Developmental Science 10:2 (2007), pp 159–171 DOI: 10.1111/j.1467-7687.2007.00549.x

# TARGET ARTICLE WITH COMMENTARY AND RESPONSE

# Listening to language at birth: evidence for a bias for speech in neonates

## Athena Vouloumanos[1] and Janet F. Werker[2]

*1. Department of Psychology, McGill University, Canada*
*2. Department of Psychology, University of British Columbia, Canada*

For a commentary on this article see Rosen and Iverson (2007).

### Abstract

*The nature and origin of the human capacity for acquiring language is not yet fully understood. Here we uncover early roots of this capacity by demonstrating that humans are born with a preference for listening to speech. Human neonates adjusted their high amplitude sucking to preferentially listen to speech, compared with complex non-speech analogues that controlled for critical spectral and temporal parameters of speech. These results support the hypothesis that human infants begin language acquisition with a bias for listening to speech. The implications of these results for language and communication development are discussed.*

## Introduction

Learning how to communicate depends crucially on the ability to select meaningful signals from the environment. For human infants, this requires selectively attending to those auditory (or visual) units that carry communicative content, a problem made complex by the richness of the infant's world. Many animals filter their rich acoustic world through a general predisposition for the vocalizations of conspecifics (members of the same species), a selectivity which, in some cases, is evident even at birth (e.g. Marler, 1990). Do humans show a similar early bias for listening to speech? A bias for listening to speech would provide a potential sieve through which newborns could glean the acoustic signals important for communication.

At birth, infants already have a remarkable facility for discriminating and categorizing many aspects of human language. For example, newborns are sensitive to word boundaries (Christophe, Dupoux, Bertoncini & Mehler, 1994), distinguish between rhythmically dissimilar languages (Mehler, Jusczyk, Lambertz, Halsted, Bertoncini & Amiel-Tison, 1988; Nazzi, Bertoncini & Mehler, 1998; Ramus, Hauser, Miller, Morris & Mehler, 2000), distinguish between stress patterns of multisyllabic words (Sansavini, Bertoncini & Giovanelli, 1997), categorically discriminate lexical versus grammatical words (Shi, Werker & Morgan, 1999), and differentiate between good and poor syllable forms (Bertoncini & Mehler, 1981). Moreover, infants respond differentially to speech and non-speech. Neonates are able to discriminate languages from different rhythmical classes when the speech is played forwards, but not when it is played backwards (Ramus *et al*., 2000), suggesting that this ability is based on particular properties of speech, and not applicable to just any patterned complex sound. Although at least some of these perceptual abilities may not be unique to humans; for example, both rats (Toro, Trobalon & Sebastián-Gallés, 2005) and tamarin monkeys (Ramus *et al*., 2000) can discriminate languages from rhythmical classes in forward but not backwards speech, only humans learn language, suggesting that some aspect(s) of the acquisition process must be unique to humans (for candidates see Pinker & Jackendoff, 2005; Werker & Vouloumanos, 2000). Moreover, speech and non-speech are represented in different areas of the brain in humans: Neuroimaging studies demonstrate that listening to forward speech activates different areas of the infant brain than does backwards speech, in both neonates (Peña, Maki, Kovacic, Dehaene-Lambertz, Koizumi, Bouquet & Mehler, 2003) and 3-month-old infants (Dehaene-Lambertz, Dehaene & Hertz-Pannier, 2002), though the areas of differential activation differ in these two studies, suggesting that the neonatal brain already discriminates between speech and non-speech sounds.

Despite evidence for differential processing for speech and non-speech, a behavioural preference for the speech

Address for correspondence: Athena Vouloumanos, Department of Psychology, McGill University, 1205 Dr. Penfield Avenue, Montreal, QC, H3A 1B1, Canada; e-mail: athena.vouloumanos@mcgill.ca

© 2007 The Authors. Journal compilation © 2007 Blackwell Publishing Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

signal itself has yet to be demonstrated in the neonatal period. As Doupe and Kuhl (1999) note: 'In humans, there is no convincing experimental evidence that infants have an innate description of speech'. In an often-cited methodological study, neonates favoured a stimulus that included a speech component (folk music) over a non-speech condition deliberately made unappealing ( broadband white noise – the unmodulated sound of radio static) (Butterfield & Siperstein, 1970). On the basis of this study, it was widely reported that neonates prefer speech. However, the differences in spectral and temporal parameters between speech and white noise (modulated vs. invariant signals) (Eisenberg, 1976), and the choice of a 'speech' condition that includes both vocal and musical components, leave this question unanswered.

To investigate whether neonates demonstrate a bias for speech, we presented infants with isolated syllables of human speech contrasted with non-speech stimuli crafted to control for infants' sensitivity to critical spectral and temporal parameters of speech. These stimuli had been used in previous studies investigating listening preferences in infancy, in which we demonstrated that infants as young as 2 months old prefer listening to speech (Vouloumanos & Werker, 2004). The speech signal is composed of concentrations of energy at multiple frequencies that change over time (Figure 1C). Non-speech counterparts were modelled on sine-wave analogues of speech (Remez, Rubin, Pisoni & Carrell, 1981), and consisted of time-varying sinusoidal waves that track the resonant centre frequencies (formants) of natural speech to reproduce the changes in these frequency peaks across time (Figure 1B). In reproducing the main spectral and temporal changes in natural speech, these complex non-speech analogues contrast sharply with single-frequency tones (Figure 1A) and white noise (Figure 1D), two types of stimuli commonly used as non-speech conditions (e.g. Butterfield & Siperstein, 1970; Eisenberg, 1976). In the present study, we investigate whether human neonates show a bias for listening to speech by comparing neonates' contingent sucking responses in eliciting speech and complex non-speech sounds.

## Method

### *Participants*

Twenty-two neonates (1–4 days old, $M$ = 45.1 hr) were recruited from a local hospital and tested in a high amplitude sucking (HAS) procedure (Cooper & Aslin, 1990; Eimas, Siqueland, Jusczyk & Vigorito, 1971). An additional 24 infants were not included for the following reasons: falling asleep (1), failing to meet the sucking criterion (8; see below), equipment failure (1), experimenter interference (5), crying or fussing (2), rejection of pacifier (3), sucking weakly (2) and hospital fire alarm ringing during the experiment (2).

### *Stimuli*

#### Speech stimuli

Speech stimuli consisted of four tokens of a monosyllabic nonsense word ('lif') spoken by a female native English speaker. Tokens varied in intonational contour (average minimum and maximum pitches = 203 Hz, 325 Hz), and in duration (average = 665 ms). The limited variation in phonetic information minimized differences between the speech and non-speech stimulus sets.

