A computer model of auditory efferent suppression: Implications for the recognition of speech in noise

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

Robert T. Ferry and Ray Meddis
Department of Psychology, University of Essex, Colchester CO4 3SQ, United Kingdom

(Received 11 June 2009; revised 4 November 2009; accepted 20 November 2009)

The neural mechanisms underlying the ability of human listeners to recognize speech in the presence of background noise are still imperfectly understood. However, there is mounting evidence that the medial olivocochlear system plays an important role, via efferents that exert a suppressive effect on the response of the basilar membrane. The current paper presents a computer modeling study that investigates the possible role of this activity on speech intelligibility in noise. A model of auditory efferent processing [Ferry, R. T., and Meddis, R. (2007). J. Acoust. Soc. Am. **122**, 3519–3526] is used to provide acoustic features for a statistical automatic speech recognition system, thus allowing the effects of efferent activity on speech intelligibility to be quantified. Performance of the "basic" model (without efferent activity) on a connected digit recognition task is good when the speech is uncorrupted by noise but falls when noise is present. However, recognition performance is much improved when efferent activity is applied. Furthermore, optimal performance is obtained when the amount of efferent activity is proportional to the noise level. The results obtained are consistent with the suggestion that efferent suppression causes a "release from adaptation" in the auditory-nerve response to noisy speech, which enhances its intelligibility.
 [DOI: 10.1121/1.3273893]

PACS number(s): 43.64.Bt, 43.71.Rt [WPS] Pages: 943–954

I. INTRODUCTION

The detection of communication sounds against a background of environmental noise is a fundamental problem that affects many animal species. Among humans, this problem is particularly acute for listeners with impaired hearing, who frequently complain of difficulties in hearing speech in noisy places such as offices, shops, bars, and restaurants. An understanding of the mechanisms that underlie the ability of normal human listeners to recognize speech in the presence of background noise is therefore of considerable theoretical interest, and has an important practical application in the development of aids for the hearing impaired.

Our understanding of hearing is based mainly on our knowledge of the afferent system, where speech sounds are processed and passed through the auditory nervous system in the direction of the cerebral cortex. However, there have been numerous recent suggestions that the efferent system may make an important contribution (see Guinan, 1996, 2006 for reviews). The efferent system consists of nerve fibers whose direction of information flow appears to be away from the cortex. The most peripheral part of the auditory efferent system consists of fibers in the auditory nerve that project from the brainstem to the cochlea itself. It is now generally agreed that one component, the medial olivocochlear (MOC) system, indirectly exerts a suppressive influence on the response of the basilar membrane (BM) to sounds (Dallos, 1992). This has the effect of shifting the auditory-nerve (AN) rate/level function toward higher sound levels, and may be a way of extending the dynamic range of the auditory system. This effect has been recorded in small mammals at the level of the basilar membrane (Dolan *et al.*, 1997; Russell and Murugasu, 1997; Cooper and Guinan, 2006), the auditory nerve (Wiederhold and Kiang, 1970; Guinan and Stankovic, 1996), and the compound action potential (Winslow and Sachs, 1988; Dolan and Nuttall, 1988).

In addition, it has been suggested that the efferent system confers robustness to noise. For example, Dolan and Nuttall (1988) suggested that the activity of the efferent system may increase the detectability of tones in noise. They demonstrated that the compound action potential (CAP) response to a tone in noise was enhanced when the crossed olivocochlear bundle (OCB) was electrically stimulated. The mechanism responsible for this effect is complex, but is likely to involve adaptation. In prolonged background noise, the auditory nerve response becomes adapted and less responsive to new sounds. The efferent system has the potential to reduce the response to the continuous noise, reduce adaptation, and, therefore, enhance the response to a new sound presented in that background. Liberman and Guinan (1998) suggested that when the noise is continuous but the signal is transient, the MOC reflex acts to minimize the response to long-lasting stimuli while maximizing the response to novel stimuli. Other effects related to level-dependent compression may also be involved (Russell and Murugasu, 1997).

[a)]Author to whom correspondence should be addressed. Electronic mail: g.brown@dcs.shef.ac.uk

Relatively few studies give direct support to the idea that the efferent system contributes to the intelligibility of speech in noise. May and McQuone (1995) and Hienz *et al.* (1998) found that severing the olivocochlear bundle in cats reduced performance in tasks involving formant discrimination or intensity discrimination of tones in noise. Dewson (1968) showed that MOC lesions impair the ability of monkeys to discriminate vowel sounds when presented in noise, but have no effect on discrimination in silence. Giraud *et al.* (1997), using human subjects, found that contralateral noise improved speech-in-noise intelligibility in normal ears. They suggested that the crossed olivocochlear efferents were responsible. Kumar and Vanaja (2004) also found that contralateral acoustic stimulation improved speech intelligibility in noise when noise was presented to the contralateral ear, for ipsilateral signal-to-noise ratios (SNRs) of +10 and +15 dB. They showed that the same contralateral noise could suppress ipsilateral otoacoustic emissions, suggesting a role for efferent fibers. Kim *et al.* (2006) investigated the relationship between MOC efferent processing and speech intelligibility in noise for normal hearing listeners of different ages, using distortion product otoacoustic emissions (DPOAEs) as an index of efferent activity. Their findings suggest that the decline in ability to understand speech in noise with increasing age is associated with a corresponding decline in the function of the MOC efferent system.

It should be noted that not all of the evidence supports a role for the efferent system in improving the intelligibility of noisy speech. Wagner *et al.* (2008) found no correlation between efferent activity and speech intelligibility in noise, as judged by a speech reception threshold (SRT) test and measurements of contralateral suppression of DPOAEs. Other studies suggest a relatively minor role for the OCB in hearing; for example, Scharf *et al.* (1997) studied patients with sectioned crossed olivocochlear bundles and found that they had no audiological impairment. However, one of the few changes noted in these patients concerned the absence of an attention effect seen in normal listeners. In this effect, normal listeners had raised thresholds for stimulus tones at frequencies that were unexpected (i.e., had a low probability of occurrence). In their patient group, by contrast, Scharf *et al.* (1997) found no increase in threshold for unexpected stimuli. It is possible that efferent fibers were suppressing the BM response in regions most sensitive to the low probability frequencies in normal listeners. If this is the case, the patient group would be less likely to show this "attention" effect.

One way of critically assessing the claims made for the role of the efferent system in improving speech intelligibility in noise is to build and evaluate a computer model. Ghitza and co-workers (Ghitza *et al.*, 2007; Ghitza, 2007; Messing *et al.*, 2009) proposed a computer model of auditory efferent processing, and evaluated it on a speech recognition task. They described a closed-loop model of the auditory periphery in which the mechanical filtering properties of the cochlea are regulated by feedback based on short-term measurements of the dynamic range of simulated AN fibers. Ghitza *et al.* (Ghitza *et al.*, 2007; Ghitza, 2007) modeled consonant-confusions made by listeners in noise, for a diphone discrimination task that used synthetic speech stimuli with restricted phonemic variation. This was achieved by coupling the auditory model to a simple speech recognizer, which performed template-matching based on a minimum mean-squares error distance metric.

The current study also investigates the possible effect of efferent activity on speech intelligibility by using a computer model of auditory efferent processing as the "front-end" acoustic processor for an automatic speech recognition (ASR) system. Our approach differs from that of Ghitza *et al.* (2007) in several important respects. First, we use the auditory model of Ferry and Meddis (2007) which is open-loop (i.e., the amount of efferent suppression is fixed directly by the experimenter). This allows a systematic study of the effect of different amounts of suppression in order to identify the optimum level of efferent activity as a function of the level of the background noise. A similar analysis is not possible with Ghitza *et al.*'s (2007) model because of its closed-loop design [it should be noted, however, that the model of Ferry and Meddis (2007) is greatly simplified because the efferent system operates as a closed-loop system in practice].

