Same or Different? Perceptual Learning for Connected Speech Induced by Brief and Longer Experiences

Karen Banai (kbanai@research.haifa.ac.il)
University of Haifa
Hanin Karawani
University of Haifa
Limor Lavie
University of Haifa
Yizhar Lavner
Tel-Hai Collegel

Research Article

Keywords: Perceptual Learning, Connected Speech, Longer Experiences, long-lasting changes, environment, rapid learning, stabilize learning

Posted Date: October 6th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-951041/v1

Abstract

Perceptual learning, defined as long-lasting changes in the ability to extract information from the environment, occurs following either brief exposure or prolonged practice. Whether these two types of experience yield qualitatively distinct patterns of learning is not clear. We used a time-compressed speech task to assess perceptual learning following either rapid exposure or additional training. We report that both experiences yielded robust and long-lasting learning. Individual differences in rapid learning explained unique variance in performance in independent speech tasks (natural-fast speech and speech-in-noise) with no additional contribution for training-induced learning (Experiment 1). Finally, it seems that similar factors influence the specificity of the two types of learning (Experiment 1 and 2). We suggest that rapid learning is key for understanding the role of perceptual learning in speech recognition under adverse conditions while longer learning could serve to strengthen and stabilize learning.

Introduction

Connected speech recognition under adverse conditions (e.g., distortion, background noise) [1], improves substantially following brief experiences and prolonged practice [2-9]. These improvements reflect perceptual learning, defined as relatively long-lasting changes in the ability to extract information from the environment following experience or practice [10,11]. An open question is whether rapid learning following brief experiences (rapid learning) and the learning that emerges with more intensive training (training-induced learning) reflect the same type of learning. Whereas some view only training-induce learning as true perceptual learning and refer to rapid learning as procedural or task learning, others view rapid learning as perceptual as well, because it shares some characteristics with training-induced learning [12,13]. In the case of speech, the literature portrays a complex picture. Both rapid and training-induced learning of speech stimuli are usually considered perceptual [e.g., 1], but the degree to which they share characteristics like stimulus specificity and contribute to dynamic speech perception are not unanimously agreed on. Furthermore, rapid and training-induced learning were typically studied in different studies, with different training methods, stimuli and learning tests, making it hard to compare outcomes across studies.

Understanding the similarities and differences between rapid and training-induced learning has important implications for the role of perceptual learning in speech perception under challenging conditions. Specifically, if rapid learning is both generalizable and long-lasting, brief experiences or short training episodes could gradually re-shape the perception of distorted speech, leading to a general increase in perception as a function of experience. On the other hand, if rapid learning is as stimulus specific as training-induced learning, past learning of either type is unlikely to shape future speech perception because future conditions are unlikely to be an exact replication of the past. Rather, rapid learning could support perception in challenging conditions online by allowing listeners to quickly adapt to the acoustic characteristics of the current situation [14,15]. Here we focus on rapid learning of distorted (time-compressed) speech. In Experiment 1 we compared learning and retention between rapid and training-induced learning. We also compared how the two types of learning relate to perception in independent

challenging speech tasks. In Experiment 2 we compared four protocols of rapid learning to determine whether rapid learning is as stimulus specific as found in previous studies on training-induced learning [16].

The Potential Role of Perceptual Learning in Speech Recognition

Theories of both perceptual learning [17] and speech processing [18,19] suggest that encounters with speech input trigger an implicit and largely automatic process which attempts to match this input to long-held representations. However, in daily listening situations inputs do not automatically match long-term representations (e.g., due to noise or accent), and the automatic matching process can therefore fail. According to the Reverse Hierarchy Theory [RHT, 17], such failure can trigger a learning process that gradually allows listeners to resolve finer-grained acoustic details and help them recognize previously unrecognizable input. However, because learning is triggered by a specific input, learning is at least partially specific to the acoustics of the input [7,17,18]. This specificity probably constrains the role of learning in complex communication environments. One option is that intensive experience is required to yield learning that supports speech recognition. However, training-induced learning of challenging speech is often quite specific to the trained stimuli [20-23]. Therefore, it can support future speech perception only to the extent that newly encountered situations replicate the conditions encountered in training which is unlikely. Therefore, intensive training studies are not a good analogue for real life conditions when a practice period is unlikely and the acoustics can change rapidly (e.g., in a multi-talker conversation). Consistent with this view, training in groups of listeners who need them most (e.g., due to hearing impairment) often fails to yield quantifiable benefits in any untrained conditions, despite good learning on the trained ones [24-26]. Studies on learning new speech categories [e.g., 27,28] are also not a good approximation for daily environments because they usually do not use connected speech.

Another potential role of perceptual learning which we pursue here is based on rapid learning: if learning occurs rapidly, it could serve as a skill listeners can recruit whenever they encounter new acoustic challenges. Accordingly, specific learning could afford optimal adaptation to the particulars of a new acoustic challenge without more general and undesirable changes in speech perception. Rapid learning studies are more representative of real-world challenges than training studies, because they often include little stimulus repetition and connected speech materials [4,5,29-33]. Therefore, this account is more ecological than accounts based on the generalization of past learning. Consistent with the idea that perceptual learning is a general resource, recent findings show that learning is correlated across different tasks and even across modalities [34-36].

Rapid and Training-Induced Learning of Distorted Speech

Direct comparisons of learning that follows different training or exposure durations have been rare and did not include the conditions required to determine whether differences in outcomes are quantitative or qualitative [37-39]. On the one hand, improvements that follow either brief exposure or training are both maintained over time, as required by the definition of perceptual learning [16,27,40]. On the other hand, one of the hallmarks of perceptual learning is its exquisite specificity to the physical attributes of the

trained stimuli [41,42]. Whereas speech learning following training appears quite specific to the acoustics of the trained stimuli [20,22,24,33,37,38,43], learning following brief exposure is thought to generalize more broadly [32,44,45]. For example, whereas rapid learning of time-compressed speech resulted in improved recognition of natural fast speech [46], no such transfer was observed after more intensive training [16,33]. Methodological differences make the outcomes hard to compare across studies. Therefore, one goal of the current study was to test the talker specificity of rapid learning of time-compressed speech across different learning protocols (short and long) and test times (immediate and delayed).

Overview of the Current Study

We conducted two experiments using a time-compressed speech task to elicit learning. In ***experiment 1***, we compared learning and retention between rapid and training-induced learning of time compressed speech. We also asked whether the two types of learning are differentially correlated with speech recognition in independent tasks - speech in noise and natural fast speech. We report that learning of time-compressed speech is associated with the perception of natural fast speech and speech in noise, with no apparent differences between rapid and training-induced learning. ***Experiment 2*** explored the effects of stimulus repetition and talker variability on rapid perceptual learning of time-compressed speech. Outcomes were compared to those of previous studies on learning following longer training protocols to suggest that the pattern of learning and specificity does not change between brief and prolonged training.

Experiment 1

Methods

Participants

160 university students or recent graduates (ages 18-35 years, Mean = 26, SD = 3, 91 female and 69 male) participated in this experiment. Participants were volunteers and reported they were native Hebrew speakers, with normal hearing and no history of attention, learning or language deficits and no experience with time-compressed speech. The study was performed in accordance with the declaration of Helsinki. All aspects of the study were approved by the ethics committee of the Faculty of Social Welfare and Health Sciences at the University of Haifa (permit #199/12). Informed consent was obtained from all participants. Participants were tested as described below; no other tests were conducted.

Participants were divided randomly to two groups, an exposure group (to assess rapid learning) and a training group (to assess training-induced learning) as explained below. Both groups completed two test sessions on separate days, in which they performed a time-compressed speech recognition task. At the end of the first session, the training group completed an additional training phase as described below. We note that parts of the data from the exposure group were previously published as part of a conference proceedings [47], and re-analyzed for the purpose of the current study. One participant had missing data

and was not included in data analysis, so we report data from 79 listeners in the *exposure* group (age: Mean = 26, SD = 4; 38 female, 41 male) and 80 listeners in the *training* group (ages: Mean = 26, SD = 3; 52 female, 28 male).

Overall Design

The experiment comprised of two sessions, 5 to 9 days apart. On each session, participants from both groups completed three speech recognition tests – time-compressed speech, natural-fast speech and speech-in-noise, in a counterbalanced order as described below. The training group received additional training on time-compressed speech at the end of the first session. Participants completed the experiment in a quiet room on campus or in their homes. Stimuli were delivered diotically through headphones (Sennheiser HD-205 or HD-215) at a comfortable listening level, using costume software [22]. The time-compressed speech task was used to assess learning both within (rapid learning) and between (retention or consolidation) sessions. Comparison between the exposure and training groups was used to assess differences between rapid learning induced by the time-compressed speech task and training-induced learning. The other two tasks were used to determine if perceptual learning of one type of speech is related to recognition of other types of challenging speech.

Stimuli and Tasks

Stimuli. 290 simple sentences in Hebrew [based on 48], were used. Sentences were five to six words long and had a subject-verb-object grammatical structure. Half of the sentences were semantically plausible (e.g., “the talented poet wrote a poem”) and half the sentences were implausible (e.g., “the angry shopkeeper fired the rabbit”).

Stimuli for the speech-in-noise and time compressed speech tests were recorded by talker 1, a female native speaker of Hebrew with an average speech rate of 111 words/min (SD = 17). Stimuli for the natural-fast speech test were recorded by talker 2, at an average natural-fast rate of 214 words/minute (SD = 26) because pilot testing suggested that natural-fast speech by talker 1 was not fast enough to challenge university students who are native speakers of Hebrew. Sentences were recorded in a sound attenuating room at a sampling rate of 44.1 kHz, with a standard microphone and edited in Audacity® software© 2.1.3 to remove remaining noise and equate root-mean-square (RMS) amplitude across sentences.

