Hearing Research 431 (2023) 108725

Contents lists available at ScienceDirect

Hearing Research

journal homepage: www.elsevier.com/locate/heares

# Speech recognition in echoic environments and the effect of aging and hearing impairment

Nai Ding[a], Jiaxin Gao[a], Jing Wang[a], Wenhui Sun[b], Mingxuan Fang[a], Xiaoling Liu[a], Hua Zhao[a,*]

[a] College of Biomedical Engineering and Instrument Science,Department of Nursing, The Second Affiliated Hospital of Zhejiang University School of Medicine, Zhejiang University, Hangzhou, Zhejiang, China
[b] Research Center for Applied Mathematics and Machine Intelligence, Research Institute of Basic Theories, Zhejiang Lab, Hangzhou, Zhejiang, China

ARTICLE INFO

*Article history:*
Received 30 November 2022
Revised 12 February 2023
Accepted 23 February 2023
Available online 26 February 2023

*Keywords:*
Echo
Modulation spectrum
Aging
Hearing impairment

ABSTRACT

Temporal modulations provide critical cues for speech recognition. When the temporal modulations are distorted by, e.g., reverberations, speech intelligibility drops, and the drop in speech intelligibility can be explained by the amount of distortions to the speech modulation spectrum, i.e., the spectrum of temporal modulations. Here, we test a condition in which speech is contaminated by a single echo. Speech is delayed by either 0.125 s or 0.25 s to create an echo, and these two conditions notch out the temporal modulations at 2 or 4 Hz, respectively. We evaluate how well young and older listeners can recognize such echoic speech. For young listeners, the speech recognition rate is not influenced by the echo, even when they are exposed to the first echoic sentence. For older listeners, the speech recognition rate drops to less than 60% when listening to the first echoic sentence, but rapidly recovers to above 75% with exposure to a few sentences. Further analyses reveal that both age and the hearing threshold influence the recognition of echoic speech for the older listeners. These results show that the recognition of echoic speech cannot be fully explained by distortions to the modulation spectrum, and suggest that the auditory system has mechanisms to effectively compensate the influence of single echoes.



## 1. Introduction

Reverberation is prevalent in in-door communication environments, and it can strongly influence speech intelligibility. Recently, echoing, which can be viewed as a special kind of reverberation, has also received a significant amount of attention, since it frequently occurs during online communication: Speech from one speaker can be picked up by the microphone of another speaker and transmitted back, creating an echo (Zhang and Wang, 2022). Echoes can significantly reduce the quality of online conversation, and a number of contests have recently been hosted to develop algorithms to effectively cancel echoes in the field of audio signal processing (Sridhar et al., 2020;Cutler et al., 2022). Studies on reverberation have also considered the effect of single reflections and demonstrate that, in a noisy environment, an early reflection (e.g., with the delay below 25 ms) can be integrated with the direct sound and boosts speech intelligibility while a late reflection (e.g., with the delay beyond 100 ms) can reduce speech intelligibility (Warzybok et al., 2013).

The influence of reverberation on speech intelligibility is well studied and it has been demonstrated that reverberation influences intelligibility by altering the modulation spectrum of speech (Houtgast and Steeneken, 1985, 1973). The modulation spectrum is the power spectrum of the speech envelope, i.e., the temporal fluctuations of sound power (Rosen, 1992;Chi et al., 1999;). When reverberation occurs, the dynamic range of the speech envelope is compressed: For clean speech, the dynamic range of the speech envelope is broad, since speech contains high-energy segments, e.g., vowels, as well as low-energy segments and silence periods. When reverberation occurs, the speech envelope is smeared since acoustic reflections fill in the low-energy segments and silence periods of speech. Furthermore, the reduction in dynamic range depends on the modulation rate. For clean speech, the modulation spectrum has a stereotyped shape that peaks around 4–5 Hz (Fig. 2, Greenberg, 1999;Greenberg et al., 2003;Ding et al., 2017;Varnet et al., 2017). When reverberation occurs, the modulation spectrum is altered (Houtgast and Steeneken, 1973, 1985;Steeneken and Houtgast, 1980), and measures such as the speech transmission index (STI) can automatically predict the intelligibility of reverberant speech by analyzing how much the speech modulation spectrum (ranged between 0.5

* Corresponding author.
*E-mail address:* 2503125@zju.edu.cn (H. Zhao).

https://doi.org/10.1016/j.heares.2023.108725
0378-5955/

and 16 Hz) is altered by reverberation. These measures, however, are less precise when predicting speech intelligibility for hearing impaired listeners (Payton et al., 1994).

The modulation spectrum analysis is originally motivated by research on reverberated speech but it has also been demonstrated to be useful to predict intelligibility in many other challenging listening environments, suggesting that it is a fundamental feature related to speech perception (Dau et al., 1997a, 1997b;Elhilali et al., 2003;Jørgensen and Dau, 2011). When the speech envelope is directly manipulated, i.e., being processed by a digital filter, speech intelligibility also significantly reduces: A number of studies have demonstrated that temporal modulations between 1 and 16 Hz are critical for intelligibility (Chi et al., 1999, 2005). Filtering the speech envelope, however, is a challenging task since for some methods the cochlea can recover the speech envelope even when it is filtered out from the stimulus (Ghitza, 2001;Zeng et al., 2004), while other methods also degrade the fine-structure of speech on top of the envelope (Ghitza, 2001).

Reverberation can be decomposed into a superposition of echoes of different delays, and therefore individual echoes can be viewed as the basic elements of reverberation. A single echo can also strongly influence the speech envelope (Fig. 3, further illustrated in Section 2.3), and the influence is highly frequency specific. For example, if speech is contaminated by a 125-ms echo, its temporal modulations at 4 Hz are selectively attenuated. Therefore, in the current study, we used single echoes as a simple but effective way to manipulate the speech modulation spectrum and test whether speech intelligibility can be explained by the modulation spectrum. Furthermore, previous studies have shown that speech recognition in reverberant environments is influenced by age and hearing impairment, we tested both young and older listeners (Harris and Reitz, 1985;Helfer and Wilber, 1990;Marrone et al., 2008;Nábělek and Robinson, 1982). Additionally, recent studies have shown that speech recognition in adverse listening environments is a dynamic process: For some types of degraded speech, the speech recognition rate increases with more exposure, even without any feedback or instruction (Peelle and Wingfield, 2005;Cooke et al., 2022). Therefore, we employed a design that could reveal how the speech recognition accuracy might vary over time during the exposure to echoic speech (Fig. 4B).

In the following, we first demonstrate that the STI, as well as two other common speech intelligibility measures, predict that a single echo strongly reduces speech intelligibility. Next, we measure the speech recognition rate from both normal-hearing young listeners and hearing-impaired older listeners. We hypothesize that the speech recognition rate of normal-hearing young listeners matches the prediction of STI, after the exposure to a few sentences, and the recognition rate of hearing-impaired older listeners falls below the prediction of STI. The empirical data, however, contradict the hypotheses, suggesting that current speech intelligibility models such as the STI cannot well explain the effects of single echoes.