Second, the current study uses a conventional statistical ASR system that is trained on a large corpus of naturalistic speech (spoken digits from the TIDigits speech corpus; Pearce and Hirsch, 2000). This contrasts with Ghitza *et al.*'s (2007) study, which used a simple template-matching recognizer and synthetic speech. Ghitza *et al.*'s (2007) choice of speech material and recognizer was made in order to ensure that errors due to the recognizer were minimized, so that the consonant-confusions that occurred were mainly due to the auditory model. A limitation of our approach is that we are unable to discriminate errors due to the "back end" recognizer from those that originate in the front end auditory model. However, our approach also has advantages; the speech material used is naturalistic and therefore more representative of the phonemic variation that is typically encountered in speech. Also, the statistical speech recognizer that we use is typical of modern ASR systems; the results of the current study therefore indicate whether, in principle, an auditory model that incorporates efferent processing could serve as a noise-robust front-end for a practical ASR system.

Finally, we note that Ghitza *et al.* (2007) used a simplistic model of neuromechanical transduction by inner hair cells and made no reference to the role of adaptation in explaining their findings. The model of Ferry and Meddis (2007) used here incorporates a detailed model of adaptation (and recovery from adaptation) at the auditory nerve synapse, and this will be shown to be an important factor in explaining the effect of efferent suppression.

It is important to stress that the proposed model is intended purely to illustrate the principle benefits of efferent stimulation when recognizing speech in noisy backgrounds. It is not proposed as a working model of the auditory efferent system for general use. Such a model would need to operate on a closed-loop basis such as that of Ghitza *et al.* (Ghitza *et al.*, 2007; Ghitza, 2007) and take into account the considerable body of knowledge recently accumulated concerning the time constants of efferent activation and the different amounts of inhibition observed across frequency. This will be the focus of future work. However, the results of Liber-

man (1988, Fig. 12C) suggested that noise, when it is present, is the dominant influence on the amount of efferent activity compared to the influence of accompanying pure tones. It follows that the model used here might well be a useful representation of what happens in the presence of continuous background noise of unchanging level.

The remainder of the article is structured as follows. In Sec. II, the computer model is described and it is informally demonstrated that efferent suppression of the BM response leads to an improved representation of speech in noise. An analysis is then presented which concludes that the beneficial effects of efferent suppression can largely be explained in terms of release from adaptation. After a description of the ASR system and speech corpus in Sec. III, experiments are presented in Secs. IV–VI that quantify the speech intelligibility gain associated with efferent processing, and the extent to which this depends on the noise level and the speech level.

II. THE COMPUTER MODEL

A. Model description

The computer model is shown schematically in Fig. 1, and consists of two main stages (delineated in the figure by gray boxes). The first stage is a model of peripheral auditory processing, which takes a digitally sampled mixture of speech and noise as its input and produces a simulation of action potential generation in the AN. The second stage is an ASR system that uses statistical word models to decode the AN firing pattern into its corresponding word sequence.

The computer model of the auditory periphery consists of a cascade of modules representing the resonances of the outer/middle ear, the response of the basilar membrane, coupling by inner hair cell stereocilia, the inner hair cell receptor potential, calcium dynamics, and transmitter release and adaptation at the inner hair cell auditory-nerve synapse. The final stage of the model produces a probabilistic representation of firing rate in the AN. 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 model of the basilar membrane used here is a modification of the dual resonance nonlinear (DRNL) filterbank proposed by Ferry and Meddis (2007) (see also Meddis *et al.*, 2001). The DRNL receives its input (stapes velocity) from a model of the outer/middle ear, and produces an output (basilar membrane velocity) that drives a simulation of inner hair cell function. A single DRNL filter is shown in Fig. 2. The output of the DRNL is the sum of two signal pathways, which represent linear and nonlinear components of the basilar membrane response. Each pathway consists of a sequence of bandpass (gammatone) and lowpass (Butterworth) filters.

The nonlinear path also contains a compressive nonlinearity, implemented by a "broken-stick" function that compresses the input signal (i.e., stapes velocity) when it exceeds a threshold level.[1] The nonlinear path also begins with an attenuation stage, introduced by Ferry and Meddis (2007) to model the effect of efferent suppression from the MOC. The amount of attenuation is determined by the parameter ATT

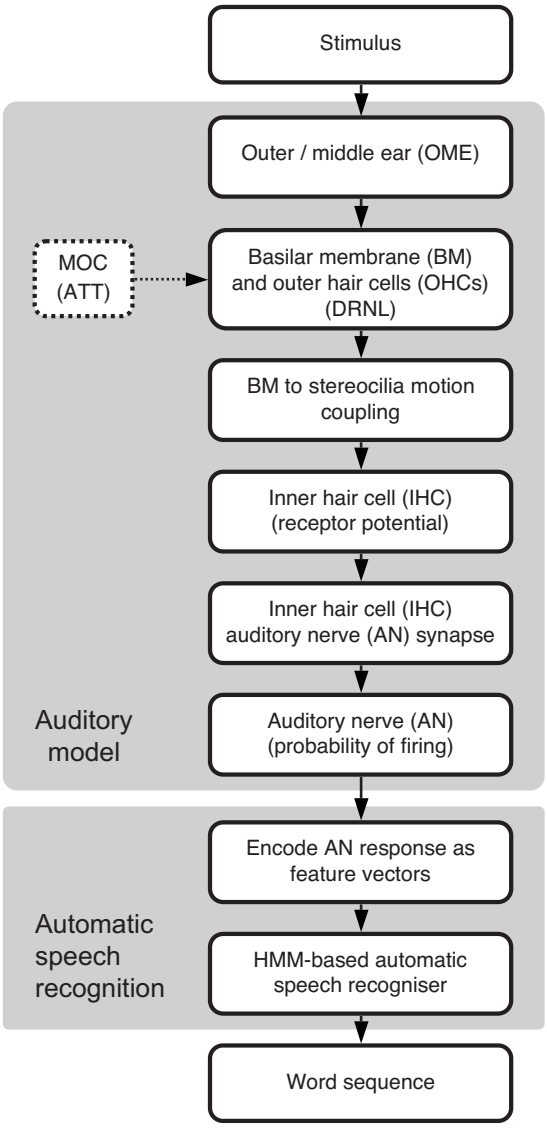

FIG. 1. Schematic of the computer model used for speech recognition experiments. The first major component is a model of the auditory periphery, which includes a stage representing efferent suppression from the MOC. The input to the auditory model is a digitally sampled stimulus (a mixture of speech and noise) and the output is the AN representation of the stimulus, in terms of firing probability. The second major component is an ASR system based on statistical word models. Data reduction is performed so that the AN response is encoded by a time-series of small feature vectors. These features are then decoded by an ASR system based on HMMs, producing a word sequence that is scored against a reference transcription.

(in decibels). This model has been shown to be in good agreement with physiological measurements of the basilar membrane, auditory-nerve and CAP responses when the value of ATT is chosen to be proportional to the amount of MOC activity (Ferry and Meddis, 2007).

In the following simulations, the parameters of the model differ from those used by Ferry and Meddis (2007). Whereas their study modeled physiological data from the guinea pig, our study addresses the representation of speech in human hearing. The outer/middle ear stage of the model was configured using data from Huber *et al.* (2001),[2] whereas the DRNL filterbank parameters were taken from Lopez-Poveda and Meddis (2001). The parameters for subsequent stages of the model were those given by Meddis

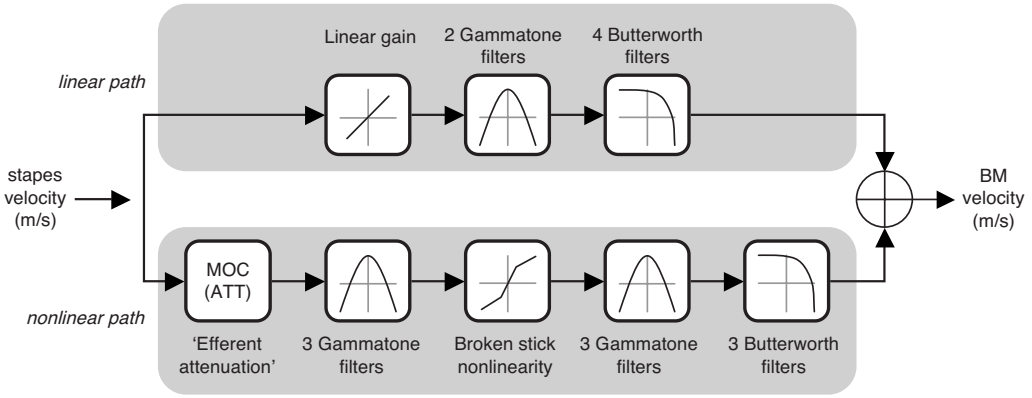

FIG. 2. Schematic diagram of the DRNL filterbank, modified to include an "efferent attenuation" stage (adapted from Ferry and Meddis, 2007). The DRNL consists of parallel linear and nonlinear signal paths, and the suppressive role of the MOC is modeled by inserting an attenuator at the start of the nonlinear path. The degree of efferent activity is determined by tuning the parameter ATT; larger values of ATT correspond to greater suppression by the MOC. The DRNL receives its input (stapes velocity) from a model of the outer/middle ear. Output from the DRNL (basilar membrane velocity) is subsequently processed by a model of inner hair cell function to give a representation of auditory-nerve activity.

