On the dichotomy in auditory perception between temporal envelope and fine structure cues[a] (L)

Fan-Gang Zeng,[b] Kaibao Nie, Sheng Liu, Ginger Stickney, Elsa Del Rio, Ying-Yee Kong, and Hongbin Chen

Hearing and Speech Research Laboratory, Departments of Anatomy and Neurobiology, Biomedical Engineering, Cognitive Sciences, and Otolaryngology-Head and Neck Surgery, University of California, Irvine, California 92697-1275

(Received 12 December 2003; revised 1 June 2004; accepted 10 June 2004)

It is important to know what cues the sensory system extracts from natural stimuli and how the brain uses them to form perception. To explore this issue, Smith, Delgutte, and Oxenham [Nature (London) **416**, 87–90 (2002)] mixed one sound's temporal envelope with another sound's fine temporal structure to produce auditory chimaeras and found that ''the perceptual importance of the envelope increases with the number of frequency bands, while that of the fine structure diminishes.'' This study addressed two technical issues related to natural cochlear filtering and artificial filter ringing in the chimaerizing algorithm. In addition, this study found that the dichotomy in auditory perception revealed by auditory chimaeras is an epiphenomenon of the classic dichotomy between low- and high-frequency processing. Finally, this study found that the temporal envelope determines sound location as long as the interaural level difference cue is present. The present result reinforces the original hypothesis that the temporal envelope is critical for speech perception whereas temporal fine structure is critical for pitch perception, but does not support the assertion regarding the temporal envelope and fine structure as the acoustic basis for the ''what'' and ''where'' mechanisms.
© *2004 Acoustical Society of America.* [DOI: 10.1121/1.1777938]

PACS numbers: 43.64.Sj, 43.66.Ba, 43.66.Hg [KWG] Pages: 1351–1354

I. INTRODUCTION

Classical dichotomies in auditory perception include pitch encoding with temporal and spectral cues (Licklider, 1951), sound localization with interaural time and level differences (Rayleigh, 1907), and listening with either the ''auditory'' or ''speech'' mode (Liberman and Mattingly, 1989). Recently, Smith *et al.* (2002) showed a new dichotomy in auditory perception between temporal envelope and fine structure cues. They first digitally filtered a wideband signal (80–8 820 Hz) into 1–64 frequency bands and then used the Hilbert transform to decompose the band-limited signal into a slowly varying envelope component and a fast varying fine-structure component. To assess relative contributions of temporal envelope and fine structure to auditory perception, they produced ''auditory chimaeras'' by mixing one sound's envelope with another sound's fine structure. By conducting listening tests with the chimaeric sounds, they found that the envelope is most important for speech reception, and the fine structure is most important for pitch perception and sound localization.

Smith *et al.*'s (2002) report has generated much interest in the field and already stimulated several new studies in speech perception (Xu and Pfingst, 2003) and in cochlear implants (Hong *et al.*, 2003; Litvak *et al.*, 2003). Here we address two technical issues in Smith *et al.*'s chimaerizing algorithm. One issue is related to an over-interpretation of the role of the fine structure cue with a few frequency bands (e.g., 1 and 2). The other issue is related to a filter ringing artifact that may have contributed to an over-interpretation of the role of the envelope cue with a large number of frequency bands (e.g., 32, 48, and 64). In addition, we take issue with the generality of Smith *et al.*'s assertion that the temporal envelope and fine structure are the acoustic basis of ''what'' and ''where'' cortical mechanisms.

II. ENVELOPES RECOVERED BY COCHLEAR FILTERING

Smith *et al.* (2002) found 70%–90% correct performance for chimaeric sounds with one- or two-band speech fine structure and noise envelope. This result was reminiscent of the classical work by Licklider and Pollack (1948) showing that infinite amplitude clipping had minimal effects on speech recognition. Because both the noise envelope in the speech-noise chimaeric sound and the envelope in the infinitely clipped speech were relatively flat, the good speech performance appeared to suggest a significant contribution of the fine structure to speech intelligibility. This suggestion is inconsistent with previous work showing that the fine structure with flattened $\frac{1}{4}$-oct band temporal envelopes contributed only about 17% speech intelligibility (Drullman, 1995). This inconsistency lies in the auditory filter's ability to recover the narrow-band speech envelope from the broad-band speech fine structure (Ghitza, 2001).

