Title: Perceptual gating of a brainstem reflex facilitates speech understanding in humans

Heivet Hernandez-Perez * [1], Jason Mikiel-Hunter [1], David McAlpine [1], Sumitrajit Dhar [2], Sriram Boothalingam [3], Jessica J.M. Monaghan [4] ‡, Catherine M. McMahon [1] ‡

[1] The Australian Hearing Hub, Macquarie University, Australia[2] Department of Communication Sciences and Disorders, Northwestern University, USA [3] University of Wisconsin-Madison, USA[4] National Acoustic Laboratories, Australia

‡ C.M.M and J.J.M.M contributed equally to this work

* Corresponding author
Heivet Hernandez-Perez.
Email: heivet.hernandez-perez@mq.edu.au

Abstract

Navigating "cocktail party" situations by enhancing foreground sounds over irrelevant background information is typically considered from a cortico-centric perspective. However, subcortical circuits, such as the medial olivocochlear reflex (MOCR) that modulates inner ear activity itself, have ample opportunity to extract salient features from the auditory scene prior to any cortical processing. To understand the contribution of auditory subcortical nuclei and the cochleae, physiological recordings were made along the auditory pathway while listeners differentiated non(sense)-words and words. Both naturally-spoken and intrinsically-noisy, vocoded speech — filtering that mimics processing by a cochlear implant—significantly activated the MOCR, whereas listening to speech-in-background noise revealed instead engagement of midbrain and cortical resources. An auditory periphery model reproduced these speech degradation-specific effects, providing a rationale for goal-directed MOCR gating to enhance representation of speech features in the auditory nerve. These results highlight two strategies co-existing in the auditory system to accommodate categorically different speech degradations.

Introduction

Cocktail-party listening, the ability to focus on a single talker in a background of simultaneous, overlapping conversations, is critical to human communication, and a long-sought goal of hearing technologies, including automated speech-recognition software. Problems listening in background noise are a key complaint of many listeners with even mild hearing loss, and a stated factor in the non-use and non-uptake of hearing devices. However, despite its importance in every-day listening tasks, and its relevance to hearing impairment, physiological mechanisms that enhance attended speech remain poorly understood. In addition to circuits in the inner ear and lower brainstem evolved to enhance automatically the neural representation of ecologically relevant sounds, it is likely that such a critical goal-directed behaviour as cocktail-party listening also relies on top-down, or attentionally-mediated, processes to emphasize perceptually relevant sounds, and suppress those that are irrelevant (Ahveninen et al., 2006; Knudsen, 2007). However, their specific roles in complex listening tasks remain to be determined.

One likely mechanistic pathway is through efferent feedback. The auditory efferent system is extensive, with multi-synaptic connections extending from auditory cortex (AC) all the way to the inner ear (Dragicevic et al., 2015; Mulders and Robertson, 2000; Xiao and Suga, 2002). In particular, reflexive activation of fibers in the medial olivocochlear reflex (MOCR) innervating the outer hair cells (OHCs)—electromotile elements responsible for the cochlea's active amplifier—is known to reduce cochlear gain (Ashmore et al., 2010), thereby increasing the overall dynamic range of the inner ear to facilitate sound encoding in high levels of background noise (Almishaal and Jennings, 2016). Ipsilateral and contralateral MOC fibers originate in medial divisions of the superior olivary complex of the brainstem and synapse directly on the basal aspects of OHCs (Warr and Guinan, 1979) to modulate mechanical (Cooper and Guinan, 2006; Murugasu and Russell, 1996) and neural sensitivity to sound (Guinan and Gifford, 1988; Winslow and Sachs, 1987). MOC neurons are also innervated by descending fibers from AC and midbrain neurons, providing a potential means of modulating cochlear activity in a task-related or attention-dependent manner (Faye-Lund, 1986; Mulders and Robertson, 2000; Suthakar and Ryugo, 2017).

Efferent-mediated changes in cochlear gain can be assessed by measuring otoacoustic emissions (OAEs), energy generated by the active OHCs and measured non-invasively as sound from the ear canal (Kemp, 1978). When transient sounds, such as clicks, are presented in the presence of noise in the opposite ear; OAE amplitudes are reduced, reflecting increased MOC efferent activity (Collet et al., 1990). Nevertheless, despite the presumed benefits of MOC activation in enhancing speech-in-noise perception, it remains unclear under which listening conditions, e.g. whether listeners are actively engaged in a task, the MOC contributes to speech perception (de Boer et al., 2012; Mertes et al., 2019; Mishra and Lutman, 2014). Although it has been speculated that MOC-mediated changes in cochlear gain could enhance speech coding in background noise (Guinan, 2017), the magnitude of OAE inhibition has been reported as either increased (Giraud et al., 1997; Mishra and Lutman, 2014), reduced (de Boer et al., 2012; Harkrider and Bowers, 2009) or unaffected (Stuart and Butler, 2012; Wagner et al., 2008) in participants with improved speech-in-noise perception. These confounding effects of auditory attention on cochlear gain could depend on factors such as task difficulty or relevance (e.g. speech *vs.* non-speech tasks) or even methodological differences in recording and analysing inner ear signatures such as OAEs (Mertes et al., 2019; Mishra and Lutman, 2014). To this end, the role of the MOCR in reducing cochlear gain during goal-directed hearing in normal hearing subjects (i.e. with physiologically normal OHCs) remains controversial.

Here, we examined the role of the auditory efferent system in active *vs*. passive listening tasks for three commonly employed speech manipulations: vocoding of 'clean' speech; speech presented in a background of multi-talker 'babble' noise and speech presented in a speech-shaped noise i.e. noise with the same long-term spectrum as speech. As well as comparing measures of listening performance across these manipulations (detecting non-words in a string of Australian-English words and non-words), we also assessed the sensitivity of the central auditory pathway including cortical auditory responses in both the active listening task and when listeners were required to ignore the task and watch a silent, non-subtitled film.

Maintaining a fixed task difficulty across speech manipulations, we found measures of hearing function at the level of the cochlea, brainstem, midbrain and cortex to be modulated differently depending on the degradation type applied to speech sounds, and on whether or not speech was

actively attended. Specifically, the MOCR, assessed in terms of the magnitude of click-evoked OAEs (CEOAEs), was activated by vocoded speech—an intrinsically degraded speech signal—but not by otherwise 'clean' speech presented in either babble-noise or speech-shaped noise. Neural activity at the first synaptic stage of central processing in the cochlear nucleus (CN)—assessed physiologically through auditory brainstem responses (ABRs)—confirmed the reduction in cochlear gain for actively attended vocoded speech, but not speech-in-noise. Conversely, neural activity generated by the auditory midbrain was significantly increased in active *vs*. passive listening for speech in babble and speech-shaped noise, but not for vocoded speech. This increase was associated with elevated cortical markers of listening effort for the speech-in-noise conditions. A model of the auditory periphery including an MOC circuit with biophysically-realistic temporal dynamics confirmed the stimulus-dependent role of the MOCR in enhancing neural coding of speech signals. Our data suggest that otherwise identical performance in active listening tasks may invoke quite different efferent circuits, requiring different levels of listening effort, depending on the type of stimulus degradation experienced.

Results

Iso-performance in three manipulations of speech intelligibility.

We first assessed speech intelligibility—specifically the ability to discriminate between Australian-English words and non-words—when speech was degraded by three different speech manipulations: 1) vocoding of the entire speech waveform; 2) the addition of 8-talker babble noise to 'clean' speech (i.e. speech-in-quiet) and 3) the addition of speech-shaped noise to 'clean' speech. In an active-listening condition, participants were asked to respond (by means of a button press) when they heard a non-word in a string of words and non-words (Figure1A).

Three levels of task difficulty were achieved by either altering the number of vocoded channels—16 (Voc16), 12 (Voc12) and 8 (Voc8) channels)—or by altering the signal-to-noise ratio (SNR) when speech was masked by babble noise—+10 (BN10) and +5 (BN5) dB SNR—or speech-shaped noise—+8 (SSN8) and +3 (SSN3) dB SNR. This was statistically confirmed in all 56 listeners (n=27 in the vocoded condition, and n=29 in the two masked conditions) who showed consistently better performance—a higher rate of detecting non-words—in the less degraded conditions i.e. more vocoded channels or higher SNRs in the masked manipulations (Repeated Measures ANOVA (rANOVA) [vocoded: [$F(3, 78) = 70.92$, $p = 0.0001$]; babble noise: [$F(3, 78) = 70.92$, $p = 0.0001$] and speech-shaped noise: [$F(2, 56) = 86.23$, $p = 0.0001$], see Figure 1B for *post hoc* analysis with Bonferroni corrections).

Iso-performance levels were achieved across speech manipulation experiments: the highest speech discrimination was observed in the two natural speech conditions—one used in the vocoding experiment, the other in the two masked conditions—and was indistinguishable [One-way ANOVA: $F(1, 54) = 0.43$, $p = 0.84$]; a similarly moderate level of performance (significantly lower than performance for natural speech) was achieved for Voc16/ Voc12 (n.s. compared to each other; $p = 0.11$), BN10 and SSN8 conditions: [$F(3, 108) = 0.67$, $p= 0.57$]. The lowest level of performance (which was significantly lower than high and moderate performance conditions) was observed for Voc8, BN5 and SSN3: [$F(3, 81) = 0.07$, $p = 0.72$]).

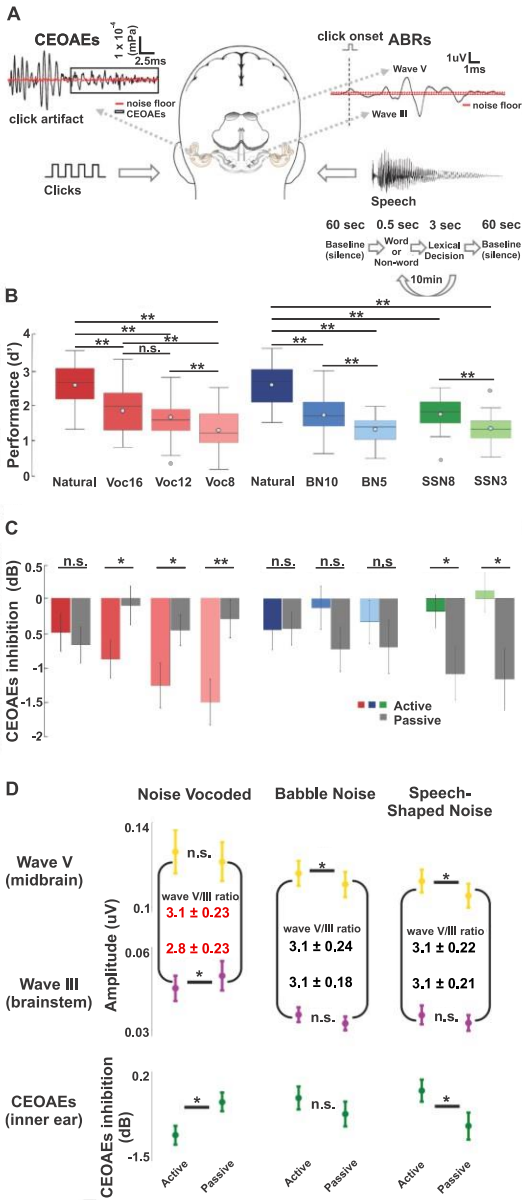

Figure 1. Behavioural and physiological measurements during active and passive listening. **A.** Schematic of the experimental paradigm (adapted with permissions (Schaette and McAlpine, 2011)) Each ear received one type of stimulus (click in the left and speech tokens in the right). Clicks were continuously presented whereas speech tokens were delivered for 10 minutes after a minute of baseline (only clicks being delivered) at the beginning and at the end of each condition. Participants had 3 sec to make a lexical decision task (discriminate words *vs.* non-words) in the active listening condition whereas in the passive condition participants were asked to ignore all the auditory stimuli and watch a movie. Upper left panel shows the CEOAEs analysis window obtained from the inner ear recordings. Upper right panel corresponds to an averaged click-evoked ABRs. **B**. Performance during the lexical decision task. Mean d' is represented in white circles, median as a horizontal line and black dots correspond to outliers. The upper and lower borders of the boxplot represent the 1st (q1) and 3rd (q3) quartiles respectively while the whiskers denote the interquartile range (IQR =q3-q1). Within speech conditions showed that highest performance was always achieved for natural speech compared to the all vocoded, babble noise and speech-shaped noise manipulations *(*$p = 0.001$). Performance was moderately-high (statistically lower than Natural speech but higher than Voc8, BN5 and SSN3) for Voc16 and Voc12 channels vocoded (non-significant differences represents n.s., $p = 0.11$) as well as for BN10 and SSN8 respectively. The lowest level of performance was predictably observed for Voc8, BN5 and SSN3 ($p =0.001$). **C.** Click-Evoked OAEs inhibition. Bars represent the mean CEOAE magnitude changes relative to the baseline and the whiskers denote the error bars represent SEM. **D.** Physiological measurements collapsed across attentional states (Active (A); Passive (P)) (** $p < 0.01$; * $p < 0.05$). Data reported as means ± SEM.

