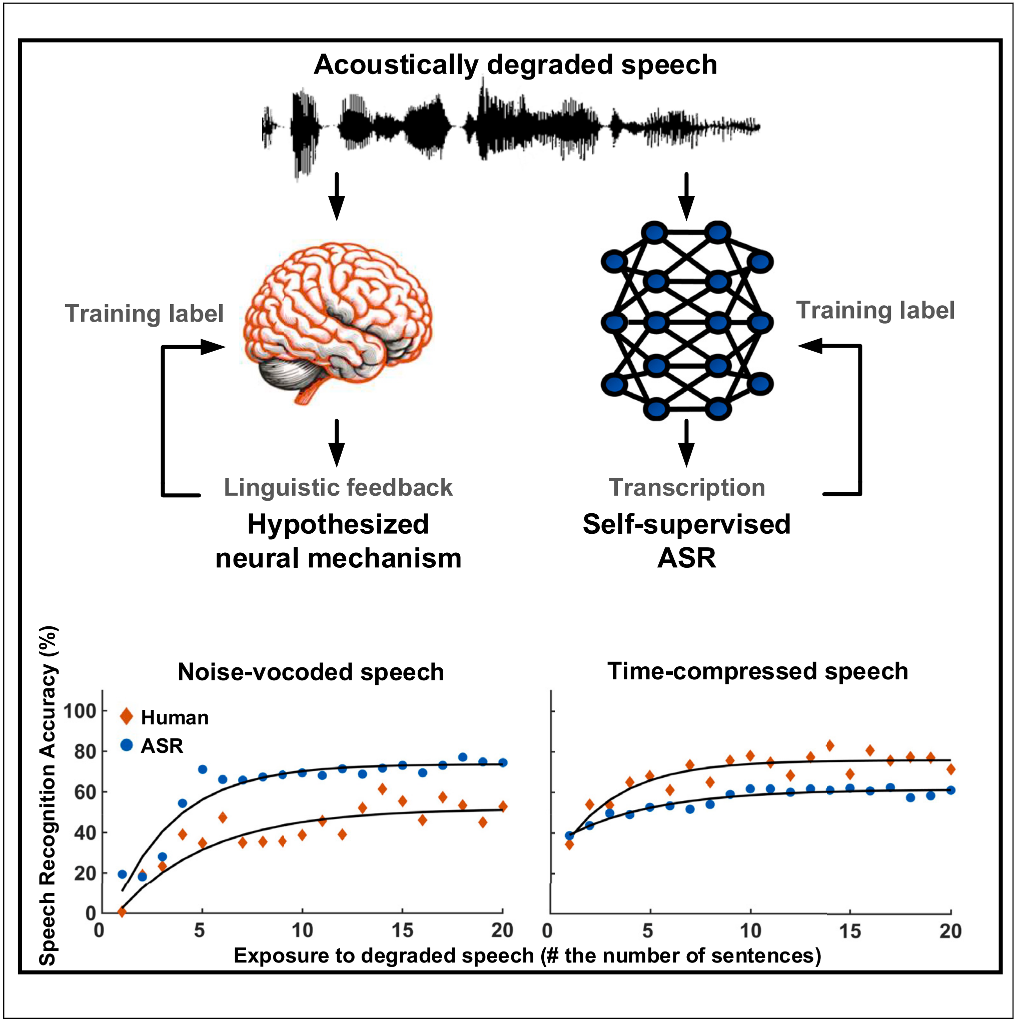
iScience

CellPress OPEN ACCESS

Article

Linguistic feedback supports rapid adaptation to acoustically degraded speech



Wenhui Sun, Jiajie Zou, Tianyi Zhu, Zhoujian Sun, Nai Ding

ding_nai@zju.edu.cn

Highlights

Provide new data and methods to compare speech adaptation in humans and ASR

Model linguistic-feedback-based speech adaptation using self-supervised learning

Linguistic feedback can support rapid adaptation to degraded speech

Without adaptation, human speech recognition is not more robust than ASR

Sun et al., iScience *27*, 110055
June 21, 2024 © 2024 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.isci.2024.110055

CellPress
OPEN ACCESS

Article

Linguistic feedback supports rapid adaptation to acoustically degraded speech

Wenhui Sun,[1] Jiajie Zou,[2] Tianyi Zhu,[2] Zhoujian Sun,[1] and Nai Ding[2,3,*]

SUMMARY

Humans can quickly adapt to recognize acoustically degraded speech, and here we hypothesize that the quick adaptation is enabled by internal linguistic feedback – Listeners use partially recognized sentences to adapt the mapping between acoustic features and phonetic labels. We test this hypothesis by quantifying how quickly humans adapt to degraded speech and analyzing whether the adaptation process can be simulated by adapting an automatic speech recognition (ASR) system based on its own speech recognition results. We consider three types of acoustic degradation, i.e., noise vocoding, time compression, and local time-reversal. The human speech recognition rate can increase by >20% after exposure to just a few acoustically degraded sentences. Critically, the ASR system with internal linguistic feedback can adapt to degraded speech with human-level speed and accuracy. These results suggest that self-supervised learning based on linguistic feedback is a plausible strategy for human adaptation to acoustically degraded speech.

INTRODUCTION

Humans have the capability to rapidly adapt to recognize acoustically degraded speech.[1–5] Critically, previous studies have shown that human listeners can better adapt to meaningful sentences than meaningless utterances, suggesting that top-down linguistic feedback is important for adaptation to degraded speech.[1] In other words, adaptation is enabled only when the brain can map degraded speech onto linguistic units, e.g., words. Furthermore, although previous research has shown that top-down linguistic feedback can modulate learning, human listeners could adapt to degraded speech without external linguistic feedback.[1,3,6] In the conditions without external linguistic feedback, it is unclear what are the factors driving adaptation. In this work, we aim to explore the potential computational strategy and hypothesize that human listeners can adapt to degraded speech based on internal linguistic feedback, i.e., what the listener recognizes the degraded speech. Additionally, linguistic feedback, i.e., recognized words, is not the only cue that can drive auditory perceptual learning, and other potential cues include prior knowledge, attention, statistical learning.[7–9] If linguistic feedback is computationally insufficient to drive speech adaptation (our alternative hypothesis), it is evident that the listeners have to integrate multiple cues during speech adaptation. In contrast, if linguistic feedback is computationally sufficient to drive speech adaptation in the conditions we tested here, it can stimulate future studies to investigate in which conditions speech adaptation cannot be fully explained by linguistic feedback and what other cues are actually used by human listeners.

