The inner ear's active process contributes to selective attention to speech in noise

Janna Steinebach [1] and Tobias Reichenbach [1,*]
[1]Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Nürnberger Strasse 74, 91052 Erlangen, Germany
[*]Correspondence: tobias.j.reichenbachfau.de

Keywords: speech processing | otoacoustic emissions | auditory attention | inner ear biology | efferent feedback

ABSTRACT

Humans are remarkably skilled at understanding speech in noisy environments. While segregation of different audio streams is mostly accomplished in the auditory cortex, it remains unclear whether the inner ear, where the sound detection occurs, already contributes to selective attention. In particular, the cochlea possesses an active process enabling it to amplify sound in a frequency-dependent manner. A physiological correlate of the active process are distortion-product otoacoustic emissions (DPOAEs) that can be measured non-invasively from the ear canal. Here we employed speech-DPOAEs that are connected to the spectral structure of voiced speech to show that these emissions are modulated by selective attention to one of two competing voices. We found that speech-DPOAEs evoked by the resolved harmonics of a voice were significantly reduced when that voice was attended as compared to when it was ignored. No such effect was observed for the unresolved harmonics of the target voice when the competing voice's harmonics in that range were unresolved as well, indicating that attentional modulation is specific to those components of voiced speech that are spectrally resolved. Our findings demonstrate for the first time that selective attention to speech in noise is already shaped by the inner ear's active process.

INTRODUCTION

Understanding speech in noisy environments is an important yet highly complex human ability. In crowded settings such as restaurants or family gatherings, selective auditory attention enables listeners to focus on a single speaker while filtering out competing voices also known as the cocktail party effect [1, 2]. This ability is essential for many aspects of social participation, but is vulnerable to hearing damage: many people with hearing impairment complain of difficulty understanding speech in background noise, even when using hearing instruments [3].

The neural machinery behind selective auditory attention, including attention to speech, has been extensively studied at the level of the cerebral cortex [4, 5, 6, 7, 8]. However, anatomical and physiological evidence points to substantial descending feedback from the auditory cortex to the auditory brainstem and to the cochlea [9, 10, 11]. Through these neural feedback loops, subcortical processing centers may contribute to accomplishing the cocktail party effect. Several studies, including from our group, have indeed found that neural responses from the brainstem, in particular frequency-following responses, are modulated by selective auditory attention [12, 13, 14, 15].

The inner ear – the sensory organ of hearing in which sound vibrations are converted into electrical signals – may already contribute to selective auditory attention as well. Also referred to as the cochlea, this fascinating organ spatially decomposes a complex sound such as speech into its individual frequency components, following a tonotopic map in which high frequencies are detected near the organ's base, and lower frequencies progressively further towards the apex [16, 17].

In addition, the inner ear possesses an active process through which it amplifies weak sounds [18, 19]. This mechanical amplification is provided by outer hair cells and can be reduced through activation of the medial olivocochlear (MOC) fibers – efferent connections that innervate the outer hair cells [20, 21]. Because each MOC fiber is tuned to a narrow frequency band, and because the innervation of the inner ear by these fibers displays a tonotopic arrangement, the gain of the active process can potentially be regulated by the brain in a frequency-dependent manner.

The active process is accompanied by a nonlinear response that gives rise to otoacoustic emissions (OAEs). These can be recorded from the ear canal and serve as a non-invasive measure of the amplification gain. OAEs have indeed been used to assess the contri-

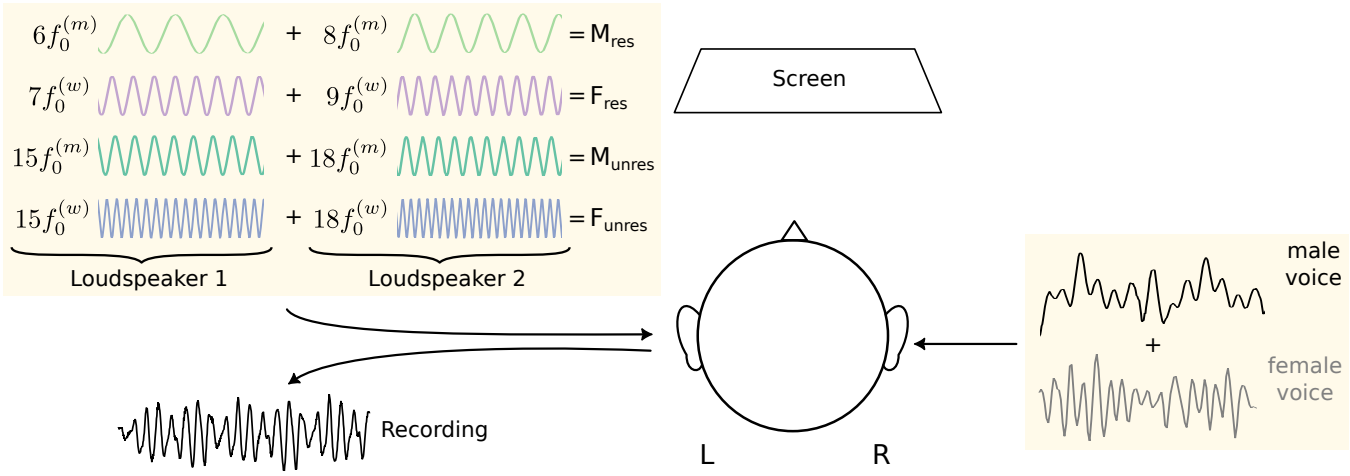

Figure 1: Experimental setup. Two audiobooks, one spoken by a women with fundamental frequency $f_0^{(w)}$ and one narrated by a man with fundamental frequency $f_0^{(m)}$, were simultaneously presented to the right ear. Speech-DPOAEs were recorded from the left ear. They were elicited by four pairs of waveforms. The first stimulus pair, M_{res}, contained the resolved harmonics of the male voice (frequencies $6f_0^{(m)}$ and $8f_0^{(m)}$). The second stimulus pair, F_{res}, consisted of resolved harmonics of the female voice (frequencies $7f_0^{(w)}$ and $9f_0^{(w)}$). The third and fourth stimulus pair, M_{unres} and F_{unres}, contained unresolved harmonics of the male resp. female voice (frequencies $15f_0^{(m)}$ and $18f_0^{(m)}$ resp. $15f_0^{(w)}$ and $18f_0^{(w)}$). Participants directed their attention either to one of the two talkers or visually to a continuously presented story on a screen.

bution of the inner ear to selective attention, although with inconclusive results [22, 23, 24]. A limitation of these studies was that they either did not involve naturalistic sounds such as speech that facilitate attention, or that they elicited OAEs in a manner that was not directly related to the auditory signal that the participants were asked to attend.

Computational models have shown that selective attention to a speech signal in noise may be supported by frequency-specific modulation of the cochlear active process [25, 26]. Most parts of speech are voiced, with the energy carried by the fundamental frequency and its many higher harmonics. The lower harmonics, up to the 10th, can be spatially resolved in the cochlea, that is, they cause peaks at significantly distinct locations [27, 28, 29, 30].

A compelling hypothesis is that the cochlear amplifier selectively enhances the resolved harmonics of a target speech and suppresses the spectral bands that lie inbetween. This mechanisms would thus reduce background noise already at the level of cochlear activity. Because unresolved harmonics cannot be spatially differentiated in the cochlea, this mechanism should not be able to operate for these.

