Behavioral/Systems/Cognitive

Psychophysiological Analyses Demonstrate the Importance of Neural Envelope Coding for Speech Perception in Noise

Jayaganesh Swaminathan[1] **and Michael G. Heinz**[1,2]

[1]Department of Speech, Language, and Hearing Sciences and [2]Weldon School of Biomedical Engineering, Purdue University, West Lafayette, Indiana 47907

Understanding speech in noisy environments is often taken for granted; however, this task is particularly challenging for people with cochlear hearing loss, even with hearing aids or cochlear implants. A significant limitation to improving auditory prostheses is our lack of understanding of the neural basis for robust speech perception in noise. Perceptual studies suggest the slowly varying component of the acoustic waveform (envelope, ENV) is sufficient for understanding speech in quiet, but the rapidly varying temporal fine structure (TFS) is important in noise. These perceptual findings have important implications for cochlear implants, which currently only provide ENV; however, neural correlates have been difficult to evaluate due to cochlear transformations between acoustic TFS and recovered neural ENV. Here, we demonstrate the relative contributions of neural ENV and TFS by quantitatively linking neural coding, predicted from a computational auditory nerve model, with perception of vocoded speech in noise measured from normal hearing human listeners. Regression models with ENV and TFS coding as independent variables predicted speech identification and phonetic feature reception at both positive and negative signal-to-noise ratios. We found that: (1) neural ENV coding was a primary contributor to speech perception, even in noise; and (2) neural TFS contributed in noise mainly in the presence of neural ENV, but rarely as the primary cue itself. These results suggest that neural TFS has less perceptual salience than previously thought due to cochlear signal processing transformations between TFS and ENV. Because these transformations differ between normal and impaired ears, these findings have important translational implications for auditory prostheses.

Introduction

One way to characterize a speech waveform is as the sum of a number of amplitude-modulated signals representing the outputs of a set of narrow frequency channels distributed across the acoustic spectrum (Flanagan, 1980). In this view, the output of each channel can be separated into a rapidly varying carrier signal that specifies the waveform temporal fine structure (TFS) and a slowly varying modulating signal that specifies its temporal envelope (ENV). It is of considerable theoretical and translational interest to understand the relative roles of ENV and TFS for robust speech perception. Findings from several psychoacoustic studies suggest that there is an important dichotomy between ENV and TFS cues; however, this remains a topic of considerable debate (Drullman, 1995; Smith et al., 2002; Oxenham and Simonson, 2009).

Envelope cues have been shown to support robust speech identification in quiet when provided in as few as four frequency bands (Shannon et al., 1995). This finding has important implications for cochlear implants (CIs), which currently only provide ENV information over a relatively small (~8) number of effective channels, and is consistent with the observation that many CI patients understand speech remarkably well in quiet (Wilson et al., 1991). However, ENV cues have been shown to be susceptible to noise degradations (Fu et al., 1998). Conversely, TFS cues have been suggested to be important for speech perception in noise (Qin and Oxenham, 2003; Zeng et al., 2005). The lack of TFS in current CI stimulation strategies has been hypothesized to underlie some of the difficulties CI patients have, e.g., understanding speech in noise (Fu and Shannon, 1999) and appreciating music (Drennan and Rubinstein, 2008). These perceptual difficulties have motivated efforts to develop novel CI strategies to provide TFS in addition to ENV (e.g., Rubinstein et al., 1999; Litvak et al., 2003; Nie et al., 2005; Zeng et al., 2005).

Many conclusions about the relative roles of ENV and TFS cues for speech perception have been derived from listening experiments using vocoders (Dudley, 1939; Flanagan and Golden, 1966) that retain only the ENV or TFS components of the acoustic waveform. The interpretation of perceptual vocoder studies relies on assuming that ENV and TFS can be isolated. However, narrowband cochlear filtering limits the ability to isolate TFS from ENV (Ghitza, 2001) due to transformations between acoustic TFS and (recovered) neural envelopes (see Heinz and Swaminathan, 2009, their Fig. 10). Speech ENV cues recovered from broadband TFS speech can be intelligible (Zeng et al., 2004) but are reduced (although not completely eliminated) for TFS speech created with narrow analysis bands (Gilbert and Lorenzi, 2006; Heinz and Swaminathan, 2009).

Received Aug. 30, 2011; revised Nov. 21, 2011; accepted Dec. 1, 2011.

Author contributions: J.S. and M.G.H. designed research; J.S. performed research; J.S. analyzed data; J.S. and M.G.H. wrote the paper.

Our research was supported by the Purdue Research Foundation, Weinberg funds from the Department of Speech, Language, and Hearing Sciences, and National Institutes of Health Grant R01-DC009838 (awarded to M.G.H.). We thank Gavin Bidelman, Jon Boley, Louis Braida, Ananth Chintanpalli, Jay Desloge, Judy Dubno, Kenneth Henry, Sushrut Kale, Charlotte Reed, and Michael Walls for helpful comments and discussions. We also thank Kenneth Henry for help with statistical analyses.

Correspondence should be addressed to Dr. Jayaganesh Swaminathan at his present address: 77 Massachusetts Avenue, Room 36-757, Massachusetts Institute of Technology, Cambridge, MA 02139. E-mail: jswamy@mit.edu.

DOI:10.1523/JNEUROSCI.4493-11.2012

In light of such transformations, the hypothesized role of acoustic TFS for speech perception in noise must be evaluated by factoring in the neural coding of ENV and TFS to various forms of vocoded speech. Here, we quantitatively related the neural coding of ENV and TFS predicted from a physiologically based auditory nerve (AN) model to speech identification scores collected from normal hearing listeners using the same set of noise-degraded speech stimuli.

Materials and Methods

Acoustic stimuli. Speech signals consisted of 16 consonants (/p, t, k, b, d, g, f, s, ʃ, v, z, j, m, n, r, l/) in an /a/-C-/a/ context spoken by four talkers (two for each gender), i.e., 64 vowel-consonant-vowel (VCV) utterances (Shannon et al., 1999). Recorded phonemes were transformed to equal duration (500 ms) and overall level [65 dB sound pressure level (SPL)]. A speech-shaped noise (generated with the same magnitude spectrum as the original phoneme, but with randomized phase in each interval) was added to each intact speech utterance before vocoding. Vocoded stimuli were created with either only fine structure (TFS speech) or envelope (ENV speech). Five stimulus versions that differed in their composition of acoustic ENV and TFS were used to evaluate the salience of TFS and ENV cues: (1) intact speech; (2) phonemic ENV speech (PHENV); (3) periodicity ENV speech (PDENV); (4) broadband TFS speech (BBTFS); and (5) narrowband TFS speech (NBTFS).

Specific stimulus generation details were the same as those in previous studies (Gilbert and Lorenzi, 2006; Lorenzi et al., 2006; also see Swaminathan, 2010). Each VCV was initially bandpass filtered into 1 or 16 spectral bands using third-order Butterworth filters. Analysis filter bands spanned 80–8020 Hz and were logarithmically spaced. The Hilbert transform (Hilbert, 1912) was applied in each band to decompose the signal into ENV (magnitude of the Hilbert analytic signal) and TFS (cosine of the Hilbert analytic signal phase).

For PHENV speech, Hilbert envelopes were extracted in each of the 16 bands and lowpass filtered at 64 Hz with a sixth-order Butterworth filter. For PDENV speech, the same procedures were used, except that the ENVs were bandpass filtered from 64–300 Hz and the first five bands (center frequencies $\leq \sim 250$ Hz) were eliminated to avoid aliasing effects. These ENV signals were used to amplitude modulate sine-wave carriers at the respective analysis filter center frequencies. Finally, signals were bandpass filtered with the original analysis filters and then summed across all bands to create ENV speech.

For TFS speech, VCVs were first bandpass filtered into 1 or 16 bands (BBTFS or NBTFS, respectively). The ENV component was discarded and the TFS in each band was scaled by the rms power of the original bandpass filtered signal. These power-adjusted TFS signals were summed over all bands to create TFS speech (Gilbert and Lorenzi, 2006; Lorenzi et al., 2006).

Subjects. Five male native speakers of American English, matched closely in age (mean = 28.8; SD = 1.9), right handed, and with normal hearing (0.25–8 kHz), participated in the study. Informed consent was obtained in compliance with an approved Institutional Review Board protocol from the Purdue University Human Research Protection Program.