Speech Recognition Tests. Sentences were randomly divided across the different tests such that on each test half the sentences were plausible and half were implausible. Different sentences were used on each test and session. Order of presentation was random but fixed across participants, with no sentence repetition. Sentence delivery was self-paced. Participants were asked to transcribe the sentences as accurately as they could, and the number of correctly transcribed words was counted for each sentence. Only perfectly transcribed words (ignoring homophonic spelling errors) were counted as correct. The proportion of correct words per sentence was used as an index of recognition accuracy.

Speech-in-Noise Tests. On each session participants had to transcribe 25 different sentences. Sentences produced by talker 1 were mixed with 4-talker babble noise [taken from 22] at a signal-to-noise ratio of −6 dB.

Natural-fast speech tests. On each session participants had to transcribe 20 different sentences produced by talker 2.

Time-Compressed Speech Tests. On each session participants transcribed 10 sentences produced by talker 1. To afford isolation of the rapid learning effects, we used the minimal number of sentences thought to yield rapid learning in the majority of participants based on previous work [15]. Sentences were compressed to 30% of their natural duration using a WSOLA algorithm [49].

Training. Three blocks of 60 sentences each produced by talker 1 were delivered. In the first block, participants had to transcribe sentences compressed to 30% of their natural duration, as described above. The additional two blocks were adaptive. For each sentence participants had to determine whether it was semantically plausible or not. Initial compression was 50%. Subsequently a 2-down/1-up staircase procedure was used to adjust compression based on participants' responses. This procedure was used to give participants extra training without overburdening them. Because the main goal of the study was to determine whether training-induced and rapid learning differed in their relationships with other types of challenging speech, data from the training phase itself was not analyzed.

Data Analysis

Recognition accuracy data were analyzed in R [50] with a series of generalized linear mixed models using the lme4 package [51].

Learning Analysis. We used data from the time-compressed speech tests to assess rapid perceptual learning within and between sessions as well as training-induced learning (see Results). Learning between the two test sessions was our main index of learning. To this end, for each participant the proportion of words correctly transcribed across all sentences within a session was averaged and the difference between the averages of the two sessions served as a learning index. For the exposure group, this is an index of the rapid learning induced by completing the tests. For the training group, the value is a mixture of the rapid learning that occurred during the tests and the additional contribution of training-induced learning. Group effects in the statistical models described in the Results were used to statistically separate rapid and training-induced learning. Within-session learning across sentences was also modeled to further assess rapid learning and how it may interact with training-induced learning.

Results

Rapid, Exposure-Induced Learning Conforms to the Definition of Perceptual Learning

Time-compressed speech recognition in the two groups and sessions is shown in Figure 1. In the exposure group, mean recognition accuracy was 0.20 (SD = 0.14) in session 1, and 0.33 (SD = 0.21) in

session 2. In the training group, mean accuracy was 0.26 (SD = 0.18) in session 1 and 0.47 (SD = 0.22) in session 2. Our first goal was to determine whether learning of time-compressed speech occurred between the two sessions and whether it differed between the two groups. Learning, defined as the amount of improvement on time-compressed speech recognition accuracy between the two sessions, is also shown in Figure 1. This figure suggests that recognition accuracy of the majority of participants in both groups improved between the two sessions.

To determine whether this learning was significant, and whether it was modulated by additional practice, mixed modelling was conducted. Random effects included random intercept for participants, as well as a sentence by participant random slope to account for the possibility that rapid learning rates (changes in accuracy over sentences) vary across participants. Fixed effects included group (*exposure*, dummy coded as 0 and *training*, coded as 1), sentence number (coded 1 to 10) and session (session 1 coded 0 and session 2 coded as 1). A binomial regression with logistic link function was used (as recommended by Dunn & Smyth, 2018 for proportion data). Three models were constructed. A model that included the random effects only (AIC = 11485), a model with additional main effects for each of the three fixed factors (AIC = 10670), and a "full" model that included all possible interaction terms between the fixed factors (AIC = 10558). Model comparison (using anova) suggested that the model with main effects fit the data significantly better than the model with random effects only ($\chi^2_{(3)} = 821$, $p \leq 0.001$) and the full model fit the data better than the model with only main effects ($\chi^2_{(4)} = 120$, $p \leq 0.001$).

The effects in the full model (see Table 1) were used to determine whether learning occurred, whether it was maintained over time and whether it differed between the two groups. As expected from previous studies, a significant main effect of sentence was present, confirming that rapid learning of time-compressed speech occurred within session. Furthermore, overall performance was more accurate in the second session (main effect of session), and between-session improvements were even larger in the training group (significant group by session interaction), suggesting that training yielded additional learning between sessions. On the other hand, the effect of rapid learning itself (that is change over sentences within a session) was smaller in the second session (significant sentence by session interaction). Although figure 2 suggests that the magnitude of decline in rapid learning between sessions could have been larger in the training group, the group by session by sentence interaction was not significant.

Table 1. Fixed effects and interactions from the full learning model

Effect	β	Standard Error	Z	p
group	0.43	0.24	1.8	0.072
session	1.34	0.13	10.69	≤ 0.001
sentence	0.19	0.02	10.89	≤ 0.001
group x session	0.61	0.17	3.61	≤ 0.001
group x sentence	-0.01	0.02	-0.53	0.594
sentence x session	-0.10	0.02	-5.00	≤ 0.001
group x session x sentence	-0.05	0.03	-1.95	0.051

To help interpret the effects from the statistical model, within session learning is presented in Figure 2. Each listener transcribed 10 (different) time-compressed sentences on each session, and for the purpose of visualization, learning was defined as the difference in transcription accuracy between the final and first 5 sentences in a session. This figure suggests that consistent with the sentence by session interaction, rapid learning rates were similar in the two groups during the first session (Mean = 0.14 and 0.15; SD = 0.14 and 0.15 in the exposure and training groups respectively), and that rapid learning in the second session was reduced in both groups. Furthermore, while not statistically significant in the full model, inspection of the within session learning data suggests that whereas rapid learning during the second session was observed in 56/79 participants in the exposure group (with a median of 0.06 and interquartile range from 0 to 0.13), only 41/80 participants in the training group continued to improve during session 2 (Median = 0.003, IQR = -0.087 to 0.087; χ^2 = 6.44, p = 0.011).

Taken together, the time-compressed speech data suggests that rapid learning occurred during the first test session, and to a lesser extent during the second session; additional training resulted in additional learning. Furthermore, rapid learning was maintained between sessions, conforming to the definition of perceptual learning.

Rapid Learning is Responsible for the Relationships between Perceptual Learning and Speech Recognition

A second goal was to determine whether perceptual learning of time compressed speech was associated with speech perception in independent tasks (natural fast speech and speech in noise), and if so, whether rapid- and training-induced learning differed in this respect.

Speech perception in the two groups and sessions is shown in Figure 3 (Natural-fast speech: Session 1: Mean = 0.86, SD = 0.10 and Mean = 0.85, SD = 0.09; Session 2: Mean = 0.89, SD = 0.09 and Mean = 0.89, SD = 0.11 in the exposure and training groups, respectively; Speech-in-noise: Session 1: Mean = 0.41, SD = 0.15 and Mean = 0.46, SD = 0.20; Session 2: Mean = 0.30, SD = 0.15 and Mean = 0.35, SD = 0.16 in the exposure and training groups, respectively). Similar performance in the two groups in session 2 would

suggest that learning on time-compressed speech did not generalize to these other tasks and thus that any associations between between-session learning on time-compressed speech and session 2 speech perception (i.e. natural-fast speech and speech in noise) could only arise due to rapid learning. Therefore, speech perception data in each of the tasks was modelled as a function of group, session and group by session interaction as fixed effects and random intercepts for participants and individual sentences. Model comparison suggests that the model with all fixed effects (AIC = 13832) is a better fit to the natural-fast speech data than the model with random effects only (AIC = 13939; $\chi^2_{(3)} = 112$, $p \leq 0.001$). The fixed effects (see Table 2) suggest that natural-fast speech recognition was more accurate in session 2, but as both the group effect and the session by group interaction were insignificant, this is not due to generalization of training-induced learning in the training group. Similarly, for speech in noise, model comparison suggests that the model with fixed effects (AIC = 25334) fit the data better than the model with random effects only (AIC = 25959; $\chi^2_{(3)} = 631$, $p \leq 0.001$). Although speech-in-noise recognition was poorer in session 2 than in session 1, there was no indication that this is due to training (see Table 2). Therefore, session 2 speech perception data were used in the following analyses to assess the association between perception and learning.

Table 2. Natural-fast speech and speech-in-noise perception as a function of group and session

	Natural-fast speech				Speech-in-noise			
Effect	β	SE	Z	p	β	SE	Z	p
group	-0.03	0.13	-0.24	0.808	0.23	0.14	1.66	0.097
session	0.37	0.05	7.59	≤ 0.001	-0.59	0.03	-18.51	≤ 0.001
group x session	-0.01	0.07	-0.21	0.830	0.07	0.04	1.58	0.11

Speech recognition is plotted in Figure 4 as a function of perceptual learning. To determine how perceptual learning contributed to speech recognition in the two tasks, data was again modelled with mixed-effects binomial regression with a logistic link function. For each speech task, the following models were constructed: (1) a “random” model with random intercepts for participant and sentence; (2) a “main effects” model which included three additional main effects: group (exposure coded as 0 and training coded as 1), perceptual learning (the difference between session 2 and session 1 as plotted in Figure 1) and baseline recognition of time-compressed speech (mean of the first 5 sentences from session 1); the two continuous predictors were scaled and (3) an “interaction” model in which the group by learning interaction was also included. Model comparisons (anova) were used to determine whether the “main effects” model fits the speech data better than the model with random effects only. Then the “main effect” and “interaction” models were compared to determine if the contribution of perceptual learning to speech perception differed between the trained and the exposure groups.

For natural-fast speech, the “main effects” model (AIC = 5991) significantly improved data fit over the random effects (AIC = 6030, $\chi^2_{(3)} = 45$, $p \leq 0.001$). Adding the group by learning interaction in the

interactions model had no significant effect (AIC = 5993, $\chi^2_{(1)} = 0.69$, p = 0.406, see Table 3 for the parameters of the best fit model).