## 2. Methods

### 2.1. Listeners

Totally, sixty adults participated in this experiment. Thirty participants were young listeners (15 males, 20–30 years old,mean age, 23 years old), while the other thirty participants were older listeners (6 males, 51–95 years old,mean age, 71 years old). Both the young and older listeners were native Mandarin listeners. Informed consents were obtained from all participants before the experiment, and the experimental protocol was approved by the Ethics Committee of the College of Biomedical Engineering & Instrument, Zhejiang University (No. 2022–001). The air conduction threshold was measured from all listeners bilaterally at octave intervals between 0.5 and 4 kHz. The threshold was averaged across ears and shown in Fig. 1. For the young listeners, the pure tone averages (PTA,average threshold from 0.5 to 4 kHz) were 5.85 dB HL (range: −1.25 dB HL – 15 dB HL) (Fig. 1). Specifically, twenty-nine listeners had hearing threshold within 20 dB HL at all audiometric frequencies, and one listener had slight hearing loss for the left ear (25 dB HL at 2 kHz, and 30 dB HL at 4 kHz). For the thirty older listeners, the bilaterally average PTAs were 31.69 dB HL (range: 11.25 dB HL – 63.13 dB HL).

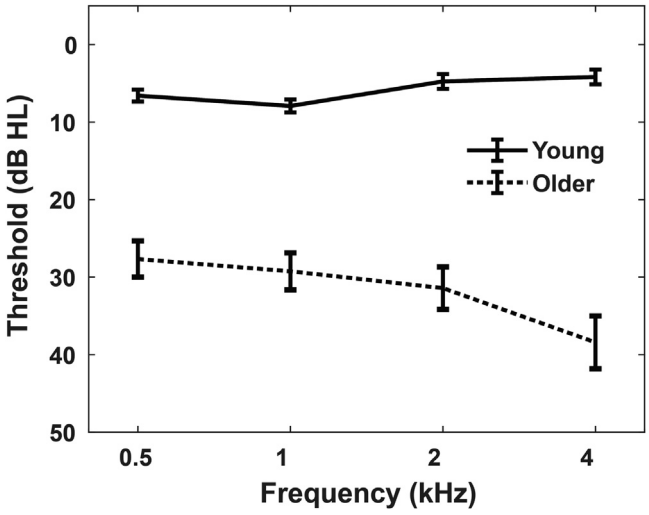


**Fig. 1.** Audiogram averaged across ears. Error bars represent one standard error of the mean (SEM).

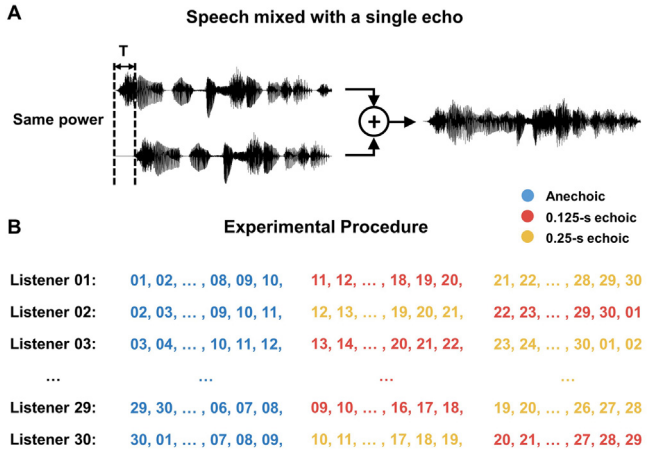


**Fig. 2.** (A), Stimulus. Echoic speech is generated by adding a delayed copy of speech to the original speech sound. (B), Experimental procedure. The order of sentences is circularly shifted by one for each listener. The color represents stimulus condition. Anechoic speech is presented first, followed by the two echoic conditions, the order of which is counterbalanced cross listeners.

### 2.2. Stimuli

We employed the Mandarin Hearing in Noise Test (MHINT) dataset to test intelligibility (Wong et al., 2007). The sentences of MHINT represented simple conversational style speech, and were easy to understand by people with varied educational background. Thirty sentences were used in this experiment, each containing ten syllables and on average lasted 2.58 s (2.20–3.23 s for individual sentences). Each sentence was spoken by a male talker and digitized at a sampling rate of 48,000 Hz.

