An Efferent-Inspired Auditory Model Front-End for Speech Recognition

Chia-ying Lee, James Glass

MIT Computer Science and
Artificial Intelligence Laboratory
Cambridge, Massachusetts 02139, USA
{chiaying, glass}@csail.mit.edu

Oded Ghitza

Boston University
Hearing Research Center
Boston, Massachusetts 02215, USA
oghitza@bu.edu

Abstract

In this paper, we investigate a closed-loop auditory model and explore its potential as a feature representation for speech recognition. The closed-loop representation consists of an auditory-based, efferent-inspired feedback mechanism that regulates the operating point of a filter bank, thus enabling it to dynamically adapt to changing background noise. With dynamic adaptation, the closed-loop representation demonstrates an ability to compensate for the effects of noise on speech, and generates a consistent feature representation for speech when contaminated by different kinds of noises. Our preliminary experimental results indicate that the efferent-inspired feedback mechanism enables the closed-loop auditory model to consistently improve word recognition accuracies, when compared with an open-loop representation, for mismatched training and test noise conditions in a connected digit recognition task.
Index Terms: efferent, auditory model, feature extraction

1. Introduction

Despite continuous progress in automatic speech recognition (ASR), the human ability to understand speech in the presence of noise is still considerably superior to current ASR technology, especially when the noise has not been previously seen, or in the case of dynamic noises [1]. There is considerable body of prior and ongoing research to develop statistical methods to compensate or adapt to new noise conditions [3, 4, 5, 6]. Much of this work incorporates well-known spectral representations such as Mel-frequency cepstral coefficients (MFCCs) or perceptual linear prediction (PLP) [7, 8]. However, it is also worthwhile to consider novel spectral representations as well. For example, as scientists learn more about the role of medial olivocochlear (MOC) efferent system in the human auditory system, it is reasonable to consider their potential use as a feedback mechanism for ASR [9]. In this paper, we investigate one such efferent-inspired auditory model, and explore its potential as a front-end representation for ASR.

Our work was inspired by the mounting evidence of the possible role of the MOC in the human cochlea [9], and experiments that integrated efferent feedback capabilities into auditory models to predict speech intelligibility in noise [10, 11, 13, 12, 14]. In [10, 11], the authors reported that by including a phenomenological representation of the MOC efferent pathway into an auditory model, they were able to match human confusion patterns for a task of speech discrimination in noise, and produce robust performance for varying levels of stationary additive noise. More recently, in [14], an auditory model, tuned manually to mimic efferent function, was tested as an ASR front-end and was shown to improve performance in the presence of noise, compared to the auditory model without efferent activity.

Despite the observed progress in previous studies, the concept of incorporating auditory neural feedback into an ASR front-end has not been widely considered. For example, the majority of auditory-based feature extraction algorithms that have been incorporated into an ASR have been feed-forward, or open-loop models [7, 8]. This observation motivated us to investigate the potential of integrating an efferent incorporated auditory model, or closed-loop model, into an ASR front-end, and measure the ability of the closed-loop model to process speech in the presence of unseen noise.

In this paper, we modified the efferent-inspired auditory model in [10, 11] to create a closed-loop feature extraction method, and we applied it as a front-end for an ASR task. Previous studies have systematically examined how efferent-incorporated auditory models improved speech intelligibility for speech in one particular type of noise with varying levels of intensities [10, 11, 14]. In this paper, we designed a connected digit recognition task that incorporated a variety of noises. We then tested the ability of the closed-loop model to cope with mismatched training and test noises for a digit recognition task. The purpose of this experimentation was to further explore the potential role for incorporating efferent-like feedback into ASR front-ends. To provide perspective, performance of the closed-loop feature extraction method was compared to that of the standard MFCC approach.

2. Nonlinear Closed-loop Auditory Model

For our research, we are using the closed-loop model described in [10], which was inspired by current evidence about the role of the efferent system in regulating the operating point of the cochlea. This regulation results in an auditory nerve (AN) representation that is less sensitive to changes in environmental conditions. In implementing the model, we use a bank of overlapping filters as cochlear channels, uniformly distributed along the equivalent rectangular bandwidth (ERB) scale, which cover the entire speech band. A block diagram of one closed-loop cochlear channel is shown in Figure 1. The upper path of the figure consists of the open-loop components, while the lower path contains the feedback mechanism. In the following sections, we describe the open-loop components, the feedback mechanism, and the feature extraction method used for subsequent recognition experiments.