Speech reception procedures. The noise-degraded VCVs and their vocoded homologues were presented at eight different signal-to-noise ratios (SNRs) (Q, and SNR = 10, 5, 0, −5, −10, −15, −20 dB, where Q represents in quiet). Intact VCVs were fixed at 65 dB SPL, and the rms noise level was chosen to achieve the desired SNR. Stimuli were presented with Tucker Davis Technologies hardware and software. Sennheiser HD 580 headphones were used to present sounds monaurally to the right ear. Listeners were tested in a double-walled, sound-attenuating chamber using a single-interval, 16-alternative forced choice procedure. Each block presented the entire set of VCV stimuli (16 consonants × 4 speakers) in random order for a single stimulus version (INTACT, BBTFS, NBTFS, PHENV, or PDENV speech) at a single SNR. One session (∼2 h, with breaks as needed) comprised two repetitions of eight blocks (in quiet followed by the seven decreasing SNR conditions in order). Each subject participated in five sessions (five stimulus versions), with the order of stimulus versions randomized across subjects. Training was provided only for the Q condition for each stimulus version, and feedback was provided only during training. The modest training was given only to familiarize subjects with the testing environment and task (Shannon et al., 1995).

Table 1. Phonetic features of the 16 English consonants used in this study

Consonant	Voicing	Manner of articulation	Place of articulation	Nasality
/p/	Unvoiced	Occlusive	Front	Non-nasal
/t/	Unvoiced	Occlusive	Middle	Non-nasal
/k/	Unvoiced	Occlusive	Back	Non-nasal
/b/	Voiced	Occlusive	Front	Non-nasal
/d/	Voiced	Occlusive	Middle	Non-nasal
/g/	Voiced	Occlusive	Back	Non-nasal
/f/	Unvoiced	Constrictive	Front	Non-nasal
/s/	Unvoiced	Constrictive	Middle	Non-nasal
/ʃ/	Unvoiced	Constrictive	Back	Non-nasal
/v/	Voiced	Constrictive	Front	Non-nasal
/z/	Voiced	Constrictive	Middle	Non-nasal
/j/	Voiced	Constrictive	Middle	Non-nasal
/m/	Voiced	Constrictive	Front	Nasal
/n/	Voiced	Constrictive	Front	Nasal
/r/	Voiced	Constrictive	Middle	Non-nasal
/l/	Voiced	constrictive	Middle	Non-nasal

Classification of consonants by acoustic phonetic features: voicing (voiced versus unvoiced), manner (occlusive versus constrictive), place (front versus middle versus back), and nasality (nasal versus non-nasal).

Speech reception analyses. For each stimulus version and SNR, a 16 × 16 confusion matrix from the 64 VCV tokens was compiled for each listener. The specific reception of phonetic features (see Table 1), e.g., voicing (voiced versus unvoiced), manner (occlusive versus constrictive), place (front versus middle versus back), and nasality (nasal versus non-nasal), was evaluated by information–transmission analyses (Miller and Nicely, 1955) on the individual confusion matrices. All statistical analyses were conducted on arcsine-transformed scores (Studebaker, 1985).

Computational auditory nerve model. Spike trains were obtained from a phenomenological AN model (Zilany and Bruce, 2006, 2007) that has been tested extensively against neurophysiological data obtained in response to both simple and complex stimuli, including tones, broadband noise, and speech-like sounds (Zhang et al., 2001; Zilany and Bruce, 2006, 2007). Many physiological properties associated with nonlinear cochlear tuning are captured by this model, including compression, suppression, and broadened tuning with increased sound level (Heinz, 2010). Ten high spontaneous rate AN fibers with characteristic frequencies (CFs) ranging from 200 Hz to 8 kHz were selected based on the acoustic characteristics of the VCVs. For each fiber, sound levels were chosen at the best modulation level (BML) for each stimulus type, where BML is defined as the sound level producing maximal ENV coding and was typically ∼15 dB above fiber threshold (Joris and Yin, 1992). Although this approach generally means that the sound levels used for the AN model were lower than those used for the perceptual measures (65 dB SPL), it does not provide a significant limitation to the conclusions from this study. The choice of BML is often used in single-fiber studies of modulation coding to reduce across-fiber variability associated with the non-monotonic effect of sound level, since the level dependence (relative to fiber threshold) of modulation coding is generally consistent across AN fibers (e.g., Palmer, 1982; Joris and Yin, 1992). Because modulation perception is generally level independent (Viemeister, 1979), the use of BML for individual fibers in the present study represents the assumption that at most sounds levels, some AN fibers are available to support robust envelope perception. Several factors support this assumption, including the wide range of AN fiber thresholds (Liberman, 1978), efferent-mediated dynamic range decompression in noise (e.g., Kawase et al., 1993), and long-duration dynamic range adaptation effects that occur in the AN and higher auditory centers (Dean et al., 2005; Wen et al., 2009).

Quantifying TFS and ENV coding in AN spike trains. Neural cross-correlation coefficients (ρ_{TFS} and ρ_{ENV}) were used to quantify the similarity between TFS or ENV components of different AN spike train

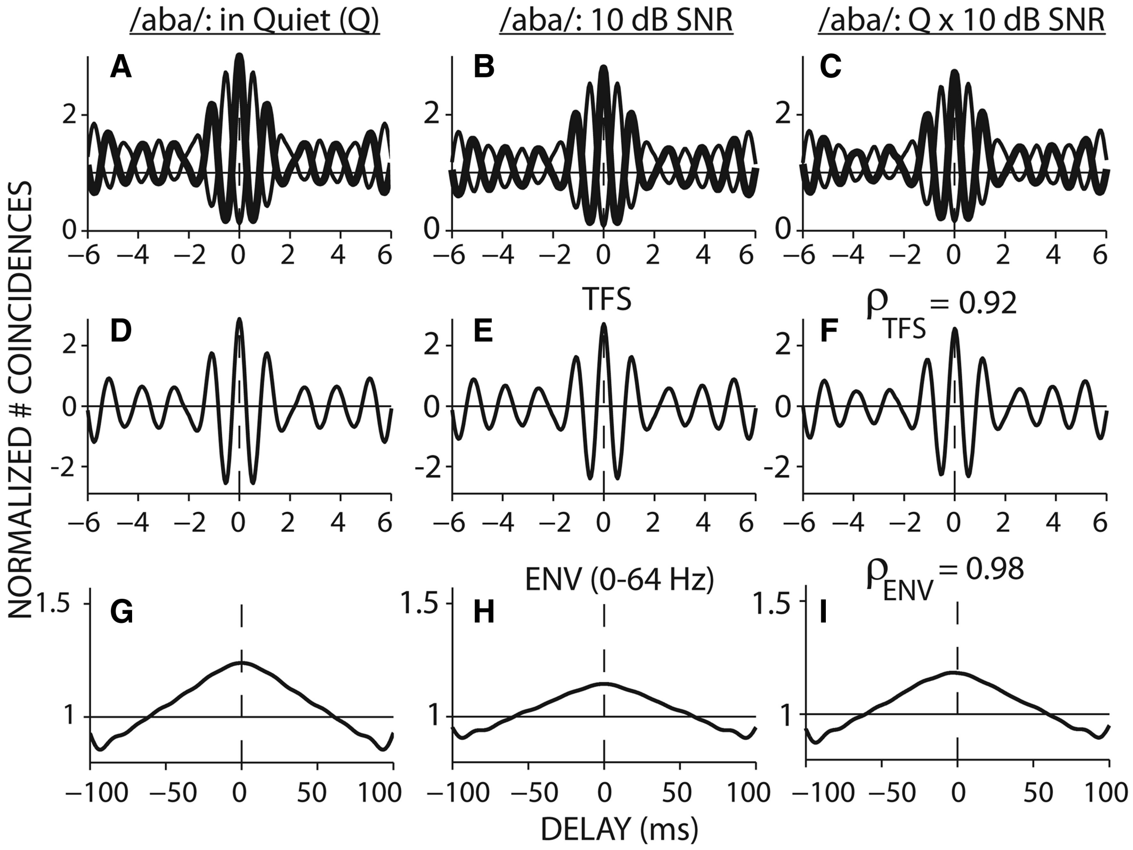

Figure 1. Correlogram analyses to quantify the neural coding of ENV and TFS in noise-degraded speech. Columns 1 and 2 show temporal coding of */aba/* in quiet (Q) and for a 10-dB SNR, respectively; column 3 illustrates the similarity in temporal coding between these two conditions. ***A***, ***B***, Normalized shuffled auto correlograms (thick line) and cross-polarity correlograms (thin line). ***C***, Shuffled cross-stimulus correlogram (thick line) and cross-polarity, cross-stimulus correlogram (thin line). ***D***–***F***, Difcors represent TFS coding, with ρ_{TFS} shown in ***F***. ***G***–***I***, Sumcors represent phonemic (0–64 Hz) ENV coding, with ρ_{ENV} shown in ***I***. Fiber CF = 1000 Hz.

responses (Heinz and Swaminathan, 2009). For each AN fiber, the salience of speech-related TFS and ENV coding following degradation (i.e., due to vocoding and/or noise) was quantified by computing ρ_{TFS} and ρ_{ENV} between intact speech in quiet (baseline) and each degraded condition.