#### Complex non-speech analogues

Non-speech stimuli consisted of time varying sinusoidal waves tracking the main regions of significant energy, namely the fundamental frequency and the first three formants of speech (stimulus creation has been reported in Vouloumanos, Kiehl, Werker & Liddle, 2001; Vouloumanos & Werker, 2004). Non-speech analogues retained the duration, pitch contour, amplitude envelope, relative formant amplitude, and relative intensity of their speech counterparts. The two stimulus types differed in voice quality (non-speech analogues have none), in naturalness or biological quality (non-speech analogues are artifacts), and in the characteristics of the source (speech has one source, the vocal tract, while non-speech analogues have four, one per sinusoidal tone). However, and crucially for the question asked in the current study, the non-speech analogues track changes across time for the peak frequencies of their speech counterparts, and in so doing, follow very closely the spectral and timing changes of natural speech. The signals were further equated for infant ears by retaining the fundamental frequency that carries information about pitch contour of the speech counterparts, because pitch contour contributes importantly to infants' preference for infant-directed speech (Fernald & Kuhl, 1987), and discrimination of their native language (Nazzi, Floccia & Bertoncini, 1998).[1]

### *Design and procedure*

Approximately 2 hours after feeding, neonates were presented with a sterilized pacifier coupled to a pressure

[1] The addition of the fundamental frequency is detrimental to the perception of traditional sine-wave analogues (Remez & Rubin, 1993). However, omitting this pitch information from the non-speech analogues would render the comparison trivial for infants, since this dimension alone would predispose the infants towards speech.

© 2007 The Authors. Journal compilation © 2007 Blackwell Publishing Ltd.

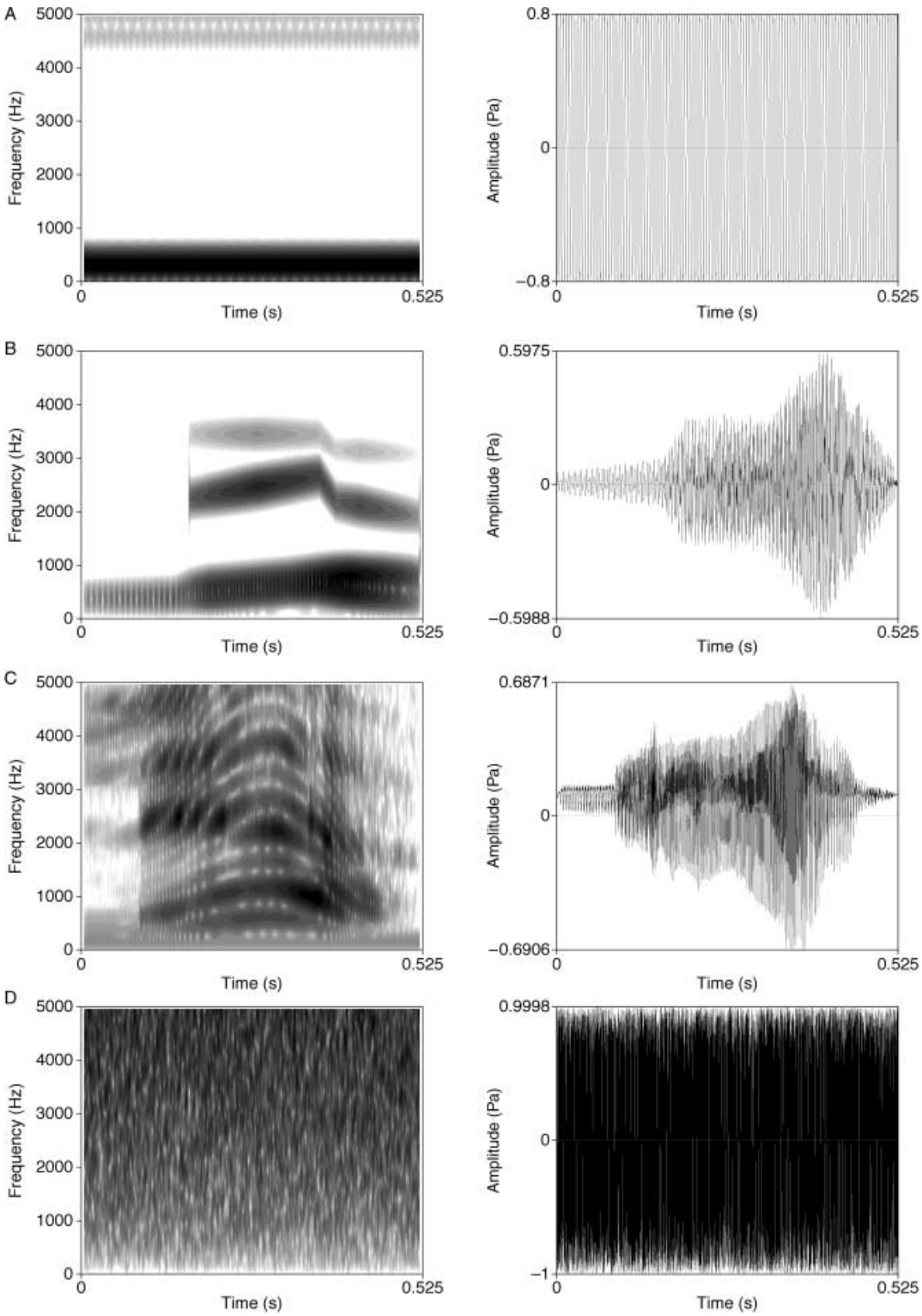


**Figure 1** *Comparison of acoustic properties of speech and non-speech stimuli. Wide-band spectrograms (left) depict the change in frequency across time, and waveform diagrams (right) illustrate the amplitude changes across time in pressure units; (A) single frequency tones, (B) sample token of complex non-speech used in this study, (C) sample speech token used in this study, and (D) white noise. The changes in frequency across time of the speech signal can also be observed in the complex non-speech sounds whereas this time-varying property is absent from the other sounds.*

© 2007 The Authors. Journal compilation © 2007 Blackwell Publishing Ltd.

transducer. Following one silent baseline minute, during which each individual infant's sucking amplitude range was established, neonates were presented with a sound stimulus every time they delivered suction in the upper 80% of their sucking amplitude range. The presentation of speech and non-speech stimuli alternated every minute, and the 8 minutes post-baseline were submitted to analysis.

To ensure that infants had enough exposure to hear the different sounds and learn the contingency, it was necessary to implement criteria for an acceptable minimum number of HA sucks. First, infants were excluded from the analysis if they had any experimental minutes in which they delivered 0 HA sucks. This ensured that newborns whose data were analyzed would hear at least one sound in each experimental minute. Second, in order to ensure that enough sounds were heard at the beginning of the study for the infants to demonstrate a potential preference, we excluded infants who delivered fewer than 10 HA sucks in each of the first four experimental minutes (Floccia, Nazzi & Bertoncini, 2000).