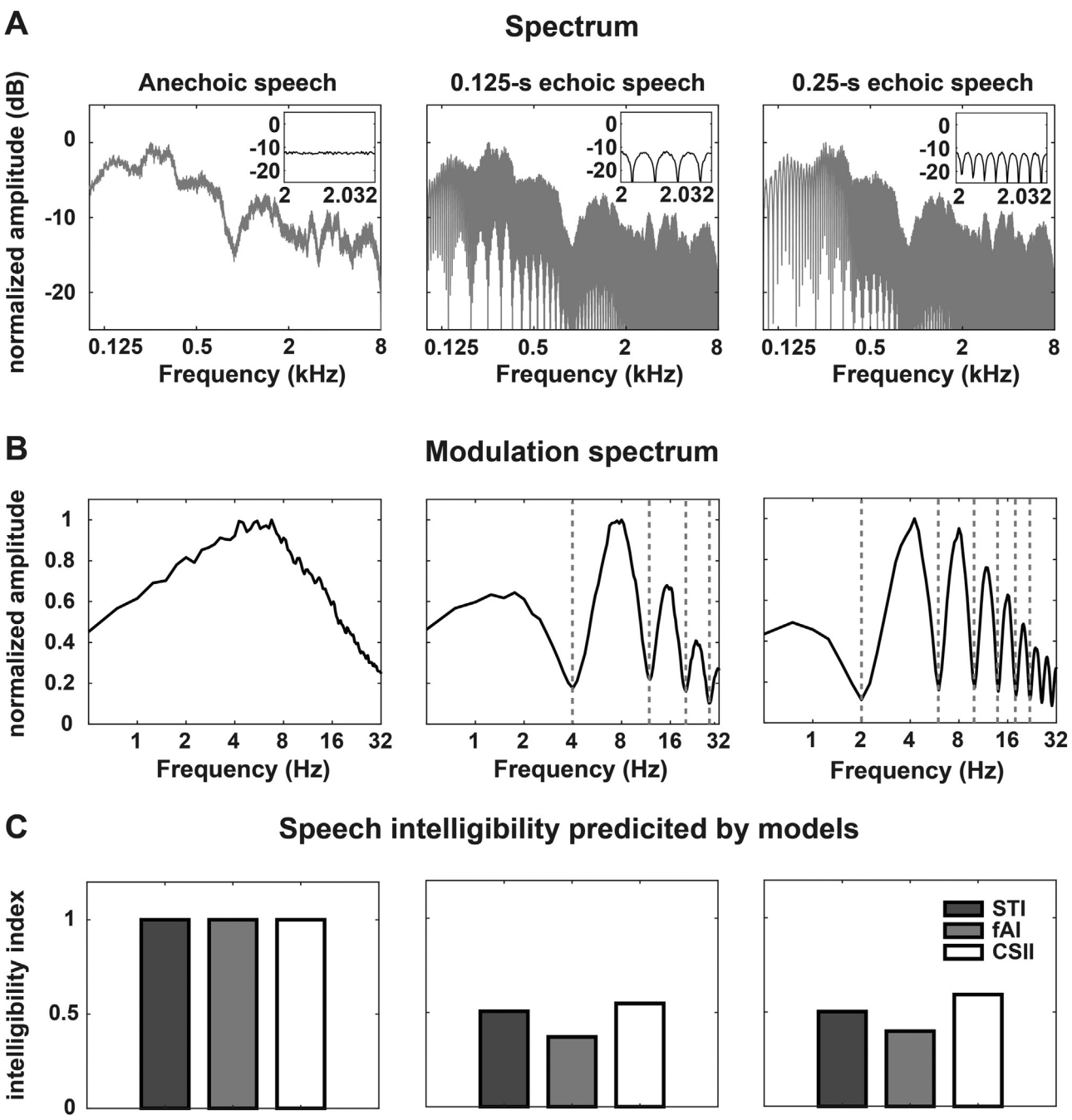


**Fig. 3.** (A), Stimulus spectrum. The inset shows a zoomed-in view of the spectrum between 2000 Hz and 2032 Hz. (B), Stimulus modulation spectrum. For speech with a 0.125-s echo, both the spectrum and the modulation spectrum are notched at $4 + 8n$ Hz, $n = 0, 1, 2....$ For speech with a 0.25-s echo, both the spectrum and the modulation spectrum are notched at $2 + 4n$ Hz, $n = 0, 1, 2....$ (C), Speech intelligibility predicted by three models, i.e., STI, fAI, and CSII. All models predict that the intelligibility of echoic speech was below 60%.

The experiment included 3 conditions, i.e., the anechoic speech condition, the 0.125-s echoic condition, and the 0.25-s echoic condition. The delay was set to 0.125 s or 0.25 s since such a delay could effectively attenuate the speech modulation rates that were critical for speech intelligibility, i.e., 4 and 2 Hz, as was shown in the following. Echoic speech was generated by adding a delayed version of the original speech, and the delay was 0.125 s or 0.25 s. The delayed version and the original version had equal amplitude (Fig. 2A). The stimulus spectrum was computed by applying the Fourier transform to each sentence after zero padding each sentence to the maximal sentence duration. The spectrum averaged over all 30 sentences were shown in Fig. 3A for each stimulus condition, which had a resolution of 0.31 Hz. The echo notched out the speech power at $1/(2T) + k/T$ Hz, $k = 0, 1, 2...$, but preserved the general shape of the stimulus. Each condition included 10 MHINT sentences, and was presented without any additional noise.

### 2.3. Modulation spectrum analyses

The modulation spectrum was the spectrum of the temporal envelope of sound and was computed following the procedure in Ding et al., 2017. In brief, a cochlear model was applied to extract the temporal envelope of speech in 128 narrow frequency bands (Yang et al., 1992), and each narrowband envelope was

transformed into the frequency domain using the Discrete Fourier Transformation (DFT). The modulation spectrum summed the DFT spectra in individual frequency bands, and its maximal amplitude was normalized to 1. The modulation spectrum for the stimulus in each condition was shown in Fig. 3B. For echoic speech, when the delay of the single echo was $T$ s, the modulation spectrum was notched at $1/(2T) + k/T$ Hz, $k = 0, 1, 2\ldots$

### 2.4. Procedures

Each stimulus condition, i.e., anechoic speech, 0.125-s echoic speech, or 0.25-s echoic speech, was presented in a block, but there was no break between blocks. The listeners were not notified about the change in condition during the experiment, but were instructed prior to the experiment that some sentences would be difficult to understand. The anechoic speech condition was always presented first, followed by the two echoic speech conditions. The order of 0.125-s and 0.25-s echoic conditions was counterbalanced across listeners for each listener group, i.e., young listeners or older listeners. The stimuli were presented via a 24-bit soundcard (Realtek$^{R}$ audio) through headphones (Sennheiser HD 280 Pro).

After hearing a sentence, the listeners were asked to verbally report what they heard. The experimenter scored how many syllables were correctly reported but did not give any feedback to the listeners. The speech recognition rate was calculated by the number of correct syllables divided by the number of total syllables (Wong et al., 2007). The speech recognition rate was calculated based on syllables instead of words, since words are not well-defined units in Chinese while each syllable corresponds to a grapheme and generally corresponds to a morpheme. Speech intelligibility was also evaluated using models, including the STI, as well as two popular measures, i.e., fractional Articulation Index (fAI) and Coherence Speech Intelligibility Index (CSII), which were not directly extracted from the modulation spectrum (Loizou, 2013). The STI, fAI, and CSII scores were evaluated based on the scripts provided by (Loizou, 2013).

Furthermore, the order of sentences in the experiment was counterbalanced across listeners through a Latin square design: The order of sentences was circularly shifted by one for each listener (Fig. 2B). Since 30 listeners were recruited for each listener group, for example, the first sentence presented to a listener included all the 30 possible sentences. The same applied to the $k^{\text{th}}$ sentence presented to a listener, where $k = 1, 2, \ldots, 30$. We employed the Latin square design since we would analyze how the speech recognition rate varied as a function of sentence exposure. Using this design, when the speech recognition rate was averaged over a listener group, the mean speech recognition rate for the sentence presented at each position was not influenced by the variance in the difficulty of individual sentences. Before the experiment, listeners were familiarized with 3 MHINT sentences (anechoic) that were not used in the formal experiment. During the familiarization session, the listeners were allowed to adjust the sound volume to their most comfortable level (Kong et al., 2004;Nabelek et al., 2006).

### 2.5. Data analysis

**Exponential curve fitting:** The change of speech recognition rate as a function of the number of sentences presented was fitted by an exponential curve. For the $i^{\text{th}}$ sentence, the speech recognition for the sentence $x_i$ was formulated as:

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

where, $A$, $B$, and $\tau$ were parameters to fit using the least squares methods. The parameters were fitted using iterative least squares estimation method, which was implemented by the *nlinfit* function in MATLAB R2020a.

