The value of auditory offset adaptation and appropriate acoustic modeling

H a Wa g[1,4,5], Da id Gelbart[2], Ha s-G¨ ter Hirsch[3], Wer er Hemmert[4,5]

[1]Infineon Technologies AG, Germany [2]International Computer Science Institute, USA
[3]Niederrhein University of Applied Sciences, Germany [4]Technische Universität München, Germany
[5]Bernstein Center for Computational Neuroscience, Germany

huan.wang@infineon.com, david.gelbart@gmail.com
hans-guenter.hirsch@hs-niederrhein.de, werner.hemmert@tum.de

Abstract

A critical step in encoding sound for neuronal processing occurs when the analog pressure wave is coded into discrete nerve-action potentials. Recent pool models of the inner hair cell synapse do not reproduce the dead time period after an intense stimulus, so we used visual inspection and automatic speech recognition (ASR) to investigate an offset adaptation (OA) model proposed by Zhang et al. [1].

OA improved phase locking in the auditory nerve (AN) and raised ASR accuracy for features derived from AN fibers (ANFs). We also found that OA is crucial for auditory processing by onset neurons (ONs) in the next neuronal stage, the auditory brainstem.

Multi-layer perceptrons (MLPs) performed much better than standard Gaussian mixture models (GMMs) for both our ANF-based and ON-based auditory features. Similar results were previously obtained with MSG (Modulation-filtered SpectroGram) auditory features[2]. Thus we believe researchers working with novel features should consider trying MLPs.

Index Terms: offset adaptation, auditory sound processing, feature extraction, acoustic modeling

1. Introduction

One of the most critical processing steps during auditory sound processing occurs at the inner hair cell (IHC) synapse: here the mechanically pre-filtered analog sound signal is converted into discrete nerve action potentials which propagate along the auditory nerve fibers (ANFs) to the brain. This conversion induces massive information loss – or to phrase it positively – information reduction. As any information lost during this process is no longer available for neuronal processing, it is important to understand and model the underlying principles correctly.

In our previous work, we have developed a model of human auditory sound processing, which codes sound signals into trains of nerve-action potentials. The model includes outer- and middle ear frequency responses, inner ear hydrodynamics, a compression stage, IHCs, and ANFs [3]. We used automatic speech recognition (ASR) tools to measure how well our model codes speech in noise [3]. When we investigated models of onset neurons (ONs) in the auditory brainstem, we discovered that we could not obtain realistic responses to amplitude modulated signals above 3 kHz. Detailed analysis of the responses revealed a flaw in the synapse model between inner hair cells and the auditory nerve: recent pool models fail to reproduce a realistic offset adaptation [4, 5].

In this paper we therefore extend our model with a model of offset adaptation following the proposal of Zhang et al. [1] and analyze how it improves the coding of speech signals. We also compare the ASR performance of Gaussian mixture models (GMMs) and multi-layer perceptrons (MLPs) to see which handles features derived from our auditory model better.

2. Modeling synaptic adaptation

In this section, we focus on the synaptic processes between IHCs and the auditory nerve.

The observed time course in the firing rate of ANF responses to tone bursts can be characterized by two exponential components [6] (see Fig. 2):

$$R_{on}(t) = A_{sus} + A_r e^{-t/\tau_r} + A_{st} e^{-t/\tau_{st}} \quad (1)$$

where A_r and A_{st} are two exponential components of rapid and short term adaptation, τ_r and τ_{st} are the respective decay time constants, and A_{sus} is a steady-state component.

We implemented a pool model as proposed by Meddis [4, 7]. The model has three reservoirs: the immediate store (q), the synaptic cleft (c), and the reprocessing store (w) (see Fig. 1). The model output is proportional to the rate of transmitter release from the immediate store (q) to the synaptic cleft (c), given by $k(t)q(t)$, where $k(t)$ is the only stimulus dependent variable. $k(t)$ describes the fusion rate of synaptic vesicles (mediated by Ca^{2+}-influx into the cell) which is specified as a function of intracellular IHC voltage [5].