For speech-in-noise, the "main effects" model (AIC = 11538) fit the data significantly better than the model with random effects only (AIC = 11586, $\chi^2_{(3)} = 54$, $p \leq 0.001$). Addition of the group by learning interaction had no significant impact on the fit (AIC = 11539, $\chi^2_{(3)} = 0.77$, p = 0.381, see Table 3 for the parameters of the best fit model). Therefore, it seems that additional training did not significantly change the contribution of rapid perceptual learning of time-compressed speech to either natural-fast speech or speech-in-noise recognition. Taken together, the current data replicates our previous findings [15] in showing that rapid perceptual learning contributes to speech recognition in independent tasks. Furthermore, the current findings suggest that this contribution does not change with training and is not attributable to the generalization of learning. Because learning was assessed across sessions, the present findings also suggest that the learning/perception correlations reflect "true" perceptual learning and not a transient effect.

Table 3. Prediction models for speech recognition as a function of perceptual learning

	Natural-fast speech				Speech in noise			
Effect	β log-odds	SE	Z	p	β log-odds	SE	Z	p
group	-0.41	0.14	-2.82	0.005	0.04	0.13	0.34	0.737
learning	0.42	0.07	6.02	≤ 0.001	0.31	0.07	4.63	≤ 0.001
baseline	0.31	0.07	4.41	≤ 0.001	0.40	0.06	6.28	≤ 0.001

Experiment 2

Both talker variability and stimulus repetition were previously suggested to influence the specificity of perceptual learning for speech [52-54]. In previous training studies on time-compressed speech, learning was specific and neither of these factors influenced it [16,38]. Experiment 2 therefore explored the effects of repetition and talker variability on rapid learning of time-compressed speech and its talker specificity.

Methods

Participants

255 native Hebrew speakers (ages 18-35 years, Mean = 27, SD = 4; 153 females, 102 males) participated in this study. All other details are as in Experiment 1. Participants were randomly divided to five groups and tested as described below; No other tests were conducted.

Overview of the Experiment and Exposure Groups

Participants were assigned randomly to one of five groups, a 'no exposure' control group and four exposure groups. The exposure groups transcribed 20 time-compressed sentences in one of the following

conditions (see Table 4): ‘baseline’ (20 different sentences presented by a single talker), ‘multi-talker’ (the same 20 sentences presented by 5 different talkers such that each talker delivered 4 different sentences), ‘multi-repetition’ (four sentences presented by a single talker, each repeated five times), and ‘single sentence’ (one sentence presented 20 times by a single talker).

Table 4. Overview of the design

Group	Session 1			Session 2
	Exposure		Immediate Test	Delayed test
	# different sentences	# different talkers		
1 no exposure (control group)	0	0	10 TCS, talker 1 10 TCS, talker 2 10 NFS, talker 2	10 TCS, talker 1 10 TCS, talker 2 10 NFS, talker 2 NFS, talker 3 SIN, talker 1 Digit span Similarities Matrices
2 baseline	20	1		
3 multi-talker	20	5		
4 multi-repetition	4	1		
5 single sentence	1	1		

TCS = time-compressed speech sentences; NFS = natural-fast speech; SIN = speech in noise

On the first session, participants in the exposure groups transcribed 20 time-compressed sentences as described below (see Table 4). After the exposure, all five groups were tested on the time-compressed and natural-fast speech tests described below in a fixed order. On the second session (~7 days after session 1) they were again tested on the same tests (with different sentences), again in fixed order. Then they completed another natural-fast speech test, a speech in noise test and the matrices, digit-span and similarities subtests from WAIS-IV [55] in counterbalanced order.

Table 5. Group characteristics – Mean (SD)

	No exposure	Baseline	Multi-talker	Multi-Repetition	Single sentence
N	53	51	51	50	50
F:M	31:22	28:23	35:15	31:19	27:23
Age	27 (4)	27 (4)	27 (4)	26 (4)	26 (4)
Education (years)	14 (2)	14 (2)	14 (2)	14 (2)	14 (2)
Days between sessions	7 (1)	7 (1)	7 (1)	7 (1)	7 (1)
Digit span (scaled score)	12.6 (3)	12.2 (3)	11.5 (3)	12.1 (3)	11.3 (3)
Similarities (scaled score)	10.9 (2)	11.2 (3)	11.3 (2)	11.0 (2)	11.2 (2)
Matrices (scaled score)	14.1 (2)	14.2 (2)	13.9 (2)	13.7 (3)	13.8 (2)

Stimuli, Exposure and Test Conditions

120 simple sentences in Hebrew were used in this study (see Experiment 1 for further details). Sentences for the exposure condition were recorded with five native speakers of Hebrew (all female), including talkers 1 and 2 from Experiment1. Natural-fast speech was recorded by talker 2 and an additional talker 3 (220 words/minute SD = 21).

Exposure Conditions. In all exposure conditions listeners had to transcribe 20 sentences compressed to 30% of their original duration as in Experiment 1. No feedback was provided.

Baseline. 20 different sentences were presented by talker 1. This condition is similar to those used in past studies to document rapid learning of time-compressed speech[15,33,40].

Multi-talker. The same sentences as in the baseline conditions were presented by five different talkers, including talker 1 and talker 2. Although we found no effect of talker variability on the perceptual learning of time-compressed speech and its generalization in the past, this condition was included because past literature on other types of challenging speech suggests that talker variability can influence the transfer of learning (for review of past studies and our previous attempt see Tarabeih-Ghanayim et al. [16]).

Multi-repetition. Four sentences were selected randomly from the baseline condition and presented five times each by talker 1 in pseudo-random order such that a single sentence could not repeat on two successive trials.

Single sentence. To further probe the effects of stimulus repetition, a single sentence was randomly selected from the baseline condition and presented 20 times by talker 1.

Test Conditions. On each test session participants had to transcribe 10 time-compressed sentences presented by talker 1, 10 time-compressed sentences presented by talker 2 and 10 natural-fast sentences presented by talker 2 (Table 4). These tests were presented in a fixed order. In session 2, after completion of those tests, two other tests were carried out: in one test 20 natural-fast sentences recorded by talker 3 were presented; in the other 20 sentences recorded by talker 1 and embedded in background noise (as in Experiment 1) were also presented. In addition, three subtests from WAIS-IV (see Table 5). Those were presented in counterbalanced order.

Sentence Transcription. Across exposure and testing, presentation was self-paced. Listeners heard each sentence, transcribed it and continued to the next sentence by pressing a “continue” button on screen using custom software [22]. Each sentence was played once, and no feedback was provided.

Data Analysis

For each sentence, the proportion of correctly transcribed words was counted as in Experiment 1 and submitted for further analysis. Data was analyzed using mixed-effects generalized linear modelling (using lme4 in R) with random intercepts for sentence and participant. Proportions of correct responses on each test were the dependent variables. Exposure condition (coded 0, 1, 2, 3, 4 for the no-exposure, baseline, multi-talker, multi-repetition and single sentence conditions, respectively) and test session (session 1, session 2) were fixed effects for the time-compressed and talker 2 natural-fast speech tests. Exposure condition was the only fixed factor for talker 3 natural-fast speech and for the speech-in-noise test which were conducted in session 2 only. For each dependent variable (talker 1, talker 2 etc.), model comparison was used to determine whether the inclusion of each of the fixed effects improved the fit of the model significantly.

Results

Test Performance as a function of Exposure

Time-compressed speech recognition accuracy is shown in Figure 5 (left and mid panels) and Table 6.

Table 6. Time-compressed speech and natural-fast speech by exposure condition and session

	Time-compressed Talker 1		Time-compressed Talker 2		Natural-fast Talker 2	
	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)
No-exposure						
Session 1	0.24 (0.16)	0.23 (0.12 – 0.35)	0.19 (0.14)	0.16 (0.11 – 0.27)	0.58 (0.14)	0.60 (0.49 – 0.66)
Session 2	0.28 (0.19)	0.24 (0.14 – 0.41)	0.35 (0.20)	0.34 (0.21 – 0.49)	0.68 (0.13)	0.70 (0.62 – 0.75)
Baseline						
Session 1	0.37 (0.20)	0.41 (0.25 – 0.50)	0.23 (0.15)	0.22 (0.10 - 0.37)	0.56 (0.16)	0.61 (0.47 – 0.68)
Session 2	0.32 (0.19)	0.31 (0.19 – 0.45)	0.42 (0.22)	0.43 (0.28 – 0.54)	0.67 (0.16)	0.72 (0.62 – 0.78)
Multi-talker						
Session 1	0.30 (0.19)	0.31 (0.13 – 0.42)	0.22 (0.14)	0.20 (0.09 – 0.32)	0.54 (0.15)	0.57 (0.41 – 0.65)
Session 2	0.25 (0.16)	0.21 (0.15 – 0.31)	0.34 (0.17)	0.32 (0.20 – 0.45)	0.65 (0.14)	0.70 (0.56 – 0.76)
Multi-repetition						
Session 1	0.29 (0.21)	0.27 (0.11 – 0.42)	0.22 (0.18)	0.16 (0.08 – 0.34)	0.50 (0.16)	0.55 (0.40 – 0.62)
Session 2	0.32 (0.20)	0.29 (0.17 – 0.45)	0.41 (0.23)	0.37 (0.24 – 0.55)	0.68 (0.18)	0.71 (0.62 – 0.81)
Single sentence						

	Time-compressed Talker 1		Time-compressed Talker 2		Natural-fast Talker 2	
	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)	Mean (SD)	Median (IQR)
Session 1	0.26 (0.18)	0.24 (0.13 – 0.37)	0.18 (0.15)	0.15 (0.06 – 0.27)	0.52 (0.16)	0.55 (0.45 – 0.60)
Session 2	0.25 (0.17)	0.22 (0.11 – 0.33)	0.34 (0.20)	0.31 (0.19 – 0.45)	0.67 (0.13)	0.68 (0.62 – 0.75)

For talker 1, each successive model fit the data better than the previous one. The model with exposure condition (AIC = 15397) fit the data better than the model with random effects only (AIC = 15399, $\chi^2_{(4)}$ = 9.54, p= 0.049). Adding session reduced AIC to 15392 ($\chi^2_{(1)}$ = 7.31, p = 0.007). The model with interactions fit the data best (AIC = 15344, $\chi^2_{(4)}$ = 55.78, p ≤ 0.001). However, this model was hard to interpret because as shown in Figure 5, performance in the no-exposure group (grey rectangles) improved between the two sessions, whereas changes in the exposure groups were variable: performance decayed in the 'baseline', 'multi-talker' and 'single-sentence' groups and somewhat increased in the 'multi-repetition' group. An inspection of the model parameters (Table 7) suggests that the condition by session interaction stems from a decrease in group difference between the baseline and the no-exposure groups, the multi-talker and the no-exposure groups and the single-sentence and the no-exposure groups, from session 1 to session 2.