**Partial correlation:** For three variables that denoted as $x$, $y$, and $z$, the partial correlation between $x$ and $y$ was calculated using the following equation.

$$corr_{partial}(x, y \mid z) = corr\left(x - \hat{x}_z, y - \hat{y}_z\right),$$

where $\hat{x}_z$ and $\hat{y}_z$ were the value of $x$ and $y$ predicted by $z$ using linear regression, respectively. In other words, we first regressed out the conditional variable $z$ from $x$ and $y$, and calculated the correlation between the residual of $x$ and the residual of $y$.

**Multivariate linear regression:** The speech recognition rate of a listener, $x$, was modeled based on the age of the listener, $y$, and the PTA, $z$.

$$x = \beta_0 + \beta_1 y + \beta_2 z + e,$$

where $\beta_k$ ($k = 0, 1, 2$) were the coefficients to fit and $e$ was the residual error. The coefficients were estimated using the least squares method. The model was evaluated by leave-one-out cross validation: Each time, a listener was selected as the test set and the remaining listeners were chosen as the training set.

### 2.6. Bootstrap

Bias-corrected and accelerated bootstrap (Efron and Tibshirani, 1994) was employed to compare the speech recognition rate between conditions. In the bootstrap procedure, data of all listeners were resampled with replacement 10,000 times. The test was two-sided: If the result in one condition was higher than the result in another condition $M$ out of the 10,000 replacements, the significance level was $2 \times \min(M + 1, 10{,}001\text{-}M) / 10{,}001$. When multiple comparisons were performed, the p-value was adjusted using the false discovery rate (FDR) correction (Benjamini and Hochberg, 1995).

## 3. Result

### 3.1. Mean speech recognition rate

The mean speech recognition rates in the three stimulus conditions, i.e., anechoic speech, 0.125-s, and 0.25-s echoic speeches, were shown in Fig. 4A. For young listeners, the speech recognition rate was near ceiling for all 3 conditions. Statistically, the speech recognition rate was significantly lower in the 0.25-s echoic condition that the anechoic condition ($p = 0.027$, paired two-sided bootstrap, FDR corrected), but the difference was subtle. There was no significant difference between the anechoic condition and 0.125-s echoic condition, or between 0.25-s and 0.125-s echoic conditions ($p = 0.059$ and 0.623, respectively, paired two-sided bootstrap, FDR corrected).

For older listeners, the speech recognition rate was close to 100% for the anechoic condition, but dropped to 74.8% and 70.9% for the 0.125-s, and 0.25-s echoic conditions, respectively. The speech recognition rate was significantly lower for the 0.125-s, and 0.25-s echoic conditions than the anechoic condition ($p = 3.0 \times 10^{-4}$ and $3.0 \times 10^{-4}$, respectively, paired two-sided bootstrap, FDR corrected), and the recognition rate was also lower for the 0.25-s echoic condition than the 0.125-s echoic condition ($p = 0.026$, paired two-sided bootstrap, FDR corrected).

### 3.2. Time course of the speech recognition rate

Previous studies have demonstrated that exposure to an adverse listening environment could improve the speech recognition rate in that listening environment (Peelle and Wingfield, 2005;Cooke et al., 2022). Therefore, in the following, we analyzed how the speech recognition rate varied when the listeners

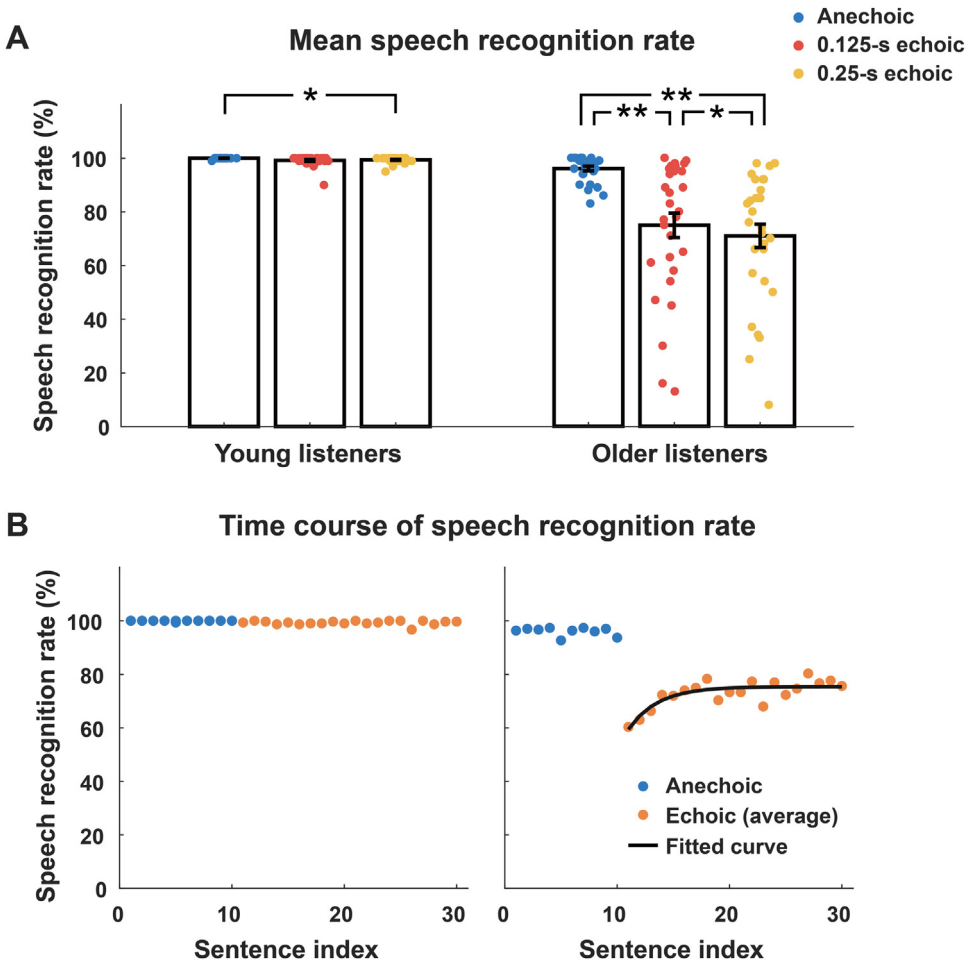


**Fig. 4.** (A), Speech recognition rate averaged within each listening condition. Each dot represents the result of a listener. Stars indicate significant differences between conditions (** $p < 0.001$, * $p < 0.05$). (B), Time course of speech recognition rate. For young listeners, the speech recognition rate is constantly high. For older listeners, the speech recognition rate drops when the stimulus switches from anechoic speech to echoic speech, i.e., at the 11th sentence, and recovers to some extent when the listener is exposed to more sentences. The recovery of speech recognition rate is fitted by an exponential function (black curve).

were exposed to more sentences. The sentence-level analysis was challenging since the sentences were not equally difficult to recognize. Nevertheless, the Latin square design allowed us to reliably analyze the time course of speech recognition rate when it was averaged over all listeners. In this analysis, the two echoic conditions were pooled since the Latin square design required all listeners to be averaged to cancel the differences between sentences and the two conditions did not differ much in their intelligibility.