The model output can be represented by Eq. 3 in the Laplace domain and solved analytically using a high-frequency tone burst as the stimulus input. The IHC voltage is assumed to be constant after the onset (denoted as k_2). $q(0^-)$,$c(0^-)$, and $w(0^-)$ are the reservoir concentrations before the onset. For more details please refer to the paper from Zhang et al. [1].

The resulting characteristic function of $q(t)$ is:

$$q(t) = \Phi_0 + \Phi_1 e^{-t/\tau_1} + \Phi_2 e^{-t/\tau_2} \quad (2)$$

where $-1/\tau_i$ are poles of $Q(s)$. The values τ_i and Φ_i can be calculated from $Q(s)$ directly.

Note that Eq. 2 has two exponential components which are the same as in Eq. 1. Therefore, the analytical relationship between model parameter and adaptation characteristics is established. Any desired adaptation responses can be achieved using appropriate model parameters, and vice versa.

$$Q(s) = \frac{(sq(0^-) + yM)(s+x)(s+l+r) + c(0^-)rxs + w(0^-)xs(s+l+r)}{s(s+x)(s+y+k_2)(s+l+r) - k_2 rxs} \quad (3)$$

Accepted after peer review of full paper

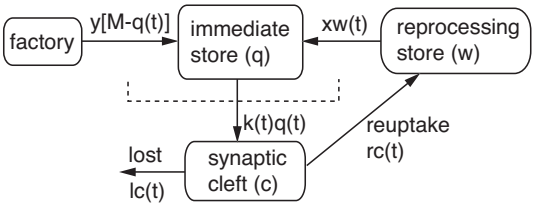

Figure 1: Schematics of the IHC-AN model. The immediate store q (maximum size: M) is refilled by the global pool at a rate of $y[M - q(t)]$ and by the reprocessing store (w) at a rate of $xw(t)$. The transmitter in the synaptic cleft c is lost at a rate of $lc(t)$ or recycled at a rate of $rc(t)$ into the reprocessing store.

The pool model uses the same characteristic function for both offset and onset adaptation. Since $q(t)$ can not be negative in Eq. 2, the model can not reproduce a "dead-time" period as found in physiological experiments. Instead, the rapid component of the model recovery function causes the synapse to recover immediately exponentially after stimulus offset.

In order to achieve a more physiologically consistent offset adaptation, Zhang et al. proposed adding a shift value A_{shift} to allow for negative output. In a next step negative outputs are set to 0, which represents the dead-time period. Therefore, the characteristic function of the output becomes:

$$R(t) = \max\{k(t)q(t) - A_{shift}, 0\}; \quad (4)$$

As we intend to improve offset adaptation while keeping onset adaptation as it is, the synaptic output becomes,

$$\begin{aligned} k(t)q(t) &= R_{on}(t) + A_{shift} \\ &= A_{shift} + A_{sus} + A_r e^{-t/\tau_r} + A_{st} e^{-t/\tau_{st}} \end{aligned} \quad (5)$$

The parameters of the model are then recalculated for the new adaptation parameters and A_{shift} is subtracted from $k(t)q(t)$ to get the output of the pool model, as in Eq. 4.

Thus, by including this shift, the same equation can be used to represent the onset and offset adaptation in the model. In summary, the new model achieved the desired offset adaptation while keeping onset adaptation untouched.

3. Feature extraction and ASR task

We derived features for ASR directly from discrete spike trains for 91 frequency channels of our model using a total of 17,200 high spontaneous rate (HSR) nerve fibers and 182 onset neurons. We temporally averaged the output of each channel with a 25 ms Hanning window advanced in 10 ms steps. The output derived from ANFs are very similar as conventional short-term spectra, whereas ONs code distinct temporal features, e.g., they are driven best by amplitude modulated stimuli. We applied a discrete cosine transform (DCT) to reduce the spectral resolution and to decorrelate the feature vectors. We kept the first 12 cepstral coefficients, including C0.