The cochlear efferent system is modulated by task engagement in a stimulus-dependent manner

To assess the effect of active *vs*. passive listening and speech manipulation on inner ear activity, CEOAEs were continuously recorded whilst participants both actively performed the lexical task, and when they passively listened to the same *corpus* of speech tokens (Figure 1A). Compared to baseline CEOAE measures obtained in the absence of speech, CEOAEs were significantly reduced in magnitude during active listening of natural speech, and in all vocoded conditions (natural: [t (24) = -2.33, p = 0.03]; Voc16: [t (23) = 3.40, p = 0.002]; Voc12: [t (24) = 3.98, p = 0.001] and Voc8: [t (25) = 5.14, p = 0.001]). Conversely, during passive listening, CEOAEs obtained during naturally-spoken, but not vocoded speech were significantly smaller than baseline: [t (25) = 2.29, p = 0.03], as were CEOAEs recorded during the two masked conditions at all SNRs (natural: [t (26) = 2.17, p = 0.04]; BN10: [t (28) = 2.80, p = 0.009] and BN5: [t (28) = 2.36, p = 0.02]; SSN8: [t (28) = 3.37, p = 0.002] and SS3: [t (28) = 3.50, p = 0.002]). This suggests that auditory efferent activity is modulated differently in active and passive listening, and by the different speech manipulation types, despite iso-performance across experiments.

We calculated the reduction in CEOAEs between baseline and experimental conditions (CEOAE inhibition)—a proxy for MOCR activation—to quantify auditory efferent control of cochlear gain for active and passive listening, and for the different speech manipulations. For vocoded speech, CEOAE inhibition was significantly greater when participants were actively engaged in the lexical task (-1.01 ± 0.18 dB) compared to when asked to ignore the auditory stimuli (-0.38 ± 0.18 dB): rANOVA: [F (1, 22) = 8.49, p = 0.008], (Figure 2A). Moreover, a significant interaction was observed between conditions and stimulus-type: [F (3, 66) = 2.80, p = 0.046], indicating that CEOAE inhibition was stronger for all vocoded conditions in which listeners needed to make lexical decisions, compared to when they were not—Voc16: ([t (23) = -2.16, p = 0.04]; Voc12: [t (24) = -2.19, p = 0.038] and Voc8: [t (25) =3.51, p = 0.002]). Task engagement did not alter CEOAE inhibition for natural speech [t (24) = 0.62, p = 0.54].

In contrast, speech-in-speech-shaped noise elicited opposite effects to vocoded speech (Figure 1C). CEOAE inhibition was significantly stronger during passive (-0.83 ± 0.26) compared to active listening (-0.16 ± 0.21): rANOVA: [F (1, 24) = 4.44, p = 0.046]. A significant interaction between condition and stimulus-type: [F (2, 48) = 4.67, p = 0.014] confirmed this at both SNRs tested for speech-shaped noise: (SSN8 [t (27) = 2.71, p = 0.01] and SSN3 [t (28) = 2.67, p = 0.012]) (Figure 1C). Intermediate effects on CEOAEs were observed for words masked by babble noise, with CEOAE magnitudes significantly different from baseline during passive, but not active, listening. In summary, cochlear gain is suppressed during active listening of vocoded speech, slightly but significantly during passive listening in babble noise, and strongly suppressed during passive listening in speech-shaped noise. Together, our data suggest that the MOCR is modulated by task engagement, and strongly depends on speech manipulation type.

Auditory brainstem activity reflects changes in cochlear gain

To determine the effects of active *vs*. passive listening on neural activity in subcortical auditory regions, we simultaneously measured auditory brainstem responses (ABRs) during OAE measurement to the same clicks used to evoke CEOAEs. Click-evoked ABRs—measured during simultaneous presentation of vocoded speech—showed task-engagement-specific effects similar to the effects observed for CEOAE measurements. Specifically, we observed a significant asymmetry in the ratio of wave V to wave III—roughly corresponding to neural activity generated by the inferior colliculus (IC) and CN respectively (Møller and Burgess, 1986)— when listeners were engaged in the lexical task (waves V/III ratio = 3.01 ± 0.23) compared to when they were asked to ignore the auditory stimuli (waves V/III ratio =2.77 ± 0.23): [F (1, 12) = 6.07, p = 0.03] (Figure 1D). An rANOVA for wave V and III amplitudes demonstrated that wave V remained unchanged [F (1, 12) =1.75, p = 0.21], whereas wave III was significantly smaller [F (1, 12) = 6.87, p = 0.02] in active, compared to passive, listening. This is consistent with the reduced magnitude of CEOAEs measured during active listening in the vocoded condition, since a reduction in cochlear gain reduces activity of auditory-nerve (AN) afferents synapsing in the CN. Conversely, although wave III was unchanged in both masked conditions for active *vs*. passive listening, wave V was significantly enhanced: [F (1, 26) = 5.67, p = 0.025 and F (1, 25) = 8.91, p = 0.006] when a lexical decision was required. Combined with evidence of reduced cochlear gain for vocoded speech but

unaltered, and even enhanced, cochlear gain in active compared to passive listening for speech-in-background noise, the ABR data suggest different brain mechanisms are tapped across the different speech manipulations in order to maintain iso-performance levels.

Cochlear efferent feedback improves the neural representation of vocoded speech, but not speech-in-noise

To understand why control of cochlear gain appears to depend on how speech is degraded, we implemented a model of the auditory periphery incorporating efferent feedback through a brainstem circuit with biophysically-realistic temporal dynamics (Meddis, 2016). Natural and degraded tokens of speech, generating iso-performance in the active task (Voc8; BN5; SSN3), were presented individually to the model at 75 dB, with and without efferent feedback included (Figure 2A), to assess how cochlear gain inhibition affects neural representation of speech in individual AN fibers. Chosen for their importance to signal detection in noise as well as their lack of saturation at high intensity levels (Costalupes et al., 1984; Kujawa and Liberman, 2015), we performed this analysis first for low-spontaneous rate, high-threshold AN fibers.

The neural cross-correlation coefficient, ρENV, was calculated for seven logarithmically-spaced frequencies between 1-2 kHz, to quantify how similar neural representations of speech envelopes were in each degraded condition, with and without efferent feedback, to a naturally-spoken control condition with efferent feedback (where 0 is independent and 1 is identical) (Figure 2B and 2C). Efferent feedback was active in the control as both CEOAE (Figure 1C) and model data (Figure 2A and Figure supplement 2) indicated benefits of suppressing the cochlear gain on the perception and neural encoding of naturally-spoken words. We therefore considered this an optimal AN representation of the speech envelope. For an example word, 'Fuzz', the three degraded versions in the absence of efferent feedback already displayed high average similarity in their neural envelopes to their natural controls (Figure 2A) (mean ρENV_{AN} for Voc8 = 0.84 ± 0.01; mean ρENV_{AN} for BN5 = 0.65 ± 0.01; mean ρENV_{AN} for SSN3 = 0.63 ± 0.01). However, the addition of efferent feedback to the different degraded conditions increased the average ρENV significantly for Voc8 (mean ΔρENV = +3.46 ± 0.86%, [$t(6) = 10.60$, $p = 0.0001$]); but reduced it significantly for BN5 (mean ΔρENV for BN5 = -3.47 ± 1.88%, [$t(6) = -4.88$, $p = 0.0028$]; and also, but not significantly so, for SSN3 (mean ΔρENV for SSN3 -1.42 ± 4.01%, $p = 0.3859$). This confirms our hypothesis that inhibition of cochlear gain via MOCR activation selectively enhances neural encoding of the modulation envelope of vocoded speech, but not of masked speech.

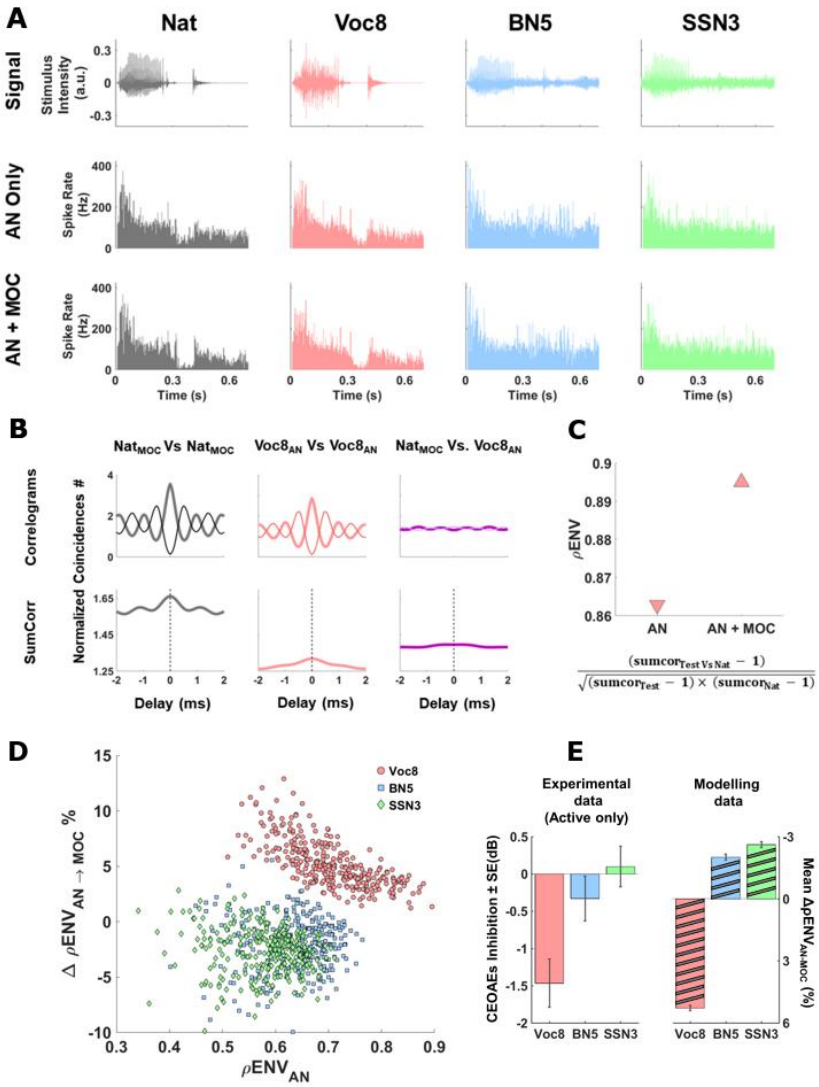

Figure 2. Model auditory periphery output with and without efferent feedback. **A.** Presentation of natural and degraded versions of the word 'Fuzz' with and without efferent feedback. 'Fuzz' waveforms for natural (*dark gray*, far left), Voc8 (*pink*, second left), BN5 (*light blue*, second right) and SSN3 (*green*, far right) conditions are shown in the top row. Post-stimulus-time-histograms (400 fibers/1ms bin-width) were calculated for low SR AN fibers (Characteristic Frequency: 1kHz) with (bottom row) and without (middle row) efferent feedback. Including efferent feedback reduced activity during quiet for natural condition (and Voc8, but less so) whilst maintaining high spiking rates at peak sound levels (e.g. at 0.075, 0.3 and 0.45ms). No changes in neural representation of signal were visually evident for BN5 and SSN3 'Fuzz'. **B & C**. Quantifying ρENV for Voc8 'Fuzz' without efferent feedback. Sumcorr plots (bottom row, B) were generated by adding Shuffled Auto-Correlograms (*thick lines*, top left/middle panels, B) and Cross-Correlograms (*thick line*, top right panel, B) to cross-polarity Correlograms (*thin lines*, top row, B) using naturally-spoken 'Fuzz' with efferent feedback (left/right columns, B) and Voc8 'Fuzz' without efferent feedback (middle/right columns, B). ρENV for Voc8 'Fuzz' without efferent feedback (AN, *downward-pointing triangle*, C) was calculated from SumCorr peaks in B. ρENV with efferent feedback (AN+MOC, *upward-pointing triangle*, C) is also displayed. **D & E**. Comparing ΔρENVs for 300 words after introduction of efferent feedback. Mean percentage changes in ρENVs (calculated across 7 frequencies between 1-2kHz) after adding efferent feedback were plotted against ρENV without efferent feedback for degraded versions of 300 words (each symbol represents one word). ΔρENVs were always positive for Voc8 words (*pink circles*, D) (Max-Min ΔρENV for Voc8: +12.92 to +1.24%), appearing largest for words with lowest ρENVs without efferent feedback. This relationship was absent for BN5 (*light blue squares*, D) and SSN3 (*green diamonds*, D) words whose ΔρENV ranges spanned the baseline (Max-Min ΔρENV for BN5: +5.57 to -9.95%; Max-Min ΔρENV for SSN3: +2.83 to -9.85%). Progression of mean ΔρENVs (± SEM) for model data (*diagonally-striped bars*, right, E) mirrored that of active-task, CEOAE data (mean ± SEM) (*solid-color bars*, left, E).