Time course of the speech recognition rate was shown in Fig. 4B. The x-axis was the index of the sentence within an experiment, and the y-axis was the speech recognition rate of each sentence. The stimulus condition altered every ten sentences: The first 10 sentences were anechoic speech, and the rest 20 sentences were echoic speech, with the echo delay being changed at the 21st sentence. For young listeners, the speech recognition rate was not affected even when the stimulus just changed from anechoic speech to echoic speech, i.e., at the 11th sentence. Similarly, no change in the speech recognition rate was observed when the delay of the echo changed, i.e., at the 21st sentence.

For older listeners, however, the speech recognition rate dropped by more than 30% when the stimulus just changed from anechoic speech to echoic speech, i.e., at the 11th sentence. After the initial drop, the speech recognition rate gradually recovered, and the trend was well fitted by an exponential function. The fitted time constant of the exponential function was 2.56, which corresponded to 6.6 s given that the mean duration of a sentence was 2.58 s. The correlation between the fitted curve and the actual curve of the speech recognition rate was 0.835. Although the change from anechoic speech to echoic speech decreased the speech recognition rate for older listeners, no clear change in the speech recognition rate was observed when the echo delay changed, i.e., at the 21st sentence.

### 3.3. Effects of age and PTA on speech recognition

Next, we analyzed whether the speech recognition rate correlated with the age and hearing threshold of individual listeners. Since the 0.125-s and 0.25-s echoic conditions did not differ much in the speech recognition rate, two conditions were pooled in this analysis. We first separately considered the correlation between age and speech recognition rate. For young listeners, there was no significant correlation between age and speech recognition rate for neither anechoic and echoic speech, since the speech recognition rate was at ceiling. For older listeners, the recognition rate for anechoic speech was negatively correlated with both age ($R = -0.523$, $p = 0.003$) and hearing threshold ($R = -0.562$, $p = 0.001$) (Fig. 5A),and the recognition rate for echoic speech was also negatively correlated with both age ($R = -0.8$, $p = 1.1 \times 10^{-7}$) and hearing threshold ($R = -0.861$, $p = 1.1 \times 10^{-9}$) (Fig. 5B).

For older listeners, however, the hearing threshold was correlated with age, and a partial correlation analysis was applied to distinguish the influences of age and hearing threshold. For

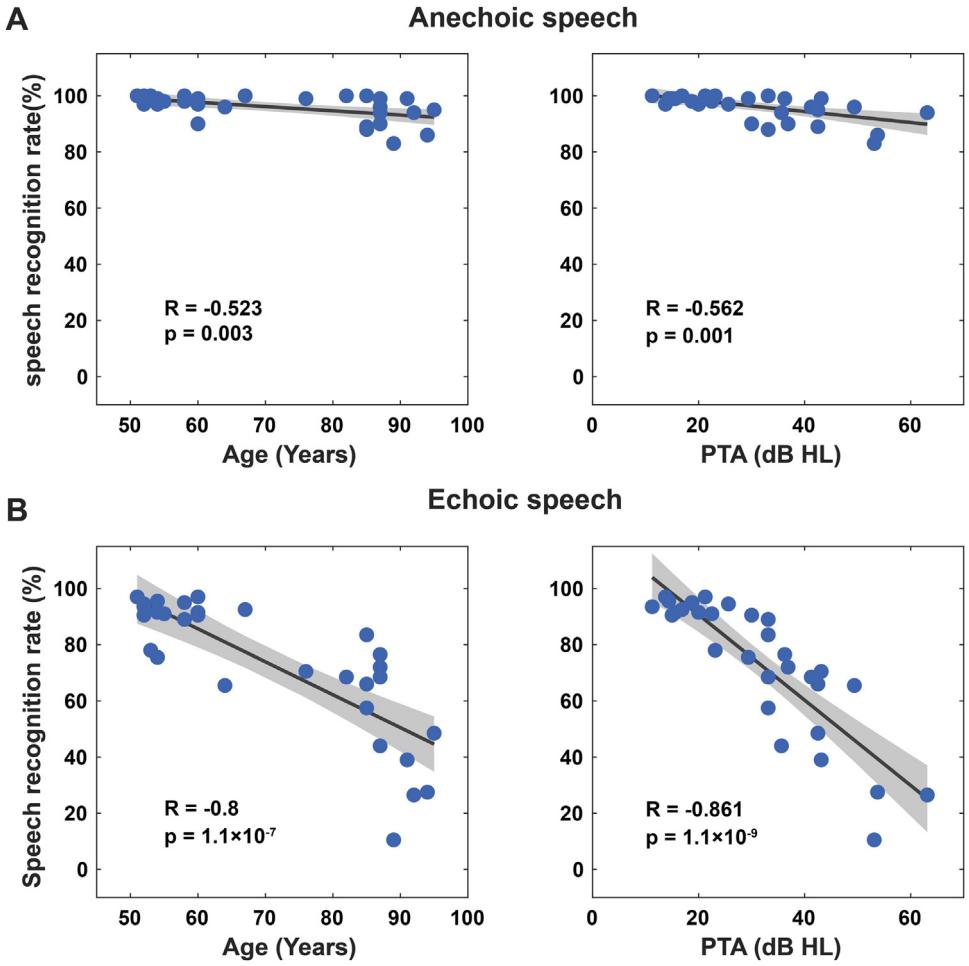


**Fig. 5.** Age and hearing threshold modulate individual speech recognition rate for older listeners. (A), Correlation between age and speech recognition rate, or between PTA and speech recognition rate for anechoic speech. (B), Correlation between age and speech recognition rate, or between PTA and speech recognition rate for echoic speech. Blue dots show results of individual listeners, and the black line shows the linear regression line, with the error bands showing 95% confidence interval across listeners.

echoic speech, after excluding the influence of hearing threshold, the speech recognition rate remained negatively correlated with age ($R = -0.398$, $p = 0.032$). Similarly, after excluding the influence of age, the speech recognition rate was remained negatively correlated with the hearing threshold ($R = -0.627$, $p = 3 \times 10^{-4}$). For anechoic speech, however, the partial correlation was not significant for neither age ($R = -0.162$, $p = 0.401$) nor hearing threshold ($R = -0.287$, $p = 0.132$). Finally, since age and hearing threshold both contributed to the speech recognition rate, we used both factors to predict the speech recognition rate of individual listeners using multivariate linear regression with leave-one-out cross-validation. The correlation between the predicted speech recognition rate and experimentally measured speech recognition rate reached 0.855.