Figure 1 illustrates the computation of these metrics for /aba/ in quiet and at 10 dB SNR for a 1 kHz CF AN fiber. Separate metrics for TFS and ENV were computed using shuffled auto-correlograms (SACs) and shuffled cross-correlograms (SCCs) (Joris, 2003; Louage et al., 2004; Joris et al., 2006). SACs are computed by tallying spike intervals across stimulus repetitions (rather than within) and yield more robust characterizations of temporal responses than classic all-order interval histograms (Ruggero, 1973). Normalized SACs are plotted as a function of delay (or inter-spike interval) and are much like auto-correlation functions (Fig. 1*A*,*B*, dark lines). TFS and ENV coding can be separated by comparing the responses to a stimulus and its polarity-inverted pair (e.g., A+ with A−) (Joris, 2003; Louage et al., 2004; Joris et al., 2006). Polarity inversion acts to invert the TFS but does not affect ENV. Cross-polarity correlograms are computed by comparing spikes from A+ and A− [e.g., SCC(A+,A−)] (Fig. 1*A*,*B*, thin lines). To emphasize TFS, difcors were computed as the difference between the SAC (original ENV, original TFS) and the cross-polarity correlogram (original ENV, inverted TFS), where the difcor peak height quantifies the strength of TFS coding. To quantify ENV coding, sumcors were computed as the average of the SAC and the cross-polarity correlogram. Here, only the neural coding of phonemic ENV information was considered by restricting the sumcor spectra to include only frequencies below 64 Hz. The third column of Figure 1 illustrates the use of SCCs to quantify the similarity between spike trains in response to intact speech in quiet (A) and degraded speech (B). Cross-stimulus correlograms [e.g., SCC(A+, B+), thick line in Figure 1*C*] and cross-stimulus, cross-polarity correlograms [e.g., SCC(A+, B−), Figure 1*C*, thin line] were computed to facilitate the separation of TFS and ENV cross-correlations by using difcors and sumcors, respectively.

Neural cross-correlation coefficients (Heinz and Swaminathan, 2009) ranging between 0 and 1 were computed by comparing the degree of response similarity (column 3 of Fig. 1) to the degree of temporal coding for each stimulus individually (columns 1 and 2 of Fig. 1). The cross-correlation coefficient for TFS was computed from the difcor peak heights as follows:

$$\rho_{\mathrm{TFS}} = \frac{\mathrm{difcor}_{\mathrm{AB}}}{\sqrt{\mathrm{difcor}_{\mathrm{A}} \times \mathrm{difcor}_{\mathrm{B}}}}. \tag{1}$$

Likewise, the neural cross-correlation coefficient for ENV was computed from sumcor peak heights (after subtracting the baseline value of 1) as follows:

$$\rho_{\mathrm{ENV}} = \frac{(\mathrm{sumcor}_{\mathrm{AB}} - 1)}{\sqrt{(\mathrm{sumcor}_{\mathrm{A}} - 1) \times (\mathrm{sumcor}_{\mathrm{B}} - 1)}}. \tag{2}$$

For the single-fiber responses in Figure 1, both ρ_{TFS} and ρ_{ENV} were high (close to 1), indicating that the temporal coding in quiet and at 10 dB SNR was quite similar.

Results

Perception of noise-degraded speech

Figure 2*A* shows identification scores for consonants as a function of SNR for intact and vocoded speech averaged across all listeners. Pooling across all SNRs, intact speech was the most intelligible, followed by PHENV and BBTFS speech (not statistically different from one another), followed by PDENV, and then

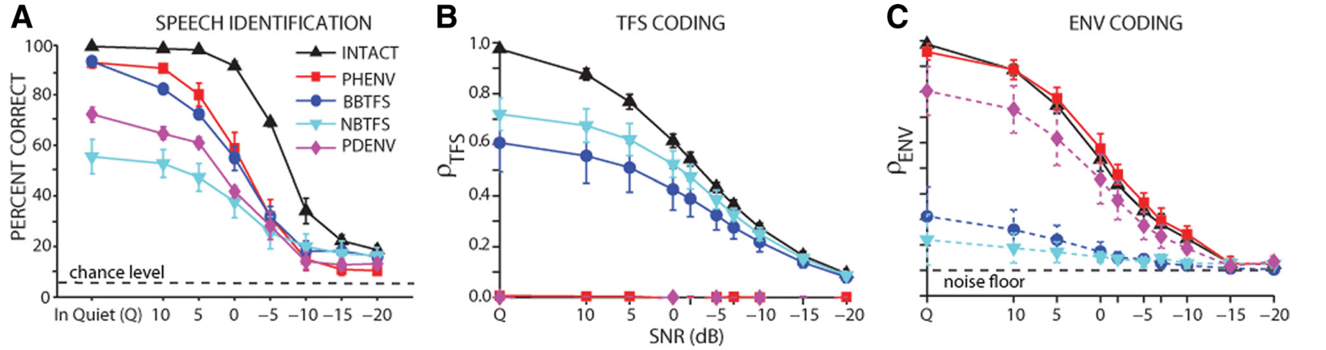

Figure 2. The effects of noise on speech identification and neural coding are compared between intact speech and four speech vocoders that differed in their composition of acoustic ENV and TFS. ***A***, Mean consonant identification scores across listeners (with SEM bars) are shown as a function of signal-to-noise ratio, SNR, with chance level (1/16) indicated by the dashed line. As noise level increased (SNR decreased), identification became more difficult in all cases; however, the relative performance across vocoders differed for positive and negative SNRs. ***B***, ***C***, Neural coding of TFS (ρ_{TFS}) and phonemic ENV (ρ_{ENV}), where the neural cross-correlation coefficients were computed between model spike train responses to the noisy vocoded speech and the intact speech in quiet. Mean ρ_{TFS} and ρ_{ENV} values across AN fibers [all eight CFs $\leq$ 2.5 kHz for TFS (Johnson, 1980); all 10 CFs $\leq$8 kHz for ENV] are plotted with SEM bars. Recovered phonemic envelope coding [e.g., from the periodicity envelope (PDENV), broadband TFS (BBTFS), and narrowband TFS (NBTFS) speech vocoders] is represented by dashed curves in *C*, whereas true phonemic envelope coding is represented by solid curves. Black dashed line at $\rho_{ENV} = 0.1$ in *C* represents the ENV noise floor; the TFS noise floor was negligible ($\rho_{TFS} = 0.01$).

by NBTFS speech. The performance with NBTFS speech was significantly worse ($p < 0.05$) than all other vocoded speech. Despite quantitative differences, the finding that NBTFS speech was less intelligible than BBTFS is consistent with previous studies (Smith et al., 2002; Gilbert and Lorenzi, 2006). The overall performance reported for speech reception of sentences in quiet with modest training (~70% for BBTFS, ~0% for NBTFS) was much lower than the present results, likely due to the increased difficulty of an open-set task (Smith et al., 2002). Performance for closed-set identification of French VCVs in quiet with significant training (~98% for BBTFS, ~90% for NBTFS) was better than the present results, particularly for the NBTFS condition (Gilbert and Lorenzi, 2006). Previous studies suggest significant recovered envelopes remain for the NBTFS condition, despite being reduced relative to BBTFS (Sheft et al., 2008; Heinz and Swaminathan, 2009). Thus, it is not clear whether the improved performance for NBTFS speech with extensive training observed in other studies is due to listeners learning to use TFS cues (as is often assumed) or from learning to use the small but significant recovered ENV cues. The fact that the neural coding of TFS for NBTFS speech is stronger than BBTFS (Heinz and Swaminathan, 2009) and that this trend is opposite from the consistent perceptual trend (Fig. 2; also see Smith et al., 2002; Gilbert and Lorenzi, 2006) questions whether the training effect is based on TFS cues alone.

Identification scores were compared across stimuli at positive and negative SNRs to determine whether the relative contributions of ENV and TFS cues differed between favorable and degraded conditions. At positive SNRs (Q, 10, 5, 0 dB) intact speech was the most intelligible followed by PHENV speech, which was slightly (but significantly) better than BBTFS speech, followed by PDENV speech and finally NBTFS speech. A two-way ANOVA on positive SNRs showed significant effects of stimulus type [$F_{(4,99)} = 975.48$, $p < 0.0001$] and SNR [$F_{(3,99)} = 385.41$, $p < 0.0001$], as well as a significant interaction [$F_{(12,99)} = 14.95$, $p < 0.0001$]. *Post hoc* Tukey–Kramer adjustments for pairwise comparisons between stimulus types revealed that the order of decreasing performance (INTACT, PHENV, BBTFS, PDENV, to NBTFS) represented all significant differences ($p < 0.05$). In contrast, at negative SNRs (−5, −10, −15, −20 dB) intact speech was the most intelligible followed by TFS speech (BBTFS comparable to NBTFS speech) and then ENV speech (PHENV comparable to PDENV speech). For negative SNRs, there were significant effects of stimulus type [$F_{(4,99)} = 110.28$, $p < 0.0001$] and SNR [$F_{(3,99)} = 218.26$, $p < 0.0001$], as well as a significant interaction [$F_{(12,99)} = 18.06$, $p < 0.0001$]. *Post hoc* comparisons revealed that intact speech was significantly better than all other types, followed by TFS speech (BBTFS not significantly different from NBTFS), which was significantly better than ENV speech (PHENV not significantly different from PDENV). These psychoacoustic results suggest that the relative salience of ENV and TFS cues differs for favorable and degraded conditions, consistent with the general consensus from perceptual studies (Shannon et al., 1995; Qin and Oxenham, 2003; Zeng et al., 2005).