The automatic speech recognition tests were carried out on a version of ISOLET database with artificially added noise (noisy ISOLET). One of eight different noise types was added to each utterance at one of six different SNRs: clean, 20 dB, 15 dB, 10 dB, 5 dB and 0 dB. We kept the original division of the ISOLET database into five subsets, and used a five-way cross-validation to increase the statistical significance of our results [3]. We used the first of the five splits to tune SPRACHcore decoder options (we found that tuning HTK decoder options had no significant effect for this task). We reported results on the remaining four splits, 6240 words total.

We used two speech recognizers: one built with Cambridge's HTK using GMMs, and one built with SPRACHcore [8] using MLPs [9] . The recognition scripts and noisy ISOLET corpus that we used are available at `www.icsi.berkeley.edu/Speech/papers/eurospeech05-onset/isolet`. With HTK we used six states per word (one state for the pause model) and eight diagonal-covariance Gaussians per state, and with SPRACHcore we used 1600 MLP hidden units. In earlier work (not using identical features), we found that increasing GMM and MLP acoustic model sizes beyond this had only a minor effect.

We augmented the feature vectors with first and second order delta coefficients, calculated over nine frames (four frames each for past and future context). When using HTK for our auditory features (not for MFCC features[1]), to make the features easier for a GMM to model, we gaussianized the feature distributions prior to delta calculation, using the SPRACHcore pfile_gaussian tool. MLPs used a five-frame context window.

We tried MLP acoustic modeling because we hoped avoiding the statistical assumptions of the GMM approach would provide useful flexibility when working with our model-based features [10]. While we did try to decorrelate and Gaussianize our features for the GMMs, we knew this might not be enough – for example, those transformations worked with global distributions while GMMs model state-conditional distributions.

4. Results

The modified IHC-AN synaptic model produces the same onset adaptation but a physiologically more realistic offset adaptation. The time constants for rapid adaptation and short term adaptation are 1 ms and 54.7 ms, respectively. Fig. 2 compares the traditional and enhanced model of adaptation with physiological data. Traditional adaptation model only showed a depression of spontaneous activity and an exponential recovery after a tone burst. For the enhanced adaptation model, ANF responses were silenced during the dead-time period after signal offset and then slowly recovered to spontaneous activity, in accordance with physiological experiments [11][2].

Analysis of the synchronization index showed that the modified auditory nerve fiber model generates more precise phase locking to amplitude modulated stimuli. The synchronization index takes values from zero, where all spikes occur randomly throughout the stimulus, to 1, where all spikes are synchronized to the stimulus.Both models achieve high synchronization indices in the low frequency region ($\leq$1 kHz) with values in the range of measurements [12]. In the original model, the synchronization index degrades drastically in the frequency range above 1 kHz and lies far below experimental data (compare Fig. 4). The model of offset adaptation greatly improved the synchronization and the fit to experimental data.

The enhanced phase-locking of ANF responses is vital for further neuronal processing stages. We connected

[1]With gaussianization ASR results slightly improved for auditory features, but worsened a little for MFCC features. See also results for MSG features [10].

[2]Note that the enhanced adaptation model adds only one parameter, A_{shift}. The model is tuned to keep the spontaneous rate of the fibre constant; the driven rate changes slightly.

