A frequency-selective feedback model of auditory efferent suppression and its implications for the recognition of speech in noise

Nicholas R. Clark
Department of Psychology, University of Essex, Colchester CO4 3SQ, United Kingdom

Guy J. Brown
Department of Computer Science, University of Sheffield, Sheffield S1 4DP, United Kingdom

Tim Jürgens and Ray Meddis[a)]
Department of Psychology, University of Essex, Colchester CO4 3SQ, United Kingdom

(Received 11 May 2012; revised 20 July 2012; accepted 21 July 2012)

The potential contribution of the peripheral auditory efferent system to our understanding of speech in a background of competing noise was studied using a computer model of the auditory periphery and assessed using an automatic speech recognition system. A previous study had shown that a fixed efferent attenuation applied to all channels of a multi-channel model could improve the recognition of connected digit triplets in noise [G. J. Brown, R. T. Ferry, and R. Meddis, J. Acoust. Soc. Am. **127**, 943–954 (2010)]. In the current study an anatomically justified feedback loop was used to automatically regulate separate attenuation values for each auditory channel. This arrangement resulted in a further enhancement of speech recognition over fixed-attenuation conditions. Comparisons between multi-talker babble and pink noise interference conditions suggest that the benefit originates from the model's ability to modify the amount of suppression in each channel separately according to the spectral shape of the interfering sounds. © *2012 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4742745]

PACS number(s): 43.64.Bt, 43.71.Rt [BLM] Pages: 1535–1541

I. INTRODUCTION

Human speech recognition is remarkably robust to the effects of background noise and the physiological mechanisms underlying this ability are only partially understood. Physiological studies (Delgutte and Kiang, 1984; Geisler and Gamble, 1989) and computer models of peripheral auditory processing (Holmberg *et al.*, 2007) have been used to investigate the effects of noise on the representation of speech in the auditory nerve, but typically, such studies have only considered afferent processing. There is increasing interest in the role of efferent auditory pathways that regulate the afferent processing of signals within the periphery. In particular, the efferent fibers of the medial olivocochlear complex (MOC) synapse on outer hair cells (OHCs) within the cochlear partition, inhibiting basilar membrane displacement when activated (Russell and Murugasu, 1997). It has been hypothesized that efferent suppression originating from the MOC may be beneficial when listening in noisy environments (see, for review, Guinan, 2010).

The effects of efferent suppression on the representation of sounds in the auditory nerve have been modeled using the dual-resonance non-linear (DRNL) model of cochlear function as a basis (Ferry and Meddis, 2007). In the DRNL model, the displacement of a point along the cochlear partition is modeled by parallel linear and nonlinear signal processing pathways. A bank of such DRNL filters can be used to model basilar membrane (BM) displacement over a frequency region of interest. The linear pathway of the DRNL simulates the passive mechanical properties of the BM, whereas the non-linear pathway models the active properties of the membrane associated with the action of the OHCs. Ferry and Meddis (2007) were able to model the effects of efferent suppression by attenuating the input to the nonlinear pathway of each DRNL filter, thus simulating a reflexive MOC-induced reduction in OHC motility. In their model, efferent suppression was applied in a frequency-independent manner (i.e., the same attenuation was applied to all DRNL filters regardless of their best frequency).

By using the model described previously as the front-end processor for an automatic speech recognition (ASR) system, Brown *et al.* (2010) predicted the potential effects of MOC efferent activity on speech intelligibility. They systematically adjusted the amount of efferent attenuation applied in the DRNL model, and found that the recognition of speech in noise was improved when efferent attenuation was applied. Relatively poor recognition performance was obtained when efferent activity was disabled. Additionally, they found that optimum recognition performance was obtained when the amount of attenuation was proportional to the background noise level. They explained their results in terms of the limited dynamic range of the auditory periphery. Background noise reduces the dynamic range available to represent a signal (such as speech) in the auditory nerve, both because it introduces a noise floor above the

[a)]Author to whom correspondence should be addressed. Electronic mail: rmeddis@essex.ac.uk

spontaneous rate, and because it causes adaptation that reduces the maximum firing rate. By introducing an efferent attenuation proportional to the noise level, the operating point of the cochlear was moved so that much of the background noise fell below threshold. This reduced the noise floor and provided a release from adaptation (Brown *et al.*, 2010), increasing the availability of dynamic range to represent the amplitude fluctuations of the speech. At positive signal-to-noise ratios (SNRs) the resulting spectro-temporal representation of speech was partially de-noised, leading to improved ASR performance.

Messing *et al.* (2008, 2009) presented an alternative model simulating MOC suppression based on Goldstein's (1990) multi-band-pass non-linear (MBPNL) model of cochlear mechanics, in which, the gain and bandwidth of the filters change with changes in input intensity. To simulate the effect of efferent suppression, Messing *et al.* made the tip gain of the MBPNL filters dependent on the level of background noise in each frequency band. This was realized by adaptively adjusting the tip gain of the cochlear filters in response to noise alone, until a predetermined amount of noise energy was measurable at the output of the auditory model. The gains were applied in a frequency dependent manner. After finding the gains for the specific noise level under test, the gains were fixed. The speech material was then added to the noise and processed by the model before subsequent evaluation. Messing *et al.* (2009) evaluated their auditory model by attempting to replicate human consonant confusions in noise. Their model including MOC suppression was able to produce a better match to the human data than the version of the model without MOC processing.