Neural coding of noise-degraded speech

The effects of noise on the neural coding of TFS for intact and vocoded speech are shown in Figure 2*B*. Mean TFS coding across CFs was better for intact than for NBTFS speech, which was itself better than BBTFS for SNR ≥ -10 dB. Not surprisingly, there was no TFS coding for PHENV or PDENV vocoders. These trends in TFS coding across stimulus types are inconsistent with the VCV identification scores (Fig. 2*A*), in particular the reversal in relative salience of NBTFS and BBTFS speech.

Figure 2*C* shows mean values of ρ_{ENV} for intact and vocoded speech as a function of SNR. As expected, neural coding of ENV was strongest (and similar) for intact and PHENV speech. However, phonemic ENV coding was also present for the three vocoder versions in which acoustic phonemic ENV was absent (i.e., "recovered" ENV coding). Consistent with results from previous perceptual modeling (Zeng et al., 2004; Gilbert and Lorenzi, 2006; Sheft et al., 2008) and neurophysiological (Heinz and Swaminathan, 2009) studies in quiet, recovered ENVs from BBTFS speech were stronger than NBTFS speech for positive SNRs; however, recovered ENVs from NBTFS and BBTFS were similar for negative SNRs. It is interesting that, in contrast to TFS coding, these trends in the relative salience of neural recovered ENVs for NBTFS and BBTFS are consistent at both positive and negative SNRs with the identification scores for TFS speech (Fig. 2*A*). In addition to recovered ENVs from TFS speech, there were recovered phonemic ENVs from PDENV speech. This result suggests that higher-rate acoustic ENV cues can be recovered as lower-rate neural phonemic ENV coding, which may have contributed to the perception of the PDENV speech that lacked acoustic envelope modulations below 64 Hz (Fig. 2*A*).

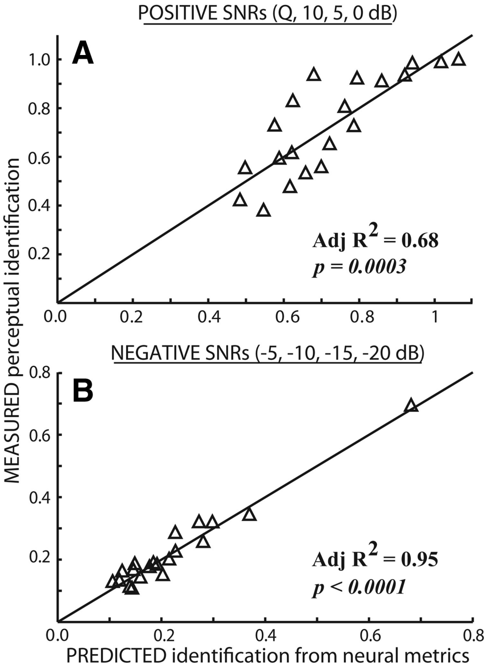

Figure 3. Neural coding of phonemic envelope and fine structure predicts VCV identification scores at both positive (***A***) and negative (***B***) SNRs. Consonant identification scores for intact and vocoded speech are plotted against predicted identification scores from linear regression models based on ρ_{ENV} and ρ_{TFS}. Data from all five stimulus types (i.e., intact, BBTFS, NBTFS, PHENV, and PDENV) are included for all SNRs. The Adj R^2 and p value are inset for each panel (regression coefficients are shown in Table 2). The diagonal line represents a perfect one-to-one match between measured and predicted identification scores.

Quantitative correlations between neural coding and perception of noise-degraded speech

A regression model based only on neural ENV and TFS coding was successful in predicting consonant identification scores for both positive and negative SNRs. Figure 3 shows measured identification scores for intact and vocoded speech conditions plotted against identification scores predicted from the neural metrics ρ_{ENV} and ρ_{TFS}. Predicted identification scores were obtained from the linear regression model

$$\mathrm{PC} = b0 + b1\rho_{\mathrm{ENV}} + b2\rho_{\mathrm{TFS}} + b3\rho_{\mathrm{ENV}}\rho_{\mathrm{TFS}}, \qquad (3)$$

where *PC* is the mean percentage correct consonant identification score (from Fig. 2*A*) normalized to a range between 0 and 1, and ρ_{TFS} and ρ_{ENV} are the mean values from Figure 2*B*,*C*. To ensure accurate evaluation of TFS contributions when ENV cues were very small (i.e., at very low SNRs), the neural noise floor was subtracted from the cross-correlation coefficients before fitting the regression model. The statistical significance of the coefficients *b1*, *b2*, and *b3* provides insight into the relative contributions of neural ENV, TFS, and their interaction, respectively, to speech intelligibility.

The overall model adjusted R^2 (Adj R^2)for predicted consonant identification at positive SNRs was 0.68 (Fig. 3*A*), which was highly significant ($p = 0.0003$). The regression coefficient corresponding to ρ_{ENV} was found to be mainly significant ($p = 0.0032$), with ρ_{TFS} just marginally significant ($p = 0.0317$) (Table 2). This result suggests that at positive SNRs, neural coding of ENV (both true and recovered) is the most significant contributor to consonant identification, consistent with previous psychoacoustic studies demonstrating that ENV cues alone are sufficient for speech perception in quiet listening conditions (e.g., Shannon et al., 1995). The finding that neural TFS cues were only marginally significant contrasts with recent psychoacoustic studies that have suggested acoustic TFS alone can support speech intelligibility in quiet (Gilbert and Lorenzi, 2006; Gilbert et al., 2007; Sheft et al., 2008); however, these apparently contrasting results can be reconciled by the fact that at positive SNRs, acoustic TFS can contribute as neural recovered ENV (Fig. 2C).

At negative SNRs (Fig. 3*B*), the overall fit of consonant identification from neural ENV and TFS coding was excellent (Adj $R^2 = 0.95$, $p < 0.0001$). In contrast to positive SNRs, ρ_{ENV}, ρ_{TFS}, and the interaction between ρ_{ENV} and ρ_{TFS} were all found to be significant, although to varying degrees (Table 2). Consistent with commonly held beliefs, neural TFS was found to be significant ($p = 0.0011$) for degraded SNRs. In fact, the interaction between neural TFS and ENV was found to be highly significant ($p = 0.0004$), suggesting that the contribution of TFS is even greater in the presence of ENV cues. However, neural ENV alone was found to be the most significant contributor ($p < 0.0001$) to speech intelligibility at negative SNRs. These findings suggest that the perceptual salience of neural ENV cues for speech perception in degraded listening conditions is equal to or even greater than that of neural TFS cues.

Reception of phonetic features in noise

Figure 4 shows mean reception of voicing, manner, place, and nasality as a function of SNR for intact and vocoded speech averaged across all listeners. As with identification scores, the reception of phonetic features was compared across stimuli at positive and negative SNRs to identify whether the relative contributions of ENV and TFS cues differed between favorable and degraded conditions. For each of the features, separate two-way ANOVAs on positive and negative SNRs showed significant effects for factor stimulus type ($p < 0.0001$) and SNR ($p < 0.0001$), as well as significant interactions between these two factors ($p < 0.01$). For the reception of voicing (Fig. 4*A*), *post hoc* comparisons between stimuli revealed that the reception of voicing was greatest for intact speech, followed by PHENV and BBTFS speech (not significantly different from one another), followed by NBTFS and PDENV speech (not significantly different from one another). Voicing reception at negative SNRs showed a different pattern across TFS and ENV speech than for positive SNRs (particularly for SNR ≤ -10 dB). Voicing reception was greatest for intact speech, followed by TFS speech (BBTFS not significantly different from NBTFS speech) and then ENV speech (PHENV not significantly different from PDENV speech). Thus, these psychoacoustic results suggest that the relative importance of TFS cues (compared to ENV) for voicing is greater at negative than at positive SNRs. For the reception of manner (Fig. 4B), *post hoc* comparisons between stimuli showed that information transmitted for manner was greatest for intact speech, followed by PHENV speech, which was better than BBTFS speech, followed by PDENV speech, and finally NBTFS speech (all significant differences, $p < 0.05$). In contrast to voicing, manner reception for negative SNRs was only different for intact speech, with no significant differences for all four vocoded speech versions, which were all extremely low for most of the degraded conditions. The reception of place (Fig. 4*C*) at positive SNRs was greatest for intact speech, followed by PHENV and BBTFS speech (not significantly different from one another), followed by PDENV speech and finally NBTFS speech. Similar to manner, the reception of place for negative SNRs was greatest for intact speech, but was not significantly different for all other vocoded speech (although it appears that both versions of TFS speech were slightly above both ENV speech versions). The reception of nasality (Fig. 4*D*) at positive SNRs was higher for intact