Here, we set out to test this hypothesis. We employed distortion product otoacoustic emissions that were evoked by certain higher harmonics in voiced speech (speech-DPOAEs). Because the fundamental frequency of natural speech varies over time, the stimuli that we used to generate speech-DPOAEs, as well as the speech-DPOAEs themselves, were not pure tones but had instantaneous frequencies that changed over time. We recently developed and corroborated this approach [31].

We elicited and recorded speech-DPOAEs from one ear while two competing talkers were presented to the contralateral ear. Subjects were instructed to attend either one of the two talkers or to read a text in front of them (visual attention). We then evaluated how the speech-DPOAEs, in particular those related to resolved and unresolved harmonics, were affected by the attentional focus.

RESULTS

We utilized a competing-speaker paradigm in which two audiobooks, one spoken by a women and the other by a man, were mixed together and presented to each participant's right ear (Fig. 1). Subjects were instructed to attend either the female or the male voice. To test effects of intermodal attention, a visual distractor was introduced and attended upon prompt, participants then read a story displayed in segments on a screen while ignoring both audio streams.

To enable clean measurement of speech-DPOAEs, these were evoked and measured from the contralateral, i.e. the left, ear. Speech-DPOAEs were elicited by pairs of stimuli that corresponded to certain harmonics of the fundamental frequency of the male and the female voice.

To assess attentional modulation of cochlear activity at both resolved and unresolved harmonics, we utilized four pairs of harmonics for stimulation: (1) stimulus F_{res}: two resolved harmonics of the female voice, (2) stimulus F_{unres}: two unresolved harmonics of the female voice, (3) stimulus M_{res}: two resolved harmonics of the male voice, and (4) stimulus M_{unres}: two unresolved harmonics of the male voice.

The recordings of the speech-DPOAEs were segmented into two-minute trials, each followed by three comprehension questions and a rating of perceived mental effort to ensure task engagement. For the auditory conditions (attending the female resp. male voice), mental effort was equivalent to the listening effort. In order to include assessment of the effort for the third, i.e. the visual, condition, the term mental effort was chosen. In the initial single-speaker scenario, subjects listened to a single voice in isolation, once to the female and once to the male one, and focused their attention on this.

In the competing speaker scenario, subjects were instructed to attend either the female voice, the male voice, or the visual distractor in each individual trial. In the following, we refer to these three attentional conditions as attended female voice (Att. F), attended male voice (Att. M), and attended visual distractor (Att. V). Each attentional condition was presented six times in randomized order, resulting in 18 competing-speaker trials per participant.

Comprehension scores and mental effort

Speech comprehension was quantified as the percentage of correct answers, and mental effort was rated using mean values on a Likert scale from 1 to 13 (low to high). Values are given as mean ± SD. Statistical significance was assessed using Wilcoxon signed-rank tests.

In the single-speaker measurements, comprehension scores were high: 96.5 ± 0.1% for the female voice and 95.6 ± 0.1% for the male voice, with no significant difference between them.

The comprehension scores were slightly lower in the competing-speaker condition: 94.5 ± 0.1% when attending the female voice, 91.1 ± 0.1% for the male voice, and 93.9 ± 0.1% when reading the text. Again, no significant differences were found, suggesting that all conditions were similarly comprehensible.

Regarding mental effort, the female and male voices were perceived as similarly demanding in the single-speaker condition, with

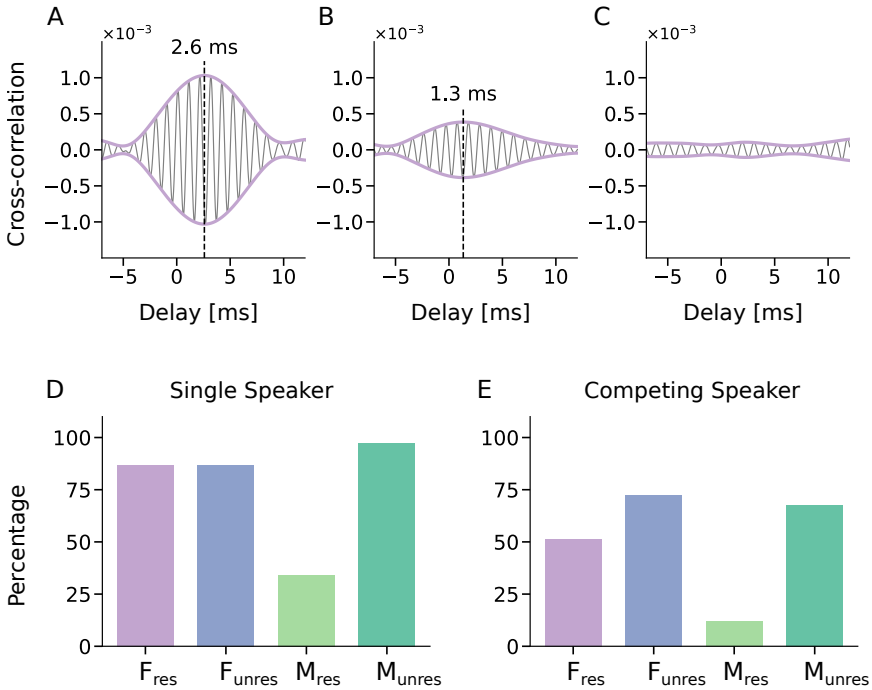

Figure 2: Recording of speech-DPOAES. Speech-DPOAEs were measured through cross-correlating the waveform that corresponded to the distortion product with the microphone signal. (A,B) The envelope (purple) of the cross-correlation (black) exhibited a peak at a short delay (dashed line) that showed the presence of the speech-DPOAE at this delay. (C) In contrast, if no peak emerged, the speech-DPOAE could not be measured. The exemplary recordings show results from a single two-minute trial in the single-speaker scenario, for the F_{res} stimulus and for different subjects. (D,E), Speech-DPOAEs could be detected in most trials both in the single-speaker and in the competing-speaker scenario, except for the stimulus M_{res}.

ratings of 4.0 $\pm$ 2.3 versus 4.4 $\pm$ 2.4 and no significant difference.

In the competing-speaker condition, perceived effort varied significantly depending on the attentional focus. Attending the visual distractor was rated easiest at 5.7 $\pm$ 2.2; lower than when attending the male voice ($p < 0.001$) and lower than when attending the female voice ($p < 0.05$). In contrast, attending the male voice was rated hardest at 7.3 $\pm$ 1.7; with $p < 0.001$ when compared to other conditions. Attending the female voice fell inbetween at a value of 6.5 $\pm$ 1.8.

Measurement of speech-DPOAEs

To measure a particular speech-DPOAE, we computed a waveform that corresponded to the lower-sideband cubic distortion product of the pair of harmonics that was used for stimulation. As an example, for the stimulus F_{res} we utilized the 7th and the 9th harmonic, yielding the 5th harmonics as the lower sideband cubic distortion product.

The speech-DPOAE waveform was then cross-correlated with the microphone recording (Fig. 2A - C). We computed the envelope of the cross-correlation to obtain, in most subjects, a single peak at a delay between $0{-}3$ ms. Each peak meant that a speech-DPOAE could be measured, with a delay that corresponded to that of the peak. Examples of trials with large and moderate peak amplitudes, as well as a trial without a significant peak, are shown in Fig. 2A-C, illustrating the variability in speech-DPOAE morphology.

