Auditory-based Automatic Speech Recognition

Werner Hemmert, Marcus Holmberg

Infineon Technologies, Corporate Research
Munich, Germany
{Werner.Hemmert,Marcus.Holmberg}@infineon.com

David Gelbart

International Computer Science Institute
Berkeley, California
gelbart@icsi.berkeley.edu

Abstract

In this paper we develop a physiologically motivated model of peripheral auditory processing and evaluate how the different processing steps influence automatic speech recognition in noise. The model features large dynamic compression (>60 dB) and a realistic sensory cell model. The compression range was well matched to the limited dynamic range of the sensory cells and the model yielded surprisingly high recognition scores. We also developed a computationally efficient simplified model of auditory processing and found that a model of adaptation could improve recognition accuracy. Adaptation is a basic principle of neuronal processing, which accentuates signal onsets. Applying this adaptation model to mel-frequency cepstral coefficient (MFCC) feature extraction enhanced recognition accuracy in noise (AURORA 2 task, averaged recognition scores) from 56.4% to 75.6% (clean training condition), a relative improvement of 41% in word error rate. Adaptation outperformed RASTA processing by more than 10%, which corresponds to a relative improvement of 31%.

1. Introduction

Auditory modeling has previously been demonstrated to improve automatic speech recognition (ASR) in noise [5, 6, 12, 15]. In this work, we created a model of peripheral auditory sound processing, interfaced it to a hidden Markov model (HMM) speech recognizer, and evaluated the impact of every processing step on ASR. We have focused on the task of ASR in noise. This choice of task is motivated by the practical importance of improving ASR in noise and by the comparatively excellent performance of human hearing in noisy conditions. Compared to previous work, we more closely model cochlear physiology and our model provides much larger dynamic compression.

Because of the computational complexity of a detailed auditory model, as well as a desire to better understand the effect of different processing principles on ASR performance, we also investigated a simplified model, and experimented with combining one part we found to be significant – adaptation – with conventional mel-frequency cepstral coefficient (MFCC) feature extraction.

2. Detailed inner ear model

2.1. Basilar membrane model

In contrast to typical speech recognition front-ends, which rely on a Fast Fourier Transform (FFT), the frequency decomposition performed in the human inner ear resembles cascaded filters. This becomes evident when the hydrodynamics of the inner ear are transformed into the equivalent electrical circuit (Fig. 1). Motions of the stapes are propagated in the inner ear in the form of a travelling wave on the basilar membrane (BM). High frequency signals reach their vibration maximum close to the basal end of the inner ear, whereas low frequency stimuli travel further apically. That way, the travelling wave causes a spectral decomposition of sound (see Fig.2). We modeled BM vibrations with a computationally efficient wave digital filter model consisting of 100 sections [13]. The response of this model is plotted in Fig. 2 (panel b, solid line).

