The role of recovered envelope cues in the identification of temporal-fine-structure speech for hearing-impaired listeners (L)

Agnès C. Léger,[a] Joseph G. Desloge, Louis D. Braida, and Jayaganesh Swaminathan[b]
Research Laboratory of Electronics, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Room 36-757, Cambridge, Massachusetts 02139

(Received 25 July 2014; revised 14 November 2014; accepted 26 November 2014)

Narrowband speech can be separated into fast temporal cues [temporal fine structure (TFS)], and slow amplitude modulations (envelope). Speech processed to contain only TFS leads to envelope recovery through cochlear filtering, which has been suggested to account for TFS-speech intelligibility for normal-hearing listeners. Hearing-impaired listeners have deficits with TFS-speech identification, but the contribution of recovered-envelope cues to these deficits is unknown. This was assessed for hearing-impaired listeners by measuring identification of disyllables processed to contain TFS or recovered-envelope cues. Hearing-impaired listeners performed worse than normal-hearing listeners, but TFS-speech intelligibility was accounted for by recovered-envelope cues for both groups. © *2015 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4904540]

[DB] Pages: 505–508

I. INTRODUCTION

One way to characterize a speech waveform is as the sum of a number of amplitude-modulated narrow frequency bands (e.g., Flanagan, 1980). Each frequency band's signal can be separated into a rapidly varying carrier [the temporal fine structure (TFS)], and a slowly varying modulation [the temporal envelope (ENV)]. To evaluate the role of TFS cues for speech perception, the Hilbert transform has been used to separate speech into TFS and ENV components for experimental presentation (e.g., Lorenzi *et al.*, 2006; Sheft *et al.*, 2008; Smith *et al.*, 2002). Lorenzi *et al.* (2006) showed that normal-hearing (NH) listeners can achieve high consonant identification scores in quiet when tested with speech processed to remove ENV cues (referred to as TFS speech), which suggests that TFS speech can carry phonetic information. They also showed that listeners with sensorineural hearing loss had reduced ability to identify TFS speech compared with NH listeners. However, both NH and hearing-impaired (HI) listeners achieved similar identification on consonants processed to remove TFS cues (referred to as ENV speech). This result has been interpreted as evidence of a TFS processing deficit in HI listeners. However, recent studies indicating that measurements of TFS-speech intelligibility may not be an accurate indicator of underlying TFS processing ability have suggested alternate interpretations of this result (e.g., Apoux *et al.*, 2013; Swaminathan *et al.*, 2014).

Several perceptual (e.g., Ghitza, 2001; Gilbert and Lorenzi, 2006; Zeng *et al.*, 2004) and neurophysiological (e.g., Heinz and Swaminathan, 2009) studies have shown that when broadband speech is filtered through a set of narrowband filters (akin to filtering in the cochlea), ENV information can be "recovered" from the TFS component. Swaminathan *et al.* (2014), using simulated cochlear filters, assessed the contributions of these recovered ENV (RENV) cues to the intelligibility of TFS speech for NH listeners. They compared the intelligibility of speech processed to contain the TFS information either as TFS (extracted from the Hilbert phase; TFS speech) or as RENV cues (extracted from the Hilbert envelope of the TFS speech filtered into narrow bands; RENV speech). After sufficient exposure/training, the intelligibility of TFS and RENV speech was similar, suggesting that ENV cues remaining in the TFS-speech signal contribute substantially to its intelligibility. Lorenzi *et al.* (2012) suggested that mild-to-moderate cochlear hearing loss may have a (modest) detrimental effect on ENV recovery. However, the influence of RENV for TFS-speech intelligibility in HI listeners remains unknown.

The goal of the current study was to determine whether RENV cues contribute to TFS-speech intelligibility for HI listeners. If so, it will suggest that the deficit observed for HI listeners with TFS speech may not be entirely attributable to an impaired ability to process TFS cues *per se*, but may arise, at least in part, due to other factors including an impaired ability to recover and use RENV cues. This impaired envelope recovery may be related to the broadened cochlear filters of HI listeners (Baskent, 2006).