Table 2. Predictive models of VCV identification and phonetic feature reception based on neural coding of ENV and TFS

	Percent Correct	Voicing	Manner	Place	Nasality
Positive SNRs					
ENV	***0.87*** (0.0032)	***1.19*** (0.0046)	***1.21*** (0.0018)	***1.07*** (0.0020)	***1.18*** (0.0042)
TFS	0.66 (0.0317)	1.14 (0.0120)	0.77 (0.0493)	0.85 (0.0194)	***1.81*** (0.0003)
E×T	−0.61 (0.1634)	−1.02 (0.1100)	−0.63 (0.2614)	−0.70 (0.1691)	***−1.73*** (0.0091)
Adj R^2	0.68	0.65	0.75	0.73	0.73
p value	(0.0003)	(0.0006)	(<0.0001)	(0.0001)	(0.0001)
Negative SNRs					
ENV	***0.76*** (<0.0001)	***0.30*** (0.0028)	NF	***0.40*** (0.0042)	0.40 (0.0560)
TFS	***0.33*** (0.0011)	***0.19*** (0.0031)	NF	0.11 (0.1705)	***0.82*** (<0.0001)
E×T	***2.65*** (0.0004)	***1.48*** (0.0017)	NF	***3.84*** (<0.0001)	***3.24*** (0.0023)
Adj R^2	0.95	0.92	NF	0.95	0.94
p value	(<0.0001)	(<0.0001)		(<0.0001)	(<0.0001)

Coefficients from the regression models (i.e., *b1*, *b2*, and *b3* in Eq. 3, corresponding to ρ_{ENV}, ρ_{TFS}, and their interaction, E×T) are shown with their *p* values in parentheses, along with the overall model goodness of fit (Adj R^2) for overall percentage correct (Fig. 3) and the reception of individual phonetic features (voicing, manner, place, and nasality; Fig. 6). Statistically significant coefficients ($p < 0.01$) are underlined and bold; marginally significant coefficients ($0.01 \leq p < 0.05$) are underlined. Top and bottom sections represent positive (Q, 10, 5, 0 dB) and negative (−5, −10, −15, −20 dB) SNRs, respectively. NF, Not fit (e.g., reception of manner was too poor to fit for negative SNRs).

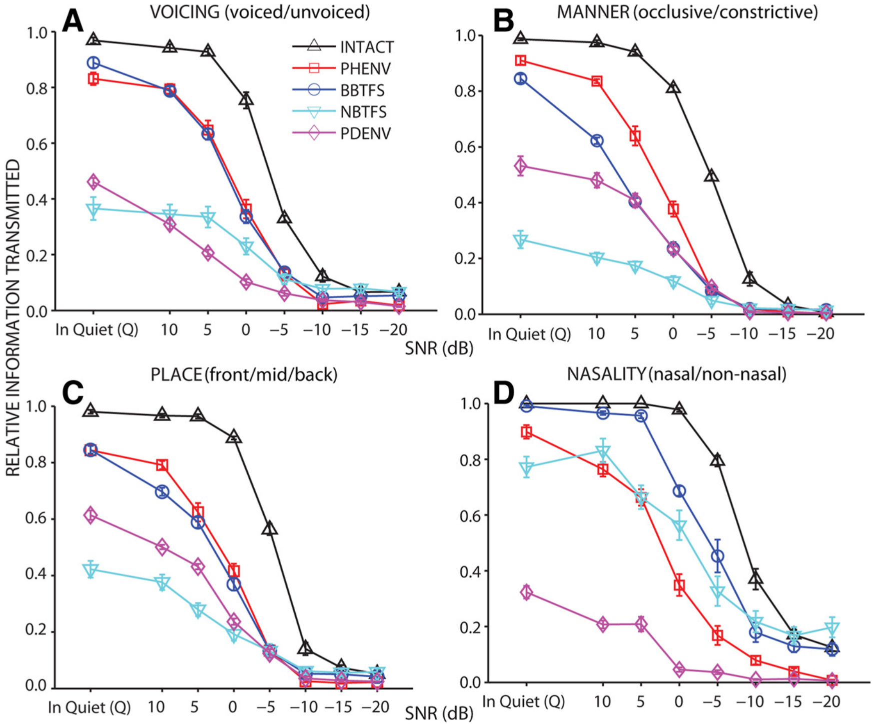

Figure 4. The effects of noise on specific reception of phonetic features for intact and vocoded speech. Mean reception of voicing (***A***), manner (***B***), place (***C***), and nasality (***D***) is plotted (with SEM bars across listeners) as a function of signal-to-noise ratio. Reception of phonetic features was measured in terms of relative information transmitted (i.e., 1.0 represents perfect reception).

speech, followed by BBTFS speech, which was better than NBTFS and PHENV speech (not significantly different from one another), followed by PDENV speech. The reception of nasality for negative SNRs was greatest for intact speech, followed by TFS speech (BBTFS and NBTFS speech were not significantly different), which was better than PHENV speech, followed finally by PDENV speech.

Figure 5 compares the reception of voicing, manner, place, and nasality for intact, ENV, and TFS speech in quiet and at a degraded listening condition (−10 dB SNR). The comparisons that were used to delineate the relative contributions of ENV and TFS cues to phonetic feature reception based on the present psychoacoustic data were primarily between PHENV and NBTFS speech, similar to previous studies (e.g., Gilbert and Lorenzi, 2006; Sheft et al., 2008) without factoring in the neural coding of ENV and TFS.

For the in-quiet condition, listeners showed good (better than 80%) reception of all four phonetic features (voicing, manner, place, and nasality) when using mainly ENV cues (i.e., for PHENV speech); the reception of manner and nasality was slightly better than voicing and place in these conditions. Shannon et al. (1995) used ENV vocoded speech to suggest that nearly perfect reception of manner and voicing (but poor reception of place) are obtained when using mainly temporal ENV cues. However, their results were based on ENV vocoders created using a maximum of four analysis bands with impoverished spectral information, which may have led to poor reception of place. Studies with 16-band ENV vocoders have shown good reception of place with ENV cues (Friesen et al., 2001; Başkent, 2006; Sheft et al., 2008), consistent with the present results from the PHENV speech that was generated with 16 analysis bands. In contrast with PHENV speech in quiet, the reception of voicing, manner, and place was much poorer (<50%) for NBTFS speech (for which TFS cues are typically assumed to be the primary cues; Gilbert and Lorenzi, 2006; Gilbert et al., 2007; Sheft et al., 2008).Together, these perceptual results suggest that the high reception of voicing, manner, and place from intact speech in quiet is largely due to contributions of ENV cues with less contribution from TFS cues. Nasality was shown to be well transmitted for both PHENV and NBTFS speech, with reception being slightly better for PHENV speech. This result is consistent with previous studies (Sheft et al., 2008) that showed good reception of nasality for both ENV and TFS speech.

The relative pattern of phonetic feature reception across TFS and ENV speech differed for positive and negative SNRs, with the transition often occurring at −5 dB SNR (Fig. 4). Thus, Figure 5 compares feature reception across stimulus types at −10 dB SNR, which is representative of the degraded listening condition beyond this transition. The reception of voicing and place at −10

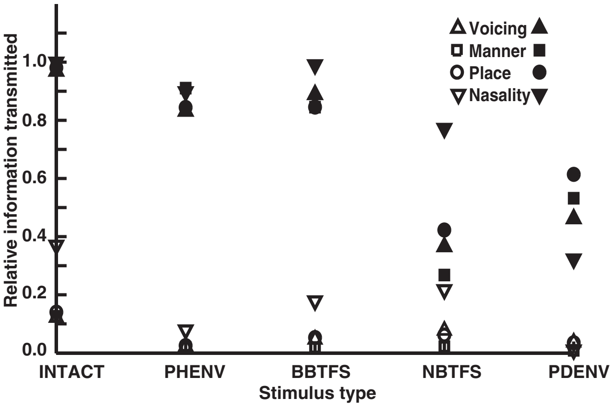

Figure 5. Specific reception of voicing, place, manner, and nasality in quiet (filled black symbols) and degraded listening condition (−10 dB SNR; open red symbols) for intact, 16-band phonemic envelope speech (PHENV), broadband TFS speech (BBTFS), narrowband TFS speech (NBTFS), and periodicity envelope speech (PDENV).

dB SNR was better for NBTFS speech than for PHENV speech, with essentially no reception of these features for PHENV speech. The reception of manner was absent for both PHENV and NBTFS speech. In contrast, the reception of nasality was highest among all features for both NBTFS and PHENV speech and was better for NBTFS speech than for PHENV speech. These psychoacoustic results suggest that the reception of voicing, place, and nasality in degraded listening conditions is primarily due to the contribution of TFS cues, with minimal if any contribution from ENV cues.

