Received February 13, 2020, accepted March 8, 2020, date of publication March 23, 2020, date of current version March 31, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2981885

Optimizing Speech Recognition Using a Computational Model of Human Hearing: Effect of Noise Type and Efferent Time Constants

IFAT YASIN[1], VIT DRGA[1], FANGQI LIU[2], ANDREAS DEMOSTHENOUS[2], (Fellow, IEEE), AND RAY MEDDIS[3]

[1]Department of Computer Science, University College London, London WC1E 6EA, U.K.
[2]Department of Electronic and Electrical Engineering, University College London, London WC1E 7JE, U.K.
[3]Department of Psychology, University of Essex, Colchester CO4 3SQ, U.K.

Corresponding author: Ifat Yasin (i.yasin@ucl.ac.uk)

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/R511638/1.

ABSTRACT Physiological and psychophysical methods allow for an extended investigation of ascending (afferent) neural pathways from the ear to the brain in mammals, and their role in enhancing signals in noise. However, there is increased interest in descending (efferent) neural fibers in the mammalian auditory pathway. This efferent pathway operates via the olivocochlear system, modifying auditory processing by cochlear innervation and enhancing human ability to detect sounds in noisy backgrounds. Effective speech intelligibility may depend on a complex interaction between efferent time-constants and types of background noise. In this study, an auditory model with efferent-inspired processing provided the front-end to an automatic-speech-recognition system (ASR), used as a tool to evaluate speech recognition with changes in time-constants (50 to 2000 ms) and background noise type (unmodulated and modulated noise). With efferent activation, maximal speech recognition improvement (for both noise types) occurred for signal-to-noise ratios around 10 dB, characteristic of real-world speech-listening situations. Net speech improvement due to efferent activation (NSIEA) was smaller in modulated noise than in unmodulated noise. For unmodulated noise, NSIEA increased with increasing time-constant. For modulated noise, NSIEA increased for time-constants up to 200 ms but remained similar for longer time-constants, consistent with speech-envelope modulation times important to speech recognition in modulated noise. The model improves our understanding of the complex interactions involved in speech recognition in noise, and could be used to simulate the difficulties of speech perception in noise as a consequence of different types of hearing loss.

INDEX TERMS Auditory, hearing, efferent, Medial OlivoCochlear (MOC), speech recognition, auditory model, time constant, SNR, amplitude-modulated noise.

I. INTRODUCTION

Much is now known of the role of ascending (afferent) neural fibers in the mammalian auditory pathway in conveying information about sound processing from peripheral auditory structures to higher brain centers. However, recently there has been renewed interest in the descending (efferent) neural pathways which convey information from higher brain centers back to peripheral auditory structures, and may play a role in humans in enhancing speech in background noise. One such efferent pathway, originating in the brainstem with connections in the Medial OlivoCochlear (MOC) system, innervates the cochlea of the inner ear [1], [2]. Within the cochlea, some of these efferent projections make synaptic contact directly at the base of outer hair cells acting to modulate the response of the basilar membrane (BM) to sound [3]. One of the main roles of MOC feedback in humans is suggested to be the improvement of speech perception in noisy environments [4], [5]. This is demonstrated by assessing the performance of individuals with lesions in the auditory

The associate editor coordinating the review of this manuscript and approving it for publication was Cesar Vargas-Rosales.

efferent pathway who display distinct difficulties with speech and vowel intelligibility in noise [6], [7].

Computational modeling of the auditory system provides a useful means of understanding the mechanisms contributing to improved speech recognition in noise. Modeling the auditory neural pathways has been informative with regards to understanding peripheral and central auditory processing [8], top-down cognitive and linguistic processing [9], changes in cortical projections due to deafness [10], and speech-recognition performance in noise [11].

Most auditory models simulating the effects of noise on speech have incorporated afferent neural processing [12]. Computational models of the auditory system which also include aspects of efferent processing [13]–[16] are able to demonstrate a marked improvement in speech intelligibility when the models have served as the front-end acoustic processors for automatic speech recognition (ASR) systems [11], [14], [16], [17]. However, most of these models used a single, relatively long, efferent decay time constant for modeling the MOC efferent effect, and/or used time constants not directly associated with human auditory processing.

Otoacoustic emission (OAE) measures from humans suggest a range of time constants associated with the MOC effect: slow (tens of seconds), medium (290-350 ms), or fast (ranging from 60-80 ms) [18], [19]. As to what purpose different co-existing time constants serve with regards to speech recognition in noise is still an open question. Previous studies using an auditory model [13], have used only long efferent time constants (e.g., 2000 ms), in order to study speech recognition in pink noise [14], [16]. In order to investigate the role of auditory efferent activation across the range of time constants recorded in humans [18], a previous study by the authors [20] investigated speech recognition in unmodulated (UM) pink noise with an auditory model [13] using shorter time constants (118, 200, 450, 1000 ms), as well as the longer time constant of 2000 ms originally used in earlier studies. A range of SNRs was used in the study (−10, −5, 0, 5, 7, 10, 12, 15 and 20 dB). The results showed that efferent time constants shorter than 2000 ms (but longer than 118 ms) can provide improved speech recognition in noise, and successive increases in efferent time constant (from 118 to 450 ms) leads to successive improvements in speech recognition in noise.

A number of studies suggest that the effect of efferent activation on the discrimination or recognition of speech in amplitude-modulated (AM) noise is different from that achieved in UM noise; the efferent effect may be most affected by modulations below 100 Hz [21], [22]. The present study is aimed at understanding the interplay between different efferent time constants and background noise type in improving speech recognition, and also extends the measurement of speech recognition performance to efferent time constants below 118 ms, i.e., 50, 70, 86, and 100 ms, (which have also been recorded in humans using OAEs [18] and in AM noise).

In Experiment 1 of the current study, speech recognition in noise was measured in UM pink noise with time constants ranging from 50 to 2000 ms. The results will allow for a direct comparison with earlier studies measuring speech recognition in UM pink noise, using an auditory model with a single long efferent time constant of 2000 ms [14], [16] or time constants within a range of 118-2000 ms [20]. In Experiment 2 of the current study, speech recognition in noise was also measured (using the same model settings) in AM pink noise using AM rates spread across 1-14 Hz (comparable to envelope fluctuation rates in speech). The results will allow for a direct comparison with Experiment 1 (which used UM pink noise), as well as comparison with earlier studies (measuring speech recognition in speech-shaped/babble noise) using auditory models with longer efferent time constants [14]–[16], [20].

