Audiovisual training is better than auditory-only training for auditory-only speech-in-noise identification

Björn Lidestam^{a)}

*Department of Behavioral Sciences and Learning, Linköping University,
SE-581 83 Linköping, Sweden
bjorn.lidestam@liu.se*

Shahram Moradi

*Linnaeus Centre HEAD, Department of Behavioral Sciences and Learning,
Linköping University, SE-581 83 Linköping, Sweden
shahram.moradi@liu.se*

Rasmus Pettersson and Theodor Ricklefs

*Department of Clinical and Experimental Medicine, Division of Speech and Language
Pathology, Linköping University, SE-581 85 Linköping, Sweden
raspe397@student.liu.se, theri524@student.liu.se*

Abstract: The effects of audiovisual versus auditory training for speech-in-noise identification were examined in 60 young participants. The training conditions were audiovisual training, auditory-only training, and no training ($n=20$ each). In the training groups, gated consonants and words were presented at 0 dB signal-to-noise ratio; stimuli were either audiovisual or auditory-only. The no-training group watched a movie clip without performing a speech identification task. Speech-in-noise identification was measured before and after the training (or control activity). Results showed that only audiovisual training improved speech-in-noise identification, demonstrating superiority over auditory-only training.

© 2014 Acoustical Society of America

PACS numbers: 43.72.Dv, 43.71.Sy, 43.71.Gv, 43.71.Es [DDO]

Date Received: May 15, 2014 **Date Accepted:** June 18, 2014

1. Introduction

Moradi *et al.* (2013) reported that prior exposure to audiovisual (AV) speech stimuli resulted in better subsequent identification of auditory-only (AO)-presented speech in noise, compared to prior AO exposure. This happened even though the speech stimuli (gated consonants, words, and final words in sentences) were not similar to the subsequent speech-in-noise test stimuli [sentences were presented at a constant 0 dB signal-to-noise ratio (SNR) in the gating tasks, but at adaptive SNR in the Hearing In Noise Test (HINT)]. Further, the talker used for the gated speech tasks was different to that used for the subsequent speech-in-noise tasks.

Some studies have reported superior effects of AV training on subsequent AO speech identification, in terms of perceptual learning of noise-vocoded speech (Bernstein *et al.*, 2013; Kawase *et al.*, 2009), but no randomized controlled study has demonstrated the advantage of AV training relative to other training methods for AO speech-in-noise identification.

Shams *et al.* (2011) suggested three explanations to illustrate how prior AV experiences might boost subsequent unisensory processing. First, multisensory

^{a)} Author to whom correspondence should be addressed.

experiences quickly recalibrate unisensory maps in the brain. Second, a new connection between unisensory cortical areas in the brain is created. Third, unisensory representations of stimuli (i.e., AO or visual representation of speech) are integrated with those stimuli in a multisensory manner.

The motivation for the present study came from the findings of [Moradi *et al.* \(2013\)](#). In the present study, we replicated and extended the experimental design described by [Moradi *et al.* \(2013\)](#), using a randomized controlled design, in order to observe whether prior exposure to AV stimuli resulted in better speech-in-noise identification than prior exposure to AO stimuli. In [Moradi *et al.* \(2013\)](#), participants performed the gating paradigm speech tasks under both silent and noisy conditions in the first session, and in the second session they were assessed with cognitive tests and the HINT (a Swedish version of the speech-in-noise test; [Hällgren *et al.*, 2006](#)).

The present study aimed to test, in one session and as an extended replication of the procedure in [Moradi *et al.* \(2013\)](#), whether AV training had any beneficial effects over AO training. To accomplish this aim, we presented only gated speech consonants and isolated words, and only in noise (i.e., the gating procedure was not as extensive, meaning that exposure to AO and AV stimuli was not as long as in [Moradi *et al.*, 2013](#)). The rationale for presenting only gated consonants and words (but not also final words in sentences, as in [Moradi *et al.*, 2013](#)) was to emphasize bottom-up processing; that is, force participants to allocate attention to signal features. Additionally, the rationale for presenting stimuli under noisy conditions only was based on [Wayne and Johnsrude \(2012\)](#), who reported that AV training with clear auditory input did not improve perceptual learning in children with cochlear implants, whereas the presentation of AV stimuli with degraded auditory input did. In addition, we wanted to shorten the training phase to allow all testing to be carried out in only one session, to prevent potential fatigue (cf. [Wayne and Johnsrude, 2012](#)), even though there was a risk of smaller training effects as a result of shorter training periods for both modalities. We also wanted to control for the possibility that both AV and AO training could improve speech-in-noise identification. We therefore included a no-training (NT) control group; this group was not exposed to speech identification training between the repeated measurements of speech-in-noise identification.

