

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

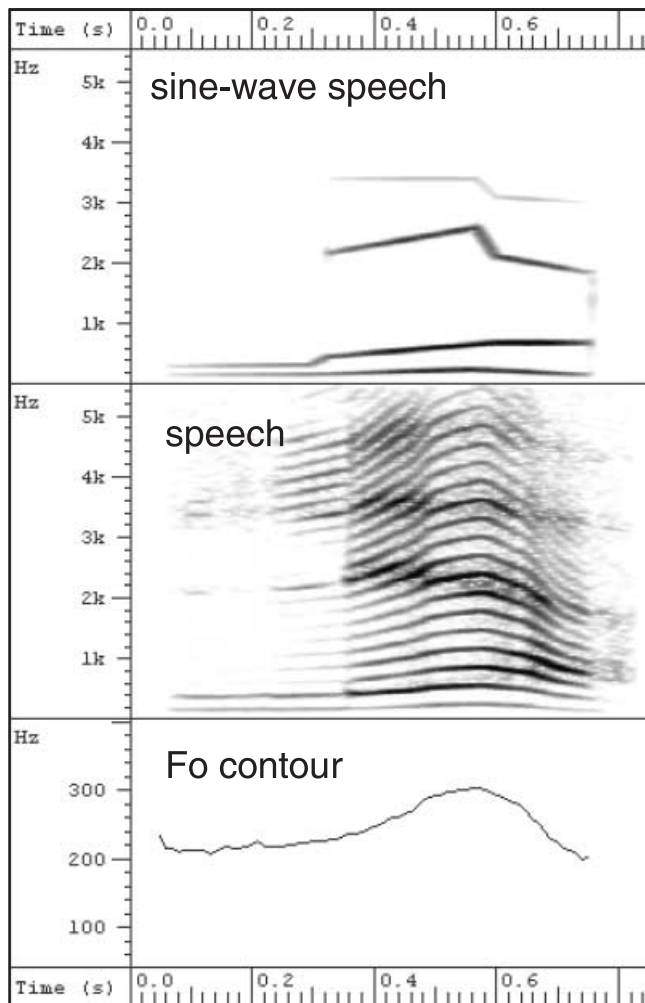


Figure 1 Examples of the sounds used by V&W in their study, along with the fundamental frequency (F_0) contour extracted directly from the speech sound. The top two panels are narrow-band spectrograms. Note the multitude of harmonics representing the fundamental frequency in the speech sound. To listen 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).

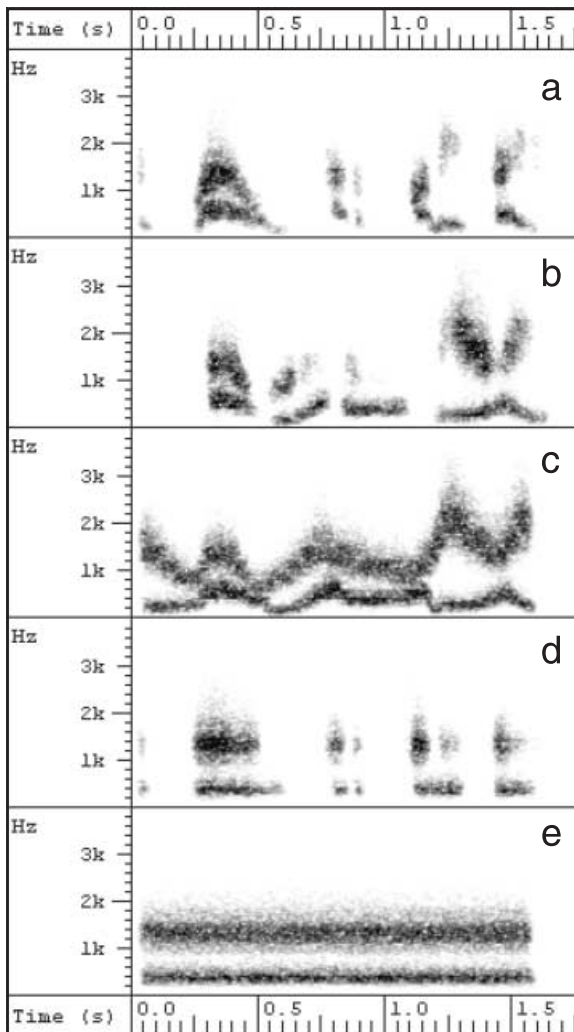


Figure 2 Spectrograms of a hierarchy of stimuli, varying in complexity and intelligibility, all constructed using the first two formants of sine-wave speech. Once the manipulations of the sine-wave formants are done, the stimuli are passed through a 16-channel noise-excited vocoder (Shannon et al., 1995) 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., Bertoincini, 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., Bertoincini, 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

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.