For the current study, a computational model of the auditory system described by [13], and further developed by [14] and [16], was used. The model represents the responses of the outer ear, middle ear, vibration of the cochlear partition, and subsequent auditory-nerve (AN) firing [13], [23]–[25]. Time constants associated with the efferent response can be modified within the model to represent the range of time constants reported in human OAE studies [18], [26] or non-human mammalian physiological studies [27]. In the present study, the efferent decay time constants tested within the model were: 50, 70, 86, and 100 ms, in order to make a comparison with the shorter range of human auditory efferent time constants recorded using OAEs [18]; 118 ms, an efferent decay constant reported by [28] using human psychoacoustical measures; 200 ms, 450 ms, and 1000 ms which are within the range of slow and medium human efferent decay constants recorded using OAEs [18]; and 2000 ms, in order to make a direct comparison with two previous studies [14], [16]. UM pink noise was used in order to make a comparison with earlier findings [14], [16], [20], and AM pink noise was used in order to investigate the interaction of noise modulation with efferent time constants. UM pink noise or AM pink noise will hereafter be referred to as UM noise or AM noise.

The auditory model was prepended to an ASR system (as described below). The ASR was used in the present context as a tool to evaluate the effect on speech recognition of altering the parameters of the auditory model. The model was presented with speech items in pink noise at a range of SNRs (−10 to 20 dB) to assess which combinations of efferent time constants (50 to 2000 ms) provided the greatest improvement in speech recognition in noise. An additional condition (no-MOC condition), in which the efferent feedback was turned off in the auditory model was also run as a comparison.

II. METHODS

A. AUDITORY MODEL

The computational auditory model used in the current evaluation is a variant of the one described by [13] and subsequently used by [14] and [16] which models the auditory system from outer ear to the AN. We shall be using the model variant used in [16]. The key components of the auditory model and associated ASR system are shown in Fig. 1. Since the

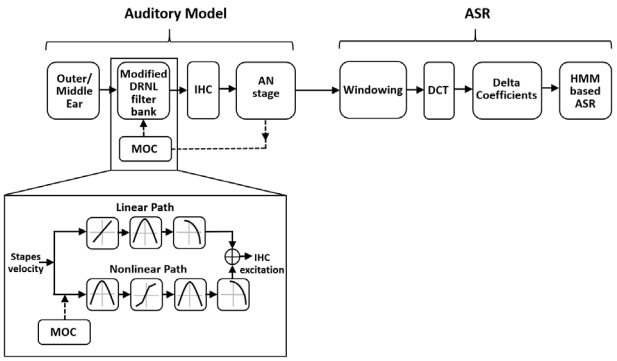

FIGURE 1. Schematic of the auditory model as front-end to the ASR, including a schematic of the modified Dual Resonance Nonlinear (DRNL) filterbank (adapted from [13]) to include Medial OlivoCochlear (MOC) feedback from the Auditory Nerve (AN) stage of the auditory model to the nonlinear path of the DRNL (DRNL component is shown in further detail in an enlarged box). The MOC feedback is modelled by inserting an attenuator at the start of the nonlinear path of the DRNL. The auditory model is composed of a model of the outer/middle ear, the output of which (as stapes velocity; m/s) is received as the input to the DRNL. The output of the DRNL (instantaneous velocity) feeds to a model of the inner hair cell (IHC) receptor potential transmitter release at the IHC and AN synapse. The AN stage of the model produces a probabilistic representation of the firing rate in the nerve. The DRNL comprises a linear path (linear gain, 2 Gammatone filters, 4 Butterworth filters) and a nonlinear path (3 Gammatone filters, compressive nonlinearity, 3 Gammatone filters, 3 Butterworth filters). The output of the auditory model (AN firing probability), is the input to the ASR system. Feature vectors are generated by first windowing the AN input, applying a discrete cosine transform (DCT) and generating the corresponding delta coefficients; these features are passed to the Hidden Markov Model (HMM)-based ASR.

model components are described in sufficient detail in these papers, only the salient components will be described here. The auditory model consists of a number of elements (modules) each representing a key stage in auditory processing such as outer ear, middle ear resonance, and response of the BM in the cochlea. The response of the BM is described by the dual resonance nonlinear (DRNL) model [23], [29], which receives outer-ear stapes displacement information as input. The DRNL model describes cochlear vibrations using parallel linear and nonlinear pathways, receiving as its input the middle-ear stapes velocity with a linear-gain path (modelled by 2 first-order Gammatone filters plus low-pass filtering by 4 second-order Butterworth filters), and a nonlinear-gain path (modelled by 3 first-order Gammatone filters, a compressive nonlinearity, 3 first-order Gammatone filters and 3 second-order Butterworth filters). AN fibers can be modelled using the approach described in [13] using the rate-level functions detailed in [30]. AN fiber characteristics are modelled by setting the calcium clearance time constants [25], [31] as used in a speech-recognition task by [14]. An auditory filterbank of 21 channels were specified in 4-octave steps from 250-8000Hz.

The output of the DRNL (instantaneous velocity) feeds forward to a model of the inner hair cell (IHC) receptor potential, transmitter release, and adaptation at the IHC and the AN synapse. The model of the AN firing rate includes responses from both high-spontaneous rate and low-spontaneous rate fibers. The current model settings used combined high-spontaneous rate and low-spontaneous rate neural fiber settings. The output of the AN component of the model produces a probabilistic representation of the firing rate in the nerve. The reduction in the BM response caused by efferent (MOC) activity is modelled by attenuating the signal at the start of the nonlinear pathway in the model. The DRNL model with the modelled efferent attenuation [13] has been shown to provide simulation data in good agreement with physiological measurements of the BM, AN, and compound action potentials, when the magnitude of attenuation (dB) is selected to be proportional to the amount of efferent activity [13], [16].

Efferent control is implemented as a feedback signal from the AN module in the model, which controls the amount of attenuation in the nonlinear path of the DRNL model (using the recent history of the AN firing response) [14], [16]. The control signal (in the feedback loop) was based on AN fiber firing rate, and was derived from a temporal smoothing of the instantaneous firing rate to provide a good fit to the rate-level functions of [32], as described in [16]. This control signal was then multiplied with a scalar rate-to-attenuation factor [16]. The AN firing rate smoothing was implemented using a first-order lowpass filter. A 10-ms lag was incorporated to account for the MOC-OHC synaptic latency (minimal lag) [32]. The output of the auditory model (AN firing probability) per auditory channel is the input to the ASR system.

B. ASR TRAINING AND EVALUATION

The auditory model output forms the input to the ASR system. A comprehensive description of the ASR which uses the auditory model as its front-end can be found in [14] and [16]. However the main steps in combining the model output with ASR input and model evaluation are provided below. A continuous hidden-density Hidden Markov Model (HMM) [33] was used. The input signal to the HMM is a sequence of feature vectors. These are generated by integrating the AN firing probability within overlapping 25-ms windows spaced at 10-ms intervals, and applying a discrete cosine transform (DCT) to yield a set of vector components. Time derivatives of the static DCT coefficients were included. In this respect, the first and second-order regression coefficients (''deltas'' and ''accelerations'') were appended to each vector to improve speech recognition performance [16], [34].