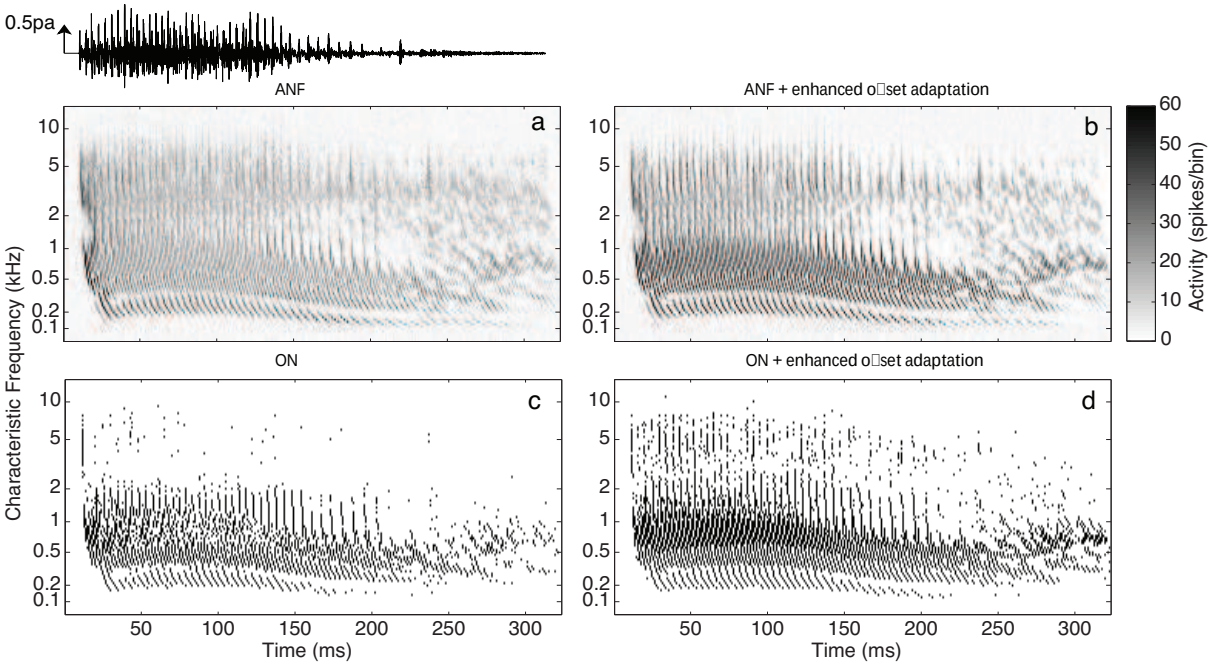

Figure 3: Responses from ANFs (upper row) and ONs (lower row) for our model of auditory sound processing with (right column) and without (left column) offset adaptation. For each frequency channel, we have plotted responses of 60 ANFs and one ON innervated by them. The stimulus was an "a" from a female speaker (ISOLET). The number of spikes falling in 1 ms time bins was represented in gray scale for the ANF response (see color bar top right).

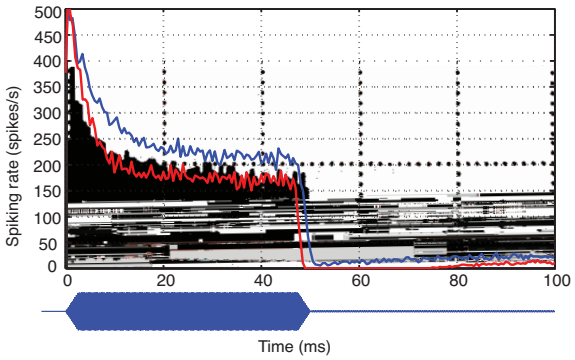

Figure 2: Response of an auditory nerve fiber (black area, from [11]) to a tone burst (characteristic frequency 10.34 kHz, sound pressure 39 dB re 20 μPa) and model output with enhanced offset adaptation (red line). After the tone burst the spontaneous activity is silenced for about 15 ms and recovers slowly thereafter, which is not predicted by the adaptation model of Sumner et al. [5] (blue line).