[Wayne and Johnsrude \(2012\)](#) reported that the superior effect of AV training for noise-vocoded speech is most probably due to the additional listening effort for comprehension of noise-vocoded speech and not due to the addition of visual cues to the auditory-input training. In order to test this hypothesis, we measured self-rated effort during AV and AO training. An effect of training modality (AO vs AV gating tasks) would suggest that listening effort (and resulting fatigue) affected subsequent AO speech-in-noise identification.

2. Methods

2.1 Participants

A total of 60 young university students from Linköping University participated in this study [27 males and 33 females, $M = 23.2$ yrs, standard deviation (SD) = 2.4 yrs]. The participants were randomly allocated to 3 groups (AV, AO, and NT) with 20 subjects in each (11 females and 9 males in each). The mean ages of participants in the AO, AV, and NT groups were 23.5, 23.2, and 22.7 yrs, respectively. Participants were rewarded with cinema tickets for their participation in this study. All were native Swedish speakers who reported having normal hearing and vision (or corrected-to-normal vision).

2.2 Design

A 3×2 mixed design with Training (AV, AO, and NT) as a between-groups variable, and Occasion (before vs after training) as a within-groups variable was used. In order to balance out potential ordering effects in the speech identification training tasks, ten presentation orders were used, each with two participants per group. Accordingly,

different lists in the HINT were used and balanced out such that all lists were presented both before and after the training phase, and two other lists were used for the familiarization trials. All three training groups were tested in a random order during the same period of time (e.g., one participant per group in one and the same day), to avoid potential time-cohort effects.

2.3 HINT

A Swedish version of the HINT (Hällgren *et al.*, 2006) was used to measure speech-in-noise identification ability. The HINT consisted of everyday sentences that were three to seven words long, presented in random noise with the same long-term spectral properties as the speech signal, in the adaptive procedure described below. The test comprised 2 practice lists (10 sentences each) and 10 experimental lists (20 sentences each). The sentences, which were prerecorded using a female talker, were presented via earphones. The first sentence in each list was presented at 67 dB (0 dB SNR). The noise presentation level varied according to each participant's response. For each sentence, participants were instructed to listen and repeat as many of the words as possible. The experimenter recorded the number of words correctly repeated on a computer terminal (which was out of sight of the participant).

An automatic, adaptive up-down procedure determined the SNR: If all words were correctly repeated, the SNR was lowered by 2 dB, and if one or more words were not correctly repeated, the SNR was raised by 2 dB. Participants were first familiarized with the test using a 10-sentence practice list. To determine the SNR for each participant, a 20-sentence list was used. This procedure was followed both before and after training, with different lists each time, to avoid repetition priming.

The HINT took approximately 10 min to complete. The HINT scores were adjusted to calculate an individual SNR for a correct response rate of 50%.

2.4 Speech identification training tasks

A detailed description of the gated stimuli used in the present study can be found in Moradi *et al.* (2013); whereas Lidestam (2014) provided details regarding the synchronization of auditory and visual input for a coherent AV speech signal used in both Moradi *et al.* (2013) and in the present study. A brief description is provided here.

For AO and AV training, participants were presented with AO or AV gated Swedish consonants and words under the same speech-spectrum matched noise as in the HINT (Hällgren *et al.*, 2006). These were presented at 0 dB SNR at 75 dB sound pressure level (long term average), in ten fixed quasi-random presentation orders that were balanced out. Because the stimuli were short in duration, there were considerable differences in sound pressure level from onset to offset (i.e., depending on how much of a specific stimulus was presented as a function of gates). This relatively loud long-term average sound pressure level was set following pilot testing of consonants. Relatively loud sound pressure level was required for avoiding potential floor effects for consonants since they were generally perceived as less loud than the surrounding vowels in the /aCa/ format.

In the AV presentation, the visual stimuli presented the whole face as well as the neck and shoulders of the talker, displaying movement of the larynx. AV asynchrony was established to be less than 1 ms (i.e., better synchrony than available from AV presentation with standard recording and playback, see Lidestam, 2014).

2.4.1 Consonants

Eighteen Swedish consonants were presented in a vowel-consonant-vowel (/aCa/) syllable format (/aba, ada, afa, aga, aja, aha, aka, ala, ama, ana, aña, apa, ara, aña, asa, afa, ata, ava/). The gate size for consonants was set at 16.67 ms. Gating started after the first vowel /a/ and at the beginning of the consonant onset. Hence, the first gate included the vowel /a/ plus the initial 16.67 ms of the consonant, the second gate provided an additional 16.67 ms of the consonant (a total of 33.33 ms), and so on. The consonant gating task took 10–15 min to complete.