For the training phase, the speech material was taken from the TIDIGITS corpus [35]. The recognizer was trained on a clean set of material (without either background noise or efferent-related attenuation) consisting of 8439 utterances. The evaluation task was to identify a connected sequence of digits in the presence of background noise (in this case, pink noise). For testing the recognizer, 452 utterances were used, each containing three connected digits from the set (''oh'', ''one'' ''two'', ''three'', ''four'', ''five'', ''six'', ''eight'' and ''nine''). Each test stimulus comprised a sample of 6s of background noise, as used by [16] (a duration sufficient enough to initiate the efferent response) preceding the combined

speech-plus-noise segment. The HMM finds the most probable sequence of digits corresponding to the input sequence of features. A correct response was classified as one in which the recognizer correctly identified all three digits in the correct position within the presented triplet of digits. The speech samples used for the training phase were distinct from the speech samples used for the subsequent testing phase.

C. EXPERIMENTAL CONDITIONS AND MODEL RUNS

For the study, speech intelligibility in noise was assessed by measuring the output of the ASR in response to speech signals mixed with pink noise to produce different SNRs. The utterances were presented at 60 dB Sound Pressure Level (SPL) within the simulation, and UM or AM noise was added to the utterances to generate the different SNRs with each noise type. Experiment 1 used UM noise. The speech-plus-noise stimuli were presented in 81 conditions of nine efferent time constants (50, 70, 86, 100, 118, 200, 450, 1000 and 2000 ms), nine SNRs (−10, −5, 0, 5, 7, 10, 12, 15, 20 dB). A comparison condition was also run using the same time constants and SNRs in which there was no MOC feedback implemented in the model.

Experiment 2 used AM noise. The same time constants and SNR conditions as Experiment 1 were tested (i.e., 81 conditions of efferent time constant and SNR). A comparison condition was also run using AM noise with no-MOC feedback, which was separate to the no-MOC conditions for the UM noise in Experiment 1. The pink noise was amplitude-modulated using an eight-component modulator (components had frequencies of 0.9. 2.7, 4.0, 5.9, 7.7, 10.3, 13.7 Hz), combined with equal amplitude and in random phase. The modulator frequencies included 4.0 Hz, but otherwise were selected to have slight offsets from integer values, so as to reduce subharmonic interactions. The frequencies were chosen to include the components suggested to be the most useful for conveying linguistic information [36], with most phonetic classes falling within this range of modulation frequencies [37]. The long-term-average power level of the AM noise was set to match the long-term-average power level of the UM noise in Experiment 1. Data from 10 model runs for each of the 81 plus no-MOC conditions for both Experiments 1 and 2 were collected. Each model run for a given data-series took approximately 24-28 hours to complete both the training and testing phases. Different pseudorandom number seeds were used per run, in order to obtain estimates of possible variability across runs.

III. RESULTS

A. COMPARING EFFECTS OF UM AND AM NOISE ON SPEECH RECOGNITION

To compare the effects of UM and AM noise on speech recognition, the data (10 runs per 81 conditions for each noise type) were averaged across runs and plotted in Fig. 2. Fig. 2 shows the plots of the averaged value (across 10 runs) of percentage word correct (%) as a function of SNR (dB) for speech presented in UM or AM noise. Logistic functions

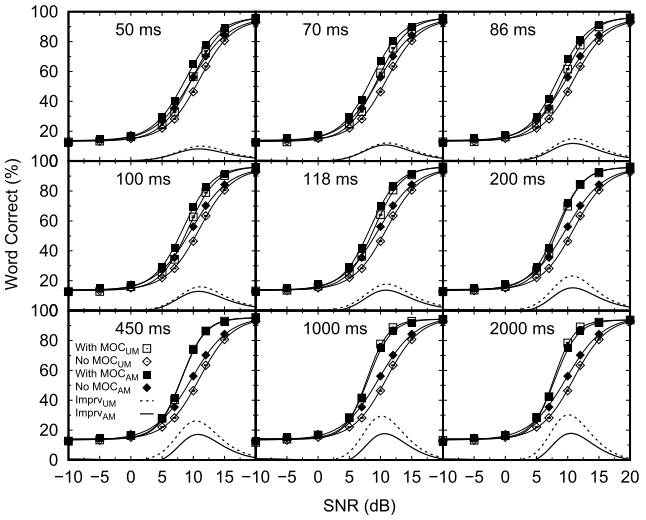

FIGURE 2. ASR performance, word correct (%) as a function of SNR (dB), obtained for speech recognition in UM and AM noise for different efferent time constants. Results with UM noise (unfilled square symbols) and AM noise (filled square symbols) are shown. The comparison plots for the data obtained in the condition where there was no efferent feedback (no-MOC) are also shown for the conditions of UM noise (unfilled diamond symbols) and AM noise (filled diamond symbols). Each panel displays the word recognition performance for a given efferent time constant (50, 70, 86, 100, 118, 200, 450, 1000 and 2000 ms). The logistic-function fit to each data series are shown by thin black lines overlaying the open symbols. Standard Error bars are also shown (mostly falling within the symbol size). The humped series at the lower right of each panel shows the net-improvement in speech recognition score with efferent activation per time constant for UM noise (dashed lines) and AM noise (solid line). The net improvement in speech recognition in noise was estimated by subtracting the word recognition scores obtained in the absence of efferent activation from the word scores obtained with efferent activation.

relating the word correct score (%) to SNR (dB) were fitted to each data series using the Bayesian procedure and *psignifit 4* software described by [38]. The lower and upper asymptotes were allowed to vary in the fitting procedure. Each panel of Fig. 2 displays the speech-recognition scores for a given efferent time constant. For each panel, logistic-function fits to the data points for the conditions in which MOC was activated are shown with overlaid logistic function fits, for the UM noise condition (with-MOC UM; unfilled squares and solid lines) and AM noise condition (with-MOC AM; filled squares and solid lines). Logistic-function fits to data points obtained with no MOC are also shown: UM noise (no-MOC UM; unfilled diamonds and dashed lines) and AM noise (no-MOC AM; filled diamonds and dashed lines). The standard errors (SE) are also displayed.

In order to estimate the net improvement in speech recognition with efferent activation, the speech-recognition scores obtained in the absence of efferent activation were subtracted from the speech scores obtained with efferent activation in UM noise (imprv_{UM}) or AM noise (imprv_{AM}). The net improvement in speech recognition with efferent activation in UM or AM noise is labelled in Fig. 2 as Improvement UM (Imprv_{UM}; dashed lines) or Improvement AM (Imprv_{AM}; solid lines).