To assess how well the speech-DPOAEs for the different stimuli could be measured, we computed the percentage of trials per stimulus type that yielded significant peaks in the cross-correlation. In the single-speaker scenario, the three stimuli F_{res}, F_{unres}, and M_{unres} all yielded significant speech-DPOAEs in over 85% of single-speaker trials (Fig. 2D). However, the stimulus M_{res} performed notably worse, with only 35% of trials showing a significant speech-DPOAE.

A similar pattern emerged in the competing-speaker scenario (Fig. 2E): the stimuli F_{unres} and M_{unres} performed well with speech-DPOAEs detectable in about 70% of the trials, the stimulus F_{res} slightly lower, with about 50% of trials yielding significant speech-DPOAEs, and M_{res} remained poor with only about 12% of trials producing a significant measurement.

To further compare the speech-DPOAEs evoked by the different stimuli, we averaged the envelopes of the cross-correlations across trials and subjects, yielding grand averages.

In the single-speaker scenario, all stimulus types yielded significant peaks, but with considerable variation in the amplitudes, that is, the values of the envelopes of the cross-correlations at the peak (Fig. 3E). The speech-DPOAE for the stimulus M_{res}, with an average frequency of 350 Hz, had the smallest amplitude. The highest amplitudes emerged for the stimuli F_{res} and M_{unres}, both of which produced speech-DPOAEs around 1 kHz.

The delays of the speech-DPOAEs also varied across the four stimulus types (Fig. 3F). The shortest delay was observed for the F_{unres} stimulus at 0 ms, while the longest delay occurred for the M_{res} stimulus at 3.1 ms. For the M_{res} stimulus, peak delays showed substantial variability across participants. This variability was likely due to the low number of significant peaks for this stimulus type (Fig. 2DE), further underscoring the limited interpretability of the corresponding results.

Due to the small amplitude and the low rate of significant peaks observed, the stimulus M_{res} was excluded from further analysis.

For competing-speaker trials, the remaining stimuli (F_{res}, F_{unres}, M_{unres}) produced significant peaks in all three attentional conditions (Fig. 4A-I). While the delays of the peaks remained stable across attentional conditions, they continued to vary between stimulus types.

Out-of-ear control measurements did not yield significant peaks in the grand average for any stimulus type in single- or competing-speaker trials, confirming the absence of measurable distortion products when the probe was placed outside the ear canal.

Attentional modulation of speech-DPOAEs

To assess effects of attentional focus on the speech-DPOAEs in the competing-speaker scenario, we characterized the latter through both the delay of the peak in the cross-correlation and the amplitude, that is, the value of the cross-correlation at the grand average peak delay.

We first quantified the influence of attention on the amplitude of the emissions. We started with the stimulus F_{res}, that is, by assessing the speech-DPOAEs evoked by resolved harmonics of the female voice (Fig. 4J). We found a significantly lower amplitude when

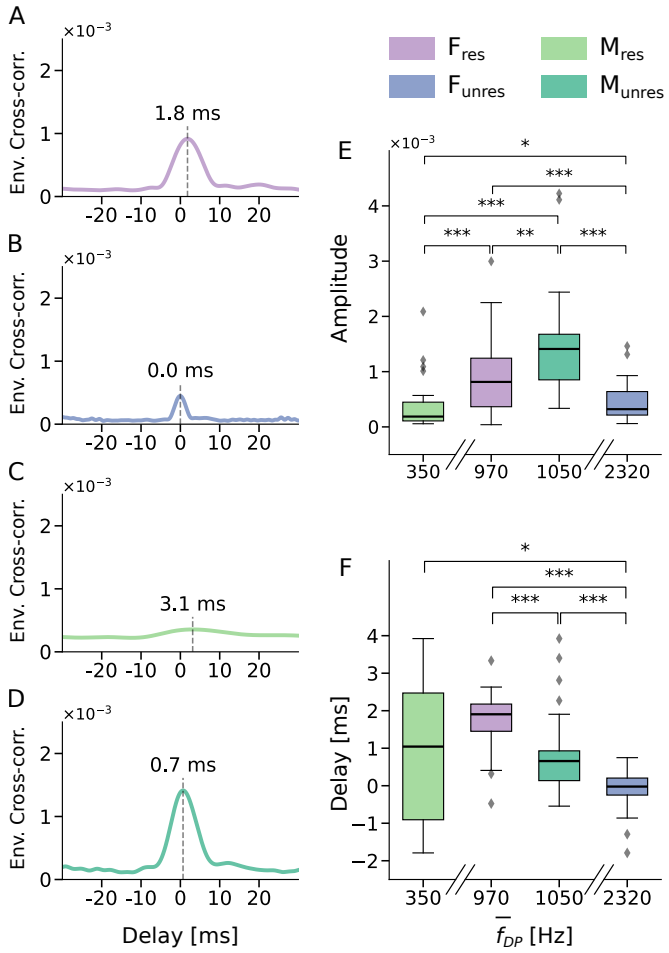

Figure 3: Speech-DPOAEs in the single-speaker scenario. (A) (D) Grand averages of the envelopes of the cross-correlations for the different stimuli. All peaks except the one for the M_{res} stimulus were clearly visible. The delays of the peaks are marked by dashed lines. E) The correlation coefficients at the peak latency varied significantly across stimulus types. They displayed an inverted U when plotted against the frequency of the speech-DPOAE, with the largest amplitudes emerging for speech-DPOAEs around 1 kHz. F) Peak delays of individually identified peaks differed significantly across stimulus types. The variation in delays for the M_{res} stimulus was high due to the low number of significant peaks. Statistical significance is indicated as $*$ ($p < 0.05$), $**$ ($p < 0.01$), $***$ ($p < 0.001$).

the female speaker was attended (Att. F) than when it was ignored, i.e. while the male speaker was attended (Att. M). The difference was highly statistically significant, with a p-value below 0.001. The ratio of the amplitudes in the two conditions, Att. F. versus Att. M., was 0.8, or -2.2 dB.

The amplitude when attending the female speaker was also significantly lower than when reading the text, that is, when attention was focused on the visual modality, with a p-value below 0.001 (Fig. 4J). In this case, the amplitude ratio was 0.7, or -2.6 dB. In contrast, no significant difference emerged when comparing the attended male voice (Att. M) condition to the attend visual condition (Att. V).

For the speech-DPOAEs elicited by unresolved harmonics of the female voice, the stimulus F_{unres}, we did not observe any difference in amplitudes across the three attentional conditions (Fig. 4K).

For the male speaker, because the stimulus M_{res} did not give reliable speech-DPOAEs, we could only assess the stimulus M_{unres} which utilized unresolved harmonics of the male voice. Significant amplitude differences emerged between all three conditions (Fig. 4L). The amplitude when attending the male voice was significantly higher than in the other two conditions. The ratio between the amplitudes when attending the male voice and when attending the female one was 1.5, or 3.8 dB. In addition, the amplitude when ignoring the male voice, i.e. attending the female voice, was smaller than when reading the text (ratio of 0.8, or -2 dB). The amplitude difference between attending the male voice and reading the text was slightly less pronounced ($p = 0.03$) with a ratio of 1.2, or 1.7 dB.