(2006). Guinan and Stankovic (1996, Fig. 1) showed six different types of rate/level functions with and without electrical stimulation, all of which have been simulated in an earlier publication (Ferry and Meddis, 2007). One of these was chosen for the present study on the basis that it gave a good representation of the speech in quiet but a poor representation of speech in noise. The model fiber had a low spontaneous rate (LSR), a threshold of 20 dB and a narrow dynamic range saturating at 50 dB sound pressure level (SPL), and simulated in all respects the fiber in their Fig. 1D. These characteristics were obtained by setting the calcium clearance time constant $\tau_{Ca}=0.75\times10^{-4}$ s, as given in Table III of Meddis, 2006 (see also Sumner *et al.*, 2002, Table II). Clearly, many different kinds of fibers or mixtures of them could have been used. However, our main purpose was to illustrate how the efferent system could benefit the recognition of speech in noise even when the dynamic range is restricted.

To provide an auditory time-frequency representation of the noisy speech stimuli, 30 frequency channels were used with best frequencies (BFs) distributed between 100 and 4500 Hz on a logarithmic scale. A detailed list of model parameters can be found in Appendix A of Ferry (2008). The model was implemented in the MATLAB programming language: the source code is available from the authors on request.

In the second stage of the computer model, shown in Fig. 1, the simulated AN firing patterns provide the input to an automatic speech recognizer. The recognizer is a conventional statistical speech recognition system in which whole words are modeled by hidden Markov models (see, for example, Gales and Young, 2008). To provide a suitable input to the recognizer, the AN firing patterns are encoded as a sequence of feature vectors, each of which describes the spectral shape of the AN response at a certain point in time. Details of the encoding strategy and recognizer architecture are given in Sec. III B.

B. Analysis

Figure 3 shows the output of the peripheral auditory model in the form of an "auditory spectrogram," obtained by integrating overlapping 25 ms Hann-windowed segments of the simulated AN firing probability within each frequency channel at intervals of 10 ms. The grayscale value corresponds to firing rate (darker tones indicate higher firing rate). Panel (a) of the figure shows the auditory spectrogram for the utterance "two eight four one" spoken by a male talker and presented at a level of 60 dB SPL. Acoustic-phonetic features that are known to be important for speech intelligibility

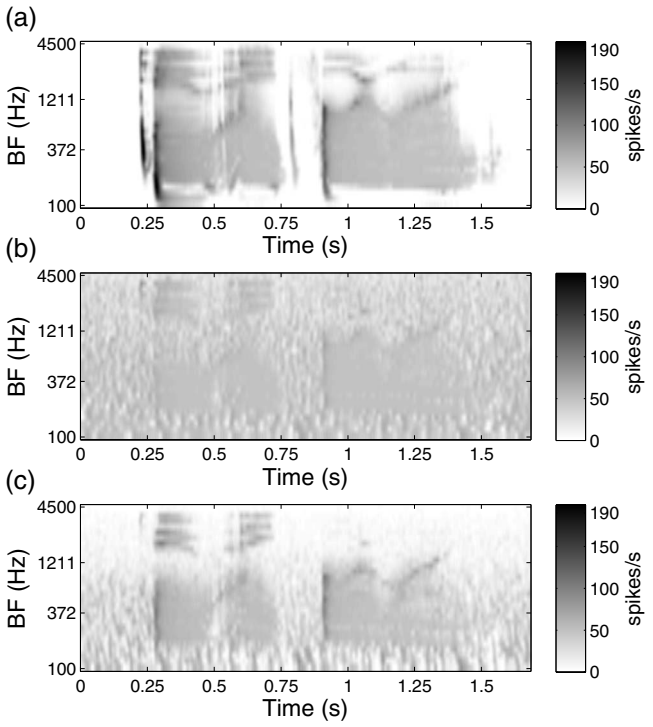

FIG. 3. Simulated auditory-nerve firing rate representations ("auditory spectrograms") for the utterance "two eight four one" spoken by a male talker. Darker regions represent higher firing rate, and the level of the speech is 60 dB SPL in all panels. (a) Clean speech with no efferent activity. (b) Speech with pink noise added at a level of 50 dB SPL (giving a signal-to-noise ratio of 10 dB), with no efferent activity. (c) As in panel (b), but with an efferent activity of 15 dB applied to the model. Efferent suppression reduces the masking effect of the noise. For clarity, the first 1 s of the auditory-nerve response has been omitted from the display; the speech was preceded by 1 s of silence in panel (a) and 1 s of noise in panels (b) and (c).