## 4. Discussion

The study investigates how speech intelligibility is affected by a single echo. The single echo effectively modulates the shape of the modulation spectrum, and the study chooses two echo delays, i.e., 0.125 s and 0.25 s, that can notch out the modulation power at 2 and 4 Hz, respectively. Although previous studies have demonstrated that the 4-Hz temporal modulations are critical for speech intelligibility (Chi et al., 1999;Drullman et al., 1994;Elliott and Theunissen, 2009) and the 2-Hz temporal modulations are also relevant for speech perception (Füllgrabe et al., 2009), the current study shows that speech intelligibility is not influenced by the echo for young listeners. For older listeners, speech intelligibility is lowered by the echo and is more severely lowered for the 0.25-s echoic condition (corresponding to a 2-Hz notch) than the 0.125-s echoic condition (corresponding to a 4-Hz notch). Further analyses reveal that individual ability to recognize echoic speech is influenced by both the hearing threshold and age, for the older listener group. These results suggest that the auditory system can effectively cancel an echo, but such ability is affected by both hearing loss and age.

### 4.1. Influence of reverberation on speech intelligibility

Previous studies demonstrate that temporal modulations between 1 and 16 Hz can enable speech recognition even when only providing very coarse spectral information (Shannon et al., 1995). In a quiet anechoic environment, the modulation spectrum of speech peaks around 4 Hz (Greenberg et al., 2003;Ding et al., 2017;Varnet et al., 2017). In reverberant environments, however, the shape of the modulation spectrum is distorted and its peak generally shifts to lower frequencies (Duquesnoy and Plomp, 1980;Houtgast and Steeneken, 1985). The distortion of the modulation spectrum can predict speech intelligibility in reverberant environments, based on, e.g., the speech transmission index (STI) (Houtgast et al., 1980;Steeneken and Houtgast, 1980;Houtgast and Steeneken, 1985;Ellis and Zahorik, 2019). It has been shown that the STI can well predict speech intelligibility for normal hearing but not hearing-impaired listeners

(Payton et al., 1994). Here, however, it is demonstrated that, with a single echo, the STI cannot well predict speech intelligibility for normal-hearing young listeners either. When the modulation spectrum is manipulated by directly filtering the speech envelope, removing modulations between 3 and 8 Hz can also strongly degrade intelligibility (Drullman et al., 1994;Chi et al., 1999;Elliott and Theunissen, 2009).

In the current study, a single echo is introduced to effectively filter the speech envelope (Fig. 3). Adding an echo is equivalent to filtering speech, including the envelope, using a comb filter (Blauert, 1971;Oppenheim et al., 1997). Although the echo strongly distorts the shape of the modulation spectrum, it does not affect speech intelligibility for young listeners, even when the young listeners listen to the first echoic sentence. Therefore, the current result provides a challenge for theories and models that explain speech intelligibility using the speech modulation spectrum (Houtgast and Steeneken, 1985;Chi et al., 1999;Elhilali et al., 2003;Jørgensen and Dau, 2011). It is likely that the auditory system possesses mechanisms to effectively cancel echoes (Fuglsang et al., 2017;Mesgarani et al., 2014;Puvvada et al., 2017), and possible candidates include synaptic depression (David et al., 2009) and an increase in inhibitory receptive field (Ivanov et al., 2022) since these mechanisms can reduce the neural response to a repeated sound. The exact neural mechanisms to cancel echoes, however, need to be determined by future studies. Note that, in the current experiment, speech is presented in a quiet environment. When speech is mixed with stationary noise, however, previous studies have found that an echo with >100-ms delay significantly reduces speech intelligibility (Warzybok et al., 2013). Therefore, although a 125/250-ms echo per se does not influence speech intelligibility, it can make speech recognition more susceptible to noise. At the neurophysiological level, human MEG studies have also demonstrated that the neural encoding of speech envelope is more susceptible to noise for reverberant speech, compared with anechoic speech (Puvvada et al., 2017).

### 4.2. Influence of age and hearing threshold and speech intelligibility

Here, older listeners have trouble recognizing echoic speech and the effect is modulated by both the hearing threshold and age. Elevated hearing threshold generally lowers the speech recognition rate in challenging listening environments (Walden and Walden, 2004;Shub et al., 2020). In a study that tests a number different listening conditions, it is found that the correlation between hearing threshold and speech recognition rate varies between −0.77 and −0.94 (Humes and Roberts, 1990). In the current study, the correlation between hearing threshold and speech recognition rate is −0.86, within the range reported by Humes and Roberts. Some other studies, however, find much weaker influence of the hearing threshold on speech recognition in challenging listening environments. For example, using structural equation modeling, a study reveals very weak correlation ($R = 0.112$) between hearing threshold and speech recognition rate in older listeners (Anderson et al., 2013). Furthermore, mid-age listeners with normal hearing threshold show great variance in their speech recognition rates in reverberant environments (Ruggles et al., 2011). Several factors may lead to the variations in the correlation between threshold and speech recognition rate, including how speech is amplified and the kind of speech materials used for testing. For example, for listeners with sensorineural hearing loss, the hearing threshold can explain 49% and 71% of the variance in speech recognition threshold for QSIN sentences and HINT sentences, respectively (Grant and Walden, 2013). It has been suggested that studies using sentence materials tend to show weaker correlation between threshold and speech recognition than studies using digits or single words as the testing material (Akeroyd, 2008).

Aging is also an important factor influencing speech recognition in challenging listening environments (Marrone et al., 2008;Xia et al., 2018). Aging is often accompanied with the elevation of hearing threshold, but is also accompanied by more central factors, such as declines in temporal processing (Fitzgibbons and Gordon-Salant, 1996;Strouse et al., 1998) and cognitive functions (Salthouse, 2012). The deficits in temporal processing is partly demonstrated by degraded central neural encoding of speech temporal features, e.g., the frequency following response (Ruggles et al., 2011;Anderson et al., 2013;Presacco et al., 2016). Here, age influences the recognition of echoic speech for older listeners but not younger listeners. Nevertheless, it should be mentioned that the younger listeners are all between 20 and 30 years old (age range < 10 years), but the ages of older listeners span the range between 51 and 95. Future studies are needed to analyze whether the age-related decline in recognizing echoic speech emerge between, e.g., age 30 and 50.

## 5. Conclusion

In summary, the current results demonstrate that young listeners are resilient to the influence of single echoes, while aging and hearing impairment renders speech recognition more susceptible to echoes. The resilience to echoes cannot be explained by current modulation-spectrum-based models of speech intelligibility, and suggest that the auditory system can effectively cancel echoes. What are the neural mechanisms to cancel echoes and why aging and hearing impairment can degrade such mechanisms need to be investigated by future studies.

