Coding of speech into nerve-action potentials

M. Holmberg[1], H. Wang[2], Michael Isik[3], W. Hemmert[4]
[1] *Oticon A/S, Smørum DK-2765, Denmark, Email: mhb@oticon.dk*
[2] *Infineon Technologies, Munich, Germany, Email: huan.wang.cn@gmail.com*
[3] *Technische Universität München, Germany, Email: michael.isik@tum.de*
[4] *Technische Universität München, Germany, Email: werner.hemmert@tum.de*

Introduction

One of the most critical processing steps during encoding of sound signals for neuronal processing happens when the analog pressure wave is coded into discrete nerve-action potentials. 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 quantitatively model the underlying principles. We have developed a detailed model of auditory processing, which codes sound signals into spike-trains of the auditory nerve fibers (ANF). We have also developed Hodgkin-Huxley models of cochlear nucleus neurons, which are driven by auditory nerve spike-trains. We analyze the quality of coding with the framework of automatic speech recognition and the temporal information processing capabilities with a method based on information theory. Our latest improvements in speech coding by introducing the effect of offset-adaptation together with an improved matching of neuronal features to the speech recognizer using an artificial neuronal network have lead to significant improvements of recognition scores, now reaching the values of successful technical feature extraction methods. Offset adaptation is also required to drive onset neurons (ON) in the cochlear nucleus, which are able to code temporal information with high precision. Our results provide quantitative insight into temporal processing strategies of neuronal processing and are highly relevant for our understanding of auditory information processing in normal hearing and hearing impaired persons. They also have important implications for hearing aids and automatic speech recognition systems.

Methods

Nonlinear basilar membrane model

We have developed a model of the peripheral auditory system which consists of a model of the outer- and middle ear, a traveling-wave model of the inner ear and nonlinear compression of inner ear vibrations. The model was especially developed to achieve the strong amplification of up to 80 dB found in physiological measurements of basilar membrane motion (reviewed in [1]), and the sharp auditory filters close to threshold as revealed in psychoacoustical measurements [2, 3]. The mechanical output (basilar membrane vibrations) of this model to three pure tones at different frequencies is plotted in Figure 1. At low levels, auditory filters are very narrow and almost symmetrical. Toward higher levels, response areas become broader and extend especially in the basal direction, an effect known as upper spread of masking. Note that the large dynamic range of the acoustic input (10–90 dB, corresponding to a ratio of 1:10000) is compressed to a much smaller increase of the basilar membrane vibration amplitude (from 1 to about 33 nm). This dynamic range compression is essential for the next steps, the mechano-electrical transduction of the sensory cells, the inner hair cells, and for the coding into nerve-action potentials, which both have only very limited dynamic ranges.

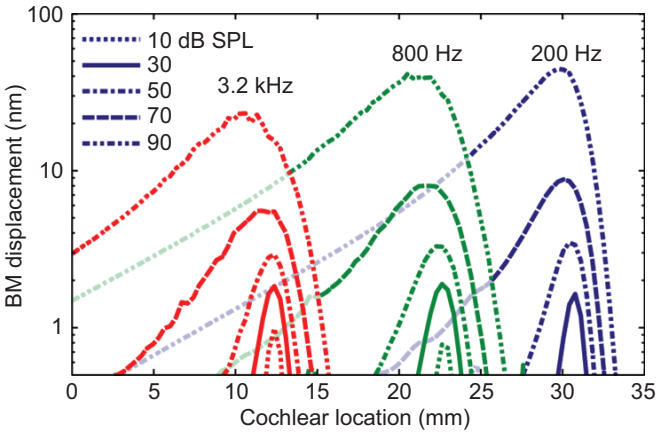

Figure 1: Modeled amplitudes of basilar membrane vibration of our inner ear model. The responses feature large dynamic range compression, narrow filter shapes at low levels and upper spread of excitation (and therefore masking) at high sound levels.

Coding into nerve-action potentials

Basilar membrane vibrations cause fluid motions which in turn drive the hair bundles of the inner-hair cells, the sensory cells in the inner ear. Inner hair cells convert mechanical stimuli into a receptor potential. At the efferent synapses of these cells, this electrical signal is converted into action potentials of the auditory nerve, which are propagated to the brain.