2.4.2 Words

We employed 23 Swedish monosyllabic words in a consonant- vowel-consonant format (all nouns). Similar to [Moradi *et al.* \(2013\)](#), the gate size for words was set at 33.3 ms. The words used in the present study had a small-to-average number of neighbors. The word-gating task took 20–25 min to complete.

2.5 NT control group

The participants in the NT control group watched a 28-min movie clip (Björnens rätta ansikte [Attacking bears]; Wildlifefilm.com), which is a documentary about the circumstances under which a bear may attack. The participants in the NT group were instructed to watch this movie clip having been informed that it was a test of memory and attention. After viewing the movie clip, participants responded to six questions about facts presented in the clip.

2.6 Procedure

Participants were tested individually in a quiet room. They sat in front of a Mitsubishi Diamond Pro 2070 SB CRT monitor (Mitsubishi Electric, Japan). The monitor was turned off during AO presentation. All auditory stimuli were delivered binaurally in mono via Sennheiser HDA200 earphones. In the HINT and gated speech identification tasks, the participants were instructed to respond aloud which item they identified (or guessed), and the experimenter recorded their verbal responses.

In the gated speech identification tasks, one stimulus at a time was presented, with increasing duration as the gate moved one step forward for each new presentation. The noise was presented with the same onset and offset as the audio speech signal. In the AV training condition, the first (still) image was presented as a cue for playback for half a second. Immediately after playback the screen turned gray and remained so between presentations. There were no visual cues in AO training. The presentation of gates continued until the target was correctly identified on six consecutive presentations; if not, the presentation continued until the entire target was disclosed. Thus, there was implicit feedback about performance.

After the training phase (i.e., the AO or AV gating tasks) the participants rated the effort required for the training tasks on a questionnaire with a visual analog scale ranging between 0 (no effort) and 100 (maximum effort). The NT control group also rated their effort, in order to keep the procedure as constant as possible for all three groups.

3. Results

The scores were first checked for potential outliers, defined as more than 3 *SDs* of the respective group means, and more than 0.5 *SD* beyond the closest score. One participant in the AV group scored more than 3 *SDs* above than the AV group mean at before training (which would have potentially exaggerated the effect of AV training). This score was therefore adjusted to 0.5 *SD* above the second poorest result within the AV group. (Note: Since the output of the HINT was SNR, a higher SNR meant poorer speech identification.)

The grand mean for speech-in-noise test performance before training was identical for the participants in the present study compared to the norm data in [Hällgren *et al.* \(2006\)](#), $M = -3.0$ dB SNR, and with similar variation ($SD = 0.86$ dB vs $SD = 1.1$ dB, $F[9, 59] = 1.64$, $p > 0.05$). This suggests that all our participants indeed had normal hearing.

Figure 1 shows the mean scores and standard errors for HINT performance, before and after the training phase, for each group. A 3 (Training [AO, AV, and NT]) $\times$ 2 (Occasion [before vs after training]) mixed analysis of variance, with repeated measures on the second factor, was conducted to examine differences between the groups on HINT scores before and after interventions.

The results showed that the main effect of Training was not significant ($F[2, 57] = 2.18$, $p = 0.12$, $\eta_p^2 = 0.07$). However, the main effect of Occasion was

Fig. 1. HINT speech-in-noise identification results ($M \pm SE$) for the three groups, before and after the speech identification training phase.

significant ($F[1, 57] = 4.38$, $p = 0.04$, $\eta_p^2 = 0.07$). There was an interaction of Training $\times$ Occasion ($F[2, 57] = 3.64$, $p = 0.03$, $\eta_p^2 = 0.11$). Tests of simple effects of Occasion per group (i.e., before vs after training) did not reach significance in either the AO ($F[1, 114] = 0.28$, $p > 0.05$) or the NT group ($F[1, 114] = 0.83$, $p > 0.05$). However, the effect of Occasion was significant in the AV group ($F[1, 114] = 6.47$, $p < 0.05$, $\eta_p^2 = 0.11$). Tests of simple effects of group on Occasion showed that the effect of group on HINT performance before training was not significant ($F[2, 114] = 0.87$, $p > 0.05$); however, the effect of group on HINT performance after training was significant ($F[2, 114] = 4.44$, $p < 0.05$, $\eta_p^2 = 0.16$). Tukey honest significant difference *post hoc* tests showed that the AV group tolerated significantly more noise than both the AO and NT groups, but the AO and NT groups did not differ from each other. Comparing mean performance after AV versus AO training, the effect size was $d = 0.72$; this was reported as $d = 1.09$ in Moradi *et al.* (2013).

Table 1 presents the results for self-ratings of effort. As a planned comparison, we tested whether self-rated effort differed between the AO and AV groups, but it did not ($t[38] = 0.48$, $p = 0.64$).

