (or even at home).

Table 3. Duration and percent of caregiver IDS within 5-min segments across infants at each age

	IDS Duration (sec) during 5 min				Percent (%) of 5 min			
	Mean	Median	Min	Max	Mean	Median	Min	Max
0 months	29.3	14.8	0	140.1	9.8	5	0	47
1 month	23.3	9.2	0	150.6	7.8	3	0	50
3 months	38.6	21.6	0	161.7	13	7.2	0	54

Temporal Structure of Caregiver IDS in Response to Protophones and Cries

Figure 2 shows proportions of IDS utterances in response either to protophones or cries in 1 second intervals referenced with regard to the offset of infant utterances after collapsing the data across ages. The vertical line in the Figure indicates the point of offset of infant utterances, “0” on the x-axis. Thus, percent of IDS utterances beginning in each interval after the offset is displayed right of the black line, and each interval in seconds is labeled “+” on the x-axis, indicating positive lag. Similarly, percent of IDS utterances beginning in each interval before the offset is displayed left of the black line, and values in seconds are labeled with a minus sign on the x-axis, indicating negative lags. The figure displays a range from <-2 sec to >+5 sec lag. Long negative lags were rare, as indicated in the figure, because infant utterances were usually not long enough to allow them. For data collapsed across all three ages, 71% of IDS responses to protophones began *after* the offset of infant utterances whereas 66% of IDS responses to cries began *before* the offset of infant utterances.

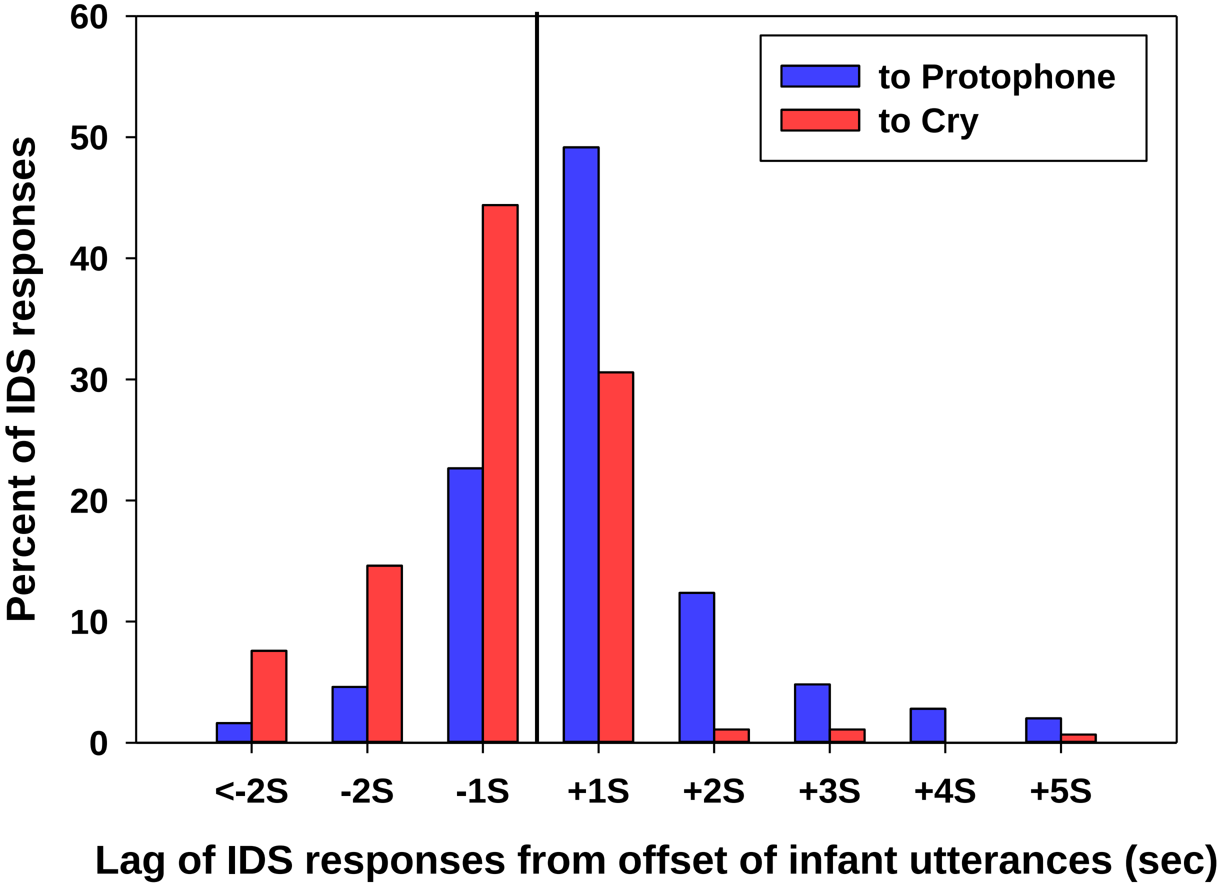


Figure 2. The black vertical line represents the offset of infant utterances, the 0 point in time. Percent of all IDS responses to cries and protophones is plotted for each 1-second interval before and after the 0 point. The display shows that IDS utterances in response to protophones tended to begin after the offset of infant utterances (positive lag), and especially in the 1 sec interval after. In contrast, IDS in response to cry tended to begin before the offset of infant utterances (negative lag), overlapping with them.

Distributions of the data in Figure 2 also show that IDS either to protophones or cries was heavily concentrated within the 1 second interval around the offset of infant utterances and became sparse as lags increased positively or negatively. Short latency of caregiver responsivity has been suggested by Papoušek and Papoušek (1987) and Keller et al. (1999), although neither prior study nor any other prior one to our knowledge has distinguished between lags of responses to protophones and cry. The present study confirms previous findings overall, but adds the clarification that a preponderance of responses occurring in the first second *after* offset of infant vocalizations applies to protophones, but *not* to cries. This pattern of results applied to all the coders in the agreement data. For the 28 segments that were coded by all of them, the mean lag for each of the coders was positive and occurred within the first second after the infant offset (in fact the first half second) for protophones, and mean lag was negative and occurred within the first second before the infant offset for cries (Table 4).

Table 4. Coder agreement on response lags to infant vocalizations

	Mean response lags to protophones (ms)	Mean response lags to cries (ms)
Coder 1	463	-133.94
Coder 2	446.25	-229.62
Coder 3	423.82	-334.7
Coder 4	362.75	-173.54
Coder 5	417.25	-229.53

Breaking the data down by age, as shown in Figure 3, a similar distribution of lags to protophones and cries was observed at each of the three ages, with higher proportion of responses near the offset of infant utterances at all ages. Caregivers responded to protophones mostly after the offset of infant utterances whereas they responded to cries mostly before the offset of infant utterances. At 0 months, IDS responses to *protophones* occurred in 71% of the cases after the offset of infant utterances, whereas IDS to cries occurred 69% before the offset of infant utterances. At 1 month, IDS to protophones occurred 74% after the offset of infant utterances, whereas IDS to cries occurred 67% before. At 3 months, IDS to protophones occurred 69% after the offset of infant utterances whereas IDS to cries (which occurred very infrequently at 3 months) occurred 57% before the offset of infant utterances.

A possible artifact in the data needs to be considered. Namely, cries in the data were more than twice as long on average as protophones[2]. Could it be that the tendency for IDS to overlap with cries more than with protophones was an artifact of this difference in mean durations? To test for this possibility we segregated the data for both cries and protophones into 500 ms bins[3], and plotted proportion of overlapped to alternating IDS (the ratio of overlapped caregiver responses to alternating caregiver responses) as shown in Figure 4. Regardless of duration of infant utterances, caregiver responses overlapped more often with cries than with protophones. The pattern applied at all ages and at all durations (Figure 4). The statistical significance of the tendency for alternation to protophones as opposed to overlap with cries was tested by chi-square, with significant findings in 9 of the 12 comparisons (Table 5). The analyses suggest that the duration differences between cries and protophones was not responsible for the differentiation in IDS lags for cries and protophones. On the other hand, duration was not irrelevant in the pattern of IDS responsivity. The maximum difference in the ratios in Figure 4 was observed for the longest utterances (>1.5 sec), both for cries and protophones, and in general there was a tendency for more overlap of IDS at longer durations. Thus the data suggest that the longer the infant utterance (whether protophone or cry), the less likely caregivers were to produce their IDS response after the infant utterance was finished.

[2] Cries in the data were more than twice as long on average as protophones (0 mo: Prot = 742 ms, Cry = 1709 ms; 1 mo: Prot = 660 ms, Cry = 1664 ms; 3 mo: Prot = 1034 ms, Cry = 1697 ms).

[3] We collapsed infant utterances into 500 ms utterance duration groupings or “bins” for the analysis in Figure 4. That is, all infant utterances less than 500 ms were collapsed into one bin, all utterances between 500 ms and 1 s were grouped together in another bin, and so on. After collapsing them into these 500 ms bins, timing of caregiver responses was determined for each bin at each age and displayed as the ratio of the number of caregiver responses overlapping with infant utterances over the number of caregiver responses alternating with (following) infant utterances. By doing this, we tested whether duration of infant utterances affected the tendency for caregivers to overlap vocalization with cries and to alternate with protophones.

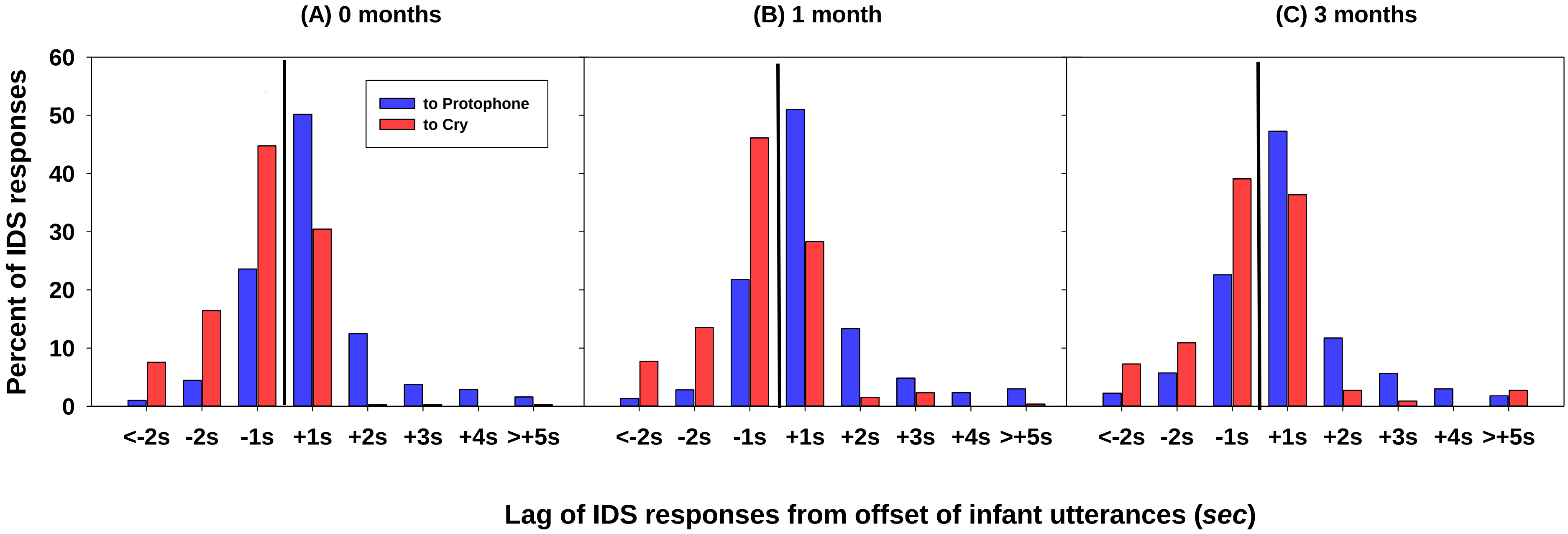


Figure 3. As in Figure 2, the black vertical line represents the offset of infant utterances. The display shows that IDS in response to protophones tends to being after the offset of infant utterances (positive lag), and especially in the 1 sec interval after. In contrast, IDS in response to cry tends to begin before the offset of infant utterances (negative lag), overlapping with them. This pattern is consistent at each age.

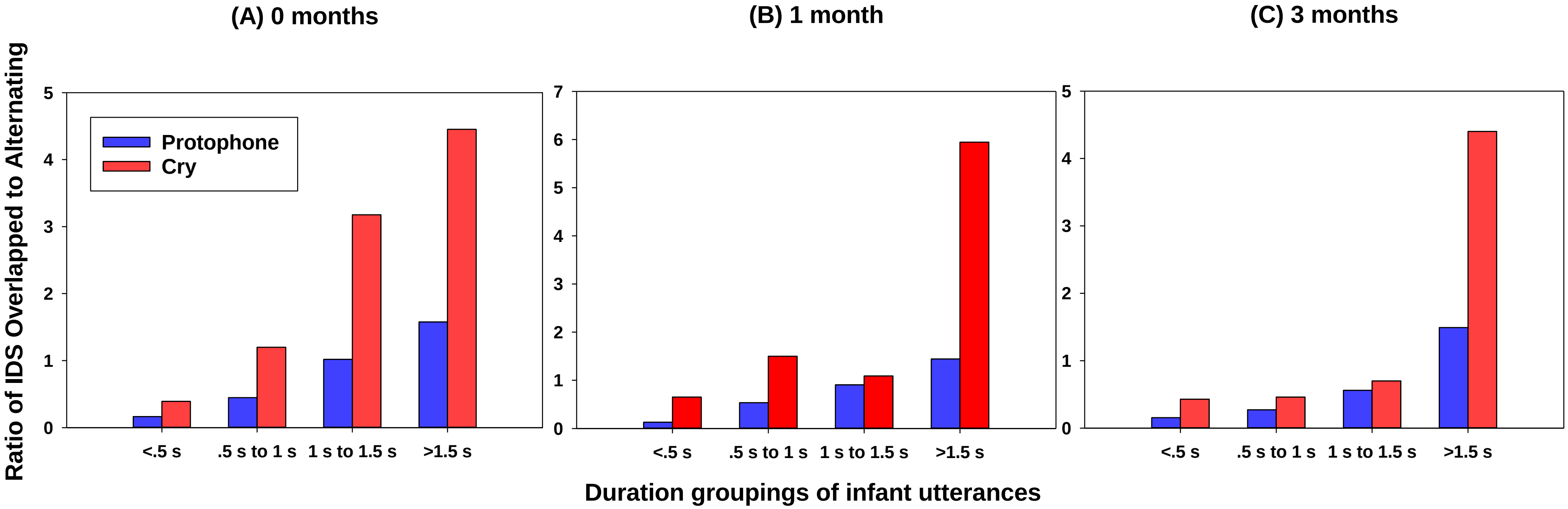


Figure 4. Degree of overlap/alternation of caregiver responses to protophones and cries in groupings of .5 sec (i.e., 500 ms bins). The display shows that regardless of duration of infant utterances, either cry or protophones, caregivers tended to respond to infant cry with more overlap (higher ratio of overlapped/alternating) than to protophones. Conversely, turn-taking (lower ratio of overlapped/alternating) tended to occur to a greater extent with protophones than with cries at all durations of utterances. The display also shows that ratios of overlap to alternation were higher at longer durations of infant utterances for both cries and protophones, with a very high ratio for long crie

Table 5. Chi-square statistics for alternation vs. overlap for protophones and cries at various durations of infant utterances

	< 500 ms	500 ms to 1s	1s to 1.5s	>1.5s
0 months	4.52*	16.44**	11.71**	14.71**
1 month	23.97**	8.84**	0.23	16.85**
3 months	4.35*	1.06	0.19	9.03**

*p < .05, **p < .01

A GEE model confirmed the predicted patterns of positive lag for IDS in response to protophones vs. negative lag for IDS in response to cries, taking account in the model for the clustered data. Among variables summarized in Table 1, age (0, 1 or 3 months), vocal types (cry vs. protophone), and duration of infant vocalizations (treated as a continuous variable) showed significant main effects, while birth order, age-vocal type interaction, and duration of caregiver IDS utterances were not significant in the model (Table 6). The GEE model predicted that as infant age increased, lag of IDS increased. With regard to vocal types, lags were positive when protophones were responded to with IDS but negative when cries were responded to. Duration of infant vocalizations showed a significant main effect in the GEE model. However, as shown in Table 6, since the coefficient of duration of infant vocalizations was extremely small, the duration of infant vocalizations was not, *in practical terms*, significantly associated with lag in the model.

Birth order did not show any significant main effect in the model. Birth order was included because prior research has suggested lower parent interaction with later borns (e.g., Downey, 1995) and because we observed that some recordings with low IDS were conducted with infants who had older sibling(s). Caregiver utterance duration was also not significant in the model. In other words, caregiver responsivity timing to infant vocalizations was independent of caregiver utterance durations.

These findings provided evidence that caregiver IDS in response to protophones showed a turn-taking pattern even at 0 months. However, caregivers responded much differently to cries, overlapping rather than taking turns. Importantly, the distinctively different interaction patterns from caregivers to cries and protophones were observed even at 0 months, and the patterns remained similar at all three ages.

Table 6. Significant parameters from the GEE analysis

	Coefficient	S.E	*p*-value
Intercept	-1.12	329.14	.997
Age	14.39	6.72	.032
Vocal Type (Protophones vs Cries)	363.59	38.41	<.0001
Duration of Infant Vocalizations	-0.64	.03	<.0001

Discussion

The development of vocal language appears to depend on both a capacity and an inclination of infants to vocalize plentifully and for caregivers to take advantage of those infant sounds to engage them in vocal interaction (Bruner, 1983; Bornstein and Bruner, 2014). Many have noticed the tendency of caregivers to interact with their infants vocally (Bell and Ainsworth, 1972; Keller et al., 1999; Richman et al., 1992), but a key opportunity to illuminate the process has not previously been exploited. The opportunity resides in the difference between cry sounds of the human infant and the precursors to speech, the protophones. We hypothesized, in agreement with Stern et al. (1975), that cry sounds should not elicit alternating caregiver vocal responses, because cry sounds are not the potential material of speech. To the extent that caregivers, even interacting with infants in the first month of life, intuitively alternate their vocalizations with protophones, but overlap their vocalizations with cries, they provide compelling evidence that human caregivers are predisposed to treat protophones as potential speech material long before infants are capable of speaking. Our results empirically confirm Stern's suspicion and our own, as caregivers were far more inclined to converse in alternating fashion with protophones than with cries.[4]

The results, we think, offer an enhancement to prior perspectives on the importance of early vocal interaction, because they illustrate that human caregivers must possess not only a capacity to recognize protophones as primitive speech material, but a predisposition to treat the protophones as such by interacting with them in a protoconversational way. The contrast in the way caregivers in our research reacted to protophones and cries highlights the fact that caregivers know, even if subliminally, that protophones offer a special opportunity to bond with the infant and to set the process of speech development on course.

[4] In the present work, we did not, in the original coding, differentiate cry types, but coded both wail cries and whimpers as cries. However, after all the data had been preliminarily analyzed, we conducted an additional round of coding just in order to differentiate wail cries and whimpers in our samples. Whimpers turned out to be a relatively small percentage of all cries at 0 and 1 months of age (17-18%) but represented ~48% of cries at 3 months. Note also in Appendix B that the occurrence rate of both wail cries and whimpers was dramatically lower at 3 months than at the other ages. The N of whimpers that were responded to by the parents was very small, <40 at each age, so the power of any analysis of them is very low. Nonetheless we computed mean lags of responses to wail cries and to whimpers and found that, collapsing data across the three ages, caregiver responses showed the expected pattern of more overlap to whimpers than to protophones, but the trend was not as strong as for wail cries. Further, the expected pattern occurred at all 3 ages for wail cries, but for whimpers it occurred only at 0 and 1 months.

We found that caregiver vocal responses to protophones were heavily concentrated in the 1 sec interval after the offset of infant protophones. This finding is consistent with the results of Keller et al. (1999), studying interactions with infants at 3 months, showing that maternal responses (verbal or nonverbal) occurred most frequently within the 1 sec after infant behaviors occurred. Papoušek and Papoušek (1987) suggested caregivers' contingent responses to infant vocalization occurred within 800 ms. Infants seem to be capable of perceiving contingency from birth (Gewirtz and Pelaez-Nogueras, 1992; Murray et al., 1985; Striano and Reid, 2006). According to Keller et al. (1999), "the experience of contingency allows the infant to develop expectations about behavioral occurrences …" (p. 475). Caregiver responses to the protophones thus appear to provide a rich learning opportunity. Of course coaction with parent and infant vocalization in unison does appear to occur on occasion even with the protophones. The pattern of coaction may reflect another function of interactivity that, although it occurs infrequently, may be of considerable importance in child development.

While many longitudinal studies have shown that protophones are foundations for speech (Koopmans-van Beinum and van der Stelt, 1986; Oller, 1980; Roug et al., 1989; Stark, 1980), some still assert that protophones develop *from* cries (Mampe et al., 2009; Takahashi et al., 2015), and thus imply that protophones are absent in the first months of life. In fact, however, infants produce *both* protophones and cries from birth (e.g., Nathani-Iyer et al., 2006). Moreover, the evidence shows that, protophones occur *more frequently* than cries, even in the first 2 months, and that the preponderance of protophones over cries increases to a ratio of perhaps 8 to 1 by 3 months and continues to expand thereafter. This evidence in itself suggests that failure to recognize the significance of protophones from birth may have misled prior theorists. The modern evidence suggests a massive endogenous tendency on the part of infants, from the beginning of life (Jhang and Oller, 2017; Nathani-Iyer et al., 2006; Oller, 2000), to explore the vocal capacity with protophones. Infant vocal exploration thus offers caregivers a basis for laying a frame for bonding and social interaction with infants and for protoconversation as an expression of the caregiver investment in the relationship with infants. Significant consequences for potential language learning seem obvious even if neither the caregiver nor the infant has any immediate awareness of the long-term significance of their interactions.

There exists persuasive empirical evidence that caregivers' intuitive interaction with these infant vocalizations is highly associated with cognitive and language development (Ainsworth and Bell, 1974; Jaffe et al., 2001;Lewis and Coates, 1980; Lewis and Goldberg, 1969). Surprisingly, however, cries and/or distress sounds have been almost entirely ignored in prior face-to-face interaction literature that has attempted to address the role of interaction in language development—responses to cries and fussing sounds have typically not been coded at all in such studies. Stern et al. (1975) had speculated that caregivers usually speak to infants simultaneously with their cries, and consequently had brought into focus the opportunity to illustrate the power of the protophones to elicit conversational reactions. But Stern's speculation requires that reactions to protophones be systematically contrasted with reactions to cries. Given his extensive influence on the literature, we are surprised that no empirical demonstration of this distinction in caregiver reactions has been made until the present work.

While the many prior results suggest that early protophones are foundations for speech, it is notable that their form is very distant from the form of speech, particularly because early protophones do not consist of well-formed ("canonical") syllables. Canonical syllables, not produced systematically until the second half year, have long been recognized as speech precursors, because there exists a clear continuity between canonical syllables and early meaningful speech—the types of syllables utilized in the two cases are very similar (Locke 1989; Oller et al., 1976; Stoel-Gammon, 1989; Vihman et al., 1985). And when the canonical stage begins, caregivers react not only by interacting with infants in protoconversation, but saliently by treating the canonical syllables as potential words (e.g., Papoušek, 1994). A canonical babble sequence [dada] can immediately be treated as "daddy", even though the infant presumably didn't intend it that way. In contrast, the early protophones are rarely if ever treated by caregivers as possible words.

As early as the 1970's the precanonical protophones were already recognized as being related to speech because of their tendency to include normal phonation (the kind of phonation that is overwhelmingly predominant in speech) and because the primitive articulation patterns that often accompany early protophones hint at a foundation for speech articulation (Stark, 1981; Oller, 1981; Zlatin, 1975). More recently, precanonical protophones have also been recognized as foundations for speech because they (unlike cries) possess functional flexibility, which is a fundamental property for all natural languages (Nathani-Iyer and Ertmer, 2014; Oller, 1981; Oller et al., 2013; Scheiner et al., 2002). The present results indicate that caregivers intuitively provide systematic conversational frames in response to precanonical protophones, even at 0 to 3 months, thus introducing the infant to the turn-taking system that characterizes most speech interaction.

In our data, caregivers responded to cries at about the same rate as to protophones (see Table 2), but there were many more protophones available for response, so the data consisted primarily of responses to protophones. A question that arises is why caregivers respond vocally to cries at all, since they are not natural speech material. Stern et al. (1975) contended that " ..mothers commonly vocalize simultaneously with the crying of their infants in order to soothe them" (p. 90). The idea finds partial support in the suggestion of Wolff (1965) that continuous sound (particularly white noise) can soothe neonates. Bell and Ainsworth (1972) reported that caregiver vocal responses (without touching the baby) to cries were the second most common responses to cries, following physical responses (pick-up and hold the baby). Interestingly, however, mere vocal responses to cries were found to be the least effective intervention to terminate cries. In the face of these results and interpretations it is not clear whether caregivers in prior work or in our own were using simultaneous speech over cries principally to soothe infants. This is a question that could be investigated productively with audio-video recorded interactions.

Another focus of our investigation is caregiver responsivity in a much more naturalistic environment than in most prior research on interaction. We found that caregivers tended to respond much less often to infant vocalizations in all-day recordings compared to prior research conducted in structured settings. On average caregivers responded in our study to 10-21% of infant vocalizations in 5-min segments. In contrast, Kärtner et al. (2010) reported that on average, mothers contingently responded to infant non-distress vocalizations at a rate of 47% in 10 min structured interactions. Gros-Louis et al. (2006) reported even higher maternal contingent response rates to infant vocalizations: 73% in 10 min play sessions. Fagan and Doveikis (2017) reported that mothers responded to about 30% of infant utterances in ordinary interaction at home, while they summarized prior literature suggesting laboratory rates in structured interactions of about 70%. Although Fagan and Doveikis did not obtain their data with all-day recordings, their motivation and results are consistent with ours. When mothers are instructed to interact, their voices often occupy a considerable portion of the total time of observation. Franklin et al. (2014) found that in face-to-face interaction with six-month olds in the "still-face" paradigm, mothers' speech occupied about 50% of the time. Dominguez et al. (2016) reported that mothers' speech occupied about 29% of the time in observations where their newborn infants were present and awake with them for 10 min. Farran et al. (2016) reported that mothers' speech occupied about 25% (during 10 min selected from home and laboratory recordings where mothers were expected to interact with their infants) in both Lebanese and American mother-infant dyads. In contrast Table 2 indicates that in our all-day home recordings only 8 to 13% of the time was occupied by caregiver responses to infant vocalizations.

