**Effect of auditory efferent time-constant duration on speech recognition in noise**

**Ifat Yasin[a], Fangqi Liu[b], Vit Drga[a], Andreas Demosthenous[b], Ray Meddis [c]**

a) Department of Computer Science, University College London, 66-72 Gower Street, London WC1E 6BT, U.K.

Email: Ifat Yasin : i.yasin@ucl.ac.uk
Email: Vit Drga : v.drga@ucl.ac.uk

b) Department of Electronic and Electrical Engineering, University College London, Torrington Place, London WC1E 7JE, U.K.

Email: Fangqi Liu : fangqi.liu.14@ucl.ac.uk
Email: Andreas Demosthenous : a.demosthenous@ucl.ac.uk

c) Department of Psychology, University of Essex Colchester, CO4 3SQ, U.K.

Email Ray Meddis : rmeddis@essex.ac.uk

Corresponding author: Ifat Yasin, Department of Computer Science, University College London, 66-72 Gower Street, London WC1E 6BT, U.K; E-Mail:i.yasin@ucl.ac.uk; Phone: ++44(0)20 31087159

Running title: Speech-in-noise: Auditory efferent time-constant duration

## ABSTRACT

The human auditory efferent system may play a role in improving speech-in-noise recognition with an associated range of time constants. Computational auditory models with efferent-inspired feedback demonstrate improved speech-in-noise recognition with long efferent time constants (2000 ms). This study used a similar model plus an Automatic Speech Recognition (ASR) system to investigate the role of shorter time constants. ASR speech recognition in noise improved with efferent feedback (compared to no-efferent feedback) for ***both*** short and long efferent time constants. For some signal-to-noise ratios, speech recognition in noise improved as efferent time constants were increased from 118 ms to 2000 ms.



Keywords: Auditory model, Efferent, MOC, Speech recognition, Time constants

## 1. Introduction

In addition to afferent neural pathways, the mammalian auditory system includes a number of efferent pathways, one of which is a brainstem-mediated pathway by way of the medial olivocochlear (MOC) system which reduces the response of the basilar membrane (BM) in the cochlea to sound (Murugasu and Russell, 1996). Physiological non-human mammalian studies and otoacoustic emission (OAE) measures from humans suggest a range of time constants associated with the MOC effect, categorised as slow (tens of seconds), medium (290-350 ms) or fast (ranging from 60-80 ms) (OAEs measured in humans: Backus and Guinan, 2006). Kim *et al*. (2001) also measured efferent time constants in humans using OAEs and described time constants falling within a fast (10-350 ms) and slow (350 ms-5.5 s) range. [Temporal descriptors of fast (short), medium, and slow (long) are typically used in the literature to describe both efferent onset and offset durations.]

Although the MOC is suggested to play a role in improving speech intelligibility in noise (Giraud *et al*., 1997), the role of a range of efferent time constants and their effect on speech recognition in noise remains unknown. The motivation for this study is to investigate the effect of different MOC time constants on speech recognition in noise by adapting an existing computational model of the auditory system (Brown *et al*., 2010). The auditory model is used as the front-end to an ASR system; the ASR is used as a tool to understand the effect of manipulating efferent time constants within the auditory model on speech recognition in noise.

Currently there is much interest in incorporating aspects of human neural "feedback" in computational models of the auditory system (serving as the front-end to ASR devices) to understand the effect of MOC feedback on speech recognition. In general, models incorporating efferent processing (in addition to afferent processing) using even a single long MOC time constant demonstrate a marked improvement in speech intelligibility in noise (Messing *et al*., 2009; Brown *et al*., 2010; Clark *et al*., 2012). Brown *et al*. (2010) used an auditory model (as the "front-end" for an ASR system) with efferent-inspired feedback (Ferry and Meddis, 2007) operating as an open-loop system with fixed amount of efferent gain reduction across signal frequencies and found that speech reception thresholds in pink noise improved by about 10 dB SNR compared to the case where there was no efferent feedback. A similar improvement for speech recognition in pink noise was demonstrated by Clark *et al*. (2012) using a variant of the same model in which the feedback signal dynamically controlled the amount of frequency-dependent attenuation; this is more representative of the physiological operation of the MOC (Guinan, 2006). The feedback (control) signal was dependent on the recent history of auditory nerve activity and was estimated from the temporally-smoothed firing rate using a 1st-order lowpass filter and a lag of 10 ms to account for the MOC-OHC synaptic minimum latency (Liberman, 1988). In the model, the rate-level function was replicated by deriving the control signal from the logarithm of the ratio of the temporally-smoothed firing rate to a firing-rate threshold. The efferent attenuation was derived from multiplying the control signal by a scalar. Further details of this stage of the model are also provided in Clark e*t al*. (2012).