Automatic speech recognition (ASR) is an algorithm that automatically transcribes speech, i.e., converts the speech waveform into text. Here, we consider ASR systems based on deep neural network (DNN), which automatically learns the mapping between speech acoustic features and words through manually transcribed speech datasets. To test our hypothesis, we utilize a computational approach by adapting the ASR system based on its own transcription, which simulates the internal linguistic feedback and it may contain both correctly recognized words and incorrectly recognized words. After an acoustically degraded sentence is transcribed by the ASR, we let the ASR system learn the mapping between degraded acoustic features and the words that the ASR system transcribes. If the ASR speech recognition accuracy significantly improves after adaptation based on its own transcription, it provides strong evidence that linguistic feedback is sufficient to drive adaptation to degraded speech.

The idea of training an ASR system based on its own transcription is closely related to the widely applied pseudo-labeling method in the field of ASR, which involves using a model's own transcription of degraded speech as a supervision signal to adapt the model.[10–13] Recent advances in DNN-based ASR systems have offered a potential tool to investigate the computational strategy underlying speech recognition behavior, as these systems have reached human-level speech recognition performance in many scenarios.[14,15] Therefore, despite dramatic differences in the implementation of ASR systems and the human brain, we utilize the ASR system to probe the computational-level principle behind the rapid human adaptation to acoustically degraded speech.[16]

[1] Research Center for Life Sciences Computing, Zhejiang Lab, Hangzhou 311121, China
[2] Key Laboratory for Biomedical Engineering of Ministry of Education, College of Biomedical Engineering and Instrument Sciences, Zhejiang University, Hangzhou 310027, China
[3] Lead contact
*Correspondence: ding_nai@zju.edu.cn

https://doi.org/10.1016/j.isci.2024.110055

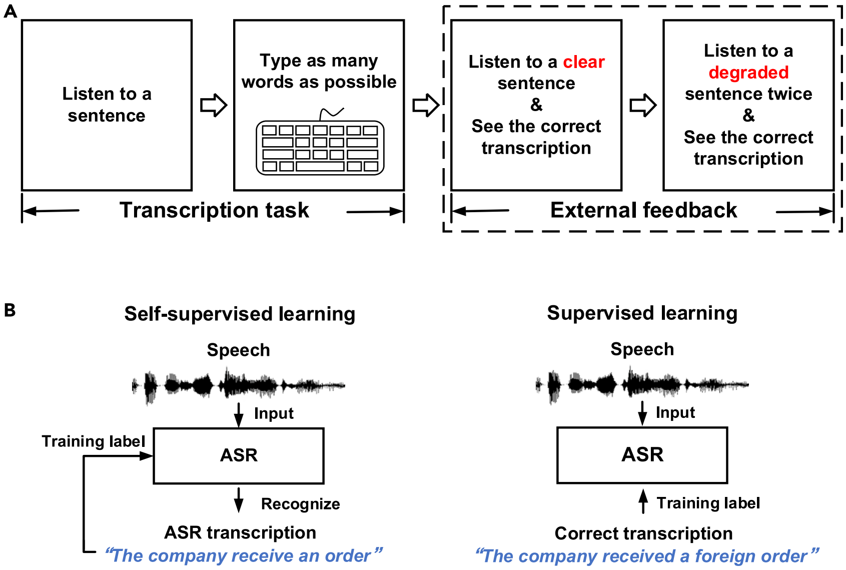

A
Listen to a sentence
Type as many words as possible
Listen to a clear sentence & See the correct transcription
Listen to a degraded sentence twice & See the correct transcription
Transcription task
External feedback
B
Self-supervised learning
Speech
Input
Training label
ASR
Recognize
ASR transcription
"The company receive an order"
Supervised learning
Speech
Input
ASR
Training label
Correct transcription
"The company received a foreign order"

Figure 1. Adaptation strategies for the human listeners and automatic speech recognition (ASR) systems

(A) Adaptation strategies for the human experiments with external feedback. In the human experiment without external feedback, the experimental procedure only involves the transcription task.

(B) Adaptation strategies for the ASR systems using self-supervised learning and supervised learning.

For the human experiment, we adopted a fine-grained experimental design that enabled us to track speech adaptation on a sentence-by-sentence basis.[3,17] Furthermore, human listeners adapt to acoustically degraded speech either with or without external linguistic feedback, i.e., the correct transcription (Figure 1). We build computational strategies that can potentially simulate human adaptation with or without external linguistic feedback. To model human adaptation without external linguistic feedback, the ASR system is fine-tuned based on its own transcription of degraded speech. To model human adaptation with external linguistic feedback, the ASR system is fine-tuned based on the correct transcription. We test how well these two computational speech-adaptation strategies can explain human adaptation to degraded speech.

Specifically, we employed a two (external feedback: without vs. with) × three (type of degraded speech: noise vocoding vs. time compression vs. local time-reversal) × two (difficulty: easy vs. hard) experimental design. Noise vocoding filters speech into 8 or 4 channels and removes spectral detail from speech.[18] Time compression alters the speech rate to 2.5 or 3.0 times, and local time-reversal impairs the time information within a 50 or 62 ms time window (see STAR Methods: Degraded speech for details). Spectrograms of clear speech and three easy cases of acoustic degradation are exhibited in Figure 2. In human experiments, we implemented a between-subject design, wherein each participant was exposed to one of the 12 possible conditions. In the ASR experiments, we employed a Conformer-based ASR system because of its proven effectiveness in ASR tasks.[19] This ASR system comprised a 12-block Conformer[19] encoder and a 6-block Bitransformer decoder.[20] It was pre-trained on a 10,000-h Mandarin dataset.[21]

RESULTS

Automatic speech recognition system using self-supervised learning achieves human-like performance