Given the variety of acoustic waveforms in our CNC word *corpus*, we expanded analysis to include 300 words from the listening task (150 stop/non-stop consonant words) (Figure 2D). Despite speech token diversity, the effects of adding efferent feedback on ρENV were consistently dependent on stimulus manipulation type. Neural encoding of speech envelopes significantly improved with efferent feedback for all vocoded words (pink circles, Figure 2D) (mean ΔρENV for Voc8= +5.30 ± 0.12%, [$t(299) = 42.71$, $p = 0.0001$]), with the largest enhancements observed for words with the lowest ρENV values in the absence of efferent feedback. In contrast, no such relationship was observed for words in BN5 (light blue squares, Figure 2D) or SSN3 (green diamonds, Figure 2D). Moreover envelope coding in both conditions was significantly impaired on average when efferent feedback was introduced (mean ΔρENV for BN5 = -2.01 ± 0.14, [$t(299) = -14.37$, $p = 0.0001$]; mean ΔρENV for SSN3 = -2.62 ± 0.12, [$t(299) = -22.03$, $p = 0.0001$]). An improvement in envelope encoding with efferent feedback for vocoded, but not for speech-in-noise conditions, was also observed for high-spontaneous rate, low-threshold AN fibers (Figure supplement 3), despite their poorer dynamic range at 75dB SPL likely impacting their overall envelope encoding ability.

The progression in neural coding enhancement by including MOC feedback across degraded conditions (diagonally-striped bars, Figure 2E) mirrored that of CEOAE inhibition (solid-color bars, Figure 2E), and ABR data (Figure 1D) for corresponding active listening conditions. In summary, where efferent feedback was beneficial to envelope encoding for vocoded speech, significant CEOAE inhibition (indicating reduced cochlear gain) was observed. This was not the case for babble or speech-shaped noise—conditions in which implementing efferent feedback generated poorer neural representations of the stimulus envelope than when absent.

Cortical evoked potentials are enhanced during active speech perception

The MOCR's lack of contribution during active listening to speech masked by speech-like sounds (i.e. babble and speech-shaped noise) suggested other compensatory brain mechanism/s were contributing to maintain iso-performance with vocoded speech. One possibility was that higher brain centres provided top-down, perhaps attention-driven, enhancement of speech processing in background noise. The significant increase observed in ABR wave V for active speech-in-noise conditions was potential evidence of this, indicating greater activity in the IC—the principal midbrain nucleus—which receives efferent feedback from auditory cortical areas. It is therefore possible that levels of cortical engagement could differ across speech manipulations, despite similar overall task performance. To determine the degree of cortical engagement in the active-listening task, we recorded cortical evoked potentials simultaneously to the CEOAEs and ABRs measurements from all 56 participants using a 64-channel, EEG-recording system. Grand averages of the speech-onset responses (event-related potentials, ERPs, Figure 3A and Figure supplement 4) were analysed to test the hypothesis that greater cortical engagement occurred when listening to speech-in-noise compared to vocoded speech despite matching task difficulty. We performed a One-way ANOVA for each of the ERP components generated in response to the most demanding speech manipulations: Voc8, BN5 and SSN3.

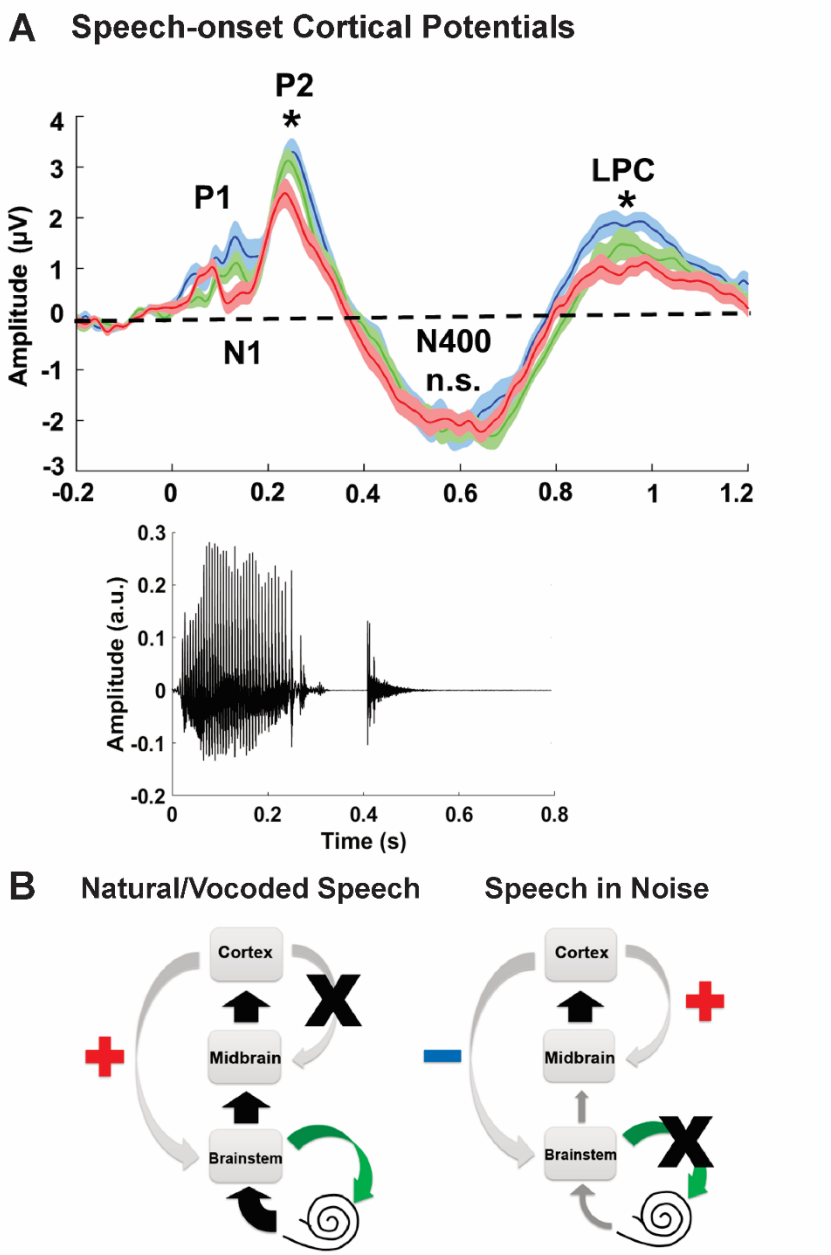

Figure 3. Cortical contribution and proposed mechanisms for active listening to vocoded and masked speech. **A**. ERP component (from electrodes: FZ, F3, F4, CZ, C3, C4, TP7, TP8, T7, T8, PZ, P3, P4) magnitudes and time courses during the active listening of Voc8, BN5 and SSN3. Lower panel shows an example word ("Fuzz") for time-course comparisons with the ERPs. Thick lines and shaded areas represent mean and the standard errors of the mean respectively. Both noise and speech were of equal duration. **B**. Proposed auditory efferent mechanisms for speech processing. Here vocoded speech is processed in a mostly feed-forward manner (as should be the case for natural speech) (*black arrows*, left, B). The activation of the MOCR (*green arrow*) improves the auditory nerve representation of speech-envelope and this information passes up the auditory centres without much need to 'de-noise' the signal (represented as thick black arrows from the cochlea-brainstem-midbrain). Given the trend for increasing cochlear gain inhibition with task difficulty, we include the possibility for increased MOCR drive from higher auditory regions (*red plus sign* and *downward gray arrow*). For babble and speech-shaped noise (right, B), decreasing cochlear gain via the MOCR negatively affects envelope encoding of speech signals (thin forward gray arrows cochlea-brainstem-midbrain). We therefore propose that an efferent loop suppresses the MOCR (*blue minus symbol* and *black cross over green arrow*) leaving greater responsibility for speech signal extraction to the midbrain, cortex and the efferent loop therein (*red plus symbol* and *downward gray arrow*). This ultimately leads to equal behavioural performance across Voc8, BN5 and SSN3.

Early auditory cortical responses (P1 and N1) are largely driven by acoustic features of the stimulus (Getzmann et al., 2015; Grunwald et al., 2003). Here, Voc8 words elicited well-defined P1 and N1 components compared to the speech-in-noise conditions despite words and noises having similar onsets to vocoded tokens. This observation reflects the relatively high precision of vocoded speech's envelope components at stimulus onset compared to those of the masked conditions. Later ERP components, such as P2, N400 and the Late Positivity Complex (LPC), have been linked to speech- and task-specific, top-down (context-dependent) processes (Getzmann et al., 2015; Potts, 2004). Speech masked by babble or speech-shaped noise (BN5 and SSN3) elicited significantly larger P2 components during active listening compared to the Voc8 condition, but not significantly different between themselves ([$F(2,79) = 4.90$, $p = 0.01$], *post hoc* with Bonferroni corrections for multiple comparisons*:* BN5 *vs*. Voc8 ($p = 0.014$); SSN3 *vs.* Voc8 ($p = 0.046$); BN5 *vs.* SSN3 ($p = 1.00$)). Similarly, the magnitude of the LPC—thought to reflect the involvement of several cognitive resources during speech processing such as memorization and understanding (Juottonen et al., 1996) and post-decision closure (Stuss et al., 1992)—differed across conditions: [$F(2,79) = 4.15$, $p = 0.019$]. Specifically, LPCs were greater during active listening to speech-in-babble noise compared to vocoded speech, with LPCs obtained for active listening to speech-in-speech-shaped noise intermediate to both (n.s.). Consistent with the concept that effortful listening varied across speech-in-noise manipulations even when iso-performance was maintained (Francis et al., 2016; Pichora-Fuller et al., 2016), the speech manipulation generating the clearest signature of cortical engagement was speech-in-babble noise, which is considered the most difficult of the masked conditions where iso-performance was achieved at +5 dB SNR for speech-in-babble noise (BN5) compared to only +3 dB SNR for speech-in-speech-shaped noise (SSN3). In contrast to P2 and the LPC, the N400 component of the ERP—which has been associated with the processing of meaning (Kutas and Federmeier, 2011)—did not differ between conditions [$F(2,78) = 0.16$, $p = 0.85$]. This is unsurprising, given that participants were equally able to differentiate non-words in Voc8, BN5 and SSN3 conditions.

Our ERP data were consistent with differential cortical contributions to the processing of vocoded and masked speech, being larger in magnitude for speech manipulations in which the MOCR was less efficiently recruited i.e. babble and speech-shaped noise, and largest overall for the manipulation requiring most listening effort—speech-in-a-multi-talker babble background. In summary, our data across auditory regions revealed categorical differences between the processing of vocoded speech and speech-in-noise, consequently, two different mechanisms or strategies may exist in the auditory system to achieve similar performance levels (Figure 3B).

Discussion

We assessed the role of attention in modulating the contribution of the cochlear efferent system in a lexical task—detecting non-words in a stream of words and non-words. Employing three speech manipulations to modulate task difficulty—vocoding of words, or words masked by multi-talker babble or speech-shaped noise (i.e. noise with the same long-term spectrum as speech)—we find that these manipulations differentially activate the efferent system to modulate cochlear gain. Activation of the cochlear efferent system is also dependent on whether listeners are actively performing the lexical task or are not required to engage whilst watching a silent, stop-motion film. Specifically, with increasing task difficulty (i.e. fewer vocoded channels), noise-vocoding increasingly activates the MOCR in active, compared to passive, listening. The opposite is true for the two masked conditions, where words presented at increasingly lower SNRs more strongly activate the MOCR during passive, compared to active, listening. By adjusting parameters of the three speech manipulations—number of vocoded channels, or SNR— to maintain iso-performance in the task, we find that lesser MOCR engagement is accompanied by greater engagement of cortical resources, to maintain task performance. A computational model of the auditory periphery incorporating efferent feedback demonstrates that improvements in neural representation coding of the amplitude envelope of vocoded speech and speech-in-noise provides a rationale for suppressing and maintaining the cochlear gain, respectively. Our data suggest that a network of brainstem and higher brain circuits is involved in maintaining performance in an active listening task, and that different aspects of this network, including reflexive circuits in the lower brainstem, and the relative allocation of attentional resources, are invoked differently depending on specific features of the listening environment. Our data also suggest that neural, and other physiological, markers of listening effort may not adequately represent the full range of brain processes required to maintain identical levels of performance in a listening task.