## Results

Based on previous studies with neonates using this HAS procedure in which differences emerged in the latter part of the experiment (Floccia *et al*., 2000; Sansavini *et al*., 1997), the experimental phase was examined as two blocks. A 2 (sound type: speech vs. non-speech) × 2 (stimulus block: first 4-min block vs. second 4-min block) × 2 (order: speech first vs. non-speech first) mixed analysis of variance (ANOVA) indicated no main effect of sound type and no main effect of order. A main effect of block ($F(1, 20) = 4.70$, $p = .042$) revealed a higher number of sucks in the first block ($M = 74.4$, $SE = 4.52$) than the second block ($M = 68.2$, $SE = 3.51$) overall. There was no main effect of sound type; however, there was a significant interaction between sound type and experimental block ($F(1, 20) = 6.68$, $p = .018$); planned comparisons on sucking rates in each of the two blocks showed that neonates sucked significantly more to listen to speech ($M = 73.0$, $SE = 3.49$) than to complex non-speech analogues ($M = 63.5$, $SE = 4.31$) in the second experimental block ($t(21) = 2.84$, $p = .010$) (Figure 2). Means in the first block were not significantly different from each other ($t(21) = 1.18$, *ns*). The emergence of the effect of interest within the second block is consistent with other HAS studies.

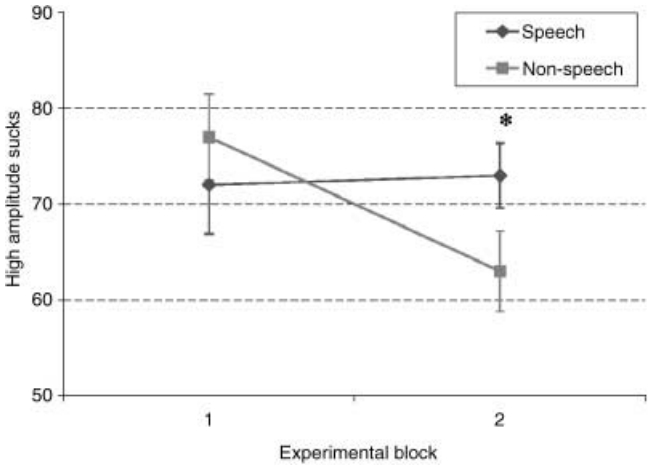


**Figure 2** *Neonates' HAS for speech and complex non-speech sounds. Neonates sucked significantly more to listen to speech than to complex non-speech analogues in the second experimental block.*

## Discussion

This study provides the first demonstration that human neonates are biased to listen to speech. This bias is consistent with initial proclivities, widespread in the animal kingdom, that direct animals towards particular types of auditory and visual information (e.g. Gould & Marler, 1987; Johnson, Bolhuis & Horn, 1992; Lorenz, 1965; Marler, 1990; Ryan, Phelps & Rand, 2001). Previous research had shown that infants process speech differently than non-speech; for example, they are better able to discriminate languages when speech is played forwards rather than backwards (Mehler *et al*., 1988; Ramus *et al*., 2000) even when the speech is low pass filtered (Mehler *et al*., 1988; Nazzi *et al.*, 1998), or when every syllable is replaced with the consonant-vowel sequence /sa/ (Ramus, 2002), and they recruit differential neural resources for speech and non-speech processing (Dehaene-Lambertz *et al*., 2002; Peña *et al*., 2003). The present research focuses on a different aspect of speech and non-speech processing: whether infants show a behavioural bias for listening to speech. We compare neonates' listening preferences for speech to complex non-speech stimuli, and find a bias for listening to speech. This bias may confer an adaptive advantage by tuning humans to the communication signal of their conspecifics, and hence facilitate more in-depth processing and rapid learning of the specific attributes of the native language.

More than simply conferring an advantage, there is some evidence that a bias for speech in infancy may be essential for developing normal language abilities, as recent studies demonstrate that children with autism spectrum disorder (ASD) fail to show a preference for speech when it is compared with either an unresolvable stimulus composed of superimposed voices (Klin, 1991) or a complex non-speech stimulus composed of three sine-waves, similar to the non-speech counterpart of the current study (Klin, 1991; Kuhl, Coffey-Corina, Padden & Dawson, 2005). Instead, children

© 2007 The Authors. Journal compilation © 2007 Blackwell Publishing Ltd.

with ASD seem to show a preference for non-speech, the degree of which correlates significantly with ASD symptomology, especially with respect to expressive language abilities (Kuhl *et al*., 2005). Although this evidence is necessarily correlational, it is suggestive of the important role a bias for speech may play in normal language development.

The discovery of a neonatal bias for speech suggests several important questions. First, whence originates this neonatal bias for speech: is the bias rooted in prenatal experience with speech, or is it experience-independent? Previous studies have revealed neonatal preferences that are clearly experience-based, such as an attraction to the mother's voice (DeCasper & Fifer, 1980) and to the native language (Mehler, Bertoncini & Barriere, 1978; Moon, Cooper & Fifer, 1993). Though these specific preferences are unequivocally experience-based, a more general bias for speech, in its potential human universality, may not be. Indeed, evidence for innate conspecific preferences in other species suggests that a bias for speech might be a tantalizing candidate for an experience-independent human bias.

Second, what aspect of speech is the bias based on? In the case of duckling preference for conspecific calls, specific spectral and temporal aspects of the duck call are crucial. When conspecific calls are contrasted with a heterospecific foil similar in repetition rate and fundamental frequency, ducklings show no preference (Gottlieb, 1997). The neonatal speech preference could be based on a number of dimensions, ranging from low-level acoustic properties to higher-level abstract properties. For example, the sheer complexity of an acoustic stimulus can drive newborn and foetal physiological arousal for sounds; stimuli rich in spectral characteristics or patterned in temporal properties elicit greater changes in EMG (Hutt, Hutt, Lenard, van Bernuth & Muntjewerff, 1968), EEG (Lenard, von Bernuth & Hutt, 1969), and heart rate (Clarkson & Berg, 1983; Groome, Mooney, Holland, Smith, Atterbury & Dykman, 2000). Like other biologically special stimuli, such as faces or biological motion, attempts to create non-biological analogues must by necessity eliminate at least some of the characteristics of the original stimulus. Questions always remain as to whether the particular characteristics that were eliminated were the most important to maintain, or whether in mimicking the biological signal, one particular cue was highlighted over others. Though our complex non-speech stimuli preserve many of the spectral and temporal aspects of speech, they were composed of narrow frequency bands and thus necessarily lacked the broadband frequency information of speech. The relative acoustic complexity of speech may thus contribute to the neonatal preference observed in the current study. However, a preference for speech might stem from higher-level aspects of the speech stimulus, such as its human source, its biological origin or its intention to communicate. Studies are under way to investigate these possibilities.