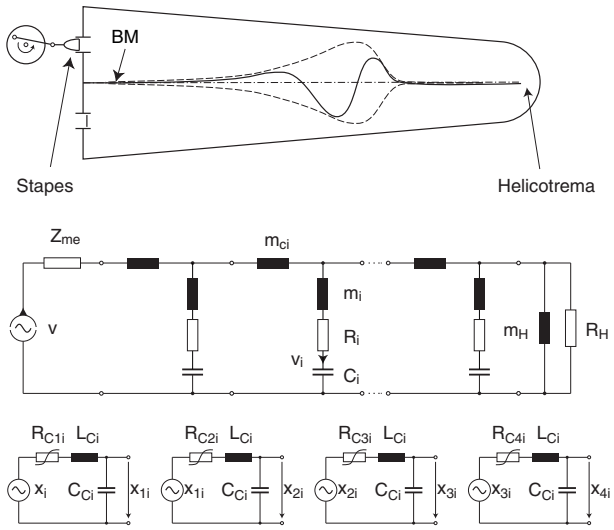

Figure 1: *Electrical equivalent circuit of inner ear hydrodynamics (v-I/p-U analogy). Each section* i *takes the form of a second-order resonator (*$\mathrm{m_i, C_i, R_i}$*), which are coupled by fluid masses* $\mathrm{m_{ci}}$*. Vibrations are injected by the middle ear vibration (velocity* v*, source impedance* $\mathrm{Z_{me}}$*) and the transmission line is closed by the impedance of the helicotrema (*$\mathrm{m_H||R_H}$*). The electrical circuit representation of the nonlinear "amplification" and compression stage is shown only for one section. The first stage is driven by the BM displacement derived from the velocity output* $\mathrm{v_i}$ *of the transmission line model. Following stages are driven by the displacement of the previous stage. The compressed output is available at the last stage (*$\mathrm{x_{4i}}$*). The frequency map of the model was adjusted according to Greenwood's [2] map for the human inner ear.*

We know that inner ear processing is highly nonlinear. Low-level sounds are mechanically amplified, probably by the outer hair cells. This active, nonlinear amplification stage both boosts the vibration amplitudes and significantly sharpens the travelling wave at low levels. The amplification saturates at medium to high levels, causing compression of the dynamic range. Measurements have shown that the maximum amplification is more than thousand-fold (> 60 dB, [10]), however, it is still unknown how the inner ear reaches this extremely high amplification without becoming unstable. To achieve stable "amplification" and compression, we im-

plemented resonators with time-variable quality-factors (Q-values). This technique has previously been used within the hydrodynamic model to achieve BM nonlinearity (e.g. [13, 12]). However, to obtain the high amplification found in nature (>60 dB) Q-values as high as 1000 would be required. A resonator with such a high Q-value has an extremely narrow bandwidth and causes excessively long ringing. Here, the amplification stage was realized by adding multiple time-variable second-order resonators at the outputs of the cochlear filter bank (compare Fig. 1, a similar implementation was proposed by [9]). Quality factors were adjusted at every sample depending on the instantaneous displacement of each resonator, using a Boltzmann function similar to the sensitivity function of the OHCs. By cascading four resonator stages and modulating their quality factors, we realized large amplification and compression (modulating Q-factors between ten and one achieves maximum compression of 80 dB) together with approximately correct filter shapes. In the high frequency range of the inner ear, larger amplification is required to model physiological data than in the low-frequency range. Note that the resonators amplify only the vibration amplitude, not the energy of the vibration, as it is the case in the living cochlea. For the recognition task chosen in this paper (signals had telephone bandwidth), we modelled the human cochlea only up to a characteristic frequency of 4 kHz; the high–frequency region (distance from stapes: 0 – 12 mm) was truncated. To keep processing times at a minimum, we set the sampling rate to 8 kHz for the whole model (no over-sampling).

The excitation pattern of the BM along the whole length of the cochlea for a tone complex (pure tones with frequencies 500 Hz, 1 kHz and 2 kHz; levels as indicated in the legend) is shown in Fig. 2a. The cochlea map roughly corresponds to a logarithmic frequency scale, with the basal part reacting predominantly to high-frequency signals. Accordingly, the 2 kHz tone causes activity at a distance of approximately 15 mm (measured from the stapes), the so-called characteristic location (CL) of that frequency. At low levels (Fig. 2a, 20 dB), filter shapes are narrow and almost symmetrical. The amplitudes at the most sensitive location are greatly amplified, so that even faint sounds cause excitation above threshold (which is around 1 nm, [10]). At high signal levels, responses share the characteristics of a passive travelling-wave with a gradual build-up and a sharp roll-off after the relatively broad maximum is reached (80 dB line). For increasing intensities, the excitation pattern grows highly asymmetrical, in accordance with the well known phenomenon of upper spread of masking. At CL, the growth function (the ratio of BM displacement to signal amplitude) is greatly compressed. Compression varies with location along the cochlea; it is largest in the most basal part. Here, the growth function follows approximately a cube root law at medium signal levels. At more apical locations, growth functions are almost linear (compare responses to 2 kHz tone, CL 15 mm, and 500 Hz tone, CL 25 mm). Despite the huge (nonlinear) compression, distortions are surprisingly low (visible only around 17 mm in the 60 dB trace of 2a).