Quantitative correlations between neural coding and reception of phonetic features in noise

The regression model (Eq. 3) also successfully predicted the reception of individual phonetic features (e.g., voicing, manner, place, and nasality; see Fig. 6) based only on the neural metrics ρ_{ENV} and ρ_{TFS} (Fig. 2*B*,*C*). For the reception of voicing, manner, and place at positive SNRs (Fig. 6*A*–*C*), ρ_{ENV} was found to be mainly significant, with ρ_{TFS} only marginally significant (Table 2). These results suggest that the reception of these three phonetic features at positive SNRs depends most significantly on the neural coding of ENV alone and much less so on neural TFS alone. This finding is consistent with previous psychoacoustic studies suggesting that listeners can use ENV cues for accurate reception of manner (Shannon et al., 1995; Gilbert and Lorenzi, 2006), place [when ENV speech is created with a large number of analysis bands (Fu et al., 1998; Sheft et al., 2008)], and voicing (Shannon et al., 1995). However, the present result that neural ENV is the most significant contributor to the reception of place and voicing is inconsistent with suggestions based on TFS speech that listeners can use TFS alone to achieve accurate reception of place and voicing in quiet (Gilbert and Lorenzi, 2006). In contrast to voicing, manner, and place, the reception of nasality at positive SNRs (Fig. 6*D*) was best fit by a regression model with significant coefficients for ρ_{ENV}, ρ_{TFS}, and their interaction, E×T, although only the TFS term was highly significant ($p < 0.001$, Table 2). These results suggest that the reception of nasality at positive SNRs is primarily determined by the neural coding of TFS (and to a lesser degree by its interaction with ENV or by ENV alone). This finding is consistent with previous results suggesting listeners can use TFS cues for nasality reception in quiet; however, it was also suggested that ENV cues can be used (Sheft et al., 2008). The present results suggest that neural ENV can contribute to nasality at positive SNRs, but mainly as a secondary cue to neural TFS.

Regression models fit to the data for negative SNRs were quite accurate (Fig. 6*E*–*G*, Table 2), and suggest that the relative salience of ENV and TFS cues for phonetic feature reception was different between favorable and degraded listening conditions. For voicing (Fig. 6*E*), ρ_{ENV}, ρ_{TFS}, and their interaction, E×T, were all found to be significant (Table 2). This suggests that the reception of voicing at degraded SNRs benefits from both neural ENV and TFS alone, but also from having neural ENV and TFS together (as indicated by the positive and significant coefficient for the E×T term). At negative SNRs, the reception of manner was near the noise floor for most conditions (Fig. 4*B*) and hence was not fit with the regression model. For place at negative SNRs (Fig. 6*F*), both ρ_{ENV} and the interaction between ρ_{ENV} and ρ_{TFS} were found to be significant, with the E×T term being extremely significant (Table 2). This result suggests that the reception of place in degraded listening conditions is most dependent on the interaction between ENV and TFS (i.e., both cues being available together), but that neural ENV alone is also a significant contributor; TFS alone was not a significant contributor. Similar to positive SNRs, the regression results (Fig. 6*G*, Table 2) suggest that the reception of nasality at negative SNRs is primarily determined by the neural coding of TFS alone; the neural coding of ENV does contribute at negative SNRs, but only through an interaction with neural TFS and not by itself.

Integrating the present data with previous studies, it is clear that the apparent contributions of ENV and TFS to phonetic feature reception differ between acoustical, psychoacoustical, and psychophysiological analyses based on linking neural coding directly to perceptual results with identical stimuli. In a seminal study describing the acoustic structure of speech based on temporal information, Rosen (1992) suggested that ENV provided mainly phonetic cues to manner, whereas TFS contributed primarily to place and, to a smaller extent, voicing. Subsequent detailed psychoacoustic studies with ENV and TFS vocoded speech in quiet (e.g., Shannon et al., 1995; Sheft et al., 2008; Ardoint and Lorenzi, 2010) were largely consistent (with a few discrepancies) with Rosen's acoustic analyses, suggesting important and complementary contributions for both ENV and TFS for speech perception under favorable listening conditions. Previous psychoacoustical support for the importance of TFS for listening in noise has been mainly limited to studies comparing the perception of intact and ENV speech in noise (e.g., Qin and Oxenham, 2003). Psychoacoustic analyses of the present vocoder data, extended to include TFS speech in noise, also suggested that the reception of all phonetic features in degraded listening conditions is primarily due to the contribution of TFS cues, with minimal contribution from ENV cues (Figs. 4 and 5). In contrast, the present psychophysiological analyses (Fig. 6, Table 2) demonstrate that consideration of cochlear signal processing suggests the following: (1) a primary contribution of ENV coding (as both true and recovered ENV) for all features (except nasality) at positive SNRs; (2) an equal or greater contribution of neural ENV alone compared to neural TFS alone at negative SNRs for all features (except nasality); (3) a significant interaction between neural ENV and TFS at negative SNRs suggesting that the contribution of neural TFS is greater in the presence of neural ENV; and (4) a primary role of neural TFS coding alone for nasality.

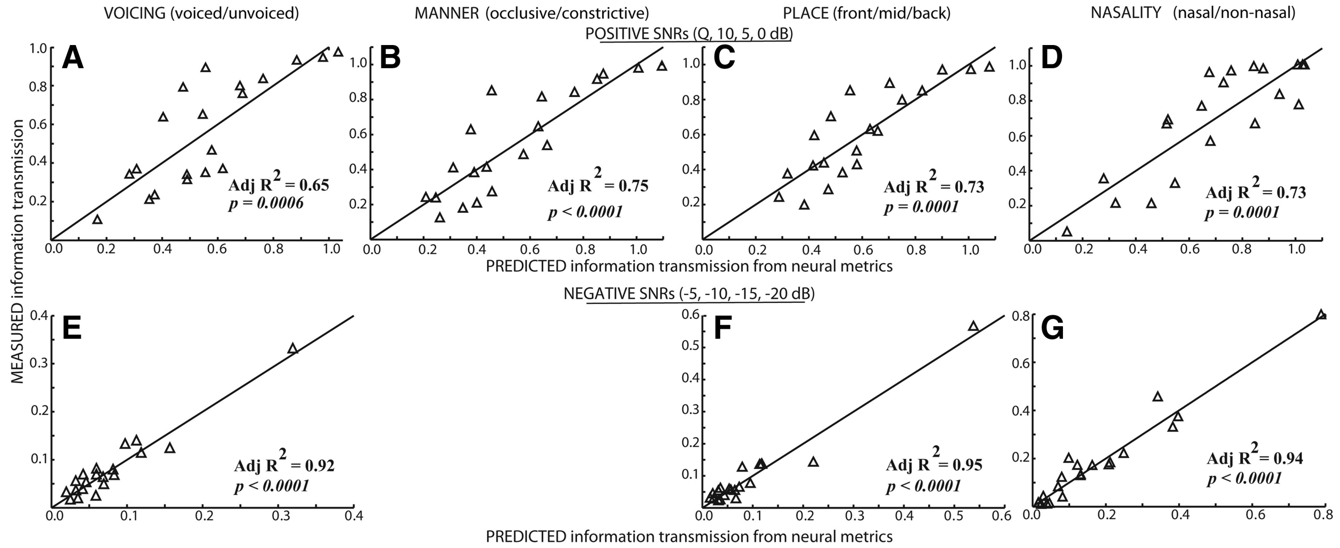

Figure 6. Neural coding of phonemic envelope and fine structure predicts (Eq. 3, Table 2) measured reception of phonetic features at positive (***A–D***) and negative (***E–G***) SNRs. Measured information transmission from psychoacoustical data (from Fig. 4) is plotted against predicted information transmission from neural coding of phonemic envelope and fine structure (ρ_{ENV} and ρ_{TFS} from Fig. 2 *B*, *C*) for voicing, manner, place, and nasality. At negative SNRs the reception of manner was negligible for most of the vocoder conditions, and thus the regression model was not fit. The adjusted R^2 and *p* value of the regression models are inset for each panel. The diagonal lines represent a perfect one-to-one match between measured and predicted reception scores.