Both Clark *et al*. (2012) and Brown *et al*. (2010) used a single, relatively long efferent time constant for modelling the MOC efferent effect; 2000 ms in duration. The present study investigates whether there is a difference between the effects of short- to medium-duration efferent time constants and longer time constants on speech recognition in noise using the closed-loop model described by Clark *et al*. (2012). For the purpose of this study only pink noise was used in order to allow a direct comparison with the results of Brown *et al*. (2010), Clark *et al*. (2012) as well as Lee *et al*. (2011) who measured speech recognition in pink noise using an alternative auditory model with efferent-inspired feedback.

## 2. Methods

### *2.1 Auditory Model*

The computational auditory model used in the current study is the one described by Ferry and Meddis (2007) and subsequently used by Brown *et al*. (2010) and Clark *et al*. (2012). Since the model components are described in sufficient detail in these papers, only the salient components will be described here. The auditory model represents the responses of the outer ear, middle ear, and basilar membrane (BM) in the cochlea, coupling of BM response to inner hair cell (IHC), IHC transmitter release and auditory-nerve (AN) firing. The Dual Resonance NonLinear (DRNL) model is used to describe the mechanical BM response, with a linear and nonlinear pathway. BM response attenuation by way of efferent feedback is represented by an attenuation stage at the start of the nonlinear pathway (the feedback control signal is received from the recent history of the AN firing response) (Ferry and Meddis, 2007; Clark e*t al*., 2012). A schematic of the model is shown in Fig. 2 of Brown *et al*. (2010). In the present study the efferent activation and decay time constants tested within the model were 2000 ms (in order to make a direct comparison with Brown *et al*., 2010 and Clark *et al*., 2012), 1000 s, 450 ms, 200 ms [within the range of slow and medium efferent time constants reported by Backus and Guinan (2006) using OAE measures], and 118 ms [efferent time constant reported by Yasin *et al.* (2014) using psychoacoustical measures]. The model also includes processing by both high- and low-spontaneous rate fibers, although for the modelling described here only the high spontaneous rate fibers were used in order to make comparisons with previous studies (e.g., Clark *et al*., 2012).

### *2.2 ASR Training and Evaluation*

The input signal to the Hidden Markov Model (HMM) is a sequence of feature vectors generated by integrating AN firing probability at 10-ms intervals; a discrete cosine transform is applied to yield a set of components. The first fourteen coefficients were retained. Since the main steps in training the ASR are described in detail in Brown *et al.*, (2010) and Clark *et al*. (2012), only a summary is provided here. A continuous hidden-density HMM toolkit (Young *et al*., 2009) was used. The speech material was taken from the TIDIGITS corpus (Leonard, 1984). The recogniser was trained on a clean set of material (without either background noise or efferent-related attenuation) consisting of 8440 utterances. The evaluation task was to identify a connected sequence of digits in the presence of background noise (in this case, pink noise). For testing the recognizer, 358 utterances were used, each containing three connected digits from the set (“oh”, “one” “two”, “three”, “four”, “five”, “six” “eight” and “nine”). Utterances were presented in random order at 60 dB SPL and pink noise was added to the utterances at SNRs ranging from -10 dB to 20 dB (-10, -5, 0, 5, 7, 10, 12, 15 and 20 dB). Each test stimulus comprised a sample of 6 s of background noise [as used by Clark *et al*. (2012); a duration sufficient enough to initiate the efferent response] preceding the combined speech plus noise segment. The HMM finds the most probable sequence of digits corresponding to the input sequence of features. A correct response was classified as one in which the recognizer identified the correct digit in the correct position within the presented triplet of digits. For each SNR condition two values of ASR output (speech recognition score) were obtained for each time constant (model runs were randomized). The averaged values are shown in Figure 1 [plot of averaged value of percentage digits correct (%) as a function of the signal-to-noise (SNR)]. The standard errors ranged from 0.036 to 4.09.