The general trend for an increase in speech recognition scores in the absence of efferent activation with increasing

SNR is similar to that reported by [16] and [20] for SNR values up to 20 dB. Overall, for SNRs of 0 to 20 dB, there was an improvement in speech recognition with efferent activation, compared to when there was no efferent activation within the model. With successive increases in efferent time constant (from 50 to 2000 ms), the logistic function fit steepened to improved word recognition scores (compared to the case without efferent activation; the no-MOC condition). For the shortest efferent time constant of 50 ms, there was minimal speech improvement with efferent activation.

Except at the longest time constants, both the with-MOC AM and no-MOC AM performance is generally better than that for the respective results with UM noise. This is most likely due to AM noise containing brief intervals where the masker power is relatively small compared to the long-term average SNR, which for fixed-level speech, allows temporary improvements in SNR to boost performance (so-called "dip listening"). Furthermore, as the time constant increases, speech recognition performance in AM noise improves (the functions shift leftwards), but only for time constants up to 200 ms, after which performance remains steady. Overall, performance for the model when incorporating efferent processing is better under AM noise, than unmodulated noise up to the 200 ms time constant, and is worse under AM noise for time constants above 450 ms. As the SNR increases, the peak performance for net speech recognition in either UM or AM pink noise occurs at around an SNR of 10 dB. Overall, performance in UM noise shows an improvement over that achieved in AM noise for time constants longer than 200 ms.

A three-way unrelated ANOVA was conducted on the net-improvement in speech recognition scores with factors of noise type (2 levels: UM noise, AM noise), SNR (9 levels: −10, −5, 0, 5, 7, 10, 12, 15 and 20), and time constant (9 levels: 50, 70, 86, 118, 200, 450, 1000, and 2000 ms).

There was a significant effect of noise type ($F_{(1,1458)} = 1163.8$, $p < 0.001$ (two-tailed) with effect size, $\eta^2 = 0.444$), SNR ($F_{(8,1458)} = 4434.4$, $p < 0.001$ (two-tailed) with effect size, $\eta^2 = 0.96$), and time constant ($F_{(8,1458)} = 223.7$, $p < 0.001$ (two-tailed) with effect size, $\eta^2 = 0.55$). There was a significant 2-way interaction between noise type and SNR ($F_{(8,1458)} = 139.6$, with effect size, $\eta^2 = 0.43$, $p < 0.001$ (two-tailed), and a significant 2-way interaction between noise type and time constants ($F_{(8,1458)} = 67.08$, with effect size, $\eta^2 = 0.27$, $p < 0.001$ (two-tailed), and a significant 2-way interaction between SNR and time constants ($F_{(64,1458)} = 53.93$, with effect size, $\eta^2 = 0.70$, $p < 0.001$ (two-tailed). There was also a significant 3-way interaction between noise type, SNR and time constants ($F_{(64,1458)} = 8.37$, with effect size, $\eta^2 = 0.27$).

To investigate the 2-way interaction between noise type (UM and AM) and SNR, *post hoc* pairwise comparisons (Bonferroni corrected) were conducted and showed that speech scores were significantly improved, i.e., higher % correct scores ($p < 0.05$ (two-tailed), in AM noise compared to UM noise, at SNRs of −10 and −5 dB. For SNRs of 0, 5, 7, 10, 12, 15, and 20 dB, speech scores were significantly improved, ($p < 0.05$, two-tailed), in the presence of UM noise compared to AM noise. To investigate the 2-way interaction between noise type (UM and AM) and time constants, *post hoc* pairwise comparisons (Bonferroni corrected) showed that speech scores were significantly improved, i.e., higher % correct scores ($p < 0.05$, two-tailed), in UM noise compared to AM noise, at all time constants of 50, 70, 86, 100, 118, 200, 450, 1000, and 2000 ms ($p < 0.05$ (two-tailed).

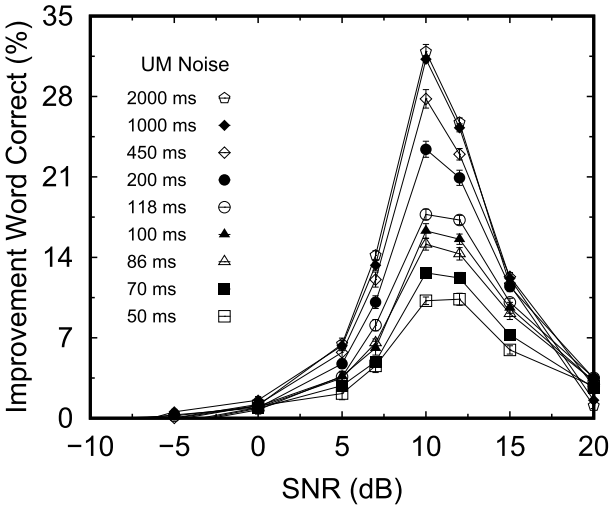

FIGURE 3. The net improvement to speech recognition in UM noise with efferent activation was estimated by subtracting the word recognition scores obtained in the absence of efferent activation from the word recognition scores obtained with efferent activation. The net-improvement in word recognition scores is shown as a function of SNR for time constants of 50, 70, 86, 100, 118, 200, 450, 1000 and 2000 ms.

To investigate the 3-way interaction between noise type (UM and AM), SNR, and time constants, *post hoc* pairwise comparisons (Bonferroni corrected) were conducted and showed that at a noise level of −10 dB, AM noise resulted in significantly higher (improved) speech scores compared to UM noise, with a time constant of 1000 ms ($p < 0.01$, two-tailed). At other noise levels of 10, 12, and 15 dB, UM noise resulted in significantly higher (improved) speech scores compared to AM noise, with low and midrange time constants of 50, 70, 86, 100, 118, 200, and 450 ms ($p < 0.01$, two-tailed). The midrange to longer time constants of 450, 1000, and 2000 ms, were significantly more effective at improving speech scores in UM noise compared to AM noise at noise levels of 7 dB (in addition to time constants of 118, 200 ms), 5 dB (in addition to time constants of 100, 118, 200 ms), and 0 dB (in addition to a time constant of 50 ms) ($p < 0.01$, two-tailed). The net improvement in speech recognition with UM or AM noise as a function of SNR, for each time constant, is shown in Fig. 3 (UM noise) and Fig. 4 (AM noise), respectively. Comparison of Figs 3 and 4 show that i) the peak in speech improvement with UM or AM noise occurs at around 10 dB, ii) the greater improvement in speech scores is achieved with UM noise compared to AM noise for SNRs above −10 dB, and iii) comparing word-correct performance in UM and AM noise, divergence in word correct