Table 7. Prediction models for time-compressed speech recognition as a function of group, talker and session (including group x session interactions)

Parameter	Talker 1		Talker 2	
	Log-odds (SE)	Z (p)	Log-odds (SE)	Z (p)
Intercept (no-exposure)	-1.44 (0.23)	-6.23 (< 0.001)	-1.86 (0.29)	-6.49 (< 0.001)
Baseline	0.75 (0.22)	3.50 (< 0.001)	0.26 (0.22)	1.21 (0.224)
Multi-talker	0.41 (0.22)	1.90 (0.058)	0.21 (0.22)	0.97 (0.334)
Multi-repetition	0.28 (0.22)	1.29 (0.197)	0.13 (0.22)	0.57 (0.566)
Single sentence	0.04 (0.22)	0.18 (0.853)	-0.12 (0.22)	-0.56 (0.577)
Session 2	0.18 (0.07)	2.73 (0.006)	1.03 (0.07)	14.84 (< 0.001)
Baseline x Session 2	-0.48 (0.09)	-5.20 (< 0.001)	0.09 (0.10)	0.90 (0.371)
Multi-talker x Session 2	-0.54 (0.09)	-5.73 (< 0.001)	-0.26 (0.10)	-2.64 (0.008)
Multi-repetition x Session 2	-0.04 (0.09)	-0.38 (0.701)	0.18 (0.10)	1.81 (0.070)
Single sentence x Session 2	-0.23 (0.10)	-2.41 (0.016)	0.05 (0.10)	0.05 (0.613)

For talker 2, exposure condition did not significantly improve the fit of the model to the time-compressed speech data (AIC_{random} = 16234, $AIC_{condition}$ = 16237, $\chi^2_{(4)}$ = 4.94, p = 0.293), but the addition of session (AIC = 15103, $\chi^2_{(1)}$ = 1135, p < 0.001) and the exposure condition by session interaction (AIC = 15089, $\chi^2_{(4)}$ = 22.32, p < 0.001) did. However, as all groups improved from session 1 to session 2, it seems that learning during the tests was sufficient for learning the time-compressed speech produced by talker 2 regardless of previous exposure (see Table 7).

Finally, there were no group differences in the recognition of either natural fast speech (talker 3) or speech in noise in the final tests in session 2 (Figure 6). For natural fast speech, adding exposure condition had no significant effect on the fit compared to a model with random effects for item and participant only (AIC_{random} = 10681, $AIC_{condition}$ = 10688, $\chi^2_{(4)}$ = 0.67, p = 0.955). The same is true for speech in noise (AIC_{random} = 13941, $AIC_{condition}$ = 13945, $\chi^2_{(4)}$ = 3.64, p = 0.456). Given the group data shown in Table 6 and Figure 6 (top panel), we decided not to model that natural-fast speech data from talker 2 for group differences.

Discussion

Active listening to 10 time-compressed sentences was sufficient for robust and long-lasting perceptual learning (Experiment 1), consistent with the available literature. This rapid learning was specific to the acoustic characteristics of the speech used to elicit learning (Experiment 2). Although additional practice resulted in more learning, the associations between perceptual learning and speech recognition were driven by rapid learning (Experiment 1). In the context of previous works these data tentatively suggest

that additional practice does not change the nature of the resulting perceptual learning. If this is the case, rapid learning is key in understanding the function of perceptual learning in speech recognition, as we discuss in the following sections.

Long-Lasting and Specific: The Outcomes of Rapid Learning are Consistent with the Characteristics of Perceptual Learning

In the current study (Experiment 1), the duration of the practice phase had quantitative but not qualitative effects on perceptual learning. Consistent with previous studies [16,40], this rapid learning was relatively long lasting, but also quite specific to the acoustics of the stimuli that elicited learning. Although natural-fast speech recognition improved between the two test sessions, the improvement cannot be attributed to the transfer of learning of time-compressed speech because while learning itself was stronger in the training than in the exposure group, improvements in the recognition of natural-fast speech did not depend on group. Therefore, it is more likely to reflect relatively rapid learning of natural-fast speech and not transfer. Likewise, in Experiment 2, even when rapid learning occurred (in the baseline group), it was not reflected in the recognition of time-compressed speech produced by a new talker, similar to findings on training-induced learning of time-compressed speech [37,38].

Second, if learning was not stimulus specific, increasing the number of talkers or reducing the number of different sentences in Experiment 2 should not have interfered with learning. Yet these manipulations prevented rapid learning in line with previous reports on the effects of talker variability [16,56,57]. For example, when listening to speech produced by talkers with atypical /s/ or /sh/ pronunciations, adaptation to the unusual sounds was faster when each speaker was presented on its own than when the two were interleaved [56]. Talker variability during learning is thought to support the transfer of learning by providing listeners with a better sample of the systemic variability in the target speech [7,58]. However, this is not necessarily true for time-compressed speech, in which talker variability was found to slow training-induced learning with no effect on learning transfer [16]. Therefore, we suggest that rapid and training-induced learning are similarly specific or general, and therefore that rapid learning of speech reflects true perceptual learning rather than merely procedural or task learning. Similar conclusions were reported for non-verbal auditory and visual learning [12,13,42]. If learning emerges once experience with novel speech has provided sufficient familiarity with the characteristics of the target speech, both brief and prolonged practice could yield specific or general learning, depending on the characteristics of the input. For time-compressed speech, we demonstrate that learning is quite talker specific, as discussed above. On the other hand, learning of noise-vocoded speech seems to generalize more broadly across talkers and stimuli [32,45].

If more training does not change the nature of learning, what does it do? One option is that multi-session training could provide further opportunities for learning to stabilize and consolidate without changing the overall nature of learning [38,59,60]. This is consistent with the outcomes of both lab-based [37,61,62] and rehabilitation-oriented [22,60] studies. For example, in speech category learning, listeners accumulate information about the acoustic characteristics of the talker over time [61,63], thus

additional experience with a talker is likely to result in additional gains. Gradual accumulation of information about the talkers and the listening context could similarly support learning of perceptually challenging speech beyond the single word level. Furthermore, additional experience gives slower-learning listeners the opportunity to 'catch-up'. Sadly, most social, educational, and professional environments are not likely to provide those opportunities. Therefore, added to the relative specificity of learning already discussed, it seems that for understanding the role of perceptual learning in speech perception in challenging 'real-world' conditions, rapid learning is the key.

Rapid Learning and Individual Differences in Speech Perception

Individual differences in rapid perceptual learning of speech account for unique variance in speech perception in independent tasks [15,33,64]. The current data essentially replicates this association for rapid learning of time-compressed speech and natural-fast speech and speech-in-noise perception. However, we focused on between-session learning rather than on within-session learning which was the focus of previous studies. This let us assess whether the contributions of rapid and training-induced learning to speech perception on the other tasks differ (Figure 4). Overall, individuals with good learning were more likely to accurately recognize both natural fast speech (odds ratio = 1.52) and speech-in-noise (odds ratio = 1.36). Additional practice on time-compressed speech by the training group had no significant contribution to speech-in-noise, and a negative contribution to the recognition of natural-fast speech. In the absence of cross-task generalization following training, these findings suggest that in the current study rapid learning makes for the bulk of the speech/learning associations, consistent with the previous findings on rapid within-session learning [15,33].

The idea that rapid perceptual learning plays an independent role in individual differences in speech perception has merit only to the extent that (rapid) perceptual learning is a general ability or capacity of an individual, at least within a domain. A few recent auditory [35] and visual [34,36] learning studies suggest that a common factor could explain learning across different tasks. Using visual and auditory discrimination tasks, Yang et al. [36] reported that despite large differences in learning rates across tasks, a common learning factor accounted for more than 30% of the variance across different learning tasks. Roark et al. [35] studied the learning of non-speech auditory and visual categories. They found that while learning rates were faster for visual than for auditory categories, categorization accuracy at the end of training was correlated between the auditory and the visual task, suggesting that individual differences in category learning are correlated across the auditory and the visual modalities. As for associations across speech learning tasks, accuracy data of the type we normally collect might be insufficient to address this issue given the analytical methods used in the studies that reported cross-task association. Furthermore, although the rapid rates of learning in some speech tasks might make it difficult to separate "perception" and "learning", the consistent replication of the contribution of rapid learning of time-compressed speech to the perception of natural fast speech and speech in noise suggests that this is not an incidental finding. Future studies should nevertheless test the hypothesis that different speech learning conditions cluster around a common factor.

Limitations & Implications

First, sample sizes were not based on a formal power analysis because it was not obvious how to conduct it based on previous data and the rapid rates of time-compressed speech learning. Nevertheless, our previous studies of time-compressed speech learning (with similar but not identical conditions) yielded significant group differences as a function of the training protocol with sample sizes of 10 to 24 per group [e.g., 16,37,38,43]. Current sample size was therefore sufficient to uncover similar or larger effects. Furthermore, for the learning/recognition associations reported here (Table 3), the effect sizes for learning (expressed as odd ratios) were similar to those reported by Rotman et al. [15] (1.52 vs 1.44 – 1.68 for natural-fast speech and 1.36 vs. 1.49 for speech-in-noise) with similar groups sizes.