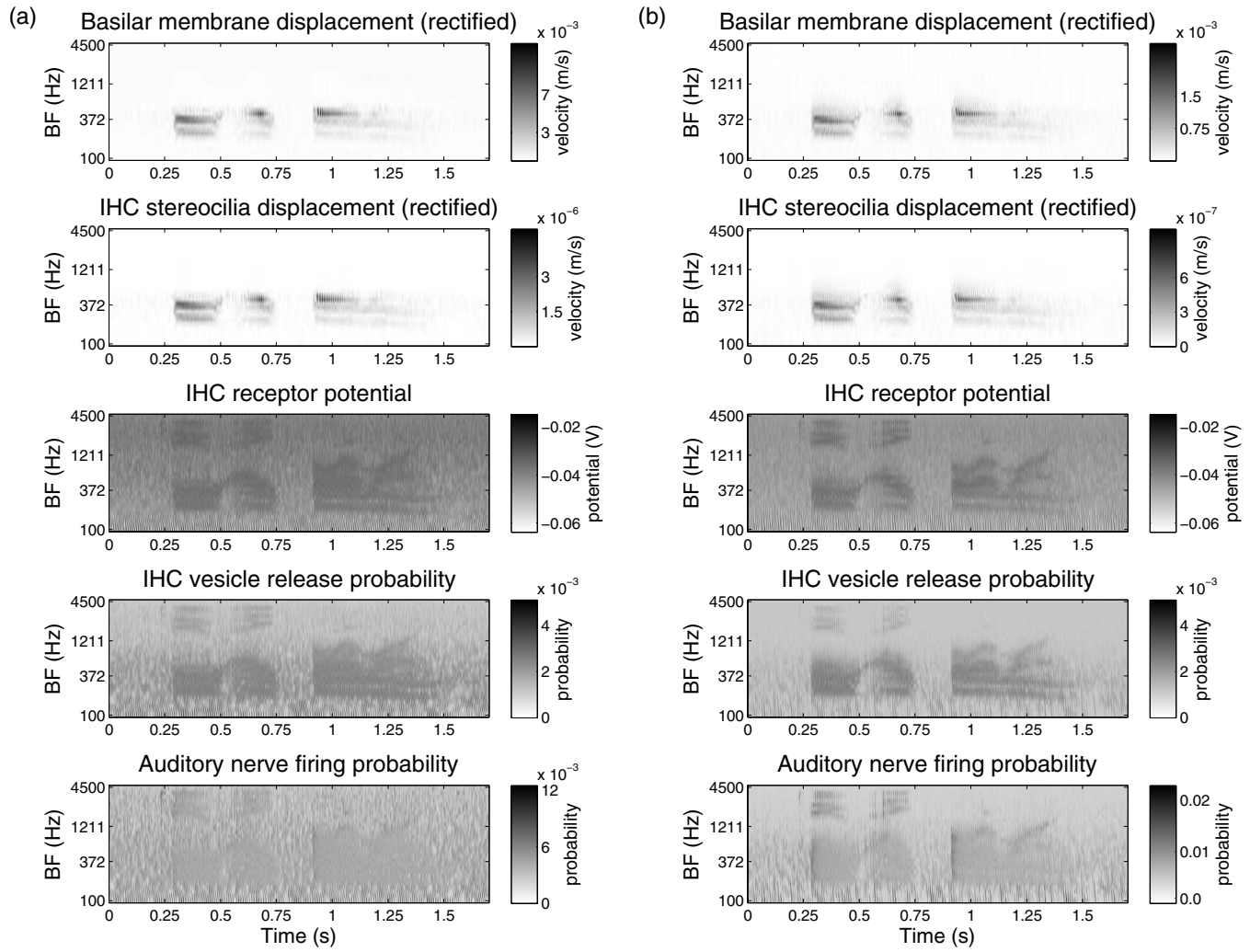

FIG. 4. Output from major stages of the auditory model, for the utterance "two eight four one" presented in pink noise. The sound levels of the speech and noise were 60 and 50 dB SPL, respectively, giving a signal-to-noise ratio of 10 dB. The response to 1 s of noise preceding the speech has been omitted from each figure. (a) No efferent attenuation (ATT=0 dB). (b) Simulated MOC stimulation giving an efferent attenuation of 15 dB (ATT=15 dB). From top to bottom, the plots in (a) and (b) show basilar membrane displacement, cilia displacement, IHC receptor potential, IHC vesicle release probability, and auditory-nerve firing probability. To improve the quality of the grayscale display, the basilar membrane displacement and IHC stereocilia displacement are full-wave rectified, and different scales are used in the left and right panels. Note that the auditory-nerve firing probability is higher in panel (b), because efferent attenuation reduces the adaptation caused by the preceding noise.

(e.g., formant transitions, release bursts, and frication) are well represented. However, panel (b) shows that much of this structure is lost when pink noise with a level of 50 dB SPL is added to the utterance (corresponding to a SNR of 10 dB). Weak time-frequency structure is masked by the noise, and high-intensity parts of the signal now drive the simulated auditory-nerve fibers close to their saturated firing rate. Panel (c) shows that the masking effect of the noise can be partially reversed by efferent suppression of the basilar membrane response. Here, efferent activity was simulated by setting ATT=15 dB, which reduces the gain in the nonlinear path of the DRNL by 15 dB (this amount of attenuation was found to be optimal for a speech level of 60 dB SPL and noise level of 50 dB SPL, as described in Sec. V). Note that the speech was preceded by 1 s of silence in panel (a), and 1 s of noise in panels (b) and (c); for clarity, the first 1 s of the auditory-nerve response has been omitted from the display.

Further insight into the effect of efferent activity can be gained by considering the representation of noisy speech in each stage of the auditory periphery model. Figure 4 shows the output from each stage of processing for the same mixture of speech and noise used in Fig. 3, for cases in which (a) no efferent attenuation is applied and (b) the efferent attenuation is 15 dB. In each case, the speech was preceded by 1 s of noise (which is not shown in the display). At stages of the model up to the inner hair cell (IHC) receptor potential, the effect of efferent activity resembles scaling by a constant factor (although constant scaling is not specifically expected due to the effect of BM compression). However, efferent activity has a more complex effect at the stage of the IHC vesicle release probability and beyond. Vesicle release due to the noise is suppressed, reducing adaptation and allowing the speech regions to elicit a larger vesicle release. This is reflected in the AN response, which shows a greater probability of firing and an increased dynamic range. The "unmasking" of noisy speech by efferent suppression can therefore be explained in terms of release from adaptation.

An alternative way of understanding this effect is to con-

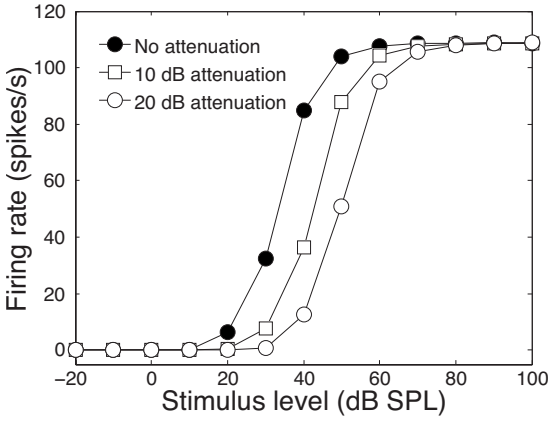

FIG. 5. Rate-level functions for a simulated auditory-nerve fiber with a best frequency of 1 kHz and a low spontaneous rate. The stimulus was a pure tone presented at BF with a duration of 100 ms and onset/offset ramps of 5 ms. Firing rate was averaged over the last 50 ms of the stimulus. Conditions are shown in which no efferent attenuation was applied (ATT =0 dB), and in which MOC stimulation was modeled by applying efferent attenuations of 10 and 20 dB (ATT=10, 20 dB). The rate-intensity curve shifts to the right as the amount of MOC stimulation increases.

sider the change in the rate-level function of a simulated auditory-nerve fiber when efferent activity is applied. Figure 5 shows a rate-level function generated by presenting brief (100 ms) pure tones to a simulated auditory-nerve fiber with a BF of 1 kHz. The frequency of the tone was set to the BF of the simulated fiber, and the firing rate was measured over the last 50 ms of each stimulus. The rate-level function has a typical sigmoidal shape, which progressively shifts to the right when increasing amounts of efferent attenuation are applied (conditions for ATT=10 dB and ATT=20 dB are shown in the figure). When speech is presented in a less intense noise background (i.e., at a positive SNR), the effect of such a shift in the rate-level function will be to reduce the AN response to the noise, since the lower-level noise will be relegated to the toe of the curve. Likewise, a shift in the rate-level function moves high-energy regions of the speech and noise mixture from the shoulder of the rate-level curve back to its linear portion. This reduces saturation and restores the dynamic range of the fiber. Note that the amount of unmasking produced by such a mechanism will depend on the level of the speech; this point is addressed later (Sec. VI).

III. AUTOMATIC SPEECH RECOGNIZER

The above discussion suggests that speech intelligibility in background noise should be improved by efferent suppression and raises the question of how the amount of unmasking is related to the speech level and noise level. The remainder of the paper investigates these issues by using the auditory model as the front-end processor for an ASR system. The speech and noise corpus and recognizer architecture are now described.

A. Corpus

Speech material for the following experiments was drawn from the Aurora 2.0 corpus (Pearce and Hirsch, 2000), which consists of sequences of between one and seven digits ("oh," "zero," and "one" to "nine") 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, 1001 utterances from the "clean1" section of Aurora test set A were used. In addition, a small development set of 200 utterances drawn from the "clean2" section of test set A was used to tune the auditory model and ASR system. The training, testing, and development sets were completely independent, and each contained an approximately equal number of recordings from male and female talkers.

The Aurora speech material was modified in two respects to suit the experiments described here. First, the neuromechanical transduction stage of the auditory model involves numerical integration that must be performed at a high sample rate. Accordingly, all utterances were upsampled to a rate of 44.1 kHz (from the Aurora sample rate of 20 kHz) using the MATLAB resample function. Second, all utterances were scaled to the same root-mean-square level (60 dB SPL) in order to minimize changes in the spectral representation obtained from the (nonlinear) auditory model due to variations in sound level.

Noisy speech was generated by adding pink noise to the test utterances at a range of SNRs between 200 dB (clean) and 0 dB. Broadband noise was employed because it is known to be a particularly effective speech masker (e.g., Miller, 1947). The pink noise was band-passed between 100 Hz and 10 kHz in order to ensure that noise energy above the Nyquist frequency of the Aurora speech signals did not influence the SNR. Prior to adding the noise, 1 s of silence was appended to the start of each utterance; this allowed the auditory model to adapt before the onset of the speech. The corresponding second of simulated auditory-nerve response was removed before speech recognition.

B. Automatic speech recognizer

Speech recognition was performed by a conventional continuous-density hidden Markov model (HMM) system (e.g., see Gales and Young, 2008). Such systems require the acoustic input to be encoded as a sequence of feature vectors, each of which (a "frame") encodes the shape of the speech spectrum over a brief time window. The goal of the recognizer is to find the most likely word sequence that corresponds to an observed sequence of feature vectors.