II. METHOD

A. Listeners

Seven HI listeners with mild to severe sensorineural hearing loss participated in the study. For each HI listener, an age-matched (within 3 years) NH listener also participated. Table I provides the gender, age and audiometric thresholds (for the tested ear) of each listener. HI listeners were tested using their better ear and NH listeners were tested using the right ear. Note that one older listener (NH7, 69 years old) had a threshold of 25 dB hearing level (HL) at 8 kHz. All listeners were native speakers of American English. All listeners provided informed consent and were paid for their participation in the study.

[a]Author to whom correspondence should be addressed. Current address: School of Psychological Sciences, University of Manchester, Manchester, M13 9PL, United Kingdom. Electronic mail: agnes.leger@manchester.ac.uk
[b]Also at: Sensimetrics Corporation, Malden, MA 02148.

TABLE I. Description of the HI and NH listeners in terms of gender, age (in years), and audiometric thresholds (in dB HL; for the tested ear, see text for details) at octave frequencies (in kHz) in the range of 0.25 to 8 kHz.

			Audiometric threshold specified for frequency					
Listener	Sex	Age	0.25	0.50	1	2	4	8
HI1	F	19	5	10	65	55	15	0
HI2	M	19	10	20	40	60	75	90
HI3	M	21	30	20	40	45	60	90
HI4	F	24	45	50	60	65	65	80
HI5	M	25	30	40	60	45	60	70
HI6	F	63	40	45	40	65	80	—
HI7	F	67	5	5	5	15	40	50
NH1	M	18	10	5	10	10	5	5
NH2	M	20	0	5	5	0	−5	−5
NH3	F	20	10	0	5	0	−5	0
NH4	F	21	5	0	0	0	0	5
NH5	M	25	5	5	5	0	10	10
NH6	F	60	5	10	10	0	15	20
NH7	M	69	10	10	5	5	20	25

B. Speech material and speech maskers

Speech stimuli were recordings of disyllables in /a/-C-/a/ format taken from the corpus of Shannon *et al.* (1999), where C represents one of sixteen consonants (/p t k b d g f s ʃ v z j m n r l/). Each disyllable was uttered once by two male and two female talkers, to form a total set of 64 stimuli. All talkers were speakers of American English (with no noticeable regional accent). The recordings were digitized with 16-bit precision at a sampling rate of 32 kHz (yielding a bandwidth of 16 kHz).

Stimuli were presented at 70 dB sound pressure level (SPL) for NH listeners and amplified using a modified half gain rule (NAL; see Dillon, 2012) for HI listeners. All listeners confirmed that the presentation level was comfortable. Independently of the measures reported here, HI listeners were asked to indicate their preferred presentation level for the (unprocessed) stimuli used in this study; for all HI listeners, the preferred level was between 65 and 70 dB SPL (pre-amplification).

C. Signal processing

Prior to presentation to the listener, the speech stimuli were vocoded as TFS speech or RENV speech. Speech processing details were similar to Swaminathan *et al.* (2014) and are briefly described here.

For TFS speech, the speech signal was first bandpass filtered into N_{TFS} bands of equal bandwidth on a log frequency scale spanning 80 to 8020 Hz. N_{TFS} was chosen to be 1, 2, 4, 8, and 16 based on previous studies addressing similar questions (e.g., Gilbert and Lorenzi, 2006; Swaminathan *et al.*, 2014). The TFS component within each band was extracted as the cosine of the phase of the Hilbert analytic signal. The ENV component was discarded and the TFS component was scaled to match the long-term average energy of the original bandpass signal. The resulting normalized TFS components were then summed to yield the TFS-speech stimulus.