Increases in task difficulty have been linked to the allocation of auditory attention and cognitive resources towards the task itself (Kahneman, 1973). Previous efforts to explore how auditory attention might affect the relationship between cochlear gain and performance in speech-in-noise tasks demonstrated increased activation of the MOCR in active, compared to passive, listening, but were unable to determine whether, or how, cochlear gain—as determined by CEOAE inhibition—was modulated with task difficulty (Kalaiah et al., 2017a; Smith and Cone, 2015). Here, employing a similar experimental design to these studies, we achieved, three levels of task difficulty for each speech manipulation. Importantly, rather than using spectrally-flat (white) noise we employed multi-talker babble noise and speech-shaped noise, both of which are more spectrally similar to speech and are therefore more effective (and ethologically relevant) maskers. We also presented vocoded words to assess intelligibility of degraded speech in the absence of additional masking noise. From this, our data comparing active and passive listening conditions highlight a categorical distinction between speech manipulations, a difference between processing a single, but degraded, auditory stream (vocoded speech) and parsing a complex acoustic scene to hear out a stream from multiple competing and spectrally-similarly, sounds (multi-talker babble and speech-shaped noise).

Task difficulty appeared to modulate the cochlear gain in a stimulus-specific manner. For instance, clean speech—the most easily discernible stimulus—was capable of suppressing CEOAEs, demonstrating that contralateral MOCR activation is not limited to tones and broadband noise (Kawase et al., 1993; Marian et al., 2018). The reduction in cochlear gain with increasing task difficulty for vocoded speech and, conversely, the preservation of cochlear gain for speech-in-noise, suggests that attentional resources might gate the MOCR differently depending on speech manipulation type. In contrast to active listening, when participants were asked to ignore the auditory stimuli and direct attention to the film, the MOCR was gated in a direction consistent with the auditory system suppressing irrelevant auditory information whilst (presumably) attending to visual streams (Puel et al., 1988; Wittekindt et al., 2014). Interestingly, MOCR activation did not depend on whether participants attended or not to natural speech. This may be explained by: 1) natural speech being particularly salient as an undegraded, ethologically-relevant stimulus and 2) the low attentional load of passively watching a film resulting in the continued perception of unattended speech (Lavie, 2005).

To explain the varying reduction in cochlear gain between vocoded and masked stimuli across active and passive conditions, both top-down and bottom-up mechanisms are candidates for

modifying MOCR activation. Bottom-up mechanisms such as the increased reflexive activation of the MOCR by wideband stimuli with flat power spectra (Lilaonitkul and Guinan, 2009) may explain a facet of our results—especially during passive listening. For instance, reduced MOCR activation observed for babble noise during passive listening may arise from it being a weaker suppressor than 'stationary' noises such as white noise or speech-shaped noise (Boothalingam et al., 2014; Kalaiah et al., 2017b). However, vocoded stimuli with fewer channels should have also activated the MOCR more effectively due to their more 'noise-like' spectra. Therefore, the absence of any CEOAE inhibition in the passive vocoded conditions as well as the pattern of MOCR activation across all active manipulations, suggests that a perceptual, top-down categorization of the stimuli is necessary to appropriately engage or disengage the MOCR.

The modelling data accounted for task-dependent MOCR activation through enhancement of speech-envelope encoding in the auditory nerve. The ability of the MOCR to 'unmask' acoustic transient signals in the presence of background noise has long hinted at a potential influence on speech processing (Guinan and Gifford, 1988; Kawase et al., 1993; Wiederhold and Kiang, 1970). By suppressing cochlear gain, the MOCR increases the dynamic range of the mechanical (Cooper and Guinan, 2006; Murugasu and Russell, 1996) and, neural (Guinan and Gifford, 1988; Winslow and Sachs, 1987) responses of the cochleae to amplitude-modulated components in speech such as syllables and phonemes, preferentially facilitating their representation over that of masking noises. The apparent benefit of suppressing cochlear gain to neural coding of vocoded words' envelopes compared to a disbenefit for words masked by background noise is consistent with vocoded words retaining relatively strong envelope modulations, and that these modulations are extracted effectively through expansion of the dynamic range as the cochlear gain is reduced. For both noise-maskers, however, any improvement to envelope coding due to dynamic range expansion applies to both speech and masker as they overlap spectrotemporally, resulting in poorer representation of the speech envelope. This does not appear dependent on whether masking noises saturate AN firing as both high and low SR AN fibers demonstrated deterioration in envelope encoding when efferent feedback was introduced to both speech-in-noise conditions (Figure 2D and Figure supplement 3).

Beyond the auditory periphery, the impact of modulating the MOCR was observed in the auditory brainstem, midbrain, and cortex. Evidence for reduced auditory periphery output to the brainstem has been demonstrated as decreased AN responses to click stimuli after electrical stimulation of MOC fibers (Galambos, 1956; Gifford and Guinan, 1987). The smaller magnitude of ABR wave III—generated by the cochlear nucleus— in the active vocoded, and passive masked, conditions is consistent with this interpretation. The reduced magnitude of wave V (midbrain) for passive masked and lower brainstem-to-midbrain gain for vocoded conditions are consistent with reductions in ABR magnitude observed during passive or unattended listening for both speech (Forte et al., 2017) and click (Brix, 1984; Lukas, 1980) stimuli. This highlights the potential for irrelevant and distracting auditory information (e.g. when watching the film) to be processed less efficiently at brainstem and midbrain levels. How this is achieved during passive listening differs for the two speech manipulation categories: reduced ABR magnitudes appear inherited from reduction of the cochlear gain for masked stimuli whereas transmission of information from brainstem to midbrain is reduced for vocoded stimuli.

At the cortical level, recorded potentials were larger for attended speech, consistent with previous reports (Fujiwara et al., 1998; Hillyard et al., 1973), suggesting greater cortical engagement in the lexical task. If a stimulus-dependent categorization is made at the level of the inner ear, brainstem and midbrain, then the auditory cortex should at least reflect this differentiation. Evidence for systemic differences in processing single streams of degraded, compared to masked speech, has been reported in the autonomic nervous system (Francis et al., 2016): where, despite maintaining similar task difficult across conditions, masked speech elicits stronger physiological reactions than single streams. Later cortical components displayed higher magnitudes for speech-in-noise than vocoded speech during the hardest manipulations. In this context, late cortical components have been associated to complex stimulus evaluation and classification (Potts, 2004) as well as mental allocation during a task (Getzmann et al., 2015). Cortical activity recorded during the lexical distinction between words and non-words likely reflects greater reliance on (or at least increased contribution from) top-down or context-dependent processes for speech masked by noise than for vocoded speech.

An overview across the auditory pathway—from the auditory periphery to the cortex—suggests that two different strategies co-exist to achieve similar levels of performance when listening to single undegraded/degraded streams compared to masked speech (Figure 3B). The first involves enhanced sensitivity to energy fluctuations through recruitment of the MOCR, followed by a gain adjustment in the afferent pathways, to generate a central representation of the stimulus sufficient and necessary for speech intelligibility (Figure 3B, left panel). The second, activated when processing speech-in-noise (Figure 3B, right panel), preserves cochlear gain to prevent deterioration of envelope encoding placing the onus of 'denoising' on midbrain and cortex, including loops between them, to maximise speech understanding.

Although normal hearing listeners appear to benefit from an MOCR that modulates cochlear gain and is amenable to top-down, attentional control, it is important to note that users of cochlear implants (CIs)—for whom normal hearing listeners listening to vocoded speech have often been a proxy in experimental studies—have no access to the MOCR. CIs bypass the mechanical processes of the inner ear, including the outer hair cells—which are non-functioning or absent completely in individuals with severe-to-profound hearing loss—to stimulate directly with electrical pulses the auditory nerve fibers themselves. Efforts have been made to incorporate MOC-like properties into CI processes, at least to provide for expanded auditory spatial cues to improve listening in bilaterally implanted CI users (Lopez-Poveda et al., 2016), but the capacity to exploit efferent feedback to aid speech understanding in CI listeners is yet to be exploited in any device.

For other hearing-impaired listeners, aided or unaided, the contribution of MOC feedback to speech processing is limited compared to normal hearing listeners as, in most cases, their hearing loss comes from damage to the OHCs, which receive direct synaptic input from MOC fibers. In hearing loss generally, the degradation or loss of peripheral mechanisms contributing to effective speech processing in complex listening environments may mean that listeners rely more heavily on attentional and other cortical-mediated processes, contributing to widely reported increases in listening effort required to achieve adequate levels of listening performance (Pichora-Fuller et al., 2016). This increase in listening effort—likely manifesting over time—may not be reflected in performance in relatively short, laboratory- or clinic-based assessments of hearing function.

Materials and Methods

This study was approved by the Human Research Ethics Committee of Macquarie University (ref: 5201500235). Each participant signed a written informed consent form and was provided with a small financial remuneration for their time.

Hearing assessment: All 66 participants (36 females, aged between 18-35 (mean: 24 ± 7 years old)) included in this study had normal pure tone thresholds (< 20 dB HL); normal middle ear function (standard 226 Hz tympanometry; normal OHCs function (assessed with Distortion Product OAEs (DPOAEs) between 0.5-10 kHz. Steady-state contralateral and ipsilateral broadband noise middle ear muscle reflexes (MEMR) were assessed and all participants had thresholds > 75 dB HL.

Experimental Protocol: Participants were seated comfortably inside a sound-proof booth (ISO 8253-1:2010) while wearing an EEG cap (Neuroscan 64 channels, SynAmps2 amplifier, Compumedics Limited). Two listening conditions (passive and active) were counterbalanced across participants. In the passive listening condition, subjects were asked to ignore the auditory stimuli and to watch a non-subtitled, stop-motion movie. To ensure participants' attention during this condition, they were monitored with a video camera and were asked questions at the end of this session (e.g. What happened in the movie? How many characters were present?). During active listening, participants performed an auditory lexical decision task, where they were asked to press a keyboard's space key each time they heard a non-word in strings of 300 speech tokens. d' was used as a measure of accuracy and calculated as: Z(correct responses) – Z(false alarm). Simultaneous to the presentation of word/non-word in one ear, CEOAEs were recorded continuously in the contralateral ear (Figure 1A). The speech tokens were presented at 75 dB SPL and the click stimulus at 75 dB p-p, therefore no MEMR contribution was expected given a minimum of 10 dB difference between MEMR thresholds and stimulus levels (ANSI S3.6-1996 standards for the conversion of dB SPL to dB HL). The ear receiving either the clicks or speech stimuli was randomized across participants.

Speech Stimuli: 423-word items were acquired from Australian-English-adapted versions of monosyllabic consonant-nucleus-consonant (CNC) word lists (Lehiste and Peterson, 1959) and were spoken by a female, native Australian-English speaker. The duration of words ranged between 420-650 ms. 328 monosyllabic CNC non-word tokens were selected from the Australian Research Council non-word database (Rastle et al., 2002). Speech stimuli were delivered using ER-3C insert earphones (Etymotic Research, Elk Grove Village, IL) and Presentation® software (Neurobehavioral Systems, Inc., Berkeley, CA, version 18.1.03.31.15) at 44.1 kHz, 16-bits. Blocks of 200 words and 100 non-words (randomly selected from the speech corpus) were presented for each of the speech manipulations (see details below). Speech tokens were counterbalanced in each block based on the presence of stop and non-stop initial consonants: 100 stop/ non-stop consonant words; 50 stop/non-stop consonants words with a maximum of 3 repeats per participant allowed. Each block had a duration of 12 minutes, (Figure 1A) and participants could take short breaks between them if needed. The order of blocks was randomized to prevent presentation order bias or training effects (Ritzer, 2007).

Noise Vocoded Speech. Twenty-seven native speakers of Australian-English (17 females: 25 right-handed, 2 left-handed) were recruited (aged between 18-35 (mean: 23 ± 5 years old). Based on the noise vocoding method and behavioural results of Shannon et al. (Shannon et al., 1995), three noise vocoded conditions (16-, 12- and 8-channels: Voc16, Voc12, and Voc8, respectively) were tested to represent 3 degrees of speech intelligibility (i.e. task difficulty). Four blocks were assessed in both active and passive listening conditions: Block 1: natural or clean speech; Block 2: Voc16; Block 3: Voc12 and Block 4: Voc8, for a total duration of 2.6 hours (including hearing assessment and EEG cap set-up).

Speech in Babble Noise. Twenty-nine native speakers of Australian-English (19 females: 28 right-handed, 1 left-handed) were recruited, aged between 20-35 (mean: 26 ± 9 years old). The BN used consisted of four females and four male talkers (Keidser et al., 2002) and was filtered to match the

long-term average spectrum of the speech corpus. Random segments from a 2-min BN recording were selected as noise maskers. Both speech tokens and the BN segments were matched in duration. Three blocks were presented in the active and passive listening conditions: Block1: natural or clean speech; Block 2: BN at +10 dB SNR (BN10) and Block 3: BN at +5 dB SNR (BN5).