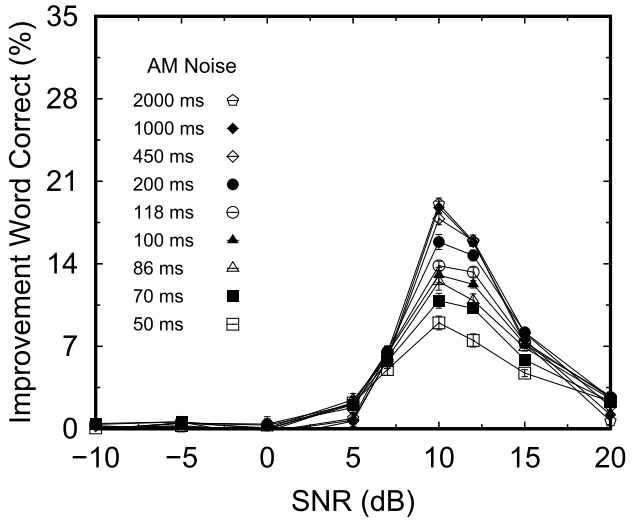

FIGURE 4. The net improvement to speech recognition in AM noise with efferent activation was estimated by subtracting the word recognition scores obtained in the absence of efferent activation from the word recognition scores obtained with efferent activation. The net-improvement in word recognition scores is shown as a function of SNR for time constants of 50, 70, 86, 100, 118, 200, 450, 1000 and 2000 ms.

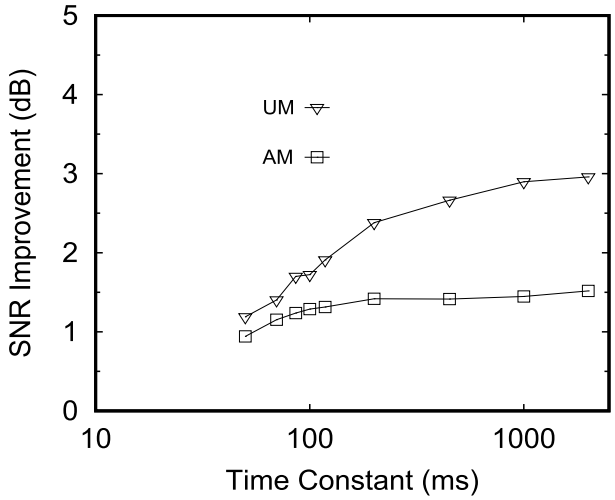

FIGURE 5. Effective improvement in SNR (dB) at the 50% speech recognition cut-off as a result of increasing efferent time constant. The effective improvement in SNR (dB) estimated as the difference in performance between the with-MOC condition and the no-MOC condition is plotted as a function of increasing efferent time constant. Unfilled downturned triangle and square symbols represent data for UM and AM conditions, respectively.

performance occurs around a time constants of 100-200 ms; with improved word correct performance in UM noise for longer efferent time constants. For Figs 3 and 4, the maximum improvement in word correct score (peak in improvement) is shifted slightly to lower SNRs as the time constant duration increases.

B. EFFECTIVE SNR IMPROVEMENT FOR DIFFERENT NOISE TYPES AND TIME CONSTANTS FOR ACHIEVING A 50% CORRECT WORD RECOGNITION SCORE

To further explore the improvement in speech recognition, the effective improvement in SNR was calculated by subtracting the no-MOC speech-correct scores from the with-MOC speech-correct scores for UM and AM noise conditions, at the 50% speech-correct point, derived from the logistic-function fit for each efferent time constant. Fig. 5 shows the plot of the SNR improvement as a function of efferent time constant at the 50% word-correct cut-off score for each fiber-type condition.

In general, the effective improvement in SNR (dB) at the 50% word-correct level in the presence of UM noise with increasing efferent time constant is most rapid as the efferent time constant increases from 50 ms to 450 ms. For further increases in time-constant duration (450 ms to 2000 ms), effective improvement in SNR increases less rapidly and remains roughly constant after about 1000 ms and at maximal level. Comparing these results with those for word recognition in AM noise, the effective improvement in SNR (dB) at the 50% speech-correct level with increasing efferent time constant in AM noise is most rapid as the efferent time constant increases from 50 ms to 100 ms. Beyond 100 ms and up to 2000 ms, effective improvement in SNR remains roughly constant and at maximal level, albeit below the level attained with UM noise.

IV. DISCUSSION

This study investigated the interaction between short- and long-duration efferent time constants on speech recognition in UM and AM noise at a range of SNRs using a computational model of the auditory system combined with an HMM-based ASR system. For SNRs of −5 to 15 dB, the results show that there is marked improvement in speech recognition with efferent feedback. The greatest increase in performance (in terms of speech-recognition score) due to the efferent effect is seen for mid-range SNRs of about 5-15 dB. This range of SNRs corresponds well with the reported range of SNRs characteristic of most real-world listening situations encountered by individuals; i.e., SNRs between 2 and 14 dB comprise approximately 63% of all SNRs typically encountered in daily life with the most common SNRs being around 8 dB [39]. Reference [11] used an efferent-inspired auditory model to assess speech discrimination in noise using the Multi-Band Non-Linear (MBPNL) model of cochlear mechanics [40]. Despite differences in realizations of cochlear mechanics between the DRNL and MBPNL models and the specifics of model efferent feedback, [11] also showed that speech discrimination was improved with efferent feedback, particularly at 60 dB SPL and an SNR of 10 dB. Reference [17] used an auditory model with a gain-control based on an ensemble interval histogram representation (based on available cat data) and showed that speech representation could be improved with efferent-inspired feedback; although [17] did not measure speech recognition performance with specific efferent time constants derived from humans.

In the present study, the word-recognition score approaches a lower asymptote value of about 11-17%, at SNRs of 0 dB and below. This is similar to the reduction of speech recognition performance to chance level shown by [14] and [20]

at 0 dB and negative SNRs, most likely due to saturating rate-level functions of the simulated auditory-nerve fibers in the model. In general, the speech recognition accuracy obtained is lower than that observable for a human listener (the human-machine speech gap), as seen in a study of human listeners' performance on the same speech stimuli (digits) [41]. Although incorporation of efferent parameters in a model of auditory processing improves speech recognition, which helps reduce the human-machine speech gap [16], there is scope for further improvement as we learn more about efferent effects such as the frequency tuning of the effect [42] and binaural efferent effects [43], [44].