To demonstrate the auditory filter's ability to recover the narrow-band speech envelope, Fig. 1 shows the output of the gammachirp auditory filters (Irino and Patterson, 2001) in response to the original speech (top panel) as well as to the

[a]Portions of this work were presented at the 26th Midwinter Meeting of the Association for Research in Otolaryngology, Daytona Beach, FL, 2003.

[b]Address correspondence to University of California, 364 Med Surge II, Irvine, CA 92697-1275. Electronic mail: fzeng@uci.edu

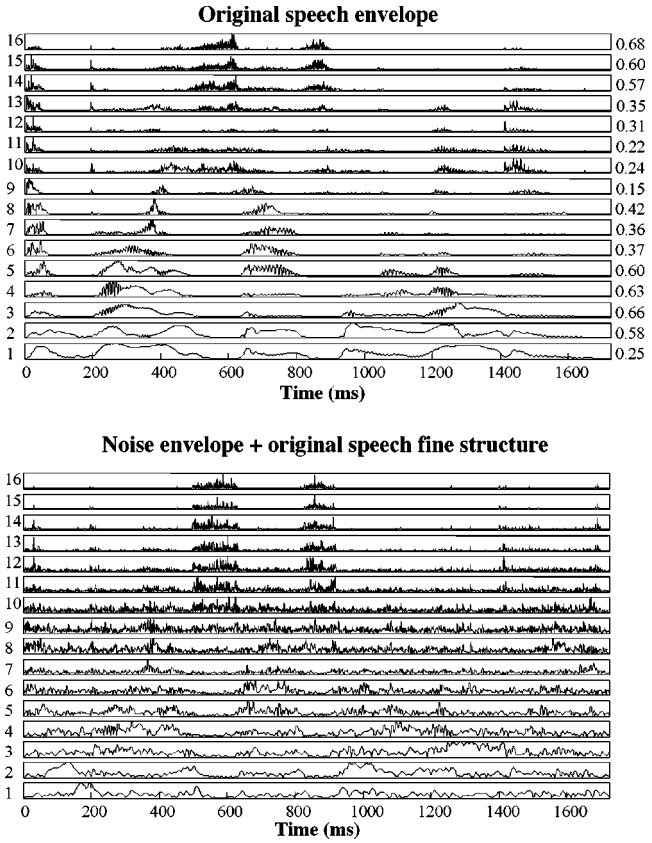

FIG. 1. Narrow-band envelopes derived from the output of 16 gammachirp filters in response to the original speech "A boy fell from the window" (top panel) and the one-band chimaeric sound with the speech fine structure and the noise envelope (bottom panel). Numbers displayed to the left of the panels represent different frequency bands. Numbers displayed to the right of the top panel are correlation coefficients between the original speech and the chimaeric sound's narrow-band envelopes.