Coding of information by onset neurons

The model of the inner hair cells and auditory nerve synaptic complex was adopted from Sumner et. al.[4], which we complemented by a model of enhanced offset adaptation as proposed by Zhang et al. [5]. In Figure 2 the responses of 60 high spontaneous rate auditory nerve fibers per frequency channel are plotted for a spoken utterance [ou] (female speaker). Please note that we

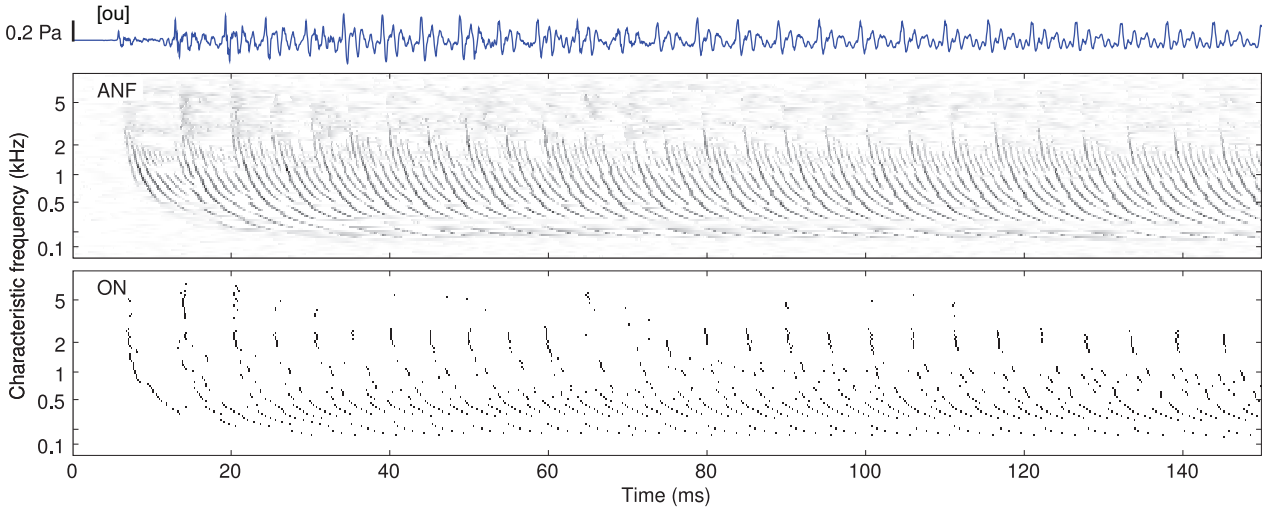

Figure 2: Modeled neuronal activity to the utterance [ou] from a female speaker (upper trace: sound pressure waveform). Responses were plotted for 60 ANFs per frequency channel (upper panel) and one ON innervated by them (lower panel). The number of spikes falling in 0.21 ms time bins was represented in gray scale for the ANF response. Note that the temporal structure of the speech sound is preserved in the activity pattern of the auditory nerve. Onset neurons lock precisely on the resolved harmonic frequencies (here up to the second harmonic at 400 Hz) and extract the pitch frequency above about 1 kHz.

took our speech samples from the ISOLET database, which is low-pass filtered with a cut-off frequency of 8 kHz. At low characteristic frequencies, the fundamental frequency and the first harmonics are resolved both in the "frequency" domain (more precise: the location in the inner ear) and in the time domain. In the frequency range above about 1 kHz, modulations corresponding to the pitch frequency of the acoustic input become more and more obvious in the neural responses. This indicates that spectral components are no longer completely resolved. At even higher frequencies, the speech formants instead of the individual harmonic frequencies are coded. Also in this frequency range, the speaker's pitch frequency is visible in the modulations of the neural response.

Modeling of onset neurons

We also modeled neurons in the auditory brainstem, which are innervated directly by primary auditory nerve fibers. We used a single-compartment model with Hodgkin-Huxley-type ion channels to model onset neurons based on the analysis of Rothman and Manis [6]. Onset neurons are coincidence detectors, we adjusted their threshold so that they fired only if at least 6 of the 60 nerve fibers innervating them fired synchronously. At the fundamental frequency and the second harmonic (approx. 200 Hz and 400 Hz, respectively), onset neurons fire during almost every cycle. At higher frequencies, they tend to fire with a high probability every pitch period, triggered by the amplitude modulations coded in the auditory nerve responses. We found that offset adaptation was essential to drive onset neurons in the frequency range above about 3 kHz with voiced speech signals (compare [7] data not shown here).

