AUDITORY INFORMATION CODING BY COCHLEAR NUCLEUS ONSET NEURONS

Huan Wang, Marcus Holmberg and Werner Hemmert

Infineon Technologies Inc, Corporate Research, Otto-Hahn-Ring 6, 81730 Munich, Germany

ABSTRACT

In this paper we use information theory to quantify the information in the output spike trains of modeled cochlear nucleus onset neurons. Onset neurons are known for their precise temporal processing, and they code the periodicity of voiced speech with high fidelity. We conclude that the maximum information transmission rate for a single neuron is close to 1000 bits/s, which corresponds to 3.26 bits/spike. For quasi-periodic signals like voiced speech, the transmitted information saturates with word duration, with 90% of the information being transmitted within 73 ms. Information theory also shows that the maximum temporal resolution of onset neurons is approximately 0.1 ms.

1. INTRODUCTION

Our auditory system performs a spectral decomposition of acoustic stimuli, and at the same term preserves temporal information. We are interested in how speech signals are encoded into spike trains by the neurons in the auditory pathway. What is the temporal resolution of this neural coding? How is the encoding affected by noise? And furthermore, how robust do neurons encode different frequencies in the speech signal? In this paper we use an auditory model and apply information theory to find answers to these questions and to evaluate the robustness of speech coding for so-called onset neurons (ON) located in cochlear nucleus, the first neuronal processing stage after the inner ear. ON have very specialized membrane properties, and respond with great precision to signal onsets. ON also extract the periodicity of voiced speech with high fidelity [1]. Information theory [2, 3, 4] provides us with quantitative tools to assess the information content of ON spike trains without making any assumption on the coding strategy.

2. MODELING

In this section we give a very brief overview of the model we use for the experiments.

2.1. Inner ear model: Coding of sound signals into trains of nerve action potentials

The model of the peripheral hearing system consists of a simplified middle ear model, a model of inner ear hydrodynamics followed by a compression stage, and sensory cells (see Fig.1). The hydrodynamics model effectively acts as a filter bank that spectrally decomposes acoustic stimuli. The compression stage models the so-called "cochlear amplifier", and gives the model up to fourth-root compression of the dynamic range and the high spectral resolution found in humans. The compression is crucial for sound coding in the sensory cell, the inner-hair cell, since it has a dynamic range of only 40 dB. The auditory nerve fibers (ANF) innervate the sensory cells and encode the stimuli in nerve action potentials or spike trains. The generation of a spike is modeled as a stochastic process. We tuned the model to reproduce recent psychoacoustic measures of frequency selectivity and compression [5].

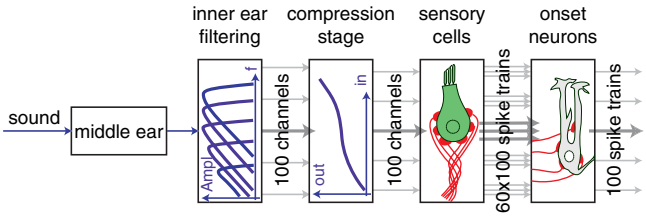

Fig. 1. Schematics of the auditory model. The model has 100 frequency channels. Each channel is coded by 60 auditory nerve fibers, which connect to one onset neuron per channel.

2.2. Model of onset neurons

We modeled Type II onset neurons [6] located in the brainstem (ventral cochlear nucleus) and connected them to 60 ANFs from our inner ear model (compare Fig. 1). Rothman and Manis [6] characterized the ion channels of the onset neurons. We used a single-compartmental model including five major Hodgkin-Huxley-type ion channels. We corrected conductance and time-constants to a body temperature of 38° and solved the differential equations in the time domain.

3. ALGORITHM FOR THE INFORMATION CALCULATION

We implemented an algorithm based on information theory for calculating information carried by the output spike trains of ON [3]. Let S denote the input stimulus and R the response of the ON; given the output spike trains of the ON, the information $I(R;S)$ they provide about the input stimulus (mutual- or transmitted information) is given by