Overall, the results suggest much lower rates of caregiver responsivity to infant vocalizations in our study than in laboratory studies, presumably because our interactions occurred in households where no experimenters instructed mothers to interact nor observed them doing it and where mothers were free to move about in various rooms in the house. We presume our results reflect more representative patterns of interaction, where caregivers in their natural environments choose convenient moments to interact with their infants, focusing on the special circumstance of interaction with protophones, fostering sociality, bonding, and laying groundwork for language.

Chapter 3: Behavioral and physiological responses of nonparent identification of infant cry and whine

Introduction

Infant Distress Sounds

Infants produce cry and cry-like vocalizations to express distress. Although both express negative affect, caregivers may respond differently to cry and cry-like sounds based on their degree of negativity. By definition cry is judged more negative than cry-like sounds. Cry and cry-like sounds should be possible to categorize on a continuum of negativity according to their distinctive acoustic features such as amplitude or spectral energy concentration (Yoo, Buder, Lee, & Oller, 2015). In a substantial literature, cry and cry-like sounds have been assumed to be self-evidently differentiable (Barr, Kramer, Boisjoly, McVey-White, & Pless, 1988; Mende, Herzel, & Wermke, 1990). Thus, there has been no systematic research investigating differentiability of cry and cry-like vocalizations. A few studies have, however, presumed differentiability without providing any clear criteria or operational definitions for cry and cry-like sounds (Barr et al., 1988; Petrovich-Bartell, Cowan, & Morse, 1982). The work that exists has primarily depended on listeners varying judgements based on their own interpretations of terms such as “cry” and “fuss”. As a result, it remains unclear whether (or how well) cry and cry-like vocalizations are reliably differentiable.

Differentiating infant cry and cry-like utterances is important because caregiver responses (which influence social and cognitive development) are surely at least partly based on the perceived degree of negativity (presumably related to urgency) of infant sounds. Investigating differentiability between cry and cry-like sounds will also be important in the future for developing an automated analysis system to differentiate types of infant vocalizations. To develop automated algorithms for detection of cry and cry-like utterances, one requires a perceptual gold-standard of human judgment for a range of distress vocalizations. Once automated tools for gauging the degree of negativity of infant sounds are in place, it should be much easier to develop clinically useful automatic assessments and perhaps even predict risk for disorders based on distress vocalizations.

In the present study, we aimed to investigate how well infant cry and cry-like vocalizations can be distinguished perceptually by naïve (nonparent) adult listeners. To our knowledge, there have been no systematic studies examining how (or even if) human listeners can reliably identify these salient infant vocalizations. In addition, in our study we aimed to develop quantitative methods for characterizing the perception of distress sounds of infants.

Importance of Perceiving Negativity of Infant Distress Sounds

Human infants seem endogenously capable of performing various behaviors that promote proximity to caregivers. These are termed attachment behaviors by Bowlby (1969). Bowlby (1969) suggested that attachment behaviors used to maintain proximity and contact with the primary caregiver contribute to building bonds between infant and caregiver. Ainsworth et al. (1978) suggested that maternal sensitivity to the infant's signals and moods highly influence the development of attachment. Such sensitivity tends to promote a secure relationship that allows the infant to balance between proximity to the mother and exploration of the environment (Ainsworth & Bell, 1970; Keller, 2003; Wolff & Ijzendoorn, 1997). The insecurely attached infant is portrayed as being limited in developing an attachment-exploration balance (Blehar, Lieberman, & Ainsworth, 1977; Crockenberg, 1983; Egeland & Farber, 1984). The reason researchers have paid so much attention to investigating the development of attachment is that secure attachment established in early life is an important predictor of later developmental outcomes (Arend, Gove, & Sroufe, 1979; Erickson, Sroufe, & Egeland, 1985; Kobak & Sceery, 1988).

Since human infants are slow in developing physical movement capabilities, vocal signaling may play a critical role in promoting proximity for the human infant by stimulating the caregiver to come nearer (Bell & Ainsworth, 1972). Among various types of vocalizations in human newborns, crying is especially powerful and is thought to be more effective in promoting proximity (especially during or shortly after crying) than other vocalizations (Murray, 1979). Considering the importance of caregiver-infant interaction in development, it seems plausible to hypothesize that when caregivers interact with their infants, they may react differently to their sounds depending on perceived negativity or aversiveness (Frodi & Senchak, 1990; Gustafson & Green, 1989; Zeskind, 1980). Since infants express distress sounds with varying degrees of negativity, caregivers' ability to differentiate degrees of negativity could play a critical role in patterns of interaction with infants.

Prior Studies on Parent Perception of Infant Cry

It is well known that adult listeners are capable of inferring affective states of speakers from their vocalizations (Frick, 1985; Scherer, 2003; Standke, 1992). A great deal of research on parental perception of infant crying has been focused on auditory differentiation of presumed types of cry (hunger cry, pain cry, anger cry, e.g., Zeskind, Sale, Maio, Huntington, & Weiseman, 1985) Researchers have also reported how individual differences in mothers (e.g., depressed vs. not depressed) can affect their perception of infant crying (Schuetze & Zeskind, 2001). Some studies have shown that acoustic features (e.g., fundamental frequency, signal energy) can predict auditory judgments of the degree of aversiveness of the various presumed types of cry (Gustafson & Green, 1989; Zeskind & Marshall, 1988). There is not full agreement in the literature about whether parental experience affects aversiveness judgements with regard to cry; Green, Jones, and Gustafson (1987) found parents and non-parents gave similar judgments, while others have found differences between parents and non-parents (Irwin, 2003; Leger, Thompson, Merritt, & Benz, 1996). More recently, researchers have attempted to investigate neurological reactions to infant cry (Kim, Feldman, Mayes, Eicher, Thompson, Leckman, & Swain, 2011; Swain, Tasgin, Mayes, Feldman, Todd Constable, & Leckman, 2008).

Only a handful of studies have attempted to investigate infant cry as opposed to cry-like sounds, the latter of which would presumably include sounds termed in the common parlance whine, fuss, moan, whimper, groan and perhaps some others. Barr et al. (1988) investigated cry and fuss and measured frequency of occurrence and duration of each category (i.e., cry and fuss). Mothers were asked to report six behaviors (i.e., sleeping, awake and content, fussing, crying, feeding, and sucking) for a day. The authors compared the parent reports on frequency and duration of cry and/or fuss episodes in their 6-week-old infants with laboratory transcribed results from recordings of the infants made on the same day. The transcriptions indicated how often "negative" vocalizations occurred (presumably collapsing across cry and fuss). They found that parent indications of cry were highly correlated with frequency of "negative" vocalizations as judged by transcribers of the recordings, but judgments of fuss were not. Based on the results the authors concluded that in a short-term period (all judgments pertained to a single day for each infant), using parent report may be useful in providing reliable data when investigating cry and fuss quantitatively.

The results of the study are, however, difficult to interpret because the definitions of cry, fuss and negative were vague. Fuss was referred to merely as: “not quite crying but not awake and content either” (Barr et al., 1988, p. 381). In addition, since the authors did not define cry in their study, the definition of fuss had no external referent, no grounding, and thus remained effectively undefined. In other words, without defining what is cry, there is no way to define fuss as distinct from it.

Another study focusing on both infant cry and “fuss” was conducted by Petrovich-Bartell et al. (1982). In the study, mothers were asked to rate their own infants’ vocalizations on a 5-point continuum ranging from fuss to cry (presumably a negativity judgment), without definition of what cry or fuss might mean. On the first visit the authors thought mothers used both acoustic cues and contexts in order to judge degrees of negativity of each sound. Three weeks later, mothers were asked to categorize randomly presented sounds from both their own infants and other infants. On this second visit, contextual cues were no longer available because parents were only presented with isolated audio stimuli. The authors found that mean ratings of their own infant’s sounds at the first and second visit were highly correlated, suggesting that mothers reliably rated vocalizations without contextual cues. They examined further whether ratings of vocalizations were associated with acoustic features. The authors found that mothers’ ratings (on a 5-point continuum ranging from fuss to cry) were correlated with acoustic measurements (e.g., intensity and mean frequency of Formant 2). This study suggested that mothers are reliable judges of a vaguely defined continuum from cry to fuss based on their infant distress sounds. By investigating acoustic features of vocalizations, the authors attempted to seek more objective factors that may influence listeners’ judgment of negativity in infant vocalizations.

A final work that is of interest even though it focused on two and three year olds was done by Green et al. (2011). The authors recorded tantrums and performed acoustic analysis on the largest number of negative vocal behaviors (scream, yell, whine, cry and fuss) that has been studied to our knowledge in childhood. The authors reported that the 5 vocalization types identified auditorily were also acoustically differentiated. Discriminant analysis with combinations of acoustic features (e.g., duration, total energy and energy band) reliably predicted the 5 categories. This study supports the idea that vocal affect expression can be perceptually differentiated. In addition, the study provides evidence that acoustic features can reflect different degrees of emotion in vocalizations. These conclusions encourage our pursuit of differentiation of negative vocalizations at even younger ages.

In our work with negative vocalizations of the first year, we suggest grounding definitions of cry and other cry-like sounds with reference to judgments of real utterances by experienced listeners, investigators of infant vocalizations. As suggested by the studies reviewed above (Barr et al., 1988; Green et al., 2011; Petrovich-Bartell et al.,1982) and additional work (Frick, 1985; Scherer, 2003; Standke, 1992), adult listeners can reliably recognize others' emotional states, and we reason that mothers as well as naïve listeners should be able to consistently rate cry and cry-like sounds selected to represent a continuum of infant vocal negativity. To provide better grounding for such research, one should provide systematic operational definitions of cry and cry-like sounds based on real utterances that can be referred to as exemplars representing the categories. Second, one should select cry and cry-like sounds that are acoustically comparable and isolated from their real contexts of occurrence, so that listeners can identify each category based on an independent negativity judgement corresponding to particular acoustic features.

In sum, prior studies have assumed that cry and cry-like sounds are self-evidently differentiable (Barr et al., 1988; Petrovich-Bartell et al, 1982). We have been unable to find in the literature any clear acoustic and/or auditory perceptual criteria for classifying cry vs. non-cry sounds. Yet, we have observed in years of research on vocal development that cry and cry-like sounds are quite diverse. We propose to begin by evaluating perceptual reactions to a relatively well-defined subset of negative sounds, which we call "wail cry" and "whine". These sounds consist of expiratory phonatory nuclei only. Glottal bursts and catch breaths, which very often accompany negative sounds and contribute to their perception, usually as onsets or offsets to cry nuclei or to very brief nasalized nuclei, are excluded in our proposed initial focus because they greatly complicate the kinds of sounds that need to be analyzed. It appears that prior work has typically included in the category "fuss" all the cry-like possibilities, with and without glottal bursts and catch breaths. By limiting the initial scope of inquiry, we presume it will be possible to make concrete progress in the delimited domain, where relatively few parameters are likely to affect perception of cry and cry-like sounds. In later work, we plan to incorporate glottal bursts and catch breaths as additional features of cry and cry-like sounds into our efforts.

Perception in Noise Environment

Normal human communication rarely occurs in completely quiet environments (Helfer & Wilber, 1990). Thus, we also aimed to assess the effects of different noise backdrops on infant cry vs. whine perception. Background noise interference is known to impair spoken word recognition (Billings, McMillan, Penman, & Gille, 2013); and the physiological processing of speech (e.g., Bidelman, 2017; Bidelman & Howell, 2016). Presumably, the ability to parse and properly identify vocalizations amidst everyday noise distractions would have important implications judging infants state as expressed by vocalizations.

Different types of noise can be present in the caregiver-infant environment (e.g., TV, radio, music) and may affect perception differently. In cognitive processing tasks, concurrent noise that is familiar to listeners (e.g., music) is often less distracting than unfamiliar acoustic interferences (e.g., Cassidy & McDonald, 2009; Etaugh & Michals, 1975). We have further shown that listening to familiar music during lexical-semantic decisions is associated with faster response times and lower frequency of mind wandering, suggesting certain types of acoustic interference are less detrimental to linguistic processing, perhaps by reducing cognitive load or listening effort (Feng & Bidelman, 2015). Other studies have reported that vocal music (with lyrics) disrupts task performance on reading comprehension (Perham and Currie, 2014). The key aspects of these studies are related to how distracting background noise can be while performing perceptual identification tasks.

The Purpose of The Study

In the present study, we first aimed to study the auditory perceptual identification of infant cry and whine vocalizations, and to determine the extent to which the selected utterances can be identified by naïve adult listeners in concordance with judgments of listeners who are experienced in the study of cry and cry-like sounds. While there are emerging investigations into the acoustic characteristics of these sound classes (Green et al., 2011; Petrovich-Bartell et al., 1982; Yoo et al., 2015), there is a surprising dearth of studies examining the perception of these important utterances with clear auditory definitions. Our study offers the opportunity to assess the importance of cry and whine sounds in caregiving—we assume that cry has been naturally selected to be the most intense (aversive, negative) expression of infant distress, and consequently that it will be easier to recognize quickly than whine. Our empirical measures are formulated as a way of quantifying the differential saliency of cry and whine, a presumable reflection of their relative intensity of expression.

Secondly, we aimed to assess how common acoustic stressors in the caregiver-infant environment (e.g., speech-babble noise, background music) affect the perceptual identification of infant cry and whine. The way acoustic interference is thought to challenge speech communication is by increasing cognitive load and/or listening effort (Andreassi, 2000; Winn, Edwards, & Litovsky, 2015; Zekveld, Kramer, & Festen, 2010, 2011). We presume that cognitive load in identifying cry and whine should be higher for whine, because it is presumably less salient than cry. We predict, thus, that reaction times should be higher to whine than to cry because of higher cognitive load in identification. It is also possible to quantify listening effort with physiological responses recorded via pupillometry during behavioral tasks to provide a better understanding of the underlying mechanisms for identifying cry and whine sounds. Eye-tracking is a useful tool to investigate cognitive load because changes in pupil diameter accompany difficult mental operations and can be recorded from the eyes (Andreassi, 2000). We hypothesized that pupil diameter would increase if adult listeners have difficulty identifying infant vocalizations, suggesting an increase in cognitive load while judging these sound categories. Complementing behavioral accuracy and reaction time measures, pupillometry also allowed us to reveal whether cry or whine vocalizations were perceptually more demanding than the other type of utterance and how environmental acoustic stressors (i.e., noise, music) modulate listening effort during infant sound categorization.

Hypotheses:

1) Noise interference will slow reaction time (RT) in identifying both cry and whine.
2) Naïve adult listeners will be able to accurately identify cry and whine sounds, as determined by significant agreement of their judgments with those of experienced listeners.
3) Noise interference will be detrimental to accuracy in identifying both cry and whine.
4) Reaction time in the identification task will be greater to whine than to cry sounds due to the lower saliency of whines.
5) Pupil dilation will be greater for whine than for cry sounds, because judgments of whines presumably impose higher cognitive load than judgments of cries.
6) Pupil dilation will correlate negatively with RT, providing two simultaneous measures of cognitive load in identifying cry and whine.

Methods

Participants

Eleven adults (4 males, 7 females) with self-reported normal hearing and no history of neurological or cognitive deficits participated in the study. All were speakers of American English and five had experience with at least one additional language. Participants varied in age from 25 to 51 years (*M* = 30.8, *SD* = 7.6 years). All were nonparents at the time of testing. Participants completed a written consent form in compliance with a protocol approved by the Institutional Review Board at the University of Memphis.

Cry and whine stimuli

The utterances were collected between 0 and 10 months of age (*M* = 3.9, *SD* = 3.9 for cry, *M* = 2.7, *SD* =3 for whine, see Appendix). Isolated utterances were extracted from infant recordings of typically developing Caucasians from English-speaking, mid-to-low socioeconomic families. Inclusion criteria for selecting an utterance were as follows: 1) We required each utterance to be highly audible and discernible without overlay with another speaker's vocalizations or background noises; 2) an utterance was only selected when the two experienced listeners independently agreed on the categorization based on our established criteria utilized in the University of Memphis Infant Vocalizations Laboratory. The experienced listeners were the first and second authors of the paper. The first author is a 4th year PhD student working in infant vocal development, and the second author is an investigator of infant vocalizations for over 40 years.

Utterances included a mixture of laboratory-based recordings and recordings obtained in naturalistic environments using LENA (Language ENvironment Analysis) all-day battery-powered recorders. Laboratory recordings were conducted in a quiet room with high-fidelity equipment and digitized at 44 kHz with the wireless microphone worn by the infant in a vest to maximize signal-to-noise ratio. LENA recordings were obtained from infants in their home settings with the device placed inside the pocket of the infant's clothing, ~7-10 cm from the infant's mouth, and digitized at 16 kHz. For more details about LENA, see (Oller et al., 2010).

Cry and whine definitions. Infant crying sounds have been considered self-evidently differentiable and presumably easy to identify. There are currently no formal, generally agreed upon, auditory-perceptual and/or acoustic definitions of what constitutes a "cry" or "whine" in the literature. Indeed, working towards such definitions is one of the motivations for the present study. In response to the lack of agreed on formal criteria, we have focused attention on distinctive auditory markers of cry (as we identify it intuitively) such as high amplitude nuclei (the phonatory period of the cry), dysphonation (which can consist of several kinds of non-normal phonation during the nucleus of a cry), ingressive catch breaths (which often occur at the very end of an intense cry nucleus), and glottal bursts (which can occur at either the beginning or the end of cry nuclei) (some of these criteria have been discussed in our prior work, e.g., Oller, Buder, Ramsdell, Warlaumont, Chorna, & Bakeman, 2013; Yoo et al., 2015). Cries can be quite complicated when they consist of several of these features within a single utterance, and one of the most frequently occurring negative, cry-like utterances (which we term "whimper") consists of an obligatory glottal burst plus a short (usually nasalized) nucleus.

For the present study, we resolved to simplify the initial comparison of cry and cry-like sounds by selecting only utterances *without* glottal bursts and catch-breaths or by eliminating the glottal bursts and catch-breaths that occurred in some of the selected cases. The utterances were selected to represent a rough continuum from cry utterances, consisting of a nucleus only, to whine utterances, also consisting of a nucleus with no bursts or catch breaths. The nucleus of a whine conveys negative affect but is typically not as intense in negativity as the nucleus found in cries. We matched the cry and whine utterances for duration by selecting the first 350 ms of the nucleus of each of the selected utterances.

A total number of 20 single utterances were selected by the first author from an acoustic database developed from our ongoing longitudinal studies on infant vocal development. 10 were identified as cries and 10 as whines by the experienced listeners, forming the gold standard for the judgments of the naïve listeners.

Behavioral identification task

For the tasks, listeners were seated in an IAC sound-treated booth to minimize external noise interference. The stimuli were presented by computer inside the booth with no need for interaction with experimenters.

Following a brief familiarization phase and task instructions, participants performed a speeded identification task classifying cry and whine sounds. On each trial, listeners heard a single utterance and were asked to judge whether the sound they heard was a “cry” or “whine” as fast and accurately as possible. Response collection was achieved using a custom MATLAB® GUI (e.g., Bidelman, Jennings, & Strickland, 2015). Stimuli were delivered via supra-aural headphones (Sennheiser HD 280) at a comfortable listening level. Each 350 ms stimulus utterance was intensity normalized to 70 dB SPL. The stimulus set consisted of 400 trials (=20 stimuli x 20 presentations each) that were randomly ordered and equally distributed in cry and whine frequency.

This same stimulus set was presented in three separate randomly ordered blocks with different types of concurrent acoustic interference: none (no concurrent noise), music (Mozart’s *Eine Kleine Nachtmusik*), or multi-talker noise babble (Bidelman & Howell, 2016; Killion, Niquette, Gudmundsen, Revit, & Banerjee, 2004). This specific music clip was selected given that it is highly familiar even to musically naïve listeners and our previous work demonstrating that music has a differential effect on listening effort and cognitive load depending on its familiarity (Feng & Bidelman, 2015). Both interferences (noise, music) were presented at a fixed signal-to-noise ratio (SNR) of +15 dB SNR. This SNR was favorable enough to avoid total masking of the speech utterances (e.g., Bidelman & Howell, 2016) yet poor enough to yield meaningful changes in listening effort and thus pupil responses (Zekveld et al., 2010). Each noise block took ~15 min to complete. Participants were allowed short breaks between each block as needed. Collectively, participants heard 1200 infant cry and whine utterances presented in three different types of acoustic degradation.

Behavioral data analysis (accuracy, RTs, and response bias)

Listeners' behavioral identification responses were recorded in MATLAB. Raw accuracy scores (i.e., count of correct responses) and reaction times (RTs) were computed separately for each stimulus and noise condition. Accuracy was determined as the number of utterances correctly identified (in accord with the gold standard judgments) as either "cry" or "whine" on cry and whine trials, respectively. RTs were computed as the median time lapse between the onset of stimulus presentation and the listener's behavioral response (i.e., button press).

Physiological measures of cognitive listening effort: Pupillometry

Physiological changes related to listening effort were tracked via measurement of listeners' pupil dilation response (e.g., Winn et al., 2015). Pupillometry was measured using a Gazepoint GP3 eyetracker. This device provides precise measurement of the location of ocular gaze and pupil diameter with an accuracy of ~1 degree visual angle via an infrared, desktop mounted camera. Continuous eye data were collected from the left and right eyes every 16.6 ms (i.e., 60 Hz sampling rate). Data from the GP3 were logged via an API interface to MATLAB. Pupillometry is affected by a number of factors in addition to cognitive effort including the pupillary light reflex produced by the sympathetic nervous system (Andreassi, 2000). Consequently, the IAC booths' lights remained off during the auditory identification task.

Participants were allowed to wear corrective lenses in the form of contacts; two subject's pupil data were discarded due to excessive noise from the use of glasses.

Continuous eye data were recorded online as participants performed the behavioral identification task. Time stamps were triggered in the data file marking the onset of each stimulus presentation. This allowed us to analyze time-locked changes in the pupil response for each stimulus condition akin to an evoked potential in the EEG literature (Beatty, 1982). Continuous recordings were filtered using a passband of 0.01—15 Hz, epoched using a -50-3500 ms window (where *t*=0 is the stimulus onset), baseline corrected to the prestimulus period, and averaged in the time domain to obtain the evoked pupil dilation response to each stimulus per subject. Right and left eye responses were averaged prior to quantification. This resulted in six waveforms per subject (=3 noise conditions x 2 vocalization types). Blinks were automatically logged by the eye tracking system and epochs contaminated with these artifacts were discarded prior to averaging.

Statistical analysis

Unless otherwise noted, two-way repeated measures (rm)ANOVAs were conducted separately on each dependent measure IBM SPSS (v.23). Vocalization type (2 levels: cry vs. whine) and noise interference (3 levels: clean, noise, music) functioned as within-subject factors. Following omnibus analysis, post hoc multiple comparisons were employed using Bonferroni corrections to control Type I error inflation. An *a priori* alpha level was set at α= 0.05 for all statistical tests.

To evaluate the correspondence between behavioral identification of infant vocalizations and physiological measures of listening effort, we conducted Spearman correlational analyses between listeners' peak pupil diameter and their (i) accuracy score in the identification task and

(ii) RTs. Peak pupil responses were taken as the maximum dilation measured in the 2100-2500 ms time window within each waveform. This analysis window was selected based on initial visual inspection of waveforms which revealed prominent noise condition effects in this time range (see Fig. 3).

Results

Perceptual identification of infant cry and whine

Figure 1 shows RTs and accuracy scores for identifying cry vs. whine infant vocalizations in clean, music, and noise-degraded listening conditions. An rmANOVA revealed a main effect of interference type on RTs [$F(2, 20)=15.44$, $p < 0.0001$]. Multiple comparisons revealed that RTs under clean conditions (*M* = 1433.13, *SD* =52.52) were significantly longer than both RTs under music (*M* = 1296.06, *SD* = 56.89) and noise (*M* = 1304.9, *SD* = 47.5) conditions, confirming hypothesis 1. Music and noise RTs did not differ and no interaction between vocalization type and interference was found. These results indicate that listeners took considerably longer in identifying infant vocalizations when they were *not* in background noise, irrespective of whether they were cry or whine sounds. While seemingly counterintuitive at first glance, faster RTs in noise are indicative of "fast guessing" which is known to occur when perceptual decisions are overly difficult (e.g., Binder, Liebenthal, Possing, Medler, & Ward, 2004; Grice, Nullmeyer, & Spiker, 1982; Yellott, 1971).

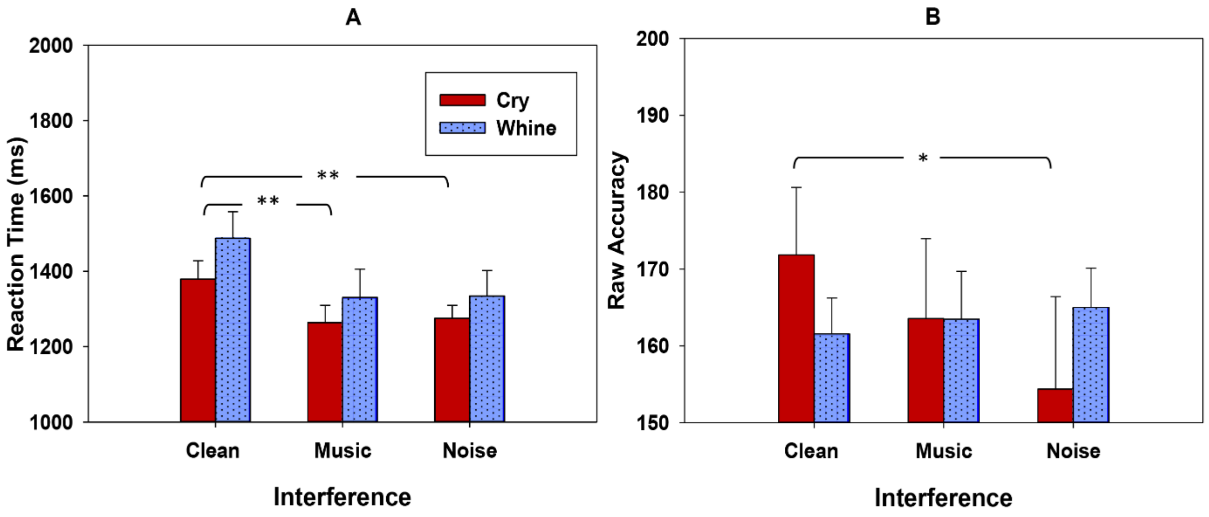


Figure 1. Reaction times (RTs) (A) and accuracy (B) for identifying cry and whine out of 200 trials for either cry or whine under clean, music, and noise conditions. Errorbars= ±1 s.e.m. ** $p < 0.01$, * $p < 0.05$

The accuracy of identification of cry and whine by the naïve listeners was well above chance (= 100/200) for all three conditions as seen in Figure 1B, confirming hypothesis 2. An rmANOVA conducted on accuracy of identification scores revealed a vocalization type x interference interaction [F(2, 20)=4.01, p = 0.032] that was not predicted by hypothesis 3. This suggests that listeners' accuracy in identifying cries vs. whines depended on the specific acoustic stressor. To parse this interaction, we conducted a one-way ANOVA twice, one with cry and the other with whine, and found a main effect of interference with cry [F(2, 20) = 4.97, p = 0.018]. Raw scores of cry identification under clean conditions (M = 171.81, SD = 29.16) were better than raw scores under the speech- babble noise condition (M = 154.36, SD = 39.93), although this contrast did not survive Bonferroni adjustment (p =.07 after adjustment). No differences were found among whine identification scores.