Discussion

Implications for robust speech perception

Evidence for the importance of TFS and ENV cues for speech perception in quiet (Shannon et al., 1995; Gilbert and Lorenzi, 2006; Sheft et al., 2008) and in noise (Qin and Oxenham, 2003; Nie et al., 2005; Zeng et al., 2005; Hopkins and Moore, 2009) has come primarily from psychoacoustic studies. Thus, these perceptual results should be interpreted mainly in terms of the role of acoustic ENV and TFS cues, and with caution in translating these results to conclusions about neural ENV and TFS. For complex stimuli, neural ENV and TFS at the output of the cochlea may differ from acoustic ENV and TFS at the input to the ear (Zwicker, 1962; Saberi and Hafter, 1995; Ghitza, 2001). For example, acoustic TFS can produce useful temporal coding in the AN in two ways: (1) responses that are synchronized to the stimulus fine structure itself (i.e., "true TFS"); and (2) responses that are synchronized to stimulus-related ENV (i.e., "recovered envelopes") (see Heinz and Swaminathan, 2009, their Fig. 10). Hence, interpretation of the perceptual salience of acoustic ENV and TFS must carefully consider the physiological transformations that these cues can undergo through cochlear filtering.

A more complete understanding of the relative roles of ENV and TFS for speech perception can be achieved only by taking a synergistic approach of quantifying the salience of ENV and TFS in both the perceptual and neural domains, as employed here. Relating neural coding to measured speech identification demonstrated that: (1) neural ENV (comprising both acoustic ENV coded as neural ENV and acoustic TFS coded as recovered neural ENV) is a primary cue for speech perception, even in degraded listening conditions; and (2) neural TFS does contribute in degraded listening conditions (less by itself and more through an interaction with ENV), but never as the primary cue (except for nasality). The present finding that the perceptual salience of neural TFS is less than previously thought arises due to consideration of cochlear signal processing that transforms TFS and ENV coding in normal hearing ears and is consistent with the general conclusions of several recent psychoacoustic studies (Oxenham and Simonson, 2009; Bernstein and Brungart, 2011).

An important consideration for the present conclusions is that they are based on relating speech perception measured from human listeners to predicted neural coding from a computational AN model that was validated against physiological data from animal studies. One issue that could affect the present neural predictions is the potential difference in the degree of ENV recovery between humans and cats. Based on suggestions that humans have sharper tuning than cats (Shera et al., 2002, 2010; Joris et al., 2011; however, also see Ruggero and Temchin, 2005), it has been predicted that ENV recovery is greater in humans than in cats (Ibrahim and Bruce, 2010). Thus, the present results may actually provide a conservative estimate of the contribution of ENV recovery to speech perception in noise.

Implications for hearing aids and cochlear implants

Any perceptual benefit of acoustic TFS that arises from recovered ENVs in normal hearing listeners (sharp cochlear tuning) may not be restored with auditory prostheses designed to enhance TFS coding in listeners with sensorineural hearing loss (SNHL; broadened cochlear tuning) or in CI listeners (no cochlear tuning) because their ability to recover ENV cues is severely degraded or completely absent. The theoretical framework used in the present study (also see Heinz and Swaminathan, 2009) suggests that it may be possible to overcome this apparent limitation by focusing hearing aid and CI signal processing strategies on the restoration of neural ENV cues, rather than acoustic TFS cues.

Listeners with SNHL have particular difficulty understanding speech in noise and have been shown to have a correlated deficit in their ability to use acoustic TFS cues (Lorenzi et al., 2006). The straightforward interpretation of this perceptual TFS deficit is that it arises due to a degradation in the fundamental ability of AN fibers to phase lock to TFS; however, neurophysiological recordings have demonstrated that the strength of phase locking is not degraded following noise-induced hearing loss (Kale and Heinz, 2010). Thus, efforts to develop hearing aids to enhance the neural coding of TFS directly would appear to be misguided. Although other factors may contribute [e.g., degraded spatiotemporal cues (Heinz et al., 2010; Kale, 2011)], the present data suggest that the loss of recovered ENVs could also contribute to the reduced ability to use acoustic TFS cues following SNHL and is expected to occur due to the broadened cochlear tuning and loss of tonotopicity often observed in impaired AN responses (Liberman and Dodds, 1984; Miller et al., 1997; Heinz and Swaminathan, 2009; Kale, 2011). The fact that both perceptual and neurophysiological studies suggest that envelope coding in quiet is not degraded following SNHL (Lorenzi et al., 2006; Kale and Heinz, 2010) provides promise for hearing aid strategies focused on restoring neural envelopes, rather than enhancing acoustic TFS, to improve speech perception in noise.

The implications of the present results for CIs can be understood by considering each of the regression model terms (Eq. 3) with respect to current CI technology. The present finding that neural ENV was a primary contributor to speech perception in both favorable and degraded listening conditions is extremely promising, because CIs are very successful at providing ENV information directly to the AN; however, an important distinction is that CIs currently provide acoustic ENV rather than neural ENV. The present results suggest that CI speech perception in noise could be improved by incorporating a physiologically based AN model that captures the non-linearities associated with healthy cochlear signal processing (e.g., Zilany and Bruce, 2007) as a front end in CI speech processors so that acoustic ENV coded as neural ENV and acoustic TFS recoded as (recovered) neural ENV can be transmitted. The present finding that neural TFS alone was rarely the primary cue for speech perception, even in degraded listening conditions, is promising in the sense that current CI technology is unable to provide neural TFS. Furthermore, if neural TFS alone could someday be provided, the present results suggest that speech perception would likely be no better than that for a CI that provided neural ENV alone, at least in steady-state noise. A key empirical question would then be whether neural ENV alone is good enough, given the vast redundancies of speech and the fact that robust speech perception in noise likely also depends on across-channel ENV coding, not only within-channel ENV coding (Crouzet and Ainsworth, 2001; Swaminathan and Heinz, 2011). Of course, it is likely that some improvements in speech perception in noise (and in sound localization) could be achieved by adding TFS. However, the highly significant interaction term (E×T) for overall speech perception in noisy conditions (Table 2) implies that if TFS is able to be provided in future technology (e.g., Middlebrooks and Snyder, 2010), an important design constraint must be that TFS be provided in a way that does not disrupt neural ENV coding.

Implications for neural coding: are recovered envelopes a vocoder artifact or robust biological signal processing?

Recovered ENV from acoustic TFS has mainly been interpreted as an outcome of signal processing "artifacts" from vocoder implementations (Zeng et al., 2004; Gilbert and Lorenzi, 2006; Sheft et al., 2008). However, it is possible that the ability to extract ENV from acoustic TFS in the speech waveform represents a useful form of robust biological signal processing (e.g., especially for noise-degraded speech) that is available to normal hearing listeners with narrow cochlear filters, but not to listeners with SNHL or CI patients. It has been well documented (theoretically and perceptually) that FM provides an SNR improvement over AM (e.g., Crosby, 1937; Zeng et al., 2005). It could be that the auditory system is designed around this principle such that TFS is able to faithfully carry/transmit neural ENV by aiding in its recovery at peripheral and/or higher auditory stages when true ENV is degraded by noise. Moreover, it has been shown that ENV coding is enhanced at central stages compared to the periphery (Joris, 2003; Agapiou and McAlpine, 2008), suggesting the possibility of a central hierarchy of ENV recovery from TFS starting at the cochlea.

References

Agapiou JP, McAlpine D (2008) Low-frequency envelope sensitivity produces asymmetric binaural tuning curves. J Neurophysiol 100:2381–2396.

Ardoint M, Lorenzi C (2010) Effects of lowpass and highpass filtering on the intelligibility of speech based on temporal fine structure or envelope cues. Hear Res 260:89–95.

Başkent D (2006) Speech recognition in normal hearing and sensorineural hearing loss as a function of the number of spectral channels. J Acoust Soc Am 120:2908–2925.

Bernstein JG, Brungart DS (2011) Effects of spectral smearing and temporal fine-structure distortion on the fluctuating-masker benefit for speech at a fixed signal-to-noise ratio. J Acoust Soc Am 130:473–488.

Crosby GM (1937) Frequency modulation noise characteristics. Proc Inst Radio Eng 25:472–514.

Crouzet O, Ainsworth WA (2001) On the various influences of envelope information on the perception of speech in adverse conditions: an analysis of between-channel envelope correlation. Paper presented at One-Day Workshop on Consistent and Reliable Cues for Sound Analysis, Aalborg, Denmark. September.

Dean I, Harper NS, McAlpine D (2005) Neural population coding of sound level adapts to stimulus statistics. Nat Neurosci 8:1684–1689.

Drennan WR, Rubinstein JT (2008) Music perception in cochlear implant users and its relationship with psychophysical capabilities. J Rehabil Res Dev 45:779–790.

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

Dudley HW (1939) The Vocoder. Bell Labs Rec 18:122–126.

Flanagan JL (1980) Parametrics coding of speech spectra. J Acoust Soc Am 68:412–430.

Flanagan JL, Golden RM (1966) Phase Vocoder. Bell Syst Tech J 45:1493–1509.

Friesen LM, Shannon RV, Baskent D, Wang X (2001) Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. J Acoust Soc Am 110:1150–1163.