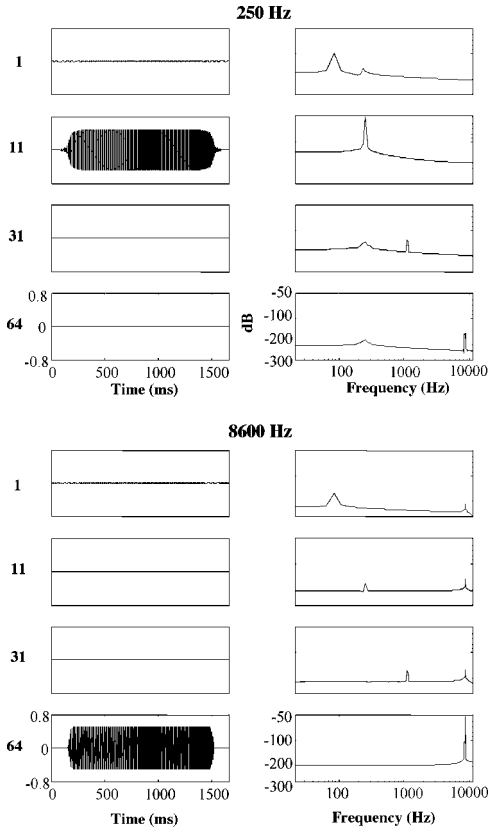

FIG. 2. Waveforms (left panels) and their Fourier amplitude spectra (right panels) at the output of selective bands (labeled by the number displayed to the left of the waveform panels) in a 64-band processor in response to two sinusoids at 250 (top panels) and 8600 Hz (bottom panels). The spectral peak at 250 Hz in the top-right four panels represents the 250-Hz stimulus component while additional peaks shifted systematically from low to high frequencies representing the ringing at each band's center frequency. Similarly, the spectral peak at 8600 Hz in the bottom-right four panels represents the stimulus while the additional peaks represent ringing artifacts.

one-band chimaerized stimulus with the noise envelope but the speech fine structure (bottom panel). The chimaeric sound's narrow-band envelopes were significantly correlated with the original speech envelopes with the mean coefficient of 0.44 from all 16 bands and a range from 0.68 at the high frequency band to 0.15 at the intermediate frequency band, suggesting that the auditory filters could at least partially recover the original narrow-band speech envelope from the one-band speech fine structure. To quantify the contribution of the recovered narrow-band speech envelopes, we additionally amplitude modulated them to noise having the same bandwidth as the original gammachirp filters (Shannon *et al.*, 1995) and measured their performance using the HINT sentences (Nilsson *et al.*, 1994) in four normal-hearing subjects. The mean percent correct score was 79% (SD=9%) for the original one-band chimaerized sound, similar to the 70% score obtained in the Smith *et al.* study. However, we found a mean score of 40% (SD=12%) for the additionally synthesized sound, suggesting that the recovered envelope could account for at least half of the performance from the one-band chimaerized sound that contained the speech fine structure and the noise envelope. Therefore, taking cochlear filtering into account, the present result has effectively removed the number of bands as a significant factor and reinforced the original idea that the temporal envelope is critical for speech recognition (Van Tasell *et al.*, 1987; Drullman, 1995; Shannon *et al.*, 1995).

III. ARTIFICAL FINE STRUCTURE INTRODUCED BY FILTER RINGING

Smith *et al.* (2002) observed a contribution of the envelope to pitch recognition only with a large number of frequency bands (e.g., 48 and 64). We found this result counter intuitive. Consider the following simplistic case in pitch perception, namely chimaerizing two sinusoids. In this case, the Hilbert envelope of a sinusoid is the amplitude and its Hilbert fine structure is the cosine of the phase of the analytic signal $[\cos(2\pi ft)]$, with frequency being the only difference between the two sinusoids. To the extent that pitch is related to frequency, amplitude (i.e., temporal envelope) should not determine the pitch of a sinusoid.[1] Why then could Smith *et al.* come to the conclusion that pitch is determined by the envelope with the large number of bands?

Figure 2 shows the waveform and its Fourier spectrum at the output of a few selected bands (the total number of bands=64) in response to a sinusoid at either 250 Hz (top panels) or 8600 Hz (bottom panels). In the top panels, the 11th band had a center frequency of 258 Hz and expectedly produced the largest response to the 250-Hz sinusoidal

stimulus. The other bands' responses were much smaller (hardly seen in the waveform) but contained both the leaked stimulus at 250 Hz (the peak aligned to that in the 11th band) as well as the ringing located at the filter's center frequency (the additional peak whose position systematically increased with the band number). The same pattern can be observed for the filter responses to the 8600-Hz sinusoidal stimulus.

When chimaerizing between the two sinusoids, the large envelope in the band that contained the stimulus (e.g., the 64th band containing the 8600 Hz sinusoid) would be combined with the fine structure in the same band in response to the 250-Hz sinusoid, which had a dominant component at the ringing frequency of 8565 Hz. This ringing frequency was close enough to 8600 Hz, resulting in an impression that the envelope determines pitch when the number of frequency bands is large as reported by Smith *et al.* (2002). With a large number of bands, the difference is small between the ringing frequency and the original signal frequency, making the differentiation between Smith *et al.*'s envelope argument and the present ringing argument subtle in a practical sense. However, the theoretical differentiation is important because the ringing is primarily determined by the filter properties, such as the center frequency and the bandwidth, having nothing to do with the original signal; we suggest that the data showing dependence of pitch perception on the temporal envelope obtained by Smith *et al.* at these large frequency bands (32–64) are an artifact of filter ringing.