The recognizer represents speech units (e.g., words) by trained HMMs that model the speech as a sequence of stationary states. Each state is characterized by a multivariate Gaussian mixture distribution over the observed acoustic feature vectors. During training, the Baum–Welch algorithm is used to learn the parameters of the HMMs from a large corpus of annotated speech. During testing, the VITERBI algorithm is applied to find the most likely sequence of HMM states (and hence words) given an observed sequence of feature vectors and the trained HMMs. For an accessible review of the Baum–Welch and Viterbi algorithms, see Rabiner, 1989. Here, a modified version of the Aurora baseline recognizer was used (Pearce and Hirsch, 2000), in which observations were modeled by Gaussian mixtures with diagonal covariance. Gaussian mixtures with seven components were

used, as these were found to give good performance on the development set.

It was necessary to perform some data reduction in the output of the auditory model in order to obtain feature vectors that were suited to the HMM speech recognizer. As previously shown in Fig. 3, 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 (i.e., successive windows overlapped by 60%), giving a temporal resolution that is typical for ASR systems. However, the resulting spectral features are not well modeled by a small number of Gaussian mixture components with diagonal covariance, because features from adjacent frequency channels are highly correlated. Accordingly, further data reduction was performed by applying a discrete cosine transform (DCT) to each frame, giving feature vectors that contain approximately independent components (Oppenheim *et al.*, 1999). The first 14 DCT coefficients were retained. To improve performance, time derivatives of the static DCT coefficients were also included; specifically, first-order and second-order regression coefficients (referred to as "deltas" and "accelerations") were appended to each vector, to give a total of 42 features per frame. A similar approach has been used in numerous other studies that employ auditory models as acoustic front-end processors for ASR systems (e.g., Jankowski *et al.*, 1995; Holmberg *et al.*, 2007).

HMMs with 16 emitting states were trained for each word in the Aurora corpus (i.e., the digits zero, oh, and one to nine). Models were also trained for silence (three states) and short pauses (one state). To reduce the number of insertion errors, a simple grammar was used to constrain all hypotheses so that they started and ended with the silence model. The hidden Markov model toolkit (HTK) was used to train the models and perform decoding (Young *et al.*, 2009). The ASR system was always trained on clean speech (i.e., without added noise) and no efferent attenuation was applied during training (ATT=0 dB).[3]

Word sequences produced by the recognizer were scored using the HTK HResults tool, which compares the transcript produced by the recognizer with a hand-labeled reference transcription. Recognition accuracy is computed as $(H-I)/N\times 100\%$, where H is the number of correct words ("hits"), I is the number of incorrectly inserted words ("insertions"), and N is the total number of words in the reference transcription.

IV. EXPERIMENT I: EFFECT OF EFFERENT ACTIVITY

It is well known from experiments with human listeners that speech intelligibility declines in the presence of broadband noise (e.g., Miller, 1947). The same is true of ASR systems, and hence the performance of an ASR system in noise can be taken as indicative of human speech intelligibility in noise. The comparison is only a qualitative one, however, because the error rate of an ASR system is typically an order of magnitude greater than that of a human listener under the same conditions (Lippmann, 1997). Additionally, humans and ASR systems differ in the rate at which their performance falls in the presence of increasing amounts of noise; ASR systems degrade much earlier, and much quicker, as the SNR worsens.

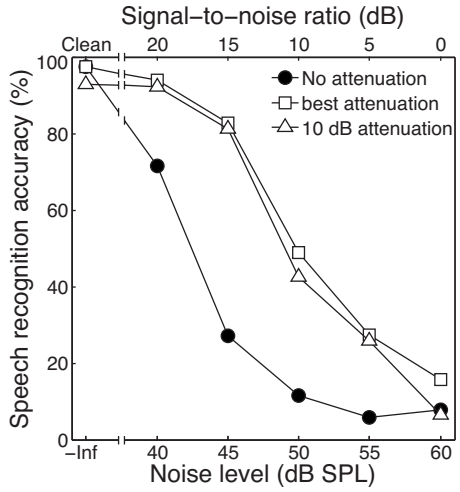

FIG. 6. Speech recognition accuracy (in percent) as a function of noise level, for conditions in which no efferent attenuation is applied, when 10 dB of efferent attenuation is applied, and when the "optimal" efferent attenuation for each noise level is used. Each point represents an average over the test set of 1001 utterances. Speech stimuli were presented at a sound level of 60 dB SPL during training and testing. Efferent activity gives a substantial performance gain in some noisy conditions. Note, however, that efferent activity degrades performance in the clean condition, suggesting that less activity is required in clean conditions and more in noisy conditions; this can be achieved by applying an optimal attenuation that is proportional to the noise level.

Noise is detrimental to ASR performance because it introduces a discrepancy between the training and testing conditions (i.e., there is a mismatch between the statistical models that are derived from clean speech during training, and the noisy speech features that are encountered during testing). A first question is whether efferent activity is able to compensate for this mismatch, by providing a representation of noisy speech that is closer to the ideal clean-speech models.

Figure 6 shows speech recognition accuracy for a range of noise levels, obtained from the auditory model and ASR system as described above. When efferent activity is disabled by setting ATT=0 dB (recall Fig. 2), speech recognition accuracy is high (97.5%) for clean speech but declines sharply with increasing noise level. For noise levels between 40 and 55 dB SPL, a substantial improvement in recognition accuracy is obtained by introducing an efferent activity of 10 dB (i.e., ATT=10 dB). This result confirms that for a speech level of 60 dB SPL, efferent suppression serves to reduce the effects of the noise, yielding acoustic features that more closely resemble those of clean speech. Experiment III (Sec. VI) investigates whether this conclusion holds across a range of different speech levels.

A notable feature of Fig. 6 is that recognition accuracy of clean speech declines (to 92.4%) when an efferent attenuation of 10 dB is introduced. This suggests that efferent activity is undesirable when noise is absent, because it warps the auditory representation of the speech away from the clean-speech models. More generally, this raises the issue of whether there is an "optimal" amount of efferent activity that

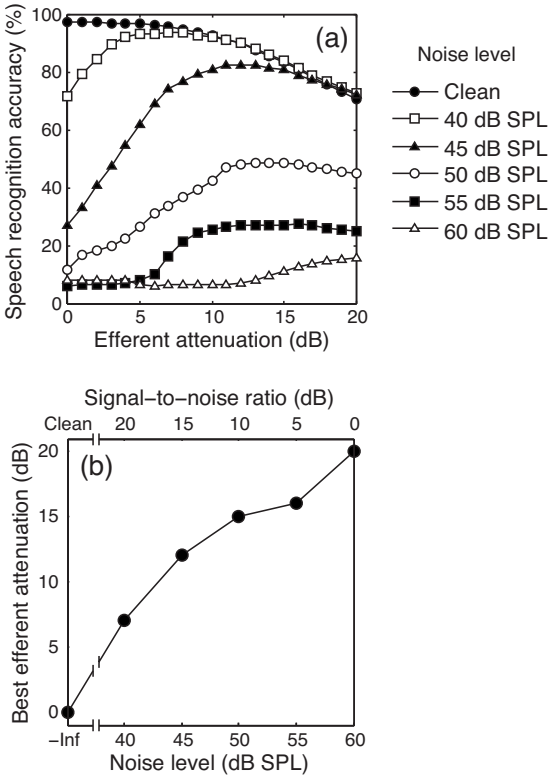

FIG. 7. (a) The effect of efferent activity on speech recognition accuracy. Each curve corresponds to a different noise level, with the speech level fixed at 60 dB SPL. (b) Best efferent attenuation as a function of noise level. The best efferent attenuation corresponds to the peak of each curve in panel (a); i.e., it is the efferent attenuation that maximizes speech recognition accuracy for the given noise level. Speech stimuli were presented at a sound level of 60 dB SPL during training and testing of the speech recognizer.

maximizes speech recognition accuracy, and whether this depends on the noise level. Similarly, an efferent attenuation of 10 dB gives little performance gain when the noise level is 60 dB SPL, and it is possible that recognition accuracy could be improved in this condition by increasing the amount of efferent activity. These issues are addressed in the following experiment.