Regarding the delays of the speech-DPOAEs, we did not find any significant differences between the three attentional conditions,

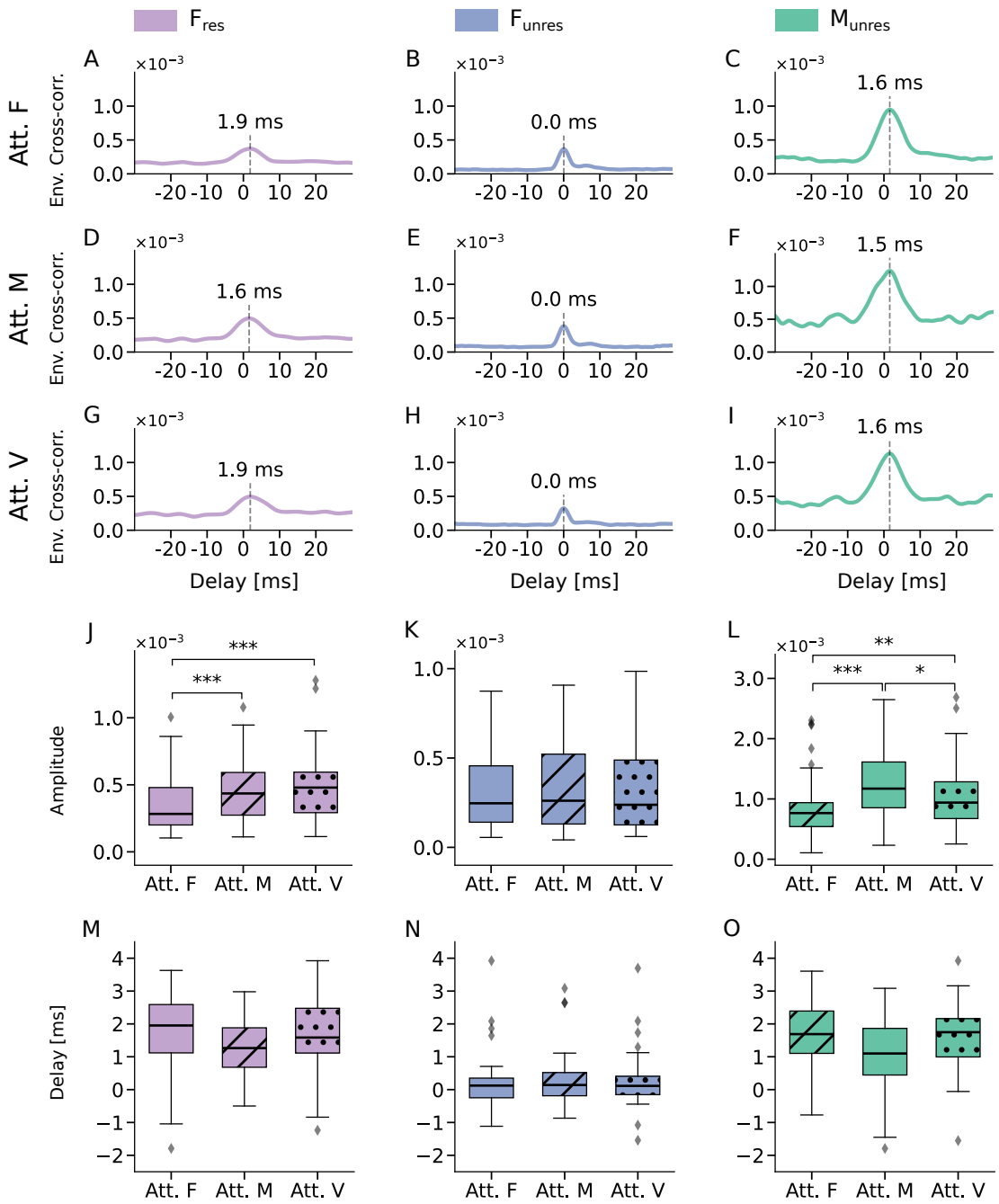

Figure 4: Attentional modulation of speech-DPOAEs. (A) - (I) Envelopes of the cross-correlations of speech-DPOAEs evoked by the three different stimuli F_{res}, F_{unres}, and M_{unres}, in the three different attentional conditions attended female voice (Att. F), attended male voice (Att. M), and attended visual (Att. V.). All showed clear peaks at short delays (dashed lines). (J) - (L) Comparison of the peak amplitudes yielded attentional effects for the F_{res} stimulus and the M_{unres} stimulus, but not for the stimulus F_{unres}. (M) - (O) The peak delays were not affected by the attentional focus, for neither of the three stimulus types. Statistical significance is indicated as $*$ ($p < 0.05$), $**$ ($p < 0.01$), $***$ ($p < 0.001$).

in neither of the three stimuli (Fig. 4M-O).

DISCUSSION

This study examined the contributions of the inner ear to selective attention to speech in noise. We employed a paradigm with two competing speakers as well as a visual distractor. Inner-ear activity was assessed through speech-DPOAEs elicited by resolved as well as unresolved harmonics in the voiced parts of the speech signals.

Regarding the behavioral results, we found that the comprehension scores did not differ significantly across attentional conditions, confirming that task difficulty was well-matched across conditions. Small differences in self-reported mental effort were likely attributable to speaker-specific factors such as intonation.

Our evaluation of the amplitude of the speech-DPOAEs in the single-speaker scenario revealed a pronounced dependency on the emission frequency (Fig. 3E). Consistent with previous findings on pure-tone DPOAEs, amplitudes were strongest for stimulus frequencies around 1 kHz [32]. Both lower and higher emission frequencies resulted in lower amplitudes. Because noise in electronics increases at lower frequencies, the signal at the lowest emission, around 350 Hz for the stimulus M_{res}, could not be detected in most trials. This stimulus was therefore excluded from further analysis.

Regarding attentional effects, we did observe attentional modulation of the amplitudes of the speech-DPOAEs for the stimuli F_{res} and M_{unres}, but not for F_{unres}.

Resolved harmonics, such as those employed for the stimulus F_{res}, produce spatially distinct excitation peaks along the basilar membrane [28, 27]. Vibrations at these locations can be selectively enhanced through higher gain of the active process, which can help to enhance the neural representation of a target speech. In contrast, unresolved harmonics yield overlapping peaks along the basilar membrane, making such a spatial filter unfeasible. These considerations likely explain the presence of attentional modulation for the stimulus F_{res} together with its absence for the stimulus F_{unres}.