Speech in Speech-Shaped noise. SSN was generated to match the long-term average spectrum long-term average of the speech corpus. Segments from a 2 min SSN were randomly selected as noise maskers. Both BN and SSN manipulations were presented in the same session, therefore, Block1 was the same for both manipulations, Block 2: SSN at +8 dB SNR (SSN8) and Block 3: SSN at +3 dB SNR (SSN3). A total of 10 blocks (5 blocks in the active and 5 blocks in the passive condition) of 12 min duration were presented to the 29 subjects who participated in BN and SSN experiments. All blocks were presented in a unique session that lasted 3 hours.

CEOAEs acquisition and analysis. Non-filtered click stimuli, with a positive polarity and 83 µsec of duration were digitally generated using RecordAppX (Advanced Medical Diagnostic Systems) software. The presentation rate was 32 Hz in all conditions. Both the generation of clicks and OAE recordings were controlled via an RME UCX soundcard (RME, Haimhausen, Germany), and delivered/collected to and from the ear canal through an Etymotic ER-10B probe connected to ER-2 insert earphones with the microphone pre-amplifier gain set at 20 dB. Calibration of clicks was performed using a sound level meter (B&K G4) and microphone IEC 60711 Ear Simulator RA 0045 (BS EN 60645-3:2007). This setup was also used to calibrate the speech stimuli. In addition, clicks were in-ear calibrated using forward equivalent pressure level (FPL) ensuring accurate stimulus levels (Souza et al., 2014). The OAE's probe was re-positioned, re-calibrated and the block re-started if participants moved or touched it.

CEOAE data were analysed off-line using custom Matlab scripts (available upon request) (Boothalingam and Purcell, 2015) (Figure 1A). The averaged RMS magnitudes of CEOAE signals were analysed between 1-2 kHz given maximal MOC effects in this frequency band (Hood et al., 1996). Only binned data for averaged CEOAEs displaying a SNR ≥ 6 dB and with > 80% of epochs retained (i.e. had RMS levels within the two standard deviations limit), were selected as valid signals for further analysis. The initial and final 60 secs of CEOAE recordings of each block, in the absence of speech tokens (Figure 1A), were used as baseline. As no significant differences were observed between CEOAE baseline magnitudes within participants ($p > 0.05$) across experimental conditions, all baselines were pooled within participants. This allowed for an increase in both SNR and reliability of the CEOAEs individually. After baseline recording, CEOAEs were continuously obtained for 10 minutes during the presentation of the speech tokens. The inhibition of CEOAE magnitude relative to the baseline was calculated as follows and reported as means and standard errors of the mean:

$$\text{CEOAE inhibition} = \text{CEOAE}_{\text{speech presentation (average across minutes)}} - \text{CEOAE}_{\text{baseline (first 60 sec)}}$$

EEG acquisition and analysis. EEG measurements and the CEOAE setup were synchronized using a Stimtracker (https://cedrus.com/). EEG data were acquired according to the 10–20 system. Impedance levels were kept below 5 kΩ for all electrodes. Signals were sampled at a rate of 20 kHz in the AC mode with a gain of 2010 and an accuracy of 0.15 nV/LSB. Early and late ERP components were analysed offline using Fieldtrip-based scripts (Oostenveld et al., 2011). Data were re-referenced to the average of mastoid electrodes. Trials started 200 ms before and ended 1.2 sec after speech onset. Components visually identified as eye blinks and horizontal eye movement were excluded from the data as well as trials with amplitude >75 µV. The accepted trials (60-80% per condition) were band-pass filtered between 0.5 to 30 Hz with transition band roll-off of 12 dB/octave. Trials were baseline-corrected using the mean amplitude between -200-0 ms from speech onset. Baseline-corrected trials were averaged to obtain ERP waveforms. Analysis windows centred on the grand average-ERP component maximums were selected (Woodman, 2010): P1 (100-110 msec) and N1(145-155 msec); P2 (235-265 msec), N400 (575-605 msec) and LPP (945-975 msec). Mean amplitude for each component within the analysis window was calculated for each participant and experimental condition.

EEG. Auditory Brainstem Responses (ABRs). ABR signals were extracted from central electrodes (FZ, FCZ, CZ). 15 msec-duration ABR analysis windows (5 ms prior and 10 ms after click onset) were selected. A total of 19200 trials (click rate of 32 Hz across 10 minutes per block) were band-pass filtered between 200-3000 Hz. Averaged ABR waveforms were obtained using a weighted-averaging method (Don and Elberling, 1994). Amplitude of waves III and V (Figure 1A) were visually determined for each subject across blocks and conditions when appropriate (wave amplitudes above the residual noise, therefore a positive SNR). Due to stimulus level restrictions (<= 75 p-p dB SPL to avoid MEMR activation), wave I could not be extracted from the EEG residual noise.

Statistical Analysis. Variables were tested for normality (Shapiro-Wilk test), residuals outliers preventing normal distribution of the data in each experimental condition were removed from the final data set. ANOVA and t-tests (alpha=0.05, with Bonferroni corrections for multiple comparisons) were performed to assess differences between manipulations. Sample size estimation was computed according to the statistical test employed by using G*Power (Faul et al., 2007) (Effect size f = 0.4; α err prob = 0.05; Power (1-β err prob) = 0.8).

Auditory Nerve and Brainstem Simulations. The Matlab Auditory Periphery and Brainstem (MAP-BS) computational model was used to simulate auditory nerve (AN) responses with and without an efferent feedback loop at seven frequencies (1000 Hz, 1122 Hz, 1260 Hz, 1414 Hz, 1587 Hz, 1782 Hz, 2000 Hz). A binaural brainstem circuit (Figure supplement 1A) using Hodgkin-Huxley models was implemented to calculate within-channel efferent attenuation of the cochlear gain with biophysically realistic temporal dynamics (Meddis, 2016) (Figure supplement 1B). The current balance equation for all single-compartment, Hodgkin-Huxley models was described as follows:

$$\begin{aligned} C_m \frac{dV}{dt} &= -I_{Na} - I_{KLT} - I_{KHT} - I_H - I_L - I_{syn}(t) \\ &= -\bar{g}_{Na}\, m^3 h(V - E_{Na}) - \bar{g}_{KLT}\, w(V - E_K) - \bar{g}_{KHT} n(V - E_K) - \bar{g}_H r(V - E_H) - \bar{g}_L(V - E_L) \\ &\qquad - A\bar{g}_{syn}(t)(V - E_{syn}) \end{aligned}$$

Where $\bar{g}_x$ is the maximum channel conductance and E_x is the corresponding reversal potential for conductance *x* where *x* was either: *Na* (fast sodium), *KLT* (low threshold potassium), *KHT* (high threshold potassium), *H* (hyperpolarization-activated mixed-cation), *L* (leak) and *K* (all potassium conductances) (see Table supplement 1 for relative values in cell types). $\bar{g}_{syn}(t)$ is the unitary synaptic input at time, *t*, and *A* is an input activity-dependent factor with a predetermined, maximum value for each cell type (*Max Synaptic Current*, Table supplement 1). E_{syn} represents the synaptic reversal potential which was either 0mV for excitation or -70mV for inhibition (Figure supplement 1A and Table supplement 1). An absolute voltage threshold of -20mV was used for suprathreshold responses. The dynamics of gating variables (*m*, *h*, *w*, *n* and *r*) were described by the following differential equation:

$$\frac{dx}{dt} = \frac{x_\infty(V) - x}{\tau_x(V)}$$

Where $\tau x(V)$ is the time constant of the variable *x* and $x\infty(V)$ is the steady state value of *x* at voltage, *V*. See 'HH5lookupInt.m' file in MAP-BS (Meddis, 2016) for values of $\tau x(V)$ and $x\infty(V)$ for individual conductances. The spiking output of MOC neurons was used to calculate inhibition of the basilar membrane displacement's nonlinear path (Clark et al., 2012) as follows:

$$ATT(t) = [1 + MOC_1(t) + MOC_2(t)]^{-1}$$

Where the formula for calculating activation and exponential decay of MOC_x at time, t, was expressed as:

$$MOC_x(t) = \left[\left(MOC_x(t-1) \times exp^{-dt/\tau_x}\right) + \left(Fw_x \times Spikes_{l/r}(t-td)\right)\right]$$

Where τ_x is the decay time constant (55ms for MOC_1 and 1000ms for MOC_2 (Backus and Guinan, 2006) and *Spikes(t-delay)* is spikes observed across 'left' and 'right' MOC populations at a time, *t-td* (where *td* equals a 10ms lag associated with synaptic latencies from MOC fibers to OHCs (Liberman, 1988). *F* is a temperature- and sample rate-dependent factor attenuating *Spikes(t-delay)* appropriately (0.02 here for a default temperature of 38 C and a sample rate of 100 kHz). w_x is a weighting factor determining the relative contribution of MOC_1 and MOC_2 to attenuation (0.9 and 0.1 for MOC_1 and MOC_2 respectively).

Word Presentation. 300 words (150 stop/non-stop consonant words), were chosen at random from the speech corpus and were degraded using the most demanding speech manipulations (Voc8, BN5 and SSN3). Normal, Test+ and polarity-inverted, Test-, versions of each manipulation were presented to the MAP-BS model at 75 dB SPL both with and without efferent feedback. Natural words (both normal, Nat+, and polarity-inverted, Nat-) were also presented to the MAP-BS model with efferent feedback enabled as a reference AN output (Figure supplement 2).

Shuffled Auto-Correlogram Analysis. Comparative analysis of AN coding of amplitude-modulated (AM) envelope between Voc8/BN5/SSN3 conditions and the reference natural condition was performed using shuffled auto- and cross- correlograms (SACs and SCCs respectively) (Joris et al., 2006) Normalized all-order histograms were calculated using the spike trains of 400 low SR, AN fibers with a coincidence window of 50µs and a delay window ± 25 ms centred on zero (Heinz and Swaminathan, 2009). No correction for triangular shape was required given brevity of delay window relative to stimulus length (between 420-650 ms) (Joris et al., 2006). A neural cross-correlation coefficient, *ρENV*, quantifying AM envelope encoding similarity between conditions was generated as follows (Liberman, 1988):

$$\rho ENV = \frac{(sumcor_{Test\ Nat} - 1)}{\sqrt{(sumcor_{Test} - 1) \times (sumcor_{Nat} - 1)}}$$

Where $sumcor_{Nat}$ (natural word reference) and $sumcor_{Test}$ (Voc/BN/SSN conditions) are the averages of SACs (Normalized all-order histograms for Nat+ *vs.* Nat+/Test+ *vs.* Test+ for $sumcor_{Nat}$ /$sumcor_{Test}$ respectively) and cross-polarity histograms (Normalized all-order histograms for Nat+ *vs.* Nat-/Test+*vs.*Test- for $sumcor_{Nat}$ /$sumcor_{Test}$ respectively). $Sumcor_{Test\ Nat}$ is the average of the SCC (Average of normalized all-order histograms for Nat+ *vs.* Test+ and Nat- *vs.* Test-) and the cross-polarity correlogram (Average of normalized all-order histograms for Nat- *vs.* Test+ and Nat+ *vs*. Test-) between natural and Voc8/BN5/SSN3 conditions. All high-frequency oscillations (> test-frequency), associated with fine-structure leakage, were removed from sumcors (Heinz and Swaminathan, 2009). *ρENV* values ranged from 0 to 1 where 0 represents completely dissimilar spike trains and 1 represents identical spike patterns (Heinz and Swaminathan, 2009).

Analysis of modelling and statistics. Percentage changes in *ρENV* due to efferent feedback inclusion in MAP_BS were calculated for each test frequency and Voc8/BN5/SSN3 condition as follows:

$$\Delta\rho\mathrm{ENV}_{freq} = \left(\frac{\rho ENV_{eff} - \rho ENV_{no\ eff}}{\rho ENV_{no\ eff}}\right) \times 100$$

Where $\Delta\rho ENV_{freq}$ is the percentage change in *ρENV* at a test-frequency for a manipulated word. $\rho ENV_{no\ eff}$ and ρENV_{eff} are measures of *ρENV* with and without efferent feedback enabled respectively. An average $\Delta\rho ENV_{freq}$ was calculated across test-frequencies for each word and manipulation. A one-sample t-test was then performed to confirm whether average $\Delta\rho ENV_{freq}$ for all words differed from zero for each speech manipulation. Data are reported as means and standard errors of the mean.

Acknowledgements

H.H.P. was supported in this study by an International Macquarie University Excellence Scholarship and the Hearing CRC. The authors thank Prof. David Poeppel for his contributions during experimental design. We thank Prof. David Ryugo for his comments on the manuscript. In addition, we thank Ronny Ibrahim, Jaime Undurraga and Greg Stewart for their technical support regarding equipment and EEG analysis; Andrew Brughera for his help with figures and Nicholas Clark for his assistance with MAP-BS.

Competing interests

The author has no financial interests to declare. The other authors declare that no competing interests exist.