It is possible that shorter efferent time constants such as below 100 ms may play a predominant role in de-noising and improving speech recognition in noise (at SNRs around 10-12 dB), particularly since these temporal durations fall within the range of durations characteristic of the modulation of the speech envelope, crucial for intelligibility (4-32 Hz, i.e., durations of 31-250 ms) [45]. A range of efferent time constants have been measured from humans using OAEs [18], [46] ranging from fast (about 50-80 ms) to slow (tens of seconds). The purpose of such a range of different auditory efferent time constants within the human auditory system is still largely unknown. It has been suggested that different efferent time constants may be required in different listening situations [27]. A number of studies suggest that efferent activation may play a role in the perception of amplitude modulation in general [47] and in human speech perception [48], [49], although the effect of different background noise-type and time constants has not been explicitly assessed. Recent studies with a binaural cochlear implant sound coding strategy with efferent-inspired feedback also demonstrate improved speech intelligibility in noise with very short efferent time constants [50].

Studies measuring efferent time constants in humans (using OAEs or psychophysical methods) have not explicitly assessed the contribution of differing efferent time constants to speech perception. The present study attempted to bridge this gap in knowledge by using a computational auditory model combined with an ASR to assess the effect of efferent time constants derived from human OAE/psychoacoustical studies. Based on the current findings it is sufficient to say that: i) the optimal performance (in terms of speech-recognition score) due to efferent activation is observed for SNRs of around 5-15 dB, characteristic of most real-world speech-listening situations encountered by individuals [46]; ii) efferent time constants with temporal durations falling within the range of durations characteristic of the modulation of the speech envelope may be important for enhancing speech perception; and iii) when comparing speech recognition performance in unmodulated versus modulated sounds, the range of efferent time constants (up to about 200 ms) required to achieve maximal SNR improvement for achieving 50% word correct score in modulated noise is smaller than for unmodulated noise (up to about 1000 ms), when using the currently described settings of the current model.

A number of studies suggest that aspects of the efferent effect observed in the discrimination or recognition of speech in modulated noise is different from that in UM noise. In such studies, human OAE data suggest that the efferent effect in the presence of modulated sounds may be most affected by modulations below or around 100 Hz [22], [51]. Reference [52] suggested that it is the cross-frequency profile of low-frequency neural fluctuation amplitudes that is important in encoding complex sounds such as speech. In this proposal by [52] the efferent system acts to increase differences between fluctuation amplitudes, and it is fluctuation differences between speech and ongoing noise that is important.

The current findings show that in AM noise, the range of efferent time constants required to achieve maximal speech recognition is smaller in AM noise compared to UM noise when using the current described settings of the auditory model. Possibly, the longer efferent time constants are not so beneficial in modulated noise (comprising of relatively low modulation rates). Modulated noise has been shown to have a variable effect on efferent activity, although it has often been shown to result in a reduced efferent response [21], [22], [53]. Whilst the current results show marked improvement in speech recognition with increasing time constant, for speech recognition in UM and AM noise it is possible that with either different types of modulated noise or signal level, differing patterns of speech recognition enhancement with time-constants associated with efferent activation may be observed and would be useful to characterize in future studies.

V. CONCLUSION

Maximal speech improvement for either UM or AM noise with efferent activation is at an SNR of around 10 dB.

For AM noise, the net speech improvement with efferent activation is predominant for shorter time constants ($<$200 ms). It is possible that efferent time constant durations that fall within the range of durations characteristic of the modulation of the speech envelope play the most important role for speech intelligibility in noise. Longer efferent time constants possibly are not so beneficial in AM noise (comprising of relatively low modulation rates).

The range of efferent time constants (up to about 200 ms) required to achieve maximal SNR improvement for achieving 50% word correct score in AM noise is smaller than the range for UM noise (up to about 1000 ms), when using the currently described settings of the current model.

ACKNOWLEDGMENT

During completion of this paper, our colleague and friend Ray Meddis passed away. The authors greatly valued and appreciated his insight, scientific discourses and generosity of spirit.

REFERENCES

[1] G. L. Rasmussen, "The olivary peduncle and other fiber projections of the superior olivary complex," *J. Comparative Neurol.*, vol. 84, no. 2, pp. 141–219, Apr. 1946.

[2] W. B. Warr, "Olivocochlear and vestibular efferent neurons of the feline brain stem: Their location, morphology and number determined by retrograde axonal transport and acetylcholinesterase histochemistry," *J. Comparative Neurol.*, vol. 161, no. 2, pp. 159–181, May 1975.

[3] I. J. Russell and E. Murugasu, "Medial efferent inhibition suppresses basilar membrane responses to near characteristic frequency tones of moderate to high intensities," *J. Acoust. Soc. Amer.*, vol. 102, no. 3, pp. 1734–1738, Sep. 1997.

[4] R. L. Winslow and M. B. Sachs, "Single-tone intensity discrimination based on auditory-nerve rate responses in backgrounds of quiet, noise, and with stimulation of the crossed olivocochlear bundle," *Hearing Res.*, vol. 35, nos. 2–3, pp. 165–189, Sep. 1988.

[5] T. Kawase, B. Delgutte, and M. C. Liberman, "Antimasking effects of the olivocochlear reflex. II. Enhancement of auditory-nerve response to masked tones," *J. Neurophysiol.*, vol. 70, no. 6, pp. 2533–2549, Dec. 1993.

[6] A. L. Giraud, S. Garnier, C. Micheyl, G. Lina, A. Chays, and S. Chéry-Croze, "Auditory efferents involved in speech-in-noise intelligibility," *NeuroReport*, vol. 8, no. 7, pp. 1779–1783, May 1997.

[7] F.-G. Zeng and S. Liu, "Speech perception in individuals with auditory neuropathy," *J. Speech, Lang., Hearing Res.*, vol. 49, no. 2, pp. 367–380, Apr. 2006.

[8] E. Aguilar, P. T. Johannesen, and E. A. Lopez-Poveda, "Contralateral efferent suppression of human hearing sensitivity," *Frontiers Syst. Neurosci.*, vol. 8, p. 251, Jan. 2015.

[9] V. Marian, T. Q. Lam, S. Hayakawa, and S. Dhar, "Top-down cognitive and linguistic influences on the suppression of spontaneous otoacoustic emissions," *Frontiers Neurosci.*, vol. 12, p. 378, Jun. 2018.

[10] B. E. Butler, J. K. Sunstrum, and S. G. Lomber, "Modified origins of cortical projections to the superior colliculus in the deaf: Dispersion of auditory efferents," *J. Neurosci.*, vol. 38, no. 16, pp. 4048–4058, Apr. 2018.

[11] D. P. Messing, L. Delhorne, E. Bruckert, L. D. Braida, and O. Ghitza, "A non-linear efferent-inspired model of the auditory system; matching human confusions in stationary noise," *Speech Commun.*, vol. 51, no. 8, pp. 668–683, Aug. 2009.

[12] M. Holmberg, D. Gelbart, and W. Hemmert, "Speech encoding in a model of peripheral auditory processing: Quantitative assessment by means of automatic speech recognition," *Speech Commun.*, vol. 49, no. 12, pp. 917–932, Dec. 2007.