However, for speech-DPOAEs elicited by the stimulus F_{res}, the resolved harmonics of the female voice, we counterintuitively observed *lower* amplitudes when attention was directed at the female voice versus when the male voice or the text was attended. Taken at face value, this result suggests that the inner ear amplifies the harmonic structure of the target voice *less* than the background noise. Although unexpected, such a reduced response for a target signal has been found in other parts of the nervous sys-

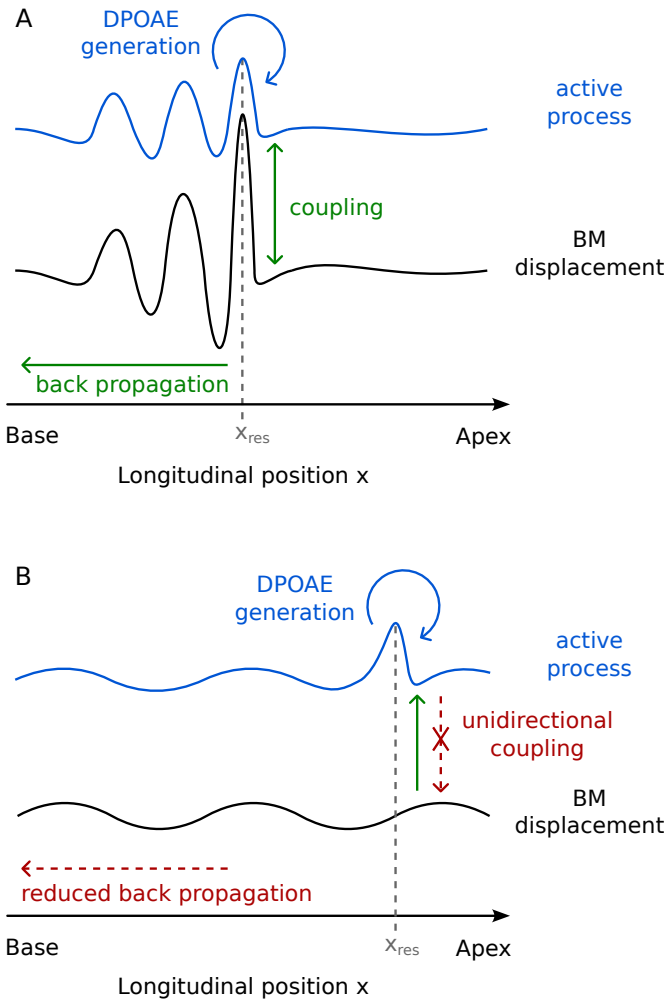

Figure 5: Basal versus apical cochlear mechanics. (A) In the basal portion of the cochlea, critical-layer absorption leads to waves on the basilar-membrane that peak at a frequency-specific point and then decline sharply. Basilar membrane motion can elicit the cochlear active process and be greatly amplified by it (bidirectional coupling). Distortion produced by the active process can hence propagate back to the middle ear through waves on the basilar membrane. (B) In the apical part of the inner ear, low-frequency waves traverse all the way to the helicotrema. The active process presumably enables an independent resonance in hair cells that may not couple back to the basilar membrane, reducing the back-propagation of distortion resulting from the active process.

tem. As an example, the neural tracking of a speech signal can decrease when speech intelligibility surpasses a certain level [33].

We believe, however, that an alternative interpretation – one that accounts for the unique mechanics of the cochlea in the low-frequency range – may be more likely. Classical theories of cochlear mechanics describe the basal, high-frequency range through the mechanism of critical-layer absorption [34, 16, 17]. A traveling wave on the basilar membrane, elicited by an incoming sound, moves toward its resonant position, where it peaks and then rapidly declines (Fig. 5A). The cochlear active process greatly amplifies the basilar-membrane motion in the peak region, yielding a nonlinear response and generating distortion that can propagate back on the basilar membrane.

In contrast, between 1.5 and 4 kHz, the tuning properties of the cochlea transition from basal to apical behavior [35, 36, 37, 38]. For frequencies below about 1.5 kHz, cochlear mechanics differs substantially from the one in the basal region. The waves elicited by these comparatively low frequencies presumably traverse the entire cochlea until the helicotrema. The active process likely yields an independent resonance at a frequency-dependent location (Fig. 5B).

Previous studies proposed that this can be achieved through unidirectional coupling from the basilar membrane to the active process: basilar-membrane movement elicits the active process in the outer hair cells, but the active process does not feed back on basilar-membrane motion [39, 40]. DPOAEs propagating back on the basilar membrane may therefore be suppressed. Attention to the resolved harmonics of a voice could then lead to a larger gain of the active process, which would be accompanied by more fine-tuned decoupling from the basilar membrane and hence smaller speech-DPOAEs. This effect could explain our observation of lower speech-DPOAEs for the stimulus F_{res} when the female voice is attended as compared to when it is ignored.

The speech-DPOAEs evoked by the F_{res} stimuli did not differ between the attended male voice condition (equivalent to ignoring the female voice) and the attended visual condition. This suggests that inner ear activity at the resolved harmonics of an ignored speaker – in this case, the female voice – remains the same in these two conditions. One possibility, in line with the above considerations regarding apical cochlear mechanics, is that the resolved harmonics of a target speaker are enhanced through the active process, but yield lower speech-DPOAEs due to unidirectional coupling. When this voice is not attended, the harmonics are less enhanced, independent of whether attention is directed towards another auditory stream or the visual signal.

Because unresolved harmonics do not produce distinct peaks along the basilar membrane, the attentional modulation observed for the M_{unres} stimulus was unexpected. However, it is important to note that the M_{unres} stimulus was presented against a background of resolved harmonics of the female voice (Fig. 6). We hypothesize that attention to the female voice enhances cochlear activity in the regions corresponding to its resolved harmonics (purple shading in Fig. 6). This enhanced activity will also affect the responses to certain unresolved harmonics of the male voice, namely those that peak in the same section of the basilar membrane. These included the waveforms used in the M_{unres} stimulus.When the attentional focus changed from the female to the male voice, we hypothesize that cochlear activity in these regions became smaller, again affecting the corresponding unresolved harmonics of the male voice, including the M_{unres} stimulus. The attentional modulation for the M_{unres} stimulus was thus likely caused by changes in cochlear activity for the resolved harmonics of the female voice.

When combined with the hypothesis of unidirectional amplification, and thus higher cochlear gain in the apical region leading to lower speech-DPOAEs, a consistent picture emerges. Attention to the resolved harmonics of the female voice leads to larger gain in the corresponding sections of the basilar membrane. This results in smaller speech-DPOAEs for both the F_{res} and the M_{unres} stimulus. The speech-DPOAEs evoked by the stimulus M_{unres} are hence smaller when attention is directed towards the female speaker (thereby ignoring the male voice) than when it is directed at the male speaker or at the text. Our results are in line with these expectations.

As expected, the delays of the speech-DPOAEs did not depend on the focus of selective attention (Fig. 4M-O). However, the delays differed between the different stimulus types as assessed in the single-speaker scenario. Speech-DPOAEs evoked by the F_{res} stimulus peaked around 2 ms, corresponding to approximately 2 cycles on the basilar membrane. Because the traveling wave, for low frequencies, accumulates about one cycle between the base and the peak location, this delay matches with the expectation of the stimuli propagating towards the peak location, the active process generating distortion there, and the latter propagating back