$$I(R;S) = H(R) - H(R|S) \tag{1}$$

where the (overall) entropy $H(X)$ of a discrete random variable X is defined by

$$H(X) = -\sum_x p(x) \log_2 p(x) \tag{2}$$

and the conditional entropy $H(X|Y)$ of X given Y by

$$H(X|Y) = -\sum_y p(y) H(X|Y=y) \tag{3}$$

We recorded the timing of individual spikes with a sampling function $T(R)$ to represent the output spike trains. According to the data processing inequality, we have $I(R;S) \geq I(T(R);S)$. We repeatedly presented the same stimulus to our auditory model to build up a binary code book, where each row is a bit sequence that represents the output spike train in response to one repetition. We transformed the binary code book with different word length (L_w) into decimal words. The overall entropy is calculated from the whole decimal code book. The conditional entropy is calculated only from the words that are synchronized in time. In order to approximate $H(R|S)$ for infinite number of trials, we first use a small fraction of the total trials, and then gradually increased the fraction [3]. In this way we were able to extrapolate $H(R|S)$ from a finite number of trials. The transmitted entropy, $I(R;S)$, is the difference between overall- and conditional entropy (Equation 1).

3.1. Results

We present the utterance */ei/* with a sound pressure level of 70 dB(A) (female speaker, ISOLET fcmc0-A1) as input stimulus for all our figures. ISOLET recordings are band-limited to 8 kHz. The response of the ANF and ON along the whole length of the inner ear are plotted in Figure 2. Notice that the inner ear provides a spectral decomposition with approximately logarithmic resolution. ANF spike trains code both spectral- and temporal features of sounds. Frequency regions with high energy – like the formants – are coded with higher spike rates. The temporal fine structure is preserved in the precise spike timing. The increasing delay of the neuronal responses towards lower frequencies is due to propagation of the traveling-wave from the base to the apex of the inner ear. ON enhance the periodicity of voiced speech; they extract the pitch frequency very reliably in the CF region from 200 Hz - 2.5 kHz. In our model, they hardly fire for CFs above 2.5 kHz. In this frequency region the phase-locking of ANFs is lost and our model predicts ON to fail. We will discuss the reasons for this failure in a future paper and propose a modification in the spike generation of ANFs to solve this problem.

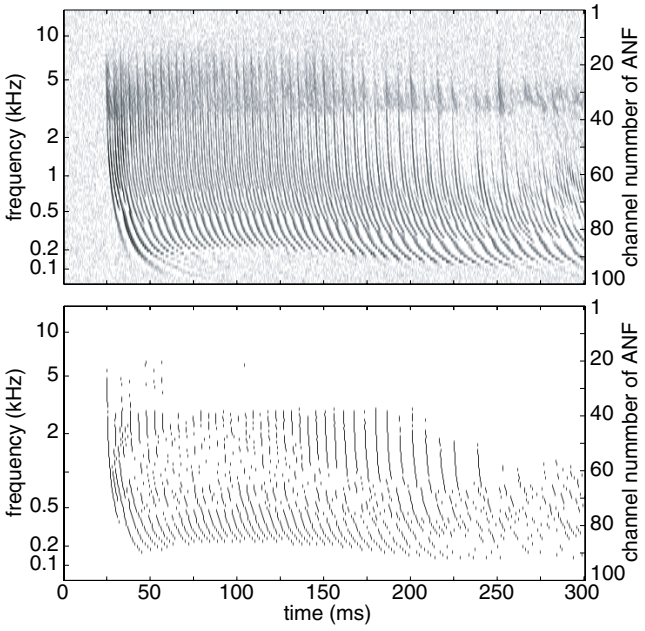

Fig. 2. Coding of speech (utterance */ei/*, female speaker, 70 dB(A)) into trains of nerve-action potentials of the auditory nerve (upper panel) and onset neurons (lower panel).

3.1.1. Information content in spike trains

Figure 3a plots entropy and conditional entropy of ON spike trains for the frequency channel number 47 with a characteristic frequency of 2 kHz. The conditional entropy, $H(R|S)$, increases linearly with word duration, as expected. To our surprise – and in contrast to other publications which assume constant entropy rates [3, 4] – we found an initial steep increase of the entropy $H(R)$, which levels off for longer word durations. As a consequence, the information rate decreases with increasing word duration. We can understand this behaviour both from a theoretical and from an intuitive point of view. Our input signal is a vowel which is a quasi-periodic signal, repeating itself – at least approximately – every pitch period. When a pitch period is coded with sufficient precision, no further information is transmitted. Intuitively, we know that we can classify vowels independent of how long they are pronounced – if their duration

just exceeds a minimum length. For a quantitative analysis, we fit the conditional entropy $H(R|S)$ as a linear function of word duration $D = L_w \cdot \Delta T$ (L_w is the word length in bits and ΔT the temporal quantization) (Fig. 4):