Subsequently, Lee *et al.* (2011) modified the auditory model of Messing *et al.* (2008, 2009) to extract features suitable for ASR. This enabled the model to be evaluated on a digit recognition task, similar to that used by Brown *et al.* (2010). Different noise backgrounds were added to the digits at a range of SNRs to assess the performance of their model. The spectral characteristics of the noises resulted in different tip gain profiles, which were fixed for each noise separately (Lee, 2010). Their feature extraction method based on MOC suppression gave good recognition accuracy across all types of noise tested, when compared to other front ends that did not include MOC-type processing.

The aim of the current study is to address some of the limitations of these previous models, in order to more closely simulate the physiology and better understand the effects of efferent suppression on speech perception in noisy environments. A limitation of the study by Brown *et al.* (2010) is that their model was "open loop," i.e., the amount of efferent attenuation was determined by the experimenter rather than estimated from the noisy signal. In reality, the efferent system is a continuously operating, closed-loop system involving afferent fibers from the cochlea and MOC efferents (Guinan, 2006).

In the model used by Brown *et al.* (2010), efferent suppression was fixed by the experimenter and applied uniformly across frequency. Hence, their model cannot effectively regulate the cochlear operating point in each frequency channel when the noise background is non-stationary. Conversely, the model of Messing *et al.* (2008) based on the MBPNL implements efferent suppression in a frequency selective manner. However, their model does not represent a fully closed-loop system, in which efferent suppression is mediated through a continuous feedback process. As a result, neither model would function optimally if the noise-spectrum, or overall noise-level were to change (i.e., if the simulated listener were to step from a noisy room into a quiet one).

In the current study, we present a new model in which efferent suppression is regulated by a continuous feedback process, based on the physiological data measured by Liberman (1988). The model is evaluated using an ASR procedure broadly similar to that used by Brown *et al.* (2010). It will be shown that the ASR performance of the new (closed-loop) model is similar to that obtained by the open-loop model of Brown *et al.* when the noise background is stationary (pink noise). We emphasize that ASR is used here as a means of approximating the effects of putative efferent processing mechanisms on human speech recognition. Our aim is not to build a noise-robust ASR system per se, but to better understand the role of reflexive MOC suppression by closely modeling the underlying physiology.

II. METHOD

A computer model of the auditory periphery was modified to simulate efferent suppression of the BM response. The amount of suppression was calculated on the basis of the rate of firing of model auditory nerve fibers in the corresponding best-frequency channel. The new model was evaluated by using it as a front-end to a standard automatic speech recognition system. The recognizer was trained in silence and tested in noise of different intensities. The aim was to evaluate the potential benefit of a closed-loop, efferent system operating on a within-channel basis for recognizing speech against a background of competing sounds.

To understand how this benefit comes about, we compared the model performance using two types of interfering backgrounds; multi-talker babble, and pink noise. It was expected that the babble would show greater benefit because its excitation pattern contains distinctive spectral peaks and valleys that are largely absent from pink noise.

Previous studies have used "optimal" amounts of suppression based on the results of many trials. This artificial arrangement demonstrates the general principle of benefit as a consequence of efferent suppression. The closed-loop, multi-channel arrangement to be tested below is an attempt to show how this optimum suppression might be achieved in a biological system. The size of the additional benefit from the new model was, therefore, assessed by comparing it with an additional condition using an optimal fixed-level applied to all channels. A separate optimal level of suppression was found for each SNR condition using exhaustive search. The aim of this comparison was to allow the study to identify whether a self-regulating system could find a solution at least as good as a "hand-tuned" version.

A final comparison condition used a fixed 10 dB attenuation applied equally across all channels. This provides a

comparison with an earlier study and offered an overall indication of improved speech recognition in noise as a result of the increased sophistication of the biological simulation.

III. COMPUTER MODEL

The computer model of the auditory periphery used in this study consists of a cascade of modules representing the resonances of the outer/middle ear (OME), the response of the BM, coupling by inner hair cell (IHC) stereocilia, the IHC receptor potential, calcium dynamics, and transmitter release and adaptation at the IHC auditory-nerve synapse. Detailed discussions regarding the implementation and evaluation of each of these stages can be found in Meddis *et al.* (2001); Lopez-Poveda and Meddis (2001); Sumner *et al.* (2002); Sumner *et al.* (2003a, 2003b); Meddis (2006). The OME stage of the model was configured using data from Huber *et al.* (2001). A MATLAB implementation of the model is available upon request. The final stage of the model produces a time by channel representation of firing rates in the auditory nerve (AN).