In the first human experiment, listeners adapted to degraded speech without any external feedback (refer to Figure 1A). The experiment involved 180 participants divided into six groups of 30, each exposed to single type of acoustic degradation. In the experiment, participants transcribed a set of 30 different sentences. The first 10 were clear, while the following 20 were subjected to a type of acoustic degradation (see STAR Methods: Human experiment for more details). The same experiment was also applied to the ASR system. Initially, the ASR system adapted to 10 clear sentences. Then, starting from the 11th sentence, it began to adapt to degraded sentences. Throughout the process, the ASR system adapted based on its own transcription (self-supervised learning). We measured speech recognition accuracy from both human listeners and ASR systems. The speech recognition accuracy, changing as a function of the number of exposed degraded sentences, was fitted by an exponential curve (see STAR Methods: Performance evaluation for details). We analyzed the fitted speech recognition accuracy at the first and final degraded sentence positions, as well as adaptation speed, to characterize the adaptation process.

Time course of the speech recognition accuracy is shown in Figure 3. The x axis was the index of the sentence within an experiment, and the y axis displayed both the averaged speech recognition accuracy across a group of participants and the accuracy of an ASR system. The adaptation of the self-supervised ASR system mirrored that of humans without external feedback, with both showing substantial improvements in

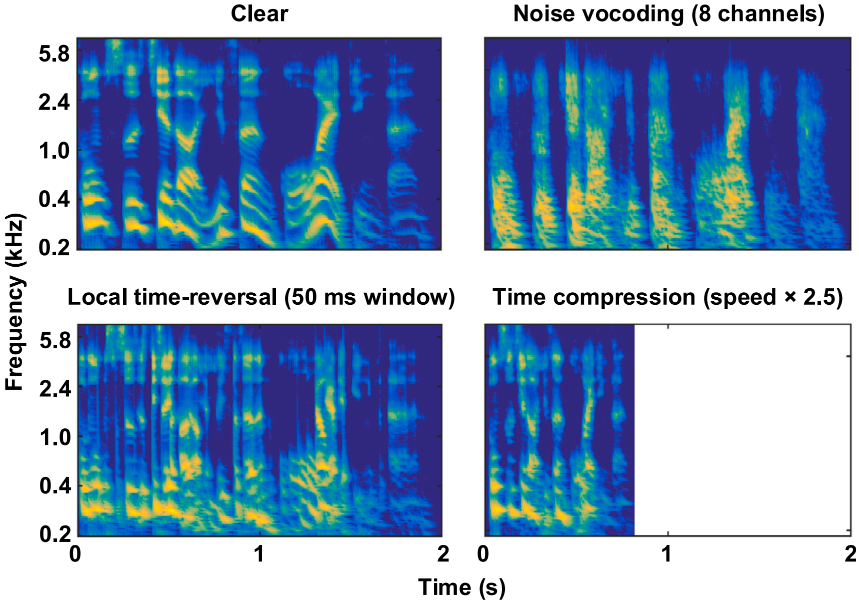


Figure 2. Spectrograms for clear speech and degraded speech conditions for an example sentence, "公司接到一份国外订单" ("The company received a foreign order")

speech recognition accuracy. Specifically, both the ASR system and humans demonstrate near-perfect accuracy in clear speech, with a noticeable drop and subsequent recovery in degraded conditions (Figure 3). In all three easy cases of acoustic degradation, the fitted speech recognition accuracy at the final degraded sentence position of human participants and the ASR system is comparable, as Figure 3's top row details. However, a notable distinction lies in the adaptation speed. The time constant of the exponential curve, indicative of the speed of adaptation (see STAR Methods: Performance evaluation for details), varies between humans without external feedback and the self-supervised ASR system, as shown in the "τ" rows of Table 1. Despite this, after adapting to 20 degraded speeches, their speech recognition accuracies converge, suggesting a similarity in overall adaptation capacity.

In hard cases involving noise vocoding (4 channels) and time compression (speed × 3.0), both human participants and the ASR system each exhibit distinct strengths in speech recognition accuracy, as detailed in Figure 3's bottom row. Specifically, differences in accuracy are notable at the final degraded sentence position. The adaptation speeds, as indicated by the "τ" rows in Table 1, are found to be comparable between the self-supervised ASR system and humans without external feedback. In the hard case of local time-reversal (62 ms window), the ASR system demonstrates no discernible signs of adaptation, whereas human listeners exhibit minor adaptation responses (refer to Figure 3).

External linguistic feedback is not critical for human adaptation to acoustically degraded speech

This study further assessed the influence of external linguistic feedback on human adaptation to acoustically degraded speech. In a second experiment, human participants received external linguistic feedback (refer to Figure 1). The experiment also included 180 participants, divided into six groups of 30 participants, each group corresponding to a single type of acoustic degradation. These participants, after typing their responses to the degraded speech, received the correct transcription of the speech as external feedback. They were then able to listen to the degraded speech again after knowing the correct transcription (see Figure 1).

Unpaired *t*-test analysis reveals that there is no significant difference between the speech recognition accuracy of humans with or without receiving external feedback ($p > 0.05$, unpaired *t*-test, FDR corrected) in all six tested conditions. This result indicates that human adaptation to acoustically degraded speech is not affected by external linguistic feedback (Figure 4).

Automatic speech recognition system benefits more from supervised learning than humans do

We subsequently investigated whether the ASR system gains more from supervised learning using correct transcription than humans do, given the humans' minimal improvement from external linguistic feedback. Statistical analysis reveals significant differences between self-supervised and supervised learning of ASR systems in four conditions (noise vocoding (4 channels), $p = 0.03$; time compression (speed × 3.0), $p = 0.01$; local time-reversal (50 ms window), $p = 9.60 \times 10^{-6}$; local time-reversal (62 ms window), $p = 1.09 \times 10^{-14}$, unpaired *t*-test). For the other two types of degraded speech, the statistical analysis shows no significant difference between these two learning strategies (noise vocoding (8 channels), $p = 0.26$; time compression (speed × 2.5), $p = 0.14$, unpaired *t*-test).

