Enhancement of speech intelligibility in reverberant rooms: Role of amplitude envelope and temporal fine structure

Nirmal Kumar Srinivasan^{a)} and Pavel Zahorik^{b)}

Department of Psychological and Brain Sciences, University of Louisville, Louisville, Kentucky 40292

nirmal.srinivasan@va.gov, pavel.zahorik@louisville.edu

Abstract: The temporal envelope and fine structure of speech make distinct contributions to the perception of speech in normal-hearing listeners, and are differentially affected by room reverberation. Previous work has demonstrated enhanced speech intelligibility in reverberant rooms when prior exposure to the room was provided. Here, the relative contributions of envelope and fine structure cues to this intelligibility enhancement were tested using an open-set speech corpus and virtual auditory space techniques to independently manipulate the speech cues within a simulated room. Intelligibility enhancement was observed only when the envelope was reverberant, indicating that the enhancement is envelope-based.

© 2014 Acoustical Society of America

PACS numbers: 43.66.Lj, 43.71.Gv [QJF]

Date Received: November 22, 2013

Date Accepted: April 10, 2014

1. Introduction

Speech signals contain two forms of temporal information within each frequency band: Temporal envelope (ENV), which describes the slow variations in the amplitude of the speech signal over time, and temporal fine structure (TFS), which describes the rapid variations with rate equivalent to the center frequency of the band. From a signal processing stand point, the ENV can be considered as the modulating signal and the fine structure can be considered as the carrier signal. When speech is presented in a room, the acoustical properties of the room can result in measurable distortions to both ENV and TFS components of the speech signal at the location of the listener. For normal-hearing listeners, speech intelligibility is surprisingly robust to these physical distortions, however. Measureable reductions in intelligibility are typically observed for normal-hearing listeners only when the room is extremely reverberant, or when noise is also present (Nabelek and Pickett, 1974).

One possible explanation for this robustness is that the normally functioning auditory system may be capable of adapting to listening environments with moderate amounts of reverberation (Brandewie and Zahorik, 2010, 2013; Srinivasan and Zahorik, 2011, 2013; Watkins, 2005a,b). There is further evidence that the adaptation processes may depend primarily on the ENV portions of the speech signal, since Watkins *et al.* (2011) have demonstrated similar reverberation adaptation effects for both normal speech sounds and noise vocoded speech, which removes TFS cues but preserves ENV. Because the Watkins *et al.* (2011) study was conducted using phonetically-limited speech materials (syllable continuum from “sir” to “stir”), it is of

^{a)}Current address: National Center for Rehabilitative Auditory Research, Portland VA Medical Center, Portland, OR. Also at: Department of Otolaryngology, Oregon Health and Science University, Portland, OR.

^{b)}Author to whom correspondence should be addressed. Also at: Division of Communicative Disorders, Department of Surgery, University of Louisville School of Medicine, KY.

interest to determine whether this effect generalizes to more natural and phonetically diverse speech materials.

The goal of this study is to examine the role of ENV and TFS cues to speech intelligibility in a reverberant room and to determine whether prior listening exposure to the room can improve the intelligibility of everyday speech. Sentences from the Perceptually Robust English Sentence Test Open-set (PRESTO) speech corpus (Gilbert *et al.*, 2013) were used. This corpus contains sentences sampled from the TIMIT database (Garofolo *et al.*, 1993), but specifically selected to create lists of sentences that are highly variable in terms of linguistic and indexical properties of the speech, but balanced for keyword frequency and familiarity. Virtual acoustic methods identical to past work (Brandewie and Zahorik, 2010; Srinivasan and Zahorik, 2013; Brandewie and Zahorik, 2013) were used to simulate listening conditions in reverberant rooms. Methods fundamentally similar to those described by Smith *et al.* (2002) were then used to independently manipulate room effects on ENV and TFS. Performance in two presentation conditions was compared: A “blocked” condition in which the listening environment was kept constant throughout a block of sentence trials, and an “unblocked” condition in which the listening environment was varied from trial to trial. If room adaptation effects generalize to these listening situations, then one would predict improved intelligibility in the blocked condition relative to the unblocked condition, because the former provides consistent exposure over time to a single listening environment, where the latter does not.