The response of the BM is simulated using the DRNL model (Sumner *et al.*, 2003b). It receives stapes displacement as an input from a model of the OME, and produces BM displacement as its output. This, in turn, is used to drive a simulation of IHC function. The nonlinear path contains an attenuation stage, proposed by Ferry and Meddis (2007) to model the effect of efferent suppression from the MOC, followed by a broken-stick instantaneous compression operation associated with outer hair cell function. The model by Ferry and Meddis (2007) of the periphery including the modification to the DRNL was able to produce data that are in good agreement with physiological measurements of the BM, AN, and compound action potential responses when the amount of attenuation in decibels was chosen to be proportional to the amount of MOC activity. The model used in the current study extends the one proposed by Ferry and Meddis (2007) by implementing a feedback signal that dynamically controls the amount of attenuation according to the recent history of AN activity, i.e., a fully closed loop. A schematic of the new processing scheme is shown in Fig. 1.

Physiological studies (Brown, 1989; Liberman and Brown, 1986) have shown that MOC fibers have tuning curves that are only slightly wider than cochlear afferent fibers. In humans, MOC tuning curves measured using otoacoustic emission techniques (Lilaonitkul and Guinan, 2009) have shown similarly narrow frequency selectivity. To reduce the number of free parameters in the current study, we assumed equal bandwidth of MOC efferents and AN tuning curves (i.e., all efferent processing was performed on a within-channel basis). This stimulus-dependent, frequency-selective attenuation has parallels with the model proposed by Messing *et al.* (2008). However, in our model the efferent attenuation due to the MOC is updated continuously in non-overlapping frames of samples, rather than being determined in a *post hoc* analysis.

Liberman (1988) measured rate-level functions in efferent neurons in response to tone bursts presented at their characteristic frequency (CF). Their data are reproduced in

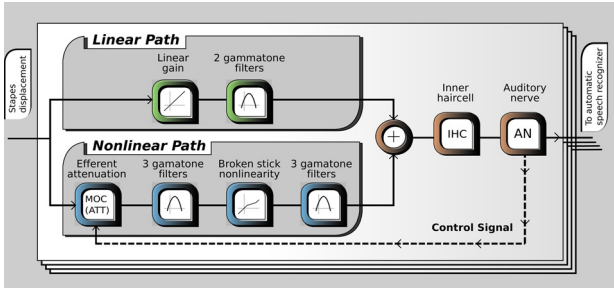

FIG. 1. (Color online) Schematic diagram of the DRNL filterbank and subsequent neural transduction, including the novel feedback mechanism introduced in the current study. The suppressive role of the MOC is modeled by inserting an attenuator at the input to the nonlinear path of the DRNL model, and the amount of attenuation is modulated by a control signal (represented by the dashed line) derived from the recent history of the AN response. The stacked panels highlight the fact that this process occurs independently within each frequency channel.

Fig. 2. The efferent units did not exhibit any spontaneous activity. However, once the tone intensity had risen above the threshold of the unit, the discharge rate increased approximately linearly with logarithmic increases in tone intensity until an upper saturation point was reached. Liberman's data show that the MOC reflex can be driven by low intensity stimuli with typical thresholds in the region of 20–40 dB sound pressure level (SPL) for CF $<$ 6 kHz. To achieve a

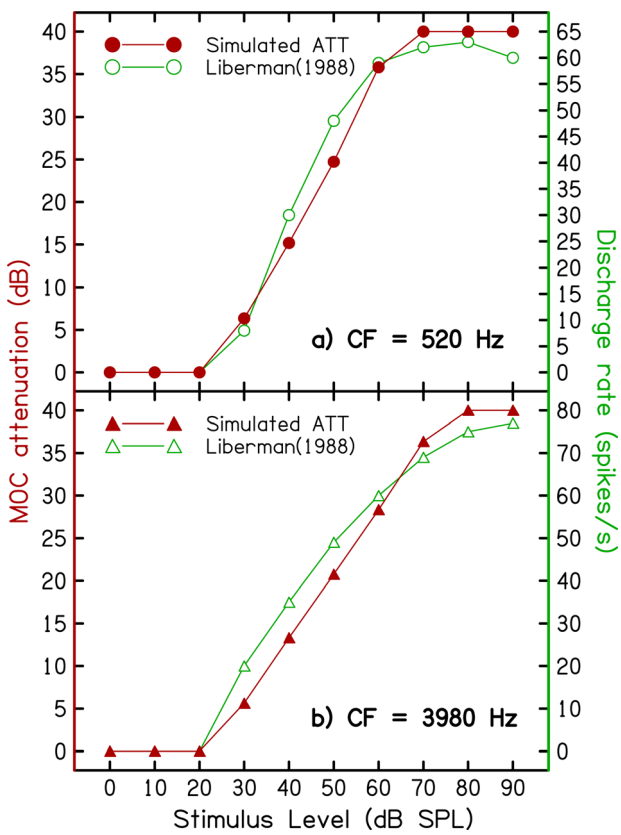

FIG. 2. (Color online) Efferent activity plotted against stimulus level for two efferent fibers with different best frequencies (Liberman, 1988) (open symbols). Model output (filled symbols) is superimposed upon experimental data. The model output is in units of decibel attenuation, plotted as a function of pure tone stimulus level. The numbers on the left-hand axes show the modeled attenuation values. In contrast, the experimental data show the firing rate measured in descending MOC fibers (right-hand axes). The model data were generated using a value of $F = 7$.

corresponding level of sensitivity in the computer model, it was assumed that the firing rate of the MOC unit could map directly to the attenuation (in decibels) applied to the nonlinear path of the DRNL.