$$H(R|S) = r \cdot D \tag{4}$$

where r denotes the rate of conditional entropy in bits/s. We can also derive Equation 4 analytically. In ON spike trains, for a given stimulus the uncertainty lies mostly in the jitter of the spiking time. The conditional entropy can be seen as a sequence of variables which are almost independent and identical. According to information theory, if variables X_1, X_2,... X_n are independent and identical, their joint entropy can be calculated: $H(X_1, \ldots X_n) = n \cdot H(X_i)$. The dependence of the overall entropy on word duration is slightly more complicated. Whereas other investigations also assume a linear increase of the overall entropy with L_w, we see that conditional- and overall entropy grow with the same slope for long word durations (see Fig. 3a). This observation is supported by our notion that the transmitted information does not increase for long word duration due to the periodicity of voiced speech. If we make the assumption that the initial increase of the overall entropy is exponential, it is described by:

$$H(R) = r \cdot D - c \cdot exp(-D/\tau) + c \tag{5}$$

$D = L_w \cdot \Delta T$ stands for the word duration. The time constant τ and the distance c between $H(R)$ and $H(R|S)$ at infinite word duration were fitted with good agreement (Fig. 3a).

The transmitted information I can then be determined using Equation 1:

$$I = c - c \cdot exp(-D/\tau) \tag{6}$$

We can see the saturation effect only for word durations larger than about 5–10 ms; for shorter durations the information rate of the conditional entropy is approximately constant. As for larger durations (equal to longer wordlenghts) the number of trials for reliable estimations of the word probabilities increase exponentially, experimental studies are usually restricted to short durations because of limited recording times. In our computational model we can afford a large number of trials (6000) so that we were able to extend the wordlength to a maximum of 18 bits. Still, to estimate word durations up to 64 ms, we had to use a coarse temporal resolution of 4 ms. For this case, we clearly see the transmitted information saturating ($\tau = 32$ ms) with 90% of the information transmitted within the initial 73 ms (derived from fit function from $H(R|S)$ with 4 ms resolution).

3.1.2. Temporal resolution

We determine the temporal resolution of ON by down-sampling spike times to multiples of the sampling frequency. Of course

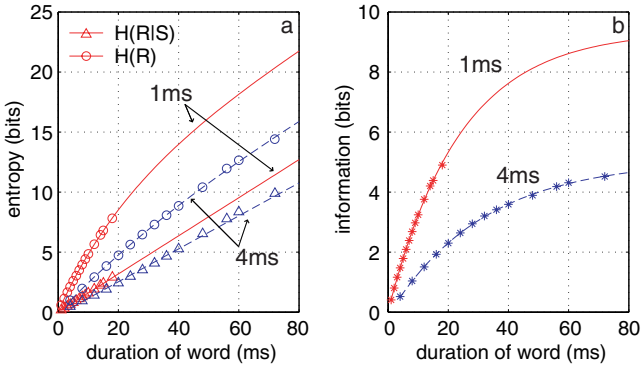

Fig. 3. Dependence of entropy, conditional entropy (left panel) and transmitted information (right panel) on word duration for the 47^{th} frequency channel (CF: 2 kHz, 6000 stimulus repetitions). With a coarser temporal resolution (4 ms), less information is transmitted but we cover longer word durations.

we have to compensate increasing sampling intervals by decreasing word lengths (in bit) to keep the absolute word duration (in ms) constant. When we calculate transmitted information and refine sampling intervals, transmitted information increases until the temporal resolution of ON is reached. Figure 4 indicates that the transmitted information rate (constant word duration of 2 ms) saturates at a value of 700 bits/s if the temporal resolution is finer than approximately 0.1 ms; 90% of the information is covered for a resolution of 0.25 ms. This value highlights the extreme temporal precision of ON – one of the purposes of these neurons is the extraction of temporal information for sound localization.

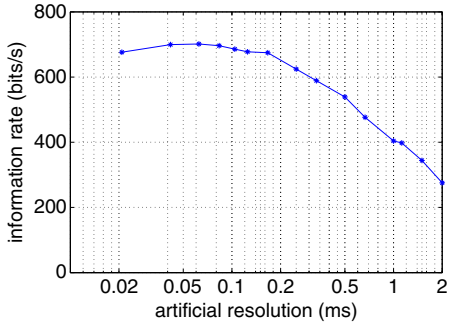

Fig. 4. Dependence of transmitted information on temporal resolution (47^{th} frequency channel, 2 kHz CF).