2.1. The Open-loop Components

As indicated in the upper path in Figure 1, the open-loop components of each channel is comprised of 1) an multi-band path

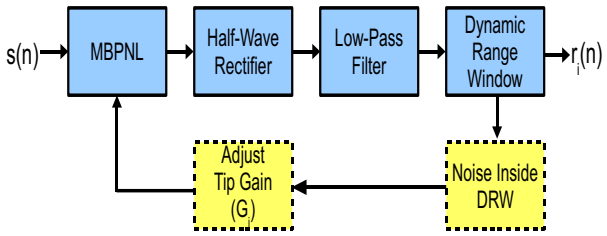

Figure 1: *The structure of one closed-loop cochlear channel.*

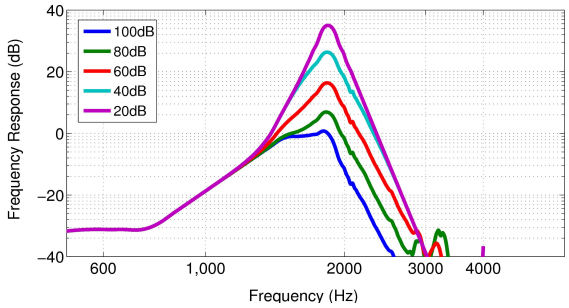

Figure 2: *The frequency response of an MBPNL filter with center frequency of 1820 Hz, for input intensities ranging from 20-100 dB.*

non-linear (MBPNL) filter, 2) a generic model of the inner hair cell (IHC), and 3) a dynamic range window (DRW). The IHC component consists of a half-wave rectifier followed by a low-pass filter, representing the reduction of synchrony with center frequency of the cochlear channel. The DRW component acts as a hard limiter, with lower and upper bounds, representing the dynamic range of the simulated IHC response, which also reflects the observed dynamic range at the AN level.

The MBPNL filter is Goldstein's model of nonlinear cochlear mechanics [15]. This model operates in the time domain and changes its gain and bandwidth with changes in the input intensity, in accordance with observed physiological and psychophysical behavior. An illustration of the changing characteristics of an MBPNL filter is shown in Figure 2, which plots the frequency response of an MBPNL filter with a center frequency of 1820 Hz and the tip gain setting to 40 dB for different input intensities [10, 11]. There are two important nonlinear characteristics of the MBPNL model that should be pointed out. First, the gain of the filter increases when the input intensity decreases. Second, the bandwidth of the model decreases as the input intensity decreases. These features are helpful for reducing the dynamic range of the output signal, and mimic the behavior of the human cochlea.

2.2. The Feedback Mechanism

In [10], the efferent-inspired feedback mechanism is introduced by modeling the effect of the medial olivocochlear efferent path. Morphologically, MOC neurons project to different places along the cochlear partition in a tonotopic manner, making synapse connections to the outer hair cells and, hence, affecting the mechanical properties of the cochlea (e.g. increasing basilar membrane stiffness) [16].

This effect-inspired feedback is realized by introducing a frequency dependent feedback mechanism which controls the tip gain, G_i, of each MBPNL filter, as shown in Figure 1. The tip gain of each cochlear channel is adjusted according to the intensity level of the sustained background noise in that frequency band, as measured at the output of the DRW. Hence, the gain per frequency channel, G_i is slowly changing, following long-term changes in the noise spectral distribution. To estimate the gain profile we assume a long enough time-window that is signal free (i.e. which contains only noise) which can be estimated from background noise. We pass the noise signal through the open-loop model with the DRW lower bound fixed. As suggested in [10, 11], we adjust the gain until the average noise energy is just above the lower bound of DRW by a prescribed value, ϵ, such as 1 dB. More specifically, suppose G_i is the gain for the i^{th} filter in the filter bank, and X_{G_i} is the noise energy we observe at the output of the i^{th} channel after the filter is multiplied with G_i. Let the lower bound of DRW be Y. Then we select G_i^*, the gain for the i^{th} filter, to be a value that satisfies the following equation:

$$G_i^* = \arg_{G_i} |X_{G_i} - Y| < \epsilon \; (dB) \tag{1}$$

Note that the value of G_i^* is not unique. Based on our experimental results, G_i^* can be set to any value that satisfies Equation 1, since different valid values of G_i^* did not cause significant performance differences.

