ForumAcusticum
2 0 0 5

Calibration of speech perception to room reverberation

Kanako Ueno[1,2], Norbert Kopčo[1,3], Barbara Shinn-Cunningham[1]

[1] Hearing Research Center, Boston University, 677 Beacon St. Boston, MA 02215, USA, **shinn@cns.bu.edu**
[2] Institute of Industrial Science, Tokyo University, Komaba 4-6-1, Meguroku,Tokyo,153-8505, Japan, **ueno@iis.u-tokyo.ac.jp**
[3] Dept of Cybernetics and AI, Technicka Univerzita, Letna 9, 04001 Košice, Slovakia, **kopco@bu.edu**

Past studies of sound perception often assumed that our auditory sensory processes are relatively static, rather than plastic. However, in everyday environments, we naturally and fluidly compensate for interfering effects of background noise and room reverberation. In order to investigate how listeners calibrate auditory perception to such acoustic interference, a listening experiment was performed to measure the effect of sudden changes of reverberation on the identification of consonants. Binaural room impulse responses (BRIRs) measured in real rooms were convolved with speech tokens to simulate natural interference caused by reverberant energy. In the experiment, listeners identified the consonant present in a vowel-consonant target. On most trials, the target was presented following a carrier phrase (in a control condition, there was no preceding carrier). In some cases, the target and carrier phrase were processed by the same BRIRs while in others the BRIRs processing target and carrier differed in their types of reverberation. Results suggest that presenting a carrier and target with matching BRIRs improves accuracy of target consonant identification compared to cases in which the reverberation of the preceding carrier does not match that of the target.

1 Introduction

Many studies of auditory perception assume that our sensory processes are relatively static and fixed, dependent only on the input signals rather than on the state of the listener. However, in everyday environments, we naturally and fluidly compensate for many interfering effects of background noise and room reverberation. The few previous studies that explored such "room calibration" hint that experience alters both how we localize and how we interpret content of auditory signals in rooms (e.g., see [1,2]).

The current study was designed to test whether consistent room experience improved speech intelligibility. Listening experiments were performed to measure the effect of sudden changes of reverberation on the identification of consonants in vowel-consonant (VC) pairs. We hypothesized that because the reverberant context can bias the interpretation of phonemes (as shown for vowels and stop consonants in [2]), overall consonant identification will be better when listeners hear consistent room cues just prior to a test stimulus.

2 Stimuli

2.1 Speech source

In ordinary speech communication, linguistic and contextual cues play an important role in overall speech intelligibility. In order to factor out these effects and to measure only changes in speech perception due to non-linguistic processing, simple VC speech syllables were used. Sixteen consonants (k, t, p, f, g, d, b, v, ð, m, n, ŋ, z, θ, s, and ʃ) were tested, each proceeded by the same vowel /a/. For each VC, three tokens were spoken by three talkers (two males and one female, 16 VCs x 3 talkers x 3 utterances = 144 in total). A male recording was taken from CUNY-NST corpus [3] and both a male recording and a female recording were taken from the corpus described in [4]. Level differences across talkers were removed by equalizing the average energy levels of all VCs to the same energy (squared and time-integrated sound pressure).

2.2 Binaural impulse responses

To simulate natural interference caused by reverberant energy, binaural room impulse responses (BRIRs) were used. To measure the BRIRs, sound was presented from an omni-directional (up to 2 kHz) dodecahedral loudspeaker system (TS-12M) located at a representative position in the tested environments. Received sound was measured at the ears of a manikin head (Head Acoustics, HMM2) placed to face the sound source from a position in the audience area. Impulse responses were measured using the swept-sine method.

BRIRs from two different large rooms were used, denoted as R1 and R2. Both R1 and R2 noticeably disrupt consonant identification; however, the two BRIRs contain different types of reverberation. R1 was measured in an elliptical church with the manikin relatively close to the sound source (distance = 12 m). R2 was measured in a large concert hall (2,020 seats)

with the manikin located in the second balcony, 33 m from the speaker system.

Figure 1 shows the acoustic properties of BRIRs. Early time-domain portions of the responses in one ear are shown in Fig. 1a. Because of its elliptic room shape, R1 has a large echo around 60 ms after the direct sound, seen prominently in the 500 Hz octave band. Fig. 1b shows that R1 has longer reverberation times (T_{60}) than R2 in all frequencies. Fig. 1c shows that the ratio of the early energy (0-50 ms) to the late energy (beyond 50 ms), C_{50}, is lower in R1 than in R2, especially in the mid-frequency bands (250-1000 Hz). This analysis suggests that R1 should be more disruptive to speech than R2; however, the STI index [5] is similar for both settings (Fig. 1d).