Results

As onset neurons are coincidence detectors, they are able to increase their spike-timing relative to auditory nerve fibers and they code especially temporal aspects of sound signals. To quantify their capability to code information, we used the framework of information theory (compare [8]). We extract the information by sampling the output of a neuron into time bins – 1 in case a spike is elicited and 0 elsewhere. From these sequences, we build binary words. In this study we used a bin width of 0.25 ms and analyzed a word length of 40 bits, covering a duration of 10 ms. We calculated responses from 18,000 trials to the same sound stimulus. We estimated the noise entropy from the variations of the neurons' output at a given signal instance and the total entropy from all possible words occurring during the stimulus. The transmitted information is the difference between the total entropy and the noise entropy (for more detail please refer to [8, 9]).

Figure 3 shows the spiking rate and how information is coded by onset neurons along the frequency axes. Onset neurons lock very precisely at almost every stimulus cycle of the fundamental frequency (approximately 200 Hz) and the second harmonic (400 Hz), where they reach their highest spiking rates.

Automatic speech recognition scores

At the third harmonic frequency, they fire more sporadically (215 spikes/s at 600 Hz). In the frequency range above about 1 kHz, they lock on the pitch frequency (firing rate is up to 150 Hz), again with high temporal precision. The firing rate at the 400 Hz CF is more than twice as high compared to the rate at pitch frequency but the information rate is only slightly higher. Therefore, onset neurons reached their highest bits-per-spike value at the fundamental frequency (about 4 bits/spike). At the second harmonic, this value was already smaller (2.3 bits/spike). Around 1 kHz, 2 kHz and 5 kHz, where onset neurons lock on the pitch frequency, values were between 3.5 and 4.5 bits/spike. The largest portion

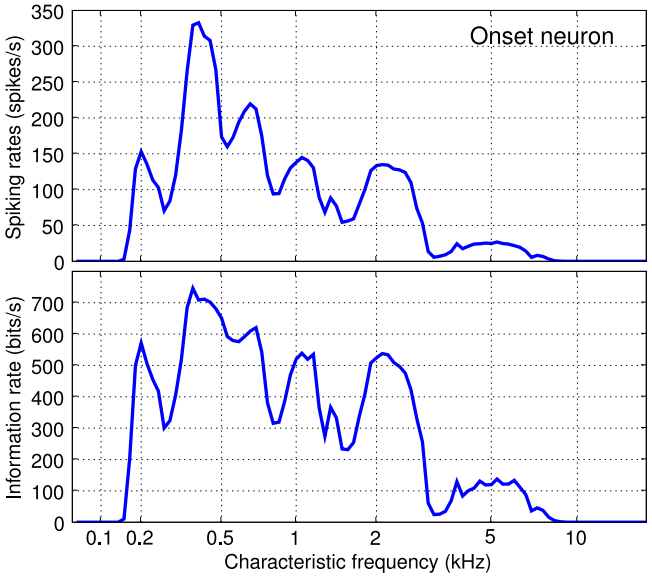

Figure 3: Average spiking rates (upper panel) and information rate of onset neurons for a vowel [ou] presented at 70 dB(A). Both spiking rate and information rate shows maxima at the fundamental frequency and its harmonics and around 1 kHz and 2 kHz, the formant frequencies of the vowel. The highest information rate was found at low frequencies (≈400 Hz), where spiking rate is highest and onset neurons phase locked on the stimulus.

of information (information rate reached values above 500 bits/s) concentrated at frequencies from about 200 Hz to 3 kHz, where the spikes precisely phase-locked to the pitch frequency or its higher harmonics. By changing the width of the time bins used for the information calculations, we found that onset neurons code the major portion of information with a temporal precision ranging from 0.2 to 4 ms (data not shown).

To complement our information theoretic approach, we also applied the framework of automatic speech recognition to assess how well speech can be discriminated using only features derived from neural activity. For speech recognition, we only exploited the rate-code: we counted the action potentials of 17,200 high spontaneous rate auditory nerve fibers (using 25 ms wide hamming windows). We applied a discrete cosine transform 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, [7]). We used two speech recognizers: one built with Cambridge's hidden Markov toolkit (HTK) using Gaussian mixture models (GMMs), and one built with SPRACHcore [10] using multi layer perceptrons (MLPs) [11]. 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. 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, to make the features easier for GMMs to model, we gaussianized the feature distributions prior to delta calculation, using the SPRACHcore pfile gaussian tool, which slightly improved automatic speech recognition (ASR) scores. The MLP used a five-frame context window.