[13] R. T. Ferry and R. Meddis, "A computer model of medial efferent suppression in the mammalian auditory system," *J. Acoust. Soc. Amer.*, vol. 122, no. 6, pp. 3519–3526, Dec. 2007.

[14] G. J. Brown, R. T. Ferry, and R. Meddis, "A computer model of auditory efferent suppression: Implications for the recognition of speech in noise," *J. Acoust. Soc. Amer.*, vol. 127, no. 2, pp. 943–954, Feb. 2010.

[15] C.-Y. Lee, J. Glass, and O. Ghitza, "An efferent-inspired auditory model front-end for speech recognition," in *Proc. 12th Ann. Conf. Int. Speech Commun. Assoc.*, 2011, pp. 49–52.

[16] N. R. Clark, G. J. Brown, T. Jürgens, and R. Meddis, "A frequency-selective feedback model of auditory efferent suppression and its implications for the recognition of speech in noise," *J. Acoust. Soc. Amer.*, vol. 132, no. 3, pp. 1535–1541, Sep. 2012.

[17] O. Ghitza, "Auditory neural feedback as a basis for speech processing," in *Proc. Int. Conf. Acoustics, Speech, Signal Process. (ICASSP)*, Apr. 1988, pp. 91–94.

[18] B. C. Backus and J. J. Guinan, Jr., "Time-course of the human medial olivocochlear reflex," *J. Acoust. Soc. Amer.*, vol. 119, no. 5, pp. 2889–2904, 2006.

[19] J. J. Guinan, "Olivocochlear efferents: Anatomy, physiology, function, and the measurement of efferent effects in humans," *Ear Hearing*, vol. 27, no. 6, pp. 589–607, Dec. 2006.

[20] I. Yasin, F. Liu, V. Drga, A. Demosthenous, and R. Meddis, "Effect of auditory efferent time-constant duration on speech recognition in noise," *J. Acoust. Soc. Amer.*, vol. 143, no. 2, pp. EL112–EL115, Feb. 2018.

[21] S. Maison, C. Micheyl, and L. Collet, "Sinusoidal amplitude modulation alters contralateral noise suppression of evoked otoacoustic emissions in humans," *Neuroscience*, vol. 91, no. 1, pp. 133–138, Jun. 1999.

[22] S. Boothalingam, D. Purcell, and S. Scollie, "Influence of 100 Hz amplitude modulation on the human medial olivocochlear reflex," *Neurosci. Lett.*, vol. 580, pp. 56–61, Sep. 2014.

[23] R. Meddis, L. P. O'Mard, and E. A. Lopez-Poveda, "A computational algorithm for computing nonlinear auditory frequency selectivity," *J. Acoust. Soc. Amer.*, vol. 109, no. 6, pp. 2852–2861, Jun. 2001.

[24] E. A. Lopez-Poveda and R. Meddis, "A human nonlinear cochlear filterbank," *J. Acoust. Soc. Amer.*, vol. 110, no. 6, pp. 3107–3118, Dec. 2001.

[25] C. J. Sumner, E. A. Lopez-Poveda, L. P. O'Mard, and R. Meddis, "A revised model of the inner-hair cell and auditory-nerve complex," *J. Acoust. Soc. Amer.*, vol. 111, no. 5, pp. 2178–2188, 2002.

[26] W. Zhao and S. Dhar, "Fast and slow effects of medial olivocochlear efferent activity in humans," *PLoS ONE*, vol. 6, no. 4, 2011, Art. no. e18725.

[27] N. P. Cooper and J. J. Guinan, "Separate mechanical processes underlie fast and slow effects of medial olivocochlear efferent activity," *J. Physiol.*, vol. 548, no. 1, pp. 307–312, Apr. 2003.

[28] I. Yasin, V. Drga, and C. J. Plack, "Effect of human auditory efferent feedback on cochlear gain and compression," *J. Neurosci.*, vol. 34, no. 46, pp. 15319–15326, Nov. 2014.

[29] E. A. Lopez-Poveda, "An approximate transfer function for the dual-resonance nonlinear filter model of auditory frequency selectivity," *J. Acoust. Soc. Amer.*, vol. 114, no. 4, pp. 2112–2117, Oct. 2003.

[30] J. J. Guinan and K. M. Stankovic, "Medial efferent inhibition produces the largest equivalent attenuations at moderate to high sound levels in cat auditory-nerve fibers," *J. Acoust. Soc. Amer.*, vol. 100, no. 3, pp. 1680–1690, Sep. 1996.

[31] R. Meddis, "Auditory-nerve first-spike latency and auditory absolute threshold: A computer model," *J. Acoust. Soc. Amer.*, vol. 119, no. 1, pp. 406–417, Jan. 2006.

[32] M. C. Liberman, "Response properties of cochlear efferent neurons: Monaural vs. Binaural stimulation and the effects of noise," *J. Neurophysiol.*, vol. 60, no. 5, pp. 1779–1798, Nov. 1988.

[33] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, and X. Liu, "The HTK book (for HTK version 3.5)," Dept. Eng., Univ. Cambridge, Cambridge, U.K., Tech. Rep. 679, 2015.

[34] L. F. De and K. Martinez, "Enhancing timbre model using MFCC and its time derivatives for music similarity estimation," in *Proc. 20th Eur. Signal Process. Conf. (AUSIPCO)*, Aug. 2012, pp. 2005–2009.

[35] R. Leonard, "A database for speaker-independent digit recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process. (ICASSP)*, Mar. 1984, pp. 328–331.

[36] N. Kanedera, H. Hermansky, and T. Arai, "On properties of modulation spectrum for robust automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process., (ICASSP)*, May 1998, pp. 613–616.

[37] S. Greenberg, J. Hollenback, and D. Ellis, "Insights into spoken language gleaned from phonetic transcription of the switchboard corpus," in *Proc. Int. Conf. Spoken Lang. Process.*, 1996, pp. 24–27.

[38] H. H. Schüt, S. Harmeling, J. H. Macke, and F. A. Wichmann, "Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data," *Vis. Res.*, vol. 122, May 2016, Art. no. 105123.

[39] Y.-H. Wu, E. Stangl, O. Chipara, S. S. Hasan, A. Welhaven, and J. Oleson, "Characteristics of real-world signal to noise ratios and speech listening situations of older adults with mild to moderate hearing loss," *Ear Hearing*, vol. 39, no. 2, pp. 293–304, 2018.

[40] J. L. Goldstein, "Modeling rapid waveform compression on the basilar membrane as multiple-bandpass-nonlinearity filtering," *Hearing Res.*, vol. 49, nos. 1–3, pp. 39–60, Nov. 1990.