For RENV speech, signal processing was used to recover narrowband ENV cues from the TFS-speech stimuli. For each N_{TFS}, the TFS stimulus was bandpass filtered into 40 bands of equal bandwidth on a log frequency scale spanning 80 to 8020 Hz (simulating a cochlear filterbank). The choice of using 40 bands, with widths smaller than 1 ERB_N (Glasberg and Moore, 1990), was made in Swaminathan *et al.* (2014) based on the findings of Shera *et al.* (2002) who suggested that the human cochlear filters are sharper than the standard behavioral measures. The envelope component within each band was extracted as the magnitude of the Hilbert analytic signal, lowpass filtered to 300 Hz (sixth order Butterworth), and used to modulate a tone carrier at the center of the corresponding frequency band. Each resulting band signal was refiltered through the corresponding bandpass filter to eliminate spectral splatter, and the resulting RENV components were summed to yield the RENV-speech stimulus.

D. Procedure

Consonant identification was measured using a single interval, 16-alternative forced-choice procedure without correct-answer feedback. One experimental run consisted of a single presentation of all 64 syllables in a random order (testing time: 2 to 5 min per run). The 16 possible responses were displayed orthographically on a computer screen and the listener was instructed to identify the consonant and select the response by computer mouse after each presentation. Listeners were first tested using two runs of intact (unprocessed) speech, to familiarize them with the speech material and the task. Listeners were then tested with TFS and RENV speech for a single N_{TFS} condition. Listeners were first tested with the easiest N_{TFS} condition (1 band) followed by $N_{TFS} = 2$, 4, 8, and finally 16 bands. For each N_{TFS}, listeners were tested using a total of eight runs of TFS speech and eight runs of RENV speech. The 16 TFS- and RENV-speech runs were interleaved, always starting with TFS speech. The ordering of N_{TFS}, the alternation of TFS and RENV speech for each N_{TFS}, and the number of runs used were chosen in order to maximize training effect (see Swaminathan *et al.*, 2014). Listeners were tested in five sessions of 90 to 120 min each.

E. Data analysis

For each 64-trial run, a stimulus-response confusion matrix and a percent-correct score were generated. For each processing condition, an overall performance score was obtained by averaging the scores from the final four runs. Performance was analyzed in two ways. First, the overall performance scores were compared by converting the scores into rationalized arcsine units (RAU; Studebaker, 1985) and performing repeated-measures analyses of variance (ANOVAs). Chance performance on the 16-item set was 6.25%-correct, which corresponds to about 2 RAU. Second, aggregate NH and HI confusion matrices were generated for each processing condition by summing across the final four runs for all listeners in a given group (NH or HI). These matrices were then submitted to a form of metric multidimensional scaling analysis (Braida, 1991) to compare their underlying response/confusion patterns.

III. RESULTS

Averaged identification scores of NH and HI listeners are shown in Fig. 1 for intact speech, and for TFS and RENV speech as a function of N_{TFS}.

A. Intact-speech scores

The mean identification scores of the HI listeners for intact speech were about 15 RAU lower than the scores of the NH listeners. A one-way ANOVA conducted with group (NH or HI) as a between-subject factor confirmed that scores of NH and HI listeners were significantly different [$F(1,12) = 11$, $p < 0.01$]. This suggests that, despite amplification of the speech, HI listeners had a poorer ability than NH listeners to identify consonants presented in quiet.

B. Processed-speech scores

For both NH and HI listeners, intelligibility of TFS and RENV speech decreased with increasing N_{TFS}. The intelligibility of RENV speech was slightly lower than that of TFS speech for all N_{TFS}. For HI listeners, intelligibility was poorer than for NH listeners for both TFS and RENV scores. The results of a three-way ANOVA with group as a between-subject factor and processing type and N_{TFS} as within-subject factors are reported below.

Scores of NH and HI listeners were different [significant effect of group: $F(1,12) = 77$, $p < 0.001$], with the average HI score (~55 RAU) being lower than the average NH score (~85 RAU).