Second, our findings suggest that perceptual learning for speech is largely acoustically specific. However, this is not to say that longer training can never be useful. Instead, training-based studies or interventions should consider the specificity of learning in their design and expected outcomes. A recent review of perceptual learning of dysarthric speech [65] suggests that this is feasible. Since learning of dysarthric speech is constrained by the dysarthria characteristics of individual patients [66] and even experienced clinicians still benefit from talker-specific training [67] it is proposed that communication partners train to improve the intelligibility of specific patients (e.g., a family member), accounting for learning specificity.

Third, speech perception under challenging conditions incorporates both stimulus (e.g., talker, input distribution; [18]) and listener (e.g., age, language, and cognition [68,69]) related factors. We now suggest that rapid learning is another meaningful listener related factor that could determine how well individual listeners adapt to new or changing auditory environments. Determining if individual differences are associated across different learning tasks or with performance in other situations requires further studies with different learning and perception tasks. Still the finding that individual differences in rapid-learning with one type of challenging speech predicts individual differences in the processing of a different type of challenging speech is telling despite the correlational nature of our work.

Declarations

Acknowledgement

This work was funded by the Israel Science Foundation Grant 206/18 and by a National Institute of Psychobiology in Israel Grant 108/2014-2015

Author Contributions

All authors designed the study, reviewed and approved the manuscript. LL, HK and YL prepared the materials for the study. KB and YL analyzed the data. KB wrote the manuscript.

Competing Interests

The authors declare no competing interests.

References

1 Mattys, S. L., Davis, M. H., Bradlow, A. R. & Scott, S. K. Speech recognition in adverse conditions: A review. *Lang. Cogn. Process.***27**, 953-978 (2012).

2 Benard, M. R. & Başkent, D. Perceptual learning of interrupted speech. *PLoS One***8**, e58149, doi:10.1371/journal.pone.0058149 (2013).

3 Borrie, S. A., McAuliffe, M. J. & Liss, J. M. Perceptual learning of dysarthric speech: a review of experimental studies. *J. Speech. Lang. Hear. Res.***55**, 290-305, doi:10.1044/1092-4388(2011/10-0349) (2012).

4 Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K. & McGettigan, C. Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *J. Exp. Psychol. Gen.***134**, 222-241, doi:10.1037/0096-3445.134.2.222 (2005).

5 Dupoux, E. & Green, K. Perceptual adjustment to highly compressed speech: effects of talker and rate changes. *J. Exp. Psychol. Hum. Percept. Perform.***23**, 914-927, doi:10.1037//0096-1523.23.3.914 (1997).

6 Green, T., Rosen, S., Faulkner, A. & Paterson, R. Adaptation to spectrally-rotated speech. *J. Acoust. Soc. Am.***134**, 1369-1377, doi:10.1121/1.4812759 (2013).

7 Greenspan, S. L., Nusbaum, H. C. & Pisoni, D. B. Perceptual learning of synthetic speech produced by rule. *J. Exp. Psychol. Learn. Mem. Cogn.***14**, 421-433, doi:10.1037//0278-7393.14.3.421 (1988).

8 Stacey, P. C. & Summerfield, A. Q. Comparison of word-, sentence-, and phoneme-based training strategies in improving the perception of spectrally distorted speech. *J. Speech. Lang. Hear. Res.***51**, 526-538, doi:10.1044/1092-4388(2008/038) (2008).

9 Orr, D. B. & Friedman, H. L. Effect of massed practice on the comprehension of time-compressed speech. *J. Educ. Psychol.***59**, 6 (1968).

10 Goldstone, R. L. Perceptual learning. *Annu. Rev. Psychol.***49**, 585-612, doi:10.1146/annurev.psych.49.1.585 (1998).

11 Green, C. S., Banai, K., Lu, Z.-L. & Bavelier, D. Perceptual learning in *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (ed J.T. Wixted) 1-47 (2018).

12 Hawkey, D. J., Amitay, S. & Moore, D. R. Early and rapid perceptual learning. *Nat. Neurosci.***7**, 1055-1056, doi:10.1038/nn1315 (2004).

13 Ortiz, J. A. & Wright, B. A. Contributions of procedure and stimulus learning to early, rapid perceptual improvements. *J. Exp. Psychol. Hum. Percept. Perform.***35**, 188-194, doi:10.1037/a0013161 (2009).

14 Banai, K. & Lavie, L. Rapid perceptual learning and individual differences in speech perception: The good, the bad, and the sad. *Auditory Perception & Cognition***3**, 201-211, doi:10.1080/25742442.2021.1909400 (2021).

15 Rotman, T., Lavie, L. & Banai, K. Rapid perceptual learning: A potential source of individual differences in speech perception under adverse conditions? *Trends Hear***24**, 2331216520930541, doi:10.1177/2331216520930541 (2020).

16 Tarabeih - Ghanayim, M., Lavner, Y. & Banai, K. Tasks, talkers, and the perceptual learning of time-compressed speech. *Auditory Perception & Cognition***3**, 33-54, doi:10.1080/25742442.2020.1846011 (2020).

17 Ahissar, M., Nahum, M., Nelken, I. & Hochstein, S. Reverse hierarchies and sensory learning. *Philos. Trans. R. Soc. Lond. B Biol. Sci.***364**, 285-299, doi:10.1098/rstb.2008.0253 (2009).

18 Kleinschmidt, D. F. & Jaeger, T. F. Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychol. Rev.***122**, 148-203, doi:10.1037/a0038695 (2015).

19 Rönnberg, J. *et al.* The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances. *Front. Syst. Neurosci.***7**, 31, doi:10.3389/fnsys.2013.00031 (2013).

20 Barcroft, J., Spehar, B., Tye-Murray, N. & Sommers, M. Task- and talker-specific gains in auditory training. *J. Speech. Lang. Hear. Res.***59**, 862-870, doi:10.1044/2016_jslhr-h-15-0170 (2016).

21 Green, T., Faulkner, A. & Rosen, S. Computer-based connected-text training of speech-in-noise perception for cochlear implant users. *Trends Hear***23**, 2331216519843878, doi:10.1177/2331216519843878 (2019).

22 Karawani, H., Bitan, T., Attias, J. & Banai, K. Auditory perceptual learning in adults with and without age-related hearing loss. *Front. Psychol.***6**, 2066, doi:10.3389/fpsyg.2015.02066 (2016).

23 Song, J. H., Skoe, E., Banai, K. & Kraus, N. Training to improve hearing speech in noise: biological mechanisms. *Cereb. Cortex***22**, 1180-1190, doi:10.1093/cercor/bhr196 (2012).

24 Burk, M. H. & Humes, L. E. Effects of long-term training on aided speech-recognition performance in noise in older adults. *J. Speech. Lang. Hear. Res.***51**, 759-771, doi:10.1044/1092-4388(2008/054) (2008).

25 Henshaw, H. & Ferguson, M. A. Efficacy of individual computer-based auditory training for people with hearing loss: a systematic review of the evidence. *PLoS One***8**, e62836, doi:10.1371/journal.pone.0062836 (2013).

26 Saunders, G. H. *et al.* A randomized control trial: Supplementing hearing aid use with Listening and Communication Enhancement (LACE) auditory training. *Ear Hear.***37**, 381-396, doi:10.1097/AUD.0000000000000283 (2016).

27 Eisner, F. & McQueen, J. M. Perceptual learning in speech: stability over time. *J. Acoust. Soc. Am.***119**, 1950-1953, doi:10.1121/1.2178721 (2006).

28 Samuel, A. G. & Kraljic, T. Perceptual learning for speech. *Atten. Percept. Psychophys.***71**, 1207-1218, doi:10.3758/APP.71.6.1207 (2009).

29 Bradlow, A. R. & Bent, T. Perceptual adaptation to non-native speech. *Cognition***106**, 707-729, doi:10.1016/j.cognition.2007.04.005 (2008).

30 Casserly, E. D. & Pisoni, D. B. Auditory learning using a portable real-time vocoder: Preliminary findings. *J. Speech. Lang. Hear. Res.***58**, 1001-1016, doi:10.1044/2015_jslhr-h-13-0216 (2015).

31 Hervais-Adelman, A. G., Davis, M. H., Johnsrude, I. S., Taylor, K. J. & Carlyon, R. P. Generalization of perceptual learning of vocoded speech. *J. Exp. Psychol. Hum. Percept. Perform.***37**, 283-295, doi:10.1037/a0020772 (2011).

32 Huyck, J. J., Smith, R. H., Hawkins, S. & Johnsrude, I. S. Generalization of Perceptual Learning of Degraded Speech Across Talkers. *J. Speech. Lang. Hear. Res.***60**, 3334-3341, doi:10.1044/2017_JSLHR-H-16-0300 (2017).

33 Manheim, M., Lavie, L. & Banai, K. Age, hearing, and the perceptual learning of rapid speech. *Trends Hear***22**, 2331216518778651, doi:10.1177/2331216518778651 (2018).

34 Dale, G., Cochrane, A. & Green, C. S. Individual difference predictors of learning and generalization in perceptual learning. *Atten. Percept. Psychophys.***83**, 2241-2255, doi:10.3758/s13414-021-02268-3 (2021).

35 Roark, C. L., Paulon, G., Sarkar, A. & Chandrasekaran, B. Comparing perceptual category learning across modalities in the same individuals. *Psychon. Bull. Rev.***28**, 898-909, doi:10.3758/s13423-021-01878-0 (2021).

36 Yang, J. *et al.* General learning ability in perceptual learning. *Proc. Natl. Acad. Sci. U. S. A.***117**, 19092-19100, doi:10.1073/pnas.2002903117 (2020).

37 Banai, K. & Lavner, Y. The effects of training length on the perceptual learning of time-compressed speech and its generalization. *J. Acoust. Soc. Am.***136**, 1908-1917, doi:10.1121/1.4895684 (2014).

38 Banai, K. & Lavner, Y. Effects of stimulus repetition and training schedule on the perceptual learning of time-compressed speech and its transfer. *Atten. Percept. Psychophys.***81**, 2944-2955, doi:10.3758/s13414-019-01714-7 (2019).