The R1 BRIRs were also processed (by a 5-ms time window) to remove most reverberant energy, generating "pseudo-anechoic" (AE) BRIRs. The resulting three BRIRs (R1, R2, and AE) were equalized for overall energy.

2.3 Presentation of the stimuli

Speech tokens were convolved with the different BRIRs at a 25 kHz sampling rate to generate binaural reverberant stimuli. Test signals were presented from MATLAB through a D/A converter (TDT RP2) and headphone amplifier (TDT HB7) driving insert headphones (Etymotic Research, ER1) at a comfortable listening level (adjusted by the experimenter). All tests were conducted in a sound treated room.

3 Experiment 1

3.1 Design

Figure 2 illustrates the design of the stimuli. On all trials, listeners were instructed to report the consonant in the final VC pair presented in the trial. On most trials, a carrier consisting of other VC syllables preceded the target (in the control condition, there was no preceding carrier). VCs within a trial were separated by an inter-stimulus interval of 0.8 s. The number of VCs preceding the target could be zero, two, or four; in order to encourage listeners to actively attend to the target, these different-length trials were randomly ordered within each session.

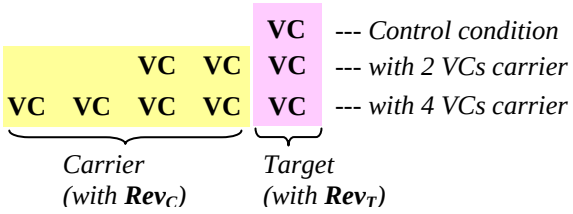
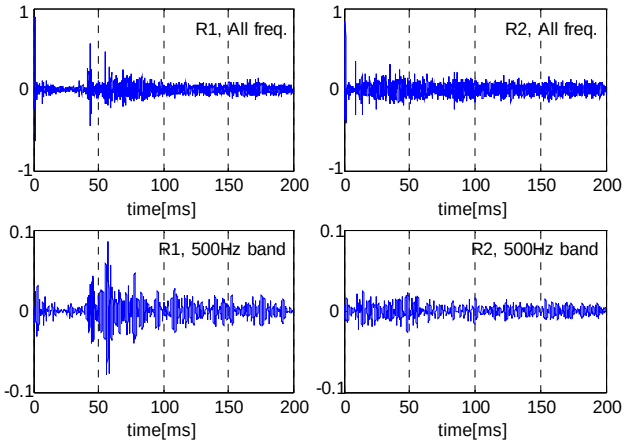
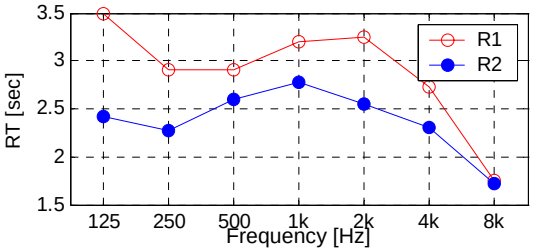


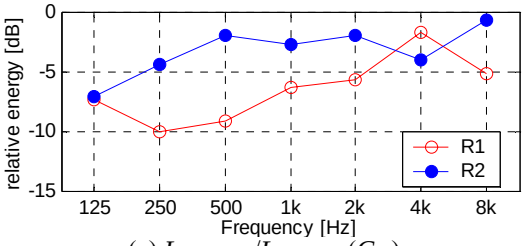
Figure 2: Composition of stimuli



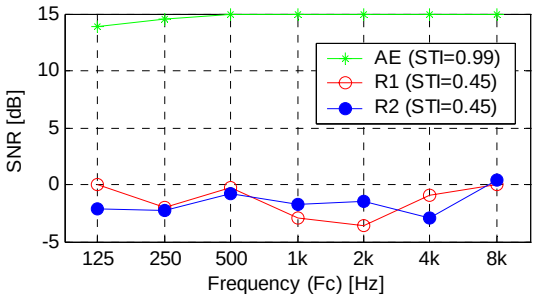
(a) Echo diagram (Lch.)