## 3. Results and Discussion

Fig. 1 presents the speech recognition scores with efferent activation (for a range of efferent time constants from 118-2000 ms) and in the absence of efferent activation. In general, the trend for an increase in speech recognition scores in the *absence of efferent activation* with increasing SNR is similar to that reported by Clark *et al*. (2012) for SNR values up to 20 dB. Efferent activity resulted in improved recognition scores for all of the time constants studied but some time constants were more effective than others. In the region above 50% correct identification, the greatest improvements were associated with longer time constants. At about 10 dB SNR the benefit to speech recognition with efferent activation (compared to no efferent activation) improved as efferent time constants were increased (there is a corresponding steepening of the sigmoidal function); there was a successive improvement in speech recognition with increasing efferent time constant (118 ms, 200 ms, 450 ms) averaging about 19 dB, 23 dB and 27 %, respectively. However, there appears to be no additional benefit as the time constant was increased from 1000 to 2000 ms. For more challenging conditions where the percent correct value fell below 50%, the shorter time constants sometimes showed greater improvement.

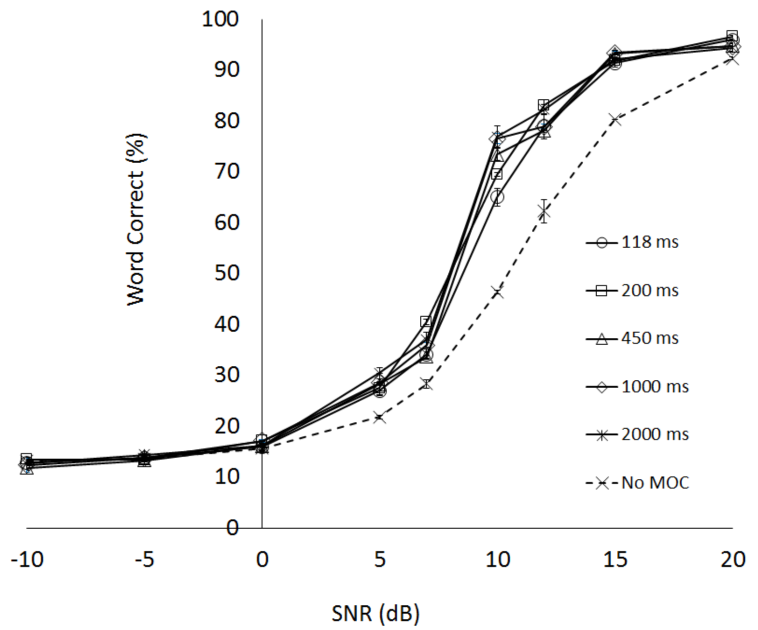


**Fig. 1.** ASR performance [digits correct (%)] as a function of SNR (dB), obtained for pink noise for efferent time constants of 118 ms, 200 ms, 450 ms, 1000 ms and 2000 ms. The comparison plot for the data obtained in the condition where there was no efferent feedback (no MOC) is depicted by the dashed line plus cross symbols.

For a positive SNR of 10 dB there is a successive improvement in speech recognition as the efferent time constant is increased from 118 to 2000 ms.  This is because the response to lower-level noise is represented at the bottom of the shifted sigmoidal rate-level function, whilst the speech response is moved from the saturated part of the curve to the steeper region. Therefore efferent activation in cases of positive SNR confers an advantage to speech recognition in noise. It can also be seen that for negative SNRs, speech recognition remains poor even with efferent feedback; this is similar to the findings of both Brown *et al*. (2010) and Clark *et al*. (2012) for both pink noise and babble noise. This is because in negative SNR conditions the response to the less intense speech is shifted to the bottom of the rate-level response curve whilst the response to the more intense noise is moved from the saturated to the steeper region of the rate-level response curve, providing little benefit to speech recognition.