The control signal in the feedback loop was based on high spontaneous rate AN fibers (spontaneous rate 56 spikes/s). To replicate the shape of the measured rate-level functions, the control signal was derived from the logarithm of the ratio of the temporally smoothed firing rate (x), to a firing rate threshold (T) as given by

$$\mathrm{ATT}(t) = \begin{cases} F \times 20\log_{10}\left(\dfrac{x(t-\tau)}{T}\right), & x(t-\tau) > T \\ 0, & x(t-\tau) \leq T. \end{cases} \quad (1)$$

T was set to 85 spikes/s throughout in order to provide the fit to the Liberman data. The instantaneous AN firing rate was smoothed using a first-order lowpass filter with a time constant of 2 s (see discussion). The lag (τ) of 10 ms was introduced to account for MOC-OHC synaptic minimum latencies estimated by Liberman (1988) to be between 5 and 40 ms. The dB attenuation applied (ATT) at a point in time (t) was calculated by multiplying the resulting control signal calculated τ ms previously with a scalar rate-to-attenuation factor (F). The value of F was derived from the assumption that the maximum attenuation was 40 dB. Negative values of ATT (occurring when the firing rate was below threshold) were set to zero.

This function was able to produce a good qualitative fit to Liberman's data (Fig. 2). The differences between the model functions at CFs of 520 and 3980 Hz are emergent properties of the model, that are accounted for by a combination of outer-middle ear processing and the dependence of BM displacement on frequency. It is important to note that the attenuation applied to the input of the nonlinear path is not the same as the resulting attenuation of the output. When compression is at work, the reduction in output is considerably less (up to 5 times less for a compression of 5:1).

IV. EVALUATION

A. Corpus

The task chosen for evaluation of the new model was identification of connected-digit triplets in the presence of background noise, similar to that employed by Brown *et al.* (2010). This allowed us to compare the model results with human performance on the same task (Robertson *et al.*, 2010). Speech material for the following experiments was drawn from the TIDIGITS corpus (Leonard, 1984), which consists of sequences of between one and seven digits spoken by male and female talkers. Three sets of utterances were used. The recognizer was trained on the "clean" training set, which consists of 8440 utterances. For testing the recognizer, we used 358 utterances, each containing three connected digits from the set "oh," "one," "two," "three," "four," "five," "six," "eight," and "nine." The training and testing sets were completely independent, and each contained an approximately equal number of recordings from male and female talkers.

All utterances were scaled to a rms level of 60 dB SPL. Noisy speech was generated by adding either 20-talker babble or pink noise to the test utterances at a range of SNRs between 20 and −10 dB, in steps of 5 dB. The noise waveforms were band-limited to frequencies between 100 Hz and 4 kHz in order to ensure that the SNR was not influenced by noise energy at frequencies outside the speech range. Each test stimulus consisted of 6 s of background followed by speech plus noise through to the end of the speech; this allowed the auditory model to adapt to the background before the onset of the speech. The response to the initial period of noise alone was removed from the AN representation before it was passed to the recognizer.

B. Automatic speech recognizer

The closed-loop MOC model was evaluated using a continuous-density hidden Markov model (HMM) system, implemented using the hidden Markov model TOOLKIT HTK (Young *et al.*, 2009). The HMM requires the input signal to be encoded as a temporal sequence of features. The ultimate goal of the recognizer is to find the most probable sequence of digits that correspond to the observed sequence of features.

Features were generated using a similar approach to various other studies that have employed auditory models as acoustic front-end processors for ASR systems (e.g., Jankowski *et al.*, 1995; Holmberg *et al.*, 2007; Brown *et al.*, 2010). Each feature is a vector of coefficients representing the spectrum of the stimulus within a fixed time window. To generate a sequence of feature vectors with a temporal resolution that is typical for ASR systems, the auditory-nerve firing probability emanating from each channel of the model was integrated over a 25 ms Hann window at intervals of 10 ms (60% overlap). The output of adjacent frequency channels from the auditory model are highly correlated due to the spectral overlap of the DRNL filters, so a discrete cosine transform was applied to each frame to yield a vector containing approximately independent components. Retaining only the first 14 coefficients reduced the amount of data. First- and second-order regression coefficients, referred to as "deltas" and "accelerations," respectively, were appended to each vector in order to improve recognition performance.