References

Ahveninen J, Jääskeläinen IP, Raij T, Bonmassar G, Devore S, Hämäläinen M, Levänen S, Lin F-H, Sams M, Shinn-Cunningham BG, Witzel T, Belliveau JW. 2006. Task-modulated "what" and "where" pathways in human auditory cortex. *Proc Natl Acad Sci U S A* **103**:14608–14613. doi:10.1073/pnas.0510480103

Almishaal A, Jennings SG. 2016. Effects of a precursor on amplitude modulation detection are consistent with efferent feedback. *J Acoust Soc Am* **139**:2155–2155. doi:10.1121/1.4950377

Ashmore J, Avan P, Brownell WE, Dallos P, Dierkes K, Fettiplace R, Grosh K, Hackney CM, Hudspeth AJ, Jülicher F, Lindner B, Martin P, Meaud J, Petit C, Santos-Sacchi J, Sacchi JRS, Canlon B. 2010. The remarkable cochlear amplifier. *Hear Res* **266**:1–17. doi:10.1016/j.heares.2010.05.001

Aubanel, V., Davis, C. Kim, J. 2017. The MAVA corpus. doi:10.4227/139/59a4c21a896a3

Backus BC, Guinan JJ Jr. 2006. Time-course of the human medial olivocochlear reflex. *J Acoust Soc Am* **119**:2889–2904. doi:10.1121/1.2169918

Boothalingam S, Purcell D, Scollie S. 2014. Influence of 100Hz amplitude modulation on the human medial olivocochlear reflex. *Neurosci Lett* **580**:56–61. doi:10.1016/j.neulet.2014.07.048

Boothalingam S, Purcell DW. 2015. Influence of the stimulus presentation rate on medial olivocochlear system assays. *J Acoust Soc Am* **137**:724–732. doi:10.1121/1.4906250

Brix R. 1984. The influence of attention on the auditory brain stem evoked responses. Preliminary report. *Acta Otolaryngol* **98**:89–92. doi:10.3109/00016488409107538

Clark NR, Brown GJ, Jürgens T, Meddis R. 2012. A frequency-selective feedback model of auditory efferent suppression and its implications for the recognition of speech in noise. *J Acoust Soc Am* **132**:1535–1541. doi:10.1121/1.4742745

Collet L, Kemp DT, Veuillet E, Duclaux R, Moulin A, Morgon A. 1990. Effect of contralateral auditory stimuli on active cochlear micro-mechanical properties in human subjects. *Hear Res* **43**:251–261. doi:10.1016/0378-5955(90)90232-e

Cooper NP, Guinan JJ Jr. 2006. Efferent-mediated control of basilar membrane motion. *J Physiol* **576**:49–54. doi:10.1113/jphysiol.2006.114991

Costalupes JA, Young ED, Gibson DJ. 1984. Effects of continuous noise backgrounds on rate response of auditory nerve fibers in cat. *J Neurophysiol* **51**:1326–1344. doi:10.1152/jn.1984.51.6.1326

de Boer J, Thornton ARD, Krumbholz K. 2012. What is the role of the medial olivocochlear system in speech-in-noise processing? *J Neurophysiol* **107**:1301–1312. doi:10.1152/jn.00222.2011

Don M, Elberling C. 1994. Evaluating residual background noise in human auditory brain-stem responses. *J Acoust Soc Am* **96**:2746–2757. doi:10.1121/1.411281

Dragicevic CD, Aedo C, León A, Bowen M, Jara N, Terreros G, Robles L, Delano PH. 2015. The olivocochlear reflex strength and cochlear sensitivity are independently modulated by auditory cortex microstimulation. *J Assoc Res Otolaryngol* **16**:223–240. doi:10.1007/s10162-015-0509-9

Faul F, Erdfelder E, Lang A-G, Buchner A. 2007. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods* **39**:175–191. doi:10.3758/bf03193146

Faye-Lund H. 1986. Projection from the inferior colliculus to the superior olivary complex in the albino rat. *Anat Embryol* **175**:35–52. doi:10.1007/bf00315454

Forte AE, Etard O, Reichenbach T. 2017. The human auditory brainstem response to running speech reveals a subcortical mechanism for selective attention. *Elife* **6**:e27203. doi:10.7554/eLife.27203

Francis AL, MacPherson MK, Chandrasekaran B, Alvar AM. 2016. Autonomic Nervous System Responses During Perception of Masked Speech may Reflect Constructs other than Subjective Listening Effort. *Front Psychol* **7**:263. doi:10.3389/fpsyg.2016.00263

Fujiwara N, Nagamine T, Imai M, Tanaka T, Shibasaki H. 1998. Role of the primary auditory cortex in auditory selective attention studied by whole-head neuromagnetometer. *Brain Res Cogn Brain Res* **7**:99–109. doi:10.1016/s0926-6410(98)00014-7

Galambos R. 1956. Suppression of auditory nerve activity by stimulation of efferent fibers to cochlea. *J Neurophysiol* **19**:424–437. doi:10.1152/jn.1956.19.5.424

Getzmann S, Falkenstein M, Wascher E. 2015. ERP correlates of auditory goal-directed behavior of younger and older adults in a dynamic speech perception task. *Behav Brain Res* **278**:435–445. doi:10.1016/j.bbr.2014.10.026

Gifford ML, Guinan JJ Jr. 1987. Effects of electrical stimulation of medial olivocochlear neurons on ipsilateral and contralateral cochlear responses. *Hear Res* **29**:179–194. doi:10.1016/0378-5955(87)90166-3

Giraud AL, Garnier S, Micheyl C, Lina G, Chays A, Chéry-Croze S. 1997. Auditory efferents involved in speech-in-noise intelligibility. *Neuroreport* **8**:1779–1783. doi:10.1097/00001756-199705060-00042

Grunwald T, Boutros NN, Pezer N, von Oertzen J, Fernández G, Schaller C, Elger CE. 2003. Neuronal substrates of sensory gating within the human brain. *Biol Psychiatry* **53**:511–519. doi:10.1016/s0006-3223(02)01673-6

Guinan JJ Jr. 2018. Olivocochlear efferents: Their action, effects, measurement and uses, and the impact of the new conception of cochlear mechanical responses. *Hear Res* **362**:38–47. doi:10.1016/j.heares.2017.12.012

Guinan JJ Jr, Gifford ML. 1988. Effects of electrical stimulation of efferent olivocochlear neurons on cat auditory-nerve fibers. I. Rate-level functions. *Hear Res* **33**:97–113. doi:10.1016/0378-5955(88)90023-8

Harkrider AW, Bowers CD. 2009. Evidence for a cortically mediated release from inhibition in the human cochlea. *J Am Acad Audiol* **20**:208–215. doi:10.3766/jaaa.20.3.7

Heinz MG, Swaminathan J. 2009. Quantifying envelope and fine-structure coding in auditory nerve responses to chimaeric speech. *J Assoc Res Otolaryngol* **10**:407–423. doi:10.1007/s10162-009-0169-8

Hillyard SA, Hink RF, Schwent VL, Picton TW. 1973. Electrical signs of selective attention in the human brain. *Science* **182**:177–180. doi:10.1126/science.182.4108.177

Hood LJ, Berlin CI, Hurley A, Cecola RP, Bell B. 1996. Contralateral suppression of transient-evoked otoacoustic emissions in humans: intensity effects. *Hear Res* **101**:113–118. doi:10.1016/s0378-5955(96)00138-4

Joris PX, Louage DH, Cardoen L, van der Heijden M. 2006. Correlation index: a new metric to quantify temporal coding. *Hear Res* **216-217**:19–30. doi:10.1016/j.heares.2006.03.010

Juottonen K, Revonsuo A, Lang H. 1996. Dissimilar age influences on two ERP waveforms (LPC and N400) reflecting semantic context effect. *Brain Res Cogn Brain Res* **4**:99–107. doi:10.1016/0926-6410(96)00022-5

Kahneman D. 1973. Attention and effort. Prentice-Hall.

Kalaiah MK, Nanchirakal JF, Kharmawphlang L, Noronah SC. 2017a. Contralateral suppression of transient evoked otoacoustic emissions for various noise signals. *Hearing, Balance and Communication* **15**:84–90. doi:10.1080/21695717.2017.1311504

Kalaiah MK, Theruvan NB, Kumar K, Bhat JS. 2017b. Role of Active Listening and Listening Effort on Contralateral Suppression of Transient Evoked Otoacousic Emissions. *J Audiol Otol* **21**:1–8. doi:10.7874/jao.2017.21.1.1

Kawase T, Delgutte B, Liberman MC. 1993. Antimasking effects of the olivocochlear reflex. II. Enhancement of auditory-nerve response to masked tones. *J Neurophysiol* **70**:2533–2549. doi:10.1152/jn.1993.70.6.2533

Keidser G, Ching T, Dillon H, Agung K, Brew C, Brewer S, Fisher M, Foster L, Grant F, Storey L. 2002. The National Acoustic Laboratories' (NAL) CDs of Speech and Noise for Hearing Aid Evaluation: Normative Data and Potential Applications. *Australian and New Zealand Journal of Audiology* **24**:16–35. doi:10.1375/audi.24.1.16.31112

Kemp DT. 1978. Stimulated acoustic emissions from within the human auditory system. *J Acoust Soc Am* **64**:1386–1391. doi:10.1121/1.382104

Knudsen EI. 2007. Fundamental components of attention. *Annu Rev Neurosci* **30**:57–78. doi:10.1146/annurev.neuro.30.051606.094256

Kujawa SG, Liberman MC. 2015. Synaptopathy in the noise-exposed and aging cochlea: Primary neural degeneration in acquired sensorineural hearing loss. *Hear Res* **330**:191–199. doi:10.1016/j.heares.2015.02.009

Kutas M, Federmeier KD. 2011. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). *Annu Rev Psychol* **62**:621–647. doi:10.1146/annurev.psych.093008.131123

Lavie N. 2005. Distracted and confused?: selective attention under load. *Trends Cogn Sci* **9**:75–82. doi:10.1016/j.tics.2004.12.004

Lehiste I, Peterson GE. 1959. Linguistic Considerations in the Study of Speech Intelligibility. *J Acoust Soc Am* **31**:280–286. doi:10.1121/1.1907713

Liberman MC. 1988. Response properties of cochlear efferent neurons: monaural vs. binaural stimulation and the effects of noise. *J Neurophysiol* **60**:1779–1798. doi:10.1152/jn.1988.60.5.1779

Lilaonitkul W, Guinan JJ Jr. 2009. Human medial olivocochlear reflex: effects as functions of contralateral, ipsilateral, and bilateral elicitor bandwidths. *J Assoc Res Otolaryngol* **10**:459–470. doi:10.1007/s10162-009-0163-1

Lopez-Poveda EA, Eustaquio-Martín A, Stohl JS, Wolford RD, Schatzer R, Wilson BS. 2016. A Binaural Cochlear Implant Sound Coding Strategy Inspired by the Contralateral Medial Olivocochlear Reflex. *Ear Hear* **37**:e138–48. doi:10.1097/AUD.0000000000000273

Lukas JH. 1980. Human auditory attention: the olivocochlear bundle may function as a peripheral filter. *Psychophysiology* **17**:444–452. doi:10.1111/j.1469-8986.1980.tb00181.x

Marian V, Lam TQ, Hayakawa S, Dhar S. 2018. Spontaneous Otoacoustic Emissions Reveal an Efficient Auditory Efferent Network. *J Speech Lang Hear Res* **61**:2827–2832. doi:10.1044/2018_JSLHR-H-18-0025

Meddis R. 2016. MAP-BS� a Matlab Auditory Processing software platform for studying Auditory BrainStem activity. doi:10.13140/RG.2.2.30627.45603

Mertes IB, Johnson KM, Dinger ZA. 2019. Olivocochlear efferent contributions to speech-in-noise recognition across signal-to-noise ratios. *J Acoust Soc Am* **145**:1529. doi:10.1121/1.5094766

Mishra SK, Lutman ME. 2014. Top-down influences of the medial olivocochlear efferent system in speech perception in noise. *PLoS One* **9**:e85756. doi:10.1371/journal.pone.0085756

Møller AR, Burgess J. 1986. Neural generators of the brain-stem auditory evoked potentials (BAEPs) in the rhesus monkey. *Electroencephalogr Clin Neurophysiol* **65**:361–372. doi:10.1016/0168-5597(86)90015-8

Mulders WH, Robertson D. 2000. Evidence for direct cortical innervation of medial olivocochlear neurones in rats. *Hear Res* **144**:65–72. doi:10.1016/s0378-5955(00)00046-0

Murugasu E, Russell IJ. 1996. The effect of efferent stimulation on basilar membrane displacement in the basal turn of the guinea pig cochlea. *J Neurosci* **16**:325–332. doi:10.1523/JNEUROSCI.16-01-00325.1996