However, it still remains an open question as to how the auditory system benefits from multiple co-existing time constants. Fast and slow effects of efferent activation appear to emanate from different underlying mechanisms (Cooper and Guinan, 2006), but their roles in perception are not too clear. Efferent effects with different time constants may be required in different listening situations, perhaps dependent on the type and duration of the ongoing background noise. The present results show that, at least with high-spontaneous rate fibers, efferent time constants shorter than 2000 ms (particularly between 118 ms to 450 ms) also bring about incremental increases in the improvement in speech recognition in noise at some SNRs. Recent studies with a binaural cochlear implant sound coding strategy with efferent-inspired feedback also demonstrate improved speech intelligibility in noise with short time constants (Lopez-Poveda *et al*., 2016; Lopez-Poveda *et al*., 2017).

Future work to evaluate the effect of efferent activation on speech recognition in noise could look into the relative contributions of different types neural fibers (low- and high-spontaneous rate) and their respective roles in the linearization of the compression applied to the signal response during efferent activation (Yasin *et al*., 2013; 2014).



## CONCLUSIONS

**1.** Efferent time constants shorter than 2000 ms can also provide improved ASR speech recognition in noise.

**2.** In the region above 50% correct, speech identification (around 10 dB SNR), successive increases in efferent time constant (118-450 ms) leads to successive improvements in speech recognition in noise.

**3.** The greatest improvements in ASR speech recognition performance were associated with the longer time constants.

**ACKNOWLEDGMENTS**

We thank Action on Hearing Loss for an International Project Grant (G70) for supporting this project. We thank the Editor and two anonymous reviewers for their helpful comments.

**REFERENCES**

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

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

Clark, N. R., Brown, G. J., Jürgens, T., and Meddis, R. (**2012**). "A frequency-selective feedback model of auditory efferent suppression and its implications for the recognition of speech in noise," J. Acoust. Soc. Am. **132**, 1535-1541.

Cooper, N. P., and Guinan, J. J. (**2006**). "Separate mechanical processes underlie fast and slow effects of medial olivocochlear efferent activity," J. Physiol. **548**, 307-312.

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.

Giraud, A. L., Garnier, S., Micheyl, C., Lina, G., Chays, A., and Chery-Croze, S. (1997). "Auditory efferents involved in speech-in-noise intelligibility", Neuroreport, 8, 1779-1783.

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

Kim, D. O., Dorn, P. A., Neely, S. T., and Gorga, M. P. (**2001**). “Adaptation of distortion product optoacoustic emission in humans,” J. Assoc. Res. Otolaryngol. **2**, 31-40.

Lee, C-Y., Glass, J., and Ghitza, O. (2011). “An efferent-inspired auditory model front-end for speech recognition,” Interspeech, Florence, Italy, 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, C. M. (**1988**). “Response properties of cochlear efferent neurons: Monaural versus binaural stimulation and the effects of noise,” J. Neurophysiol. **60**, 1779-1789.

Lopez-Poveda, E. A., Eustaquio-Martin, A., Stohl, J. S., Wolford, R. D., Schatzer, R., and Wilson, B. S. (2016). “A binaural cochlear implant sound coding strategy inspired by the contralateral medial olivocochlear reflex,” Ear and Hear. 37, e138-148.

Lopez-Poveda, E. A., Eustaquio-Martin, A., Stohl, J. S., Wolford, R. D., Schatzer, R., Gorospe, J. M., Ruiz, S. S. C., Benito, F., and Wilson, B. S. (2017). “Intelligibility in speech maskers with a binaural cochlear implant sound coding strategy inspired by the contralateral medial olivocochlear reflex,” Hear Res. 348, 134-137.

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.

Murugasu, E., and Russell, I. J. (**1996**). “The effect of efferent stimulation on basilar membrane displacement in the basal turn of the guinea pig cochlea,” J. Neurosci. **16**, 325–332

Yasin, I., Drga, V. and Plack, C. J. (**2013**). "Estimating peripheral gain and compression using fixed-duration masking curves," J. Acoust. Soc. Am. **133**, 4145-4155.

Yasin, I., Drga, V. and Plack, C. J. (**2014**). "Effect of human efferent feedback on cochlear gain and compression," J. Neurosci. **34**, 15319-15326.

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 21/07/2016).