2.3. Feature Generation

The feature generation method used for our recognition experiments follows the standard MFCC extraction process [7]. The MFCCs were generated every 10ms via short-time fourier transform (STFT) using a sliding 25ms Hamming window. Mel-frequency spectral coefficients were generated by summing STFT magnitudes weighted by triangular Mel-spaced filters, and then converted to dB. Final MFCCs were created via the discrete cosine transform (DCT) to produce a 13 dimensional vector. With the exception of the Mel-warping step, the auditory model outputs were processed in a similar fashion to also produce a 13 dimensional feature vector every 10ms.

Figure 3 illustrates the responses of the output of the open-loop and the closed-loop MBPNL models to an input digit sequence under a variety of noise conditions including speech-shaped noise, white noise, pink noise, train noise, and subway noise. From the figure, we can see that the features generated by the closed-loop MBPNL model appear to be more consistent than those generated by the open-loop model.

3. Experimental Methodology

The speech recognition experiments we performed to assess the closed-loop MBPNL model were based on connected digits contaminated with a variety of noises. We expanded the experimental setup in [14] to include conditions where we could examine the capability of the closed-loop model for handling various noises. We synthesized noisy digits and created training and test sets with five different noise types at the following conditions: 1) the noise levels are held fixed at 70 dB SPL, 2) speech is added at a particular SNR level relative to the background noise SPL, and 3) a 300ms interval of background noise precedes the digit sequence to enable the closed-loop model to adapt to the background noise. The 300ms noise interval was used to automatically compute the gains, G_i, for each channel.

The TIDigits corpus was partitioned into a 6,752 utterance training set, and a 1,001 utterance test set. There was no over-