IV. "WHERE" VERSUS "WHAT": A NEW OR OLD DICHOTOMY?

Smith *et al.* (2002) showed a dichotomy in speech recognition and sound localization by presenting chimaeric speech sounds with either a 700-μs delayed envelope or fine structure. With the 16- and 32-band conditions (their Fig. 4), they clearly demonstrated that the envelope cue is important for speech recognition and the fine-structure cue is important for sound localization, revealing a possible acoustic basis for the hypothesized "what" and "where" pathways in the auditory cortex (Tian *et al.*, 2001). This result is provocative but lacks the necessary generality to support its assertion.

First, to examine what acoustic cues were responsible for the observed dichotomy in the Smith *et al.* results, we used an identical algorithm and procedure to measure the intelligibility and localization of an exemplary condition in the Smith *et al.* (2002) study, in which one sentence's envelope was mixed via a 16-band chimaerizer with another sentence's fine structure that had a 700-μs ITD (leading in the right ear). In addition, we processed those stimuli by the classic low- and high-pass filtering (sixth-order Butterworth filters). Figure 3 shows averaged lateralization (top panel) and intelligibility (bottom panel) data as a function of the cutoff frequency for low-pass (filled circles) and high-pass (open triangles) filtering of the chimaeric sound in three normal-hearing listeners. Note first in the low-pass condition that the subjects used mostly the ITD cue in the fine structure to lateralize to the right ear. Note second in the high-pass condition that the subjects did not use the ITD cue in the fine structure but rather the ILD cue in the envelope to lateralize the sound (i.e., a center position in the head). This was true

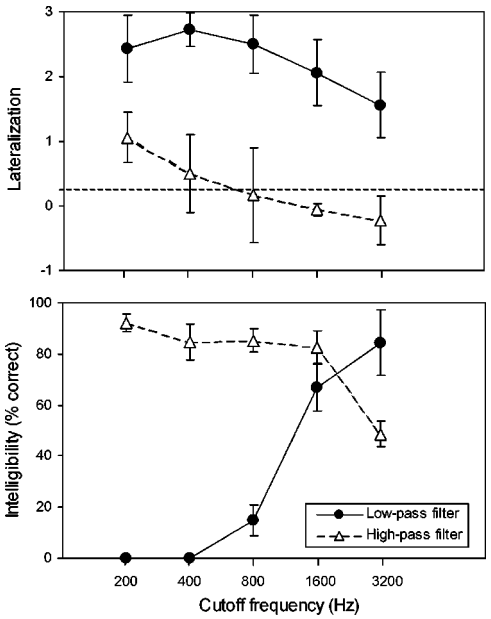

FIG. 3. Lateralization (top panel) and intelligibility (bottom panel) results from 16-band chimaerized sounds as a function of cutoff frequency for low-pass (filled circles) and high-pass (open triangles) filters. The averaged data from three normal-hearing subjects are represented by symbols while the error bars represent plus or minus one standard deviation. In the top panel, the "3" on the *y* axis represents the subject's complete lateralization to the right ear while the "0" represents the center position. The intelligibility scores are a percentage of the key words correctly identified.

even with the 200-Hz high-pass condition, in which the subjects only slightly lateralized (scale=1) to the right ear. A repeated-measure ANOVA confirmed that the difference between the low- and high-pass conditions was significant [$F(1,2)=135.32; p<0.01$]. This result suggests that lateralization using the ITD cue in the present chimaeric sound is essentially a low-frequency phenomenon.