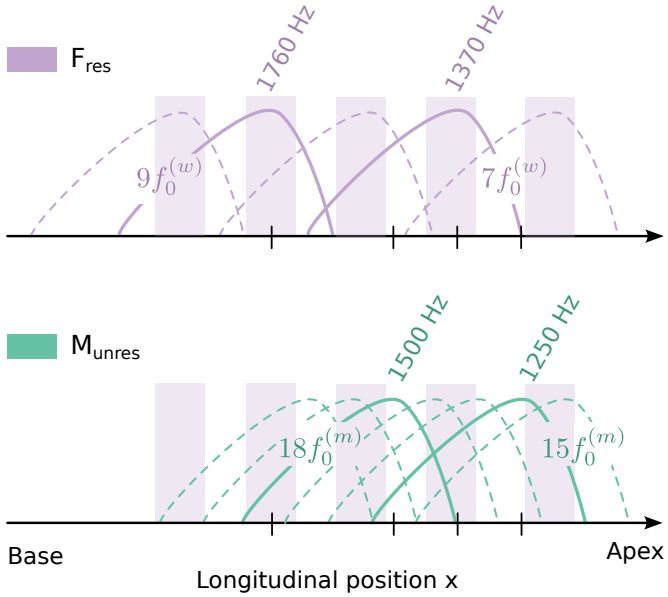

Figure 6: Resolved and unresolved harmonics. In the part of the cochlea where the F_{res} and the M_{unres} stimuli have their characteristic places, the harmonics of the female voice are resolved (top) while those of the male voice are unresolved (bottom). The cochlear active process may enhance the response at the locations where the resolved harmonics of the female voice peak (purple shading). This enhancement will then also affect the responses to the unresolved harmonics of the male voice at these locations.

through a basilar-membrane wave [41, 42, 32, 43, 16].

In contrast, the speech-DPOAEs evoked by the F_{unres} and the M_{unres} stimuli occurred at shorter latencies, 0 ms respectively 0.7 ms, corresponding to approximately 0 and 0.7 cycles. This suggests that these distortions are created basal to the peak of the corresponding basilar-membrane waves, and that they might propagate back from their generation site through a fast wave such as a compression wave [44, 45, 46].

The M_{res} stimulus did not give reliable responses and was hence excluded from more detailed analysis. It only produced a low percentage of significant peaks (Fig. 2D,E), weakly pronounced grand average peaks (Fig. 2D-E and Fig. 3C), as well as large variation in delays (Fig. 3F).

CONCLUSION

Our study demonstrates for the first time that cochlear activity related to the harmonics of voiced speech is modulated by selective attention to speech in noise. We find effects of selective attention both for resolved and unresolved harmonics, although we argue that the observed attentional effects for the unresolved harmonics originate in the attentional modulation of the resolved harmonics of the competing speaker. The attentional effects that we observe are consistent between switching attention from the target voice to a competing speaker and between switching attention from the target to a visual stimulus.

The direction of the attention effect is, however, unexpected: we find that attention reduces rather than enhances the speech-DPOAE for the target speaker. We argue that this may reflect the peculiar cochlear mechanics in the low-frequency region from which the emissions likely originate: unidirectional coupling between the basilar membrane and the active process can mean that a higher cochlear gain results in better decoupling from the basilar membrane and hence in lower otoacoustic emissions.

Comparison to attentional effects at higher frequencies, where cochlear mechanics involves critical-layer absorption, may allow to further clarify these aspects in the future. More generally, we hope that the speech-DPOAEs that we have employed here will allow to increasingly elucidate the role of the active inner ear in the complex cognition of speech.

MATERIALS AND METHODS

Participants

Speech-DPOAE measurements were conducted with $N = 40$ participants (21 female, 19 male), aged 1831 years (mean age $\pm$ SD: 25 ± 3 years). Inclusion criteria were right-handedness, native German proficiency, and the absence of neurological or hearing im-

pairments. One subject was excluded since their comprehension score was below chance level, and another due to a faulty microphone recording. 38 participants were thus included in the final analysis.

All procedures were approved by the ethics board of the University Hospital Erlangen (registration 133-12B) and were conducted in accordance with institutional regulations. Informed consent was obtained from all participants.

Speech signals

Two audiobooks were synthesized with a text-to-speech engine (ElevenLab, U.S.A.) using the voice Matilda for the female speaker and the voice Brian for the male one [47]. We chose voice parameters that yielded a large separation between the fundamental frequencies of the two voices, which resulted in an average fundamental frequency of $\bar{f}_0^{(w)} = 195 \pm 40$ Hz for the female voice and $\bar{f}_0^{(m)} = 90 \pm 20$ Hz for the male voice (mean $\pm$ SD).

The texts for the audiobooks were taken from Eine Frau erlebt die Polarnacht" [48] (Book A) and from Darum" [49] (Book B). Each text was synthesized both with the female and the male voice.

Visual distractor

The visual distractor consisted of text excerpts from a third book, Frau Ella" [50] (Book C). The text was rendered as a video in which short paragraphs appeared word by word at a comfortable reading pace and were displayed centrally on a computer screen.

Stimuli for eliciting speech-DPOAEs

Pure-tone DPOAEs were elicited by two primary frequencies, f_1 and f_2. The lower-sideband cubic distortion product $2f_1 - f_2$ is the strongest, and is maximal at a ratio of $f_2/f_1 \approx 1.2$. For its measurement we employed $f_1 = 1$ kHz and $f_2 = 1.2$ kHz.

The stimuli to elicit speech-DPOAEs were computed using an approach that we developed recently [31]. Specifically, from a particular speech signal with a time-dependent fundamental frequency $f_0(t)$, we computed narrow-band waveforms $w_n(t)$ and $w_m(t)$ ($n < m$) such that, at each time instance t, the instantaneous frequency of the waveforms was $nf_0(t)$ and $mf_0(t)$, respectively. The waveforms $w_n(t)$ and $w_m(t)$ thus corresponded to the nth and mth harmonics of the fundamental frequency. The resulting lower-sideband cubic distortion product thus had an instantaneous frequency of $(2n-m)f_0(t)$, and the corresponding waveform $w_{2n-m}(t)$ was calculated as well. Details of the computations of these waveforms are given in our previous publication [31]. The speech-DPOAE was then assessed through the cross-correlation of the waveform $w_{2n-m}(t)$ and the microphone recording.

The harmonic numbers n and m used for each of the four stimuli F_{res}, F_{unres}, M_{res}, and M_{unres} are shown in Table 1.

Table 1: The harmonic numbers n and m for the different stimuli, with the harmonic index $2n - m$ of the resulting lower-sideband cubic distortion product.

	n	m	$2n - m$
F_{res}	7	9	5
F_{unres}	15	18	12
M_{res}	6	8	4
M_{unres}	15	18	12

Experimental Setup

Experiments were conducted in a sound-proof, semi-anechoic chamber. Stimulus presentation and data acquisition were automated through PsychoPy [51]. Instructions were displayed on a screen; responses were given via mouse click.

Stimuli were presented at 44.1 kHz using a high-performance sound card (RME Fireface 802) and delivered through an extended-bandwidth otoacoustic measurement system (ER10X, Etymotics, U.S.A.), equipped with one microphone and three speakers per ear. Custom ear tips ensured optimal probe fit. Audiobooks were presented to the right ear while stimuli presentation and speech-DPOAE recordings were conducted in the left ear. For each stimulus, the two waveforms $w_n(t)$ and $w_m(t)$ were played through different speakers to avoid hardware-induced distortion.

Stimuli were presented directly to the ear canal at 37 dB SPL. This level was reported as comfortable by all participants. It was chosen to be low enough to neither evoke the medial olivocochlear (MOC) reflex nor the middle ear muscle reflex [52, 53].

Experimental Routine

Each story segment lasted about two minutes. After each such two-minute trial, participants answered three comprehension questions and rated the perceived mental effort on a 13-point Likert scale.