## Author statement

**Nai Ding:** Conceptualization, Methodology, Writing. **Jiaxin Gao:** Methodology, Investigation, Visualization, Writing. **Jing Wang:** Investigation, Writing. **Wenhui Sun:** Methodology, Writing. **Mingxuan Fang:** Methodology, Writing. **Xiaoling Liu:** Investigation, Writing. **Hua Zhao:** Conceptualization, Resources, Writing.

## Declaration of Competing Interest

No potential conflict of interest was reported by the authors

## Data availability

Data will be made available on request.

## Acknowledgments

We thank Dr. Jing Xia for insightful comments. This work was supported by the National Natural Science Foundation of China (32222035) and Key R & D Program of Zhejiang (2022C03011).

## References

Akeroyd, M.A., 2008. Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults. Int. J. Audiol. 47, S53–S71. doi:10.1080/14992020802301142.

Anderson, S., White-Schwoch, T., Parbery-Clark, A., Kraus, N., 2013. A dynamic auditory-cognitive system supports speech-in-noise perception in older adults. Hear. Res. 300, 18–32. doi:10.1016/j.heares.2013.03.006.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Methodol. 57, 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.

Blauert, J., 1971. Localization and the law of the first wavefront in the median plane. J. Acoust. Soc. Am. 50, 466–470. doi:10.1121/1.1912663.

Chi, T., Gao, Y., Guyton, M.C., Ru, P., Shamma, S., 1999. Spectro-temporal modulation transfer functions and speech intelligibility. J. Acoust. Soc. Am. 106, 2719–2732. doi:10.1121/1.428100.

Chi, T., Ru, P., Shamma, S.A., 2005. Multiresolution spectrotemporal analysis of complex sounds. J. Acoust. Soc. Am. 118, 887–906. doi:10.1121/1.1945807.