Note a totally different pattern of results for the intelligibility data, in which those low-frequency components (<800 Hz) that dominated lateralization contributed essentially nothing to speech intelligibility (see the three leftmost filled circles representing 200-, 400-, and 800-Hz low-pass cutoff frequencies). On the contrary, it was the high-frequency components that contributed to speech intelligibility. Thus the present low-pass and high-pass filtering data clearly demonstrate that the perceptual dichotomy between the temporal envelope and fine structure cues observed in the Smith *et al.* (2002) study was due to the well-known difference between the use of the low-frequency ITD cue for lateralization and the use of the high-frequency cue for intelligibility (Rayleigh, 1907; French and Steinberg, 1947; Oppenheim and Lim, 1981).

Second, to directly test the generality of Smith *et al.*'s assertion relating the temporal envelope and fine structure to the "what" and "where" mechanisms, we constructed two chimaeric stimuli that contained either a 15-dB ILD cue or conflicting ITD and ILD cues extracted from head-related transfer functions.[2] The ILD cue reflects an overall level difference that is inherently embedded in the temporal envelope, whereas the ITD cue in naturally embedded in the temporal fine structure. Figure 4 shows the effect of the ILD cue on lateralization with the ITD cue towards either left (top

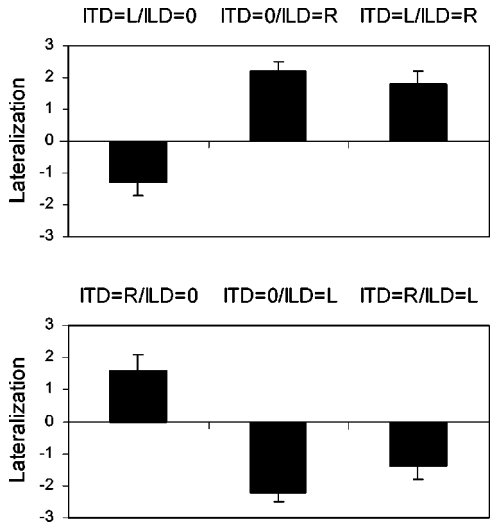

FIG. 4. Lateralization result for the ITD-only cue (left bars), the ILD-only cue (middle bars), and the conflicting ITD and ILD cue (left bars). The top panel shows results from a condition where the ITD cue is either left or neutral whereas the bottom panel shows the result from a condition where the ITD cue is either right or neutral. The data are averaged from five normal-hearing subjects with the error bars representing plus or minus one standard deviation.

panel) or right (bottom panel). Three experimental conditions included the original ITD-only control in Smith *et al.*'s study (left bars), the ILD-only condition (middle bars), and the conflicting ITD and ILD condition (right bars). Although the conflicting ITD cue slightly reduced the ILD-alone effect (0.5 to 1 on the −3 to 3 scale used), the result clearly showed that whenever the ILD cue is present, the temporal envelope, rather than the fine structure, largely determines sound location.

V. FINAL REMARKS

Smith *et al.* (2002) presented an innovative algorithm (''auditory chimaeras'') to assess the relative contribution of temporal envelope and fine structure to auditory perception. Here we addressed two technical issues that one should be aware of when using their chimaerizing algorithm and in interpreting the results derived from auditory chimaeras. First, one should be aware of the ear's natural ability to recover the narrow-band envelope with broad-band processing for a small number of frequency bands (e.g., 1 and 2). Second, one should be concerned about filter ringing artifacts with narrow-band processing for a large number of bands (e.g., 32, 48, and 64). In addition, we performed a classic filtering manipulation on the chimaerized sounds and found clear evidence suggesting that the dichotomy revealed by the auditory chimaeras is an epiphenomenon of classic duplex perception between low- and high-frequency pathways. Finally, we provided two counter examples showing that the temporal envelope via an ILD cue determines sound location for both conditions where the ITD cue was either neutral or in conflict with the ILD cue. The present result reinforces the previously proposed hypothesis that the temporal envelope is critical for speech recognition whereas the temporal fine structure is critical for pitch perception. However, the present result does not support the assertion regarding the temporal envelope and fine structure as the acoustic basis for ''what'' and ''where'' mechanisms.