V. EXPERIMENT II: EFFECT OF NOISE LEVEL

In this experiment, the relationship between noise level, level of efferent activity, and speech recognition accuracy was systematically investigated. Speech recognition accuracy was obtained for different configurations of the auditory model in which the efferent attenuation was set to a value between 0 and 20 dB in steps of 1 dB. The upper bound on the efferent activity was set in accordance with the physiological study of Liberman and Guinan (1998) in the cat, which found that the maximum suppression obtained with sound-evoked activity was approximately 20 dB. For each configuration of the model, speech recognition accuracy was evaluated in a range of noise conditions (clean speech, and speech with pink noise added at levels between 40 and 60 dB SPL).

Figure 7(a) shows the results from this experiment. For clean speech, recognition accuracy is highest when there is no efferent attenuation (ATT=0 dB). In the remaining conditions, the curve relating efferent activity to speech recognition accuracy generally shows a broad peak that indicates the optimal efferent attenuation for that noise level. Figure 7(b) plots these optimal efferent attenuation values against noise level. The results suggest that the best efferent attenuation is proportional to the noise level (i.e., if the level of the speech is held constant and noise is added, higher noise levels require more efferent activity).

Speech recognition accuracy is plotted in Fig. 6 (open squares) when the optimal efferent attenuation is used for each noise level. The resulting performance curve is indicative of the speech recognition accuracy achievable by a system that adjusts the efferent activity to its optimum value based on an assessment of the noise level. However, the optimal efferent attenuation is likely to depend on speech level (and overall sound level) in addition to the noise level. These factors are considered in the next experiment.

VI. EXPERIMENT III: EFFECT OF SPEECH LEVEL

In the two previous experiments, the speech was presented at a sound level of 60 dB SPL during training and testing of the ASR system. A further question is whether the benefits of efferent activity observed at a speech level of 60 dB SPL are also apparent at other speech levels. To address this, the ASR system was trained on clean speech and tested on mixtures of speech and pink noise for which the level of the speech was varied between 40 and 80 dB SPL in steps of 10 dB.

It should be noted that the aim of this experiment was to investigate the likely benefit of efferent activity at different sound levels, rather than to determine the robustness of the ASR system to discrepancies between the level of the speech in the training set and test set. Because the auditory model is nonlinear, it provides the recognizer with acoustic features that are level-dependent; hence, recognizer performance declines if there is a difference in speech level between the training and test sets. To avoid this confound, the sound level of the speech was always held the same during training and testing of the recognizer.

The results from this experiment are shown in Fig. 8. Speech recognition accuracy was determined using the optimum value of the efferent attenuation for each noise level, which was obtained using the procedure described in Sec. V. Substantial benefits of efferent activity are obtained at speech levels between 40 and 70 dB SPL. However, speech recognition accuracy is poor when the speech is presented at a level of 80 dB SPL, even with efferent activity. This can be explained by the rate-level function shown in Fig. 5. At a sound level of 80 dB SPL, the simulated auditory-nerve fibers are driven close to their saturated firing rate, even when a substantial efferent attenuation is applied. As a result, the efferent suppression does not result in a reduced firing rate during the noise and no reduction in adaptation is achieved. Conversely, speech recognition accuracy improves at lower sound levels, because the simulated auditory-nerve fibers respond in the linear portion of their rate-level functions. It should be noted, however, that the model does not currently include a simulation of the acoustic reflex, which would be active at sound levels above 75 dB SPL (Liberman and

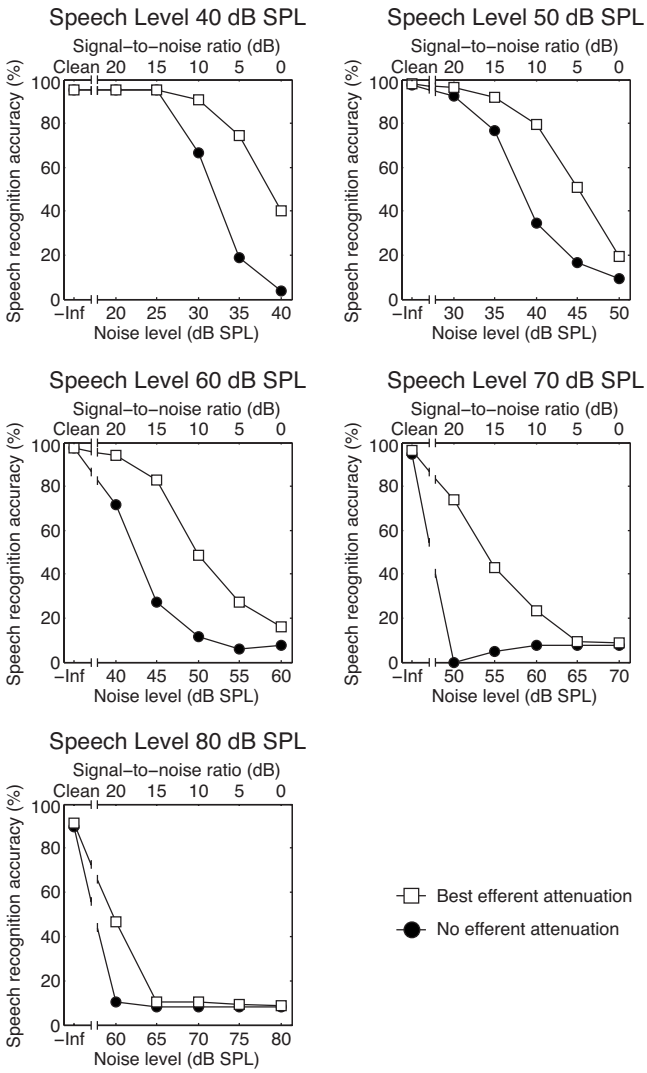

FIG. 8. Effect of speech level and noise level on word recognition accuracy. The plots show performance on the 1001-word test set when the recognizer was trained and tested on speech presented at sound levels between 40 and 80 dB SPL. The figures show performance without efferent activity and when the best efferent attenuation is used. The best efferent attenuation was determined separately for each experimental condition (i.e., for each combination of speech level and noise level).

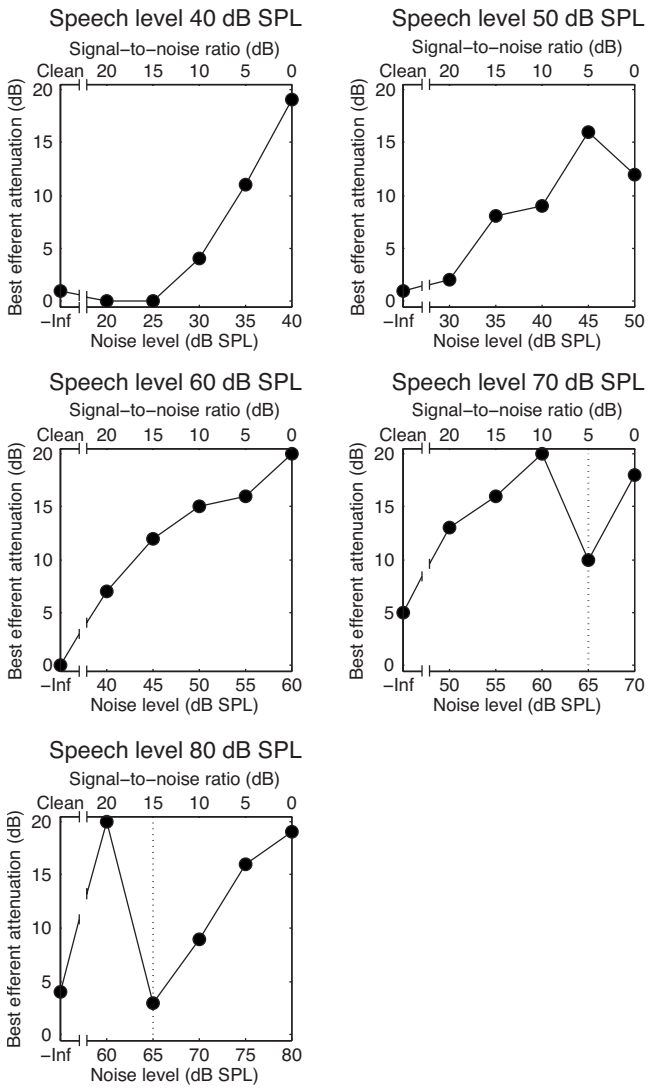

FIG. 9. Effect of speech level and noise level on best efferent attenuation. The plots show the best efferent attenuation for speech levels between 40 and 80 dB SPL. The best efferent attenuation was determined separately for each experimental condition (i.e., for each combination of speech level and noise level). At speech levels of 60 dB SPL and below, the best efferent attenuation generally increases with increasing noise level. For speech levels above 60 dB SPL, best efferent attenuation generally increases with increasing noise level until the performance of the recognizer falls to chance level. The point at which speech recognition degrades to chance performance is marked by a vertical dotted line in the panels for speech levels of 70 and 80 dB SPL.