Fu QJ, Shannon RV (1999) Phoneme recognition by cochlear implant users as a function of signal-to-noise ratio and nonlinear amplitude mapping. J Acoust Soc Am 106:L18–L23.

Fu QJ, Shannon RV, Wang X (1998) Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing. J Acoust Soc Am 104:3586–3696.

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.

Gilbert G, Lorenzi C (2006) The ability of listeners to use recovered envelope cues from speech fine structure. J Acoust Soc Am 119:2438–2444.

Gilbert G, Bergeras I, Voillery D, Lorenzi C (2007) Effects of periodic interruptions on the intelligibility of speech based on temporal fine-structure or envelope cues. J Acoust Soc Am 122:1336.

Heinz MG (2010) Computational modeling of sensorineural hearing loss. In: Computational models of the auditory system (Meddis R, Lopez-Poveda EA, Popper AN, Fay RR, eds), pp 177–202. New York: Springer.

Heinz MG, Swaminathan J (2009) Quantifying envelope and fine-structure coding in auditory nerve responses to chimaeric speech. J Assoc Res Otolaryngol 10:407–423.

Heinz MG, Swaminathan J, Boley JD, Kale S (2010) Across-fiber coding of temporal fine-structure: Effects of noise-induced hearing loss on auditory-nerve responses. In: The neurophysiological bases of auditory perception (Lopez-Poveda EA, Palmer AR, Meddis R, eds), pp 621–630. New York: Springer.

Hilbert D (1912) Grundzüge einer allgemeinen theorie der linearen integralgleichungen. Leipzig: B. G. Teubner.

Hopkins K, Moore BC (2009) The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise. J Acoust Soc Am 125:442–446.

Ibrahim RA, Bruce IC (2010) Effects of peripheral tuning on the auditory nerve's representation of speech envelope and temporal fine structure cues. In: The neurophysiological bases of auditory perception (Lopez-Poveda EA, Palmer AR, Meddis R, eds), pp 429–438. New York: Springer.

Johnson DH (1980) The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones. J Acoust Soc Am 68:1115–1122.

Joris PX (2003) Interaural time sensitivity dominated by cochlea-induced envelope patterns. J Neurosci 23:6345–6350.

Joris PX, Yin TC (1992) Responses to amplitude-modulated tones in the auditory nerve of the cat. J Acoust Soc Am 91:215–232.

Joris PX, Louage DH, Cardoen L, van der Heijden M (2006) Correlation index: a new metric to quantify temporal coding. Hear Res 216-217:19–30.

Joris PX, Bergevin C, Kalluri R, Mc Laughlin M, Michelet P, van der Heijden M, Shera CA (2011) Frequency selectivity in Old-World monkeys corroborates sharp cochlear tuning in humans. Proc Natl Acad Sci U S A 108:17516–17520.

Kale S (2011) Temporal coding in auditory-nerve fibers following noise-induced hearing loss. Unpublished PhD dissertation, Purdue University.

Kale S, Heinz MG (2010) Envelope coding in auditory nerve fibers following noise-induced hearing loss. J Assoc Res Otolaryngol 11:657–673.

Kawase T, Delgutte B, Liberman MC (1993) Antimasking effects of the olivocochlear reflex. II. Enhancement of auditory-nerve response to masked tones. J Neurophysiol 70:2533–2549.

Liberman MC (1978) Auditory-nerve response from cats raised in a low-noise chamber. J Acoust Soc Am 63:442–455.

Liberman MC, Dodds LW (1984) Single-neuron labeling and chronic cochlear pathology. III. Stereocilia damage and alterations of threshold tuning curves. Hear Res 16:55–74.

Litvak L, Delgutte B, Eddington D (2003) Improved neural representation of vowels in electric stimulation using desynchronizing pulse trains. J Acoust Soc Am 114:2099–2111.

Lorenzi C, Gilbert G, Carn H, Garnier S, Moore BC (2006) Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. Proc Natl Acad Sci U S A 103:18866–18869.

Louage DH, van der Heijden M, Joris PX (2004) Temporal properties of responses to broadband noise in the auditory nerve. J Neurophysiol 91:2051–2065.

Middlebrooks JC, Snyder RL (2010) Selective electrical stimulation of the auditory nerve activates a pathway specialized for high temporal acuity. J Neurosci 30:1937–1946.

Miller GA, Nicely PE (1955) An analysis of perceptual confusions among some English consonants. J Acoust Soc Am 27:338–352.

Miller RL, Schilling JR, Franck KR, Young ED (1997) Effects of acoustic trauma on the representation of the vowel /ε/ in cat auditory nerve fibers. J Acoust Soc Am 101:3602–3616.

Nie K, Stickney G, Zeng FG (2005) Encoding frequency modulation to improve cochlear implant performance in noise. IEEE Trans Biomed Eng 52:64–73.

Oxenham AJ, Simonson AM (2009) Masking release for low- and high-pass filtered speech in the presence of noise and single-talker interference. J Acoust Soc Am 125:457–468.

Palmer AR (1982) Encoding of rapid amplitude fluctuations by cochlear-nerve fibres in the guinea-pig. Arch Otorhinolaryngol 236(2):197–202.

Qin MK, Oxenham AJ (2003) Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. J Acoust Soc Am 114:446–454.

Rosen S (1992) Temporal information in speech: acoustic, auditory, and linguistic aspects. Philos Trans R Soc Lond B Biol Sci 336:367–373.

Rubinstein JT, Wilson BS, Finley CC, Abbas PJ (1999) Pseudospontaneous activity: stochastic independence of auditory nerve fibers with electrical stimulation. Hear Res 127:108–118.

Ruggero MA (1973) Response to noise of auditory nerve fibers in the squirrel monkey. J Neurophysiol 36:569–587.

Ruggero MA, Temchin AN (2005) Unexceptional sharpness of frequency tuning in the human cochlea. Proc Natl Acad Sci U S A 102:18614–18619.

Saberi K, Hafter ER (1995) A common neural code for frequency- and amplitude-modulated sounds. Nature 374:537–539.

Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M (1995) Speech recognition with primarily temporal cues. Science 270:303–304.

Shannon RV, Jensvold A, Padilla M, Robert ME, Wang X (1999) Consonant recordings for speech testing. J Acoust Soc Am 106:L71–74.

Sheft S, Ardoint M, Lorenzi C (2008) Speech identification based on temporal fine structure cues. J Acoust Soc Am 124:562–575.

Shera CA, Guinan JJ Jr, Oxenham AJ (2002) Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements. Proc Natl Acad Sci U S A 99:3318–3323.

Shera CA, Guinan JJ Jr, Oxenham AJ (2010) Otoacoustic estimation of cochlear tuning: validation in the chinchilla. J Assoc Res Otolaryngol 11:343–365.

Smith ZM, Delgutte B, Oxenham AJ (2002) Chimaeric sounds reveal dichotomies in auditory perception. Nature 416:87–90.

Studebaker GA (1985) A “rationalized” arcsine transform. J Speech Hear Res 28:455–462.

Swaminathan J (2010) The role of envelope and temporal fine structure in the perception of noise degraded speech. Unpublished PhD dissertation, Purdue University.

Swaminathan J, Heinz MG (2011) Predicted effects of sensorineural hearing loss on across-fiber envelope coding in the auditory nerve. J Acoust Soc Am 129:4001–4013.

Viemeister NF (1979) Temporal modulation transfer functions based upon modulation thresholds. J Acoust Soc Am 66:1364–1380.

Wen B, Wang GI, Dean I, Delgutte B (2009) Dynamic range adaptation to sound level statistics in the auditory nerve. J Neurosci 29:13797–13808.

Wilson BS, Finley CC, Lawson DT, Wolford RD, Eddington DK, Rabinowitz WM (1991) Better speech recognition with cochlear implants. Nature 352:236–238.

Zeng FG, Nie K, Liu S, Stickney G, Del Rio E, Kong YY, Chen H (2004) On the dichotomy in auditory perception between temporal envelope and fine structure cues. J Acoust Soc Am 116:1351–1354.

Zeng FG, Nie K, Stickney GS, Kong YY, Vongphoe M, Bhargave A, Wei C, Cao K (2005) Speech recognition with amplitude and frequency modulations. Proc Natl Acad Sci U S A 102:2293–2298.

Zhang X, Heinz MG, Bruce IC, Carney LH (2001) A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression. J Acoust Soc Am 109:648–670.

Zilany MS, Bruce IC (2006) Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery. J Acoust Soc Am 120:1446–1466.

Zilany MS, Bruce IC (2007) Representation of the vowel /ε/ in normal and impaired auditory nerve fibers: model predictions of responses in cats. J Acoust Soc Am 122:402–417.

Zwicker E (1962) Direct comparison between the sensations produced by frequency modulation and amplitude modulation. J Acoust Soc Am 34:1425–1430.