As shown in Figure 5, the supervised ASR system improves speech recognition accuracy under all tested conditions, even in the hard case of local time-reversal (62 ms window) that the self-supervised ASR system fails. In this difficult scenario, the supervised ASR system improves

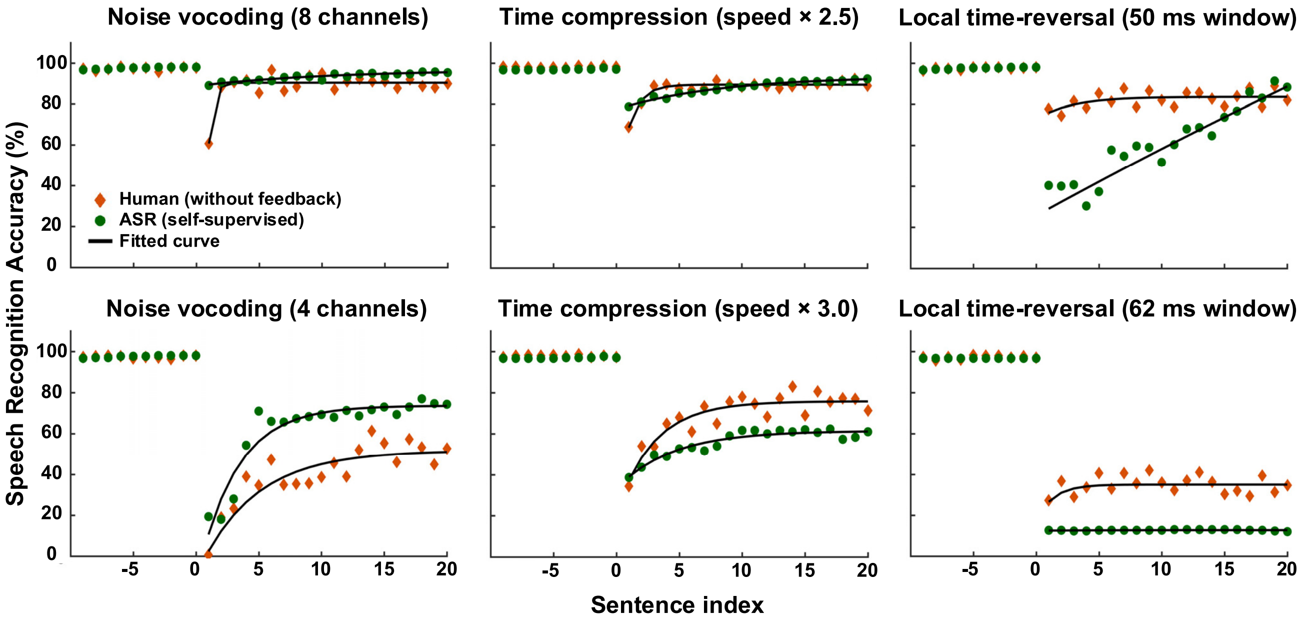


Figure 3. Speech recognition accuracy of humans without external feedback and the self-supervised ASR system, as a function of sentence index

Sentence #1 represents the first degraded speech that the human participants hear. The experiment involved two levels of difficulty in degraded conditions, namely easy (top row) and hard (bottom row). The speech recognition accuracy drops when the stimulus switches from clear speech to degraded speech, and recovers to some extent when the listener is exposed to more sentences. The recovery of speech recognition accuracy is fitted by an exponential function (black curve).

accuracy by more than 65% compared to the self-supervised ASR system. In comparison, human participants who received external feedback only achieve approximately a 20% improvement over those who did not. Additionally, the supervised ASR system adapts more rapidly than the self-supervised ASR system in the hard case of noise vocoding (4 channels) and easy case of local time-reversal (50 ms window), achieving a similar level of accuracy while requiring fewer sentences (Figure 5).

DISCUSSION

The current study employed a DNN-based ASR system to explore how humans recognize unfamiliar degraded speech at the computational level, following Marr's three levels of analysis.[16] Arguably, the DNN-based ASR system has a similar computational goal as the human speech recognition system.[22,23] We have devised and assessed two computational strategies to emulate human adaptation to acoustically degraded speech, involving with or without external linguistic feedback. We want to test whether linguistic feedback alone is computationally "sufficient" to drive perceptual learning. It certainly does not indicate that linguistic feedback is the only cue to drive perceptual learning or that it is the only cue that is used by human listeners. It concerns the computationally capacity of linguistic feedback – In the absence of other feedback, can linguistic feedback alone lead to improvement in speech recognition rate that is comparable to what is observed for human listeners? As the study shows, the answer to this question is yes.

Our results demonstrate that the self-supervised learning based on linguistic feedback is a plausible principle for human adaptation to acoustically degraded speech. Humans can quickly learn to recognize acoustic degradation after exposure to just a few sentences, and external linguistic feedback is not critical for their rapid adaptation. By using self-supervised learning, the ASR system can adapt to acoustically degraded speech in most cases. The adaptation speed and speech recognition accuracy are similar to those of humans. The ASR system's own transcription of speech functions as the top-down linguistic feedback that drives learning. Furthermore, the ASR system benefits more from supervised learning than humans, indicating that humans are less sensitive to external feedback compared to the ASR system.

Human speech recognition is robust against diverse forms of acoustic degradation.[1,3,24,25] Consistent with previous research,[1,3] our results show that humans can quickly adapt to acoustically degraded speech after exposure to a few sentences. Furthermore, our results suggest that the adaptation speed of human listeners varies between easy and hard conditions. Apart from locally time-reversed speech, humans adapt more rapidly in the easy condition than in the hard condition, as shown by the fitted time constant (see Table 1). Conversely, ASR systems adapt more rapidly in the hard conditions. Since speech with mild degradation is frequently encountered in daily life, it is possible that the human speech processing system has learned to adapt to these mild changes. Based on the reverse hierarchy theory,[26] humans may only adapt basic auditory processing in the easy conditions but may have to adapt higher-level phonetic processing in the hard conditions, and it is much more time consuming to adapt higher-level speech processing. Unlike the human speech processing system, current

Table 1. Fitted parameters of the exponential function of adaptation to acoustically degraded speech