The experiment began with a two-minute pure-tone DPOAE measurement to verify DPOAE detectability. DPOAEs could be recorded in all participants.

Next, speech-DPOAEs were recorded in the single-speaker scenario, using both the male and the female voice in isolation. Speech-DPOAEs to both the resolved and the unresolved harmonics of the corresponding voice were measured simultaneously to confirm that both speech-DPOAEs could be measured concurrently.

The main part of the experiment consisted of the competing-speaker scenario in which both the female and the male voice were presented simultaneously. Participants were instructed to focus their attention either on the female speaker (Att. F), on the male speaker (Att. M), or on the visual task (Att. V). The target audio thereby always contained the story of Book A, either spoken by the male or the female voice, so participants could continue with the same story when switching attention between the speakers. In the Att. V. condition, participants ignored both audio streams and focused on reading Book C, presented word-by-word on a monitor.

To verify that the measurement equipment did not contribute to the speech-DPOAEs, out-of-ear control measurements were conducted. The probe was placed outside the ear canal in the center of the recording room, with all reflective surfaces avoided to minimize acoustic feedback. No speech-DPOAEs emerged in that case.

Analysis of speech-DPOAEs

Hardware-induced delays were estimated per trial by cross-correlating the stimulus waveforms with the microphone recording. The recordings were corrected for the delays before further analysis.

Speech-DPOAEs were then computed by cross-correlating the corresponding waveform with the microphone recording. Grand averages were computed by averaging the envelopes of the cross-correlations across all two-minute trials, for each of the three attentional conditions and each of the four stimulus types. A peak in the grand average was considered significant if it exceeded the maximum of the noise. The noise level was thereby determined from the values of the envelope of the correlation coefficients at time lags of -750 to -70 ms and from 70 to 750 ms, that is, at delays at which no speech-DPOAEs should occur.

To detect speech-DPOAEs at the level of individual two-minute trials, the expected window for peak delays was defined as 1 ± 3 ms (mean $\pm$ SD), based on the grand averages of F_{res} and F_{unres} stimuli. Stimuli corresponding to the male voice were excluded, as the grand average for the M_{res} stimulus did not yield reliable results, and to avoid introducing an imbalance in peak widths between resolved and unresolved harmonics the M_{res} stimulus was disregarded as well. A peak within this window was considered significant if it exceeded the 97th percentile of the noise. The percentage of significant trials was then computed for each stimulus type.

To compare speech-DPOAEs across attentional conditions, the cross-correlation coefficients at the grand average peak delay were extracted for each two-minute trial, matched per stimulus type and attentional condition, and averaged per participant. Single-speaker comparisons used unpaired t- or MannWhitney U tests. For the data obtained from the competing-speaker scenario, paired t-tests or Wilcoxon signed-rank tests were used, with outlier exclusion when justified.

To evaluate differences between speech-DPOAEs evoked by resolved and unresolved harmonics, peak delays per individual two-minute trials were extracted for all significant peaks, matched per stimulus type and condition and averaged per participant. Statistical comparisons used unpaired t- or MannWhitney U tests.

All p-values were corrected for multiple comparisons using the False Discovery Rate correction.

DATA AVAILABILITY

All recorded data, speech signals, and speech-DPOAE stimuli from this study are openly available on Zenodo at https://zenodo.org/records/16837673.

CODE AVAILABILITY

All custom code used for stimulus generation, speech-DPOAE analysis via cross-correlation, and statistical evaluation is openly available on GitHub at https://github.com/janna-stb/dpoae_attention_study.git under the MIT License.

ACKNOWLEDGMENTS

This work was funded by the Deutsche Forschungsgemeinschaft (German Research Foundation) through grant 514955521 (to T.R.).

AUTHOR CONTRIBUTIONS

J.S. and T.R. designed research and analyzed data; J.S. performed research and wrote the paper; T.R. revised the paper.

AUTHOR COMPETING INTERESTS

The authors declare no competing interests.

REFERENCES

[1] Josh H. McDermott. The cocktail party problem. *Current Biology*, 19(22):R1024–R1027, 2009.

[2] E. Colin Cherry. Some Experiments on the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*, 25(5):975–979, 1953.

[3] Reinier Plomp. Auditory handicap of hearing impairment and the limited benefit of hearing aids. *The Journal of the Acoustical Society of America*, 63(2):533–549, 1978.

[4] Kenneth R. Pugh, Bennett A. Shaywitz, Sally E. Shaywitz, Robert K. Fulbright, Dani Byrd, Pawel Skudlarski, Donald P. Shankweiler, Leonard Katz, R. Todd Constable, Jack Fletcher, Cheryl Lacadie, Karen Marchione, and John C. Gore. Auditory Selective Attention: An fMRI Investigation. *NeuroImage*, 4(3):159–173, 1996.

[5] Peter Lakatos, Gabriella Musacchia, Monica N. OConnel, Arnaud Y. Falchier, Daniel C. Javitt, and Charles E. Schroeder. The Spectrotemporal Filter Mechanism of Auditory Selective Attention. *Neuron*, 77(4):750–761, 2013.

[6] Nima Mesgarani and Edward F. Chang. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, 485(7397):233–236, 2012.

[7] Nai Ding and Jonathan Z. Simon. Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, 109(29):11854–11859, 2012.

[8] Cort Horton, Michael D'Zmura, and Ramesh Srinivasan. Suppression of competing speech through entrainment of cortical oscillations. *Journal of Neurophysiology*, 109(12):3082–3093, 2013.

[9] Russell F. Huffman and O. W. Henson. The descending auditory pathway and acousticomotor systems: connections with the inferior colliculus. *Brain Research Reviews*, 15(3):295–323, 1990.

[10] James O Pickles. *An introduction to the physiology of hearing*, volume 2. Academic press London, 1988.

[11] Jeffery A. Winer, David T. Larue, James J. Diehl, and Brenda J. Hefti. Auditory cortical projections to the cat inferior colliculus. *Journal of Comparative Neurology*, 400(2):147–174, 1998.

[12] Gary C. Galbraith, Darlene M. Olfman, and Todd M. Huffman. Selective attention affects human brain stem frequency-following response. *NeuroReport*, 14(5):735, 2003.

[13] Antonio Elia Forte, Octave Etard, and Tobias Reichenbach. The human auditory brainstem response to running speech reveals a subcortical mechanism for selective attention. *eLife*, 6:e27203, 2017.

[14] Octave Etard, Mikolaj Kegler, Chananel Braiman, Antonio Elia Forte, and Tobias Reichenbach. Decoding of selective attention to continuous speech from the human auditory brainstem response. *NeuroImage*, 200:1–11, 2019.

[15] Daniel J. Strauss, Farah I. CoronaStrauss, Adrian Mai, and Steven A. Hillyard. Unraveling the effects of selective auditory attention in ERPs: From the brainstem to the cortex. *NeuroImage*, 316:121295, 2025.

[16] Luis Robles and Mario A. Ruggero. Mechanics of the Mammalian Cochlea. *Physiological Reviews*, 81(3):1305–1352, 2001.

[17] Tobias Reichenbach and A J J Hudspeth. The physics of hearing: fluid mechanics and the active process of the inner ear. *Rep. Prog. Phys.*, 77:45, 2014.

[18] P Dallos. The active cochlea. *The Journal of Neuroscience*, 12(12):4575–4585, 1992.