ACKNOWLEDGMENTS

We thank Bertrand Delgutte, Andrew Oxenham, Zachary Smith, Ken Grant, Robert Shannon, Van Summers, and an anonymous reviewer for helpful comments on earlier versions of this manuscript. We also thank Smith, Delgutte, and Oxenham for providing the chimaerizing algorithm, Ruth Litovsky for providing the head related transfer function, and G. Bruce Henning for referencing to Oppenheim and Lim's study. Preparation of this manuscript was supported by a grant from the National Institutes of Health (RO1-DC-02267).

[1]It is noted that actual pitch perception of a pure tone is dependent on frequency as well as amplitude (Stevens, 1935). However, this amplitude-dependent change in pitch is generally small (~10%) and requires a large change in amplitude (~40 dB).

[2]The head-related transfer function (HRTF) was provided by Ruth Litovsky and recorded in a human manikin. One sentence was convolved with a HRTF recorded from the right ear whereas another sentence was convolved with a HRTF recorded from the left ear. Chimearizing these two HRTF-convolved sentences produced the stimuli used in Fig. 4, in which they contained conflicting ITD and ILD cues.

Drullman, R. (**1995**). ''Temporal envelope and fine structure cues for speech intelligibility,'' J. Acoust. Soc. Am. **97**, 585–592.

French, N. R., and Steinberg, J. C. (**1947**). ''Factors governing the intelligibility of speech sounds,'' J. Acoust. Soc. Am. **19**, 90–119.

Ghitza, O. (**2001**). ''On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception,'' J. Acoust. Soc. Am. **110**, 1628–1640.

Hong, R. S., Rubinstein, J. T., Wehner, D., and Horn, D. (**2003**). ''Dynamic range enhancement for cochlear implants,'' Otol. Neurotol. **24**, 590–595.

Irino, T., and Patterson, R. D. (**2001**). ''A compressive gammachirp auditory filter for both physiological and psychophysical data,'' J. Acoust. Soc. Am. **109**, 2008–2022.

Liberman, A. M., and Mattingly, I. G. (**1989**). ''A specialization for speech perception,'' Science **243**, 489–494.

Licklider, J. C. R. (**1951**). ''A Duplex Theory of Pitch Perception,'' Experientia **7**, 128–134.

Licklider, J. C. R., and Pollack, I. (**1948**). ''Effects of differentiation, integration, and infinite peak clipping on the intelligibility of speech,'' J. Acoust. Soc. Am. **20**, 42–51.

Litvak, L. M., Delgutte, B., and Eddington, D. K. (**2003**). ''Improved temporal coding of sinusoids in electric stimulation of the auditory nerve using desynchronizing pulse trains,'' J. Acoust. Soc. Am. **114**, 2079–2098.

Nilsson, M., Soli, S. D., and Sullivan, J. A. (**1994**). ''Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise,'' J. Acoust. Soc. Am. **95**, 1085–1099.

Oppenheim, A. V., and Lim, J. S. (**1981**). ''The importance of phase in signals,'' Proc. IEEE **69**, 529–541.

Rayleigh, L. (**1907**). ''On our perception of sound direction,'' Philos. Mag. **13**, 214–232.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). ''Speech recognition with primarily temporal cues,'' Science **270**, 303–304.

Smith, Z. M., Delgutte, B., and Oxenham, A. J. (**2002**). ''Chimaeric sounds reveal dichotomies in auditory perception,'' Nature (London) **416**, 87–90.

Stevens, S. (**1935**). ''The relation of pitch to intensity,'' J. Acoust. Soc. Am. **6**, 150–154.

Tian, B., Reser, D., Durham, A., Kustov, A., and Rauschecker, J. P. (**2001**). ''Functional specialization in rhesus monkey auditory cortex,'' Science **292**, 290–293.

Van Tasell, D. J., Soli, S. D., Kirby, V. M., and Widin, G. P. (**1987**). ''Speech waveform envelope cues for consonant recognition,'' J. Acoust. Soc. Am. **82**, 1152–1161.

Xu, L., and Pfingst, B. E. (**2003**). ''Relative importance of temporal envelope and fine structure in lexical-tone perception,'' J. Acoust. Soc. Am. **114**, 3024–3027.