To test hypothesis 4, regarding accuracy of identification of cries vs. whines, we began by considering the possibility of implicit listener bias toward one vs. the other type of vocalization. To test this possibility, we applied signal detection theory to quantify response bias (c), computed as $c = -0.5[z(H)+z(FA)]$, where H and FA are the hit and false alarm rates for detecting cries and $z(.)$ is the z-transform (D. M. Green & Swets, 1966; Macmillan & Creelman, 2005). We conducted an rmANOVA and found no difference ($p = .134$) in response bias at each noise level, indicating that listeners were not inherently biased toward responding “cry” vs. “whine”, *per se.*

Because each pair of cry and whine utterances was extracted from a single infant, we next asked whether listeners’ perceptual identification differed *within* each infant’s vocalizations. For each infant’s cry-whine stimulus pairs, we computed each participant’s differential response time (ΔRT) by subtracting their cry from whine identification speeds. This paired comparison allowed us to assess how perceptual identification of vocalizations differed when using each infant’s utterances as their own controls. Positive ΔRTs denote longer (slower) identification speed for whines compared to cries (i.e., $RT_{whine} > RT_{cry}$).

Figure 2 shows the median ΔRT for identifying individual infants’ cry vs. whine across the 10 infant samples per noise condition. For the clean condition, whine RTs were 151.22 ms ($SD = 46.59$) longer than for cry identification. With interfering music, whine RTs were 59.43 ms ($SD = 33.24$) longer than cry. In speech-babble noise interference, whine identification RTs were 60.19 ms ($SD = 35.41$) longer than cry. One-sample t-tests (i.e., test against $\Delta RT=0$) showed marginal effects in clean ($p = 0.051$) and music ($p = 0.057$) conditions and no effect in the noise condition.

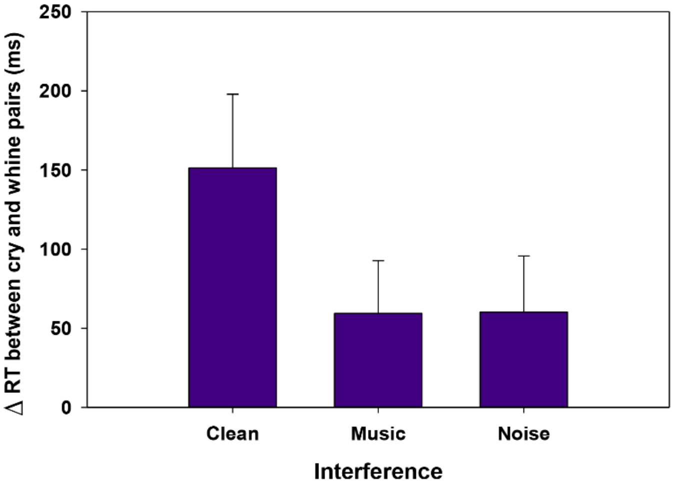


Figure 2. Differential reaction times (ΔRTs) computed by subtracting cry identification RTs from whine identification RTs when stimuli are paired within each infant. Positive values denote $RT_{whine} > RT_{cry}$ Errorbars= ±1 s.e.m.

Pupillometry

Time-locked pupil dilation responses are shown during cry and whine identification for each noise condition in Figure 3. Waveforms were stereotyped by local constrictions and dilation of the pupil diameter over ~3000 ms after the initiation of the speech utterance. Initial visual inspection of responses indicated that prominent differentiation of cry and whine activity peaked between 2100-2500 ms post stimulus onset, whereby whine responses evoked an increase in pupil diameter compared to cry tokens. Formal analyses did not reveal stimulus-related differences in pupillometry for the clean (Fig. 3A; $t(8) = 1.22$, $p = 0.26$, paired-samples *t*-test) nor music (Fig. 3B; $t(8) = 0.85$, $p = 0.43$) conditions, a finding at variance with the prediction of hypothesis 5. However, peak pupil responses were larger for whine compared to cry sounds when tokens were being identified amidst speech-babble noise [Fig. 3C; $t(8) = 4.20$, $p = 0.0026$]. Still, peak pupil response did not differ across the music and speech-babble noise conditions when identifying either cry [$F(2, 15) = 0.51$, $p = 0.61$] or whine [$F(2, 15) = 0.49$, $p = 0.62$]

vocalizations. Nevertheless, given that pupillometry is an objective physiological marker assumed to reflect listening effort (cf. Winn et al., 2015; Zekveld et al., 2010, 2011), these findings demonstrate that the identification of infant vocalizations is more perceptually demanding when identifying whines compared to cries, particularly when faced with the additional acoustic stressor of speech-babble.

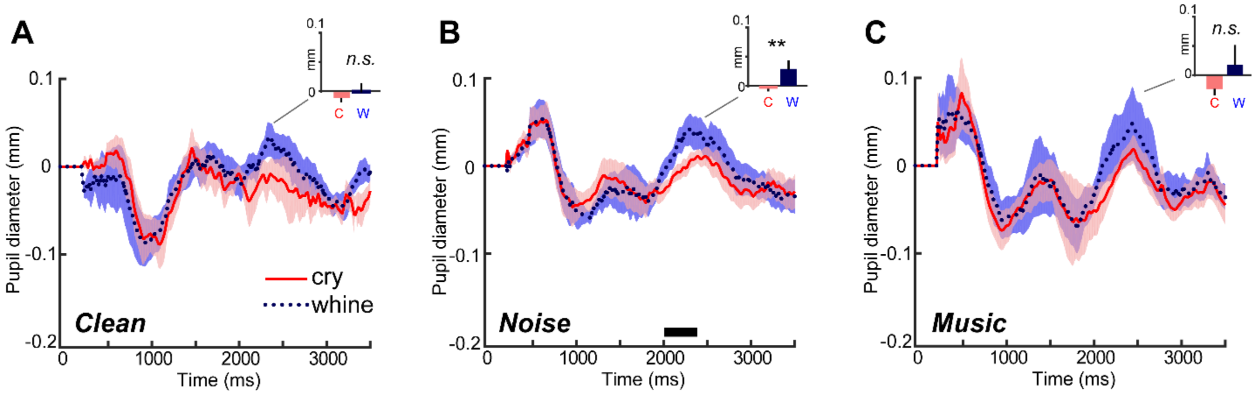


Figure 3. Pupil dilation responses to infant vocalizations reveal increased listening effort for identifying cry vs. whine sounds. Traces show the physiological change in pupil diameter (averaged across eyes) time-locked to the stimulus presentation (*t*=0) when identifying speech sounds under clean (A), speech-babble noise (B), and music (C) interferences. Peak differentiation in pupil responses occurred at ~2500 ms (*insets*) where responses were much stronger when identifying whine compared to cry vocalization, particularly in noise babble. Larger pupil response is indicative of increased listening effort. $^{**}p < 0.001$; shading= ±1 s.e.m.

Brain-behavior correspondences

Correlations between physiological measures of listening effort (peak pupil response) and listeners' behavioral performance (RTs, accuracy scores) were used to assess brain-behavior correspondences underlying the perceptual identification of infant cry and whine vocalizations. For these analyses, we pooled cries and whines in order to achieve adequate sample size for regression analysis (N > 10 observations; Babyak, 2004). Behavioral accuracy scores did not correlate with pupil responses for any of the interference conditions (all *p*s > 0.16). In contrast, we found that physiological responses strongly predicted behavioral identification speeds (RTs) when listeners were classifying cries and whines in speech-babble noise [r_s= -0.73, p=0.0009]. The negative sign of this relation indicates that larger pupil responses were associated with faster behavioral RTs. Given that pupil dilation is thought to reflect listening effort and/or cognitive demand (cf. Winn et al., 2015; Zekveld et al., 2010, 2011), faster identification in noise likely represents "fast guessing," which is known to occur in speech conditions of high difficulty (Binder et al., 2004; Grice et al., 1982; Yellott, 1971). Whine responses seemed to drive this correlation. Indeed, when considering each vocalization type alone, pupil responses were correlated with faster RTs for whine trials [r_s= -0.88, p = 0.0031] but only marginally for cry trials [r_s= -0.70, p = 0.043]

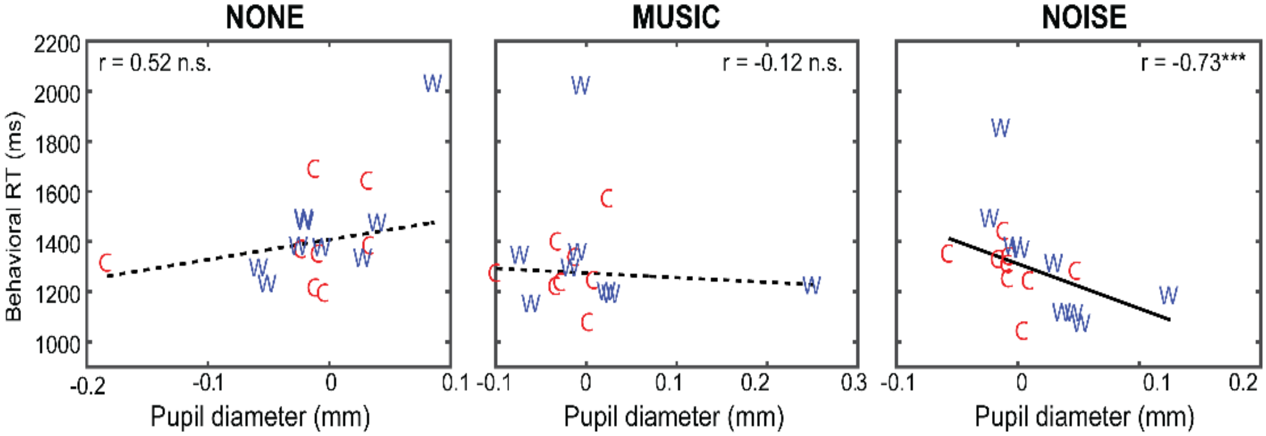


Figure 4. Correlations between physiological measures of listening effort (peak pupil dilation) and behavioral RTs during the perceptual identification of infant vocalizations. Solid lines, significant relation; dotted lines, insignificant relation. W: whine response; C: cry response. Physiological measures predict behavioral identification speeds under speech-babble noise listening only; larger pupil dilation indicating increased listening effort is associated with faster RTs, indicative of fast guessing. $^{***}p < 0.0001$

Discussion

A primary goal of the present study was to investigate the extent to which infant cry and whine were perceptually distinguishable by naïve (nonparent) adult listeners. In addition, we examined pupillary responses and the effects of noise interference during perceptual identification of cry vs. whine to better understand the underlying mechanisms for identifying infant sounds and how environmental acoustic stressors challenge this process. The effort has the advantage of providing new tools for evaluating perceptual reactions to infant vocalizations.

The present study is the first to our knowledge to systematically test whether infant cry and whine were perceptually distinguishable. We found high accuracy scores in identifying cry and whine suggesting that these sounds were identifiable despite the fact that both these sound classes are known to express distress and negative affect. Our signal detection analysis (d-prime > 1; response bias ≈ 0) also suggested that naïve (nonparent) adult listeners can reliably identify cry from whine and do so in a relatively unbiased manner (i.e., they do not implicitly favor one or the other sound class). We also observed a time-accuracy trade off when listeners attempted to identify infant vocalization amidst acoustic interferences. RTs were significantly longer under clean conditions than under music or speech-babble noise conditions. At the same time, accuracy scores in identifying cries were higher under clean condition than under speech-babble noise. In other words, adult listeners responded faster but performed more poorly in the presence of speech-babble noise interference. Presumably, this pattern of responses could reflect “fast guessing” which occurs in cases of difficult or ambiguous percepts (e.g., Binder et al., 2004; Grice et al., 1982; Yellott, 1971) and because listening in noise may impose higher cognitive load during identification (e.g., Winn et al., 2015; Zekveld et al., 2010, 2011). Reaction time differential comparisons of cry/whine judgments for each infant seemed to confirm this time-accuracy trade off (Fig. 2). Although marginally significant, listeners tended to be slower in judging whines than cries. This implies that identification of whine sounds might have been perceptually more demanding and/or induced more listening effort than identifying cry sounds at the behavioral level. This difference makes practical sense: Cry expresses more urgency and so presumably has been naturally selected to include acoustic features that elicit more rapid responses.

In this regard, pupillary responses were useful in revealing the underlying mechanisms of the behavioral findings. Supporting behavioral results, we found that the specific sound class modulated pupil responses and was larger when identifying whines than cries in speech-babble noise (Fig. 3C). Since the pupil response is a reliable and objective indicator of cognitive effort (Andreassi, 2000), our data suggest that listeners experienced increased listening effort when identifying whines, particularly when they were heard amidst additional acoustic stressors (i.e., speech-babble). Increased pupil responses were also associated with shorter RTs (i.e., fast guesses) (Fig. 4), again suggesting that whines are perceptually more demanding than cries. Collectively, our behavioral and physiological results show that identifying whine vocalizations might be more cognitively demanding than identifying cry, again consistent with the evolutionary interpretation that cry is designed to express greater urgency and elicit faster and less effortful responses.

Increased listening effort for identifying whines could be associated with their lower perceptual salience. Indeed, infant whines are found to have lower F0 peak than cry (Yoo et al., 2015) and are usually less perceptually salient to trained listeners than cry. It is known that F0 provides an important acoustic cue for speech processing in noise (Bidelman, 2016; Bidelman & Krishnan, 2010). Differences in saliency may account for why listeners showed greater dilation in pupillometry when identifying whines as opposed to cries in speech-babble noise (Figs. 3C and 4).

Effects of speech perception in noise have been investigated in relation to listener's cognitive abilities including working memory (Akeroyd, 2008; Anderson, White-Schwoch, Parbery-Clark, & Kraus, 2013; Füllgrabe & Rosen, 2016). Speech-babble as noise or music-with-singing as noise (both of which contain competing speech information) have shown detrimental effects on tasks which tap semantic processing (Feng & Bidelman, 2015; Marsh, Hughes, & Jones, 2009; Perham & Currie, 2014). Vocal music (sung by male and female signers) was found to be more disruptive than instrumental music on retrieval tasks (Salamé & Baddeley, 1989). We asked listeners to identify infant vocalizations (non-speech sounds), which places demands on the perceptual system as well as decision-related cognitive resources. Although our study investigated the perception of pre-speech infant vocalizations (i.e., cry and whine), adult listeners nevertheless seemed to be distracted by speech-babble noise when judging infant sounds, despite the fact that these utterances do not contain lexical-semantic information. Our results suggest that the differentiation of infant utterances—albeit non-speech signals—is nevertheless challenged by acoustic interference. This may account for the larger physiological differentiation of infant cries and whines we also found in decision-related pupil responses (de Gee, Knapen, & Donner, 2014). Our results also corroborate previous findings that show (i) music does not always disrupt linguistic decisions (e.g., Feng & Bidelman, 2015), but that (ii) speech-babble noise seems more distracting regardless of tasks (Kozou et al., 2005; Prodi, Visentin, & Feletti, 2013).

In conclusion, infant cry and whine were reliably identifiable by adult (naïve) listeners. Reaction times of identification were significantly influenced by noise (i.e., speech-babble sounds) interference, showing correction for fast-guessing and speech-accuracy tradeoff. Pupillometry responses also supported high cognitive demands when identifying whine amidst the speech-babble interference, showing increased pupil dilation. This study provides the first empirical evidence that classification of cry and whine was perceptually reliable and measurable with adult listeners reaching high agreement with identifications of experienced listeners. The results provide building blocks, both methodological and empirical, for further research on infant vocal expression of varying degrees of negativity.

Chapter 4: Acoustic differentiation of infant speech-like vocalizations from infant cries: A foundation for research in language origins

Introduction

Overview

Early infant vocalizations are under investigation in the search for origins of language (Locke, 1993; Oller, Griebel, & Warlaumont, 2016). In such research, a differentiation between cry sounds and speech-like vocalizations has been necessary, yet widely accepted criteria for implementing the differentiation are not available. If we can more precisely characterize/discriminate early infant vocalizations with clear auditory/acoustic criteria, it would surely enhance the inquiry into language development and the origins of language. To that end, the present study aims to determine acoustic predictors of ratings of infant vocalizations along a distress continuum ranging from cry to speech-like sounds in the first two months of life. An additional goal is to determine the extent to which infant vocalizations reliably transmit differences in level distress as reflected in agreement among listeners in rating level of distress in a range of infant sounds. Finally, we evaluate the extent to which raters use the acoustic predictors in similar or different ways in making their judgments of distress.

Infant Speech-like and Distress Vocalizations

Infants produce various kinds of vocalizations, including vegetative (e.g., coughs, burps, etc.), distress, and speech-like sounds (i.e., protophones, Oller, 2000). Child development researchers often assume that cry is predominant in the first months of life (Lester, 1992; Hoff, 2014; Várallyay & Benyó, 2007), and that protophones occur less frequently. Many have claimed that protophones develop *from* cries and thus assume that protophones do not occur until

2-3 months of life (Takahashi et al., 2015). However, there is solid empirical evidence that human infants produce endogenous protophones from birth (Dominguez, Devouche, Apter, & Gratier, 2016; Nathani-Iyer, Ertmer, & Stark, 2006; Oller et al., 2016). Nathani et al. (2006) reported that from the first months of life the proportion of protophones is dominant in comparison to cry, and that this predominance increases with age. The frequency of non-speech vocalization types (i.e., cries and vegetative sounds) reduced significantly with age whereas protophones increased in frequency (approximately 66% at 0–2 months of age to 99% of all vocalizations by 16–20 months of age).

Dominguez et al. (2016) evaluated infants at 2-4 days investigating the vocal turn-taking capability of newborns. The authors excluded distress vocalizations, focusing on protophones. Newborns in the study produced 2.7 protophones per minute (range from 0.1 to 10.4), showing that protophones are plentifully present from the first days of life. Laufer & Horii (1977) investigated fundamental frequency (*f*o) of infant speech-like (non-distress) vocalizations from 1 to 24 weeks. The authors found that mean *f*o of these vocalizations was around 335 Hz with little variation across ages. These are examples of studies focusing on very early *protophones* while excluding *cries*.

Actually most research on vocalization in the first months has focused on *cries* to the exclusion of *protophones* (for a review see Wasz-Höckert, Michelsson, & Lind, 1985). Even if researchers study protophones or distress sounds separately, they surely need systematic criteria for discriminating the types. Surprisingly, however, we know of not a single paper that provides either explicit auditory or acoustic criteria for discriminating protophones from cries. For decades, two literatures (cry and protophone literatures) have been pursued (Green, Jones, &

Gustafson, 1987; Koopmans-van Beinum & van der Stelt, 1986; LaGasse, Neal, & Lester, 2005; Michelsson, Järvenpää, & Rinne, 1983; Michelsson, Raes, Thoden, & Wasz-Höckert, 1982; Michelsson & Michelsson, 1999; Oller, 1980; Roug, Landberg, & Lundberg, 1989; Wasz-Höckert, Michelsson, & Lind, 1985; Wolff, 1969), with neither literature providing an auditory or acoustic explanation for how they segregate the sounds. Instead, it seems that researchers have mostly relied on situations (e.g., immediately following a needle prick) when defining cry, and in the absence of immediate indicators suggesting pain or discomfort, they have assumed accompanying vocalizations were protophones, and they sometimes referred to these as "comfort" sounds (Stark, Rose, & Benson, 1978).

Developing Criteria to Discriminate Protophones from Distress Vocalizations

A few studies (e.g., Nathani-Iyer et al., 2006; Oller et al., 2013), have attempted to examine both protophones and distress vocalizations in an integrated way. These studies have tended to count sounds by breath-groups or utterances—the terms utterance and vocalization have typically been used interchangeably. Stark, Bernstein, & Demorest (1993) investigated vocalizations from 51 infants (a mixed cross-sectional and longitudinal design) aged birth to 18 months. This study reported age-related effects and individual differences on infant vocalizations according to communicative contexts. In some regards the authors gave quite detailed descriptions of the coding system. However, they did not explicitly explain how fussing or cry vocalizations were differentiated from protophones or non-fussing sounds. Oller et al. (2013) investigated both infant protophones and fixed signals (i.e., cry and laugh) across the first year of life, conceding that for their vocal type coding, "no definition was given [to the coders] for cry or laugh, since it was assumed that these terms would be applied appropriately without training.

However, coders *were* given a ‘reflexivity’ instruction—cries and laughs were to be coded only if the coder perceived (intuitively of course) the infant to have produced the sound reflexively” (p. 31 in the article’s Supporting Information Appendix). Fuller & Horii (1986) investigated f_o, jitter and shimmer of four types of infant vocalizations (i.e., pain cry, hunger cry, fussing, and cooing). Again, no explicit auditory or acoustic criteria were provided to differentiate one type from the other. Instead, the authors defined vocal types based on situation. In general it appears that studies of cries and protophones have tended to rely on intuitive or situational judgments by coders to differentiate infant vocal types.

Researchers have often acknowledged the difficulty in developing clear criteria for coding infant vocalizations (Kent & Murray, 1982; Lynch, Oller, Steffens, & Buder, 1995; Nathani & Oller, 2001). For example, Nathani & Oller (2001) addressed the fact that some fussy vocalizations (a category deemed intermediate between cry and protophone) have substantial speech-like quality (e.g., as in the case of fussy canonical babbles), and as such these utterances should be typically treated as protophones. The authors noted that Stark (1989) had also argued for treating sounds as protophones to the extent that they had speech-like characteristics, even if they also had fussy or distress characteristics. Kent & Murray (1982) emphasized widespread disagreement among researchers on distinctions between speech-like and non-speech-like qualities in infants. Infant vocalizations can be produced in a graded way (e.g., fussing sounds with varying degrees of speech-like quality). The point is highlighted in work by Green, Gustafson, & McGhie (1998) who examined acoustic characteristics of sequences of cries (first 5 cries vs. last 5 cries in a bout). The authors found significant changes in cry sounds in a long cry bout, some seeming more cry-like than others. Porter, Miller, & Marshall (1986) and Thoden & Koivisto (1980) also showed changes in intensity of cry across a bout. Although they did not categorize any of the sounds as protophones, the results support the idea that infant vocalizations are not uniform across time in a vocalization bout but show gradations of features related to cry and protophones, utterance to utterance.

Even within utterances, differentiation of cries from protophones and from intermediate fussy sounds can be difficult. We have proposed that a special category “whine” be used to designate a subcategory of fussing vocalizations. In whines, there is continuous phonation (Appendix c) whereas many other fussy vocalizations include glottal bursts (Stark, Rose, & McLagen, 1975; Truby & Lind, 1965; see Appendix d). An additional category, which we term “whimper” includes at least one glottal burst, and so does not consist of a continuous phonatory event. A glottal burst consists of a sharply produced egress, that sounds like a cough when isolated (see Appendix d). We define a third category of “wail” (a subcategory of cry; see Appendix b) to consist of an intensely distressful continuous phonation. Cries can, though they are not required to, include glottal bursts and/or ingressive, spasmodic “catch breaths” (Stark et al., 1975; Truby & Lind, 1965; see Appendix e). The fact that whimpers and cries can include a wide variety of combinations (within utterance) of these features (continuous phonation, glottal bursts and catch breaths) creates considerable complexity within cry utterances. Even within a continuous phonatory event, major variation can occur in terms of shifts in vibratory regimes, from modal to loft to subharmonic to chaotic, etc. (Buder, Chorna, Oller, & Robinson, 2008). These different regimes have been referred to in an early literature. Truby & Lind (1965), for example, categorized cry as containing a variety of phonatory patterns within utterances, including normal phonation (voiced sounds), dysphonation (significant alterations included) and hyperphonation (sounds in very high pitch) based on acoustic characteristics. The authors provided many varied cry exemplars in spectrograms. Early researchers in cry (e.g. Wasz-

Höckert et al., 1985) argued that these variations were so substantial, research should be focused on a consistent selection criterion to limit the variation (e.g., always selecting the first cry in a bout).

Protophones are at least as complex as cries. They can include various phonatory patterns, including modal (normal), loft (falsetto), pulse (glottal fry), subharmonic, biphonation and chaotic regimes (Buder et al., 2008). Vocants (or vowel-like sounds) are produced with modal phonation in the mid-range of fundamental frequency (f_o) for each individual (Oller, 2000; see Appendix a). "Squeals" are usually produced in high pitch while growls are defined by rough and/or harsh vocal quality. Even though inter-coder agreement on infant vocal types is typically high, sometimes reaching 0.8 (Oller et al., 2013; Yoo, Franklin, Bene, Jhang, & Oller, 2014), there are many incidences of regime variations within utterances. For example, periods of both loft and fry can occur in a single infant utterance identified as a vocant. In this case, coders are instructed to consider the most salient vocal characteristics in choosing among the categories vocant, squeal and growl to characterize the utterance.

In addition to variations of phonatory events *across* protophones, a wide variety of interruptions in phonation can occur *within* protophones. Articulations (movements of the supraglottal tract during phonation) often interrupt phonation or creates rhythmicity akin to syllabification even as early as 1 to 4 months (the Primitive Articulation Stage, Oller, 2000). The most distinctive sounds that infants produce during this stage are called gooing, involving the back of the tongue coming into contact with the back of the throat or palate, thus producing a consonant-like and syllabification effect. Both primitive articulation (i.e., gooing) and well-formed canonical syllables can occur while producing vocants, whines, and even cries. Whiny sounds can be speech-like if they include articulations, even while they express distress.

It is thus clear a broad continuum (from no-distress vocants to high distress wail cry), occurs and that it is complex from a variety of perspectives, with many combinations of events composing cry, protophones, and intermediate distressful sounds. Thus developing explicit criteria based on auditory/acoustic (not contextual) factors for categorizing infant vocalizations is a challenge that must be met directly as we move forward in research on infant vocal communication.

Intuitive Identifiability of Cries and Protophones

Interestingly, in spite of the complexities, coders of recordings intuitively perceive differences in infant vocalizations and show fairly consistent judgments differentiating cries and protophones (Oller et al., 2013). With somewhat more explicit instructions than have been given in most prior research, especially specifying that the cry category encompasses whimpers, Yoo, Bowman, & Oller al. (in submission) found very high (r >.9) agreement among five coders asked to count protophones and cries in 28 five-minute recording segments. Further the data showed that caregivers were significantly more likely to take vocal turns with the coded protophones and significantly more likely to vocally overlap with the coded cries from the first months of life. Thus both laboratory coders and parents consistently treated infant protophones and cries differently, a clear indication that infant signals contain reliable acoustic information indicating distress or lack of it. Non-parent adult listeners were also able to identify cries vs whines selected from recordings of infants (Yoo, Oller, & Bildelman, 2016). Parents may also have subtle awareness of their *individual* infant's cry as indicated by research showing that parents were able to identify cries from their own infant among other cries of similar aged infants (Wiesenfeld, Malatesta, & Deloach, 1981).