4. Discussion

The results of the present study showed that prior presentation with AV training resulted in better performance in a speech-in-noise test; this effect was not observed following AO training. This finding corroborates the results of Moradi *et al.* (2013) and supports the notion that preceding AV training results in better speech-in-noise identification when compared with AO training. Most importantly, the findings from the present study and Moradi *et al.* (2013) highlight that the superiority of AV training in improving speech-in-noise identification is independent of the idiosyncrasy of talkers in

Table 1. Group means and standard deviations for self-ratings of effort.

Training group	Self-rated effort	
	<i>M</i>	<i>SD</i>
AO	70.7	14.9
AV	73.3	19.3
NT	40.2	26.5

AO = auditory only; AV = audiovisual; NT = no training.

pre-post training, and of task similarity. Further, there was no difference in self-rated effort, which suggests that the superior effect for the AV training group was not due to fatigue or larger processing demands for the AO training group. In contrast to [Wayne and Johnsrude \(2012\)](#), our results indicate that the superior effect of AV training on subsequent speech-in-noise identification is solely due to the addition of visual cues to the auditory stimuli, and not to higher cognitive processing demands. In addition, [Moradi *et al.* \(2013\)](#) reported that gated AV speech identification in noise was not correlated with cognitive capacities of participants (e.g., working memory and attention).

The results fit well with the explanations suggested by [Shams *et al.* \(2011\)](#): Bimodal AV exposure to speech may reorganize and activate networks in the brain, allowing faster and stronger lexical activations for subsequent unimodal speech cues. An alternative explanation as to why this relatively short AV training resulted in improved speech-in-noise identification may be that we presented the AV speech stimuli in a gating paradigm format that helped to tune and allocate listeners' attention to the phonetic features of speech (see also the "reverse hierarchy theory," discussed in [Bernstein *et al.*, 2013](#)). The smaller effect size for the post-training speech-in-noise identification results in the present study compared to [Moradi *et al.* \(2013\)](#) may be partly due to a shorter training phase (i.e., shorter time on the gating tasks), or to lack of exposure to sentence stimuli (since the criterion was perception of sentences).

Further studies are needed to examine and test specific hypotheses about the processes behind cross-modal AV facilitation, and to explore how long the facilitatory effect lasts and determine whether the effects vary depending on the population tested (e.g., younger vs older, normal-hearing vs hearing-impaired).

The present study had some limitations as the hearing thresholds, lip reading ability, and auditory working memory were not assessed. Even though participants were randomized to training groups and sampled from a relatively homogeneous population, individual differences in such capacities may skew samples and thereby drive interactions.

Finally, the results of the present study cannot readily be generalized to other populations than young and normally hearing individuals. However, the superior AV training effect for improved speech-in-noise identification may inspire speech pathologists or audiologists to use AV training in aural rehabilitation for children with cochlear implants or individuals with hearing loss who have difficulty with speech perception in noisy conditions, preferably as clinical trials for potential publication.

Acknowledgments

This research was supported by the Swedish Research Council (Grant No. 2006-6917). The authors thank Björn Lyxell and Jerker Rönnberg for valuable comments on the manuscript.

References and links

- Bernstein, L. E., Auer, E. T., Jr., Eberhardt, S. P., and Jiang, J. (2013). "Auditory perceptual learning for speech perception can be enhanced by audiovisual training" *Front. Neurosci.* **7**, 34.
- Hällgren, M., Larsby, B., and Arlinger, S. (2006). "A Swedish version of the Hearing In Noise Test (HINT) for measurement of speech recognition," *Int. J. Audiol.* **45**, 227–237.
- Kawase, T., Sakamoto, S., Hori, Y., Maki, A., Suzuki, Y., and Kobayashi, T. (2009). "Bimodal audio-visual training enhances auditory adaptation process," *NeuroReport* **20**, 1231–1234.
- Lidestam, B. (2014). "Audiovisual presentation of video-recorded stimuli at a high frame rate," *Behav. Res. Methods*. **46**, 499–516.
- Moradi, S., Lidestam, B., and Rönnberg, J. (2013). "Gated audiovisual speech identification in silence vs. noise: Effects on time and accuracy," *Front. Psychol.* **4**, 359.
- Shams, L., Wozny, D. R., Kim, R., and Seitz, A. (2011). "Influences of multisensory experience on subsequent unisensory processing," *Front. Psychol.* **2**, 264.
- Wayne, R. V., and Johnsrude, I. S. (2012). "The role of visual speech information in supporting perceptual learning of degraded speech," *J. Exp. Psychol. Appl.* **18**, 419–435.
- WildlifeFilm.com. "Björnens rätta ansikte [Attacking bears]", Välkommen till wildlifeFilm.com, <http://www.wildlifeFilm.com> (Last viewed July 10, 2014).