The recognizer represents digits using trained HMMs, where each digit is modeled as a sequence of stationary states. Each state is characterized by a multivariate Gaussian mixture distribution with seven components and diagonal covariance. During training, the Baum–Welch algorithm is used to learn the parameters of the HMMs from a large corpus of annotated files containing digits. During testing, the Viterbi algorithm is applied to find the most likely sequence of HMM states given an observed sequence of feature vectors, thus, returning the most likely sequence of digits. HMMs with 16 emitting states were trained for each word in the corpus. The models for zero and seven were discarded in the testing phase, preventing the recognizer from identifying digits absent from the testing corpus. Three-state models were also trained for non-speech acoustic stimulation. To reduce the number of insertion errors, a simple grammar was

used to constrain all hypotheses so that they started and ended with the non-speech model. The ASR system was always trained on clean speech (i.e., without added noise) and no efferent attenuation was applied during training.

A simplified scoring metric was applied that did not involve use of the HRESULTS tool bundled with HTK. The transcript produced by the recognizer was compared with labels generated from the file names, which identify the digit sequences contained within each file. A correct response was registered only when the recognizer identified the correct digit in the correct position in the triplet, allowing results to be directly compared with a recent psychophysical study (Robertson *et al.* 2010). Results are reported as % correct for each individual digit (i.e., scores are awarded for partially correct triplets).

V. RESULTS

The results in the form of % digits correct for all conditions and for both types of interfering noise are given in Fig. 3. The performance of the original model (with no efferent simulation of any type) is shown as open squares. For both multi-talker babble and pink noise, performance is almost perfect for speech tested in silence. However, in background noise performance is poor with 50% digits correct obtained at relatively low noise levels. This compares adversely with human performance for the same test stimuli, shown as the open circles at the extreme left of the upper panel Robertson *et al.* (2010). The difference at the 50% point of the function is approximately 25 dB SNR.

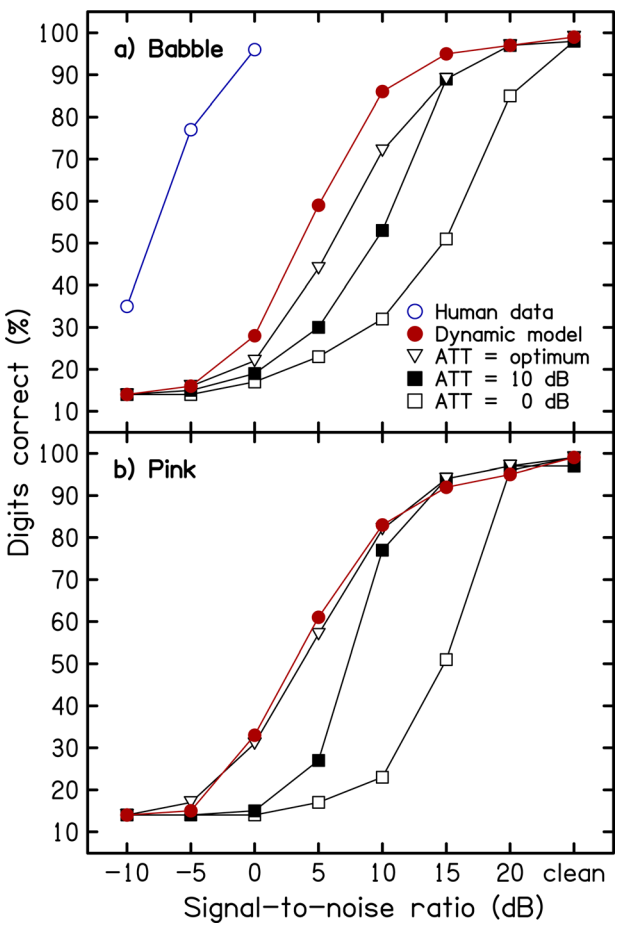

FIG. 3. (Color online) Data showing ASR performance (% correct) as a function of SNR, where the parameter is the type of efferent suppression used in the auditory front end. The noise used for the data shown in the top panel was babble, whereas the noise used for the data shown in the bottom panel was pink. Human data from Robertson *et al.* (2010) for the same task (monaural presentation of spoken digits in a background of 16-talker babble) are shown as open circles to highlight the current human–machine performance gap.

A fixed 10 dB efferent attenuation (closed squares) results in an improvement of ~10 dB SNR at 50% correct for pink noise. This is a replication of the same effect demonstrated by Brown *et al.* (2010). The effect is greater for pink noise than for babble. Pink noise produces a relatively flat excitation pattern when compared with multi-talker babble, which has distinct peaks and valleys at different frequencies. A fixed efferent attenuation across all channels will, therefore, offer a better match to pink noise.

The open inverted triangles show further improvements when the fixed across-channel attenuation is optimized by exhaustive search to find the best attenuation for each SNR value. Brown *et al.* (2010) had shown that the optimum attenuation was sensitive to overall speech and noise levels. These data give a further indication of this effect. Once again, performance is best in pink noise and this presumably reflects the fact that interfering-noise with a flat excitation pattern is most effectively dealt with by a fixed, across-channel attenuation.