2.2. Sensory cell model

Basilar membrane vibrations drive the hair bundles (HB) of the sensory cells (the inner hair cells, IHCs), by fluid motion. As a first approximation, fluid friction and HB stiffness form a first-order high-pass filter, i.e. HB displacement is proportional to BM velocity at low frequencies and to BM displacement at high frequencies. The corner frequency is taken to decrease (2000 – 200 Hz) along the length of the cochlea as a result of increasing length and decreasing stiffness of the bundles. This high-pass filter provides some additional sharpening of the shallow low-frequency slope of the tuning curves. Figure 3a shows the basilar membrane vibration for a 70 dB

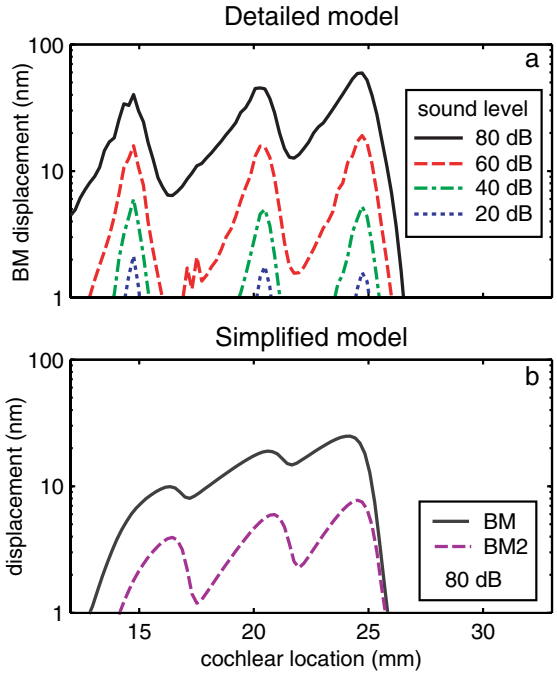

Figure 2: *Excitation pattern (RMS) of the basilar membrane for a tone complex (pure tones with frequencies of 0.5, 1 and 2 kHz) at various sound levels as indicated in the legend (modeled human cochlea). Panel a (top): Detailed model. Levels in the legend refer to the level of each single tone. Panel b (bottom): Displacement of the simplified basilar membrane model (BM, solid line) and with sharpened responses by three additional high-pass filters (BM2, dashed line, compare section 6). Note that our model covered only frequencies up to 4 kHz; it therefore started at a distance of 12 mm measured from the stapes.*

200 Hz tone burst at its characteristic location (29 mm); the vibration amplitude is about 50 nm. Notice that the BM vibration is delayed relative to the tone onset due to the traveling time of the wave in the inner ear. Hair bundle displacement is plotted in Fig. 3b, its amplitude is about 3 dB lower than that of the BM. Upon deflection of the HB, ion channels at its tip open. The open probability of the mechano-electrical transduction channels follow a second-order Boltzmann function [8]. This function resembles a half-wave rectification; channels open only for deflections in excitatory direction (compare Fig. 3c). For displacements in the opposite direction, the activity (open probability) of the transduction channels is depressed slightly from their activity at rest. The saturating nonlinearity of the Boltzmann function deforms the receptor potential significantly and gives the sensory cells a limited dynamic range of less than 40 dB. Therefore mechanical compression before the transduction is essential to process sounds with large dynamic ranges. The voltage difference between the fluid space into which the HBs extrude (scala media, +90 mV), and the membrane potential of the IHCs (V_R, -45 mV at rest), drives a transduction current which depolarizes the IHC membrane (capacity: 10 pF, conductivity: 60 nS). These electrical properties of the membrane are modeled as a first order low-pass filter with a corner frequency of approximately 1 kHz.