If adult listeners are intuitively able to differentiate infant cry and protophone vocalizations, it is obvious that there are acoustic features that contribute to perception of the differences. As far as we know, however, in spite of extensive research on cries and extensive research on protophones, *there has been no attempt to directly account for how acoustic parameters play a role in the distinction between cries and protophones.* Furthermore, since caregivers seem to intuitively judge varying *degrees* of distress in infant sounds in daily life, more systematic research is needed to investigate the link between perception of level of distress and acoustic correlates of perception along a continuum from cry to protophones. This line of work could lay important foundations for studies both of the development of speech infrastructure and for clinical studies focused on cry and speech-like vocalizations.

Rationale and Goals for the Present Study

Research has so far failed to establish auditory and/or acoustic criteria that define cry as distinct from protophones. Further, no prior research has attempted to address directly the whole continuum of phonatory phenomena (from cry to protophones) that make it possible to make reliable judgments about the level of distress in infant vocalizations. The present study is the first to investigate acoustic features that contribute to categorizing cry, cry-like vocalizations (whines), and protophones. To provide a full description of infant vocal development and the origin of language, including both protophones and sounds intermediate between cries and protophones is critical. This study could help develop foundations for more powerful automated analysis (Gilkerson et al., 2017; Xu, Richards, & Gilkerson, 2014) to differentiate types of infant vocalizations.

In the present study, we investigated perception and acoustic properties of prototypical protophones, prototypical cries, and vocalizations of intermediate levels of distress (whines). We confined the analysis to utterances that were phonatory only in order simplify this initial investigation, and we sought to determine 1) reliability of vocal distress signaling in the first months as indicated by the extent to which adult listeners agreed on level of distress for the selected stimuli, 2) the acoustic parameters that best account for perception of level of distress, and 3) the extent to which various listeners may have used the acoustic parameters in similar or in different ways to make their judgments of vocal distress. Based on this present study, we aim to conduct a series of future projects to investigate the complex nature of the infant vocal distress continuum.

Methods

Participants

Infants: Recordings from 7 infants at both 0 and 1 month(s) of age were used. All infants had been recruited for a longitudinal study in the Infant Vocalizations Laboratory at the University of Memphis. All infants had normal hearing and no known developmental impairments. All parents completed an informed consent for the recordings, approved by the Institutional Review Board at the University of Memphis.

We included only data from newborn protophones and distress vocalizations in this initial phase of our research on this topic for the following reasons: 1) The rate of occurrence of infant wail cries is at the highest in the newborn period and decreases dramatically after 2 months (Nathani et al., 2006; Wolff, 1969)—in order to conveniently compare protophones and cry, the newborn period offers a particularly good balance of vocal types; 2) significant neurological development occurs at around 2 months that may have impact on the form of infant vocalizations (Rochat, 1998); 3) given that cries and protophones change under the influence of development and learning (Koopmans-van Beinum & van der Stelt, 1986; Stark, 1980; Wilder & Baken, 1978), newborn vocalizations may represent prototypical forms; and 4) in research on the origin of language, it is sensible to begin studies from as soon as infants can vocalize (Oller et al. 2016).

Listeners: Participants were 39 adults (37 females and 2 males) with an average age of 27.4 years (SD = 5.1; range = 21 ~ 38 years). By self-report all had normal hearing and no history of neurological or cognitive deficits. One listener was the first author, who has been researching infant vocal development for several years and who also conducted the stimulus selection. Her data will be included and in some cases presented as a standard of comparison. All the participants spoke English. 34 main participants were native speakers of American English. The remainder spoke various other languages as well as English (Korean, Spanish, Hungarian, Hindi, Telugu, and Arabic). Four participants were parents at the time of participation. 19 of the listeners had been given systematic training in coding of infant vocalizations, including differentiating cry, whimper, and protophones. The remaining 20 listeners had not been given any training in infant vocalizations. All completed an informed consent for the experiment that was approved by the Institutional Review Board at the University of Memphis.

Distress Level Judgement Task and Acoustic Analysis Procedures

Recordings. All utterances were extracted from the all-day LENA recordings that had been made within the longitudinal study of infant vocal development. The archive of recordings made it possible to extract naturally occurring infant vocalizations from 5-minute periods that had previously been coded by trained human listeners (Yoo, Buder, Lee, & Oller, 2015).

The LENA recorder is small enough to fit in a vest pocket of clothing for infants. The distance from infants' mouth to microphone is usually 5-10 cm. The sampling rate is 16kHz, providing adequate quality recoding for human coding and acoustic analysis. (For details on LENA recording, see Xu, Richards, & Gilkerson, 2014).

From each of the LENA recordings on the 7 infants, thirty-four 5-min segments had been coded—from each all-day recording, 24 segments had been randomly selected, and 10 had been selected as those with the highest infant vocalization rates according to the LENA automated analysis (Xu, Richards, & Gilkerson, 2014). The human listener coding provided the most reliable indications for each of the segments regarding numbers of infant vocalizations (e.g., protophones and cry) contained in them. Selection of the 42 utterances that were used as stimuli in the present study took advantage of the prior human coding. By listening to 5-min segments with high rate of occurrence of infant vocalizations including cries, as indicated by the prior coding, the first author was able to select utterances meeting the below inclusion/exclusion criteria.

Rationale for focusing on a restricted set of infant sounds. As indicated above, both protophones and cries are highly complex, and consequently we limited the stimulus utterances in this initial study to a manageable range of types, including phonation only. In addition to wail cries and protophones, we selected whines to represent utterances displaying an intermediate level of distress. By selecting phonatory segments only, we focused on the most prototypical exemplars of all three types of utterance. In the case of protophones, we included no-distress or very low-distress vocants (see Appendix). Vocants are far more frequent in occurrence than squeals or growls, the other most prominent protophones of the first months (Oller et al., 2013), and are differentiated from squeals and growls by consisting overwhelming of modal phonation (Buder, Warlaumont, & Oller, 2013; see Appendix a). This is the typical phonatory pattern in mature speech and in protophones—vocants account for ~70% of infant protophones. Squeals were excluded both because they occur much less frequently than vocants, and because they are not typically produced in the default pattern of phonation (modal). In addition, squeals are produced at very high pitch, and we have opted, for simplicity's sake, to exclude very high-pitched sounds from all three types of utterances selected as stimuli (wails, whines, and protophones). Growls were excluded for similar reasons (pitch and vocal quality). The pattern of phonation in growls (including a significant period in either pulse or rough phonation, e.g., subharmonic) is relatively uncommon for protophones. In addition, we excluded any vocalizations (wail, whine or vocant) including significant supraglottal articulation corresponding to perception of multiple syllables (see Appendix f), because such articulation is atypical of protophones and cries in the first months.

In the case of wail cries, only intense nuclei were included, intense enough to justify the intuitive label "cry". Glottal bursts and catch breaths were excluded, leaving phonatory periods of wailing only (Appendix b). In accord with our definitional criteria, some wail cries are either preceded or followed by a glottal burst (within the breath group). Nonetheless, the wail nucleus was treated as the primary distress indicator. Thus the wails we selected as stimuli for the present study did not include glottal bursts.

Whimpers, in our system, are *obligatorily* preceded or followed by a glottal burst, usually accompanied by a brief nucleus that cannot be a wail (Appendix d). Whines are interpreted intuitively as being distressful, but less so than wail (Appendix c). If they are accompanied by any glottal burst, they are categorized in our system as whimpers. However, since we exclude glottal bursts from the present study, whimpers are not included. Whines thus represent utterances with phonation only, presenting an intermediate level of distress between wails and protophones.

This selection method substantially restricts the set of utterances to be considered in the present study. The focus is on three prototypical vocal types—the most intense distress sounds in their simplest form (wail cries), the least distressful vocalizations in their simplest and most prototypical form (vocants), and the intermediate distress class of whines. None of the three included glottal bursts, catch breaths, or syllabifying supraglottal articulations. In subsequent research we intend to evaluate effects of the many ways that changes across time within an utterance or in a vocalization bout (e.g., glottal bursts, catch breaths, supraglottal articulation, etc.) can affect perception of varying degrees of distress and speech-likeness.

The present study will evaluate how vocal distress is expressed acoustically. The work addresses signals from high distress to lack of distress (or non-distress). This single dimension does not include the presumable opposite of distress (joy or positivity), in part because positivity is not reliably identifiable in very early infant vocalizations (Jhang & Oller, 2017).

Utterance selection. At the first step, utterances were selected for each prototypical vocalization type (i.e., vocant, whine, and wail) for each of the 7 infants at each age (0 and 1 months). In order to obtain the most prototypical and most acoustically analyzable exemplars, we included infant utterances only when they were 1) highly audible and discernible and 2) produced without overlay by caregiver vocalizations or background noises. We excluded any utterance 1) perceived as so low in intensity that we deemed it would not tend to be noticed by caregivers, 2) shorter than 400 ms or longer than 2000 ms, and 3) any utterance that would have been deemed a squeal (any utterance with very high pitch) or growl.

We found a total of 422 utterances (~10 utterances for each type, infant and age) meeting these criteria. The utterances were then rated by the first and last authors according to prototypicality for each designated category using a 10-point Likert-type scale. The most prototypical utterances within each infant at each age were selected for the perception task and acoustic evaluation (7 infants * 3 types *2 ages = 42 utterances)—thus each infant contributed one wail, one whine, and one cry at each of the two ages.

Listener judgments for distress level. On each trial, participants were asked to intuitively judge the level of distress for each infant utterance. They were *not* asked to categorize the sounds, and the terms cry, wail, whine, protophone, etc. were *not* mentioned in the instructions. A customized high-resolution slider scale was implemented in AACT (Action Analysis, Coding, and Training, Delgado, 2018). Wave files displayed in TF32 (Milenkovic, 2018), the acoustic analysis system invoked by AACT, were presented on a computer monitor, and the rating scale tool (from 0: no-distress to 100: very high distress) appeared on another. The task was to click with the mouse on the rating scale to judge a value from no distress to high distress. Participants read detailed written instructions for the experiment, and the first author also verbally summarized the procedure.

After a practice session with nine infant utterances that were not part of the test set and without feedback, participants performed the actual judgments, with 420 trials (42 randomly ordered utterances x 10 blocked presentations for each of the 42). The ratings were obtained in a quiet room, and participants wore a headset to further minimize noise during the ratings. Participants were allowed to take breaks as needed.

Acoustic Feature Determination

Overview. By restricting stimuli to phonation only, we simplified the initial task of seeking acoustic determinants of vocal distress in human infants. Both a review of the literature on cries and protophones and considerable scouting by the first and last authors provided initial expectations about likely determiners of distress signaling in phonatory segments of infant sounds. The scouting consisted of examination of hundreds of exemplars of infant sounds displayed in TF32 and reviewed jointly with regard to duration, amplitude, f_o and spectral properties. This work was also informed by the recognition of the importance of phonatory regime shifts in infant vocalizations. We selected a set of likely predictor acoustic features (see below), choosing to analyze them regime specifically.

Rationale for vibratory regime analysis. Considerable research has been devoted to showing that the assumption of linearity of source and filter in vocalization (Fant, 1960; Stevens, 1999) is not generally valid, particularly not with child vocalizations (Titze, Riede, & Popolo,

2008) . According to Titze et al. (2008), source-filter interactions can produce violations of linearity. Interaction of glottal airflow with acoustic vocal tract pressures can result in non-linearities reflected in distorted harmonic frequencies. Nonlinearity without source-filter interaction can be associated with subharmonics and biphonation.

A regime is a vibratory pattern of vocal folds (Buder et al., 2008). There are three common registers for speaking: modal, pulse and loft (Hollien, Girard, & Coleman, 1977), each of which corresponds to a vibratory regime in the coding scheme to be utilized here. In these regimes vocal folds vibrate regularly and thus generate periodic waveforms. The modal regime is used in most adult speech. Bifurcations in voice, i.e., sharp breaks from one regime to another, if they are not produced intentionally, are often considered pathological in adults (e.g., they can result from polyps). While it was in fact common for early researchers to treat non-modal phonation types as indicative of neurological or structural pathology, recent studies have made clear that nonlinear phenomena occur regularly in vocalizations (cry and non-cry) of typically developing infants and children (Buder et al., 2008; Robb & Saxman, 1988; Mende, Herzel, & Wermke, 1990). These studies have illustrated that a wide variety of regimes can occur within infant utterances. Since these regimes substantially change harmonic patterns and energy distribution, we view it as necessary to account for regimes in our attempt to account for vocal features of distress.

Regime segmentation of each utterance. Segmentation was performed within each utterance to designate vibratory regimes (see below for a list of regime types). Narrow (10-30 Hz bandwidth) and wide band spectrographic (300-500 Hz bandwidth) displays were used to determine variations in regimes within each selected utterance. When identifying regimes, we used both visual (i.e., spectrographic) and auditory perception, the two modalities complementing each other. For example, if subharmonics appeared in a very short segment of a spectrogram (< 50 ms), but we did not hear the distinctive period doubling (a sort of rough quality) that typically accompanies subharmonics, we did not label that brief segment as subharmonic.

For the purposes of the present study, we modified a regime scheme utilized previously in our laboratory. The following are brief descriptions of the 5 regimes (numbers 1-5; for more details see Buder et al. 2008) and 2 bifurcations (6 and 7; for more details see Buder & Strand, 2003) that actually played a role in the acoustic predictions.

1. Modal: The modal regime is the typical phonatory pattern of speech. The modal regime indicates regular vocal fold vibration, showing harmonics at regular multiples of the f_o.
2. Aperiodic: The code name implies non-harmonic or harmonically unclear periods (i.e., chaos) in a segment or non-periodic extra harmonics (i.e., biphonation).
3. Subharmonic: This category is defined "by the abrupt appearance in the narrow band spectrogram of intervening harmonic, doubling, tripling or even higher integer multiples in relation to the surrounding set" (Buder et al., 2008, p. 7). Prior research in infant cry has often reported appearance of subharmonics (Truby & Lind, 1965).
4. Pulse: The pulse regime is associated with low f_o and often low intensity. Pulse is defined "by the appearance of very closely spaced harmonics often resulting in temporal resolution of individual glottal pulses in the waveform and sometimes also the spectrogram, and a clear perception of a low 'zipper-like' quality" (Buder et al., 2008, p. 6).

5. Break: Sudden changes where pitch or amplitude changed abruptly without a regime shift were coded as instantaneous breaks. These breaks were included in regime segment counts (see below).
6. Trilling: This does not refer to tongue or lip trilling, but to an effect generated at or near the glottis at modulation frequencies similar to those of tongue or lip trills.
7. Flutter: This category was used to indicate modulations in f_o, amplitude or both, occurring at rates faster than syllables but slower than jitter/shimmer. Buder & Strand (2003) have reported three different types of modulations (tremor, flutter, or wow) and provided more details in description of flutter. In the present study, we recently added flutter into our coding scheme, partly because we found it occurring in infant cry and protophones.

Hypothesized predictive acoustic parameters. There are many possible ways that can be imagined for vocal distress to be signaled. We began by evaluating 35 possibilities based on our own prior work (e.g., Yoo, Buder, Lee, & Oller, 2015; Oller et al. 2013) and that of other researchers (e.g., Green et al., 1987; Gustafson & Green, 1989; Leger, Thompson, Merritt, & Benz, 1996). We evaluated the Pearson correlations between each possible predictor measured in TF32 (Milenkovic, 2018) and the mean ratings of the 39 listeners on the 42 utterances. In this way we culled down the predictors of vocal distress to a relatively small number to submit to multiple regression. In this culling down we kept predictors that seemed conceptually independent, and we eliminated those that were either conceptually closely tied to others (presumably redundant with them) or were not highly correlated with the ratings.

In Table 1 we present the original set of acoustic 14 measurement types which were subjected to evaluation in several ways resulting in 35 possible parameters for our analysis. The 14 parameters were evaluated across entire utterances (the *unweighted* method), but we also evaluated 13 them (all but Duration, the one that was incompatible with the *weighted* method) to account for their contributions *within each regime segment.* For example, if an utterance consisted of two regimes—unmarked (modal) and pulse— f_o was measured twice. After obtaining two f_o values, a weighted f_o could be calculated by multiplying each f_o by the durational proportion of each regime in the utterance and adding these two values. Correlations associated with the weighted approach were evaluated for all parameters in Table 1 except utterance duration.

Further, some of the parameters were evaluated by considering the maximum and/or minimum values across all the regime segments *within* an utterance (segment-specific). This approach was evaluated for 8 parameters (numbers 2, 3, 6, 7, 10, 11, 12, and 13). Table 1 gives descriptions of how the raw measures were obtained. If there were regime variations within an utterance, the parameters were measured within each segment.

After all the evaluations were completed for the 35 parameters, we selected 10 (listed and explained in Results, Figure 3) as a set for statistical evaluation as predictors of the distress ratings.

Table 1: 14 unweighted acoustic parameters measured for the analysis across whole utterances. 13 of these were also examined when they were weighted (see text) by the proportion of regime segments in each utterance. 8 of the measures (see text) were also assessed in terms of regime segment specific maxima or minima. These 35 parameters were culled down to 10 for the final analysis (see Fig. 3) of how listeners determine level of distress in newborn infant vocalizations based on acoustic parameters.

No.	Parameter	Description
1	Duration	Duration was measured from the onset to the offset of each utterance by placing cursors in TF32, using waveform displays primarily and not including breathy offsets to utterances. TF32 returns a ms accurate duration value.
2-5	Fundamental frequency (*f*o) mean, max, and sd	*f*o was measured by determining in kHz the frequency of the first harmonic of each utterance. TF32 adapted for AACT traces f_o using an automated algorithm (autocorrelation) and provides mean, sd, min, and max of f_o. In cases where the algorithms failed to trace f_o accurately, the first author corrected the f_o trace using special facilities of TF32. For example, if the trace disappeared or showed values that were transparently incorrect, we adjusted up to 6 parameters (e.g., the LPC Inverse Filter Detuning Bandwidth) to invoke a more appropriate tracing, and if the trace remained inappropriate, we manually modified it to the correct values.
6-9	Root-mean-square amplitude (RMS), mean, max and sd	RMS was used to determine average energy in volts of each utterance and each segment as provided automatically by TF32 in AACT. RMS was measured at each segment in cases where variations in regime occurred, and weighted values were obtained as appropriate.
10	Cepstral peak prominence (CPP)	CPP has been known to be a useful measure for periodicity, particularly in dysphonic speech (e.g., Heman-Ackah et al., 2003). In order to measure CPP in TF32 using LENA recordings (sampling rate 16 kHz), a special updated version of TF32 was developed by Milenkovic. CPP was measured at a typical point of periodicity in each regime segment. The values in dB (high values representing high periodic with respect to aperiodic energies) were computed without high frequency pre-emphasis in order to maximize comparability with commercially available cepstral analysis tools (Awan, 2011)
11	Low-versus high spectral energy ratio (L/H ratio)	This factor has also been show to help explain dysphonation in speech (Awan, Roy & Dromey, 2009; Hillenbrand & Houde, 1996; Awan, Watts & Awan, 2011). A ratio (in dB) was obtained with two different boundary frequencies (i.e., 4 kHz and 2 kHz). Thus we calculated the ratio between the average energy below 4 kHz (or 2 kHz) and the energy above 4 kHz (or 2 kHz).
12-13	Spectral moments of the long-term average spectrum (LTAS), mean and sd	The first and second spectral moments (mean and sd) are useful in obtaining overall shapes instead of focusing on fine structure of the original spectrum (Forrest, Weismer, Milenkovic, & Dougall, 1988). By selecting mmT in TF32, and turning off pre-emphasis, spectrum plots and moment values for a selected period were generated in the 0-8 kHz frequency range. In the present study, we measured both spectral moments with LTA. Mean and standard deviation of spectral moments were measured at each regime segment. After obtaining spectral moment values for each regime segment, weighted values were calculated to adjust for the proportion of each regime type occurring within an utterance.
14	Regime segments	The number of segments designated within the utterance was simply counted (= number of shifts plus one or number of regime tokens, no types)

Statistical analysis

The extent to which raters agreed with each other (inter-rater agreement) on the ratings of vocal distress was assessed by comparing correlations between mean ratings for the 42 stimuli across the coders. The extent to which raters were consistent in their own ratings across 10 trials for the 42 utterances (intra-rater agreement) was assessed by comparing correlations across the 10 trials for each listener. Family-wise multiple comparisons (at $p < .05$ using a Bonferroni correction) were made in R on distress level judgments to assess possible differences between the inexperienced and experienced listeners on ratings of levels of distress in the infant sounds.

Multiple linear regression was used to determine the most predictive acoustic features for distress level judgments, and to provide perspective on possible unique strategies of listeners in judging distress level of vocalizations based on acoustic factors.

To determine whether the 39 listeners varied with respect to each other in how they relied on the acoustic parameters to make their judgments of vocal distress, we conducted permutation tests on each parameter, evaluated for significance by the Kolmogorov-Smirnov test. The products of the Kolmogorov-Smirnov test were subjected to chi-square tests to determine if the raters significantly differed from each other on how they rated the utterances with regard to each acoustic parameter.

For the permutation test, we first determined correlations between each rater and each acoustic parameter. We sampled from all possible pairings of subgroupings of the 39 raters. To achieve this sampling, we began by selecting for an acoustic parameter, say Duration, a random integer between 2 and 38, say n. We then drew a random sample of size n from the integers 1 through 39. We then split the 39 correlations into two groups, one corresponding to the n raters in the sample of integers and the other corresponding to all the remaining raters. For example, suppose n = 2 (randomly chosen with equal probability). Then we would choose a sample of size 2 from the integers 1 through 39; say the values 8 and 17 were chosen. Then we would split the correlations of the 39 listeners (with the acoustic parameter, say Duration) into a sample containing the correlations of the 8th and 17th raters and another sample containing all the correlations of all the raters except the 8th and the 17th (i.e., the 37 other raters) Once these two samples were split, we conducted a non-parametric test for whether the two compared groups of correlations came from the same population (Kolmogorov-Smirnov test). We did this 10,000 times (for 10,000 randomly selected groupings of raters' correlations) for each acoustic parameter and tabulated the p-value for each test. We computed the p-value (p. < .05) for the permutation test as the proportion of times in the 10,000 trials that the test failed to reject the null hypothesis that the samples were from the same population. For example, a value of .85 means that 85% of the tests did not reject the null. Also in this case, 1 minus .85 or 15% of the tests show significant ($p < .05$) differences between the randomly selected subgroupings of raters.

In order to determine whether the observed proportion of 10,000 trials rejecting the null was significantly different from chance, we used a chi-square test on each of the parameters. For example, if 15% of the trials differed from chance at $p < .05$ using the Kolmogorov-Smirnov test on acoustic parameter X, then a two by two chi-square test would compare chance (500 rejections compared to 9500 failures to reject at $p < .05$) against the obtained number of trials where the .05 criterion rejected the null hypothesis for parameter X (1500 compared to 8500), and would determine that the chi-square difference from chance is highly significant ($p < .00001$). We could then conclude that raters differed significantly from each other (showed significant inter-rater variation) in the correlations of their ratings and parameter X.

In addition, for acoustic parameters that showed fewer than 500 out of 10,000 trials meeting the .05 criterion under the Kolmogorov-Smirnov test, we evaluated, also by chi-square, whether those parameters were significantly different from any or all of the parameters where the number of significant differences out of 10,000 was greater than chance by the Chi-Square test (as in the case of 15% differences across the 10,000 trials). For example if only 550 of the 10,000 trials rejected the null hypothesis on acoustic parameter Y, we could test parameter X (1500 compared to 8500) against parameter Y (550 compared to 9450), and would determine that parameter Y showed significantly ($p < .00001$) less inter-rater variation than parameter X.

To determine whether the 39 listeners varied across the 10 trials in how they relied on the acoustic parameters to make their judgments of vocal distress, we computed the 10 correlations each coder's distress ratings with *each* acoustic parameter and then used a Cox and Stuart (CS) test for trend (Cox & Stuart, 1955). This analysis produced an 11 (10 acoustic parameters plus age) by 39 (raters) matrix that contained the p-values of the corresponding CS-test for trend on each acoustic parameter. The null hypothesis of the CS-test for trend was that there was no trend, and the alternative was that there was a monotonic trend (in either direction) for that rater on that parameter. A low p-value (< .05) on any acoustic parameter indicated there was a reliable trend, indicating variation across the ten trials in the raters' judgments with regard to an acoustic parameter. We then determined the proportion of raters who differed from chance by the CS-test on each parameter. Finally we conducted a chi-square on each proportion of raters for each parameter. Thus for example if 8 raters differed from chance on parameter X, we would test 8 compared to 31 vs the chance expectation of 1.7 compared to 37.3, yielding $p = .032$.

To test whether experienced and inexperienced listeners differed with respect to how they relied on the acoustic parameters to make their judgments of vocal distress, we computed mean correlations for each of the listeners and compared the correlations of the experienced and inexperienced listeners using a Wilcoxon test on each of the acoustic parameters. This is a non-parametric test that is preferable in this case to t-tests, given violations of the distributional assumptions of the latter.

Results

Ratings for perception of the level of distress

Reliability of infant vocal distress signaling. The primary way we evaluated the strength of the signal of vocal distress in the 42 infant utterances was by agreement as indicated by Pearson correlations across the listeners in judging level of distress. The average correlation between all possible pairings of the 39 coders was very high, 0.92 (range 0.78-0.98). Even the lowest of these inter-rater correlations was significant at *p* < .00001. Perhaps more meaningful as a measure of inter-rater agreement was the correlation between the ratings of each one of the listeners and all the other listeners: this mean correlation was 0.92, and even the lowest (range 0.87-0.94) corresponded to p < .00001. The individual listeners also showed high consistency across the 10 blocks of 42 trials, with mean intra-rater agreement for all possible pairings of the 10 trials at 0.85 (range 0.56-0.94), and again the lowest correlation was highly statistically significant (*p* < .0002).

A secondary point about agreement in these data concerns how the individual listeners used the rating scale and the extent to which they differed in rating utterances at high or low levels. The mean rating for the 39 listeners across the 42 utterances was 42.3 (sd = 6.7, range = 29.0-59.3, coefficient of variation 6.7/42.3 = 0.16). If we take the mean intra-rater coefficient of variation (COV) across the 10 trials (= .08) as an indicator of rating noise, then it would appear that there remains discernible bias across coders exceeding the rating noise, because the inter-rater COV was twice as high as the intra-rater COV.