Finally, the closed circles show the performance of the closed-loop, efferent-feedback model with separate levels of attenuation in each channel. Little further improvement is evident for the pink noise interference condition when compared with the optimal, fixed level condition. However, further improvement can be seen for the multi-talker babble condition. This result is consistent with the suggestion that there is some advantage to be gained from a system that dynamically modifies its pattern of attenuation across channels to match the spectral profile of interfering background sound.

The full benefit of the within-channel variable attenuation algorithm (closed circles) can be seen by comparing with the "no-efferent" condition (open squares). This amounts to a benefit of 10 dB reduction in SNR at the 50% correct point.

VI. DISCUSSION

This paper has described and evaluated a new model of auditory efferent processing. The new model represents an advance over the previous model of Ferry and Meddis (2007), in that efferent attenuation is controlled dynamically by a feedback loop. This allows context-dependent control of efferent attenuation, such that the amount of efferent suppression depends on the preceding acoustic stimulation. Further, efferent attenuation is applied independently within each frequency channel of the model.

Evaluation of the model on an ASR task (recognition of digit triplets in pink noise and babble) led to two main findings. First, the feedback mechanism in the model allows it to autonomously determine a near-optimal value of efferent attenuation (i.e., a value that maximizes ASR performance). Previously, Brown *et al.* (2010) showed that the amount of efferent attenuation required to maximize speech recognition

performance at each SNR tested was proportional to the level of the noise. The amount of attenuation applied by the new model is proportional to the level of the noise by virtue of its feedback design, thus satisfying the criterion for good performance. Second, the new model provides an advantage when the background noise is non-stationary. The ASR performance of the closed-loop model was similar to that obtained by Brown *et al.* (2010) in a pink noise background, but the closed-loop model gave a notable increase in performance when the background was speech babble. The babble noise used in the current study has a more varied spectral profile than pink noise. A single broadband attenuation value, as used by Brown *et al.* (2010), represents a compromise regarding over- or under-attenuating certain channels in order to yield a satisfactory overall performance. The frequency specificity of the new model avoids this compromise, since an optimal efferent attenuation is derived independently in each frequency channel.

When evaluating the model, it was necessary to choose one of many scenarios concerning the level of noise for training and testing. During development of the model we considered training the ASR system in multiple noise conditions (so-called multi-condition training), or training and testing in matched noise conditions. Neither was found to be satisfactory. Multi-condition training led to better performance in noise, but poorer recognition of clean speech, which should be the easiest condition for a human listener. Training and testing in exactly the same noise conditions led to an overall improvement in ASR performance, but was held to be a poor model of human speech recognition (as ultimately, a different set of acoustic models would be required for each possible SNR and noise type). We therefore trained the recognizer with clean speech. It is doubtful whether this choice affected our final conclusions because all the models considered in this study were evaluated in the same way.

Our approach is functional, but where possible the details of the new model have been based on physiological data. The function relating the amount of efferent attenuation to AN firing rate was derived from the physiological study of Liberman (1988), and the model shows a close match to his data recorded from efferent fibers. As a result, the current study supports the idea that the MOC reflex contributes towards our understanding of speech in adverse listening conditions.

The physiological realism of the model could be further improved by incorporating efferent mechanisms with different time scales. It is thought that efferent suppression operates over distinct fast (10–100 ms) and slow (10–100 s) time scales, which may originate from separate mechanical processes (Cooper and Guinan, 2003), whereas more recent human data (Backus and Guinan, 2006) suggest that there might be three distinct time scales in the regions of 70 ms, 330 ms, and between 11 and 39 s. A more detailed model could have included all three. However, our interest centered on slowly changing background noise levels where a longer time constant would be more appropriate. The chosen time constant of 2 s was a compromise. A large value (i.e., 11 s) would have required longer runs. Whenever longer time constants were tried, they yielded no further improvement in performance. Shorter time constants also, as expected, yielded little benefit in the noise conditions. It is acknowledged that many other improvements could be made to fit the model details more closely to the available physiological data but this must remain a project for the future. There are also many physiological details that remain unknown where the modeler must simply make a reasonable choice. For example, the efferent activity is driven directly from the AN high spontaneous rate response. In practice, brainstem neural activity will intervene and low spontaneous rate fibers may also make a contribution (Winslow *et al.*, 1987; Sachs *et al.*, 2006). A more comprehensive model with spiking neurons was investigated but the computational effort when processing many thousands of speech tokens was so great as to render this approach impractical for the current study. Similarly, not all physiological recordings from efferent fibers show the same rate/level functions and the modeler, again has to choose one representative function. In the event, we chose one of the simplest. The model also assumes that the amount of efferent activity is linearly related in a time-invariant way to the amount of suppression but this is not necessarily so (Ballestero *et al.*, 2011). Further, Lilaonitkul and Guinan (2009) have recently demonstrated the tuning to be offset (i.e., the largest MOC effects were from elicitors half an octave below the probe frequency). Finally, the model assumes that the maximum efferent attenuation is a constant function of frequency, which is at odds with the physiological data (Guinan and Gifford, 1988). All of these and more would increase the verisimilitude of the model and may reveal interesting details of how signal processing is organized at the peripheral level. However, it is unlikely that they would affect the final conclusion that is primarily concerned with the functional benefit of within-channel, dynamic suppression of the peripheral response when speech is presented in continuous background noise.