2.3. Synaptic mechanisms

Depolarization of the membrane activates voltage dependent calcium channels located close to the cell's synaptic terminals and

Ca^{2+}-ions enter the cell. Ca^{2+}-channel activation also follows a Boltzmann function. Elevated Ca^{2+}-levels cause fusion of synaptic vesicles with the cell membrane. The neurotransmitter contained in the vesicles diffuses across the synaptic cleft (not modeled), binds to the receptors in the postsynaptic membrane, causes the postsynaptic membrane to depolarize, and a nerve action potential is propagated along the auditory nerve fiber towards the brain. Neurotransmitter release is dominated by a so-called readily releasable pool (RRP) of vesicles, which are located in the synaptic region closely to the cell membrane. A large stimulus causes fusion of many vesicles, depleting the RRP. Due to the depletion, the transmitter concentration in the synaptic cleft (TR) and therefore also spiking probability of the auditory nerve is reduced in the following few tens of milliseconds, an effect known as adaptation. We modeled RRP refill according to recent measurements [7] with a single time constant of 67 ms. It has to be stated that this one-pool model only partly approximates physiological data [16]; more precise models still have to be developed. The transmitter rapidly degrades in the synaptic cleft (time constant $\approx$ 0.1 ms) causing TR to decrease (see Fig. 3d). The generation of a nerve action potential depends on TR and on the refractoriness of the auditory nerve fiber (not modeled in this paper). TR exhibits a significant difference compared to the preceding steps: responses are emphasized at signal onset and decay thereafter. However, the decay is not complete and steady state signals do elicit tonic activation of the auditory nerve. If we consider only the signal envelopes in Fig. 3, TR resembles the sum of the receptor potential V_R with a high-pass filtered version of V_R.

3. Interfacing to the speech recognizer

The outputs of each section of the model – basilar membrane displacement, hair-bundle displacement, receptor potential or transmitter concentration in the synaptic cleft – provide frequency-selective features at the full sampling rate of the model. We temporally integrated the root-mean-square energy of each frequency-selective section using a Hanning window (25 ms width) which we advanced in steps of 10 ms. We then spectrally integrated using a second Hanning window (width covering 17.5 cochlear sections, advanced in steps of 7 sections) to reduce the frequency resolution from 100 channels to 12 feature vectors. This procedure resulted in a signal representation which was more appropriate for the speech recognition back end.

Since the diagonal-covariance Gaussian mixture models we used for ASR acoustic modeling have a limited ability to model correlations between different features in the feature vector, we applied a Karhunen-Loeve Transform (KLT) to decorrelate the features. (We determined the KLT transform using only the clean training data.) The KLT is approximated by the more commonly used discrete-cosine transform (DCT) if certain assumptions are satisfied; since we are working with novel features we felt safer using the KLT. We completed our feature vector by adding delta and delta-delta features; this can be seen as a way to include temporal context without introducing strongly correlated features as would happen if we appended the feature vectors of adjacent frames.

Features are plotted (Fig. 4) for the spoken digits "one three". The spectral and especially the temporal resolution is severely blurred because of the data reduction. Whereas BM and HB displacement show only minor differences from each other, the receptor potential V_R is significally altered due to the saturating nonlinearity of the mechano-electrical transduction (compare also Fig 3c and section 2.2). The transmitter concentration in the synaptic cleft TR strongly emphasizes onsets of sounds.

4. Recognition task and recognizer back end

We used the Aurora 2 speech recognition task (connected digits in noise, bandpass-filtered to telephone bandwidth) defined in [4]. This task presents the recognizer with a variety of signal-to-noise ratios (SNR) and noise types. We obtained results on this task using a hidden Markov model (HMM) back end used by many other researchers for this task. The back end, based on Cambridge's HTK recognizer, uses word-level digit models with 16 states per word. Mixtures of diagonal-covariance Gaussians are used for modeling each state. We used the "complex" version of the back end (http://icslp2002.colorado.edu/special_sessions/aurora/) in which the number of Gaussians per state is increased to twenty.