(b) Reverberation time (T_{60})



(c) $L_{E(0\text{-}50ms)}/L_{E(50\text{-}ms)}$ (C_{50})



(d) SNR and STI

Figure 1: Acoustic properties of BNIRs

In some trials, the target and carrier were convolved with the same BRIRs ($\boldsymbol{Rev_C}$= $\boldsymbol{Rev_T}$) while in others the target and carrier were processed by different BRIRs ($\boldsymbol{Rev_C}$≠$\boldsymbol{Rev_T}$). Each subject performed roughly 200 trials with carrier-target reverberation matching (half with R1 as the target, half with R2), 300 trials with unmatching reverberation (half with R1 as the target, half with R2), and 150 control trials (a third each with R1, R2, and AE targets; see Table 1). Within each trial, all VCs were taken from the same talker, who was randomly selected. All stimuli (15 x 48 = 720 trials in

total) were blocked into 3 sessions (240 trials each) that took approximately 30 minutes to complete. On each trial, the subject indicated the perceived target VC by clicking with a computer mouse on one of 16 graphical buttons labeled with the VCs.

Table 1: Number of trials for each condition in Exp.1. Matched reverberation condition (Rev_C=Rev_T) were highlighted.

		Rev_T		
		R1	R2	AE
No carrier (control cond.)		48	48	48
Rev_C (2 and 4VCs)	R1	**48x2**	48x2	-
	R2	48x2	**48x2**	-
	AE	48x2	48x2	-

3.2 Results

Fourteen native English speakers with normal hearing participated in the experiment. Percent-correct identification scores were calculated for each condition and subject. Large individual differences were seen. To test for significance of the results, an independent groups one-way ANOVA (with condition as the factor) was performed. The result showed that differences across conditions were highly significant ($p<0.001$). The significance of differences between pairs of conditions was tested by multiple comparisons of condition means. Figure 3 shows the result of these pair-wise comparisons, with error bars showing 95% confidence intervals. Performance for anechoic VCs (AE), whose mean was 93.5% and significantly different from all of the other conditions, is excluded. Matched reverberation conditions, which we hypothesized would produce better performance due to the benefit of room calibration, are highlighted in Fig. 3.

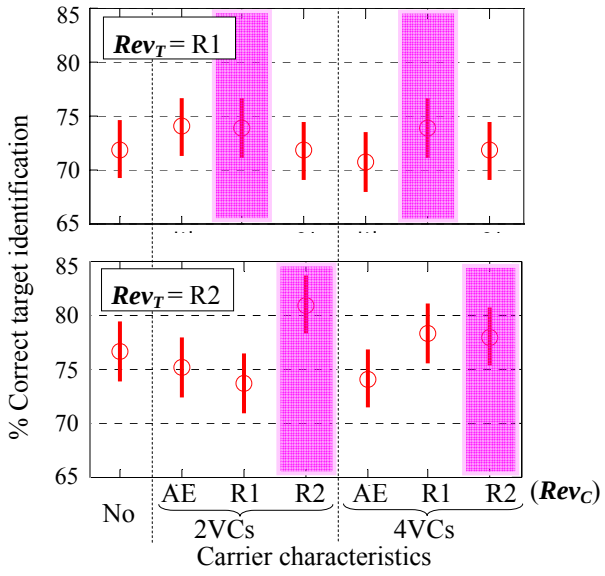


Figure 3: Result of the experiment 1

Across all conditions in which the target was R1 (Rev_T=R1), there were no significant differences. On the other hand, some significant differences were found when R2 was the target (Rev_T=R2). Specifically, when R2 was the target, the effect of carrier reverberation depended on the number of VCs in the carrier (2 VCs or 4 VCs). With 2 VCs in the carrier, performance with matched reverberation (Rev_C=Rev_T=R2) was significantly higher than both of the unmatched conditions (Rev_C=AE and R1), supporting our original hypothesis that matched reverberation can improve speech intelligibility. On the other hand, with 4 VCs in the carrier, there were no significant effects of carrier reverberation.

4 Experiment 2

We hypothesized that the inconsistent results in Experiment 1 might have been caused by subjects using different listening strategies when the carrier was long (four VCs) compared to when it was short and listeners did not know which VC was the target until after the presentation ended (two VCs). In order to both confirm the effects seen in Experiment 1 and determine if the random-length trial influenced results, a second experiment was conducted.

4.1 Design

We focused on the cases where significant differences were seen in Experiment 1 and held the number of VCs in the carrier fixed throughout blocks of trials. To reduce "wasted" test time, six of the original consonants (k, t, n, z, s and ʃ) were not presented as the target, as they were nearly perfectly identified in Experiment 1 (percent-correct was higher than 90% in all conditions). However, the same graphical interface with 16 buttons was used to collect subject responses (i.e., subjects were allowed to respond with a target consonant that was never presented). Reverberation of the target was fixed to R2 (Rev_T=R2); the carrier could be AE, R1, or R2. The carrier VCs could be any of the 16 original VCs (i.e., they were not restricted to the 10 more difficult VCs).