3.1.3. Information distribution over frequency channels

We investigated how information is distributed over frequency and how robust ON code speech in noisy conditions. In clean conditions, the initial information rate (Fig. 5, 2 ms

word duration, 0.125 ms temporal resolution, information calculated over the first 50 ms of the signal) reaches its highest values in the regions of the speech formants, where ON spike most frequently. The absolute maximum (964 bits/s) is at low frequencies, where the spikes are precisely phase-locked to the pitch frequency of the speaker. Still, also in the 2 kHz region, the information rate reaches comparable values, which indicates that the temporal precision of spikes is maintained: ON still lock reliably on each pitch period of voiced speech. The absence of information coded in the frequency range above 3.5 kHz is obvious, as in our model ON do not fire in this frequency region.

With added pink noise we evaluated the robustness of speech coded in ON spike trains. As we only want to consider the information inherent in speech and not in noise, we added randomly regenerated noise for every trial. For a SNR of 0 dB, pink noise corrupted about 84% of the information at the 286 Hz location and 90% at the 2.5 kHz location.

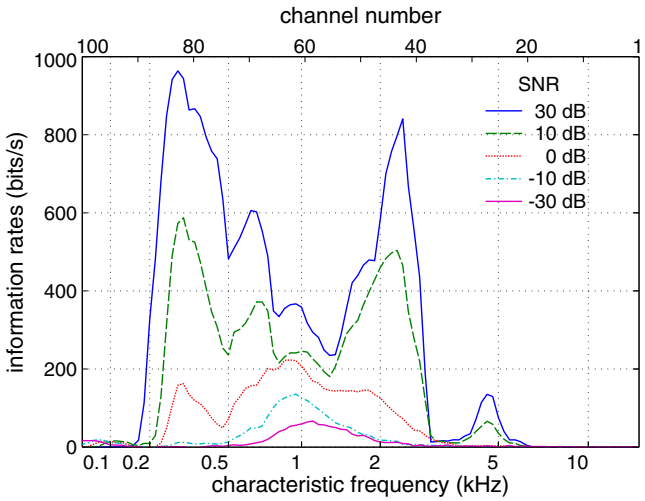

Fig. 5. Information carried across frequency channels for signals with different SNR. The signal is the vowel */ei/*, 70 dB(A), added with different levels of pink noise. Word duration is 2 ms with 0.125 ms resolution.

4. SUMMARY AND CONCLUSION

Onset neurons located in the cochlear nucleus are known for their distinct temporal processing capabilities. In this paper we analyzed their performance in the sense of information transmission. Our results show that the temporal resolution of neural coding was approximately 0.1 ms, and the maximum initial information transmission rate was close to 1000 bits/s, which corresponds to a very high information of 3.26 bits/spike. At a resolution of 1 ms, we lose half of the information carried by the spike trains compared to the finest temporal resolution. We also investigated the robustness of neuronal coding to noise. For pink noise with 0 dB SNR, the information content of voiced speech decreased by a factor of 8 compared to clean speech. In contrast to previous investigations, which were restricted to time limited neural recordings, we found that the information rate is not constant for increasing word lengths. Instead, transmitted information saturates, with 90% of the information transmitted within the first 73 ms. The time constant depends on signal level; for louder signals the information is transmitted faster (data not shown). Information theory also allows us to quantify information loss in noisy environments. Our model predicts that in pink noise with the same A-weighted level as speech, corrupts up to 90% of the information. Technical applications like automatic speech recognition rely on signal representations with a very coarse temporal resolution (usually 10 ms). If we extrapolate our data (Fig. 4), we expect that this process destroys at least 90% of the information coded by the human auditory system. We therefore suspect that fine-grained temporal information might improve ASR systems especially in noisy conditions.

Acknowledgements

This work was funded by the German Federal Ministry of Education and Research (reference number 01GQ0443).

5. REFERENCES

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

[2] T. Cover and J. Thomas, *The elements of information theory*, New York: Plenum Press., 1991.

[3] S.P. Strong, R.R. de Ruyter van Steveninck, W. Bialek, and R. Koberle, "On the application of information theory to neural spike trains," *Pac Symp Biocomput*, pp. 621–32, 1998.

[4] A. Borst and F.E. Theunissen, "Information theory and neural coding," *Nature Neuroscience*, vol. 2, pp. 947–957, 1999.

[5] M. Holmberg and W. Hemmert, "An auditory model for coding speech into nerve-action potentials," in *Proc. Joint Congress CFA/DAGA'04*, Strasbourg, France, 2004, pp. 773–4.

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