The aim of the study was to build on the work of Messing *et al.* (2009) and Brown *et al.* (2010) to investigate a general principal that could be at work in human speech recognition. Intriguingly, this principle suggests that the auditory system employs a noise-reduction process that is similar in many respects to spectral subtraction techniques used in noise-robust ASR systems. In such approaches the long-term spectrum of the noise is estimated, and then subtracted from the speech/noise spectrum in order to obtain a cleaner estimate of the target speech signal (Boll, 1979).

For speech recognition in babble noise, the new closed-loop model gives an improvement in ASR performance compared to the system of Brown *et al.* (2010). However, Fig. 3(a) indicates that the gap between human and machine performance is still large. Improvements could be made to the auditory model in order to further close this performance gap. For example, improvements could be gained by encoding the speech signal with spike timing information, instead of (or in addition to) firing rate. A number of studies (e.g., Kim *et al.*, 1999; Sheikhzadeh and Deng, 1998; Brown *et al.*, 2011) have shown that speech is coded more robustly by timing information in noisy conditions than by average firing rate, or other spectrally based features such as Mel frequency cepstral coefficients. If these principles are investigated further it might be possible to move the speech in

noise performance of the model closer to human levels and to develop deeper insights into the underlying mechanisms.

VII. CONCLUSIONS

A novel, fully closed-loop model of MOC function was presented that provides a step towards physiological realism. The model presented outperforms the model used in Brown *et al.* (2010) when evaluated using an automatic speech recognition task in a background of babble noise. This was taken as further evidence to support of the idea that the reflexive efferent system plays an important role when listening in noisy environments. Finally, we note that the proposed efferent model is well-suited to real time implementation, since it does not require hand tuning (c.f. Ferry and Meddis, 2007) or off-line calibration (c.f. Messing *et al.*, 2008, 2009).

ACKNOWLEDGMENT

This work was supported by the EPSRC Grant No. EP/H02828/1.

Boll, S. F. (**1979**). "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust. Speech, Signal Process. **27**, 113–120.

Backus, B. C., and Guinan, J. J., Jr. (**2006**). "Time-course of the human medial olivocochlear reflex," J. Acoust. Soc. Am. **119**, 2889–2904.

Ballestero, J., Zorrilla de San Martin, J., Goutman, J., Elgoyhen, A. B., Fuchs, P. A., and Katz, E. (**2011**). "Short-term synaptic plasticity regulates the level of olivocochlear inhibition to auditory hair cells," J. Neurosci. **31**, 14763–14774.

Brown, G. J., Ferry, R., and Meddis, R. (**2010**). "A computer model of auditory efferent suppression: Implications for the coding of speech in noise," J. Acoust. Soc. Am. **127**, 943–954.

Brown, G. J., Jürgens, T., Meddis, R., Robertson, M., and Clark, N. R. (**2011**). "The representation of speech in a nonlinear auditory model: Time-domain analysis of simulated auditory-nerve firing patterns," *Proceedings of Interspeech*, Italy, pp. 2453–2456.

Brown, M. C. (**1989**). "Morphology and response properties of single olivocochlear fibers in the guinea pig," Hear. Res. **40**, 93–110.

Cooper, N. P., and Guinan, J. J., Jr. (**2003**). "Separate mechanical processes underlie fast and slow effects of medial olivocochlear efferent activity," J. Physiol. **548**, 307–312.

Delgutte, B., and Kiang, N. Y. (**1984**). "Speech coding in the auditory nerve: V. Vowels in background noise," J. Acoust. Soc. Am. **75**, 908–918.

Ferry, R., and Meddis, R. (**2007**). "A computer model of medial efferent suppression in the mammalian auditory system," J. Acoust. Soc. Am. **122**, 3519–3526.

Geisler, C. D., and Gamble, T. (**1989**). "Responses of 'high-spontaneous' auditory-nerve fibers to consonant-vowel syllables in noise," J. Acoust. Soc. Am. **85**, 1639–1652.

Goldstein, J. (**1990**). "Modeling rapid waveform compression on the basilar membrane as multiple-bandpass-nonlinearity filtering," Hear. Res. **49**, 39–60.

Guinan, J. J., Jr. (**2006**). "Olivocochlear efferents: Anatomy, physiology, function, and the measurement of efferent effects in humans," Ear Hear. **27**, 589–607.

Guinan, J. J., Jr. (**2010**). "Cochlear efferent innervation and function," Curr. Opin. Otolaryngol. Head Neck Surg. **18**, 447–453.

Guinan, J. J., Jr., and Gifford, L. (**1988**). "Effects of electrical stimulation of efferent olivocochlear neurons on cat auditory-nerve fibers. III. Tuning curves and thresholds," Hear. Res. **37**, 29–46.