2. Methods

2.1 Subjects

Thirty normal-hearing (pure-tone air-conduction thresholds ≤ 20 dB hearing level from 250 to 8000 Hz) adults (ages ranged from 18.1 to 29.7 yrs with an average age of 21.7 yrs) participated in this experiment. All the subjects were paid for their participation. None had prior experience with the speech corpus materials used in the study. All procedures involving human subjects were reviewed and approved by the University of Louisville Institutional Review Board.

2.2 Room simulation

Virtual acoustic techniques were used to simulate the seven listening environments in this study. The environments were seven simulated rooms (R_0 , R_1 , R_2 , R_3 , R_4 , R_5 , and R_6) with broadband (125 to 4000 Hz) reverberation times (T_{60}) of 0 (anechoic), 0.20, 0.30, 0.49, 0.70, 1.22, and 2.32 s, respectively. All the rooms had identical dimensions of 5.7 m (length) $\times$ 4.3 m (width) $\times$ 2.6 m (height) but varied in the absorptive properties of the reflective surfaces. Within each simulated room, the speech source was positioned at a distance of 1.4 m in front of the listener. Room simulation techniques were based on methods described by Zahorik (2009). Briefly, the simulation methods used an image model to compute directions, delays, and attenuations of the early reflections, which were then, along with the direct-path, spatially rendered using non-individualized head-related transfer functions. The late reverberant energy was simulated statistically using exponentially decaying independent Gaussian noise samples in octave bands from 125 to 4000 Hz for each ear. Overall, this simplified method of room simulation has been found to produce binaural room impulse responses (BRIRs) that are reasonable physical and perceptual approximations of those measured in real rooms (Zahorik, 2009).

2.3 Stimuli

Speech stimuli consisted of eight individual lists from the (PRESTO) speech corpus (Gilbert *et al.*, 2013). Each list contained 18 sentences of varying length and syntactic structure with a total of 76 key words. Within a list, no sentence or talker was repeated.

Speech signals were first filtered into sixteen 0.4-octave-wide contiguous frequency bands using sixth-order zero-phase Butterworth filters with center frequencies

ranging from 80 to 6010 Hz (Gilbert and Lorenzi, 2006). The ENV and TFS of the individual bands were obtained using Hilbert transform techniques (Smith *et al.*, 2002), and then independently subjected to the acoustical effects of the reverberant rooms as follows. In one stimulus condition, “reverb-ENV,” the ENV was extracted from speech signals convolved with BRIRs from the different reverberant rooms ($R1$ – $R6$) and the TFS was extracted from the same signal, but convolved with an anechoic ($R0$) BRIR. The results were then summed to produce a hybrid speech signal in which only the ENV was distorted by the effects of a given reverberant room, not the TFS. In a second complementary stimulus condition, “reverb-TFS,” the TFS was extracted from speech convolved with reverberant BRIRs and the ENV from the same speech convolved with an anechoic BRIR. Results were again summed to produce a hybrid speech signal in which only TFS cues were distorted by room reverberation, but ENV cues were not.

2.4 Design

Out of the 30 subjects, 15 subjects participated in phase 1 of the experiment which presented both reverb-TFS and reverb-ENV speech stimuli. Another 15 subjects in phase 2 of the experiment were presented with only reverb-ENV speech stimuli, but using different PRESTO sentence lists in order to determine whether any observed intelligibility enhancement effects were list dependent. The PRESTO sentence lists for the different experimental conditions during phase 1 and phase 2 of the experiment are shown in Table 1.

All the listeners were tested in two presentation conditions: One was blocked by simulated room and another was unblocked by room. In the unblocked condition, the room environment was selected at random from trial to trial from one of the six simulated rooms ($R1$ – $R6$) that differed in their reverberation characteristics (see Sec. 2.2). This was done to limit the consistent listening exposure to any one room. The randomization procedure was further refined to insure that the same room did not occur on successive trials. In the blocked condition, the listening environment (either $R2$ or $R4$) was fixed within a block of trials. This provided consistent exposure to a single room (either $R2$ or $R4$) over the trial block. The type of presentation (reverb-TFS and reverb-ENV) was always kept fixed within a block of trials.