Oostenveld R, Fries P, Maris E, Schoffelen J-M. 2011. FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput Intell Neurosci* **2011**:156869. doi:10.1155/2011/156869

Pichora-Fuller MK, Kramer SE, Eckert MA, Edwards B, Hornsby BWY, Humes LE, Lemke U, Lunner T, Matthen M, Mackersie CL, Naylor G, Phillips NA, Richter M, Rudner M, Sommers MS, Tremblay KL, Wingfield A. 2016. Hearing Impairment and Cognitive Energy: The Framework for Understanding Effortful Listening (FUEL). *Ear Hear* **37 Suppl 1**:5S–27S. doi:10.1097/AUD.0000000000000312

Potts GF. 2004. An ERP index of task relevance evaluation of visual stimuli. *Brain Cogn* **56**:5–13. doi:10.1016/j.bandc.2004.03.006

Puel JL, Bonfils P, Pujol R. 1988. Selective attention modifies the active micromechanical properties of the cochlea. *Brain Res* **447**:380–383. doi:10.1016/0006-8993(88)91144-4

Rallapalli VH, Heinz MG. 2016. Neural Spike-Train Analyses of the Speech-Based Envelope Power Spectrum Model: Application to Predicting Individual Differences with Sensorineural Hearing Loss. *Trends in Hearing* **20**:2331216516667319. doi:10.1177/2331216516667319

Rastle K, Harrington J, Coltheart M. 2002. 358,534 nonwords: the ARC Nonword Database. *Q J Exp Psychol A* **55**:1339–1362. doi:10.1080/02724980244000099

Ritzer G, editor. 2007. Experimental DesignThe Blackwell Encyclopedia of Sociology. Oxford, UK: John Wiley & Sons, Ltd. doi:10.1002/9781405165518.wbeose087.pub2

Schaette R, McAlpine D. 2011. Tinnitus with a normal audiogram: physiological evidence for hidden hearing loss and computational model. *J Neurosci* **31**:13452–13457. doi:10.1523/JNEUROSCI.2156-11.2011

Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M. 1995. Speech recognition with primarily temporal cues. *Science* **270**:303–304. doi:10.1126/science.270.5234.303

Smith SB, Cone B. 2015. The medial olivocochlear reflex in children during active listening. *Int J Audiol* **54**:518–523. doi:10.3109/14992027.2015.1008105

Souza NN, Dhar S, Neely ST, Siegel JH. 2014. Comparison of nine methods to estimate ear-canal stimulus levels. *J Acoust Soc Am* **136**:1768–1787. doi:10.1121/1.4894787

Stuart A, Butler AK. 2012. Contralateral suppression of transient otoacoustic emissions and sentence recognition in noise in young adults. *J Am Acad Audiol* **23**:686–696. doi:10.3766/jaaa.23.9.3

Stuss DT, Picton TW, Cerri AM, Leech EE, Stethem LL. 1992. Perceptual closure and object identification: electrophysiological responses to incomplete pictures. *Brain Cogn* **19**:253–266. doi:10.1016/0278-2626(92)90047-p

Suthakar K, Ryugo DK. 2017. Descending projections from the inferior colliculus to medial olivocochlear efferents: Mice with normal hearing, early onset hearing loss, and congenital deafness. *Hear Res* **343**:34–49. doi:10.1016/j.heares.2016.06.014

Wagner W, Frey K, Heppelmann G, Plontke SK, Zenner H-P. 2008. Speech-in-noise intelligibility does not correlate with efferent olivocochlear reflex in humans with normal hearing. *Acta Otolaryngol* **128**:53–60. doi:10.1080/00016480701361954

Warr WB, Guinan JJ Jr. 1979. Efferent innervation of the organ of corti: two separate systems. *Brain Res* **173**:152–155. doi:10.1016/0006-8993(79)91104-1

Wiederhold ML, Kiang NY. 1970. Effects of electric stimulation of the crossed olivocochlear bundle on single auditory-nerve fibers in the cat. *J Acoust Soc Am* **48**:950–965. doi:10.1121/1.1912234

Winslow RL, Sachs MB. 1987. Effect of electrical stimulation of the crossed olivocochlear bundle on auditory nerve response to tones in noise. *J Neurophysiol* **57**:1002–1021. doi:10.1152/jn.1987.57.4.1002

Wittekindt A, Kaiser J, Abel C. 2014. Attentional modulation of the inner ear: a combined otoacoustic emission and EEG study. *J Neurosci* **34**:9995–10002. doi:10.1523/JNEUROSCI.4861-13.2014

Woodman GF. 2010. A brief introduction to the use of event-related potentials in studies of perception and attention. *Atten Percept Psychophys* **72**:2031–2046. doi:10.3758/APP.72.8.2031

Xiao Z, Suga N. 2002. Modulation of cochlear hair cells by the auditory cortex in the mustached bat. *Nat Neurosci* **5**:57–63. doi:10.1038/nn786

Supplemental Information

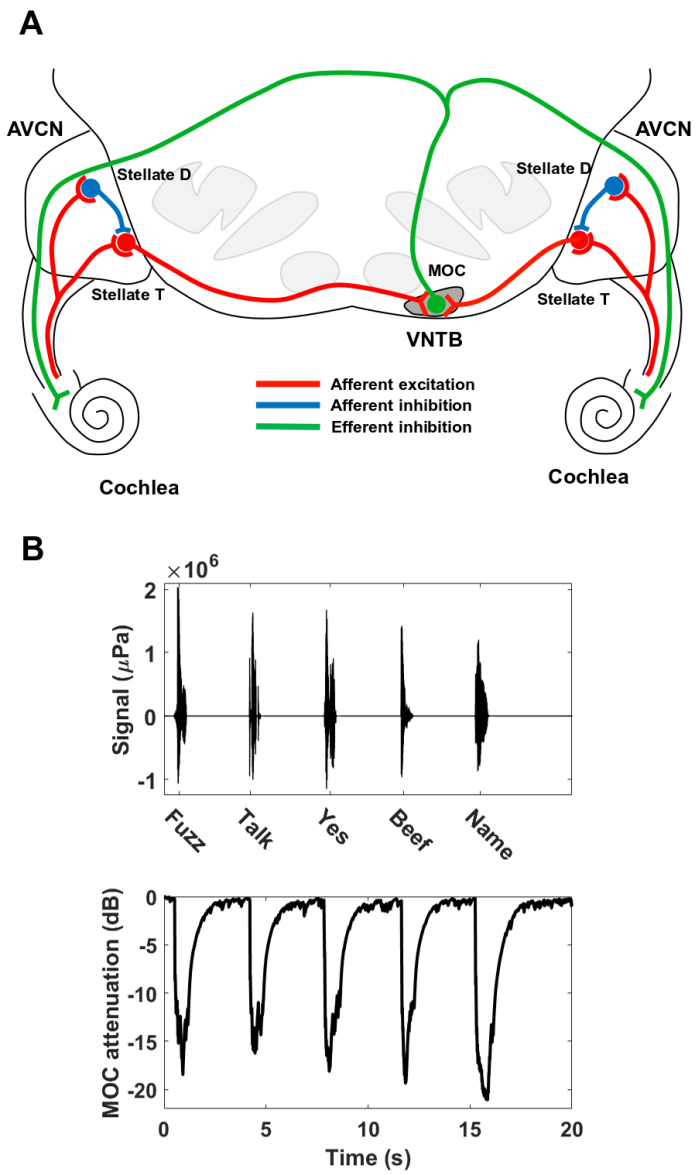

Figure supplement 1. Brainstem circuit and MOC attenuation of MAP-BS. **A**. Overview of brainstem circuit in MAP-BS. Bilateral afferent excitation (*red lines*) via high spontaneous rate auditory nerve (AN) fibers drives Stellate T (within-channel AN input) and D neurons (wideband AN input) in anteroventral cochlear nucleus (AVCN). Stellate D neurons provide ipsilateral, feedforward inhibition to Stellate T neurons in the same frequency channel (*blue lines*). Ventral nucleus of the trapezoid body (VNTB) neurons receive bilateral excitation from Stellate T neurons; their output is combined across 'left' and 'right' brainstems to calculate within-channel, efferent attenuation to both cochleae (*green lines*). Emulating inhibition of cochlear amplification, VNTB output impacts the non-linear pathway of dual-resonance nonlinear model of cochleae with a short delay (Liberman 1998, Clark et al., 2012). **B**. Temporal dynamics of efferent feedback in MAP-BS. Upper panel shows waveforms of naturally-spoken word tokens presented diotically at 75 dB to MAP-BS model with a 3 sec pause between tokens as in the psychoacoustic paradigm (Figure 1A). Lower panel presents MOC attenuation calculated across 'left' and 'right' VNTB populations. MOC attenuation did not summate across series of word tokens but recovered to baseline; single-shot presentation of word tokens was therefore implemented for the main results.

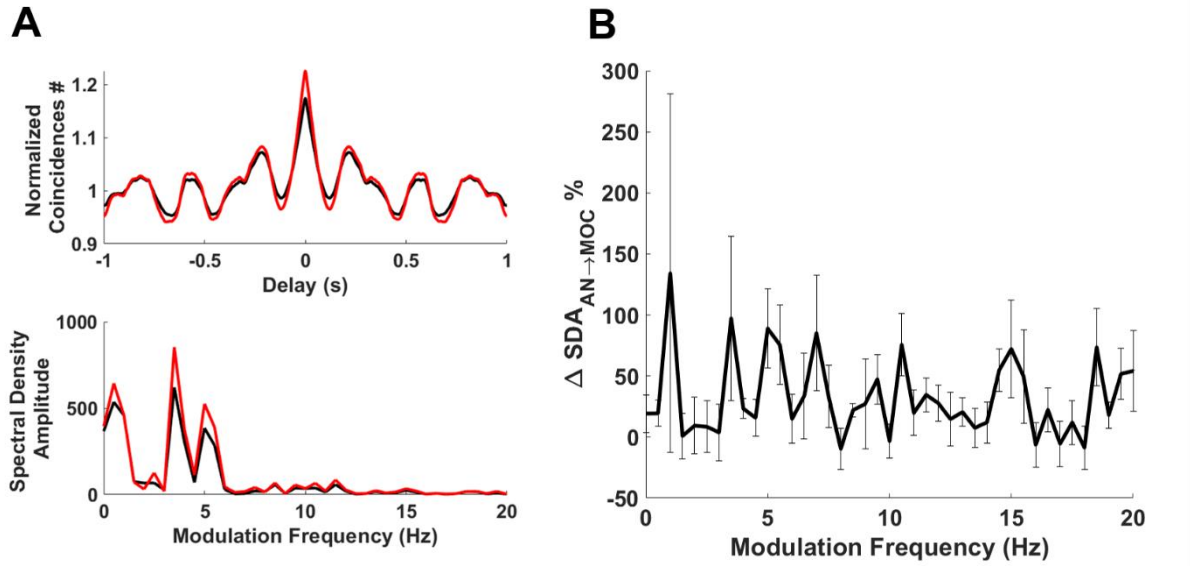

Figure supplement 2. Efferent feedback improves envelope encoding for naturally-spoken sentences. **A**. Shuffled-Correlogram Sumcors (upper panel) were calculated for the naturally-spoken sentence, 'the steady drip is worse than the drenching rain' (s86; The MAVA corpus (Aubanel et al., 2017)), using low spontaneous rate auditory nerve fiber output between 1-2kHz with (*red line*) and without (*black line*) efferent feedback. A longer, 1 sec delay-window was used compared to the single word presentation; in addition, inverted triangular compensation was implemented to compensate for large delays relative to signal length (Rallapalli and Heinz 2016). The envelope power spectral density (lower panel) was computed both with (*red line*) and without (*black line*) efferent feedback by computing Fourier transforms of the above Sumcors with a <1Hz spectral resolution. Efferent feedback was conducive to larger envelope responses, especially at low modulation frequencies associated with words and syllables. **B**. Mean percentage change (±SEM) in power spectral density amplitude ($\Delta SDA_{AN \rightarrow MOC}$) with efferent feedback included was computed using 6 sentences (s7, s26, s37, s42, s86, s164; The MAVA corpus) for modulation frequencies below 20 Hz. As for the single example in A, adding efferent feedback improved envelope encoding across most modulation frequencies.