	Type	Human (without feedback)	Model (self-supervised)	Human (with feedback)	Model (supervised)
τ	NV (8 channels)	0.37	10.52	0.60	10.71
	NV (4 channels)	4.49	3.11	2.79	0.78
	TC (speed × 2.5)	0.98	8.26	0.37	2.41
	TC (speed × 3.0)	2.97	4.43	1.60	5.39
	LR (50 ms window)	2.76	155.92	2.93	2.40
	LR (62 ms window)	1.35	4.17	1.16	5.77
A	NV (8 channels)	90.36	96.69	91.26	96.45
	NV (4 channels)	51.72	73.65	56.44	78.08
	TC (speed × 2.5)	89.48	93.61	86.87	92.03
	TC (speed × 3.0)	75.84	61.45	69.62	69.84
	LR (50 ms window)	83.63	549	86.85	92.28
	LR (62 ms window)	35.02	12.74	48.15	85.73
B	NV (8 channels)	451.48	7.95	171.38	9.27
	NV (4 channels)	61.70	86.92	76.87	206.92
	TC (speed × 2.5)	58.59	16.27	369.10	22.53
	TC (speed × 3.0)	53. 95	28.07	72.47	32.87
	LR (50 ms window)	11.16	523.36	26.11	64.03
	LR (62 ms window)	17.68	0.29	48.97	78.15

NV, noise vocoding; TC, time compression; LR, local time-reversal. "*A*," "*B*," and "τ" are the fitted parameters of the exponential curve for a function of the form $z = A - B \exp(-i/\tau)$, where τ indicates the time constant, *i* represents the sentence position, and *z* is the fitted speech recognition accuracy at that position. The time constant is inversely related to the speed of the rapid adaptation: the smaller the time constant, the faster the adaptation speed. See also Table S1.

DNN-based ASR systems do not have a clearly defined processing hierarchy and the samples in the hard condition deviate more from clear speech and therefore may drive stronger learning effects in DNN.[23]

A previous study demonstrates that performance was enhanced through the utilization of external feedback in adaptation to six-channel noise-vocoded speech.[1] A critical difference between the current study and the previous study by Davis et al.[1] is that we presented very

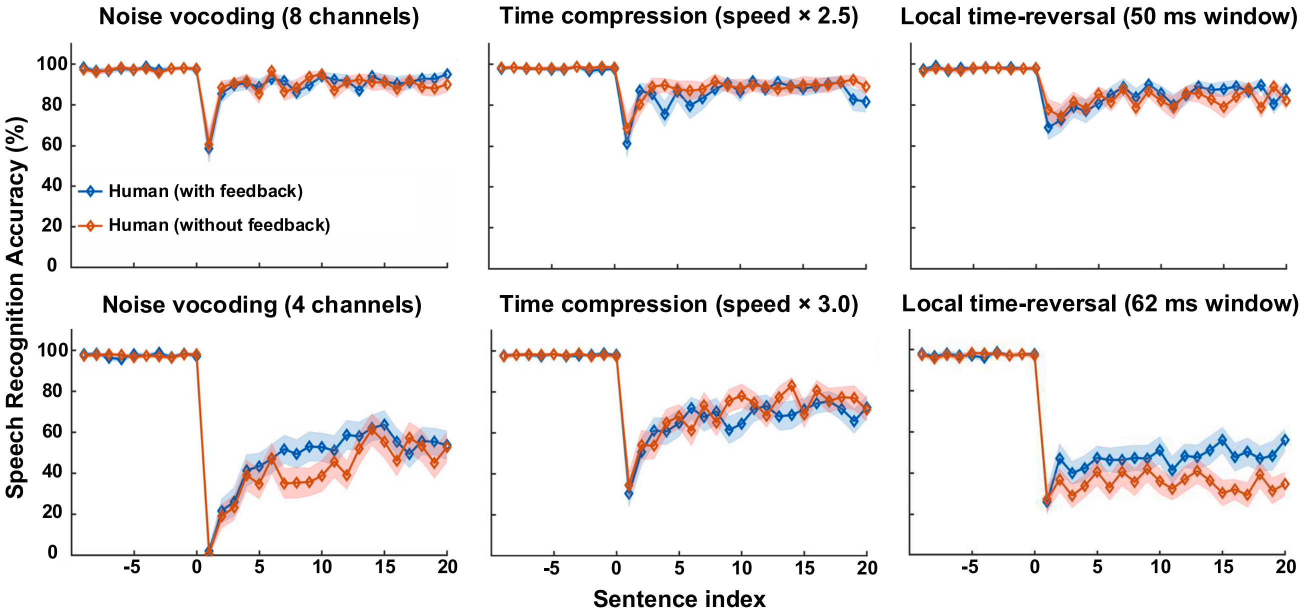


Figure 4. Speech recognition accuracy of humans as a function of sentence index, showing minimal effects from external linguistic feedback

Lines and shaded error bars indicate means ± SE. Same convention as in Figure 3.