Despite these remaining questions, a neonatal bias for speech is one important tool available at birth for learning language. Initial biases may be elaborated by experience to refine the perceptual preferences of developing organisms (e.g. Gottlieb, 1997; Werker & Tees, 1992). Indeed, the attraction to speech persists into the first few months of life (Vouloumanos & Werker, 2004) and may include communicative gestures in other modalities such as signed language (Krentz & Corina, in press). A speech bias, combined with established experience-based preferences for the mother's voice and native language, could provide human neonates with powerful tools for selecting and learning about communication signals from their rich environment.

## Acknowledgements

This research was funded by fellowships from the Natural Science and Engineering Council (NSERC) of Canada, the Killam Foundation, and the Canadian Institutes of Health Research to Athena Vouloumanos, and grants from NSERC and the Human Frontiers Science Program, as well as a Canada Research Chair to Janet F. Werker. We thank M. Bhatnagar for her seminal role in implementing the project and her help with testing infants. Thanks to S. Bird and G. Carden for synthesis of non-speech stimuli, to T. Burns, E. Moon, and A. Valji for their help in testing infants, to the staff of the BC Women's and Children's Hospital for their cooperation, and to C. Narayan for assistance with the figure. We thank G. Marcus for invaluable discussions and comments on earlier versions. We especially thank all the parents and newborns who participated in our research.

## References

Bertoncini, J., & Mehler, J. (1981). Syllables as units in infant speech perception. *Infant Behavior and Development*, **4** (3), 247–260.

Butterfield, E.C., & Siperstein, G.N. (1970). Influence of contingent auditory stimulation upon non-nutritional suckle. In J.F. Bosma (Ed.), *Third Symposium on Oral Sensation and Perception: The Mouth of the Infant* (pp. 313–334). Springfield, IL: Charles C. Thomas.

Christophe, A., Dupoux, E., Bertoncini, J., & Mehler, J. (1994). Do infants perceive word boundaries? An empirical study of the bootstrapping of lexical acquisition. *Journal of the Acoustical Society of America*, **95** (3), 1570–1580.

Clarkson, M.G., & Berg, W.K. (1983). Cardiac orienting and vowel discrimination in newborns: crucial stimulus parameters. *Child Development*, **54** (1), 162–171.

Cooper, R.P., & Aslin, R.N. (1990). Preference for infant-directed speech in the first month after birth. *Child Development*, **61**, 1584–1595.

© 2007 The Authors. Journal compilation © 2007 Blackwell Publishing Ltd.

DeCasper, A.J., & Fifer, W.P. (1980). Of human bonding: newborns prefer their mothers' voices. *Science*, **208** (4448), 1174–1176.
Dehaene-Lambertz, G., Dehaene, S., & Hertz-Pannier, L. (2002). Functional neuroimaging of speech perception in infants. *Science*, **298** (5600), 2013–2015.
Doupe, A.J., & Kuhl, P.K. (1999). Birdsong and human speech: common themes and mechanisms. *Annual Review of Neuroscience*, **22**, 567–631.
Eimas, P.D., Siqueland, E.R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, **171** (968), 303–306.
Eisenberg, R.B. (1976). *Auditory competence in early life*. Baltimore, MD: University Park Press.
Fernald, A., & Kuhl, P.K. (1987). Acoustic determinants of infant preference for motherese speech. *Infant Behavior and Development*, **10** (3), 279–293.
Floccia, C., Nazzi, T., & Bertoncini, J. (2000). Unfamiliar voice discrimination for short stimuli in newborns. *Developmental Science*, **3** (3), 333–343.
Gottlieb, G. (1997). *Synthesizing nature–nurture: Prenatal roots of instinctive behavior*. Mahwah, NJ: Lawrence Erlbaum Associates.
Gould, J.L., & Marler, P. (1987). Learning by instinct. *Scientific American*, **256** (1), 74–85.
Groome, L.J., Mooney, D.M., Holland, S.B., Smith, Y.D., Atterbury, J.L., & Dykman, R.A. (2000). Temporal pattern and spectral complexity as stimulus parameters for eliciting a cardiac orienting reflex in human fetuses. *Perception and Psychophysics*, **62** (2), 313–320.
Hutt, S.J., Hutt, C., Lenard, H.G., van Bernuth, H., & Muntjewerff, W.J. (1968). Auditory responsivity in the human neonate. *Nature*, **218**, 888–890.
Johnson, M.H., Bolhuis, J.J., & Horn, G. (1992). Predispositions and learning: behavioural dissociations in the chick. *Animal Behaviour*, **44** (5), 943–948.
Klin, A. (1991). Young autistic children's listening preferences in regard to speech: a possible characterization of the symptom of social withdrawal. *Journal of Autism and Developmental Disorders*, **21** (1), 29–42.
Krentz, U.C., & Corina, D.C. (in press). Preference for language in early infancy: the human language bias is not speech specific. *Developmental Science*.
Kuhl, P.K., Coffey-Corina, S., Padden, D., & Dawson, G. (2005). Links between social and linguistic processing of speech in preschool children with autism: behavioral and electrophysiological measures. *Developmental Science*, **8** (1), F1–F12.
Lenard, H.G., von Bernuth, H., & Hutt, S.J. (1969). Acoustic evoked responses in newborn infants: the influence of pitch and complexity of the stimulus. *Electroencephalography and Clinical Neurophysiology*, **27** (2), 121–127.
Lorenz, K. (1965). *Evolution and modification of behavior*. Chicago, IL: University of Chicago Press.
Marler, P. (1990). Innate learning preferences: signals for communication. *Developmental Psychobiology*, **23** (7), 557–568.
Mehler, J., Bertoncini, J., & Barriere, M. (1978). Infant recognition of mother's voice. *Perception*, **7** (5), 491–497.
Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, **29** (2), 143–178.
Moon, C., Cooper, R.P., & Fifer, W.P. (1993). Two-day-olds prefer their native language. *Infant Behavior and Development*, **16** (4), 495–500.
Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance*, **24** (3), 756–766.
Nazzi, T., Floccia, C., & Bertoncini, J. (1998). Discrimination of pitch contours by neonates. *Infant Behavior and Development*, **21** (4), 779–784.
Peña, M., Maki, A., Kovacic, D., Dehaene-Lambertz, G., Koizumi, H., Bouquet, F., & Mehler, J. (2003). Sounds and silence: an optical topography study of language recognition at birth. *Proceedings of the National Academy of Sciences of the United States of America*, **100** (20), 11702–11705.
Pinker, S., & Jackendoff, R. (2005). The faculty of language: what's special about it? *Cognition*, **95** (2), 201–236.
Ramus, F. (2002). Language discrimination by newborns: teasing apart phonotactic, rhythmic, and intonational cues. *Annual Review of Language Acquisition*, **2**, 85.
Ramus, F., Hauser, M.D., Miller, C., Morris, D., & Mehler, J. (2000). Language discrimination by human newborns and by cotton-top tamarin monkeys. *Science*, **288** (5464), 349–351.
Remez, R.E., & Rubin, P.E. (1993). On the intonation of sinusoidal sentences: contour and pitch height. *Journal of the Acoustical Society of America*, **94** (4), 1983–1988.
Remez, R.E., Rubin, P.E., Pisoni, D.B., & Carrell, T.D. (1981). Speech perception without traditional speech cues. *Science*, **212** (4497), 947–949.
Rosen, S., & Iverson, P. (2007). Constructing adequate non-speech analogues: what *is* special about speech anyway? *Developmental Science*, **10** (2), 165–169.
Ryan, M.J., Phelps, S.M., & Rand, A.S. (2001). How evolutionary history shapes recognition mechanisms. *Trends in Cognitive Sciences*, **5** (4), 143–148.
Sansavini, A., Bertoncini, J., & Giovanelli, G. (1997). Newborns discriminate the rhythm of multisyllabic stressed words. *Developmental Psychology*, **33** (1), 3–11.
Shi, R., Werker, J.F., & Morgan, J.L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition*, **72** (2), B11–B21.
Toro, J.M., Trobalon, J.B., & Sebastián-Gallés, N. (2005). Effects of backward speech and speaker variability in language discrimination by rats. *Journal of Experimental Psychology: Animal Behavior Processes*, **31** (1), 95–100.
Vouloumanos, A., Kiehl, K.A., Werker, J.F., & Liddle, P.F. (2001). Detection of sounds in the auditory stream: event-related fMRI evidence for differential activation to speech and nonspeech. *Journal of Cognitive Neuroscience*, **13** (7), 994–1005.
Vouloumanos, A., & Werker, J.F. (2004). Tuned to the signal: the privileged status of speech for young infants. *Developmental Science*, **7** (3), 270–276.
Werker, J.F., & Tees, R.C. (1992). The organization and reorganization of human speech perception. *Annual Review of Neuroscience*, **15**, 377–402.
Werker, J.F., & Vouloumanos, A. (2000). Language: who's got rhythm? *Science*, **288** (5464), 280–281.