Aurora 2 has two training sets: a clean training set with no added noise and a multi-condition training set with various noises added at various SNRs. There are three test sets: in test set A the noises are the same as in the multicondition training, in test set B the noises are different, and in test C the noises are different and also a different bandpass filter is used. We report our performance separately for each training condition, as percentage word recognition accuracy averaged over the three test sets and five SNR conditions between 20 and 0 dB. (We will place a more detailed performance breakdown online at http://www.icsi.berkeley.edu/Speech/papers/sapa04-hemmert.html. We may use the same location for updates or other additional information about our work.)

5. Recognition results with detailed inner ear model

Averaged recognition scores using the features derived from the inner ear are shown in Table 1. Please note that these scores relate to performance under noisy conditions; recognition scores for "clean" conditions were close to 99%. Recognition scores of all features, except TR, are similar to the results obtained using a MFCC feature vector (cepstral coefficients C0-C12 along with deltas and delta-deltas, see Table 3, first column). Given all the nonlinearities of the detailed inner ear model, the high recognition scores are surprising. Apparently, compression and nonlinearities were well matched so that no important information was destroyed. The drop of recognition for TR indicates that significant information was lost at this stage. It may be that the activation of the Ca^{2+}-channels and probably also the one-pool model for adaptation are not sufficient and that more sophisticated modeling is required.

Table 1: *Recognition accuracy for the detailed model.*

	BM	HB	V_R	TR
clean	55.20	55.47	55.68	46.26
multi	85.31	85.57	87.56	80.56

6. Simplified inner ear model

The detailed, physiological model described above is much more complex than conventional ASR front ends. For this reason, and to gain insight into the relationship between model properties and ASR performance, we also performed experiments with a simplified model.

6.1. Simplified basilar membrane model

The simplified model used the same passive transmission line as the detailed model. As this provides no amplification, we simply