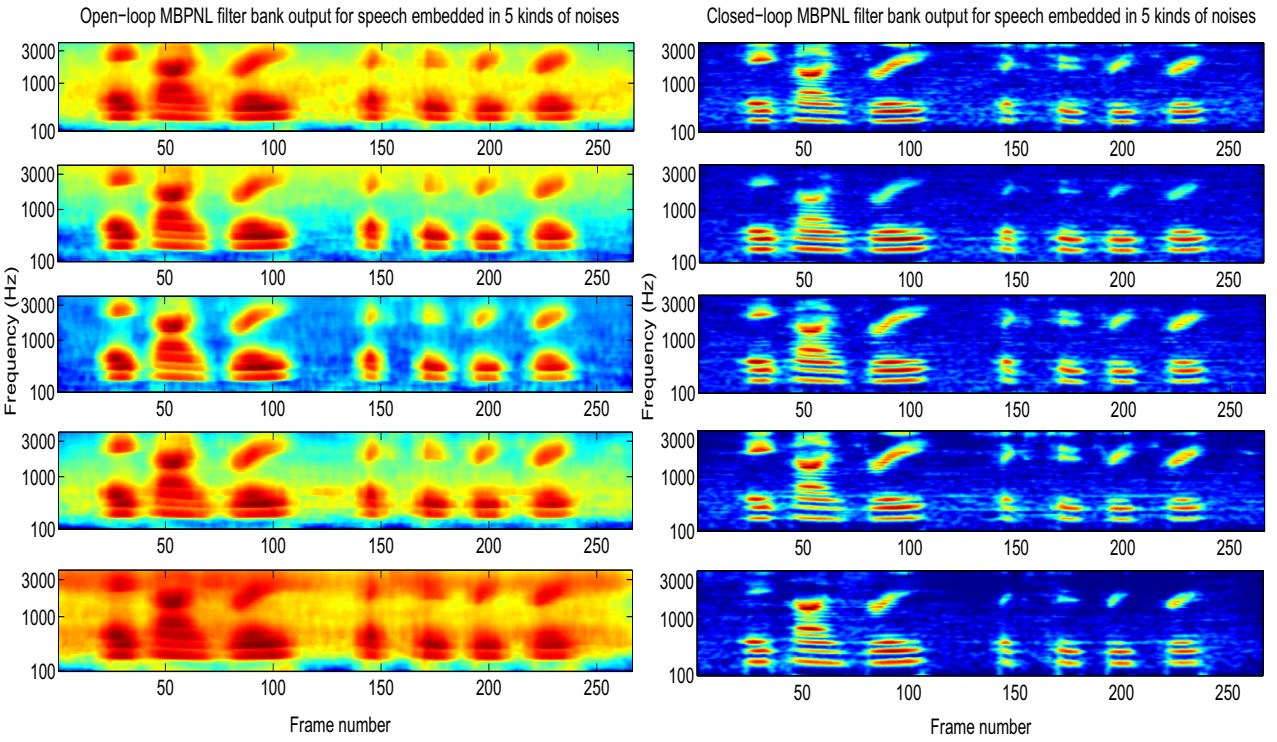

Figure 3: *The spectral representations generated by the open-loop MBPNL model baseline (left) and the closed-loop MBPNL model (right) for a digit sequence contaminated with (from top to bottom) speech-shaped, white, pink, train, and subway noise at 10 dB SNR level.*

lap between any training and testing utterances. Five different types of background noise were used for contaminating the digit recordings. In addition to pink noise and speech-shaped noise that were investigated in prior research with efferent-inspired auditory models [10, 14], we also investigated white noise, as another stationary noise for our experiment [18]. In addition to stationary noises, two non-stationary noises, train and subway noise, were randomly picked from the Aurora2 [2] database and included in our experimental setup.

In order to investigate the potential of the closed-loop for coping with various kinds of noises, we performed a series of jackknifing experiments with mismatched training and test conditions. For each experimental condition, one of the noise types was designated as the test condition. The training data was then partitioned into sixteen subsets that represented the cross-product between the four remaining noise types and four different SNRs of 5, 10, 15, and 20 dB. The resulting training set thus represented different noises and multiple SNR conditions. For each selected noise test condition, four different SNR conditions were tested (5, 10, 15, 20 dB SNR). In this case, the entire test set was configured with the test noise condition and a fixed SNR level. Thus, we performed a total of 20 (5 noise types x 4 SNR conditions) recognition experiments. ITU software was used to determine noise and signal energy levels and the appropriate signal gain needed for a particular SNR level [19].

Once the training and testing datasets had been created, the remainder of the experiments followed the standard Aurora convention [2]. For the recognition tasks, both the open-loop and the closed-loop models generated a 42-dimensional feature vector made up of energy and 13 DCT coefficients, as well as their first- and second-order time derivatives. The standard Aurora HMM-based speech recognizer was used for these experiments.

To provide a baseline comparison to the open-loop and closed-loop models, the standard MFCC representation was also subjected to the same training and testing conditions. To generate the standard MFCC, the synthetic digit data was first normalized to the maximum utterance value for each utterance as is common practice.

4. Experimental Results

Tables 1- 3 report digit correctness for the twenty (5x4) mismatched recognition scenarios described in Section 3 for the three different feature representations. Although we report correctness in this paper, digit accuracies for this series of experiments hold the same trend. Each table column specifies a test noise condition (i.e., that was not used for training), while each row indicates the test SNR condition. Table 4 summarizes the twenty experimental results for each of the three representations by showing the average and the standard deviation of the ASR correctness obtained for all twenty test conditions.

From the recognition performances shown in Tables 1- 4, the closed-loop model provided the best correctness and the smallest performance variance compared to the other two representations over all testing conditions. Specifically, the closed-loop model achieved a 25% and 41% relative improvement in word error rate compared to the MFCC baseline and the open-loop model baseline, respectively. While the open-loop model performed worse than the MFCC baseline, which we believe was due to speech normalization used in MFCC, the closed-loop version clearly benefited from a feedback loop. The closed-loop model also clearly showed a robust tolerance to mismatched conditions, as its worst-case correctness was considerably superior to either of the other two models. The results show the po-

Condition	sp-shaped	white	pink	train	subway	Avg
20 dB SNR	95	94	95	95	94	95
15 dB SNR	95	93	94	94	94	94
10 dB SNR	91	89	92	93	91	91
5 dB SNR	81	77	86	86	82	82
Avg	90	88	92	92	90	90

Table 1: *Digit recognition correctness (%) produced by the MFCC baseline representation for mismatched training and test noise conditions. Each entry represents a test condition; the corresponding training conditions are the crossproduct between the four remaining noise types and all SNRs. Sp-shaped stands for speech-shaped noise.*

Condition	sp-shaped	white	pink	train	subway	Avg
20 dB SNR	96	93	90	96	89	93
15 dB SNR	95	90	87	96	88	91
10 dB SNR	94	83	82	93	85	87
5 dB SNR	88	71	71	86	79	79
Avg	93	84	82	93	86	88