Holmberg, M., Gelbart, D., and Hemmert, W. (**2007**). "Speech encoding in a model of peripheral auditory processing: Quantitative assessment by means of automatic speech recognition," Speech Commun. **49**, 917–932.

Huber, A., Linder, T., Ferrazzini, M., Schmid, S., Dillier, N., Stoeckli, S., and Fisch, U. (**2001**). "Intraoperative assessment of stapes movement," Ann. Otol. Rhinol. Laryngol. **110**, 31–35.

Jankowski, C. R. J., Vo, H.-D., and Lippmann, R. P. (**1995**). "A comparison of signal processing front ends for automatic word recognition," IEEE Trans. Speech Audio Proc. **3**, 286–293.

Kim, D.-S., Lee, S.-Y., and Kil, R.-M. (**1999**). "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," IEEE Trans. Speech Audio Proc. **7**, 55–69.

Lee, C. (**2010**). "Closed-loop auditory-based representation for robust speech recognition," M.Sc. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, pp. 1–96.

Lee, C., Glass, J., and Ghitza, O. (**2011**). "An efferent-inspired auditory model front-end for speech recognition," *Proceedings of Interspeech*, Florence, Italy, pp. 49–52.

Leonard, R. G. (**1984**). "A database for speaker-independent digit recognition," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, San Diego, pp. 328–331.

Liberman, M. C. (**1988**). "Response properties of cochlear efferent neurons: Monaural versus binaural stimulation and the effects of noise," J. Neurophys. **60**, 1779–1789.

Liberman, M. C., and Brown, M. C. (**1986**). "Physiology and anatomy of single olivocochlear neurons in the cat," Hear. Res. **24**, 17–36.

Lilaonitkul, W., and Guinan, J. J., Jr. (**2009**). "Reflex control of the human inner ear: A half-octave offset in medial efferent feedback that is consistent with an efferent role in the control of masking," J. Neurophys. **101**, 1394–1406.

Lopez-Poveda, E. A., and Meddis, R. (**2001**). "A human nonlinear cochlear filterbank," J. Acoust. Soc. Am. **110**, 3107–3118.

Meddis, R. (**2006**). "Auditory-nerve first-spike latency and auditory absolute threshold: A computer model," J. Acoust. Soc. Am. **119**, 406–417.

Meddis, R., O'Mard, L., and Lopez-Poveda, E. (**2001**). "A computational algorithm for computing nonlinear auditory frequency selectivity," J. Acoust. Soc. Am. **109**, 2852–2861.

Messing, D. P., Delhorne, L., Bruckert, E., Braida, L., and Ghitza, O. (**2008**). "Consonant discrimination of degraded speech using an efferent-inspired model closed-loop cochlear model," *Proceedings of Interspeech*, Brisbane, Australia, pp. 1052–1055.

Messing, D. P., Delhorne, L., Bruckert, E., Braida, L., and Ghitza, O. (**2009**). "A non-linear efferent-inspired model of the auditory system; matching human confusions in stationary noise," Speech Commun. **51**, 668–683.

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

Russel, I. J., and Murugasu, E. (**1997**). "Medial efferent inhibition suppresses basilar membrane responses to near characteristic frequency tones of moderate to high intensities," J. Acoust. Soc. Am. **102**, 1734–1738.

Sachs, M. B., May, B. J., Prell, G. S. L., and Heinz, R. D. (**2006**). "Adequacy of auditory-nerve rate representations of vowels: Comparison with behavioural measures in cat," in *Listening to Speech: An Auditory Perspective*, edited by S. Greenberg and W. A. Ainsworth (Lawrence Erlbaum Associates, Hillsdale, NJ), pp. 115–127.

Sheikhzadeh, H., and Deng, L. (**1998**). "Speech analysis and recognition using interval statistics generated from a composite auditory model," IEEE Trans. Speech Audio Proc. **6**, 90–94.

Sumner, C., Lopez-Poveda, E., O'Mard, L., and Meddis, R. (**2002**). "A revised model of the inner-hair cell and auditory-nerve complex," J. Acoust. Soc. Am. **111**, 2178–2189.

Sumner, C., Lopez-Poveda, E., O'Mard, L., and Meddis, R. (**2003a**). "Adaptation in a revised inner-hair cell model," J. Acoust. Soc. Am. **113**, 893–901.

Sumner, C., O'Mard, L., Lopez-Poveda, E., and Meddis, R. (**2003b**). "A nonlinear filterbank model of the guinea-pig cochlear nerve: Rate responses," J. Acoust. Soc. Am. **113**, 3264–3274.

Winslow, R. L., Barta, P. E., and Sachs, M. B. (**1987**). "Rate coding in the auditory nerve," in *Auditory Processing of Complex Sounds*, edited by W. A. Yost and C. S. Watson (Lawrence Erlbaum Associates, Hillsdale, NJ), pp. 212–224.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2009). "The Hidden Markov Model Toolkit (HTK)," Engineering Department, University of Cambridge, Cambridge, UK, http://htk.eng.cam.ac.uk/ (last viewed 08/05/12).