[41] M. Robertson, G. J. Brown, W. Lecluyse, M. Panda, and C. M. Tan, "A speech-in-noise test based on spoken digits: Comparison of normal and impaired listeners using a computer model," in *Proc. Interspeech*, Makhuhari, Japan, 2010, pp. 2470–2473.

[42] V. Drga, C. J. Plack, and I. Yasin, "Frequency tuning of the efferent effect on cochlear gain in humans," in *Physiology, Psychoacoustics and Cognition in Normal and Impaired Hearing*, P. van Dijk, D. Baskent, E. Gaudrain, E. de Kleine, A. Wagner, and C. Lanting, Eds. Heidelberg, Germany: Springer-Verlag, 2016, pp. 477–484.

[43] J. L. Verhey, M. Kordus, V. Drga, and I. Yasin, "Effect of efferent activation on binaural frequency selectivity," *Hearing Res.*, vol. 350, pp. 152–159, Jul. 2017.

[44] J. L. Verhey and I. Yasin, "Effect of duration and gating of the signal on the binaural masking level difference for narrowband and broadband maskers," *J. Acoust. Soc. Amer.*, vol. 142, no. 3, pp. EL258–EL263, Sep. 2017.

[45] R. Drullman, "Temporal envelope and fine structure cues for speech intelligibility," *J. Acoust. Soc. Amer.*, vol. 95, no. 5, p. 3009, May 1994.

[46] A. L. James, R. J. Mount, and R. V. Harrison, "Contralateral suppression of DPOAE measured in real time," *Clin. Otolaryngol. Allied Sci.*, vol. 27, no. 2, pp. 106–112, Apr. 2002.
[47] M. I. Marrufo-Pérez, A. Eustaquio-Martín, L. E. López-Bascuas, and E. A. Lopez-Poveda "Temporal effects on monaural amplitude-modulation sensitivity in ipsilateral, contralateral and bilateral noise," *J. Assoc. Res. Otolaryngol.*, vol. 19, pp. 147–161, Mar. 2018.
[48] C. Abdala, S. Dhar, M. Ahmadi, and P. Luo, "Aging of the medial olivocochlear reflex and associations with speech perception," *J. Acoust. Soc. Amer.*, vol. 135, no. 2, pp. 754–765, 2014.
[49] S. Maruthy, U. A. Kumar, and G. N. Gnanateja, "Functional interplay between the putative measures of rostral and caudal efferent regulation of speech perception in noise," *J. Assoc. Res. Otolaryngol.*, vol. 18, no. 4, pp. 635–648, Aug. 2017.
[50] E. A. Lopez-Poveda and A. Eustaquio-Martín, "Objective speech transmission improvements with a binaural cochlear implant sound-coding strategy inspired by the contralateral medial olivocochlear reflex," *J. Acoust. Soc. Amer.*, vol. 143, no. 4, pp. 2217–2231, Apr. 2018.
[51] S. Maison, C. Micheyl, and L. Collet, "Medial olivocochlear efferent system in humans studied with amplitude-modulated tones," *J. Neurophysiol.*, vol. 77, no. 4, pp. 1759–1768, 1997, doi: 10.1152/jn.1997.77.4.1759.
[52] L. H. Carney, "Supra-threshold hearing and fluctuation profiles: Implications for sensorineural and hidden hearing loss," *J. Assoc. Res. Otolaryngol.*, vol. 19, no. 4, pp. 331–352, Aug. 2018.
[53] S. Maison, J. Durrant, C. Gallineau, C. Micheyl, and L. Collet, "Delay and temporal integration in medial olivocochlear bundle activation in humans," *Ear Hearing*, vol. 22, no. 1, pp. 65–74, Feb. 2001.

IFAT YASIN received the Ph.D. degree in psychology (on the topic of psychoacoustics) from the University of Essex, U.K. He is currently an Associate Professor with the Department of Computer Science, University College London, U.K. His current research interests include applications of psychoacoustic principles for signal processing applications, audio device design, understanding perception, and virtual environments.

VIT DRGA received the Ph.D. degree in psychology from the Victoria University of Wellington, New Zealand, in 2000. He held academic research positions at the University of Essex, U.K., and St Andrews University, U.K. He is currently a Senior Research Fellow with the Computer Science Department, University College London, U.K. His research interests include psychophysics, human signal detection, and sensory processing. He is also involved in modeling auditory cochlear processes.

FANGQI LIU received the B.Sc. degree from the Beijing University of Chemical Technology, Beijing, China, in 2012, the M.Sc. degree from the University of Leicester, Leicester, U.K., in 2014, and the Ph.D. degree in electronic and electrical engineering from University College London (UCL), London, U.K., in 2019. His research focused on audio signal processing with UCL. He is currently a Postdoctoral Research Fellow with the Department of Electronic and Electrical Engineering, UCL. His current researches focus on audio signal processing and biomedical interface system design.

ANDREAS DEMOSTHENOUS (Fellow, IEEE) received the B.Eng. degree in electrical and electronic engineering from the University of Leicester, Leicester, U.K., the M.Sc. degree in telecommunications technology from Aston University, Birmingham, U.K., and the Ph.D. degree in electronic and electrical engineering from University College London (UCL), London, U.K., in 1992, 1994, and 1998, respectively. He is currently a Professor with the Department of Electronic and Electrical Engineering, UCL. He leads the Analog and Biomedical Electronics Group. He has made outstanding contributions to improving safety and performance in integrated circuit design for active medical devices, such as spinal cord and brain stimulators. He has numerous collaborations for cross-disciplinary research, within the U.K. and internationally. He has authored over 300 articles in journals and international conference proceedings, several book chapters, and holds several patents. His research interests include analog and mixed-signal integrated circuits for biomedical, sensor, and signal processing applications. He is a Fellow of the Institution of Engineering and Technology and a Chartered Engineer. He was a co-recipient of a number of best paper awards and has graduated many Ph.D. students. He has served on the technical committees for a number of international conferences, including the European Solid-State Circuits Conference (ESSCIRC) and the International Symposium on Circuits and Systems (ISCAS). He was an Associate Editor, from 2006 to 2009, the Deputy Editor-in-Chief of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—II: EXPRESS BRIEFS, from 2014 to 2015, and the Editor-in-Chief of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: REGULAR PAPERS, from 2016 to 2019. He is an Associate Editor of the IEEE TRANSACTIONS ON BIOMEDICAL CIRCUITS AND SYSTEMS. He serves on the International Advisory Board of Physiological Measurement.

RAY MEDDIS received the Ph.D. degree in psychology from the University of London, U.K., in 1969. He held academic posts at London University, Loughborough University, and the University of Essex, U.K. He was the Head of the Psychology Department and the Director of the Hearing Laboratory, University of Essex. His primary interest lay in developing computational models of the auditory system as a basis for evaluating theories of human hearing.