Figure 1 shows mean perception ratings for each of the 42 stimuli. The shading illustrates the relation between the mean ratings and the vocal type labels that had been assigned by the first and last authors during stimulus selection. The mean ratings produced three groups corresponding precisely to the vocal type designations, even though the listeners were never instructed to consider identifying the stimuli, and did not encounter the terms wail, whine, vocant, or any other label other than "distress" during the rating task. The mean ratings of the 39 listeners showed all 14 stimuli that had been labeled as vocants were rated as showing less distress than any of the 14 whines, and all 14 stimuli labeled whines were rated as displaying less distress than any of the 14 stimuli labeled wails.

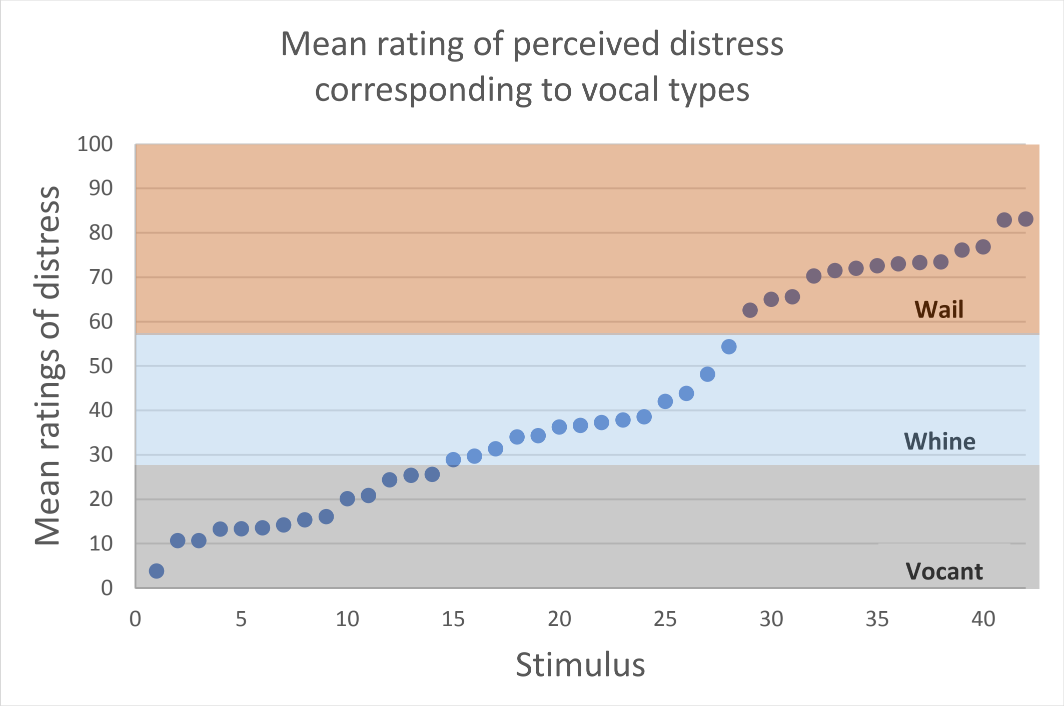


Fig 1. Mean ratings of distress level of 42 stimuli from all listeners. Shading illustrates that the ratings corresponded without exception to the three vocal types (vocant, whine, and wail) as designated during stimulus selection.

The role of experience in coding on the distress ratings. 19 of the 39 listeners had experienced some infant vocalization training and had coded prior samples (identifying vocal types), while 20 listeners were inexperienced in infant vocalization research. By a family-wise (Bonferroni corrected) comparison, there was no statistically significant difference between ratings of the experienced and inexperienced listeners (see Figure 2). 4 listeners were parents, and although we did not find significant differences in mean ratings with regard to the non-parents, it may be worth noting that the mean ratings of the parents included the one with the highest mean rating and the one with the lowest mean rating in the entire group.

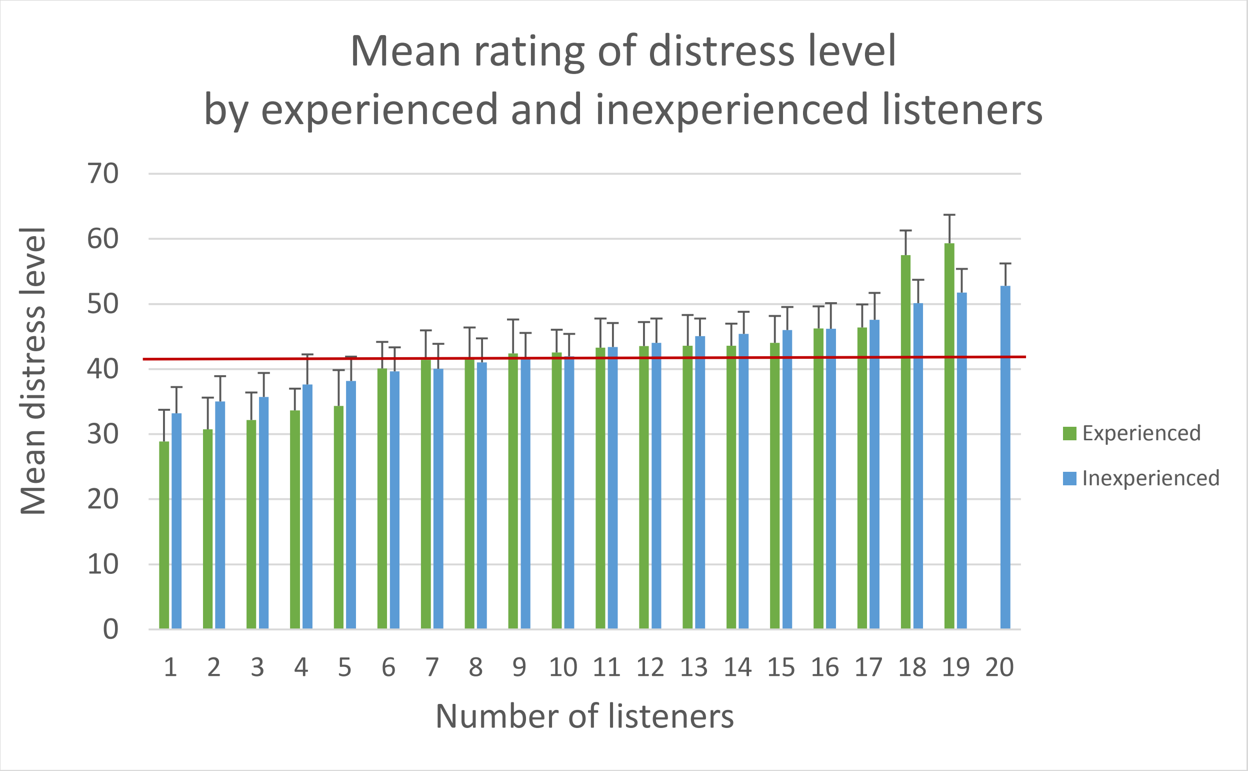


Fig 2. Mean perception ratings of distress level by experienced listeners (from 1 to 19) in green and inexperienced listeners (from 1 to 20) in blue. The red line indicates grand mean (M=42.3) of all listeners (N=39). There was no statistically significant difference between ratings of the experienced and inexperienced listeners. Errorbars= ±1 s.e.m.

Acoustic parameters predicting level of distress

After examining 35 proposed predictors for the ratings of distress as indicated above in methods, we settled on a regression approach implemented in R that started with a set of 10 of parameters plus age as possible predictors—these were the parameters showing high correlations with ratings and conceptual independence from the other proposed predictors or in a few cases lower correlations with ratings but presumed conceptual importance in infant vocalization perception. Pearson correlations between the selected parameters and the mean ratings for the 10 acoustic parameters are displayed in Figure 3.

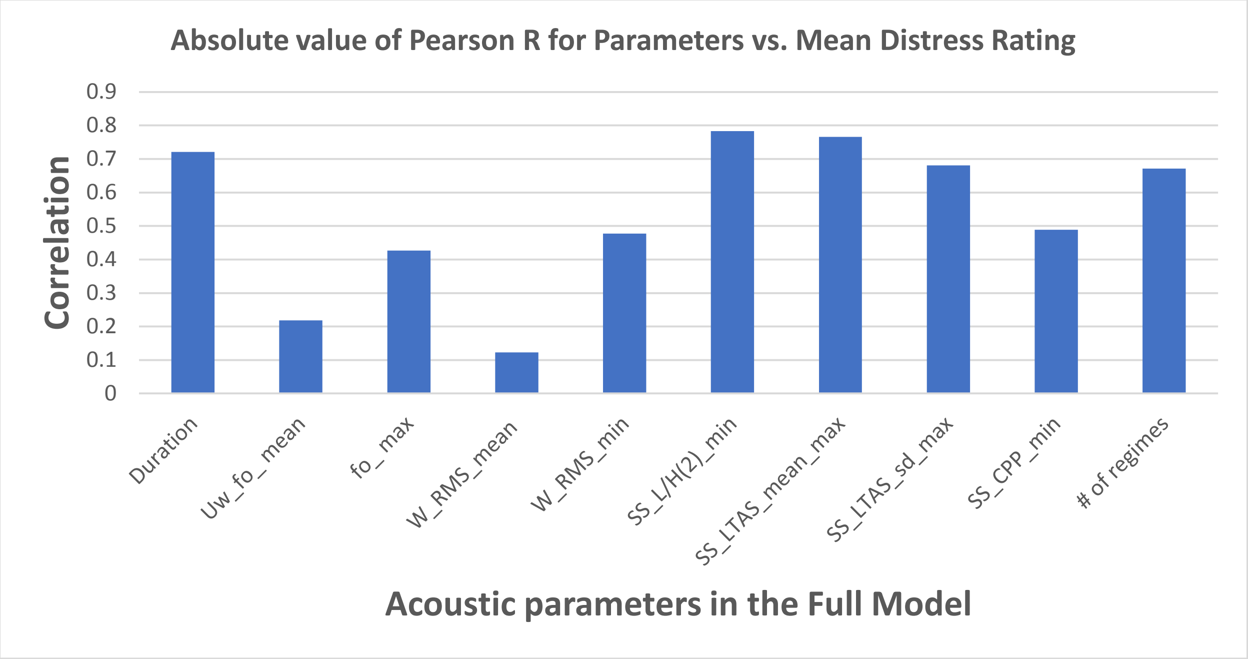


Fig 3. Pearson correlations between each of 10 acoustic parameters and the mean perception ratings of distress level. UW_f_o_mean (unweighted f_o) represents mean fundamental frequency within each utterance. f_o max represents the maximum f_o within each utterance. W_RMS_mean (weighted root-mean-square amplitude) represents mean amplitude (in volts) weighted to reflect proportions of regime durations within each utterance. W_RMS_min represents the minimum amplitude of utterances weighted to reflect proportions of regime durations within each utterance. SS_L/H(2)_min (segment-specific low-versus high spectral ratio) represents the ratio of spectral energy below 2 kHz to above 2 kHz in the regime segment with the minimum ratio for each utterance. SS_LTAS_mean_max (segment-specific spectral moments of the long-term average spectrum) represents the maximum mean of spectral concentration in kHz across the regime segments in an utterance. SS_LTAS_sd_max (segment-specific spectral moments of the long-term average spectrum) represents the maximum standard deviation of spectral concentration in kHz across the regime segments in an utterance. SS_CPP_min (segment-specific cepstral peak prominence) represents the minimum CPP in dB across the regime segments in an utterance. # of regimes represents the number of regime segments within each utterance.

In seeking a small number of optimal acoustic predictors to submit to regression, we compared the Pearson correlations between each acoustic parameter and the mean perception ratings. As can be seen in Figure 3, high correlations were obtained for several parameters. A Full Model analysis (including age as an additional predictor) indicated, however, that only duration was a significant predictor of the mean distress ratings. Evaluation of collinearity suggested that some of the Full Model parameters were extremely highly correlated with each other. For example, SS_LTAS_mean_max was correlated at > 0.8 with SS_L/H(2)_min and SS_ LTAS_sd_max. This collinearity encouraged us to consider systematic ways to reduce and optimize the number of predictors.

Table 2 shows the relative contribution of each predictor/covariate to the Full Model regression. Since the predictors were not orthogonal, the contributions of each to the R^2 did not add to the total R^2. Instead the relative contribution of each predictor can be measured as the contribution to the total R^2 when the predictor is the last one added to the model. This contribution is shown in the table labeled **relative R^2.** Another (perhaps better) means of assessing the relative importance of each predictor is to compare the standardized coefficients (subtracting the mean and dividing by the standard deviation). In this way the magnitudes of the predictors are not considered in the estimates, and the comparative effect of each predictor can be compared directly. Standardized coefficients represent the mean change in the response given a one standard deviation change in the predictor.

Table 2. Standardized coefficients and relative contribution of each predictor to the Full Model

Predictors	**Standardized. Coefficient**	**Relative R^2**
Intercept	42.351	
Age1m	-2.364	.0071
Duration	6.719	.0433
UW_f_o_mean	1.691	.0010
f_o max	4.740	.0065
W_RMS_mean	-2.583	.0056
W_RMS_min	4.772	.0156
SS_L/H(2)_min	-5.261	.0089
SS_LTAS_mean	7.961	.0130
SS_LTAS_sd	-1.561	.0007
SS_CPP_min	0.805	.0003
# of regimes	3.798	.0063

Statistics for the parameters of the best model in terms of lowest AIC (Akaike information criterion, which estimates relative quality of alternative models) are shown in Table 2. The R^2 for the best model was 0.84. A Backward Selection method was used to select the best model—thus, not every possible subset was considered. In the Backward Selection method the linear model was fit to all the available predictors, with predictors being omitted from the model one at a time based on lowest AIC. This Backward Selection method resulted in elimination from consideration of some of the highly intercorrelated predictors of the Full Model. Of particular salience in the Backward Selection outcome (Table 3) is the predictor SS LTAS mean max, which showed both the highest level of statistical significance under Backward Selection as well as the highest standardized coefficient (i.e., the highest effect size) in the Full Model. In addition, Duration, Number of Regime segments, and f_o max showed statistical significance at the $p < .05$ level in the Backward Selection method, and both age and weighted RMS min showed values approaching significance.

Table 3. Parameters used in the Backward Selection method model for predicting the level of distress

	Mean (SD)	*p*-value
Intercept	NA	0.0031
Age 1m	NA	0.0525
Duration(ms)	1030.6 (450.9)	0.0042
f_o_max	479.8 (66)	0.0037
W_RMS_min	0.42 (0.26)	0.0699
SS_LTAS_mean	1.77 (1.3)	< 0.0001
# of regimes	2.4 (1.4)	0.0468

The predictive power of the 5 acoustic factors that performed best as predictors under Backward Selection can also be seen to have yielded segregation of the stimuli into the categories wail, whine, and vocant as they had been designated by the stimulus selectors. Fig 4 shows descriptive results in terms of means and standard errors for the acoustic parameters within each vocal type (vocant, whine, and wail). Wails were about twice as long as vocants. f_o max was higher in whines and wails than in vocants. The most powerful factor in differentiating the vocal types appears to have been the same one that was most salient in the Backward Selection regression method, the segment-specific long-term-average spectral concentration, which was highest in wails and lowest is vocants. This factor indicated that more energy was concentrated at higher frequencies in wails than in the other vocal types. Wails also showed the highest RMS min, indicating minimum amplitude of wails was greater than in vocants or whines. Finally, numbers of regime segments was highest in wails.

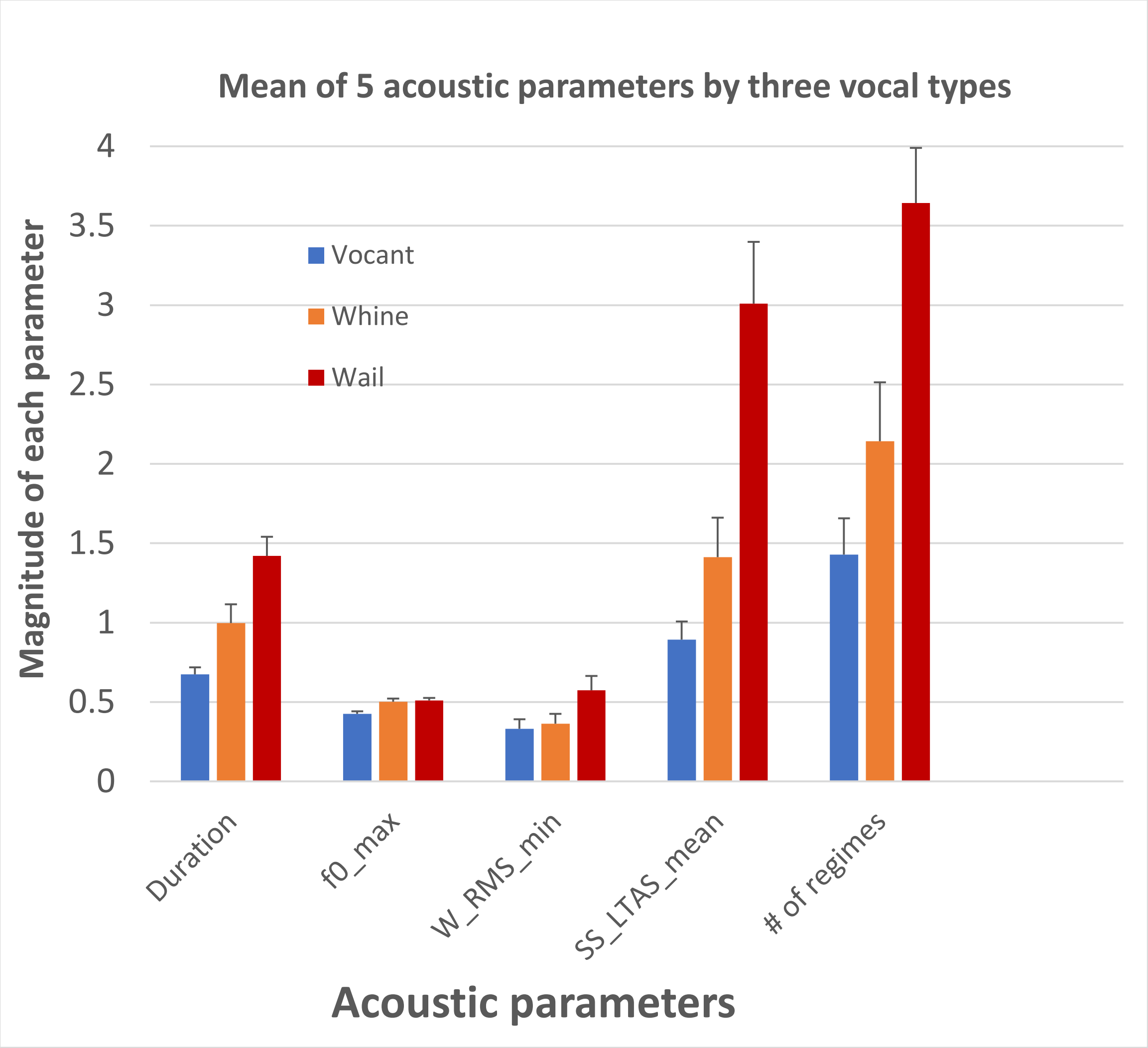


Fig 4. The acoustic parameters differentiated the three vocal types (vocant, whine and wail). Each bar presents mean and standard error. Duration is given for utterances in seconds. f_o_max represents maximum f_o within each utterance in kHz. W_RMS_min represents the minimum RMS in volts across regime segments within each utterance weighted by the proportion of each utterance accounted for by each regime segment within the utterance. SS_LTAS_mean (segment-specific spectral moments of the long-term-average spectrum) represents the highest mean spectral concentration within any regime segment of each utterance in kHz. Number of regimes represents the number of regime segments within each utterance (number of regime shifts + one).

Possible differences within and across listeners in how the acoustic parameters were used to rate distress

Table 4 shows the extent to which the 39 listeners varied with respect to each other in how they relied on the acoustic parameters to make their judgments of vocal distress, based on the permutation test described in Methods. So, for example, as indicated in column 2 of Table 4, the acoustic parameter Duration corresponded to no cases where the null hypothesis (that the randomly selected groups did not differ in their correlations with the acoustic parameters) was rejected at alpha = 0.05, whereas for the parameter SS_L/H(2)_min, there were 4,826 cases (1 - 0.5174 = 0.4826) where the null hypothesis was rejected out of the 10,000 permuted samples. These data indicate very strong differences across raters on utilization of some of the parameters, namely highly significant differences in the correlations between raters' judgments and the parameters Age, f_o _max, W_RMS_mean, SS_L/H(2)_min, SS_LTAS_sd, SS_CPP_min, and # of regimes. In all cases of significantly different usage of the parameters, chi-square tests showed p <.00001, indicating inter-rater variation was highly significant. Also all the other parameters (Duration, UW_f0_mean, W_RMS_min, and SS_LTAS_mean) differed significantly from those listed above in that they showed *fewer* differences from chance, indicating a lesser tendency for inter-rater variation in how they used the acoustic parameters.

Table 4. Statistical tests for inter-rater differences. Column 2: Proportion of cases (out of 10,000) for the permutation test failing to show .05 level differences in correlations between distress ratings for the infant utterances and the acoustic parameters across randomly selected rater groupings. Column 3: *p*-values from chi-square tests indicating whether each acoustic parameter was used differently among raters. ^ indicates that while the proportion for the indicated parameter was not significantly lower than would be expected by chance, it *was* significantly higher than the proportion for any of the cases that *were* significantly different from chance. Thus 7 of the 11 parameters showed significantly more inter-rater differences than expected by chance, and the remaining 4 showed significantly fewer inter-rater differences than the other 7.

Acoustic Parameter	Proportion of trials failing to reject the null hypothesis in the permutation test	*p*-value for chi-square test for inter-rater variation
Age	0.8957[a]	.00001
Duration	1.0000[b]	^
UW_ f_o_mean	0.9754 [b]	^
f_o _max	0.9004 [a]	.00001
W_RMS_mean	0.9272 [a]	.00001
W_RMS_min	1.0000 [b]	^
SS_L/H(2)_min	0.5174 [a]	.00001
SS_LTAS_mean	0.9724 [b]	^
SS_LTAS_sd	0.5690 [a]	.00001
SS_CPP_min	0.8688 [a]	.00001
# of regimes	0.8741 [a]	.00001

^ These 4 parameters showed significantly less inter-rater variation than the other 7 parameters.

To determine whether the 39 listeners varied the extent to which they relied on the 11 parameters (10 acoustic parameters and infant Age) across their 10 ratings of each utterance, 10 correlations with each acoustic parameter were computed, and a Cox and Stuart test for trend was conducted. For description of the procedure see Methods. Four listeners were very consistent across the 10 trials for the 10 acoustic parameters, showing no case where they significantly varied across trials on any parameter. But 35 listeners varied significantly in how they made their judgments based on the acoustic parameters across the 10 trials (i.e., at least 1 parameter showed a monotonic trend in one or the other direction). W_RMS_min was found to be the parameter where listeners tended to change most in the way they made their judgments across the trials, with 12 out of 39 listeners showing statistically significant changes. A chi-square test showed a clear difference from chance (*p* < .003) for W_RMS_min (Table 4 column 3). Other parameters with *p* < .05 were Age, UW_f0_mean, W_RMS_mean, W_RMS_min, SS_CPP_min, and number of Regimes. Notably parameters that did not show statistically reliable differences in how the acoustic parameters were used by the raters across trials included all three of the parameters associated with spectral concentration (SS_LH2_min, SS_SLTA_mean_max, and SS_SLTA_sd_max) and Duration, parameters that were all particularly highly correlated with the mean distress ratings. chi-square tests showed that raters varied significantly across the 10 trials on Age, UW_ f_o_mean, W_RMS_mean, W_RMS_min, SS_CPP_min, and # of regimes.

Table 5. Statistical tests for intra-rater differences and differences in rating experience. Column 2: Number of raters out of 39 who showed a significant trend, changing correlations between distress ratings and acoustic parameters across 10 trials by Cox and Stuart tests. Column 3: *p*-values from chi-square tests indicating significance levels for intra-rater variation across 10 trials in the degree to which their distress ratings correlated with each acoustic parameter. Column 4: *p*-values indicating degrees of difference between the experienced and inexperienced listeners on their correlations between ratings and acoustic parameters.

Acoustic Parameter	Number of raters out of 39 with significant trends of variation across 10 trials	*p*-value for chi-square test for intra-rater variation across 10 trials	*p*-value for Wilcoxon test, experienced vs inexperienced raters
Age	9	0.01	0.19
Duration	5	NS	0.99
UW_ f_o_mean	8	0.03	0.77
f_o _max	7	0.06	0.13
W_RMS_mean	10	0.009	0.01
W_RMS_min	12	0.002	0.13
SS_L/H(2)_min	6	0.1	0.0001
SS_LTAS_mean	6	0.1	0.03
SS_LTAS_sd	4	NS	0.0012
SS_CPP_min	9	0.02	0.25
# of regimes	8	0.03	0.15

In a final test we separated the experienced and inexperienced listeners into two groups. For each acoustic parameter, we computed the correlations between that acoustic parameter and the listeners within each group (see Fig. 5). To determine whether the mean correlations differed, we did a Wilcoxon test (column 4, Table 5). So, for example, the p-value for the difference in mean correlations between mean ratings and Age, between experienced listeners and inexperienced listeners was 0.19, indicating no evidence of a significant difference between the correlation of experienced listeners and unexperienced listeners with Age. As can be seen, four of the parameters (W_RMS_mean, SS_L/H(2)_min, SS_LTAS_mean, SS_LTAS_sd) showed reliable differences across the groups.