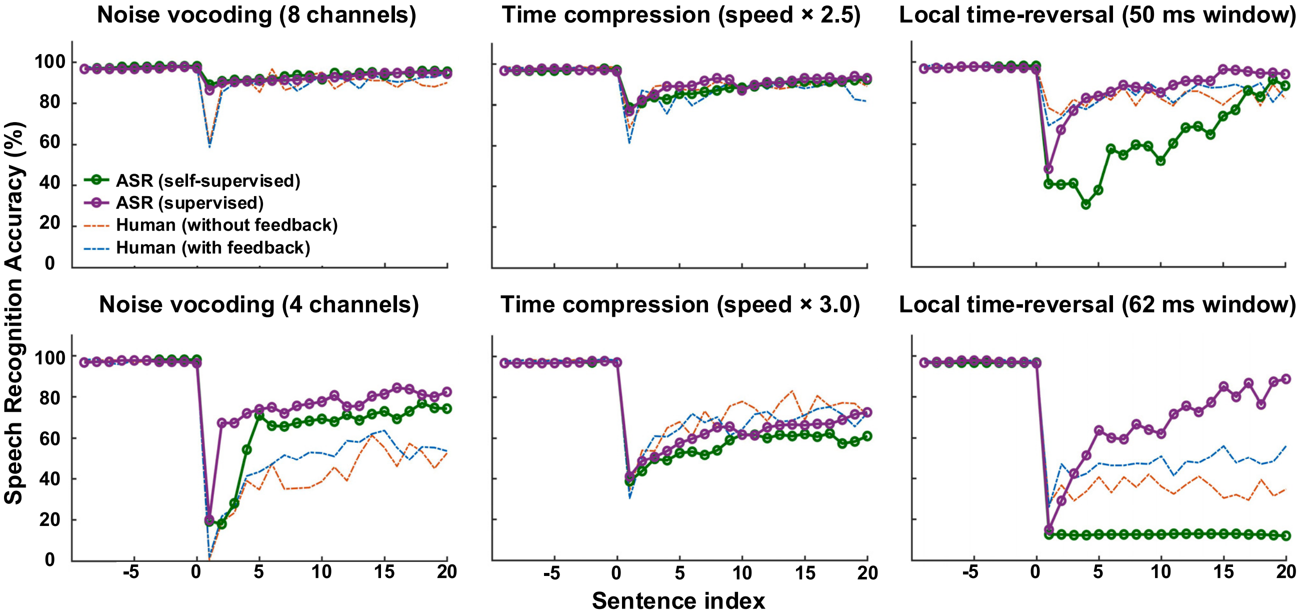


Figure 5. Speech recognition accuracy of the ASR systems as a function of sentence index

Human data (dashed lines) are shown for comparison. Same convention as in Figure 3.

high-context sentences and each sentence has 10 syllables while Davis et al.[1] presented unambiguous sentences, each of which has a variable number of words. It is possible that contextual information can strongly modulate the effects of external feedback. For high-context sentences, the listeners can use contextual information to infer words they cannot recognize based on just auditory information, rendering the external feedback highly redundant. Therefore, it is quite likely that the benefit of external feedback is weaker for high-context sentences compared with low-context sentences or word lists. Additionally, the language can also make a difference. Chinese, the language tested here, has a much smaller number of syllables compared with English, the language tested by Davis et al.,[1] the smaller pool of syllables can also provide a kind of contextual information.

Limitations of the study

Our study only tested Chinese and three types of acoustic degradation. Future studies can test whether the conclusions here generalize to other languages and other types of speech degradation. Similarly, future studies can test different types of ASR systems and different ways to fine-tune the systems, and even attempt to design brain-inspired methods that can better explain the human adaptation effects. Additionally, human speech perception is highly complex and the transition from clear to degraded speech may have violated the listener's expectation and therefore distract their attention. Future studies have to analyze whether attention and other factors can influence human performance. Here, we only used high-context sentences, and the adaptation rate of both humans and ASR systems may vary when listening to low-context sentences or even word lists. Future studies can explore how context modulates human and ASR adaptation to degraded sentences.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - ○ Participants
- METHOD DETAILS
  - ○ Degraded speech
  - ○ Human experiment
  - ○ ASR experiment

- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Performance evaluation
  - ○ Statistical analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.110055.

ACKNOWLEDGMENTS

We thank Yang Guo, Yaoyao Wang, and Cheng Luo for valuable comments on earlier versions of this article, Wenyuan Yu and Yunyi Qi for thoughtful discussions. This work was supported by the National Key Research and Development Program of China (No.2021ZD0201501).

AUTHOR CONTRIBUTIONS

W.S.: Data curation, formal analysis, investigation, visualization methodology, writing-original draft, and writing-review and editing. J.Z.: Methodology, visualization, and writing-original draft. T.Z: Investigation. Z.S.: Methodology and visualization. N.D.: Conceptualization, methodology, supervision, writing-original draft, and writing-review and editing.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 27, 2023
Revised: March 15, 2024
Accepted: May 17, 2024
Published: May 22, 2024