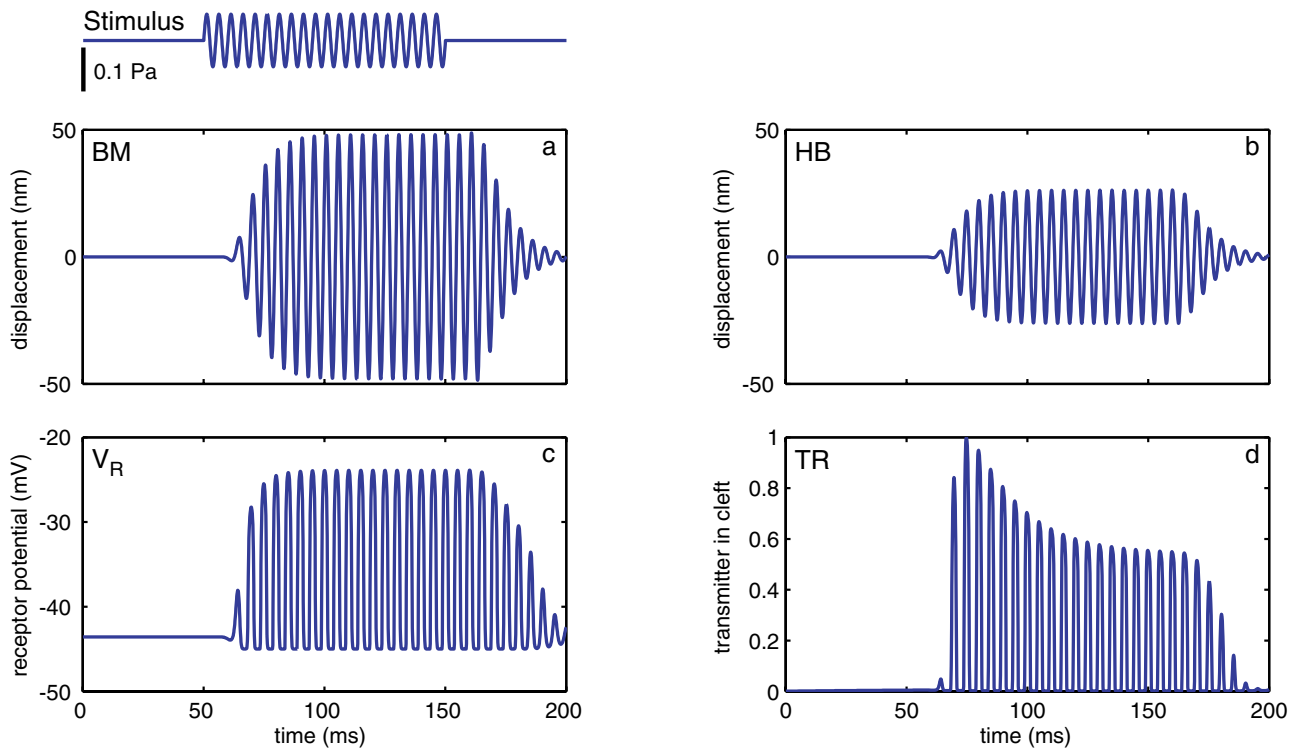

Figure 3: *Excitation caused by a 200 Hz tone burst (sound pressure: 70 dB peak amplitude, stimulus on–time: 50 –150 ms, detailed model) at its characteristic location (29 mm). a) BM: displacement of the basilar membrane, b) HB: hair bundle displacement of the sensory cells, c) V_R: receptor potential, d) TR: normalized transmitter concentration in synaptic cleft.*

multiplied the input signals by a factor of 1000 (60 dB). To derive speech recognition features, we calculated the logarithm of the features after the dimensionality reduction described in section 3. Compared to the detailed inner ear model, the logarithm has the advantage that there can be no information loss due to saturation. Speech recognition scores show, however, that the performance of this model is much worse than the detailed model (BM in Table 2). As the low-frequency slopes of the linear basilar membrane filter are much flatter than in the detailed model and physiological data, we added high-pass filtering at the output of each section, with corner frequency matched to the corresponding cochlear location's best frequency. Using three cascaded first-order high-pass filters, basal slopes of the excitation patterns were sharpened and close to these observed at medium levels (60-80 $\mathrm{dB_{SPL}}$, see Fig. 2) of the detailed model. Recognition accuracy improved greatly (BM2 in Table 2).

6.2. Simplified sensory cell model

In the simplified model, mechano-electrical transduction of the sensory cells (IHCs) was realized with a simple half-wave rectification. At this stage we also implemented compression with a logarithmic function (instead of applying the logarithm after the dimensionality reduction as in section 6.1) because we wanted to realize adaptation in the logarithmic domain. To avoid negative or zero arguments to the logarithm, we added a threshold value (1 nm) to the rectified bundle displacement. We low-pass filtered this signal with a corner frequency of 1 kHz, which is equivalent to the electrical properties of the cell's membrane.

This processing (V_R in Table 2) degraded recognition scores (presumably because of the half-wave rectification), especially if training was only performed with clean speech.

6.3. Simplified adaptation

In the simplified model, the effect of adaptation (see section 2.3) was modeled by summing V_R with two highpass filtered versions of V_R. The highpass filters were realized as first-order infinite impulse response filters, with corner frequencies of 2.65 Hz and 53 Hz. The output of the 53 Hz highpass was scaled by a factor of 6. These parameters were chosen to match physiological measurements of [16]. In contrast to the drop in performance going from V_R to TR with the detailed model, this processing scheme greatly improved clean training set recognition accuracy for the simplified model (TR in Table 2).

Table 2: *Recognition accuracy for the simplified model (f is the corner frequency of highpass filter in A2).*

	BM	BM2	V_R	TR	BM2+A2 f=2.65Hz	BM2+A2 f=1Hz
clean	46.4	57.4	46.1	56.1	65.6	75.0
multi	84.6	89.7	87.4	87.1	90.8	91.3

For comparison, we also applied the frame-based adaptation processing (A2) described in section 7 to the features of our previous simplified model (BM2), after the integration into 25 ms frames described in section 3. The results are shown in the final two columns of Table 2.