[19] A. J. Hudspeth. Integrating the active process of hair cells with cochlear function. *Nature Reviews Neuroscience*, 15(9):600–614, 2014.

[20] John J. Jr Guinan. Olivocochlear Efferents: Anatomy, Physiology, Function, and the Measurement of Efferent Effects in Humans. *Ear and Hearing*, 27(6):589, 2006.

[21] Enrique A. Lopez-Poveda. Olivocochlear Efferents in Animals and Humans: From Anatomy to Clinical Relevance. *Frontiers in Neurology*, 9, 2018.

[22] C. Meric and L. Collet. Visual attention and evoked otoacoustic emissions: A slight but real effect. *International Journal of Psychophysiology*, 12(3):233–235, 1992.

[23] Patricia T. Michie, Eric L. LePage, Nadia Solowij, Michael Haller, and Lyn Terry. Evoked otoacoustic emissions and auditory selective attention. *Hearing Research*, 98(1):54–67, 1996.

[24] Kyle P. Walsh, Edward G. Pasanen, and Dennis McFadden. Changes in otoacoustic emissions during selective auditory and visual attention. *The Journal of the Acoustical Society of America*, 137(5):2737–2757, 2015.

[25] David P. Messing, Lorraine Delhorne, Ed Bruckert, Louis D. Braida, and Oded Ghitza. A non-linear efferent-inspired model of the auditory system; matching human confusions in stationary noise. *Speech Communication*, 51(8):668–683, 2009.

[26] Nicholas R. Clark, Guy J. Brown, Tim Jürgens, and Ray Meddis. A frequency-selective feedback model of auditory efferent suppression and its implications for the recognition of speech in noise. *The Journal of the Acoustical Society of America*, 132(3):1535–1541, 2012.

[27] Joshua G Bernstein and Andrew J Oxenham. Pitch discrimination of diotic and dichotic tone complexes: Harmonic resolvability or harmonic number? *J. Acoust. Soc. Am*, 113:3323–3334, 2003.

[28] Christopher J. Plack, Andrew J. Oxenham, Richard R. Fay, and Arthur N. Popper. *Pitch: neural coding and perception*, volume 24. Springer, 2005.

[29] Christophe Micheyl and Andrew J. Oxenham. Across-frequency pitch discrimination interference between complex tones containing resolved harmonics. *The Journal of the Acoustical Society of America*, 121(3):1621–1631, 2007.

[30] Samuele Carcagno and Christopher J. Plack. Pitch discrimination learning: specificity for pitch and harmonic resolvability, and electrophysiological correlates. *Journal of the Association for Research in Otolaryngology*, 12(4):503–517, 2011.

[31] Marina Saiz-Alía, Peter Miller, and Tobias Reichenbach. Otoacoustic Emissions Evoked by the Time-Varying Harmonic Structure of Speech. *eNeuro*, 8(2), 2021.

[32] Rudolf Probst, Brenda L. Lonsbury-Martin, and Glen K. Martin. A review of otoacoustic emissions. *The Journal of the Acoustical Society of America*, 89(5):2027–2067, 1991.

[33] Anne Hauswald, Anne Keitel, Ya-Ping Chen, Sebastian Rösch, and Nathan Weisz. Degradation levels of continuous speech affect neural speech tracking and alpha power differently. *European Journal of Neuroscience*, 55(11-12):3288–3302, 2022.

[34] James Lighthill. Energy flow in the cochlea. *Journal of Fluid Mechanics*, 106:149–213, 1981.

[35] Christopher A. Shera, John J. Guinan, and Andrew J. Oxenham. Otoacoustic estimation of cochlear tuning: Validation in the chinchilla. *Journal of the Association for Research in Otolaryngology*, 11(3):343–365, 2010.

[36] Jonathan H. Siegel, Amanda J. Cerka, Alberto Recio-Spinoso, Andrei N. Temchin, Pim Van Dijk, and Mario A. Ruggero. Delays of stimulus-frequency otoacoustic emissions and cochlear vibrations contradict the theory of coherent reflection filtering. *The Journal of the Acoustical Society of America*, 118(4):2434–2443, 2005.

[37] Andrei N. Temchin, Nola C. Rich, and Mario A. Ruggero. Threshold Tuning Curves of Chinchilla Auditory-Nerve Fibers. I. Dependence on Characteristic Frequency and Relation to the Magnitudes of Cochlear Vibrations. *Journal of Neurophysiology*, 100(5):2889–2898, 2008.

[38] Jonathan Ashmore. Cochlear Outer Hair Cell Motility. *Physiological Reviews*, 88(1):173–210, 2008.

[39] Tobias Reichenbach and A. J. Hudspeth. A ratchet mechanism for amplification in low-frequency mammalian hearing. *Proceedings of the National Academy of Sciences*, 107(11):4973–4978, 2010.

[40] Tobias Reichenbach and A. J. Hudspeth. Unidirectional Amplification as a Mechanism for LowFrequency Hearing in Mammals. *AIP Conference Proceedings*, 1403(1):507–512, 2011.

[41] Christopher A. Shera and John J. Guinan. Evoked otoacoustic emissions arise by two fundamentally different mechanisms: A taxonomy for mammalian OAEs. *The Journal of the Acoustical Society of America*, 105(2):782–798, 1999.

[42] Arnold Tubis, Carrick L. Talmadge, and Christopher Tong. Modeling the temporal behavior of distortion product otoacoustic emissions. *The Journal of the Acoustical Society of America*, 107(4):2112–2127, 2000.

[43] Richard D. Knight and David T. Kemp. Wave and place fixed DPOAE maps of the human ear. *The Journal of the Acoustical Society of America*, 109(4):1513–1525, 2001.

[44] Paul Avan, Pascal Magnan, Jacek Smurzynski, Rudolf Probst, and Armand Dancer. Direct evidence of cubic difference tone propagation by intracochlear acoustic pressure measurements in the guinea-pig. *European Journal of Neuroscience*, 10(5):1764–1770, 1998.

[45] Ales Vetesnik, Renato Nobili, and Anthony Gummer. How Does the Inner Ear Generate Distortion Product Otoacoustic Emissions? *ORL*, 68(6):347–352, 2006.

[46] Wenxuan He, Anders Fridberger, Edward Porsov, Karl Grosh, and Tianying Ren. Reverse wave propagation in the cochlea. *Proceedings of the National Academy of Sciences*, 105(7):2729–2733, 2008.

[47] ElevenLabs. ElevenLabs text-to-speech engine, 2025.

[48] Christiane Ritter. *Eine Frau erlebt die Polarnacht*. Refinery, 2016.

[49] Daniel Glattauer. *Darum*. Paul Zsolnay Verlag, 2018.

[50] Florian Beckerhoff. *Frau Ella*. dotbooks, 2022.

[51] Jonathan Peirce, Jeremy R. Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1):195–203, 2019.

[52] Skyler G. Jennings. The role of the medial olivocochlear reflex in psychophysical masking and intensity resolution in humans: a review. *Journal of Neurophysiology*, 125(6):2279–2308, 2021.

[53] Monica Trevino, Andie Zang, and Edward Lobarinas. The middle ear muscle reflex: Current and future role in assessing noise-induced cochlear damage. *The Journal of the Acoustical Society of America*, 153(1):436–445, 2023.