Guinan, 1998). If present in the model, the acoustic reflex would reduce the effective level of the 80 dB SPL stimulus, leading to reduced adaptation and improved speech recognition accuracy.

Figure 9 illustrates the effect of speech level and noise level on best efferent attenuation. For speech levels of 40, 50, and 60 dB SPL, there is generally a monotonic increase in best efferent attenuation as the noise level is increased. Additionally, the best efferent attenuation rises more steeply with increasing noise level as the speech level is increased, indicating that the best efferent attenuation is also influenced by overall sound level. For speech levels of 70 and 80 dB SPL, the best efferent attenuation increases with increasing noise level only up to the point where the speech recognition performance degrades to chance level (approximately 9%; see Fig. 8). Beyond this point (indicated by a vertical dotted line in Fig. 9), efferent attenuation is unable to counteract the saturation caused by the high overall sound level, and many different values of ATT will give the same (poor) speech recognition performance.

VII. DISCUSSION

The aim of this modeling study was to investigate the potential role of the auditory efferent system in improving the intelligibility of speech that is masked by broadband noise. By using the auditory model as the front-end for an ASR system, it has been shown that speech recognition accuracy is improved by attenuating the response of the simulated basilar membrane when noise is present. Efferent activity has the effect of shifting the rate-level curve of the model auditory-nerve fibers to the right, which improves the dy-

namic range of the firing rate response and provides a release from adaptation. It was found that the optimum efferent attenuation was proportional to the noise level in most experimental conditions. This finding is compatible with a model in which the efferent system adjusts itself to the background noise level in order to prevent excessive adaptation and therefore optimizes speech recognition. Like the previous studies of Ghitza and colleagues (Ghitza *et al.*, 2007; Ghitza, 2007), our results therefore support the notion that the auditory efferent system contributes to the robustness of speech perception in adverse acoustic conditions.

The model suppresses the auditory response to broadband noise, providing a better spectro-temporal representation of speech components. Figure 6 shows that the effect of efferent activity in the model declines at high noise levels, and is marginal when the SNR is 0 dB. The ability of human listeners to recognize speech in noise is better than that of our ASR system. However, human speech intelligibility is poor below a SNR of 0 dB when the speech is masked by broadband noise (word intelligibility less than 50%; see Miller, 1947). We hypothesize that negative SNRs are particularly difficult for human listeners because the efferent system is less effective in unmasking the speech under such conditions. Although efferent activity will still reduce adaptation caused by noise when the speech level lies below the noise level, the model simulation suggests that the benefit will be small.

A limitation of the current study is that it only considers a pink noise masking sound. Human listeners are able to exploit temporal fluctuations in the envelope of a masker in order to "listen in the dips" or "glimpse" the speech (Miller and Licklider, 1950; Cooke, 2006). The effect of efferent suppression on "glimpsing" mechanisms remains an interesting issue for further study. Without efferent suppression, the availability of glimpses may be reduced because the peripheral auditory system is too adapted to respond during dips in the temporal envelope of the background noise. Listeners with impaired hearing have a reduced ability to listen in the dips (e.g., Hopkins *et al.*, 2008). Hence, a deficiency in efferent suppression (caused by outer hair cell damage or by a deficiency in the efferent system itself) is one factor that could contribute to the difficulty that hearing-impaired listeners experience when listening to speech in fluctuating background noise.

In the current model, the same level of efferent activity is applied at all frequencies. In fact, physiological data from the cat suggest that efferent suppression is greatest at frequencies above 2 kHz (Guinan and Gifford, 1988; Liberman and Guinan, 1998). However, the physiological data also show that efferent suppression is effective over a wide frequency range; for low spontaneous rate fibers (as used in the computer model), Fig. 4 of Guinan and Gifford (1988) indicates that the maximum efferent attenuation is at least 12 dB over the range of best frequencies used in the computer model (i.e., between 100 and 4500 Hz). The current model is therefore a reasonable approximation, although there is scope for more detailed modeling of the frequency-dependent effects of efferent activity.

Conventional front-end processors for ASR provide acoustic features that encode the spectral shape of speech but are largely independent of sound level. Typical approaches are cepstral mean subtraction (Liu *et al.*, 1993) and RASTA filtering (Hermansky and Morgan, 1994), both of which aim to remove slowly varying channel characteristics. Without the efferent attenuation component, the DRNL filterbank exhibits level-dependent behavior that complicates its use as a front-end for ASR; in particular, the best frequency associated with a filterbank channel is subjected to change at high sound levels, and the filter bandwidths broaden (Meddis *et al.*, 2001). Efferent activity moderates this behavior to some degree and extends the effective dynamic range of the auditory model (as would a simulation of the acoustic reflex, which is currently absent from the model). However, the results of Experiment III (Fig. 8) show that the performance of the model is still level-dependent, due to the saturating rate-level function of the simulated auditory-nerve fibers. For this reason, the speech recognition accuracy obtained with the auditory model is below that typically obtained with state-of-the-art signal processing front-ends (e.g., Cui and Alwan, 2005).

A related issue is the use of low spontaneous rate (LSR) auditory-nerve fibers in the computer model. In the simulations reported here, the calcium clearance time constant τ_{Ca} (Meddis, 2006; Sumner *et al.*, 2002) was configured to give fibers with a low spontaneous rate. The choice of a LSR fiber type to illustrate the consequences of efferent suppression was determined by the availability of relevant physiological data showing how efferent activity affects a fiber's response. Guinan and Stankovic (1996) showed the effect of efferent suppression on six different fibers, and all of these showed low spontaneous rates. All six fibers were simulated in the modeling study of Ferry and Meddis (2007). By choosing one of these fibers, it was possible to assume some physiological realism. Unfortunately, the most common type of fiber in small mammals shows high spontaneous rates. To compensate for this discrepancy, it was decided to use the fiber with a low threshold (20 dB SPL) and the narrowest dynamic range (20 dB) found in Fig. 1D of Guinan and Stankovich's (1996) report. In these two respects, at least, the fiber used in this study was similar to typical HSR fibers. In contrast, most LSR fibers have high thresholds and wide dynamic ranges.

The relative contribution of low and high spontaneous fiber types to the representation of speech sounds is of considerable interest but beyond the scope of this study (see, however, Winslow *et al.*, 1987; Sachs *et al.*, 2006). Our primary purpose here was to demonstrate improvements in speech-in-noise performance when efferent suppression is added to the model. This demonstration is particularly important with respect to fibers with narrow dynamic ranges. At first sight, auditory-nerve fibers that saturate at levels above 40 dB SPL would appear to be ill-suited to represent speech presented at levels of 60 dB SPL and above, especially when presented against a background of noise at similar levels. The modeling study described above has shown that this apparently unpromising approach can nevertheless give useful representations of speech in quiet up to 80 dB SPL (Fig.

8). Moreover, the poor performance of the system in background noise is considerably ameliorated when efferent suppression is applied at strengths related to the intensity of the noise.

VIII. CONCLUSIONS

It has been shown that efferent suppression improves the intelligibility of speech masked by broadband noise in a model that combines auditory efferent processing with an ASR system. Optimum speech intelligibility is achieved in the model using a level of efferent attenuation that is proportional to the noise level, other than when the noise is so intense that the recognizer degrades to chance performance. Unmasking due to efferent suppression occurs across a wide range of sound levels. However, the amount of unmasking depends both on the speech level and the noise level.

ACKNOWLEDGMENTS

This work was supported by grants from the Royal National Institute for Deaf People (RNID) to Guy Brown and Ray Meddis, and by a grant to Ray Meddis from the UK Engineering and Physical Sciences Research Council (Grant No. EP/E064590/1). The authors would like to thank Oded Ghitza and two anonymous reviewers for their helpful comments.