39 Wright, B. A., LeBlanc, E. K., Little, D. F., Conderman, J. S. & Glavin, C. C. Semi-supervised learning of a nonnative phonetic contrast: How much feedback is enough? *Atten. Percept. Psychophys.***81**, 927-934, doi:10.3758/s13414-019-01741-4 (2019).

40 Altmann, G. & Young, D. Factors affecting adaptation to time-compressed speech, in *EUROSPEECH'93,* 333-336, (1993).

41 Fiorentini, A. & Berardi, N. Perceptual learning specific for orientation and spatial frequency. *Nature***287**, 43-44, doi:10.1038/287043a0 (1980).

42 Hussain, Z., McGraw, P. V., Sekuler, A. B. & Bennett, P. J. The rapid emergence of stimulus specific perceptual learning. *Front. Psychol.***3**, 226, doi:10.3389/fpsyg.2012.00226 (2012).

43 Banai, K. & Lavner, Y. Perceptual learning of time-compressed speech: more than rapid adaptation. *PLoS One***7**, e47099, doi:10.1371/journal.pone.0047099 (2012).

44 Lehet, M. I., Fenn, K. M. & Nusbaum, H. C. Shaping perceptual learning of synthetic speech through feedback. *Psychon. Bull. Rev.***27**, 1043-1051, doi:10.3758/s13423-020-01743-6 (2020).

45 Loebach, J. L., Pisoni, D. B. & Svirsky, M. A. Effects of semantic context and feedback on perceptual learning of speech processed through an acoustic simulation of a cochlear implant. *J. Exp. Psychol. Hum. Percept. Perform.***36**, 224-234, doi:10.1037/a0017609 (2010).

46 Adank, P. & Janse, E. Perceptual learning of time-compressed and natural fast speech. *J. Acoust. Soc. Am.***126**, 2649-2659, doi:10.1121/1.3216914 (2009).

47 Banai, K. & Lavie, L. Perceptual learning and speech perception: A new hypothesis, in *Proceedings of the International Symposium on Auditory and Audiological Research, Vol.7: Auditory Learning in Biological and Artificial Systems* Vol. 7 (eds A. Kressner *et al.*), 53-60 (The Danavox Jubilee Foundation, 2020).

48 Prior, A. & Bentin, S. Differential integration efforts of mandatory and optional sentence constituents. *Psychophysiology***43**, 440-449, doi:10.1111/j.1469-8986.2006.00426.x (2006).

49 Verhelst, W. & Roelands, M. in *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing* Vol. 2 554-557 (IEEE, 1993).

50 R: A language and environment for statistical computing. R Foundation for Statistical Computing (Vienna, Austria, 2019).

51 Bates, D., Maechler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software***67**, 1-48 (2015).

52 Lively, S. E., Logan, J. S. & Pisoni, D. B. Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *J. Acoust. Soc. Am.***94**, 1242-1255, doi:10.1121/1.408177 (1993).

53 Stacey, P. C. & Summerfield, A. Q. Effectiveness of computer-based auditory training in improving the perception of noise-vocoded speech. *J. Acoust. Soc. Am.***121**, 2923-2935, doi:10.1121/1.2713668 (2007).

54 Loebach, J. L., Bent, T. & Pisoni, D. B. Multiple routes to the perceptual learning of speech. *J. Acoust. Soc. Am.***124**, 552-561, doi:10.1121/1.2931948 (2008).

55 Wechsler, D. *Wechsler Adult Intelligence Scale (WAIS-III) Administration and Scoring Manual*. 4 edn, (The Psychological Corporation, 2008).

56 Luthra, S., Mechtenberg, H. & Myers, E. B. Perceptual learning of multiple talkers requires additional exposure. *Atten. Percept. Psychophys.***83**, 2217-2228, doi:10.3758/s13414-021-02261-w (2021).

57 Wade, T., Jongman, A. & Sereno, J. Effects of acoustic variability in the perceptual learning of non-native-accented speech sounds. *Phonetica***64**, 122-144, doi:10.1159/000107913 (2007).

58 Baese-Berk, M. M., Bradlow, A. R. & Wright, B. A. Accent-independent adaptation to foreign accented speech. *J. Acoust. Soc. Am.***133**, EL174-180, doi:10.1121/1.4789864 (2013).

59 Molloy, K., Moore, D. R., Sohoglu, E. & Amitay, S. Less is more: latent learning is maximized by shorter training sessions in auditory perceptual learning. *PLoS One***7**, e36929, doi:10.1371/journal.pone.0036929 (2012).

60 Tye-Murray, N., Spehar, B., Barcroft, J. & Sommers, M. Auditory training for adults who have hearing loss: a comparison of spaced versus massed practice schedules. *J. Speech. Lang. Hear. Res.***60**, 2337-2345, doi:10.1044/2017_jslhr-h-16-0154 (2017).

61 Tzeng, C. Y., Nygaard, L. C. & Theodore, R. M. A second chance for a first impression: Sensitivity to cumulative input statistics for lexically guided perceptual learning. *Psychon. Bull. Rev.***28**, 1003-1014, doi:10.3758/s13423-020-01840-6 (2021).

62 Wright, B. A., Wilson, R. M. & Sabin, A. T. Generalization lags behind learning on an auditory perceptual task. *J. Neurosci.***30**, 11635-11639, doi:10.1523/JNEUROSCI.1441-10.2010 (2010).

63 Nygaard, L. C. & Pisoni, D. B. Talker-specific learning in speech perception. *Percept. Psychophys.***60**, 355-376, doi:10.3758/bf03206860 (1998).

64 Karawani, H., Lavie, L. & Banai, K. Short-term auditory learning in older and younger adults, in *Proceedings of the International Symposium on Auditory and Audiological Research, Vol. 6: Adaptive Processes in Hearing* Vol. 6 (eds S. Santurette *et al.*) 1-8 (The Danavox Jubilee Foundation, Nyborg, Denmark, 2017).

65 Borrie, S. A. & Lansford, K. L. A perceptual learning approach for dysarthria remediation: An updated review. *J. Speech. Lang. Hear. Res.***64**, 3060-3073, doi:10.1044/2021_jslhr-21-00012 (2021).

66 Borrie, S. A., Lansford, K. L. & Barrett, T. S. Generalized adaptation to dysarthric speech. *J. Speech. Lang. Hear. Res.***60**, 3110-3117, doi:10.1044/2017_JSLHR-S-17-0127 (2017).

67 Borrie, S. A., Lansford, K. L. & Barrett, T. S. A clinical advantage: Experience informs recognition and adaptation to a novel talker with dysarthria. *J. Speech. Lang. Hear. Res.***64**, 1503-1514, doi:10.1044/2021_jslhr-20-00663 (2021).

68 Kennedy-Higgins, D., Devlin, J. T. & Adank, P. Cognitive mechanisms underpinning successful perception of different speech distortions. *J. Acoust. Soc. Am.***147**, 2728, doi:10.1121/10.0001160 (2020).

69 McLaughlin, D. J., Baese-Berk, M. M., Bent, T., Borrie, S. A. & Van Engen, K. J. Coping with adversity: Individual differences in the perception of noisy and accented speech. *Atten. Percept. Psychophys.***80**, 1559-1570, doi:10.3758/s13414-018-1537-4 (2018).

Figures

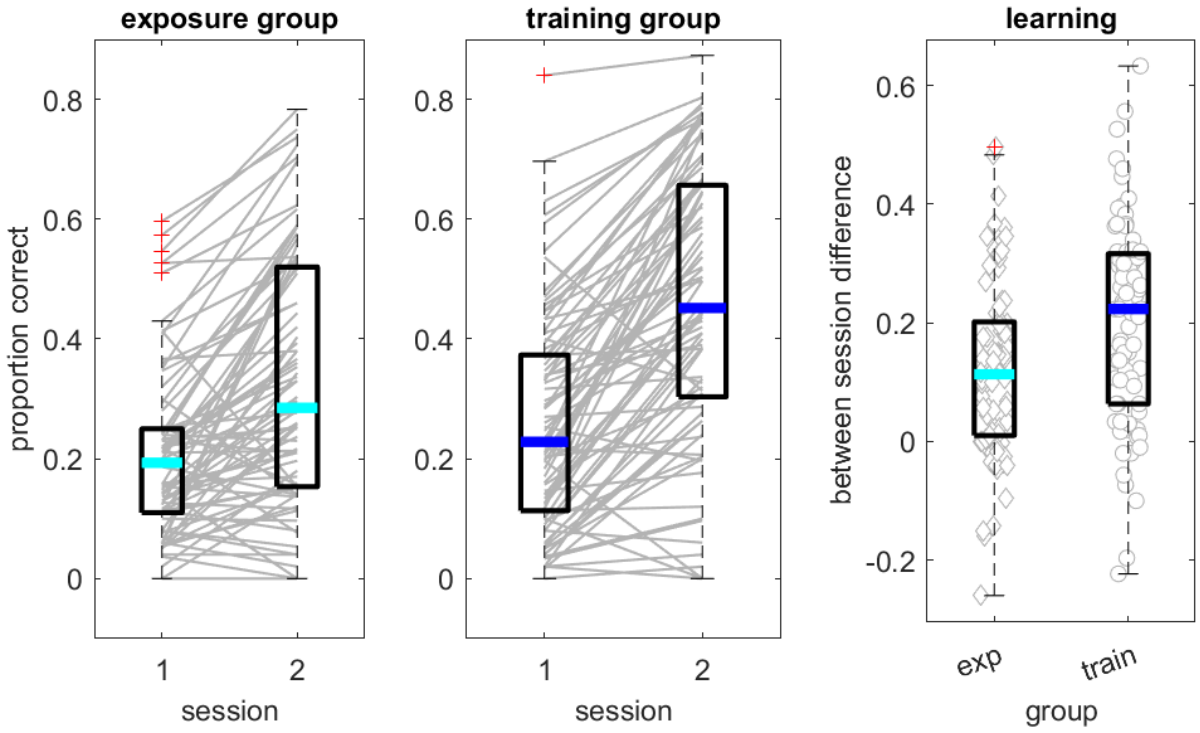

Figure 1

Time compressed speech recognition and learning in the exposure and training groups. The left and center panels show performance averaged within each participant in each session. Lines show individual data. The rightmost panel show learning, expressed as the difference in recognition accuracy between the two sessions in the exposure group (exp) and in the training group (train). Background symbols denote individual data. The thick line within each boxplot shows group median; box edges mark the interquartile

range; whiskers are 1.5 times the interquartile range; + signs are values outside the 1.5 x interquartile range.