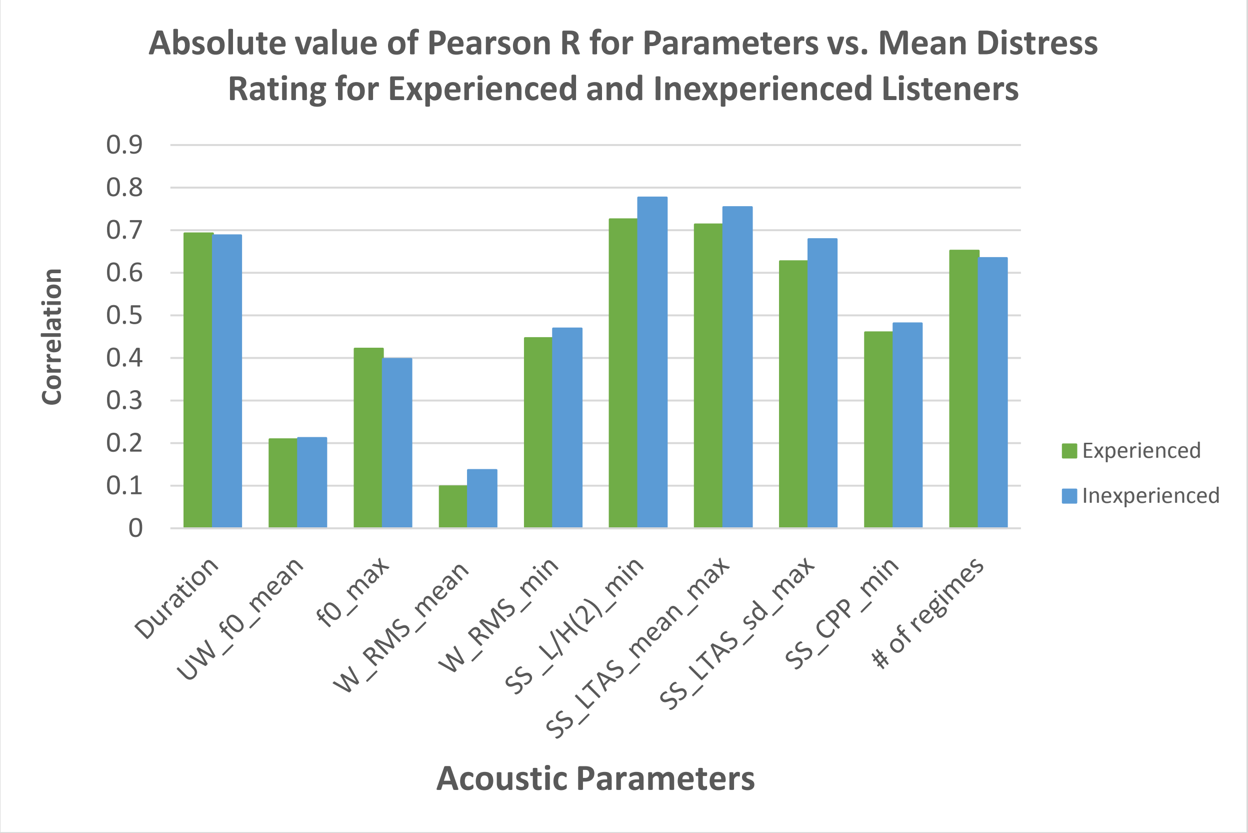


Fig 5. Pearson correlations between each of 10 acoustic parameters and the mean ratings of distress level between experienced and inexperienced listeners. W_RMS_mean, SS_L/H(2)_min, SS_LTAS_mean, SS_LTAS_sd were significantly different across groups (Table 5).

Discussion

Our study attempted to provide an expanded view of how vocal distress is expressed in human infancy and how well it can be recognized, which is to say, how stably the infant provides the signal of distress. But our intentions have been driven by interests in the origin of language, and consequently we have addressed infant vocalizations that both do and do not express distress. The protophones in particular are sounds that infants can produce with or without any sign of distress. They presumably manifest a capacity for voluntary vocalization that lays a foundation for language. In studying vocal distress, then, we address a continuum from sounds that show maximum distress (cries) to sounds that show minimum distress (vocants, the most common type of protophone). We also included in our study sounds that are intermediate in distress, referring to these sounds as whines. No prior study has ever addressed this whole continuum of infant vocal distress in either perceptual or acoustic investigation.

The outcomes of our evaluations yielded notable new information and surprises. We determined that human listeners, whether experienced in research on infant vocalizations or not, showed very high levels of agreement in judging the degree of distress across 42 carefully selected infant vocalizations. Mean ratings of distress level corresponded precisely to the three vocal types (vocant, whine and wail) designated by the first and last authors during segment selection, although the listeners were only asked to judge the level of distress of each utterance. Again the results suggest that the infant's signaling of level of distress is quite reliable, and that humanity has been evolved to have strong intuitive awareness of vocal distress in infants, a capability that would seem to be critical in intuitive parenting (Papoušek & Papoušek, 1987).

Notably, while listeners agreed with each other highly in judging level of distress in the infant sounds, they differed in how they used the acoustic parameters to achieve those judgments, and especially notably the experienced and inexperienced listeners differed in how they utilized the acoustic parameters. This latter finding amazes us, because the training and experience of the "experienced listeners" was never explicitly focused on the acoustic parameters evaluated here. And the listeners changed across time, even very short periods of time (the 10 trials of the study occurred within about a half an hour) in how they made their judgments in terms of the acoustic parameters. This finding suggests human listeners, even without feedback, engage in judgments of distress variably, as if exploring possible ways of making distress judgments, and thus of understanding infant needs.

The research also sought to determine primary acoustic indicators of infant vocal distress, in the first direct comparison of prototypical cries (wails), whines, and protophones (vocants). In our initial acoustic explorations, we developed a speculation that wailing (that is, crying without glottal bursts or catch breaths) is signaled primarily by a sort of spectral concentration, such that energy levels are relatively high above 2 kHz in cry and relatively low above 2 kHz in protophones. Gustafson & Green (1989) previously suggested that more energy at higher frequencies contributed to judgments of greater aversiveness among cries—protophones were not evaluated. Notably this spectral concentration feature of cries in our explorations often occurred *within a single vibratory regime segment* within the utterance, and did not necessarily characterize the utterance as a whole. This speculation inspired the consideration of parameters that would reflect spectral concentration especially *within particular regime segments* and perhaps serve as best predictors of ratings of vocal distress across the utterances selected for the present study.

We statistically evaluated a wide variety of possible predictors directly, including among them, the ratio of energy below and above 2 kHz within the regime segment with the lowest ratio. But the strongest predictor of the ratings of vocal distress turned out to be based on a moments analysis of the spectrum as a whole, in particular the spectral centroid (mean) of the long-term average spectrum within the regime with the highest spectral centroid for the utterance. This factor showed the highest effect size (standardized coefficient in the regression model, ~8) and the highest level of statistical significance (*p* < .0001) using a backward selection regression approach. Utterances with higher spectral concentrations were more likely to be judged as distressful.

Several other factors were also highly predictive, one of them being duration, where cries were longer than whines, which were longer than vocants. Importantly duration could have been even more important as a determiner of distress judgments if we had included the whole range of possible durations—we artificially restricted the stimuli in all three vocal types to the duration range of 400 to 2000 ms. Lest one think, however, that duration always determines the distinctions, there were two cries in the sample that were shorter than two of the vocants, and yet the listeners unambiguously rated the distress levels of the short cries with the other cries and the long vocants with the other vocants ($p < .001$).

Another factor that contributed clearly to the prediction of vocal distress was f_o maximum, which appears to correspond with a prior finding that cries with higher pitch are judged to be more aversive (e.g., Zeskind & Marshall, 1988). Ours is, however, the first direct comparison of acoustic features in cries and protophones and thus suggests for the first time directly that f_o predicts vocal distress across the whole continuum of infant vocalizations. Yet the present study only responds to part of the relevant question, because we again artificially restricted the f_o range by not including squeals among the protophones nor loft (hyperphonation) among the cries. In a subsequent study we plan to address the role of loft (and of pulse as well other growly features of some protophones) in the perception of vocal distress in infancy.

The number of vibratory regime tokens also contributed significantly to judging the degree of distress. Gustafson & Green (1989) showed that adult listeners perceived infant cry as more aversive as a function of the amount of dysphonation. We specifically accounted for dysphonation within each utterance by coding segments as aperiodic or subharmonic in our regime coding scheme. The number of regimes, therefore, indirectly reflected the presence of dysphonation.

Our opinion emphasizing the complexity of vocal distress expression in human infants has been amplified by the experience of studying these utterances individually, which were carefully selected to represent a narrow range of parameters characterizing the three vocal distress categories (high, medium, none). The acoustic analysis, which included segmentation of each utterance into vibratory regimes, suggested, as summarized above, that several parameters are involved in judgments of distress, even in this restricted set of phonatory-only vocalizations. The complexity of the possibly determining parameters may be even greater than we have been able to show with the analysis presented here. For example, for every utterance preselected as wail, there was at least one regime segment >200 ms (and often several) for which the judgment "wail" did not apply according to the two main judges (first and last authors) when the segments were played back in isolation. Instead the judges deemed these segments to be in modal voice (thus corresponding to vocant-like phonation). Often there were multiple such segments within a wail, and sometimes they were >500 ms. Also in all but one of the 14 wails, there was a notable regime segment marked by the acoustic analyst either as aperiodic or as including subharmonics, designations that presumably would have been called dysphonation in most of the earlier literature. These regime segments did not, however, by themselves determine a judgment of wail—i.e., if the regime segment was played back in isolation, it usually was not judged by either of the two main judges as wail—rather these segments sounded strained or growly, but not cry-like. The most common pattern of wail included both beginnings and endings of >100 ms that were judged in isolation unambiguously to consist of modal voice, i.e., they sounded like vocants; during the intervening regime segments, there was typically at least one dysphonated segment, that did not sound like wail in isolation, but in combination with the adjacent segments was judged as unambiguous wail. Consequently, it can be said that the great majority of the wail utterances were characterized by a strong contrast between at least one regime segment of dysphonation and surrounding segments of modal phonation (vocant).

An interesting possibility involving another change across time within distress utterances is suggested by work of Wermke, Mende, Manfredi, & Bruscaglioni (2002), who described cry as often including a rise-fall contour. The two main coders evaluated the utterances preselected as wails regarding this factor and found that only half of the 14 wails showed a rise-fall pattern. The remainder showed flat, complex, or rise-then-flat patterns. 5 of the 14 utterances that had been preselected as vocant also showed a rise-fall pattern, with the remainder showing flat, complex, or rise-then-flat patterns. Thus the hypothesis that a rise-fall pattern would be a strong predictor of wail was not straightforwardly supported in this small sample of utterances. This impression is fortified by Várallyay & Benyó (2007) whose data suggested that only about a third of cry utterances have the rise-fall contour. However, the possibility remains that melody contour may play a significant role in distress prediction. A much larger study will be needed to evaluate the possibility. For the present it would appear that overall contours are much less influential in determining judgments of distress than the number of and the nature of individual vibratory regimes.

We hasten to emphasize that much of the pattern of results depends upon the vibratory regime analysis. Prior research has not taken this approach in comparing cries and protophones. While prior research has *noticed* regimes, it has not taken systematic account of them in assessing predictive power of factors such as aversiveness of or distress manifest in cries and protophones. We think all future research on acoustic markers of distress should take account of vibratory regimes. If there had been no regime analysis in the present work, we would have been driven to the conclusion that duration was the overwhelmingly important factor in the judgment of distress. No indications of differences across raters, across trials, or across experience levels would have been revealed. Duration was the only significant factor in the initial regression analysis where no regime segment specific factors were considered and appeared to be the only important predictor.

However, after including the regime segment-specific factors, much more varied and interesting influences were revealed: 1) Segment-specific spectral concentration proved to be an especially strong predictor, even stronger than duration; 2) number of regimes itself was revealed as a significant factor along with maximum f0, 3) raters proved to differ in their degree of correlation between ratings and the acoustic parameters, 4) raters differed across ten 10 trials (within only about half and hour) on how highly their judgments correlated with particular acoustic parameters, and 5) experience proved to have a notable effect on how rating judgments were made. All these patterns would have been invisible with analysis that ignored regimes.

The study illustrates that human listeners come well-prepared to judge vocalizations of human newborns as being either speech-like, cry-like or in between. Such a capability is surely relevant to the intuitive parenting task of engaging infants in protoconversation (with protophones) while treating whines and cries as signals of need. Early interactions between parents and infants can only be laid if parents recognize the material of potential speech in infant sounds.

Since this is the first study comparing acoustic features directly across the continuum from cry to protophones, it provides new evidence about how the system of infant vocal distress and speech-like vocalization is structured. In particular it appears that spectral concentration plays a salient role, inasmuch as the nuclei of speech–like sounds have energies concentrated at low frequencies, while more distressful sounds have nuclei concentrated at higher frequencies in at least one regime segment. Such insights may be relevant to development of better automated tools for the assessment of infant vocalizations across a broad range. Data driven approaches to developing algorithms for automated detection of vocal types (Xu et al., 2014) yield little insight into how the acoustics of vocalizations drive detection. Our study should offer suggestions to modelers about how acoustic features of infant sounds contribute to success in detection.

Future Directions

37 out of 39 participants were females. Prior research has produced mixed results on gender effects in perceiving infant cry. Zeskind, Sale, Maio, Huntington & Weiseman (1985) found both similarities and differences between males and female raters when the raters were asked to judge the utterances on four dimensions (i.e., urgent, arousing, aversive, and sick) and on three cry segments (i.e., initial, middle, and final segments). The authors found that *both* males and females showed a tendency toward increased ratings from the initial segment to the final segment in hunger cry on the dimensions aversive and arousing. The authors also reported *differences* in perception between males and females. Females perceived the final segment of hunger cry as more aversive than males. Frodi, Lamb & Donovan (1978) investigated physiological responses (e.g., skin conductance and blood pressure) of mothers and fathers to infant cries and smiles (presenting by video). The mothers and fathers also filled out a mood adjective checklist (MACL) after the video presentation. The authors found that the mothers and fathers did not differ in their physiological responses to infant cries. However, the mothers reported more extreme mood than the fathers. Some researchers have suggested that females are innately predisposed to respond to infants more than males (Klaus, Trause, & Kennell, 1975;

Money & Tucker, 1975). Other researchers have also suggested that females are more encouraged to express their feelings than males and that females may thus be more responsive in answering questions about infant vocalizations (Frodi, Macaulay, & Thome, 1977; Maccoby & Jacklin, 1974). In future studies it would be useful to contrast male and female raters of vocal distress.

Another line of productive future research might seek to control stimulus parameters in additional ways. Some such work might manipulate real infant vocalizations artificially, for example by reversing them in time, by repeating regime segments or portions of regime segments, or by deleting portions of them. Synthetic stimuli could offer even more flexible possibilites for testing of acoustic factors and their influence on distress judgments.

Because our effort here excluded squeals and growls as representatives of the protophones, we think it might be useful to conduct similar studies where all the stimuli (wails, whines and protophones) include high f0 (a strategy that would bring in hyperphonated wails and squeals as protophones), and where all stimuli include low f0 (which would include wails with low frequency regime segments and growls as protophones). We also excluded glottal bursts which are extremely common in distress sounds that we call whimper. An additional line of research should address whimpers and vocants with varying amplitudes and durations of glottal bursts.

It remains unclear how to deal with the extreme complexity of all the relevant categories, cries, whimpers, whines and protophones. The complexity owes to recombinability within breath groups of all the markers that pertain to each of the categories: glottal bursts, catch breaths, nuclei of at least 7 vibratory regime types, and a vast array of possible supraglottal articulations. We have seen examples where a single utterance includes segments that would in isolation be judged to represent all four categories, cry, whimper, whine and protophone. At the same time there are prototypical versions of each category to form a basis for initial efforts; our approach is to start with comparisons of relatively limited complexity, building our investigations of more complicated utterances gradually.

An additional direction for future research would evaluate all these kinds of sounds at the level of sequences of utterances. It seems clear that caregivers are heavily influenced by sequences that yield a more stable impression than individual utterances of the degree of comfort or distress expressed by infants. We have started at the level of utterances in our analyses, but eventually it will be necessary to address the organization of distress vocalizations and protophones in larger rhythmic units.

Chapter 5: General conclusion

Study 1 found that caregivers tended to take vocal turns with infant protophones, suggesting that they use the protophones to facilitate protoconversation from the first months of life. On the other hand, they were more likely to overlap their responsive vocalizations with cries, suggesting that they calm/soothe the crying infant, but do not converse with cries. This is the first study to provide empirical evidence showing that caregivers' timing responses to protophones and to cries are systematically different. This kind of evidence is important because it supports that protophones and cries are functionally different, and that caregivers are intuitively aware of the difference, an awareness that can help foster language development of infants even from the beginning of life.

Study 2 showed that adult listeners were reliably able to identify wails and whines (> 80% accuracy) in behavioral (reaction times) and physiological tasks (pupillometry) regardless of noise interference. The listeners identified wails better (higher accuracy) in a no-noise condition and they identified wails faster (shorter reaction times) in both music and speech-babble noise conditions. This result corresponds to the "fast guessing" principle, meaning that with more cognitive load (more noise), people tend to judge faster. High cognitive load in identifying sounds with noise interference was also supported by the pupillometry results. The listeners showed significantly greater pupil dilation when judging whines (which presumably require more cognitive resources to identify) in the speech-babble noise condition.

Study 3 found high agreement on judging the level of distress in infant vocalizations (wails, whines and vocants) among adult listeners. This high agreement was not significantly affected by experience in infant vocalization coding/training. Duration of an utterance, f_o max, moments of the long-term average spectrum, and numbers of vibratory regimes were the best predictors of judgments of distress level. Although both experienced and inexperienced adult listeners highly agreed with each other on rating varying degrees of distress, they used the acoustic parameters for judgments in significantly different ways. Mean of RMS (amplitude), minimum of the low-high spectral ratio (at 2 kHz), moments of the long-term average spectrum, and number of regimes were used to different extents by the experienced and inexperienced groups.

Conclusions regarding the three studies: Infants produce sounds varying from protophones to cries. These sounds are reliably differentiable in terms of perceived distress level, even though both cries and protophones are extremely diverse, creating a challenge for research to explain how human listeners recognize and judge them. Both caregivers and non-parent adults judge the distress levels of these seemingly complex infant sounds with high reliability. This kind of capability supports the idea that humans are naturally capable of intuitive parenting, an ability deemed critical in infant cognitive, social, and language development.

We see the potential value of further research on caregivers' differential response timing to various subcategories of vocalizations of infants, for example, the three prototypical categories of distress, wail cry, whine, and whimper, along with the many recombinations that occur within real infant utterances of regime segments corresponding to these types. Similarly we know nothing yet of possible timing differences of parent responses to subcategories of protophones, including single and multiple syllable canonical and non-canonical utterances. It makes sense that parents may fine tune their timing of responses to different infant sounds depending on complexity, emotional content, and speech-likeness. Considering the critical roles of caregivers in infant development, such research could lay foundations for a better understanding of patterns of interaction presumed to correlate with cognitive and emotional development.

It would also be of value to examine acoustic features of infant-directed speech in response to infant protophones and cries. Such research could provide better understanding of how caregivers coordinate and adjust differently to different types of infant vocalizations. Better understanding of early vocal/speech development and social interaction (including cross-cultural differences) is critical for speech and language development. The kind of research reported in this dissertation is now ripe for evaluation across genders, cultures, and across families with infants at risk for significant social and language disorders.

References

Abney, D. H. (2016). The complexity matching hypothesis for human communication. Dissertattion. Merced(CA): University of California, Merced

Ainsworth, M., & Bell, S. M. (1970). Attachment, exploration, and separation: Illustrated by the behavior of one-year-olds in a strange situation. *Child development*, *41*(1), 49-67.

Ainsworth, M. S., and Bell, S. M. (1974). "Mother-Infant interaction and the development of competence", in The Growth of Competence, ed. K. J. Connolly and J. Bruner (New York, NY: Academic Press), 131-164.

Ainsworth, M., Blehar, M., Waters, E., & Wall, S. N. (1978). *Patterns of attachment: A psychological study of the strange situation*. Hove: Psychology Press.

Ainsworth, M. S. (1979). Infant-mother attachment. Am. Psychol. 34:10. doi: 10.1037/0003-066X.34.10.932

Akeroyd, M. A. (2008). Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults. *International Journal of Audiology, 47 Suppl 2,* S53-71. doi:10.1080/14992020802301142

Anderson, S., White-Schwoch, T., Parbery-Clark, A., & Kraus, N. (2013). A dynamic auditory-cognitive system supports speech-in-noise perception in older adults. *Hearing Research, 300C,* 18-32. doi: S0378-5955(13)00074-9 [pii]10.1016/j.heares.2013.03.006

Andreassi, J. L. (2000). Pupillary response and behavior *Psychophysiology: Human Behavior & Physiological Response* (pp. 289-307): Lawrence Erbaulm Associates, Inc.

Arend, R., Gove, F. L., & Sroufe, L. A. (1979). Continuity of individual adaptation from infancy to kindergarten: A predictive study of ego-resiliency and curiosity in preschoolers. *Child development*, *50*(4), 950-959.

Awan, S. N., Roy, N., & Dromey, C. (2009). Estimating dysphonia severity in continuous speech: application of a multi-parameter spectral/cepstral model. Clinical linguistics & phonetics, 23(11), 825-841.

Awan, S. N. (2011). Analysis of dysphonia in speech and voice (ADSV): An application guide. Montvale, NJ: KayPentax.

Babyak, M. A. (2004). What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine, 66*(3), 411-421.

Bakeman, R., and Brown, J. V. (1980). Early interaction: Consequences for social and mental development at three years. Child Dev. 51:2, 437–447.

Bateson, M. C. (1975). Mother-infant exchanges: the epigenesis of conversational interaction. Ann. N. Y. Acad. Sci. 263:1, 101–113.

Barr, R., Kramer, M., Boisjoly, C., McVey-White, L., & Pless, I. (1988). Parental diary of infant cry and fuss behaviour. *Archives of Disease in Childhood, 63*(4), 380-387.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychol Bull, 91*, 276–292.

Beckwith, L., Cohen, S. E., Kopp, C. B., Parmelee, A. H., and Marcy, T. G. (1976). Caregiver-infant interaction and early cognitive development in preterm Infants. Child Dev. 47:3. doi: 10.2307/1128171

Beebe, B., Alson, D., Jaffe, J., Feldstein, S., & Crown, C. (1988). Vocal congruence in mother-infant play. J Psycholinguist Res. 17:3, 245–259.

Bell, S. M., & Ainsworth, M. D. S. (1972). Infant crying and maternal responsiveness. *Child development, 43*(4), 1171-1190.

Bidelman, G. M. (2016). Relative contribution of envelope and fine structure to the subcortical encoding of noise-degraded speech. *Journal of the Acoustical Society of America, 140*(4), EL358-363.

Bidelman, G. M. (2017). Communicating in challenging environments: Noise and reverberation. In N. Kraus, S. Anderson, T. White-Schwoch, R. R. Fay & A. N. Popper (Eds.), *Springer Handbook of Auditory Research: The frequency-following response: A window into human communication*. New York: Springer Nature.

Bidelman, G. M., & Howell, M. (2016). Functional changes in inter- and intra-hemispheric auditory cortical processing underlying degraded speech perception. *Neuroimage, 124*, 581-590.

Bidelman, G. M., Jennings, S. G., & Strickland, E. A. (2015). PsyAcoustX: A flexible MATLAB® package for psychoacoustics research. *Frontiers in Psychology, 6*(1498), 1-11.

Bidelman, G. M., & Krishnan, A. (2010). Effects of reverberation on brainstem representation of speech in musicians and non-musicians. *Brain Research, 1355*, 112-125.

Billings, C. J., McMillan, G. P., Penman, T. M., & Gille, S. M. (2013). Predicting perception in noise using cortical auditory evoked potentials. *Journal of the Association for Research in Otolaryngology, 14*(6), 891-903. doi: 10.1007/s10162-013-0415-y

Binder, J. R., Liebenthal, E., Possing, E. T., Medler, D. A., & Ward, B. D. (2004). Neural correlates of sensory and decision processes in auditory object identification. *Nature Neuroscience, 7*(3), 295-301. doi: 10.1038/nn1198nn1198 [pii]

Blehar, M. C., Lieberman, A. F., & Ainsworth, M. D. S. (1977). Early face-to-face interaction and its relation to later infant-mother attachment. *Child development*, *48*(1), 182-194.

Bloom, K., Russell, A., and Wassenberg, K. (1987). Turn taking affects the quality of infant vocalizations. J Child Lang. 14:2, 211–227.

Bögels, S., and Levinson, S. C. (2017). The brain behind the response: insights into turn-taking in conversation from neuroimaging. Res Lang Soc Interac, *50*(1), 71–89.

Bornstein, M. H., and Bruner, J. S. (2014). Interaction in human development. Hove: Psychology Press.

Bowlby, J. (1969). *Attachment : Attachment and Loss*. New York: Basic Books.

Crockenberg, S. (1983). Early mother and infant antecedents of Bayley Scale Performance at 21 months. *Developmental Psychology, 19*(5), 727.

Brown, P. (1998). Conversational structure and language acquisition: The role of repetition in Tzeltal. J Linguist Anthropol. 8:2, 197–221.

Bruner, J. (1983). Child's talk. New York: Oxford University Press.

Buder, E. H., Chorna, L. B., Oller, D. K., & Robinson, R. B. (2008). Vibratory Regime Classification of Infant Phonation. *Journal of Voice*, *22*(5), 553–564.

Buder, E. H., & Strand, E. A. (2003). Quantitative and Graphic Acoustic Analysis of Phonatory Modulations. *Journal of Speech Language and Hearing Research*, *46*(2), 475-490.

Clayman, S. E. (2013). "Turn-constructional units and the transition-relevance place", in The Handbook of Conversation Analysis, ed. J. Sidnel1 and T. Stivers (Hoboken, NJ: Wiley-Blackwell), 51–166.

Coates, J. (1994). "No gap, lots of overlap; Turn-taking patterns in the talk of women friends", in Researching Language and Literacy in Social Context: A Reader, ed. D. Graddol, J. Maybin, and B. Stierer (Philadelphia: Multilingual Matters), 177-192.

Coates, J. (1997). "The construction of a collaborative floor in women's friendly talk", in Conversation: Cognitive, communicative and social perspectives, ed. T. Givón (Philadelphia: Benjamins), 55-89.

Cox, D. R., & Stuart, A. (1955). Some quick sign tests for trend in location and dispersion. Biometrika, 42(1/2), 80-95.

Delgado, R. E. (2018). AACT (Action Analysis Coding and Training). Miami, FL: Intelligent Hearing Systems.

de Gee, J. W., Knapen, T., & Donner, T. H. (2014). Decision-related pupil dilation reflects upcoming choice and individual bias. *Proceedings of the National Academy of Sciences, 111*(5), E618-E625.

Dominguez, S., Devouche, E., Apter, G., & Gratier, M. (2016). The Roots of Turn-Taking in the Neonatal Period. *Infant and Child Development*, *25*(3), 240–255.

Downey, D. B. (1995). When bigger is not better: Family size, parental resources, and children's educational performance. American Sociological Review, 60(5), 746-761.

Egeland, B., & Farber, E. A. (1984). Infant-mother attachment: Factors related to its development and changes over time. *Child development*, *55*(3), 753-771.

Elias, G., Hayes, A., and Broerse, J. (1986). Maternal control of co-vocalization and inter-speaker silences in mother-infant vocal engagements. J. Child Psychol. Psychiatry. 27:3, 409–415.

Erickson, M. F., Sroufe, L. A., & Egeland, B. (1985). The relationship between quality of attachment and behavior problems in preschool in a high-risk sample. *Monographs of the society for research in child development*, *50*(1/2)147-166.

Fant, G. (1960) Acoustic theory of speech production. The Hague: Mouton.

Fagan, M. K., and Doveikis, K. N. (2017). Ordinary interactions challenge proposals that maternal verbal responses shape infant vocal development. J Speech Lang Hear Res. 60:10, 2819–2827.

Farran, L. K., Lee, C. C., Yoo, H., and Oller, D. K. (2016). Cross-cultural register differences in infant-directed speech: An initial study. PLoS One, 11(3), e0151518.