For both NH and HI listeners, TFS and RENV scores varied with N_{TFS} [significant effect of N_{TFS}: $F(4,48) = 135$, $p < 0.001$]. The overall trend in these variations were different for NH and HI listeners [significant interaction between groups and N_{TFS}: $F(4,48) = 15$, $p < 0.001$]. For NH listeners, scores decreased with increasing N_{TFS} (~95 RAU for $N_{TFS} = 1$; ~70 RAU for $N_{TFS} = 16$). For HI listeners, scores also decreased with increasing N_{TFS} (~75 RAU for $N_{TFS} = 1$; ~30 RAU for $N_{TFS} = 16$), but the rate of decrease in scores was larger than that observed with NH listeners (a decline of ~45 RAU for HI listeners versus ~25 RAU for NH listeners as N_{TFS} increased from 1 to 16).

Scores for TFS and RENV speech were different [significant effect of processing type: $F(1,12) = 69$, $p < 0.001$], with RENV scores being lower (2 to 6 RAU) than the corresponding TFS scores. This difference was independent of group and N_{TFS} (no significant interaction between processing type, and group [$F(1,12) < 1$], N_{TFS} [$F(4,48) = 3$, $p = 0.054$], or both [$F(4,48) < 1$]).

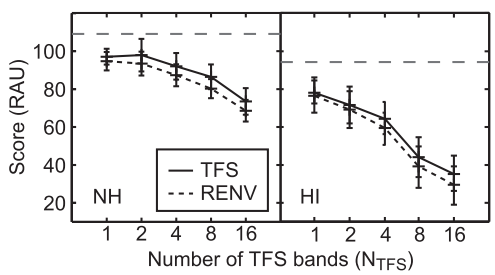

FIG. 1. Averaged speech identification scores (in RAU) for NH and HI listeners as a function of the number of TFS bands. The error bar shows the standard deviation about the mean. The horizontal dotted line shows the score obtained with intact speech.

C. Relationship between speech scores

The confusion matrices obtained for TFS and RENV speech were compared using a metric multidimensional scaling (see Braida, 1991, for details), and the results are presented in Fig. 2. For each condition and for each group, the aggregate confusion matrix was used to derive a set of 16 "stimulus centers" representing each consonant. The distances between these stimulus centers (d') represented the confusability of each pair of consonants. The resulting sets of d''s were used to compare the response/confusion patterns for corresponding TFS and RENV conditions. Significant correlations were observed between these d''s for all N_{TFS} (all $p < 0.001$). Correlations were weaker for $N_{TFS} = 1$ than for other conditions, which may be due to ceiling effects in this condition. This shows that, for both groups and for all N_{TFS}, consonant pairs that were easily distinguished when presented as TFS speech were also easily distinguished when presented as RENV speech, and vice versa, which in turn indicates similar response/confusion patterns for the two types of speech. Note that significant Pearson correlations were also observed between RENV and TFS scores (all $p < 0.001$), confirming the results of this multidimensional scaling analysis.

IV. DISCUSSION

The goal of this study was to determine whether RENV cues contribute to TFS-speech intelligibility for HI listeners, as it has been shown for NH listeners (e.g., Swaminathan *et al.*, 2014).

For NH listeners, RENV-speech scores were similar (although consistently lower by 2–6 RAU) to TFS-speech scores. This trend is consistent with findings reported by Swaminathan *et al.* (2014), although TFS-speech scores for $N_{TFS} = 16$ were higher in the current study (~70 RAU) than in Swaminathan *et al.* (2014; ~50 RAU) using the same stimuli. This may be attributable to greater opportunity for training