Over the course of the experiment, each participant in phase 1 ($n=15$) was presented with a total of 288 sentences split into 10 sets of trials. The first four sets (1–4) each consisted of 18 sentences (one full list) and was blocked by listening environment and stimulus condition. The remaining sets were all unblocked by listening environment. In sets 5, 7, and 9, listeners were presented with 6 sentences each from reverb-TFS (fine structure convolved with $R1$, $R2$, $R3$, $R4$, $R5$, and $R6$ BRIRs) with presentation order randomized within the set. Sets 6, 8, and 10 were analogous to sets 5, 7, and 9, but involved presentation type of reverb-ENV (ENV convolved with $R1$, $R2$, $R3$, $R4$, $R5$, and $R6$ BRIRs).

Each participant in phase 2 ($n=15$) was presented only with reverb-ENV speech stimuli. The participants in phase 2 listened to a total of 144 sentences split up

Table 1. Individual lists used in different experimental conditions for phase 1 and phase 2. B denotes blocked presentation and U denotes unblocked presentation.

	Phase 1				Phase 2	
	Reverb-ENV		Reverb-TFS		Reverb-ENV	
	B	U	B	U	B	U
$R2$ ($T_{60}=0.3$ s)	List 1	List 2	List 22	List 13	List 3	List 10
$R4$ ($T_{60}=0.7$ s)	List 4	List 5	List 20	List 16	List 12	List 14

into 5 sets of trials. The first two sets were blocked and the remaining three sets were unblocked.

The sentences in different listening environments of interest (*R2* and *R4*) were taken from the same list so that the number of key words was equated. No speech material was repeated during the entire study. Each listener completed the experiment in a single session which lasted approximately 1.5 h.

2.5 Procedure

The listeners were seated in a sound-attenuating chamber (Acoustic Systems, Austin, TX, custom double wall) and listened to speech stimuli via headphones (Beyerdynamic - DT990 Pro, Beyerdynamic Inc. USA, Farmingdale, NY). The speech stimuli were presented to the listeners at an A-weighted sound level of approximately 68 dB sound pressure level at the ear. The listeners were instructed to listen to the speech stimulus presented and type all the words they understood. No feedback was provided regarding the number of words they correctly entered during each trial. The listeners were encouraged to guess if they were unsure about the auditory input. Data collection was self-paced, and the listeners were instructed to take breaks whenever they felt fatigued. All stimulus presentation and data collection was implemented using MATLAB software and statistical analyses were performed using Statistical Package for Social Sciences.

2.6 Scoring

The percentage of key words correctly reported was scored as a measure of speech intelligibility for each combination of listening environments (*R2* and *R4*), stimulus conditions (reverb-ENV and reverb-TFS), and presentation conditions (blocked and unblocked). The percent correct speech intelligibility scores were transformed to rationalized arcsine units (RAUs) (Studebaker, 1985) prior to subsequent statistical analyses. Separate repeated-measures analysis of variance (ANOVA) tests were conducted to evaluate differences in RAU intelligibility scores across different experimental conditions.

3. Results and discussion

Two-sample *t*-tests were performed on the data from the reverb-ENV condition for phase 1 and phase 2 listeners. No statistically significant differences in speech intelligibility were observed between the two phases which suggests that the effects observed do not depend upon the particular PRESTO list presented in different conditions [Blocked presentation, *R2*: $t(28)=0.466$, $p=0.645$; Unblocked presentation, *R2*: $t(28)=0.305$, $p=0.762$; Blocked presentation, *R4*: $t(28)=0.202$, $p=0.841$; Unblocked presentation, *R4*: $t(28)=1.305$, $p=0.203$]. Data from the two phases for reverb-ENV conditions were thus pooled together for all subsequent analyses.

The across-subject average of percent key words correct (in RAU units) is shown in Fig. 1(a) for reverb-TFS and reverb-ENV stimulus conditions in both *R2* and *R4* listening environments during blocked and unblocked presentation conditions. Of primary interest is the improvement in performance between the unblocked and the blocked conditions for the two stimulus conditions, which were evaluated using matched-pair *t*-tests. Statistically significant improvement was observed in the reverb-ENV stimulus condition only, at both the 0.3-s [$t(1,29)=4.815$, $p<0.001$] and the 0.7-s [$t(1,29)=9.076$, $p<0.001$] reverberation times. No statistically significant improvements were observed in the reverb-TFS condition at either 0.3-s [$t(1,14)=0.727$, $p=0.479$] or 0.7-s [$t(1,14)=1.885$, $p=0.080$] reverberation time. These results are consistent with the hypothesis that ENV cues, rather than TFS cues, support enhanced speech understanding in reverberation.