Feng, S., & Bidelman, G. M. (2015). *Music listening and song familiarity modulate mind wandering and behavioral success during lexical processing*. Paper presented at the Annual Meeting of the Cognitive science Society (CogSci 2015), Pasadena, CA.

Feldman, R. (2007a). Parent-infant synchrony: Biological foundations and developmental outcomes. Curr. Dir. Psychol. Sci. *16*:6. doi:10.1111/j.1467-8721.2007.00532.x

Feldman, R. (2007b). Parent-infant synchrony and the construction of shared timing; physiological precursors, developmental outcomes, and risk conditions. J Child Psychol Psychiatry. doi:10.1111/j.1469-7610.2006.01701.x

Fogel, A., Toda, S., and Kawai, M. (1988). Mother-infant face-to-face interaction in Japan and the United States: A laboratory comparison using 3-month-old infants. Dev. Psychol. 24:3, 398-406.

Forrest, K., Weismer, G., Milenkovic, P., & Dougall, R. N. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *The Journal of the Acoustical Society of America*, *84*(1), 115–123.

Franklin, B., Warlaumont, A. S., Messinger, D., Bene, E., Nathani Iyer, S., Lee, C.-C., … Oller, D. K. (2014). Effects of parental interaction on infant vocalization rate, variability and vocal type. Lang Learn Dev. 10:3, 279–296.

Frick, R. W. (1985). Communicating emotion: The role of prosodic features. *Psychological Bulletin*, *97*(3), 412-429.

Frodi, A., & Senchak, M. (1990). Verbal and behavioral responsiveness to the cries of atypical infants. *Child development*, *61*(1), 76-84.

Fuller, B. F., & Horii, Y. (1986). Differences in fundamental frequency, jitter, and shimmer among four types of infant vocalizations. *Journal of Communication Disorders*, *19*(6), 441–447.

Füllgrabe, C., & Rosen, S. (2016). Investigating the role of working memory in speech-in-noise identification for listeners with normal hearing. In P. van Dijk, D. Başkent, E. Gaudrain, E. de Kleine, A. Wagner & C. Lanting (Eds.), *Physiology, Psychoacoustics and Cognition in Normal and Impaired Hearing* (pp. 29-36). Cham: Springer International Publishing.

Gewirtz, J. L., and Pelaez-Nogueras, M. (1992). BF Skinner's legacy in human infant behavior and development. Am Psychol, 47:11, 1411-1422.

Ginsburg, G. P., and Kilbourne, B. K. (1988). Emergence of vocal alternation in mother-infant interchanges. J Child Lang, 15:2, 221–235.

Goldstein, M. H., King, A. P., and West, M. J. (2003). Social interaction shapes babbling: Testing parallels between birdsong and speech. Proc Natl Acad Sci. 100:13, 8030–8035.

Goldstein, M. H., Schwade, J. A., and Bornstein, M. H. (2009). The value of vocalizing: Five-month-old infants associate their own noncry vocalizations with responses from caregivers. Child Dev. 80:3, 636–644.

Gratier, M. (2003). Expressive timing and interactional synchrony between mothers and infants: Cultural similarities, cultural differences, and the immigration experience. Cogn Dev, 18:4, 533–554.

Gratier, M., Devouche, E., Guellai, B., Infanti, R., Yilmaz, E., and Parlato-Oliveira, E. (2015). Early development of turn-taking in vocal interaction between mothers and infants. Front. Psychol, 6:1167. doi: 10.3389/fpsyg.2015.01167.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Green, J. A., Jones, L. E., & Gustafson, G. E. (1987). Perception of cries by parents and nonparents: Relation to cry acoustics. *Developmental Psychology, 23*(3), 370.

Green, J. A., Whitney, P. G., & Potegal, M. (2011). Screaming, yelling, whining, and crying: categorical and intensity differences in vocal expressions of anger and sadness in children's tantrums. *Emotion, 11*(5), 1124-1133.

Green, J. A., Gustafson, G. E., & McGhie, A. C. (1998). Changes in Infants' Cries as a Function of Time in a Cry Bout. *Child Development*, *69*(2), 271–279.

Green, J. A., Jones, L. E., & Gustafson, G. E. (1987). Perception of cries by parents and nonparents: Relation to cry acoustics. *Developmental Psychology*, *23*(3), 370-382.

Grice, G. R., Nullmeyer, R., & Spiker, V. A. (1982). Human reaction time: Toward a general theory. *Journal of Experimental Psychology: General, 111*(1), 135-153. doi: 10.1037/0096-3445.111.1.135

Gros-Louis, J., West, M. J., Goldstein, M. H., and King, A. P. (2006). Mothers provide differential feedback to infants' prelinguistic sounds. Int J Behav Dev. 30:6, 509–516.

Gustafson, G. E., & Green, J. A. (1989). On the importance of fundamental frequency and other acoustic features in cry perception and infant development. *Child development*, *60*(4), 772-780.

Hayashi, M. (2013). "Turn allocation and turn sharing", in The handbook of conversation analysis, ed. J. Sidnell and T. Stivers (Hoboken, NJ:Wiley-Blackwell), 167–190.

Heldner, M., and Edlund, J. (2010). Pauses, gaps and overlaps in conversations. J Phon. 38:4, 555–568.

Helfer, K., & Wilber, L. (1990). Hearing loss, aging, and speech perception in reverberation and in noise. *Journal of Speech and Hearing Research*, *33*, 149-155.

Heman-Ackah, Y. D., Heuer, R. J., Michael, D. D., Ostrowski, R., Horman, M., Baroody, M. M., … Sataloff, R. T. (2003). Cepstral peak prominence: A more reliable measure of dysphonia. *Annals of Otology, Rhinology and Laryngology*, *112*(4), 324–333.

Hillenbrand, J., & Houde, R. A. (1996). Acoustic Correlates of Breathy Vocal Quality: Dysphonic Voices and Continuous Speech. *Journal of Speech Language and Hearing Research*, *39*(2), 311-321.

Hilbrink, E. E., Gattis, M., and Levinson, S. C. (2015). Early developmental changes in the timing of turn-taking: a longitudinal study of mother-infant interaction. Front. Psychol. 6: 1492. doi: 10.3389/fpsyg.2015.01492

Hoff, E. (2004). *Language Development*. Belmont, CA: Wadsworth Publishing.

Holler, J., Kendrick, K. H., Casillas, M., and Levinson, S. C. (2015). Editorial: Turn-taking in human communicative interaction. Front. Psychol. 6:1919. doi: 10.3389/fpsyg.2015.01919

Hollien, H., Girard, G. T., & Coleman, R. F. (1977). Vocal fold vibratory patterns of pulse register phonation. *Folia Phoniatrica et Logopaedica*, *29*(3), 200–205.

Hsu, H.-C., and Fogel, A. (2003). Social regulatory effects of infant nondistress vocalization on maternal behavior. Dev. Psychol. 39:6, 976-991.

Hsu, H.-C., Fogel, A., and Messinger, D. S. (2001). Infant non-distress vocalization during mother-infant face-to-face interaction: Factors associated with quantitative and qualitative differences. Infant Behav Dev. 24:1, 107–128.

Indefrey, P. (2011). The spatial and temporal signatures of word production components: a critical update. Front. Psychol. doi.org/10.3389/fpsyg.2011.00255

Indefrey, P., and Levelt, W. J. M. (2004). The spatial and temporal signatures of word production components. Cogn. 1, 101–144.

Irwin, J. R. (2003). Parent and nonparent perception of the multimodal infant cry. *Infancy, 4*(4), 503-516.

Ishihara, H., Yoshikawa, Y., Miura, K., and Asada, M. (2009). How caregiver's anticipation shapes infant's vowel through mutual imitation. IEEE Trans. Auton. Mental Develop. 1:4, 217–225.

Jaffe, J., Beebe, B., Feldstein, S., Crown, C. L., Jasnow, M. D., Rochat, P., and Stern, D. N. (2001). Rhythms of dialogue in infancy: Coordinated timing in development. Monogr Soc Res Child Dev. 66:2, i-149.

Jasnow, M., and Feldstein, S. (1986). Adult-like temporal characteristics of mother-infant vocal interactions. Child Dev. 57:3, 754–761.

Jhang, Y., and Oller, D. K. (2017). Emergence of functional flexibility in infant vocalizations of the first 3 months. Front. Psychol. doi: 10.3389/fpsyg.2017.00300

Kärtner, J., Keller, H., and Yovsi, R. D. (2010). Mother–infant interaction during the first 3 months: The emergence of culture-specific contingency patterns. Child Dev. 81:2, 540–554.

Kaye, K., and Fogel, A. (1980). The temporal structure of face-to-face communication between mothers and infants. Dev. Psychol. 16:5, 454-464.

Keller, T. (2003). Parental images as a guide to leadership sensemaking: An attachment perspective on implicit leadership theories. *The Leadership Quarterly, 14*(2), 141-160.

Keller, H., Chasiotis, A., Risau-Peters, J., Volker, S., Zach, U., and Restemeier, R. (1996). Psychobiological aspects of infant crying. Infant Child Dev. 5:1, 1-13.

Keller, H., Kärtner, J., Borke, J., Yovsi, R., & Kleis, A. (2005). Parenting styles and the development of the categorical self: A longitudinal study on mirror self-recognition in Cameroonian Nso and German families. Int J Behav Dev. *29*:6, 496–504.

Keller, H., Lohaus, A., Völker, S., Cappenberg, M., and Chasiotis, A. (1999). Temporal contingency as an independent component of parenting behavior. Child Dev. 70:2, 474–485.

Keller, H., and Schölmerich, A. (1987). Infant vocalizations and parental reactions during the first 4 months of Life. Dev. Psychol. 23:1. doi: 10.1037/0012-1649.23.1.62

Kent, R. D., & Murray, A. D. (1982). Acoustic features of infant vocalic utterances at 3, 6, and 9 months. *The Journal of the Acoustical Society of America*, *72*(2), 353–365.

Killion, M. C., Niquette, P. A., Gudmundsen, G. I., Revit, L. J., & Banerjee, S. (2004). Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America, 116*(4 Pt 1), 2395-2405.

Kim, P., Feldman, R., Mayes, L. C., Eicher, V., Thompson, N., Leckman, J. F., & Swain, J. E. (2011). Breastfeeding, brain activation to own infant cry, and maternal sensitivity. *Journal of child psychology and psychiatry, 52*(8), 907-915.

Kobak, R. R., & Sceery, A. (1988). Attachment in late adolescence: Working models, affect regulation, and representations of self and others. *Child development*, *59*(1), 135-146.

Koopmans-van Beinum, F. J., and van der Stelt, J. M. (1986). "Early stages in the development of speech movements", in Precursors of early speech, ed. B. Lindblom and R. Zetterstrom (New York: Stockton), 37–50.

Kozou, H., Kujala, T., Shtyrov, Y., Toppila, E., Starck, J., Alku, P., & Naatanen, R. (2005). The effect of different noise types on the speech and non-speech elicited mismatch negativity. *Hearing Research, 199*(1-2), 31-39. doi: S0378-5955(04)00234-5 [pii]10.1016/j.heares.2004.07.010

LaGasse, L. L., Neal, A. R., & Lester, B. M. (2005). Assessment of infant cry: Acoustic cry analysis and parental perception. *Mental Retardation and Developmental Disabilities Research Reviews*.

Lavelli, M., and Fogel, A. (2002). Developmental changes in mother-infant face-to-face communication: birth to 3 months. Dev. Psychol. 38:2, 288-305.

Laufer, M. Z., & Horii, Y. (1977). Fundamental frequency characteristics of infant non-distress vocalization during the first twenty-four weeks. *Journal of Child Language*, *4*(2), 171–184.

Leger, D. W., Thompson, R. A., Merritt, J. A., & Benz, J. J. (1996). Adult perception of emotion intensity in human infant cries: Effects of infant age and cry acoustics. *Child development, 67*(6), 3238-3249.

Lehtonen, J., and Sajavaara, K. (1985). The silent finn. New York: Ablex Publishing Corporation.

Lester, B. M., & Boukydis, C. Z. (1992). No language but a cry. In H. Papoušek, U. Jürgens, & M. Papoušek (Ed.), Nonverbal vocal communication: Comparative and developmental approaches (145-73). New York: Cambridge University Press.

Levinson, S. C. (2016). Turn-taking in human communication–origins and implications for language processing. Trends Cogn. Sci. 20:1, 6–14.

Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. Front. Psychol. 6:731. doi: 10.3389/fpsyg.2015.00731

Lewis, M., and Coates, D. L. (1980). Mother-infant interaction and cognitive development in twelve-week-old infants. Infant Behav Dev. 3, 95–105.

Lewis, M., and Goldberg, S. (1969). Perceptual-cognitive development in infancy: A generalized expectancy model as a function of the mother-infant interaction. Merrill-Palmer Q Beh. 15:1, 81–100.

Liang, K.-Y., and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. Biometrika. 73:1, 13–22.

Locke, J. L. (1989). Babbling and early speech: Continuity and individual differences. First Lang. 9:6, 191–205.

Locke, J. (1993). The Child's Path to Spoken Language.Boston: Harvard University Press.

Lynch, M. P., Oller, D. K., Steffens, M. L., & Buder, E. H. (1995). Phrasing in prelinguistic vocalizations. *Developmental Psychobiology*, *28*(1), 3–25.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory : A user's guide* (2nd ed.). Mahwah, N.J.: Lawrence Erlbaum Associates, Inc.

Maltz, D. N., and Borker, R. A. (1982). "A cultural approach to male-female miscommunication", in A Cultural Approach to Interpersonal Communication: Essential Readings, ed. L. Monaghan, J. E. Goodman and J. M. Robinson (Hoboken, NJ: Wiley-Blackwell), 68–185.

Mampe, B., Friederici, A. D., Christophe, A., and Wermke, K. (2009). Newborns' cry melody is shaped by their native language. Curr. Biol. doi: 10.1016/j.cub.2009.09.064

Marsh, J. E., Hughes, R. W., & Jones, D. M. (2009). Interference by process, not content, determines semantic auditory distraction. *Cognition, 110*, 23–28.

McCullagh, P., and Nelder, J. A. (1989). Generalized linear models. London, England: Chapman and Hall.

Mende, W., Herzel, H., & Wermke, K. (1990). Bifurcations and chaos in newborn infant cries. *Physics Letters A, 145*(8-9), 418-424.

Michelsson, K., Järvenpää, A. L., & Rinne, A. (1983). Sound spectrographic analysis of pain cry in preterm infants. *Early Human Development*, *8*(2), 141–149.

Michelsson, K., & Michelsson, O. (1999). Phonation in the newborn, infant cry. *International Journal of Pediatric Otorhinolaryngology* (49), 297–301.

Michelsson, K., Raes, J., Thoden, C. J., & Wasz- Hockert, 0. (1982). Sound spectrographic cry analysis in neonatal diagnostics: An evaluative study. Journal of Phonetics, 10, 79-88.

Milenkovic P. (2018). TF32 (computer software), Madison, WI: University of Wisconsin-Madison.

Miura, K., Yoshikawa, Y., and Asada, M. (2007). Unconscious anchoring in maternal imitation that helps find the correspondence of a caregiver's vowel categories. Adv Robot. 21:13, 1583–1600.

Murray, A. D. (1979). Infant crying as an elicitor of parental behavior: an examination of two models. *Psychological Bulletin, 86*(1), 191-215.

Murray, L., Fiori-Cowley, A., Hooper, R., and Cooper, P. (1996). The impact of postnatal depression and associated adversity on early mother-infant interactions and later infant outcome. Child Dev. 67:5. doi: 10.1111/j.1467-8624.1996.tb01871.x

Murray, L. and Trevarthen, C. (1985). "Emotional regulation of interactions between two-month-olds and their mothers", in Social perception in infants, ed. T. M. Field and N. A. Fox (New York: Ablex), 177–197.

Nathani-Iyer, S., and Ertmer, D. J. (2014). Relationships between vocalization forms and functions in infancy: preliminary implications for early communicative assessment and intervention. Am J Speech Lang Pathol. 23:4, 587–598.

Nathani-Iyer, S., Ertmer, D. J., & Stark, R. E. (2006). Assessing vocal development in infants and toddlers. *Clinical Linguistics & Phonetics, 20*(5), 351–369.

Nathani, S., & Oller, D. K. (2001). Beyond ba-ba and gu-gu: Challenges and strategies in coding infant vocalizations. *Behavior Research Methods, Instruments, and Computers, 33*(3), 321–330.

Oller, D. K. (1980). "The emergence of the sounds of speech in infancy", in Child phonology, Volume 1. Production, ed. G. H. Yeni-Komshian, J. F. Kavanaugh, and C. A. Ferguson (New Yok:Academic Press), 93–112.

Oller, D. K. (1981). "Infant vocalizations: Exploration and reflexivity", in Language behavior in infancy and early childhood, ed. R. E. Stark (New York: Elsevier North Holland), 85–104.

Oller, D. K. (2000). The emergence of the capacity for speech. Mahwah, NJ: Erlbaum.

Oller, D. K., Buder, E. H., Ramsdell, H. L., Warlaumont, A. S., Chorna, L., & Bakeman, R. (2013). Functional flexibility of infant vocalization and the emergence of language. *Proceedings of the National Academy of Sciences, 110*(16), 6318-6323.

Oller, D. K., Niyogi, P., Gray, S., Richards, J., Gilkerson, J., Xu, D., . . . Warren, S. (2010). Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proceedings of the National Academy of Sciences, 107*(30), 13354-13359.

Oller, D. K., Wieman, L. A., Doyle, W. J., and Ross, C. (1976). Infant babbling and speech. J Child Lang. 3:1, 1–11.

Oller, D. K., Vohr, B. R., Caskey, M., Yoo, H., Bene, E. R., Jhang, Y., Warlaumont, A. Griebel, U., and Buder, E. H. (2014). Infant vocalization and birdsong: an infrastructural view. Society for Neuroscience satellite meeting: Birdsong4: Rhythms and Clues from Neurons to Behavior, Washington, DC

Oller, D. K., Griebel, U., & Warlaumont, A. S. (2016). Vocal Development as a Guide to Modeling the Evolution of Language. *Topics in Cognitive Science*, *8*(2), 382–392.

Papoušek, H., and Papoušek, M. (1987). "Intuitive parenting: A dialectic counterpart to the infant’s integrative competence", in Handbook of infant development, ed. J. D. Osofsky (New York: Wiley), 669–720.

Papoušek M (1994) Vom ersten Schrei zum ersten Wort: Anfänge der Sprachentwickelung in der vorsprachlichen Kommunikation. Bern:Verlag Hans Huber.

Papoušek, M. (1995). "Origins of reciprocity and mutuality in prelinguistic parent-infant “dialogues” ", in Mutualities in dialogue, ed. I. Markova, C. F. Graumann, and K. Foppa (Cambridge, England: Cambridge University Press), 58-81.

Perham, N., & Currie, H. (2014). Does listening to preferred music improve reading comprehension performance? . *Applied Cognitive Psychology, 28,* 279-284.

Petrovich-Bartell, N., Cowan, N., & Morse, P. A. (1982). Mothers' perceptions of infant distress vocalizations. *Journal of Speech, Language, and Hearing Research, 25*(3), 371-376.

Porter, F. L., Miller, R. H., & Marshall, R. E. (1986). Neonatal pain cries: effect of circumcision on acoustic features and perceived urgency. *Child Development*, *57*(3), 790–802.

Prodi, N., Visentin, C., & Feletti, A. (2013). On the perception of speech in primary school classrooms: Ranking of noise interference and of age influence. *The Journal of the Acoustical Society of America*, *133*(1), 255-268.

Rabain-Jamin, J., and Sabeau-Jouannet, E. (1997). Maternal speech to 4-month-old infants in two cultures: Wolof and French. Int J Behav Dev. 20(3), 425–451.

Richman, A. L., Miller, P. M., and LeVine, R. A. (1992). Cultural and educational variations in maternal responsiveness. Dev. Psychol. 28:4, 614-621.

Robb, M. P., & Saxman, J. H. (1988). Acoustic observations in young children's non-cry vocalizations. The Journal of the Acoustical Society of America, 83(5), 1876-1882.

Rochat, P. (1998, April). The two month revolution. Symposium conductedat the 11th International Conference on Infant Studies, Atlanta, Georgia

Roug, L., Landberg, I., & Lundberg, L.-J. (1989). Phonetic development in early infancy: A study of four Swedish children during the first eighteen months of life. *Journal of Child Language, 16*(1), 19–40.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn taking in conversation. Lang. 50, 696-735.

Salamé, P., & Baddeley, A. (1989). Effects of background music on phonological short-term memory. *The Quarterly Journal of Experimental Psychology, 41*(1), 107-122.

Scheiner, E., Hammerschmidt, K., Jürgens, U., and Zwirner, P. (2002). Acoustic analyses of developmental changes and emotional expression in the preverbal vocalizations of infants. J Voice. 16:4, 509–529.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech communication, 40*(1), 227-256.

Schuetze, P., & Zeskind, P. S. (2001). Relations between women's depressive symptoms and perceptions of infant distress signals varying in pitch. *Infancy, 2*(4), 483-499.

Sidnell, J., and Stivers, T. (2012). The Handbook of Conversation Analysis. doi: 10.1002/9781118325001

Sroufe, L. A., and Waters, E. (1977). Attachment as an organizational construct. Child Dev. 48:4, 1184–1199.

Standke, R. (1992). *Methods of digital speech analysis in research on vocal communication.* Frankfurt: Peter Lang.

Stark, R. E., Rose, S. N., & McLagen, M. (1975). Features of infant sounds: The first eight weeks of life. *Journal of Child Language*, *2*(2), 205–221.

Stark, R. E., Rose, S. N., & Benson, P. J. (1978). Classification of infant vocalization. International Journal of Language & Communication Disorders, 13(1), 41-47.

Stark, R. E. (1980). "Stages of speech development in the first year of life", in Child phonology: Volume 1. Production, ed. G. H. Yeni-Komshian, J. F. Kavanaugh, and C. A. Ferguson (New Yok:Academic Press), 73–92.

Stark, R. E. (1981). Infant vocalization: A comprehensive view. Infant Ment Health J. 2:2, 118–128.

Stark, R. E. (1989). Temporal patterning of cry and non-cry sounds in the first eight months of life. *First Language*, *9*(6), 107–136.

Stark, R. E., Bernstein, L. E., & Demorest, M. E. (1993). Vocal communication in the first 18 months of life. *Journal of Speech and Hearing Research*, *36*(3), 548–558.

Stern, D. N., Jaffe, J., Beebe, B., & Bennett, S. L. (1975). Vocalizing in unison and in alternation: Two modes of communication within the mother-infant dyad. Ann. N. Y. Acad. Sci. 263:1, 89–100.

Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., … SC, L. (2009). Universals and cultural variation in turn-taking in conversation. Proc. Natl. Acad. Sci. 106(26), 10587–10592.

Stoel-Gammon, C. (1989). Prespeech and early speech development of two late talkers. First Lang. 9:6, 207–223.

Striano, T., and Reid, V. M. (2006). Social cognition in the first year. Trends Cogn Sci. 10:10, 471–476.

Stross, B. (1972). Verbal processes in Tzeltal speech socialization. Anthro Ling. 14:1, 1–13.

Swain, J. E., Tasgin, E., Mayes, L. C., Feldman, R., Todd Constable, R., & Leckman, J. F. (2008). Maternal brain response to own baby-cry is affected by cesarean section delivery. *Journal of child psychology and psychiatry*, *49*(10), 1042-1052.

Takahashi, D. Y., Fenley, A. R., Teramoto, Y., Narayanan, D. Z., Borjon, J. I., Holmes, P., & Ghazanfar, A. A. (2015). The developmental dynamics of marmoset monkey vocal production. *Science*, *349*(6249), 734–738.

Tanaka, H. (1999). Turn-taking in Japanese conversation: A study in grammar and interaction Amsterdam: Benjamins Publishing Company.

Titze, I., Riede, T., & Popolo, P. (2008). Nonlinear source–filter coupling in phonation: Vocal exercises. *The Journal of the Acoustical Society of America*, *123*(4), 1902–1915.

Thoden, C-J., & Koivisto, M. (1980). Acoustic analyses of the normal pain cry. In T. Murry & J. Murry (Eds.), *Infant communication: Cry and early speech* (pp. 124-151). Houston, TX: College-Hill.

Tomasello, M. (1992). The social bases of language acquisition. Soc Dev. 1:1. doi: 10.1111/j.1467-9507.1992.tb00135.x

Trevarthen, C. (1977). "Descriptive analyses of infant communicative behavior", in Studies in mother-infant interaction, ed. H. R. Schaffer (London: Academic Press), 227–270.

Trevarthen, C., and Aitken, K. J. (2001). Infant intersubjectivity: Research, theory, and clinical applications. J Child Psychol Psychiatry. 42:1, 3–48.

Tronick, E., Als, H., and Brazelton, T. B. (1980). Monadic phases: A structural descriptive analysis of infant-mother face to face interaction. Merrill-Palmer Q Beh. 26:1, 3–24.

Truby, H. M., & Lind, J. (1965). Cry sounds of the newborn infant. *Acta Paediatrica Scandinavica*, *Suppl.163*, 8–54 ST–Cry sounds of the newborn infant.

Vihman, M. M., Macken, M. A., Miller, R., Simmons, H., and Miller, J. (1985). From babbling to speech: A re-assessment of the continuity issue. Lang, 61:2, 397–445.

Várallyay, J., & Benyó, Z. (2007). Melody Shape – A Suggested Novel Attribute for the Biomedical Analysis of the Infant Cry. Proceedings of the 29th Annual International Conference of the IEEE EMBS.

Völker, S., Keller, H., Lohaus, A., Cappenberg, M., and Chasiotis, A. (1999). Maternal interactive behaviour in early infancy and later attachment. Int J Behav Dev. 23:4. doi: 10.1080/016502599383603

Wasz-Höckert, O., Michelsson, K., & Lind, J. (1985). Twenty-Five Years of Scandinavian Cry Research. In Boukydis & Lester (Ed.), *Infant Crying* (pp. 83-104). New York: Springer.

Wermke, K., Mende, W., Manfredi, C., & Bruscaglioni, P. (2002). Developmental aspects of infant's cry melody and formants. *Medical Engineering and Physics*, *24*(7–8), 501–514.

Wilson, M., and Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. Psychon Bull Rev. 12:6, 957–968.

Winn, M. B., Edwards, J. R., & Litovsky, R. Y. (2015). The impact of auditory spectral resolution on listening effort revealed by pupil dilation. *Ear and Hearing*, *36*(4), e153-165. doi: 10.1097/aud.0000000000000145

Wiesenfeld, A. R., Malatesta, C. Z., & Deloach, L. L. (1981). Differential parental response to familiar and unfamiliar infant distress signals. *Infant Behavior and Development*, *4*(1), 281–295.