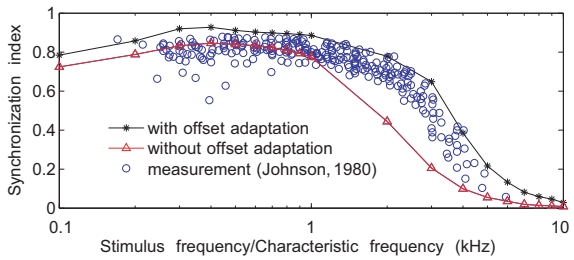

Figure 4: Synchronization index of auditory nerve action potentials. Input stimuli are pure tones at different frequencies. Synchronization indexes were calculated from output spike trains of neurons whose characteristic frequencies corresponded to the input stimuli. Modeled results are shown with and without adaptation, together with measurements from Johnson [12].

cochlear nucleus ONs, modeled with a detailed Hodgkin-Huxley model [13], to the ANFs, and found that they responded in the frequency region above 3 kHz only when offset adaptation was included in the IHC-AN model (compare panels c and d in Fig. 3). ONs require a quiet period of about $1-2$ ms before they fire [14], an effect for which offset adaptation is essential.

Fig. 5 shows speech recognition results as a function of SNR using features extracted from auditory nerve (panel a) and onset neuron (panel b) spike-trains. As onset neurons respond more strongly to voiced speech, we tested them only on the vowel subset (a, e, i,o, u and y) of ISOLET.

Using MLPs instead of HTK resulted in major performance improvements for all our auditory model-based features (see Table 1). All the improvements were statistically significant using a difference of proportions significance test ($P < 0.0001$). With MFCC features, for the full set there was no statistically significant difference between MLP and HTK, while for the vowel subset using HTK reduced word error rate (WER) considerably (this was statistically significant, $P < 0.002$).

Including the offset adaptation model resulted in very large performance improvements for features derived from ONs, and large improvements for features derived from ANFs. This shows the enhanced offset adaptation not only provided more useful input for ONs but also improved speech coding

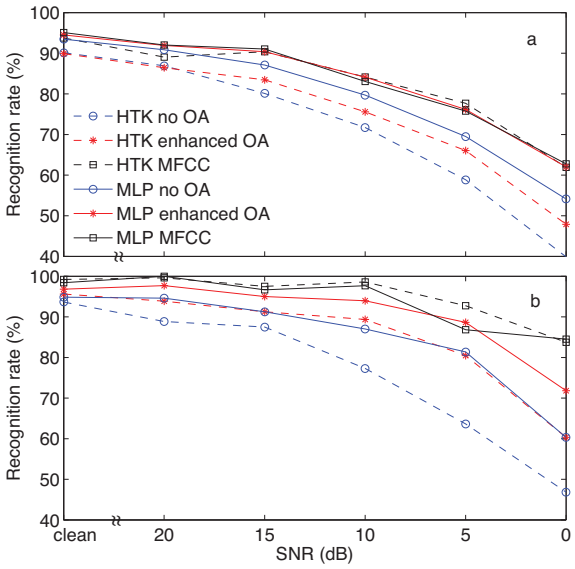

Figure 5: Speech recognition results as a function of SNR. Panel a shows results on the full noisy ISOLET task for features derived from 17,200 ANFs. Panel b shows results on the vowel subset (a, e, i, o, u and y) for features derived from 182 ONs.

		HTK	MLP
Full noisy ISOLET (17200 HSR ANFs)	MFCC	82.8	83.3
	no OA	71.3	79.1
	enhanced OA	74.9	83.2
Vowel subset (182 ONs)	MFCC	95.2	94.0
	no OA	76.3	84.9
	enhanced OA	85.1	90.7

Table 1: Recognition results (OA: offset adaptation)

per se. All the improvements were statistically significant ($P < 0.0001$). With the MLP, ANF features with offset adaptation performed similarly to MFCC features.

5. Discussion

Physiological auditory nerve measurements show offset adaptation with a "dead-time" period following the end of a tone burst. This effect is not replicated by commonly used pool models of synaptic transmission, which predict an immediate exponential recovery without a "dead-time" period.

We used an improved model of offset adaptation primarily because without it onset neurons located in the auditory brainstem were not responsive in the frequency region above 3 kHz We found that offset adaptation also improved phase locking of ANFs, and ASR results showed it improved speech coding by ANFs. We believe that these improvements in ASR performance are caused by shifting the working point of the synapse by offset adaptation especially during intense stimuli. This enhances the dynamic range of the synapse and the coding of amplitude modulations of speech sounds. As a result, the ONs responded more strongly, especially above 3 kHz, and the ASR performance of the ON features greatly improved.