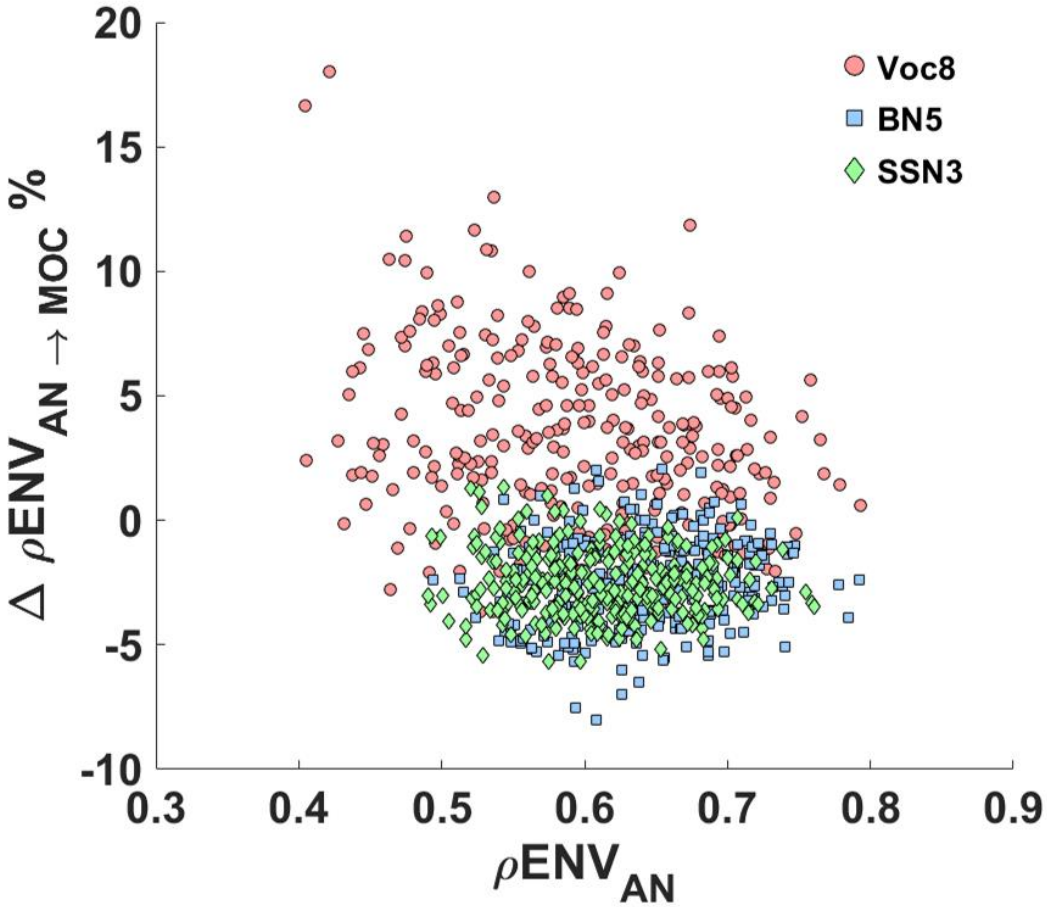

Figure supplement 3. Percentage change in envelope encoding of high spontaneous rate (HSR) auditory nerve (AN) fibers after introduction of efferent feedback. ΔρENVs for 300 words (in their three degraded forms) were calculated as in Figure 2D; however, HSR AN fibre output was used here. Although ΔρENVs for Voc8 words (*pink circles*) varied greatly and included negative values (Max-Min ΔρENV for Voc8 = +16.87 to -1.94%), mean ΔρENV was significantly positive (mean ΔρENV for Voc8 =+5.42 ± 0.18%, [t (299) = 30.17, p = 0.0001]). Note that values of ρENV_{AN} for Voc8 were significantly smaller here compared to matched values for low SR (LSR) AN fibers (mean $\rho ENV_{AN\text{-}HSR}$ for Voc8 = 0.58 ± 0.01 *vs.* mean $\rho ENV_{AN\text{-}LSR}$ for Voc8 = 0.70 ± 0.01, [t(299) = -39.21, p = 0.0001]). In contrast, the distributions of ΔρENVs for BN5 (*light blue squares*) and SSN3 (*green diamonds*) appeared more compact (Max-Min Range ΔρENV for BN5 words = +2.06 to -8.04%; Max-Min ΔρENV for SSN3 = +1.32 to -5.67 %); however, as for LSR AN fibre results, both mean ΔρENVs for HSR AN fibers were significantly negative overall (mean ΔρENV for BN5 = -2.54 ± 0.10, [t(299) = -24.92, p = 0.0001]; mean ΔρENV for SSN3 = -2.50 ± 0.08, [t(299) = -32.64, p = 0.0001]).

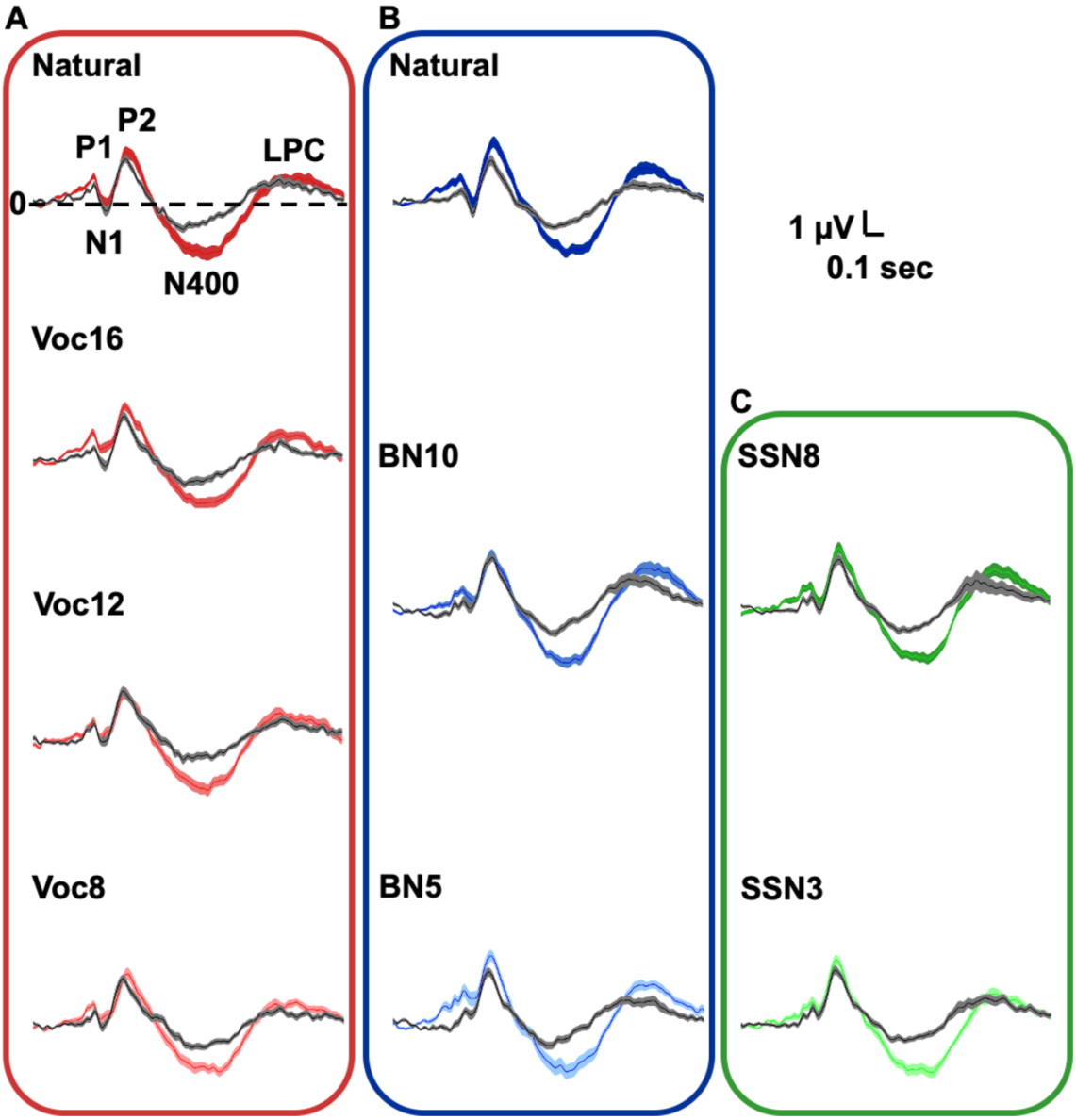

Figure supplement 4. Cortical evoked potentials during active and passive speech perception. Thick lines and shaded areas represent means and standard errors of the mean respectively. Within conditions analysis showed that, for all speech manipulations, the magnitude of P1, P2 and N400 potentials were enhanced during active (*color lines)* when compared to the passive (*gray lines*) listening conditions, whilst N1 tended to be less negative in the active task. Late Positivity Complex (LPC) magnitude was only significantly enhanced during the active listening of speech in noise. **A.** ERP components in natural and all noise vocoded manipulations: P1: [$F(1,25) = 5.99$, $p = 0.022$], N1: [$F(1, 25) = 16.00$, $p = 0.001$], P2: [$F(1, 25) = 12.48$, $p = 0.002$], N400: [$F(1, 25) = 33.74$, $p = 0.0001$], LPC: [$F(1,25)=3.38$, $p = 0.08$]. **B.** ERPs during natural (different population than noise vocoded experiment) and all babble noise manipulations: P1: [$F(1, 28) = 24.47$, $p = 0.0001$], N1: [$F(1, 28) = 10.46$, $p = 0.003$], P2: [$F(1, 28) = 10.65$, $p = 0.003$], N400: [$F(1, 28) = 62.16$, $p = 0.0001$, LPC: [$F(1,28)=10.65$, $p =0.003$]. **C.** ERP components during speech-shaped noise manipulations: P1: [$F(1, 28) = 22.98$, $p = 0.0001$], N1: [$F(1, 28) = 6.07$, $p = 0.02$], P2: [$F(1, 28) = 18.10$, $p = 0.001$] and N400: [$F(1, 28) = 60.75$, $p = 0.0001$]), LPC: [$F(1,28)=10.34$, $p =0.003$].

Table supplement 1. Hodgkin-Huxley model parameters and synaptic properties

	LSR AN Fiber	HSR AN Fiber	T Stellate	D Stellate	VNTB
Cell number	200/CF/ear	200/CF/ear	50/CF/side	50/CF/side	50/CF/side
Cell type inputs:	['ipsi' LSR syn]	['ipsi' HR syn]	['ipsi' AN (HSR); 'ipsi' VCN(chD)]	['ipsi' AN(HSR)]	['contra' VCN (ChT); 'ipsi' VCN (ChT)]
Input Bandwidth (Octaves)	[CF]	[CF]	[CF; CF]	[2]	[CF; CF]
Number of inputs	[1]	[1]	[15; 5]	[25]	[20; 5]
Max Synaptic Current (nA)	[30]	[40]	[2; 10]	[20]	[3; 3]
Synaptic Delay (ms)	[1.1]	[1.1]	[1.9; 1.9]	[1.9]	[1.9; 1.9]
LP dendritic filter (Hz)	[1000]	[1000]	[50; 50]	[500]	[70; 70]
Max Conductance (nS):					
$\bar{g}_{Na}$	[120]	[120]	[120]	[120]	[120]
$\bar{g}_{KLT}$	[5]	[10]	[20]	[25]	[5]
$\bar{g}_{KHT}$	[10]	[20]	[30]	[25]	[10]
$\bar{g}_{H}$	[0]	[5]	[0]	[0]	[0]
$\bar{g}_{L}$	[2.9]	[2.9]	[2.9]	[2.9]	[2.9]
Reversal Potentials (mV):					
E_{Na}	[+55]	[+55]	[+55]	[+55]	[+55]
E_{K}	[-80]	[-80]	[-80]	[-80]	[-80]
E_{H}	N/A	[-43]	N/A	N/A	N/A
E_{L}	[-62]	[-61]	[-59]	[-63]	[-64]
Membrane Capacitance (pF):					
C_{m}	[1]	[1]	[1]	[1]	[1]
Excitatory/Inhibitory output	N/A	Excitation	Excitation	Inhibition	MOC attenuation

LSR/HSR AN Fiber = Low/High Spontaneous Rate Auditory Nerve Fiber; *LSR/HSR syn* = Low/High Spontaneous Inner Hair Cell Synapse; *VNTB* = Ventral Nucleus of the Trapezoid Body. '*ipsi*' and '*contra*' is relative to the simulated ear or cell in question. *CF* is the characteristic frequency of AN fiber tested and represents within-channel for input bandwidth. *LP dendritic filter* represents cut-off frequency of low-pass dendritic filter used to determine unitary synaptic conductance ($\bar{g}_{syn}$) at time, *t*, as follows:

$$\tau_{SpiketoCurrent} = \frac{1}{(2\pi f_{LP\ dendritic\ filter})}$$

$$\bar{g}_{syn}(t) = \left(\frac{t - t_0}{\tau_{SpiketoCurrent}}\right) exp\left(-\frac{t - t_0}{\tau_{SpiketoCurrent}}\right)$$

Where t_0 is the onset of the synaptic event. Current per input per spike was calculated by dividing *Max synaptic current* by *Number of inputs*. The above conductance and synaptic parameters are not absolute values in the physiological sense, instead mirroring key features of the physiology with a view to reproducing the time course of spiking activity and the qualitative differences of neural response patterns.