Wilder, C. N., & Baken, R. J. (1978). Some developmental aspects of infant cry. *Journal of Genetic Psychology*, *132*(2), 225–230.

Wolff, P. H. (1965). The causes, controls, and organization of behavior in the neonate. Psychol Issues. 5:1, 1–105.

Wolff, P. H. (1969). The natural history of crying and other vocalizations in early infancy. In B. M. Foss (Ed.), *Determinants of infant behavior* (Vol. 4, pp. 81-109). London: Methuen.

Wolff, M. S., & Ijzendoorn, M. H. (1997). Sensitivity and attachment: A meta-analysis on parental antecedents of infant attachment. *Child development, 68*(4), 571-591.

Xu, D., Yapanel, U., Gray, S., Gilkerson, J., Richards, J., and Hansen, J. (2008). Signal processing for young child speech language development. Paper presented at The 1st workshop on child, computer and interaction, Chania, Crete, Greece.

Yellott, J. I. (1971). Correction for fast guessing and the speed-accuracy tradeoff in choice reaction time. *Journal of Mathematical Psychology*, *8*(2), 159-199. doi: http://dx.doi.org/10.1016/0022-2496(71)90011-3

Yoo, H, Franklin, B., Bene, E. R., Jhang, Y., and Oller, D. K. (2014). Infant vocalization in the first month of life. Seminar presented at the American Speech-Language-Hearing Association Convention, Orlando, Florida.

Yoo, H., Buder, E. H., Lee., C-C., & Oller., D. K., (2015). Acoustic properties of infant cry and non-cry: A new look at distress and non-distress sounds. Poster presented at the American Speech-Language-Hearing Association Convention, Denver, Colorado.

Yoo, H. & Bidelman, G. M. (2016). Nonparent perception of infant cry and whine. Poster presented at the American Speech-Language-Hearing Association Convention, Philadelphia, Pennsylvania.

Yoo, H. & Oller, D. K. (2016). Temporal structure of turn-taking in parent- infant dyads: A naturalistic study with 0-3-month-olds. Poster presented at the American Speech-Language-Hearing Association Convention, Philadelphia, Pennsylvania.

Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. Biometrics, 1049–1060.

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of effortful listening: the influence of sentence intelligibility. *Ear and Hearing, 31*(4), 480-490. doi: 10.1097/AUD.0b013e3181d4f251

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2011). Cognitive load during speech perception in noise: the influence of age, hearing loss, and cognition on the pupil response. *Ear and Hearing*, *32*(4), 498-510. doi: 10.1097/AUD.0b013e31820512bb

Zeskind, P. S. (1980). Adult responses to cries of low and high risk infants. *Infant Behavior and Development, 3*, 167-177.

Zeskind, P. S., Sale, J., Maio, M. L., Huntington, L., & Weiseman, J. R. (1985). Adult perceptions of pain and hunger cries: a synchrony of arousal. *Child development*, *56*(3), 549-554.

Zeskind, P. S., & Marshall, T. R. (1988). The relation between variations in pitch and maternal perceptions of infant crying. *Child development*, *59*(1), 193-196.

Zlatin, M. A. (1975). Explorative mapping of the vocal tract and primitive syllabification in infancy: The first six months. Paper presented at the American Speech and Hearing Association Convention, Washington, D.C.

Appendices

Appendices for chapter 2

Appendix A.

Infant	Gender	Birth Order	Age in months	# of 5 min segments selected for coding	
				Protophone	Cry
1	M	2	0	-	-
			1	-	-
			3	5	5
2	F	1	0	-	-
			1	-	-
			3	5	5
3	M	2	0	5	5
			1	5	5
			3	-	-
4	F	2	0	5	5
			1	5	5
			3	-	-
5	F	2	0	-	-
			1	5	5
			3	5	5
6	M	1	0	5	5
			1	5	5
			3	5	5
7	M	2	0	5	5
			1	5	5
			3	5	5
8	M	1	0	5	5
			1	5	5
			3	5	5
9	M	3	0	5	5
			1	5	5
			3	5	5
10	F	1	0	5	5
			1	5	5
			3	5	5
11	F	2	0	5	5
			1	5	5
			3	5	5
12	F	1	0	5	5
			1	5	5
			3	5	5

Appendix B

Infant	Age in months	# of infant utterances		# of caregiver utterances (IDS and ADS)	# of IDS utterances	# of IDS responses[a]		# of IDS responses/ # of infant utterances	
		Prot	Cry			To Prot	To Cry	Prot	Cry
1	0	-	-	-	-	-	-	-	-
	1	-	-	-	-	-	-	-	-
	3	719	90	383	383	237	41	.33	.46
2	0	-	-	-	-	-	-	-	-
	1	-	-	-	-	-	-	-	-
	3	356	12	337	337	111	5	.31	.41
3	0	631	29	45	44	26	2	.04	.06
	1	833	247	51	51	33	15	.04	.06
	3	-	-	-	-	-	-	-	-
4	0	659	401	305	200	107	82	.16	.20
	1	912	91	91	91	65	5	.07	.05
	3	-	-	-	-	-	-	-	-
5	0	-	-	-	-	-	-	-	-
	1	289	167	228	228	81	25	.28	.15
	3	387	204	138	138	46	19	.12	.09
6	0	730	569	302	291	90	70	.12	.12
	1	1149	684	209	207	107	60	.09	.09
	3	1013	26	392	392	243	2	.24	.08
7	0	477	90	43	22	15	1	.03	.01
	1	456	36	29	29	18	2	.04	.06
	3	288	21	112	112	37	5	.13	.24
8	0	808	287	399	399	228	66	.28	.23
	1	647	278	333	333	160	81	.25	.30
	3	868	73	393	393	275	18	.32	.25
9	0	849	31	167	6	6	0	.01	0
	1	474	1	148	148	20	0	.04	0
	3	635	58	31	31	21	1	.03	.02
10	0	142	30	266	266	77	16	.54	.53
	1	136	36	188	188	31	22	.23	.61
	3	295	100	157	157	47	2	.16	.02
11	0	477	297	261	85	63	16	.13	.05
	1	403	76	50	50	28	8	.07	.11
	3	525	50	100	100	33	12	.06	.24
12	0	537	371	403	384	166	102	.31	.27
	1	687	308	168	168	83	41	.12	.13
	3	412	115	191	191	79	6	.19	.05

IDS = Infant-Directed Speech, ADS = Adult-Directed Speech, Prot = Protophones

# of IDS utterances = total number of utterances of caregivers that were directed toward infants, some of which could be counted as responses to infant protophones or cries, and some of which were continuations of talk by the caregiver.

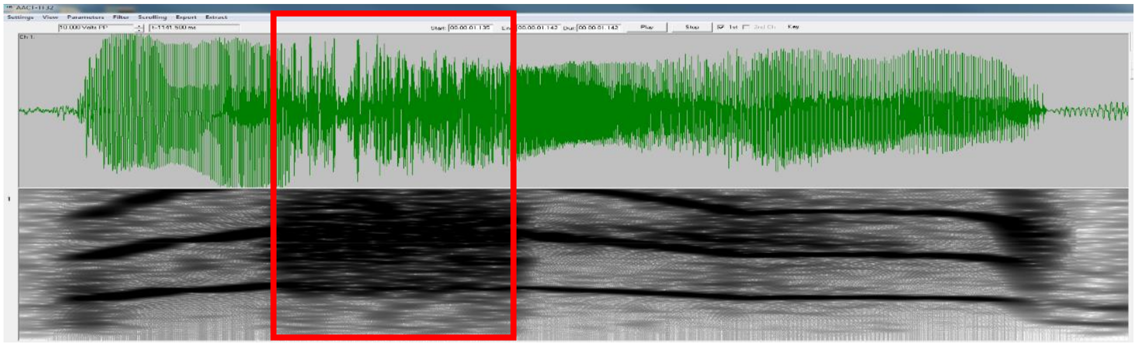
# of IDS responses = Number of caregiver IDS utterances that occurred as responses to infant protophones or cries; only the *first* IDS utterance in each caregiver sequence following the onset of an infant utterance was counted as a response. In addition any IDS utterance starting more than 5 sec after the offset of an infant utterance was not treated as a response.

[a] In 18 segments there were NO IDS responses even though there were infant protophones and/or cries and cases of IDS.

- A minus sign indicates that no recording was available for the infant at the designated age.

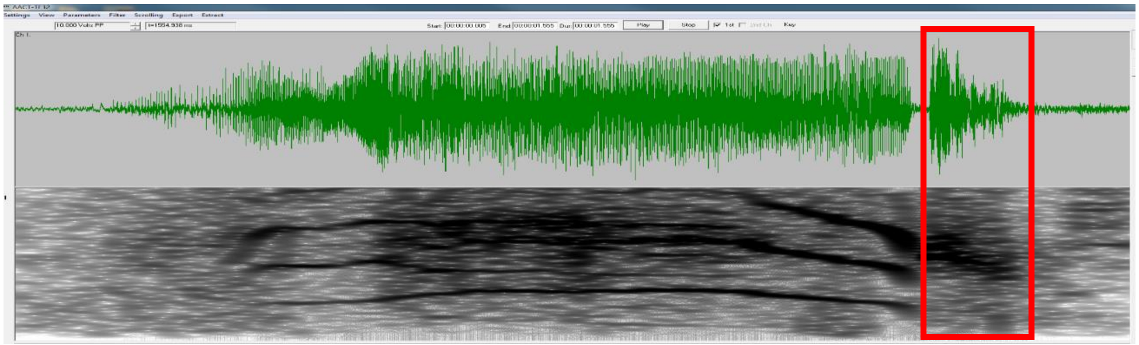
Appendix C

(A) Wail cry (high distress) with a strongly dysphonated portion of its nucleus



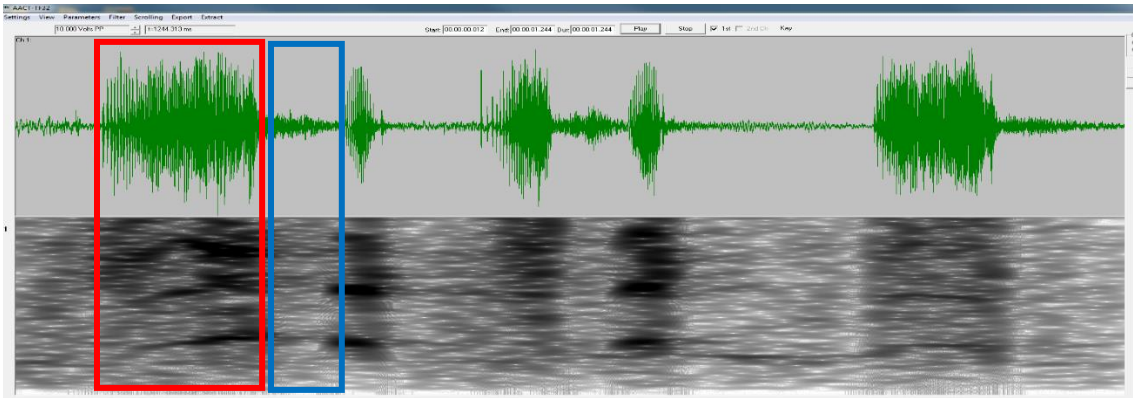
Dysphonated portion

(B) Wail cry with catch breath at the end of a nucleus including both normally phonated and dysphonated portions



Rapid ingressive catch breath

(C) Whimper (low distress): nucleus and glottal burst



Nucleus

Glottal burst

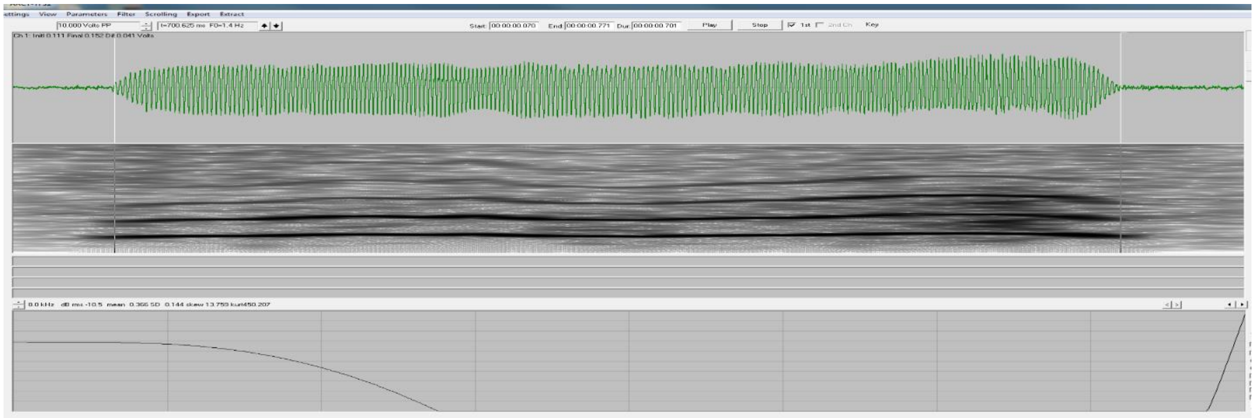
Appendix for chapter 3

Appendix. Audio files (.wav) of cry and whine

File name	Age in months
Cry 1	0m
Cry 2	0m
Cry 3	0m
Cry 4	0m
Cry 5	3m
Cry 6	3m
Cry 7	6m
Cry 8	7m
Cry 9	10m
Cry 10	10m
Whine 1	0m
Whine 2	0m
Whine 3	0m
Whine 4	1m
Whine 5	1m
Whine 6	3m
Whine 7	3m
Whine 8	3m
Whine 9	6m
Whine 10	10m

Appendix for chapter 4: Acoustic exemplars

(a) Vocant



10.000 Volts PP | I=700.625 ms F0=1.4 Hz

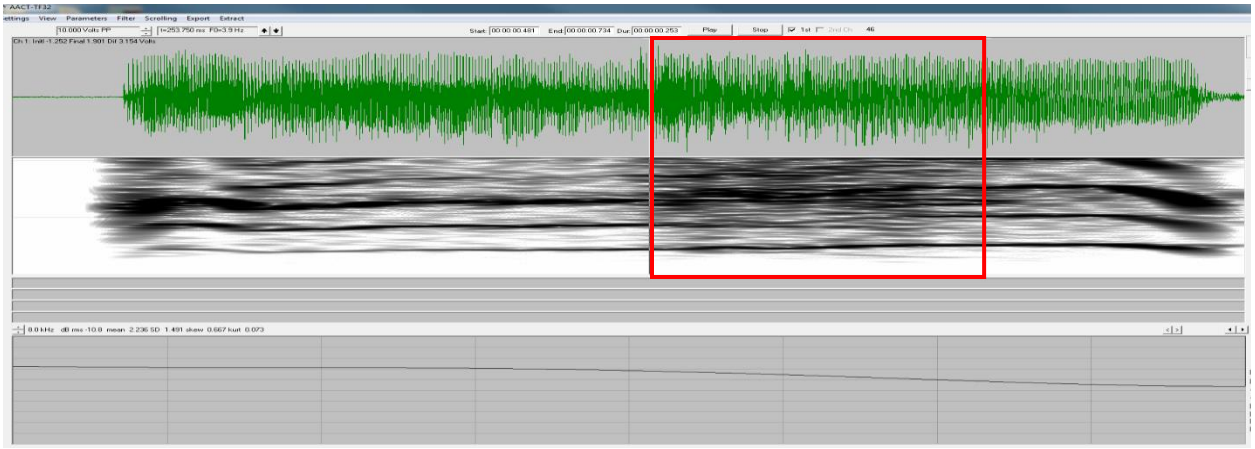
Start: 00:00:00.070 End: 00:00:00.771 Dur: 00:00:00.701

Ch 1: Init 0.111 Final 0.152 Dif 0.041 Volts

8.0 kHz dB rms -10.5 mean 0.366 SD 0.144 skew 13.759 kurt450.207

TF32 displays a waveform at the top, a type 2 spectrogram in the middle (2 kHz range), and a long-term spectral average for the period between the cursors (in this case the whole utterance) with 8 kHz range. This vocant from a 0-month-old human infant shows a single vibratory regime (Modal), with relatively evenly-spaced and easily-recognized harmonics throughout. The long-term spectral average is low in this utterance (~.36 kHz), reflecting the fact the bulk of the energy is concentrated at low frequencies.

(b) Wail



10.000 Volts PP | I=253.750 ms F0=3.9 Hz

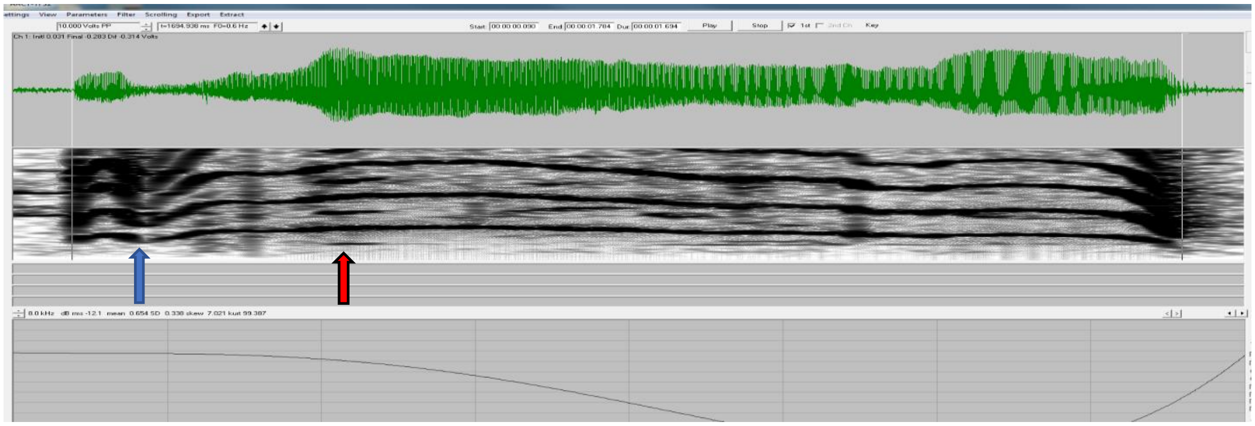
Start: 00:00:00.481 End: 00:00:00.734 Dur: 00:00:00.253

Ch 1: Init -1.252 Final 1.901 Dif 3.154 Volts

8.0 kHz dB rms -10.8 mean 2.236 SD 1.491 skew 0.667 kurt 0.073

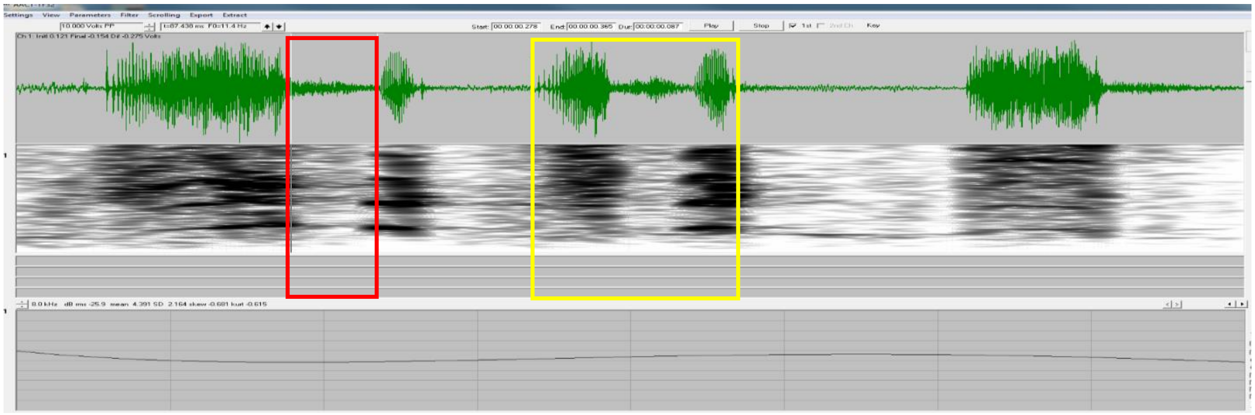
This wail cry from 0-months shows three regime segments, the middle one (Aperiodic) being indicated by the red box, with the two surrounding segments being Modal. The middle segment provides the most salient high distress information, and its long-term spectral average (~ 2.2 kHz), displayed in the bottom panel, is much higher than in typical vocants

(c) Whine



This whine shows more spectral variation than typical vocants, and the whole utterance was categorized as pertaining to the Modal regime. At the beginning of the utterance there is a brief phonatory break (blue arrow) that was counted as an instantaneous regime. Shortly thereafter a brief subharmonic segment occurs (red arrow), but it was ignored in the coding because of its brevity. The long term average spectral concentration (~.65 kHz) is intermediate between typical vocants and wails.

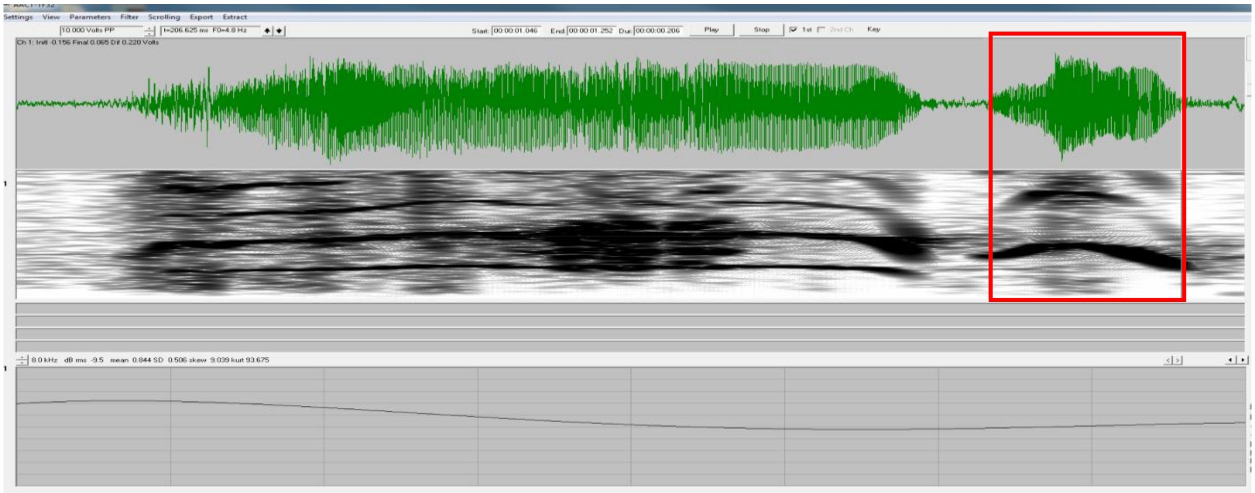
(d) Whimper



Whimpers are defined to include at least one glottal burst preceded or followed by a short nucleus that is usually somewhat nasalized. Such sequences are unambiguously heard as distressful. Often Whimpers occur in complicated sequences of events as in the example utterance above, a single breath group, including two Whimpers and adjacent whiny sounds. The red box on the left encloses a voiceless glottal burst (~80 ms) which precedes a short nasalized nucleus (~40 ms); that sequence by itself would constituted Whimper if it occurred in isolation. There is an additional sequence of glottal burst and short nucleus in the utterance to the right (yellow boxe), which also would constitute Whimper in isolation. The additional segments are typical possible adjuncts within a Whimper utterance, whiny or voiceless nuclei.

To simplify our comparisons, we did not include Whimpers among the selected stimuli for the present study. The long term average spectral concentration of the voiceless burst in the first red box was ~4 kHZ and in the nasalized nucleus thereafter ~1.3 kHz.

(e) Wail with catch breath



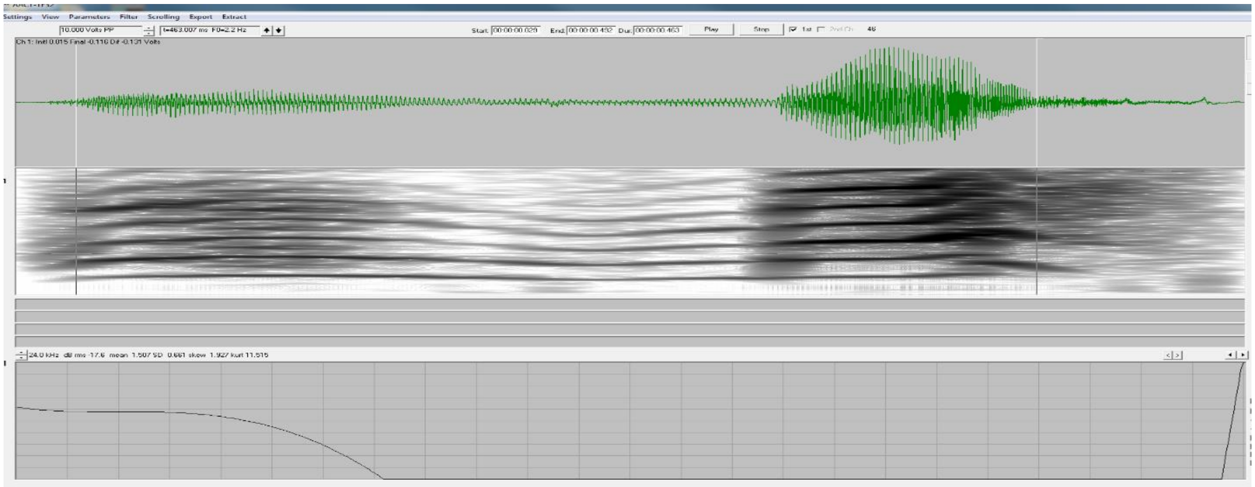
10.000 Volts PP | t=206.625 ms F0=4.8 Hz | Start: 00:00:01.046 End: 00:00:01.252 Dur: 00:00:00.206

Ch 1: Init -0.156 Final 0.065 Dif 0.220 Volts

8.0 kHz dB rms -9.5 mean 0.844 SD 0.506 skew 9.039 kurt 93.675

Another complicating factor in cry is the catch breath (in the red box), a distinctive marker for cry, defined as an abrupt inspiratory phonation that can (but often does not) occur at the end of high-distress wail. The catch breath seems spasmodic, as if the infant has used up the vital capacity with the egress and is required to inhale rapidly. To simplify our comparisons, we did not include catch breaths among the stimuli for the present study.

(f) Supraglottal Articulation : multisyllabic utterances



10.000 Volts PP | t=463.007 ms F0=2.2 Hz | Start: 00:00:00.029 End: 00:00:00.492 Dur: 00:00:00.463

Ch 1: Init 0.015 Final -0.116 Dif -0.131 Volts

24.0 kHz dB rms -17.6 mean 1.507 SD 0.661 skew 1.927 kurt 11.515

Infant vocalizations can include supraglottal articulation interrupting the phonatory pattern(s). Here a multisyllabic vocant sequence [ama] is displayed. To simplify our comparisons, we did not include utterances in any of the categories (wail, whine, or vocant) if there were syllabifying supraglottal articulations.