© 2007 The Authors. Journal compilation © 2007 Blackwell Publishing Ltd.

# COMMENTARY

# Constructing adequate non-speech analogues: what *is* special about speech anyway?

## Stuart Rosen and Paul Iverson

*Department of Phonetics and Linguistics, UCL, London, UK*

This is a commentary on Vouloumanos and Werker (2007).

### Abstract

*Vouloumanos and Werker (2007) claim that human neonates have a (possibly innate) bias to listen to speech based on a preference for natural speech utterances over sine-wave analogues. We argue that this bias more likely arises from the strikingly different saliency of voice melody in the two kinds of sounds, a bias that has already been shown to be learned pre-natally. Possible avenues of research to address this crucial issue are proposed, based on a consideration of the distinctive acoustic properties of speech.*

There has been long-standing interest in the notion that speech sounds have a privileged position in human audition, and in the extent to which auditory processing is common or distinct for speech and non-speech sounds. Much work comparing the processing of speech and non-speech has involved the construction of non-speech analogues (e.g. see Mody, Studdert-Kennedy & Brady, 1997). What has become strikingly clear, particularly in investigations of functional neuro-imaging (e.g. Scott, Blank, Rosen & Wise, 2000) is that the conclusions that can be drawn from any particular such study depend crucially on the properties of the comparison non-speech analogues. Strictly speaking then, only one claim can be supported by the results of Vouloumanos and Werker (2007) – that human neonates prefer to listen to full-blown speech sounds in comparison to sine-wave analogues. Their much more profound claim 'that human neonates are biased to listen to speech' can only be upheld to the degree to which their non-speech analogues are seen to be adequate. In fact, we believe them to be poor controls, because the original speech stimuli convey a strong and salient percept of voice melody that is very nearly absent in the non-speech analogues.

Three of the speech sounds used by V&W, and their non-speech analogues, are available in the online supplementary material (Figure S1). Even casual listening to the speech reveals the strikingly salient voice pitch of the talker, whose exaggerated melodic contours seem more appropriately aimed at a child than an adult. The non-speech analogues, on the other hand, sound much more similar to one another, with little or no sense of a melodic contour. V&W did include the voice pitch contour of the talker as a separate sinusoidal component (a departure from the standard means of constructing sine-wave speech), but only careful and analytic listening will reveal its presence. This is hardly surprising given the differences between the two sets of stimuli in the way in which voice pitch is signalled. As the spectrograms in Figure 1 show, the speech signal contains many harmonics through the entire frequency range of the speech, at multiples of the fundamental frequency (the crucial determinant of the percept of voice pitch). However, the representation of voice pitch in the non-speech analogue is only through a single component. Moreover, this component is in a low-frequency region that is relatively unimportant for speech intelligibility, and where human hearing is less sensitive compared to higher frequencies. Remez and Rubin (1984) have already noted that sine-wave sentences are perceived to have a weird intonation determined by the tone representing the first formant, even with the presence of an extra component at the fundamental frequency.

Given this crucial difference between the non-speech analogues and the speech, we might just as well claim,

Address for correspondence: Stuart Rosen, Department of Phonetics and Linguistics, UCL, 4 Stephenson Way, London NW1 2HE, UK; e-mail: stuart@phon.ucl.ac.uk

© 2007 The Authors. Journal compilation © 2007 Blackwell Publishing Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

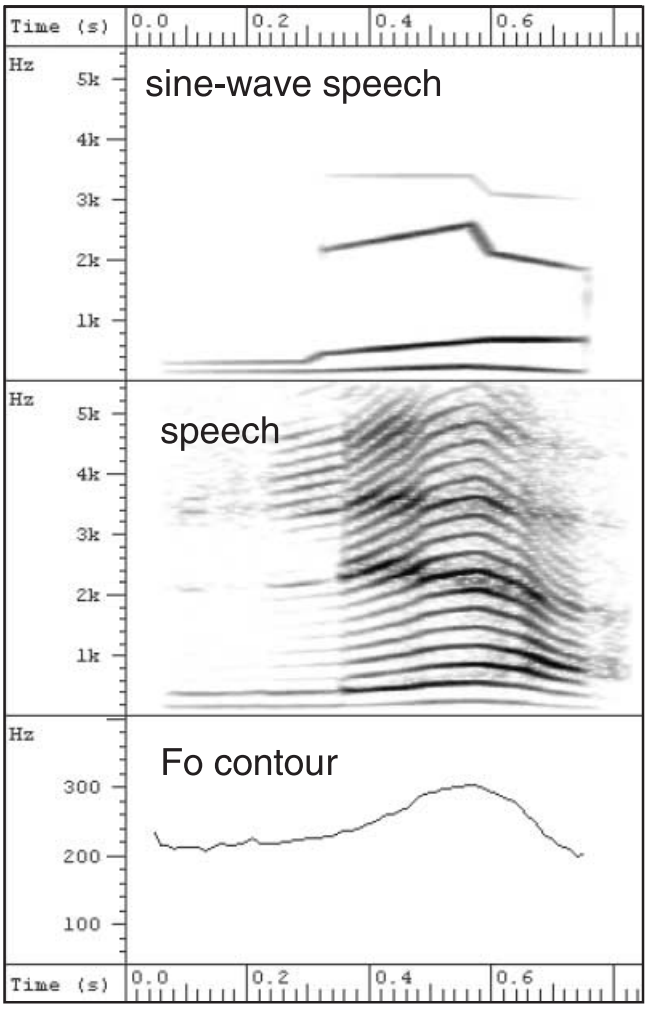


**Figure 1** *Examples of the sounds used by V&W in their study, along with the fundamental frequency (Fo) contour extracted directly from the speech sound. The top two panels are narrow-band spectrograms. Note the multitude of harmonics representing the fundamental frequency in the speech sound. To listen to these sounds, go to Figure S1 in the online supplementary material.*