Table 2: *Digit recognition correctness (%) produced by the open-loop model for mismatched training and test noise conditions.*

Condition	sp-shaped	white	pink	train	subway	Avg
20 dB SNR	97	95	96	97	96	96
15 dB SNR	97	94	97	97	95	96
10 dB SNR	95	91	96	95	92	92
5 dB SNR	84	84	92	85	82	85
Avg	93	91	95	94	91	93

Table 3: *Digit recognition correctness (%) produced by the closed-loop model for mismatched training and test conditions.*

Representation	MFCC	Open-loop	Closed-loop
Average (%)	90	88	93
Deviation (%)	5.30	7.47	5.01

Table 4: *Summary of average and deviation of digit recognition correctness shown in Tables 1- 3 produced by the MFCC, open-loop, and closed-loop representations.*

tential of the closed-loop MBPNL model for generating consistent speech representations across varying background noises.

5. Conclusion

In this work, we have explored the use of a closed-loop auditory model as a front-end for speech recognition. The model is inspired by the role of the auditory MOC efferent mechanism, which feeds back to the cochlea. In our realization, the feedback mechanism controls the gain of a nonlinear model of cochlear mechanics (MBPNL), which enables the model to dynamically adapt to changing noise conditions. After being evaluated on a noisy digit recognition task and compared to a standard MFCC baseline front-end and an open-loop version of the MBPNL model, the closed-loop model showed the best and the most consistent digit recognition performance across a variety of mismatched training and testing conditions and SNR levels. We believe these results indicate that that representations of speech that incorporate feedback show promise for generating robust speech features and are worthy of further investigation on other speech recognition tasks.

6. Acknowledgement

The authors would like to thank David Messing for useful discussions. This work is funded in part by the T-Party Project, a joint research program between MIT and Quanta Computer Inc., Taiwan, and by a grant form the Air Force Office of Scientific Research (AFOSR).

7. References

[1] R. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 1, pp. 1–15, July 1997.

[2] D. Pearce, H. Hirsch, and Ericsson Eurolab Deutschland Gmbh, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *ISCA ITRW ASR*, pp. 29–32, 2000.

[3] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 29, no. 2, pp. 254–272, Apr. 1981.

[4] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1-3, pp. 133–147, 1998.

[5] A. de la Torre, A. Peinado, J. Segura, J. Perez-Cordoba, M. Benitez, and A. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 3, pp. 355 – 366, May 2005.

[6] C. Chen and J. Bilmes, "MVA processing of speech features," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 257–270, Jan. 2007.

[7] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on, Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[8] H. Hermansky and A. Cox, "Perceptual linear predictive (PLP) analysis-resynthesis technique," *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 0_37–0_38, 1991.

[9] N. Kiang, J. Guinan, M. Liberman, M. Brown, and D. Eddington, "Feedback control mechanisms of the auditory periphery: implication for cochlear implants," *International Cochlear Implant Symposium*, 1987.

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

[11] O. Ghitza, "Using auditory feedback and rhythmicity for diphone discrimination of degraded speech," in *Proc. ICPhS*, Aug. 2007, pp. 163–168.

[12] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 1, pp. 115–132, Jan. 1994.

[13] O. Ghitza, "Auditory neural feedback as a basis for speech processing," in *Proc. ICASSP*, Apr. 1988, pp. 91 –94 vol.1.

[14] G. Brown, R. Ferry, and R. Meddis, "A computer model of auditory efferent suppression: Implications for the recognition of speech in noise," *The Journal of the Acoustical Society of America*, vol. 127, no. 2, pp. 943–954, 2010.

[15] J. Goldstein, "Modeling rapid waveform compression on the basilar membrane as multiple-bandpass-nonlinearity filtering," *Hearing Research*, vol. 49, no. 1-3, pp. 39–60, Nov. 1990.

[16] J. Guinan, "Physiology of olivocochlear efferents," *The Cochlea*, pp. 435 – 502, 1996.

[17] R. Leonard, "A database for speaker-independent digit recognition," in *Proc. ICASSP*, 1984, vol. 3, pp. 42.11.1–42.11.4.

[18] F. Nachbaur, "Audio system test files, http://www.dogstar.dantimax.dk/testwavs/," 2002.

[19] *ITU recommendation G.712, Transmission performance characteristics of pulse code modulation channels*, Nov. 1996.