REFERENCES

1. Davis, M.H., Johnsrude, I.S., Hervais-Adelman, A., Taylor, K., and McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. J. Exp. Psychol. Gen. *134*, 222–241.
2. Hervais-Adelman, A., Davis, M.H., Johnsrude, I.S., and Carlyon, R.P. (2008). Perceptual learning of noise vocoded words: effects of feedback and lexicality. J. Exp. Psychol. Hum. Percept. Perform. *34*, 460–474.
3. Cooke, M., Scharenborg, O., and Meyer, B.T. (2022). The time course of adaptation to distorted speech. J. Acoust. Soc. Am. *151*, 2636. https://doi.org/10.1121/10.0010235.
4. Rotman, T., Lavie, L., and Banai, K. (2020). Rapid Perceptual Learning: A Potential Source of Individual Differences in Speech Perception Under Adverse Conditions? Trends in Hearing *24*, 2331216520930541.
5. Bent, T., Buchwald, A., and Pisoni, D.B. (2009). Perceptual adaptation and intelligibility of multiple talkers for two types of degraded speech. J. Acoust. Soc. Am. *126*, 2660–2669.
6. Norris, D., McQueen, J.M., and Cutler, A. (2003). Perceptual learning in speech. Cogn. Psychol. *47*, 204–238. https://doi.org/10.1016/S0010-0285(03)00006-9.
7. Huyck, J.J., and Johnsrude, I.S. (2012). Rapid perceptual learning of noise-vocoded speech requires attention. J. Acoust. Soc. Am. *131*, EL236-42.
8. Sohoglu, E., and Davis, M.H. (2020). Rapid computations of spectrotemporal prediction error support perception of degraded speech. Elife *9*, e58077.
9. Neger, T.M., Rietveld, T., and Janse, E. (2014). Relationship between perceptual learning in speech and statistical learning in younger and older adults. Front. Hum. Neurosci. *8*, 628.
10. Cao, S., Kang, Y., Fu, Y., Xu, X., Sun, S., Zhang, Y., and Ma, L. (2021). Improving Streaming Transformer Based ASR Under a Framework of Self-supervised Learning. Preprint at arXiv. https://doi.org/10.48550/arXiv.2109.07327.
11. Lee, D.-H. (2013). Pseudo-Label: The simple and efficient semi-supervised learning method for deep neural networks. In Workshop on challenges in representation learning, *vol. 3* (ICML), p. 896.
12. Hwang, D., Misra, A., Huo, Z., Siddhartha, N., Garg, S., Qiu, D., Sim, K.C., Strohman, T., Beaufays, F., and He, Y. (2021). Large-Scale ASR Domain Adaptation Using Self- and Semi-Supervised Learning. In ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6627–6631.
13. Zhang, Y., and Davison, B.D. (2021). Efficient Pre-trained Features and Recurrent Pseudo-Labeling in Unsupervised Domain Adaptation. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2713–2722.
14. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. Preprint at ArXiv. https://doi.org/10.48550/arXiv.2212.04356.
15. Zhang, Y., Han, W., Qin, J., Wang, Y., Bapna, A., Chen, Z., Chen, N., Li, B., Axelrod, V., Wang, G., et al. (2023). Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages. Preprint at ArXiv. https://doi.org/10.48550/arXiv.2303.01037.
16. Marr, D. (2010). Vision: A computational investigation into the human representation and processing of visual information (MIT press).
17. Ding, N., Gao, J., Wang, J., Sun, W., Fang, M., Liu, X., and Zhao, H. (2023). Speech recognition in echoic environments and the effect of aging and hearing impairment. Hear. Res. *431*, 108725. https://doi.org/10.1016/j.heares.2023.108725.
18. Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). Speech Recognition with Primarily Temporal Cues. Science *270*, 303–304. https://doi.org/10.1126/science.270.5234.303.
19. Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., and Pang, R. (2020). Conformer: Convolution-augmented Transformer for Speech Recognition. Preprint at ArXiv. https://doi.org/10.48550/arXiv.2005.08100.
20. Zhang, B., Wu, D., Peng, Z., Song, X., Yao, Z., Lv, H., Xie, L., Yang, C., Pan, F., and Niu, J. (2022). WeNet 2.0: More productive end-to-end speech recognition toolkit. Preprint at arXiv. arXiv:2203.15455.
21. Zhang, B., Lv, H., Guo, P., Shao, Q., Yang, C., Xie, L., Xu, X., Bu, H., Chen, X., Zeng, C., et al. (2021). WENETSPEECH: A 10000+ Hours Multi-Domain Mandarin Corpus for Speech Recognition. In ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6182–6186.
22. Deng, L., and Li, X. (2013). Machine Learning Paradigms for Speech Recognition: An Overview. IEEE Trans. Audio Speech Lang. Process. *21*, 1060–1089.
23. Li, J. (2021). Recent Advances in End-to-End Automatic Speech Recognition. Preprint at ArXiv. https://doi.org/10.48550/arXiv.2111.01690.

24. Mesgarani, N., David, S.V., Fritz, J.B., and Shamma, S.A. (2014). Mechanisms of noise robust representation of speech in primary auditory cortex. Proc. Natl. Acad. Sci. USA *111*, 6792–6797.

25. Ding, N., and Simon, J.Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. Proc. Natl. Acad. Sci. USA *109*, 11854–11859.

26. Ahissar, M., and Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. Trends Cogn. Sci. *8*, 457–464.

27. Ellis, D. (2002). A Phase Vocoder in Matlab. http://www.ee.columbia.edu/∼dpwe/resources/matlab/pvoc/.

28. Wong, L.L.N., Soli, S.D., Liu, S., Han, N., and Huang, M.-W. (2007). Development of the Mandarin Hearing in Noise Test (MHINT). Ear Hear. *28*, 70S–74S.

29. Yao, Z., Wu, D., Wang, X., Zhang, B., Yu, F., Yang, C., Peng, Z., Chen, X., Xie, L., and Lei, X. (2021). WeNet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit. Preprint at arXiv. arXiv:2102.01547.

30. Graves, A., Fernández, S., Gomez, F.J., and Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd international conference on Machine learning, pp. 369–376.

31. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B *57*, 289–300.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Human and ASR data	This paper	https://github.com/1633347510/human-rapid-adaptation/tree/80e7238587f960aa598285b6b772f614fd68ccc5/Human%20and%20ASR%20data
Software and algorithms		
MATLAB R2016a	MathWorks	RRID:SCR_001622
Conformer model	WeNet	https://docs.qq.com/form/page/DZnRkVHlnUk5QaFdC
Cognition	Cognition.	https://www.cognition.run/
pvoc	Ellis[27]	https://www.ee.columbia.edu/∼dpwe/resources/matlab/pvoc/

RESOURCE AVAILABILITY

Lead contact

Further information and requests should be directed to the lead contact, Nai Ding (ding_nai@zju.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Anonymized human behavioral data and ASR experimental data have been deposited on GitHub, with the specific URL listed in the key resources table.
- The code used in this study can be accessed at GitHub (https://github.com/1633347510/human-rapid-adaptation.git).
- Any additional information required is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Participants

Three hundred and seventy-nine (197 females; mean age, 21.34 ± 2.60 years old) normal hearing native-Mandarin listeners were recruited. Experimental data of 19 human listeners were excluded due to unexpected interruptions during the experiment. The experimental protocol received approval from the Ethics Committee of the College of Biomedical Engineering and Instrument Sciences, Zhejiang University (No. 2022-001). Each participant was informed of the content of the experiment before data collection and received monetary payments after the experiment for their participation.

METHOD DETAILS

Degraded speech

Three types of degraded speech were used in our experiment, which were representatives of degradation produced by manipulating the acoustic information in spectral (noise vocoding) and temporal (time compression, local time-reversal) domains.[3,18,27] Every type of degraded speech had two levels of difficulty: easy and hard, resulting in six unique acoustic degradation scenarios. The varying difficulty levels posed different challenges to speech intelligibility, enabling us to gain a more comprehensive understanding of adaptation to acoustic degradation. Importantly, since these types of acoustic degradation, especially local time-reversed speech, are relatively uncommon for both human listeners and ASR systems, we leveraged these manipulations to study the adaptation processes of both human listeners and ASR systems.