then, that 'human neonates are biased to listen to sounds with a strong voice melody'. Once this possibility is acknowledged, then the suggestion that the bias may be innate is easily refuted. Previous work has suggested that the intonation and rhythm of a mother's voice are learned in the womb, such that newborns prefer their mother's voice over other mothers' voices (Decasper & Fifer, 1980) and prefer speech spoken in their mother's language to speech spoken in a language from a different rhythmic class (Moon, Cooper & Fifer, 1993; see also Mehler, Jusczyk, Lambertz, Halsted, Bertoncini & Amiel-Tison, 1988; Nazzi, Bertoncini & Mehler, 1998). A claim that there is an innate bias to listen to speech must thus provide a better control for learning of pitch and rhythm.

So what kinds of comparisons might prove useful in establishing whether or not infants have a bias for speech? One possible approach, based on the *source-filter* theory of speech production, is to identify what is evolutionarily innovative in the acoustics of human speech different to animal vocalizations. It is perhaps not too much to claim that the main communicative aspects of animal vocalizations concern variations in the *source* of sound production, that is, the patterning of periodic and aperiodic sounds, and the fundamental frequency when the sound is periodic. Source variations are also primarily (but not wholly) responsible for the amplitude modulations in speech. On the other hand, there is little or no evidence for the communicative use in animals of the spectral dynamics that arise from the variations in the filtering exacted by the moving vocal tract. (This is not to say that animals cannot be sensitive to filter-based aspects of spectral shape which may be indicative of size or identity, but these cues are static – see Fitch, 2000, for a review). Sensitivity to spectral dynamics can be readily argued to be the *sine qua non* of human speech perception, both necessary (Rosen, 1992) and sufficient (Shannon, Zeng, Kamath, Wygonski & Ekelid, 1995), although certainly not complete.

It might therefore be interesting to assess the preference of infants for various sounds which manipulate the presence or absence of various acoustic features. Algorithms based on sine-wave speech prove to be particularly manipulable in this regard (e.g. Scott, Rosen & Wise, 2005). Replacing the formant-tracking sine waves with bands of noise leads to sounds that cohere more readily, and hence, are more intelligible than sine-wave speech itself (and presumably, are better analogues to speech). We could then ask, for example, whether infants, in the absence of a periodic source, prefer sounds with dynamic formant variation to steady-state formants, even in the presence of natural amplitude variations. Or whether they prefer such sounds based on real sentences (hence intelligible to an adult listener), or ones which combine the formant tracks from one sentence with the amplitude variations of another, leading to speech-like, but unintelligible, sounds (Figure 2, with audio examples in the online supplementary Figure S2). Or whether sounds with amplitude variation are preferred to those with spectral modulations. One could also pit the 'attractiveness' of melodic pitch variations against amplitude and spectral envelope modulations, by exciting the formant-like spectral prominences in these sounds with a natural source of periodic and aperiodic sounds (see Figures S3 and S4 in the online supplementary material for audio examples).

© 2007 The Authors. Journal compilation © 2007 Blackwell Publishing Ltd.

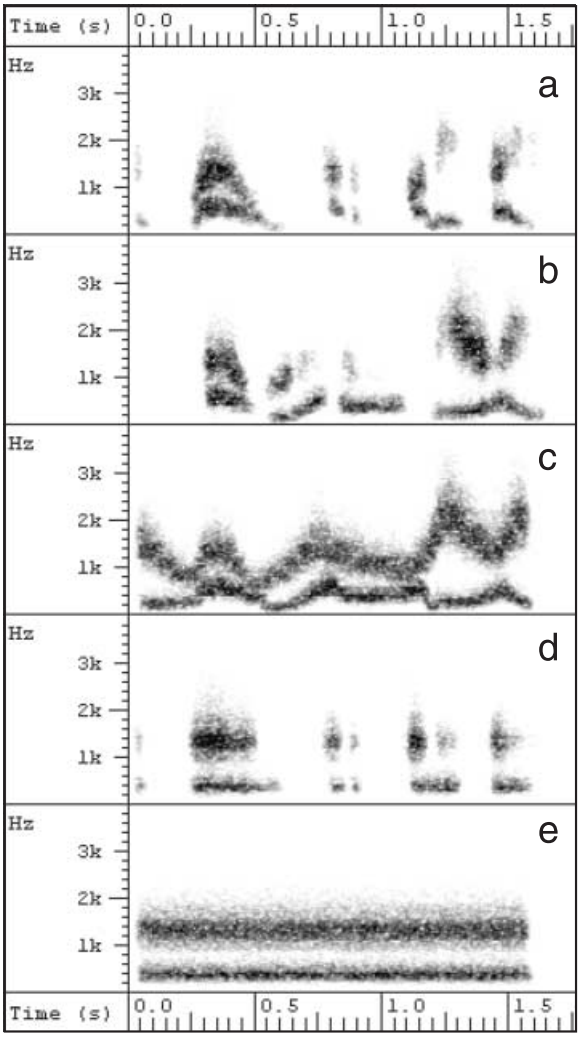


**Figure 2** *Spectrograms of a hierarchy of stimuli, varying in complexity and intelligibility, all constructed using the first two formants of sine-wave speech. Once the manipulations of the sine-wave formants are done, the stimuli are passed through a 16-channel noise-excited vocoder (Shannon* et al.*, 1995) so as to replace the sine waves with a continuous spectrum whose envelope is more reminiscent of natural speech. The common excitation also causes the two 'formants' to cohere perceptually, leading to a unitary percept. The top sentence (a) is a straightforward version of the original sentence 'The clown had a funny face', with natural formant and amplitude variations. (b) contains interpolated formant tracks from (a), as seen in (c), with the amplitude variations from another sentence imposed. This leads to a sound that is unintelligible, but has the same spectro-temporal complexity as natural speech. (d) represents steady-state formants with the natural amplitude variations, whereas (e), the simplest case, consists of two steady-state formants at a constant amplitude. To listen to these sounds, go to Figure S2 in the online supplementary material.*

In our view, any reasonable approach to unravelling the nature of infants' auditory preferences must take account, at least, of the role of modulations in these three essential features of speech: fundamental frequency, amplitude and spectral shape. It seems likely that, insofar as they are evolutionarily earlier, features associated with fundamental frequency/voice pitch and amplitude modulations are likely to be attended to first, even though apprehension of spectral modulations is essential for language acquisition. Sensitivity to voice pitch and amplitude are also essential in providing auditory feedback to the developing infant, so as to develop efficient and strong vocal fold vibration, the framework upon which speech production is built (Fourcin, 1978).