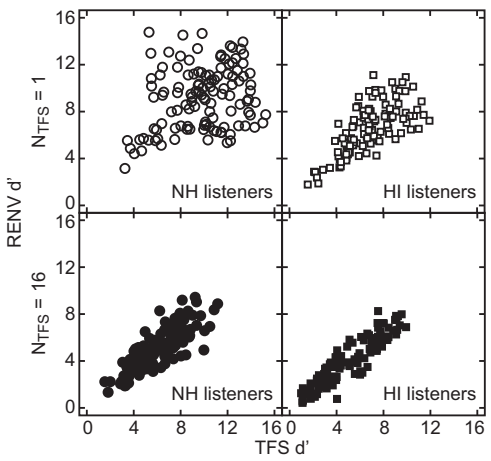

FIG. 2. Comparison of the confusion matrices obtained for TFS and RENV speech using a metric multidimensional scaling (Braida, 1991), for NH and HI listeners and for $N_{TFS} = 1$ and 16. The d''s represent the discriminability of each pair of consonants (see text for details).

in the current study where listeners were always tested in increasing order of difficulty. The discriminability of pairs of consonants was extremely similar for TFS and RENV speech, suggesting that analogous phonetic-feature information was conveyed by the two types of signals. Taken together, these results confirm that RENV cues extracted from narrow filters (<1 ERB_N) largely account for the intelligibility of TFS speech for NH listeners.

For HI listeners, identification scores and consonant-pair discriminability were also similar for RENV- and TFS-speech, which suggests that RENV cues may account for the intelligibility of TFS-speech for these listeners as well. If this is indeed the case, then the intelligibility deficits evident for HI listeners with TFS speech may not be related to a reduced ability to process TFS cues *per se*, as suggested in previous studies (e.g., Lorenzi *et al.,* 2006), but may arise instead from a different mechanism such as an impaired ability to extract and use RENVs (see below). Further, this result suggests that the Hilbert-transform-generated TFS speech used here may not be the appropriate vehicle with which to study TFS processing for NH and HI listeners, as previously suggested in other studies (e.g., Shamma and Lorenzi, 2013).

Increasing the number of bands used to generate TFS speech (N_{TFS}) led to a decrease in TFS- and RENV-speech intelligibility for both NH and HI listeners, as reported before (e.g., Apoux *et al.,* 2013; Gilbert and Lorenzi, 2006). However, at odds with previous studies (e.g., Gilbert and Lorenzi, 2006), scores for TFS- and RENV-speech remained similar for all N_{TFS}. This difference is most likely explained by the differences in the construction of RENV speech: the 40 filters used in the present study were sharper than the 30 used in Gilbert and Lorenzi (2006), which led to a better recovery of ENV from the TFS (Ghitza, 2001). It is unclear which filtering characteristic is most appropriate for the simulation of the normal and/or impaired auditory system.

Speech intelligibility was lower for HI listeners than for NH listeners for all processing conditions. The deficit of the HI listeners was larger for increasing N_{TFS}. One source of this deficit may be differences in audibility, based on different (NAL) amplification across listeners. However, the effects of the amplification were similar across N_{TFS} and therefore would not explain the observed dependence. Another source of this deficit may be an impaired ability of HI listeners to extract and use RENVs. As N_{TFS} increases, the short-time spectrum of TFS speech becomes more homogenous in both time and frequency (tending toward the long-term average spectrum) and the envelope information available for recovery decreases (e.g., Gilbert and Lorenzi, 2006). An impaired mechanism for extracting and/or using RENVs, for example, due to broadened cochlear filters (e.g., Baskent, 2006; Lorenzi *et al.,* 2012), may be more susceptible than a non-impaired mechanism to the degradations in quantity and quality of available RENV information. This would explain why HI performance decreases more rapidly than NH with an increase in N_{TFS}. It could also be the case that other factors constrained the intelligibility of both TFS and RENV speech for HI listeners, such as the presence of amplified noise in TFS and RENV speech (Apoux *et al.,* 2013; Hopkins *et al.,* 2010).