It is important to note, however, that overall intelligibility also depends on the amount of reverberation and on the ENV/TFS manipulation. Both of these results are well known (Knudsen, 1929; Lochner and Burger, 1964; Shannon *et al.*, 1995), and therefore not of primary interest in this study. They do raise the issue of the potential

Fig. 1. (Color online) Speech intelligibility results from reverb-TFS (upper panels, $n = 15$) and reverb-ENV (lower panels, $n = 30$) stimulus conditions at 0.3 s (blue symbols) and 0.7 s (green symbols) reverberation times. For each condition, results from three types of analyses are shown. (a) Mean transformed speech intelligibility (RAU) for blocked (filled bars: “B”) and unblocked (unfilled bars: “U”) trials as a function of reverberation time. Error bars show 95% confidence intervals for each mean. Results of match-pair t -tests for the effect of trial blocking are also displayed. (b) Scatter plots of speech intelligibility for 0.3 s (blue circles) and 0.7 s (green squares) reverberation times. Each point indicates data from one listener at a given reverberation time. (c) Mean transformed speech intelligibility (RAU) as a function of exposure time, measured in six-sentence epochs. Data are shown for blocked (filled points) and unblocked (unfilled points) trials for 0.3 s (blue circles) and 0.7 s (green squares) reverberation times. Error bars show 95% confidence intervals for each mean.

difficulties in comparing relative performance across quite different overall performance levels, however. Mitigating these concerns are the facts that: (1) The data were transformed to RAU prior to all analyses, and (2) similar performance levels produced very different changes in blocked/unblocked performance depending on the ENV/TFS manipulation [e.g., Fig. 1(a), Reverb-TFS at 0.7-s T_{60} versus Reverb-ENV at 0.3-s T_{60}].

The patterns of results observed in the group-averaged data [Fig. 1(a)] are also observable in nearly all individuals. Figure 1(b) shows scatter plots of transformed speech intelligibility for blocked versus unblocked presentation, for both reverb-ENV and reverb-TFS stimulus conditions. Each plot displays data from both listening environments. Although substantial individual variability may be observed, nearly all listeners demonstrated improved intelligibility in blocked versus unblocked presentation conditions for the reverb-ENV stimulus condition. In R2, 25 out of 30 listeners showed improved speech intelligibility and in R4, 14 out of 15 listeners showed improvement. Similar results were not observed for the reverb-TFS stimulus condition.

Overall, these results suggest that the enhanced intelligibility with consistent prior exposure is related to processing of the amplitude ENV of the speech signals. They are also consistent with the results of Watkins *et al.* (2011), and demonstrate that the enhancement generalizes to highly variable sentence materials. Taken together, these results may be indicative of mechanisms that compensate for the distortions to the amplitude ENV caused by room reverberation. Related work demonstrated, for example, that human detection thresholds for amplitude modulation (AM) in a reverberant sound field is 4 to 6 dB lower than would be predicted based on the modulation-transfer function of the room (Zahorik *et al.*, 2011; Zahorik *et al.*, 2012), and this enhancement in modulation sensitivity appears to be related to consistent prior exposure to the same reverberant sound field (Zahorik and Anderson, 2013). Although the precise relationship between AM sensitivity and speech perception has

not yet been demonstrated, it is clear that the amplitude ENV of the speech signal carries sufficient information for speech perception, even under degraded conditions, such as when the number of frequency channels are limited (Shannon *et al.*, 1995).

An additional and important question relates to the amount of listening exposure required for intelligibility improvement. To address this question, intelligibility was analyzed over three sentence time epochs in the experiment. The results of this analysis are shown in Fig. 1(c). Results from separate repeated-measures ANOVAs for each condition confirmed that there were no significant differences in performance in any of the conditions over time. These results suggest that the processes underlying the intelligibility improvement are relatively rapid (within the first six sentences), and once the improvements are realized, there is little additional long term improvement. This too is consistent with results from past work using homogeneous (Brandewie and Zahorik, 2010; Brandewie and Zahorik, 2013) and heterogeneous speech materials (Srinivasan and Zahorik, 2013).