Figure 4 shows speech recognition results as a function of SNR using features extracted from auditory nerve spike-trains. Using MLPs instead of HTK resulted in major performance improvements for all our auditory model-based features, while with MFCC features there was no statistically significant difference between MLP and HTK. Including the offset adaptation model resulted in large performance 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 per se. With the MLP, ANF features with offset adaptation performed as well as MFCC features.

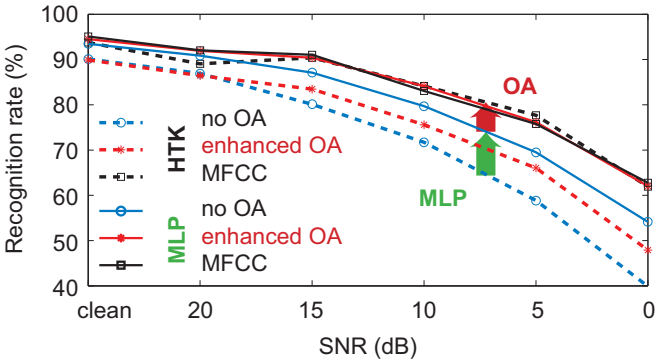

Figure 4: Speech recognition results as a function of SNR using the noisy ISOLET task for features derived from 17,200 ANFs. Offset adaptation (OA) enhances speech recognition scores considerably by 12.5% (HTK testbed) and 19.6% (MLP testbed), respectively. The MLP greatly improves interfacing of auditory-based features to the conventional ASR back end: the relative improvement in word error rate was 27.2% (no OA) and 33.1% (with OA), respectively.

We also performed similar experiments with 182 onset neurons, however, as they respond more strongly to voiced speech, we tested them only on the vowel subset (a, e, i, o, u and y) of ISOLET. Also in these experiments we saw similar improvements with offset adaptation and when we introduced the MLP acoustic modeling (data not shown).

Conclusions

We have contributed a human inner ear model that transforms arbitrary stimuli into auditory nerve action potentials. The basilar membrane model is able to reproduce the sharpness and shape of human auditory filters over a wide range of frequencies and levels. We also modeled responses of the auditory nerve and of selected neurons in the auditory brainstem. 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 therefore introduced 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. 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. To complement the investigations with automatic speech recognition, we also used the framework of information theory to calculate the information coded in spike trains of auditory neurons. Whereas with automatic speech recognition we can only investigate features with a relatively coarse temporal resolution, information theory captures all the information coded with high temporal resolution. We found that onset neurons lock on modulations and code temporal features with high precision (bit-rate). The major portion of information is coded with a temporal precision ranging from 0.2 to 4 ms. This finding is not only interesting for our understanding of neuronal sound processing but has also important implications for automatic speech recognition, where temporal information is usually processed only with a maximal precision of 10 ms.

References

[1] Mechanics of the mammalian cochlea, Physiol. Rev. **81** (2001), 1305–52

[2] Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements. Proc. Natl. Acad. Sci. USA **99(5)** (2002), 3318–3323

[3] Estimates of human cochlear tuning at low levels using forward and simultaneous masking. J. Assoc. Res. Otolaryngol. **4(4)** (2003), 541–54

[4] A revised model of the inner-hair cell and auditory-nerve complex, J. Acoust. Soc. Am. **111** (2002), 2178–88

[5] Analysis of models for the synapse between the inner hair cell and the auditory nerve, J. Acoust. Soc. Am. **118** (2005), 1540–1553

[6] The roles potassium currents play in regulating the electrical activity of ventral cochlear nucleus neurons, J. Neurophysiol. **89** (2003), 3097–3113

[7] The Value of Auditory Offset Adaptation and Appropriate Acoustic Modeling. Interspeech (2008), 902-905

[8] Auditory Information Coding by Cochlear Nucleus Onset Neurons. IEEE ICASSP (2005), 129-132

[9] Information theory and neural coding. Nature Neurosci. **2** (1999), 947-957

[10] Reference to the free software SPRACHcore. URL: `http://www.icsi.berkeley.edu/~dpwe/projects/sprach/sprachcore.html`

[11] Continuous Speech Recognition, IEEE Signal Processing Magazine **12** (1995), 24-42