Finally, providing TFS cues as narrow bands of RENV cues (simulating healthy cochlear filtering) did not provide any benefit to HI listeners. This may be related to an impaired HI mechanism for extracting and using RENVs. In the current study, 40 bands of envelopes were extracted from TFS speech and presented to the listener as modulated tone carriers for these 40 bands (RENV speech). Broadened cochlear filters may have "re-smeared" the artificially extracted RENVs and limited their use by the HI listeners. It is possible that improved methods for presenting artificially extracted RENVS, such as providing alternating-band RENVs dichotically to the listener, may improve performance (e.g., Ghitza, 2001).

ACKNOWLEDGMENTS

This research was supported by National Institutes of Health Grants Nos. R43 DC013006 and R01 DC00117.

Apoux, F., Yoho, S. E., Youngdahl, C. L., and Healy, E. (**2013**). "Can envelope recovery account for speech recognition based on temporal fine structure?," Proc. Meet. Acoust. **19**, 050072.
Baskent, 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.
Braida, L. D. (**1991**). "Crossmodal integration in the identification of consonant segments," Q. J. Exp. Psychol. A **43**, 647–677.
Dillon, H. (**2012**). *Hearing Aids*, 2nd ed. (Boomerang, New York).
Flanagan, J. L. (**1980**). "Parametric coding of speech spectra," J. Acoust. Soc. Am. **68**, 412–419.
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., and Lorenzi, C. (**2006**). "The ability of listeners to use recovered envelope cues from speech fine structure," J. Acoust. Soc. Am. **119**, 2438–2444.
Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103–138.
Heinz, M. G., and Swaminathan, J. (**2009**). "Quantifying envelope and fine-structure coding in auditory nerve responses to chimaeric speech," J. Assoc. Res. Otolaryngol. **10**(3), 407–423.
Hopkins, K., Moore, B. C. J., and Stone, M. A. (**2010**). "The effects of the addition of low-level, low-noise noise on the intelligibility of sentences processed to remove temporal envelope information," J. Acoust. Soc. Am. **128**, 2150–2161.
Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., and Moore, B. C. J. (**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.
Lorenzi, C., Wallaert, N., Gnansia, D., Leger, A. C., Ives, D. T., Chays, A., Garnier, S., and Moore, B. C. J. (**2012**). "Temporal-Envelope Reconstruction for Hearing-Impaired Listeners," J. Assoc. Res. Otolaryngol. **103**, 18866–18869.
Shamma, S., and Lorenzi, C. (**2013**). "On the balance of envelope and temporal fine structure in the encoding of speech in the early auditory system," J. Acoust. Soc. Am. **133**, 2818–2833.
Shannon, R. V., Jensvold, A., Padilla, M., Robert, M. E., and Wang, X. (**1999**). "Consonant recordings for speech testing," J. Acoust. Soc. Am. **106**, L71–L74.
Sheft, S., Ardoint, M., and Lorenzi, C. (**2008**). "Speech identification based on temporal fine structure cues," J. Acoust. Soc. Am. **124**, 562–575.
Shera, C. A., Guinan, J. J., Jr., and Oxenham, A. J. (**2002**). "Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements," Proc. Natl. Acad. Sci. U.S.A. **99**, 3318–3323.
Smith, Z. M., Delgutte, B., and Oxenham, A. J. (**2002**). "Chimaeric sounds reveal dichotomies in auditory perception," Nature **416**, 87–90.
Studebaker, G. A. (**1985**). "A 'rationalized' arcsine transform," J. Speech Hear. Res. **28**, 455–462.
Swaminathan, J., Reed, C. M., Desloge, J. G., Braida, L. D., and Delhorne, L. A. (**2014**). "Consonant identification using temporal fine structure and recovered envelope cues," J. Acoust. Soc. Am. **135**(4), 2078–2090.
Zeng, F. G., Nie, K., Liu, S., Stickney, G., Del Rio, E., Kong, Y. Y., and Chen, H. (**2004**). "On the dichotomy in auditory perception between temporal envelope and fine structure cues," J. Acoust. Soc. Am. **116**, 1351–1354.