## Acknowledgements

Many thanks to Athena Vouloumanos and Janet Werker for access to their stimuli, and AV for very useful discussions.

## References

Decasper, A.J., & Fifer, W.P. (1980). Of human bonding: newborns prefer their mothers' voices. *Science*, **208**, 1174–1176.

Fitch, W.T. (2000). The evolution of speech: a comparative review. *Trends in Cognitive Sciences*, **4**, 258–267.

Fourcin, A.J. (1978). Acoustic patterns and speech acquisition. In N. Waterson & C. Snow (Eds.), *The development of communication* (pp. 47–72). Chichester: John Wiley.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language-acquisition in young infants. *Cognition*, **29**, 143–178.

Mody, M., Studdert-Kennedy, M., & Brady, S. (1997). Speech perception deficits in poor readers: auditory processing or phonological coding? *Journal of Experimental Child Psychology*, **64**, 199–231.

Moon, C., Cooper, R.P., & Fifer, W.P. (1993). Two-day-olds prefer their native language. *Infant Behavior and Development*, **16**, 495–500.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance*, **24**, 756–766.

Remez, R.E., & Rubin, P.E. (1984). On the perception of intonation from sinusoidal sentences. *Perception and Psychophysics*, **35**, 429–440.

Rosen, S. (1992). Temporal information in speech: acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society London B*, **336**, 367–373.

Scott, S.K., Blank, C.C., Rosen, S., & Wise, R.J.S. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain*, **123**, 2400–2406.

Scott, S.K., Rosen, S., & Wise, R.J.S. (2005). Hemispheric

© 2007 The Authors. Journal compilation © 2007 Blackwell Publishing Ltd.

lateralisation in speech perception does not arise from simple acoustic properties of speech stimuli. *Assoc. Res. Otolaryngol. Abs*., **763**, 268.
Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, **270**, 303–304.
Vouloumanos, A., & Werker, J.F. (2007). Listening to language at birth: evidence for a bias for speech in neonates. *Developmental Science*, **10**, 159–164.

## Supplementary Material

The following supplementary material is available for this article, all in the form of figures with audio examples:

**Figure S1.** Examples of the sounds used by Vouloumanos & Werker (2007).

**Figure S2.** A hierarchy of stimuli, varying in complexity, intelligibility and periodicity.

**Figure S3.** Manipulations of 'pitchiness' in simple two-formant versions of speech.

**Figure S4.** Various combinations of source and filter properties in two-formant versions of speech.

This material is available as part of the online article from: http://www.blackwell-synergy.com/doi/abs/10.1111/j.1467-7687.2007.00550.x
(This link will take you to the article abstract).

Please note: Blackwell Publishing are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

© 2007 The Authors. Journal compilation © 2007 Blackwell Publishing Ltd.

DOI: 10.1111/j.1467-7687.2007.00551.x 

# RESPONSE

# Why voice melody alone cannot explain neonates' preference for speech

## Athena Vouloumanos[1] and Janet F. Werker[2]

*1. Department of Psychology, McGill University, Canada*
*2. Department of Psychology, University of British Columbia, Canada*

This is a response to the commentary on Vouloumanos and Werker (2007) by Rosen and Iverson (2007).

Are humans born with a bias for listening to the vocalizations of their species? In Vouloumanos and Werker (2007, this issue), we present data demonstrating that from birth, the human infant prefers listening to speech, compared with non-speech sounds that mimic spectral and temporal properties of speech. Rosen and Iverson (2007, this issue) criticize this interpretation, first arguing that the preference we have shown is based on voice melody rather than speech *per se*; second, they argue that such a voice melody preference likely stems from prenatal learning, rather than from an innate bias – a claim we didn't make in Vouloumanos and Werker, but that is addressed by new data we present here.

Turning to Rosen and Iverson's first point, although we agree that a voice melody account of newborns' preference for speech is not altogether impossible, we find it implausible. Voice melody, or pitch, is the subjective highness or lowness of a sound as perceived by the human ear. Although pitch extraction is not fully understood (e.g. Patel & Balaban, 2001), in natural speech, pitch is generally perceived as the fundamental frequency (F0) of an utterance, which is the frequency with which a particular speaker's vocal folds vibrate (typically around 200 Hz for a woman's voice). Because of resonance properties of sound, F0 is reflected in 'harmonics' at integer multiples of F0 (e.g. an F0 of 150Hz (itself the first harmonic), will have harmonics at 300 Hz, 450 Hz, 600 Hz, etc.), which contribute to the perception of pitch if, for example, F0 is missing. Research on infant pitch perception is limited, but has shown that 7-month-old infants demonstrate some adult-like characteristics in their perception of pitch (Montgomery & Clarkson, 1997). Even at this age, however, there is considerable variation in individual infants' abilities to recover pitch when F0 is missing (Clarkson, 1992). Pitch extraction in younger infants is currently poorly understood but is believed to differ from adult pitch perception (Bundy, Colombo & Singer, 1982; Clarkson, 1992). Though neonates are sensitive to pitch contours, discriminating, for example, high-low pitch from low-high pitch in bimoraic stimuli (Nazzi, Floccia & Bertoncini, 1998), the mechanism of pitch extraction in neonates has not been investigated.

To examine neonates' preference for speech, the non-speech sounds we used were a variant on sine-wave analogues (SWA) of speech (Remez, Rubin, Pisoni & Carrell, 1981). SWA consist of time-varying sinusoidal waves, or sine waves, that track the centre frequencies of the energy bands (formants) of natural speech to reproduce the changes in these frequency peaks across time. SWA are typically composed of three sinusoidal waves that reproduce the changes in the first three formants of speech so adroitly that under the right circumstances, adult listeners perceive SWA as intelligible (if weird) speech (Remez *et al.*, 1981). At stake here is which component of our SWA conveyed the perception of pitch. Rosen and Iverson suggest that because the first formant (F1) is usually heard as conveying pitch in SWA, even with the addition of F0 (Remez & Rubin, 1984), F1 is likely to convey perceived pitch in our stimuli as well, and thus, the voice melody perceived in our SWA is less salient compared to that in the speech set. We would argue that the F0 component in our SWA was salient, and that it, rather than F1, accounted for the perceived pitch. The key lies in the construction of the stimuli by Sonya Bird and Guy Carden, of the University of Victoria, and the University of British Columbia, respectively. While creating the SWA, they found that the first three formants (F1, F2, and F3) were virtually identical across the multiple natural speech tokens. For this reason, they selected *one* representative set of the first three formants