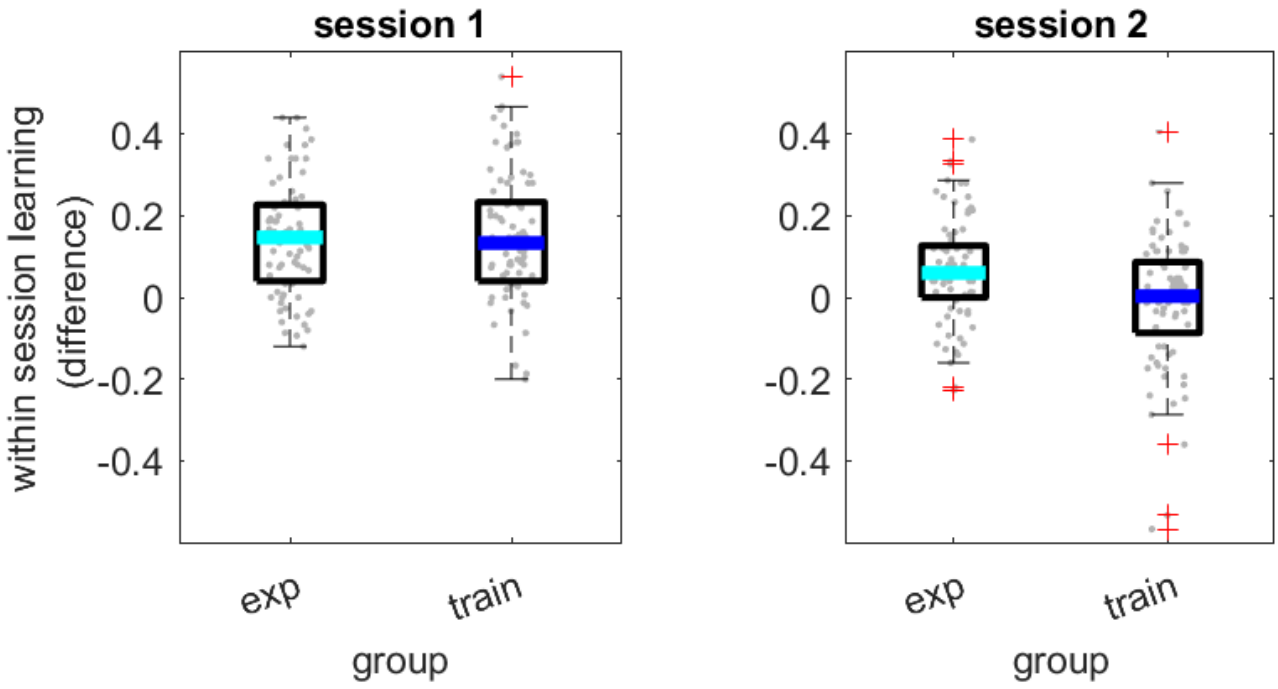

Figure 2

Rapid (within session) learning as a function of group. Rapid (within session) learning, defined as the difference in performance accuracy between the first 5 and last 5 time-compressed sentences on each session. exp = exposure group; train = training group. Background symbols denote individual data. The thick line within each boxplot shows group median; box edges mark the interquartile range; whiskers are 1.5 times the interquartile range; + signs are values outside of the 1.5 x interquartile range.

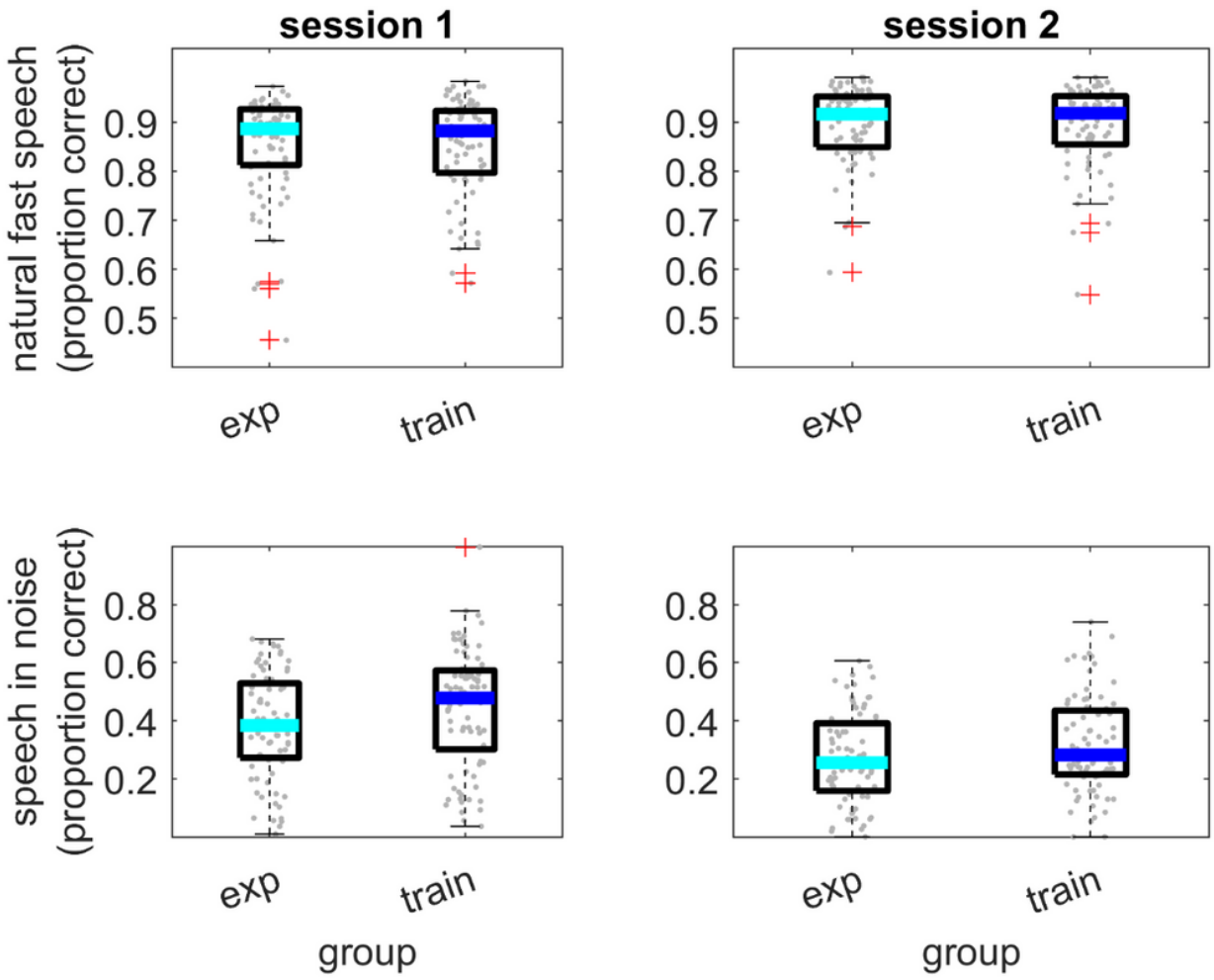

Figure 3

Perception of natural-fast speech and speech-in-noise as a function of group and session. exp = exposure group; train = training group. Background symbols denote mean individual data across sentences. The thick line within each boxplot shows group median; box edges mark the interquartile range; whiskers are 1.5 times the interquartile range; + signs are values outside of the 1.5 x interquartile range.