Cooke, M., Scharenborg, O., Meyer, B.T., 2022. The time course of adaptation to distorted speech. J. Acoust. Soc. Am. 151, 2636–2646. doi:10.1121/10.0010235.
Cutler, R., Saabas, A., Parnamaa, T., Purin, M., Gamper, H., Braun, S., Sorensen, K., Aichner, R., 2022. ICASSP 2022 acoustic echo cancellation challenge. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Singapore, Singapore, pp. 9107–9111. doi:10.1109/ICASSP43922.2022.9747215.
Dau, T., Kollmeier, B., Kohlrausch, A., 1997a. Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers. J. Acoust. Soc. Am. 102, 2892. doi:10.1121/1.420344.
Dau, T., Kollmeier, B., Kohlrausch, A., 1997b. Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration. J. Acoust. Soc. Am. 102, 2906–2919. doi:10.1121/1.420345.
David, S.V., Mesgarani, N., Fritz, J.B., Shamma, S.A., 2009. Rapid synaptic depression explains nonlinear modulation of spectro-temporal tuning in primary auditory cortex by natural stimuli. J. Neurosci. 29, 3374–3386. doi:10.1523/JNEUROSCI.5249-08.2009.
Ding, N., Patel, A.D., Chen, L., Butler, H., Luo, C., Poeppel, D., 2017. Temporal modulations in speech and music. Neurosci. Biobehav. Rev. 81, 181–187. doi:10.1016/j.neubiorev.2017.02.011, The Biology of Language.
Drullman, R., Festen, J.M., Plomp, R., 1994. Effect of reducing slow temporal modulations on speech reception. J. Acoust. Soc. Am. 85, 2670–2680.
Duquesnoy, A.J., Plomp, R., 1980. Effect of reverberation and noise on the intelligibility of sentences in cases of presbyacusis. J. Acoust. Soc. Am. 68, 537–544. doi:10.1121/1.384767.
Efron, B., Tibshirani, R.J., 1994. An Introduction to the Bootstrap. Chapman and Hall/CRC, New York doi:10.1201/9780429246593.
Elhilali, M., Chi, T., Shamma, S.A., 2003. A spectro-temporal modulation index (STMI) for assessment of speech intelligibility. Speech Commun. 41, 331–348. doi:10.1016/S0167-6393(02)00134-6.
Elliott, T.M., Theunissen, F.E., 2009. The modulation transfer function for speech intelligibility. PLOS Comput. Biol. 5, e1000302. doi:10.1371/journal.pcbi.1000302.
Ellis, G.M., Zahorik, P., 2019. A dissociation between speech understanding and perceived reverberation. Hear. Res. 379, 52–58. doi:10.1016/j.heares.2019.04.015.
Fitzgibbons, P., Gordon-Salant, S., 1996. Auditory temporal processing in elderly listeners. J. Am. Acad. Audiol. 7, 183–189.
Fuglsang, S.A., Dau, T., Hjortkjær, J., 2017. Noise-robust cortical tracking of attended speech in real-world acoustic scenes. Neuroimage 156, 435–444. doi:10.1016/j.neuroimage.2017.04.026.
Füllgrabe, C., Stone, M.A., Moore, B.C.J., 2009. Contribution of very low amplitude-modulation rates to intelligibility in a competing-speech task. J. Acoust. Soc. Am. 125, 1277–1280. doi:10.1121/1.3075591.
Ghitza, O., 2001. On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perceptiona). J. Acoust. Soc. Am. 110, 13.
Grant, K.W., Walden, T.C., 2013. Understanding excessive SNR loss in hearing-impaired listeners. J. Am. Acad. Audiol. 24, 258–273. doi:10.3766/jaaa.24.4.3.
Greenberg, S., 1999. Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation. Speech Commun. 29, 159–176. doi:10.1016/S0167-6393(99)00050-3.
Greenberg, S., Carvey, H., Hitchcock, L., Chang, S., 2003. Temporal properties of spontaneous speech—a syllable-centric perspective. J. Phon. 31, 465–485. doi:10.1016/j.wocn.2003.09.005, Temporal Integration in the Perception of Speech.
Harris, R.W., Reitz, M.L., 1985. Effects of room reverberation and noise on speech discrimination by the elderly. Int. J. Audiol. 24, 319–324. doi:10.3109/00206098509078350.
Helfer, K.S., Wilber, L.A., 1990. Hearing loss, aging, and speech perception in reverberation and noise. J. Speech Lang. Hear. Res. 33, 149–155. doi:10.1044/jshr.3301.149.
Houtgast, T., Steeneken, H.J.M., 1985. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. J. Acoust. Soc. Am. 77, 1069–1077. doi:10.1121/1.392224.
Houtgast, T., Steeneken, H.J.M., 1973. The modulation transfer function in room acoustics as a predictor of speech intelligibility. Acta Acust. United Acust. 28, 66–73.
Houtgast, T., Steeneken, H.J.M., Plomp, R., 1980. Predicting speech intelligibility in rooms from the modulation transfer function. I. general room acoustics. Acta Acust. United Acust. 46, 60–72.
Humes, L.E., Roberts, L., 1990. Speech-recognition difficulties of the hearing-impaired elderly: the contributions of audibility. J. Speech Hear. Res. 33, 726–735. doi:10.1044/jshr.3304.726.
Ivanov, A.Z., King, A.J., Willmore, B.D., Walker, K.M., Harper, N.S., 2022. Cortical adaptation to sound reverberation. Elife 11, e75090. doi:10.7554/eLife.75090.
Jørgensen, S., Dau, T., 2011. Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. J. Acoust. Soc. Am. 130, 1475–1487. doi:10.1121/1.3621502.
Kong, Y.-Y., Cruz, R., Jones, J.A., Zeng, F.-G., 2004. Music perception with temporal cues in acoustic and electric hearing. Ear. Hear. 25, 173–185. doi:10.1097/01.aud.0000120365.97792.2f.
Loizou, P.C., 2013. Speech Enhancement: Theory and Practice, Second Edition. CRC Press.
Marrone, N., Mason, C.R., Kidd, G., 2008. The effects of hearing loss and age on the benefit of spatial separation between multiple talkers in reverberant rooms. J. Acoust. Soc. Am. 124, 3064–3075. doi:10.1121/1.2980441.
Mesgarani, N., David, S.V., Fritz, J.B., Shamma, S.A., 2014. Mechanisms of noise robust representation of speech in primary auditory cortex. Proc. Natl. Acad. Sci 111, 6792–6797. doi:10.1073/pnas.1318017111.
Nabelek, A.K., Freyaldenhoven, M.C., Tampas, J.W., Burchfiel, S.B., Muenchen, R.A., 2006. Acceptable noise level as a predictor of hearing aid use. J. Am. Acad. Audiol. 17, 626–639. doi:10.3766/jaaa.17.9.2.
Náběl ek, A.K., Robinson, P.K., 1982. Monaural and binaural speech perception in reverberation for listeners of various ages. J. Acoust. Soc. Am. 71, 1242–1248. doi:10.1121/1.387773.
Oppenheim, A.V., Willsky, A.S., Nawab, S.H., Hamid, with, Hernández, G.M., 1997. Signals & Systems. Pearson Educación.
Payton, K.L., Uchanski, R.M., Braida, L.D., 1994. Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. J. Acoust. Soc. Am. 95, 1581–1592. doi:10.1121/1.408545.
Peelle, J.E., Wingfield, A., 2005. Dissociations in Perceptual Learning Revealed by Adult Age Differences in Adaptation to Time-Compressed Speech. J. Exp. Psychol. Hum. Percept. Perform. 31, 1315–1330. doi:10.1037/0096-1523.31.6.1315.
Presacco, A., Simon, J.Z., Anderson, S., 2016. Evidence of degraded representation of speech in noise, in the aging midbrain and cortex. J. Neurophysiol. 116, 2346–2355. doi:10.1152/jn.00372.2016.
Puvvada, K.C., Villafañe-Delgado, M., Brodbeck, C., Simon, J.Z., 2017. Neural Coding of Noisy and Reverberant Speech in Human Auditory Cortex. doi:10.1101/229153.
Rosen, S., 1992. Temporal information in speech: acoustic, auditory and linguistic aspects. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 336, 367–373. doi:10.1098/rstb.1992.0070.
Ruggles, D., Bharadwaj, H., Shinn-Cunningham, B.G., 2011. Normal hearing is not enough to guarantee robust encoding of suprathreshold features important in everyday communication. Proc. Natl. Acad. Sci. 108, 15516–15521. doi:10.1073/pnas.1108912108.
Salthouse, T., 2012. Consequences of age-related cognitive declines. Annu. Rev. Psychol. 63, 201–226. doi:10.1146/annurev-psych-120710-100328.
Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J., Ekelid, M., 1995. Speech recognition with primarily temporal cues. Science 270, 303–304. doi:10.1126/science.270.5234.303.
Shub, D.E., Makashay, M.J., Brungart, D.S., 2020. Predicting speech-in-noise deficits from the audiogram. Ear. Hear. 41, 39–54. doi:10.1097/AUD.0000000000000745.
Sridhar, K., Cutler, R., Saabas, A., Parnamaa, T., Loide, M., Gamper, H., Braun, S., Aichner, R., Srinivasan, S., 2020. ICASSP 2021 Acoustic Echo Cancellation Challenge: datasets, Testing Framework, and Results.
Steeneken, H.J.M., Houtgast, T., 1980. A physical method for measuring speech-transmission quality. J. Acoust. Soc. Am. 67, 318–326. doi:10.1121/1.384464.
Strouse, A., Ashmead, D.H., Ohde, R.N., Grantham, D.W., 1998. Temporal processing in the aging auditory system. J. Acoust. Soc. Am. 104, 2385–2399. doi:10.1121/1.423748.
Varnet, L., Ortiz-Barajas, M.C., Erra, R.G., Gervain, J., Lorenzi, C., 2017. A cross-linguistic study of speech modulation spectra. J. Acoust. Soc. Am. 142, 1976–1989. doi:10.1121/1.5006179.
Walden, T.C., Walden, B.E., 2004. Predicting success with hearing aids in everyday living. J. Am. Acad. Audiol. 15, 342–352. doi:10.3766/jaaa.15.5.2.
Warzybok, A., Rennies, J., Brand, T., Doclo, S., Kollmeier, B., 2013. Effects of spatial and temporal integration of a single early reflection on speech intelligibility. J. Acoust. Soc. Am. 133, 269–282. doi:10.1121/1.4768880.
Wong, L.L.N., Soli, S.D., Liu, S., Han, N., Huang, M.-.W., 2007. Development of the Mandarin hearing in noise test (MHINT). Ear Hear 28, 70S–74S. doi:10.1097/AUD.0b013e31803154d0.
Xia, J., Xu, B., Pentony, S., Xu, J., Swaminathan, J., 2018. Effects of reverberation and noise on speech intelligibility in normal-hearing and aided hearing-impaired listeners. J. Acoust. Soc. Am. 143, 1523–1533. doi:10.1121/1.5026788.
Yang, X., Wang, K., Shamma, S.A., 1992. Auditory representations of acoustic signals. IEEE Trans. Inf. Theory 38, 824–839. doi:10.1109/18.119739.
Zeng, F.-G., Nie, K., Liu, S., Stickney, G., Del Rio, E., Kong, Y.-Y., Chen, H., 2004. On the dichotomy in auditory perception between temporal envelope and fine structure cues. J. Acoust. Soc. Am. 116, 1351–1354. doi:10.1121/1.1777938.
Zhang, H., Wang, D., 2022. Neural cascade architecture for multi-channel acoustic echo suppression. IEEEACM Trans. Audio Speech Lang. Process. 30, 2326–2336. doi:10.1109/TASLP.2022.3192104.