Address for correspondence: Athena Vouloumanos, Department of Psychology, McGill University, 1205 Dr. Penfield Avenue, Montreal, QC, H3A 1B1, Canada; e-mail: athena.vouloumanos@mcgill.ca

© 2007 The Authors. Journal compilation © 2007 Blackwell Publishing Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

from *one* token, and created the F1-F2-F3 sine-wave complex from this one token. The four different SWA used in the study were then created by superimposing a sine wave tracking the F0 of the four natural infant-directed speech tokens onto this *single* F1-F2-F3 complex. Inasmuch as the listener can hear any difference between the different SWA tokens, this difference is specified *entirely* by F0, because it is the only component that differs between the tokens. Even the most casual listener presented with the different non-speech tokens nonetheless hears them as distinct (non-speech tokens can be heard at http://www.phon.ucl.ac.uk/reports/DevScience2006/). The pitch contour that conveys voice melody in the sine-wave analogues can be heard readily, requiring neither careful nor analytic listening. A preference for speech is thus unlikely to be captured entirely by a preference for the voice melody of speech.

Second, independent of whether the non-speech stimuli capture voice melody to the same extent that natural speech tokens do, the prenatal environment is unlikely to provide the kind of information that Rosen and Iverson claim it does with respect to voice melody in these two sets of sounds. While we made no claim about innateness in Vouloumanos and Werker (2007), we more carefully address the kind of information available prenatally in a follow-up (unreported) experiment. In this experiment, we low pass filtered (LPF) the speech and SWA sounds with a 400-Hz filter to emulate what a foetus would be likely to hear (Abrams & Gerhardt, 2000). This frequency range is sufficient to convey information about the mother's voice (Spence & Freeman, 1996) and about the native language (Mehler, Jusczyk, Lambertz, Halsted, Bertoncini & Amiel-Tison, 1988), and is consistent with human post-natal preferences (DeCasper & Fifer, 1980; Moon, Cooper & Fifer, 1993). We tested whether neonates could discriminate between LPF speech and LPF non-speech using the Cowan method (Cowan, Suomi & Morse, 1982), which compares infants' habituation slopes for different types of stimuli (Floccia, Nazzi & Bertoncini, 2000). When discriminable stimuli are presented in alternating minutes, newborns will maintain their high amplitude sucking rate, whereas when stimuli are non-discriminable, they are treated as a single repeating stimulus, and newborns' sucking rates decrease significantly. If newborns treat LPF speech and LPF SWA as discriminable stimuli, they should maintain their sucking rate. If, however, LPF speech and LPF SWA are not discriminable, newborns should show a significant decrease in their sucking rate. Pilot data are clear: When we present these two alternating sets of LPF sounds to newborns, their high amplitude sucking rate decreases significantly, suggesting that neonates cannot discriminate between LPF speech and LPF SWA. This suggests that the information required to discriminate between SWA and bona fide speech is contained in higher frequencies which are severely attenuated, if available at all, in the prenatal environment. In short, whatever aspect of voice melody infants are familiar with prenatally is not likely to be sufficient to discriminate between our speech and SWA tokens, and thus prenatal familiarity with voice melody *per se* is unlikely to account for neonates' preference for speech.

In addition to confirming that voice melody available prenatally is indistinguishable between speech and our sine-wave stimuli, and thus is unlikely to account for post-natal preferences, these new data reduce the range of plausible roles for human prenatal listening experience in the preference for speech over sine-wave analogues reported in Vouloumanos and Werker (2007). This suggests the intriguing possibility that human neonates' preference for speech could be innate.

## Acknowledgements

We thank Stuart Rosen and Evan Balaban for useful discussions, and Gary Marcus for comments on an earlier draft. Research was funded by Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery grant 312281-05 (AV), and NSERC Discovery grant RGP81103, the Human Frontiers Science Program, and a Canada Research Chair (JFW).

## References

Abrams, R.M., & Gerhardt, K.J. (2000). The acoustic environment and physiological responses of the fetus. *Journal of Perinatology*, **20** (8 Pt 2), S31–S36.

Bundy, R.S., Colombo, J., & Singer, J. (1982). Pitch perception in young infants. *Developmental Psychology*, **18** (1), 10–14.

Clarkson, M.G. (1992). Infants' perception of low pitch. In L.A. Werner & E.W. Rubel (Eds.), *Developmental psychoacoustics* (pp. 159–188). Washington, DC: American Psychological Association.

Cowan, N., Suomi, K., & Morse, P.A. (1982). Echoic storage in infant perception. *Child Development*, **53** (4), 984–990.

DeCasper, A.J., & Fifer, W.P. (1980). Of human bonding: newborns prefer their mothers' voices. *Science*, **208** (4448), 1174–1176.

Floccia, C., Nazzi, T., & Bertoncini, J. (2000). Unfamiliar voice discrimination for short stimuli in newborns. *Developmental Science*, **3** (3), 333–343.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, **29** (2), 143–178.

Montgomery, C.R., & Clarkson, M.G. (1997). Infants' pitch perception: masking by low- and high-frequency noises.

© 2007 The Authors. Journal compilation © 2007 Blackwell Publishing Ltd.

*Journal of the Acoustical Society of America*, **102** (6), 3665–3672.

Moon, C., Cooper, R.P., & Fifer, W.P. (1993). Two-day-olds prefer their native language. *Infant Behavior and Development*, **16** (4), 495–500.

Nazzi, T., Floccia, C., & Bertoncini, J. (1998). Discrimination of pitch contours by neonates. *Infant Behavior and Development*, **21** (4), 779–784.

Patel, A.D., & Balaban, E. (2001). Human pitch perception is reflected in the timing of stimulus-related cortical activity. *Nature Neuroscience*, **4** (8), 839–844.

Remez, R.E., & Rubin, P.E. (1984). On the perception of intonation from sinusoidal sentences. *Perception and Psychophysics*, **35** (5), 429–440.

Remez, R.E., Rubin, P.E., Pisoni, D.B., & Carrell, T.D. (1981). Speech perception without traditional speech cues. *Science*, **212** (4497), 947–949.

Rosen, S., & Iverson, P. (2007). Constructing adequate non-speech analogues: what *is* special about speech anyway? *Developmental Science*, **10** (2), 165–169.

Spence, M.J., & Freeman, M.S. (1996). Newborn infants prefer the maternal low-pass filtered voice, but not the maternal whispered voice. *Infant Behavior and Development*, **19** (2), 199–212.

Received 19 October 2005
Accepted: 1 February 2006

© 2007 The Authors. Journal compilation © 2007 Blackwell Publishing Ltd.