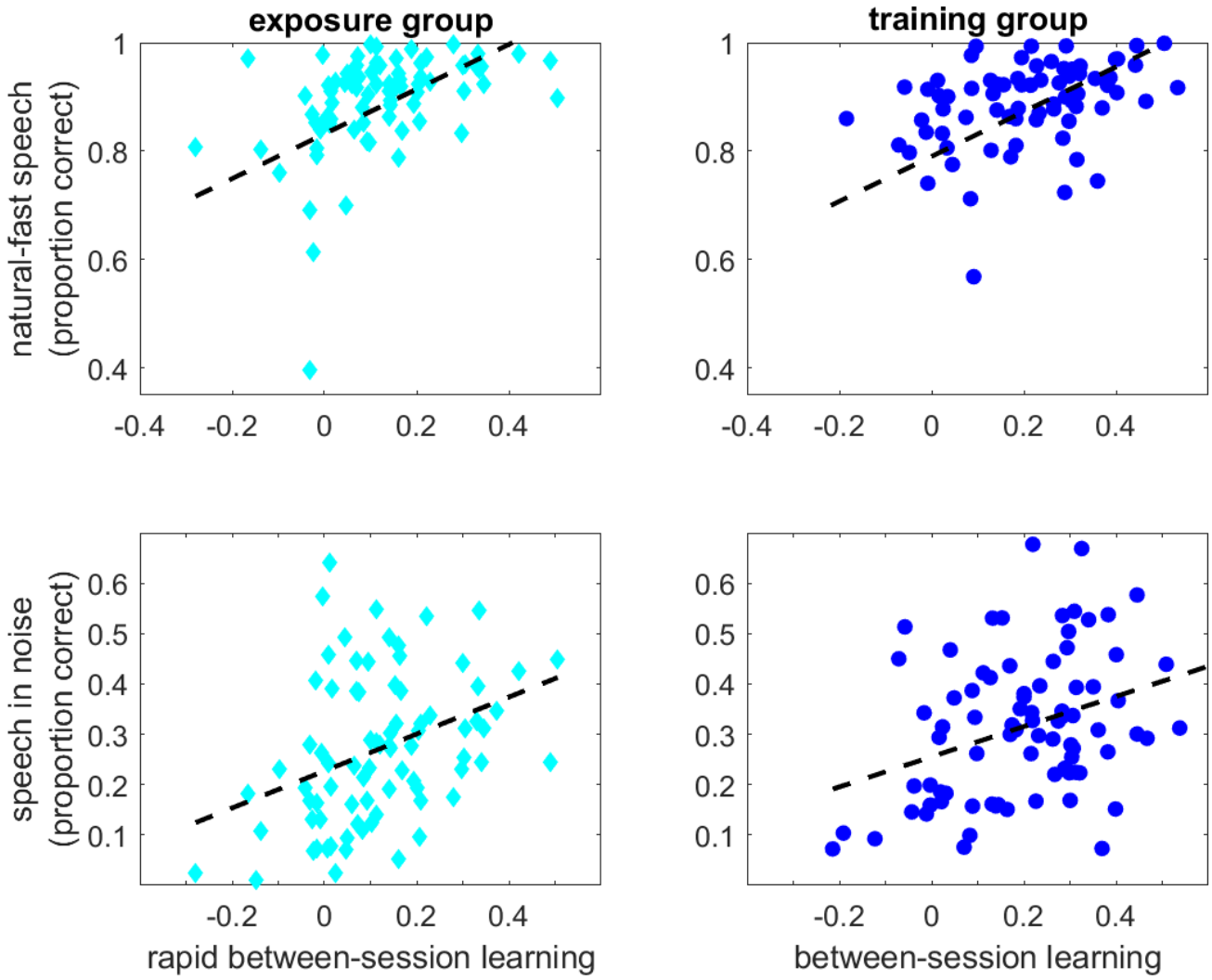

Figure 4

Speech recognition versus perceptual learning. Proportions correct in session 2 are plotted against between-session learning of time-compressed speech expressed as the difference in recognition accuracy between the two sessions for each participant. Dashed lines show linear fits. For visualization only, learning scores were adjusted to partial out the contribution of baseline recognition of time-compressed speech (as in Manheim et al., 2018). Therefore, values on the x axes are not the same as the simple difference scores shown in Figure 1.

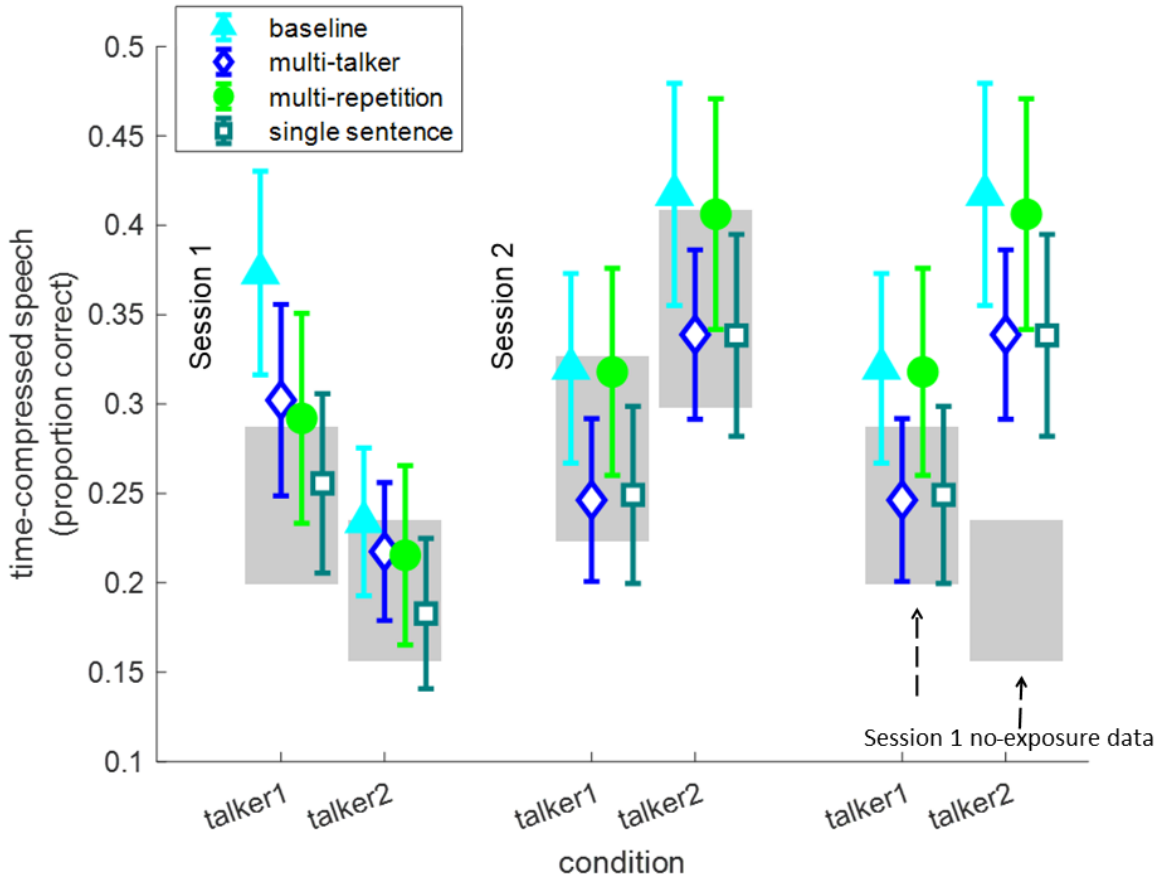

Figure 5

Time-compressed speech recognition by exposure condition in the immediate (session 1) and delayed (session 2) tests. For each exposure group Mean (across all sentences per condition) and 95% confidence interval are shown. The grey rectangles mark the 95% confidence interval of the no-exposure group who participated in testing only. The right-most section of the plot shows session 2 data for the exposure groups and session 1 data for the no-exposure group.

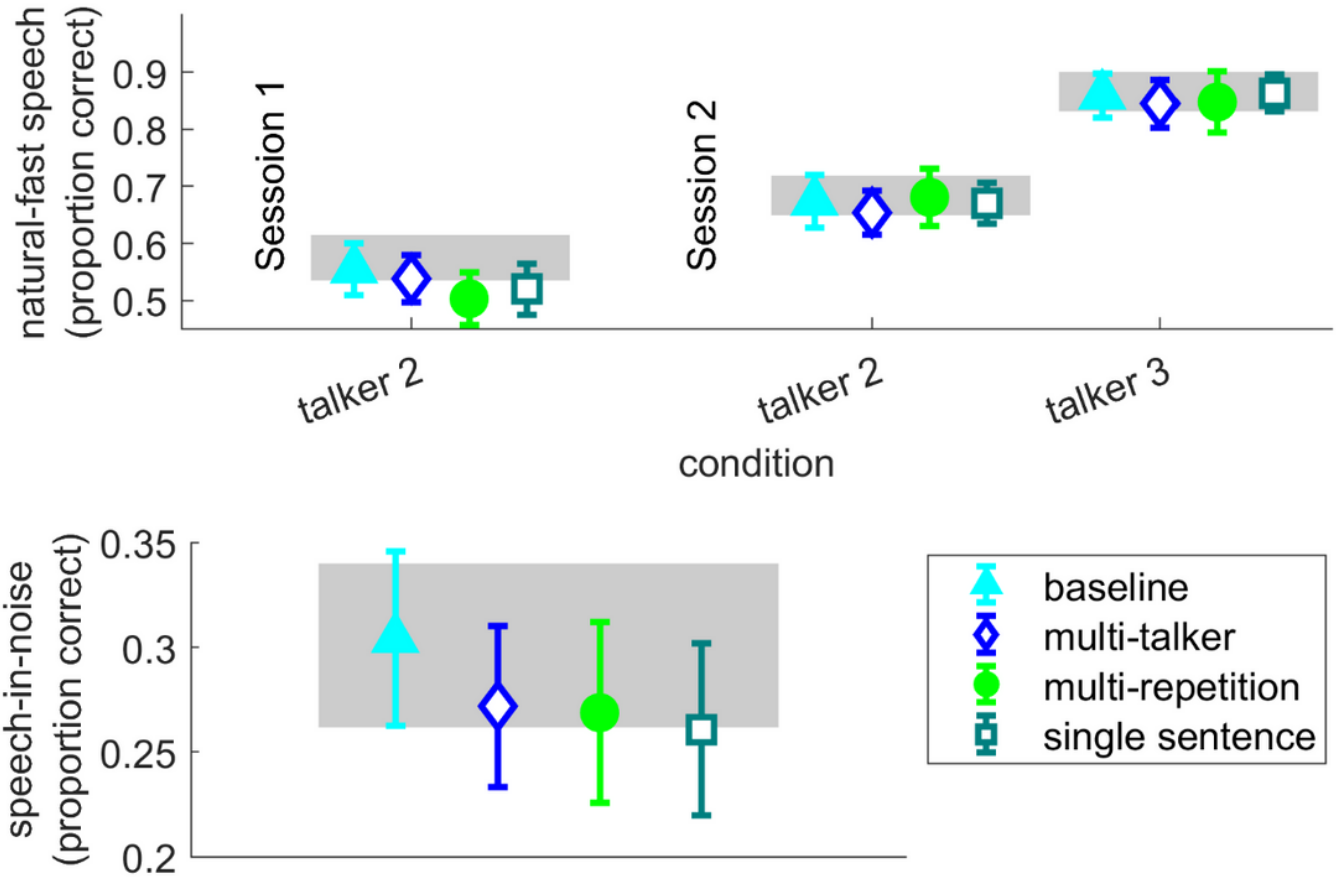

Figure 6

Natural-fast speech (top, left to right – talker 2 in session 1, talker 2 in session 2 and talker 3 in session 2) and Speech-in-Noise (bottom) recognition. For each exposure group mean (across all sentences per condition) and 95% confidence interval are shown. The gray rectangles mark the 95% confidence interval of the no-exposure group who participated in testing only.