For the noise vocoding condition,[18] speech was first decomposed into 8 or 4 frequency bands with cutoff frequencies equally spaced on the equivalent rectangular bandwidth scale between 123 and 3951 Hz. Within each frequency band, the amplitude envelope, obtained using the Hilbert transform, was used to modulate white noise filtered to the same frequency band. The modulated noise from all bands was then summed to obtain noise-vocoded speech. For the time-compressed speech, speech materials were created using the pvoc method,[27] a fast Fourier transform (FFT) based phase vocoder. The speech rate was set to 2.5 or 3.0, with an FFT window size of 512 samples. For the locally time-reversed speech, speech materials were obtained by locally time-reversing successive non-overlapping segments of speech without applying any windowing, with a segmentation duration of 50 or 62 ms. The root mean square (RMS) intensity of degraded speech was adjusted to match that of clear speech.

Human experiment

Stimulus

The recordings were extracted from the Mandarin Hearing in Noise Test (MHINT) dataset, which is widely utilized for assessing the ability to interpret speech in both quiet and noisy conditions.[28] Each sentence in the MHINT dataset contains 10 Chinese characters. The sentences are high-context sentences and are designed to be easily understood by individuals with varying educational backgrounds. The sentence difficulty in the MHINT dataset was equalized.[28] All the speeches were recorded by a single native male speaker and resampled at 16 kHz for further use. These sentences had an average duration of 2.60 s (range, 2.16–3.64; s.d., 0.20 s). Clear speech and six types of degraded speech were employed in human experiments (see STAR Methods: Degraded speech for details).

Procedure

The online experiment was conducted via a webpage, utilizing the Cognition platform (https://www.cognition.run/). Participants were informed in advance that they might encounter some degraded speech samples during the course of the experiment. They were directed to listen carefully and type as many words as they could.

Before the formal session, participants were familiarized with three clear MHINT sentences that were not used in the formal session. During the formal session, participants were presented with 30 non-repetitive sentences. The first 10 were clear sentences, and the last 20 were degraded sentences of the same type. To control for individual sentence difficulty variability, the sentences were presented following the Latin Square Design,[17] which was applied on the sentences themselves: We generated 30 distinct stimulus sequences from 30 non-repetitive sentences, applying a cyclic shift to one sentence at a time. Since each group comprised 30 participants, each individual was assigned one of the resulting 30 orderings consecutively (refer to Figure S1 in supplemental information for more details). At each sentence index, all 30 possible sentences were presented within a listener group. By employing this design, when averaging speech recognition accuracy across a listener group, the mean scores at each sentence index were not influenced by individual sentence difficulty variations.

Each sentence was preceded by a three-second countdown page, after which an input box appeared. Participants were instructed to type as many words as they could, or alternatively type “I didn’t catch that.” In the experiment involving external feedback, after participants finished the transcription task, they heard the same sentence without any degradation, followed by two repetitions of the same degraded sentence they heard before. Concurrently, the corresponding correct transcription was displayed on-screen during the audio presentation (refer to Figure 1).

ASR experiment

ASR system

The Conformer-based ASR system employed in this study was trained on the multi-domain WenetSpeech corpus[21] using the WeNet toolkit.[20,29] The dataset comprised over 10,000 hours of labeled data. The ASR system utilized a connectionist temporal classification (CTC)/attention architecture with Conformer as the Encoder.[30] Among the four decoding modes supported by this ASR system, we used the attention-rescoring mode, which typically yields the best performance.[20,29] The link to the training configuration file of the pretrained ASR model is available in the key resources table.

Data

To mimic human adaptation to acoustically degraded speech, the ASR system underwent 30 fine-tuning iterations. For each iteration, we created distinct training and test sets. Each training set contained one MHINT speech, while each test set contained 30 MHINT speeches. These training sets utilized a total of 30 MHINT sentences that differed from those used in the human experiments. The sentences used in the test set overlapped with those used in the human experiments.

Computational speech-adaptation strategy

Only the weights of encoder-layers were updated during rapid learning of degraded speech. Five epoch sizes were tested: 1, 5, 25, 125, and 625. Three learning rates were tested: 2.00×10^{-3}, 2.00×10^{-4}, and 2.00×10^{-5}. Multiple parameter combinations of learning rate and epoch size were tested, and only the one with the highest accuracy was reported.

Self-supervised learning. Speech input was initially recognized by an ASR system, and the resulting recognition result, termed ASR transcription, was paired with the original speech input to fine-tune the ASR system in a supervised manner.

Supervised learning. Both paired correct transcription and speech input were used to fine-tune the ASR system in a supervised manner.

QUANTIFICATION AND STATISTICAL ANALYSIS

Performance evaluation

The speech recognition accuracy was used to evaluate the performance of both humans and ASR systems in speech transcription tasks, and it was calculated as 100% minus character error rate (CER). The CER score indicates the percentage of characters that were incorrectly recognized. If the speech recognition accuracy was negative, it was set to zero. In the ASR experiment, multiple parameter sets were tested, and only the highest accuracy among them was reported (refer to Computational speech-adaptation strategy). For parameter comparison, the average value of the last five degraded speech samples was calculated.

The change in speech recognition accuracy, as a function of the number of degraded speeches presented, was modeled using an exponential curve. Speech recognition accuracy z at i^{th} sentence was formulated as follows:

$$z = A - B \exp(-i/\tau)$$

where, A, B, and τ were parameters to be fitted using the least squares method. The coefficient τ represents the time constant, which can be used to describe the speed at which the accuracy improves.

Statistical analysis

MATLAB R2016a (RRID:SCR_001622) was used for statistical analysis. To assess whether there was significant difference in speech recognition accuracies between human listeners without external feedback and the self-supervised ASR system, as well as between human listeners with external feedback and the supervised ASR system, we performed the single-sample *t*-test at each sentence index (from 1st to 20th in Figures S2 and S3). The speech recognition accuracies of human participants who received external feedback and those who did not, were compared using an unpaired *t*-test at each sentence index (from 1st to 20th in Figure 4). Additionally, we conducted unpaired *t*-test to compare the speech recognition accuracies of self-supervised and supervised ASR systems (from 1st to 20th in Figure 5). Multiple pairwise comparisons were corrected using the FDR method.[31] In all cases, *p* values lower than 0.05 were considered statistically significant.