4. Conclusions

Results from this study confirm that the enhancement in speech intelligibility due to prior exposure to a reverberant listening environment depends on perceptual processing related to the amplitude ENV of the speech signal. The enhancement is not specific to certain syllable continua (Watkins *et al.*, 2011) but generalizes to highly variable speech sentences that are more representative of that encountered in everyday communication situations. In addition, the enhancement process appears to be very rapid—on the order of seconds—and does not show any long-term improvement.

Acknowledgments

The authors wish to thank Noah Jacobs for his assistance in data collection. The research was supported by NIH-NIDCD (Grant No. R01DC008168).

References and links

- Brandewie, E. J., and Zahorik, P. (2010). "Prior listening in rooms improves speech intelligibility," *J. Acoust. Soc. Am.* **128**(1), 291–299.
- Brandewie, E. J., and Zahorik, P. (2013). "Time course of a perceptual enhancement effect for noise-masked speech in reverberant environments," *J. Acoust. Soc. Am.* **134**(2), EL265–EL270.
- Garofolo, J., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., and Zue, V. (1993). DARPA TIMIT acoustic-phonetic continuous speech corpus, Linguistic Data Consortium, Philadelphia.
- Gilbert, G., and Lorenzi, C. (2006). "The ability of the listeners to use recovered envelope cues from speech fine structure," *J. Acoust. Soc. Am.* **119**(4), 2438–2444.
- Gilbert, J. L., Tamati, T. N., and Pisoni, D. B. (2013). "Development, reliability, and validity of PRESTO: A new high-variability sentence recognition test," *J. Am. Acad. Audiol.* **24**(1), 26–36.
- Knudsen, V. O. (1929). "The hearing of speech in auditoriums," *J. Acoust. Soc. Am.* **1**, 56–82.
- Lochner, J., and Burger, J. (1964). "The influence of reflections in auditorium acoustics," *J. Sound Vib.* **1**(4), 426–454.
- Nabelek, A. K., and Pickett, J. M. (1974). "Monaural and binaural speech perception through hearing aids under noise and reverberation with normal and hearing-impaired listeners," *J. Speech. Hear. Res.* **17**(4), 724–739.
- 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* **416**, 87–90.
- Srinivasan, N. K., and Zahorik, P. (2011). "The effect of semantic context on speech intelligibility in reverberant rooms," *Proc Meet Acoust.* **12**, 060001.
- Srinivasan, N. K., and Zahorik, P. (2013). "Prior listening exposure to a reverberant room improves open-set intelligibility of high-variability sentences," *J. Acoust. Soc. Am.* **133**(1), EL33–EL39.
- Studebaker, G. A. (1985). "A rationalized arcsine transform," *J. Speech Hear. Res.* **28**(3), 455–462.
- Watkins, A. J. (2005a). "Perceptual compensation for effects of echo and of reverberation on speech identification," *Acta Acust. Acust.* **91**, 892–901.

- Watkins, A. J. (2005b). "Perceptual compensation for effects of reverberation in speech identification," *J. Acoust. Soc. Am.* **118**(1), 249–262.
- Watkins, A. J., Raimond, A. P., and Makin, S. J. (2011). "Temporal-envelope constancy of speech in rooms and the perceptual weighting of frequency bands," *J. Acoust. Soc. Am.* **130**(5), 2777–2788.
- Zahorik, P. (2009). "Perceptually relevant parameters for virtual listening simulation of small room acoustics," *J. Acoust. Soc. Am.* **126**(2), 776–791.
- Zahorik, P., and Anderson, P. W. (2013). "Amplitude modulation detection by human listeners in reverberant sound fields: Effects of prior listening exposure," *Proc Meet Acoust.* **19**, 050139.
- Zahorik, P., Kim, D. O., Kuwada, S., Anderson, P. W., Brandewie, E., Collecchia, R., and Srinivasan, N. K. (2012). "Amplitude modulation detection by human listeners in sound fields: carrier bandwidth effects and binaural versus monaural comparison," *Proc Meet Acoust.* **15**, 050002.
- Zahorik, P., Kim, D. O., Kuwada, S., Anderson, P. W., Brandewie, E., and Srinivasan, N. K. (2011). "Amplitude modulation detection by human listeners in sound fields," *Proc Meet Acoust.* **12**, 050005.