[1]The broken-stick nonlinearity is defined by $y(t) = \text{sign}[x(t)]\min[a|x(t)|, b|x(t)|^c]$, where a, b, and c are parameters and $x(t)$ and $y(t)$ represent the input and output signals, respectively. For details of the parameter values see Lopez-Poveda and Meddis (2001).

[2]The filtering of the human outer/middle ear is modeled by passing the acoustic signal through three first-order linear bandpass Butterworth filters arranged in series. The first filter has lower and upper cutoff frequencies of 1900 and 4200 Hz. The second and third filters have lower/upper cutoff frequencies of 4500/6300 and 8000/12 000 Hz, respectively.

[3]In practice, the efferent system is likely to be activated by speech alone, and hence it would be appropriate to apply some degree of efferent attenuation during training of the recognizer with clean speech. In this study, efferent activity was suppressed during training of the recognizer in order to simplify our study of the benefit of efferent activity during speech recognition.

Cooke, M. (**2006**). "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am. **119**, 1562–1573.
Cooper, N. P., and Guinan, J. J. (**2006**). "Medial olivocochlear efferent effects on basilar membrane response to sound," in *Auditory Mechanisms: Processes and Models*, edited by A. L. Nuttall, T. Ren, P. G. Gillespie, K. Grosch, and E. de Boer (World Scientific, Singapore), pp. 86–92.
Cui, X., and Alwan, A. (**2005**). "Noise robust speech recognition using feature compensation based on polynomial regression of utterance SNR," IEEE Trans. Speech Audio Process. **13**, 1161–1172.
Dallos, P. (**1992**). "The active cochlea," J. Neurosci. **12**, 4575–4585.
Dewson, J. H. (**1968**). "Efferent olivocochlear bundle: Some relationships to stimulus discrimination in noise," J. Neurophysiol. **31**, 122–130.
Dolan, D., and Nuttall, A. (**1988**). "Masked cochlear whole-nerve response intensity functions altered by electrical-stimulation of the crossed olivocochlear bundle," J. Acoust. Soc. Am. **83**, 1081–1086.
Dolan, D. F., Guo, M. H., and Nuttall, A. L. (**1997**). "Frequency-dependent enhancement of basilar membrane velocity during olivocochlear bundle stimulation," J. Acoust. Soc. Am. 3587–3596.
Ferry, R. T. (**2008**). "Auditory processing and the medial olivocochlear efferent system," Ph.D. thesis, University of Essex, Colchester, UK.
Ferry, R. T., and Meddis, R. (**2007**). "A computer model of medial efferent suppression in the mammalian auditory system," J. Acoust. Soc. Am. **122**, 3519–3526.
Gales, M., and Young, S. (**2008**). "The application of hidden Markov models in speech recognition," Foundations and Trends in Signal Processing **1**, 195–304.
Ghitza, O. (**2007**). "Using auditory feedback and rhythmicity for diphone discrimination of degraded speech," in Proceedings of the International Conference on Phonetic Sciences (ICPhS), Saarbrucken, Germany, pp. 163–168.
Ghitza, O., Messing, D., Delhorne, L., Braida, L., Bruckert, E., and Shondhi, M. (**2007**). "Towards predicting consonant confusions of degraded speech," in *Hearing—From Sensory Processing to Perception*, edited by B. Kollmeier, G. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp, and J. Verhey (Springer, Berlin), pp. 541–550.
Giraud, A., Garnier, S., Micheyl, C., Lina, G., Chays, A., and CheryCroze, S. (**1997**). "Auditory efferents involved in speech-in-noise intelligibility," NeuroReport **8**, 1779–1783.
Guinan, J. J. (**1996**). "Physiology of olivocochlear efferents," in *The Cochlea*, edited by P. Dallos, A. N. Popper, and R. R. Fay (Springer-Verlag, Berlin), pp. 435–502.
Guinan, J. J. (**2006**). "Olivocochlear efferents: Anatomy, physiology, function, and the measurement of efferent effects in humans," Ear Hear. **27**, 589–607.
Guinan, J. J., and Gifford, M. L. (**1988**). "Effects of electrical stimulation of efferent olivocochlear neurons on cat auditory-nerve fibers. III. Tuning curves and thresholds at CF," Hear. Res. **37**, 29–45.
Guinan, J. J., and Stankovic, K. M. (**1996**). "Medial efferent inhibition produces the largest equivalent attenuations at moderate to high sound levels in cat auditory-nerve fibers," J. Acoust. Soc. Am. **100**, 1680–1690.
Hermansky, H., and Morgan, N. (**1994**). "Rasta processing of speech," IEEE Trans. Speech Audio Process. **2**, 578–589.
Hienz, R., Stiles, P., and May, B. (**1998**). "Effects of bilateral olivocochlear lesions on vowel formant discrimination in cats," Hear. Res. **116**, 10–20.
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.
Hopkins, K., Moore, B. C. J., and Stone, M. A. (**2008**). "Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech," J. Acoust. Soc. Am. **123**, 1140–1153.
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 Process. **3**, 286–293.
Kim, S., Frisina, R. D., and Frisina, D. R. (**2006**). "Effects of age on speech understanding in normal hearing listeners: Relationship between the auditory efferent system and speech intelligibility in noise," Speech Commun. **48**, 855–862.
Kumar, U., and Vanaja, C. (**2004**). "Functioning of olivocochlear bundle and speech perception in noise," Ear Hear. **25**, 142–146.
Liberman, M. C. (**1988**). "Response properties of cochlear efferent neurons: Monaural vs. binaural stimulation and the effects of noise," J. Neurophysiol. **60**, 1779–1798.
Liberman, M. C., and Guinan, J. J. (**1998**). "Feedback control of the auditory periphery: Anti-masking effects of middle ear muscles vs. olivocochlear efferents," J. Commun. Disord. **31**, 471–482.
Lippmann, R. P. (**1997**). "Speech recognition by machines and humans," Speech Commun. **22**, 1–16.
Liu, R., Stern, R., Huang, X., and Acero, A. (**1993**). "Efficient cepstral normalization for robust speech recognition," in Proceedings of ARPA Speech and Natural Language Workshop, (Princeton, NJ), pp. 69–74.
Lopez-Poveda, E. A., and Meddis, R. (**2001**). "A human nonlinear cochlear filterbank," J. Acoust. Soc. Am. **110**, 3107–3118.
May, B. J., and McQuone, S. J. (**1995**). "Effects of bilateral olivocochlear lesions on pure-tone intensity discrimination in cats," Aud. Neurosci. **1**, 385–400.
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. D., 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.
Miller, G. A. (**1947**). "The masking of speech," Psychol. Bull. **44**, 105–129.

Miller, G. A., and Licklider, J. (**1950**). "The intelligibility of interrupted speech," J. Acoust. Soc. Am. **22**, 167–173.
Oppenheim, A. V., Schafer, R. W., and Buck, J. R. (**1999**). *Discrete-Time Signal Processing* (Pearson Education, Upper Saddle River, NJ).
Pearce, D., and Hirsch, H.-G. (**2000**). "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in Proceedings of the International Conference on Spoken Language Processing, Vol. **IV**, pp. 29–32.
Rabiner, L. R. (**1989**). "A tutorial on hidden Markov models and selected applications," Proc. IEEE **77**, 257–286.
Russell, 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 Hienz, 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.
Scharf, B., Magnan, J., and Chays, A. (**1997**). "On the role of the olivocochlear bundle in hearing: 16 case studies," Hear. Res. **103**, 101–122.
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.
Wagner, W., Frey, K., Heppelmann, G., Plontke, S. K., and Zenner, H.-P. (**2008**). "Speech-in-noise intelligibility does not correlate with efferent olivocochlear reflex in humans with normal hearing," Acta Oto-Laryngol. **128**, 53–60.
Wiederhold, M. L., and Kiang, N. Y. S. (**1970**). "Effects of electric stimulation of the crossed olivocochlear bundle on single auditory-nerve-fibers in the cat," J. Acoust. Soc. Am. **48**, 950–965.
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.
Winslow, R. L., and Sachs, M. B. (**1988**). "Single-tone intensity discrimination based on auditory-nerve rate responses in backgrounds of quiet, noise, and with stimulation of the crossed olivocochlear bundle," Hear. Res. **35**, 165–190.
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)*, Cambridge University Engineering Department, http://htk.eng.cam.ac.uk/ (Last viewed 10/06/09).