This design resulted in a total of 30 tokens (10 VCs x 3 talkers) for each of three carrier-reverberation conditions (Rev_C=AE, R1, and R2) and each of two carrier lengths (2 VCs or 4 VCs). Within each block, 90 trials (30 tokens x 3 conditions) were repeated twice in random order. Each block of 180 trials took approximately 20 - 25 min. The order of the two blocks was randomly assigned for each subject.

4.2 Results

Eleven native English speakers participated in the experiment.

Percent-correct identification scores are calculated for each condition for each repetition by each subject. The two-way ANOVA analysis was conducted with subject (with repetition) and condition as factors. The result of ANOVA showed that both subject and condition had significant effects on performance ($p<0.0001$). Figure 4 shows the result of the pair-wise comparisons for condition means.

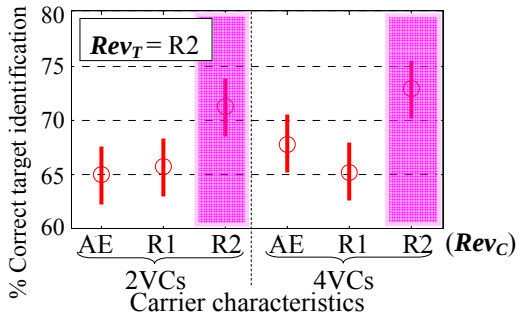


Figure 4: Result of the experiment 2

Consistent with Experiment 1, the carrier had a significant effect on percent correct performance in some cases. With two VCs in the carrier, the mean of the matched reverberation condition (Rev_C=Rev_T=R2) was significantly higher than for both unmatched conditions (Rev_C=AE and R1), consistent with results of Experiment 1. With four VCs in the carrier, the matched reverberation condition (Rev_C =R2) was also significantly better than for an R1 carrier (Rev_C=R1), and although it did not reach statistical significance, performance with the matched reverberation also tended to be better than with the anechoic carrier.

5 Discussion

In both Experiments 1 and 2, when the target had R2 reverberation, there was some benefit, in some conditions, of hearing a preceding carrier whose reverberation matched that in the target (i.e., Rev_C= Rev_T=R2). In both Experiments 1 and 2, when the carrier consisted of two VCs, performance was significantly better when the target reverberation matched the carrier than when the carrier was either anechoic or from R1. However, when the carrier consisted of four VCs, results in the two experiments differ. In Experiment 1, there were no significant differences across the three carrier conditions with 4 preceding VCs. In Experiment 2, performance was significantly better when the carrier reverberation was also from R2 than when it was from R1 and the matching condition was marginally better than when the carrier was anechoic. This result might be due to the differences in the design between Experiments 1 and 2. In Experiment 1, the listener had to pay attention to the carrier because the carrier length varied randomly on a trial-by-trial basis. In contrast, in Experiment 2, the listener could focus directly on the target VC, because the number of carrier VCs was known *a priori*.

When the target was simulated in R1, there was no effect of room calibration; performance was the same regardless of what reverberation was present in the carrier phrase. This puzzling result may be related to the fact that R1 is a more challenging environment (see Fig. 1) and yields worse consonant identification than R2 (see Fig. 2, especially for the no-carrier control condition). In this more challenging condition, it may be impossible to adjust perceptual processing to reduce the degradation caused by reverberation R1.

Further tests with additional subjects are necessary to strengthen and verify these results. However, even with this caveat, these results show that consistent experience with room reverberation can improve consonant identification, presumably by enabling a listener to calibrate to the effects of the room.

Acknowledgements

This work was supported by NSF grant SBE-0354378 and by The Kajima Foundation.

References

[1] B.G. Shinn-Cunningham, ‘Learning reverberation: Implications for spatial auditory displays.’ *Proc. Int. Conf. Aud. Display,* pp. 126-134 (2000).

[2] A. J. Watkins. "Perceptual compensation for effects of reverberation on amplitude-envelope cues to the ‘slay’-‘splay’ distinction." *Proc. Int. Congress Acoust,* Vol.14. pp. 125-132 (1992).

[3] Resnick, S.B., Dubno, J.R., Hoffnung, S., and Levitt, H. “Phoneme errors on a nonsense syllable test.” Journal of the Acoustical Society of America, 58, 114 (1075).

[4] E. W. Yund and K.M. Buckles, ‘Multichannel compression hearing aids: Effect of number of channels on speech discrimination in noise’, J. Acoust. Soc. Am. 97: 1206-1223 (1995).

[5] Steeneken, H.J.M. and Houtgast, T., ‘A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria’, Institute for Perception TNO, Soesterberg, the Netherlands. (1984)