Another important finding is that MLPs performed much better than GMMs for both ANF-based and ON-based auditory features. This is consistent with past results on MSG auditory features [10, 2]. However, when using MFCCs on the vowel subset, GMMs outperformed the MLPs. MLPs are also very easy to use in a multi-stream approach, something we hope to exploit in the future to combine features derived from different groups of neurons. And other researchers have found a tandem MLP/GMM approach to be an effective way of incorporating auditory-inspired TRAPs features into a GMM system, while still taking advantage of the GMM system's strengths such as speaker adaptation [15]. Thus we believe researchers working with novel features should consider trying MLPs. Our MLP and HTK ISOLET recognition scripts are available online for use with other features.

6. Acknowledgements

This work was funded by the German Federal Ministry of Education and Research (BCCN Munich, reference numbers 01GQ0441 and 01GQ0443 and the SmartWeb project).

7. References

[1] X. Zhang and L. H. Carney, "Analysis of models for the synapse between the inner hair cell and the auditory nerve," *J. Aco st. Soc. Am.*, vol. 118, pp. 1540–53, 2005.

[2] S. Sharma, D. Ellis, S. Kajarekar, P. Jain, and H. Hermansky, "Feature extraction using non-linear transformation for robust speech recognition on the Aurora database," in *ICASSP*, 2000.

[3] M. Holmberg, D. Gelbart, and W. Hemmert, "Speech encoding in a model of peripheral auditory processing: Quantitative assessment by means of automatic speech recognition," *Speech Comm - icatio* , 2007.

[4] R. Meddis, "Simulation of mechanical to neural transduction in the auditory receptor," *J. Aco st. Soc. Am*, vol. 79, no. 3, pp. 702–711, 1986.

[5] C. J. Sumner, E. A. Lopez-Poveda, L. P. O'Mard, and R. Meddis, "A revised model of the inner-hair cell and auditory-nerve complex," *J. Aco st. Soc. Am.*, vol. 111, pp. 2178–88, May 2002.

[6] L. A. Westerman and R. L. Smith, "Rapid and short-term adaptation in auditory nerve responses," *Hear. Res.*, vol. 15, pp. 249–260, 1984.

[7] R. Meddis, "Simulation of auditory-neural transduction: Further studies," *J. Aco st. Soc. Am*, vol. 83, no. 3, pp. 1056–1063, 1988.

[8] D. Ellis, http://www.icsi.berkeley.edu/~dpwe/projects/sprach/sprachcore.html.

[9] N. Morgan and H. Bourlard, "Continuous speech recognition," *IEEE Sig al Processi g Maga i e*, vol. 12, pp. 24–42, 1995.

[10] http://www.icsi.berkeley.edu/~dpwe/respite/multistream/aurora1999.html.

[11] N. Y. S. Kiang, T. Watanabe, E. C. Thomas, and L. F. Clark, "Discharge patterns of single fibers in the cat's auditory nerve," MIT University Press, Cambridge, MA, Tech. Rep., 1965.

[12] D. Johnson, "The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones," *J Aco st Soc Am*, vol. 68, pp. 1115–1122, 1980.

[13] J. S. Rothman and P. B. Manis, "The roles potassium currents play in regulating the electrical activity of ventral cochlear nucleus neurons," *J. Ne roph siolog* , vol. 89, pp. 3097–3113, 2003.

[14] W. Hemmert, M. Holmberg, and U. Ramacher, "Temporal sound processing by cochlea nucleus octopus neurons," *Proc. ICANN 2005, LNCS 3696*, pp. 583–588, 2005.

[15] Q. Zhu, B. Chen, N. Morgan, and A. Stolcke, "On Using MLP Features in LVCSR," in *ICSLP*, 2004.