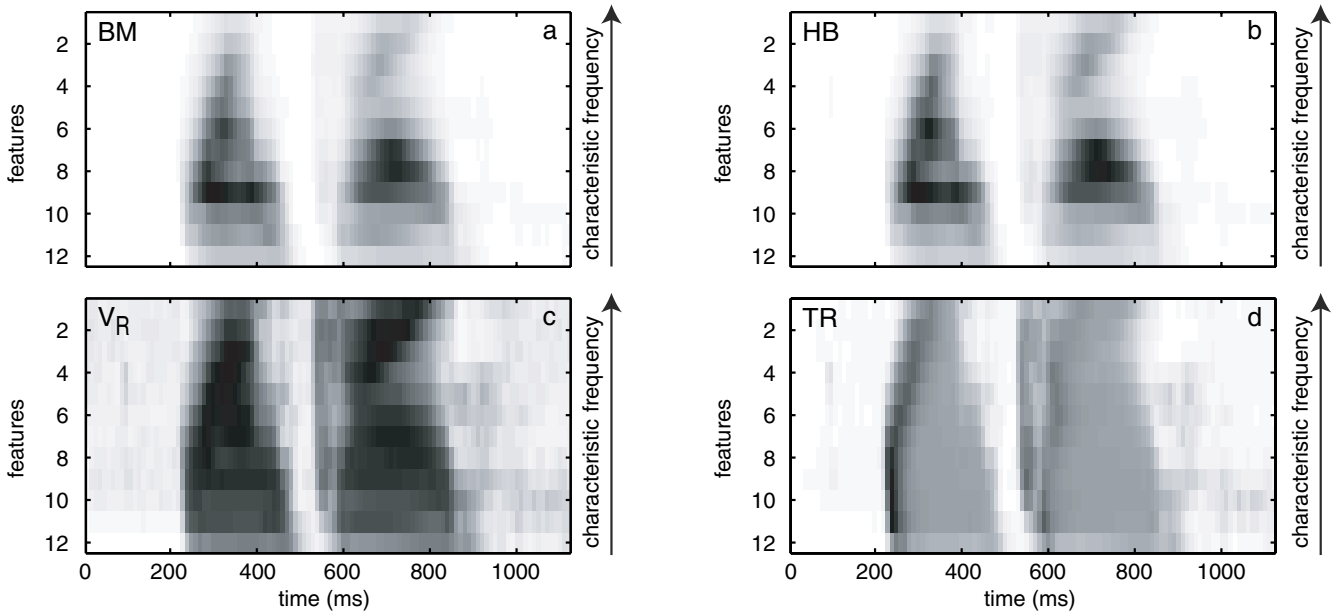

Figure 4: *Normalized features after spectral and temporal integration for the spoken digits "one three" (detailed model). a) BM: basilar membrane displacement, b) HB: hair bundle displacement, c) V_R: receptor potential, d) TR: transmitter concentration in synaptic cleft. Black areas indicate high amplitudes.*

7. Frame-level adaptation processing with FFT-based features

Inspired by the performance improvement due to the simplified adaptation processing, we tested the effect of this processing for conventional, FFT-based features. In this case the adaptation was operating at the frame rate (100 Hz) rather than the audio sampling rate, so we omitted the second (53 Hz) high-pass filter. Thus the "adapted" features were the sum of the original features and a temporally high-pass filtered version. We will refer to this frame-level adaptation processing as A2. We applied the adaptation processing to logarithmic mel-spectra, just prior to the discrete cosine transform in MFCC calculation. After some experimentation with the high-pass filter corner frequency value we found performance was improved using a corner frequency of 1 Hz instead of 2.65 Hz. ASR results are shown in the MFCC+A2 columns of Table 3.

The A2 processing is related to the commonly used RASTA [3] and cepstral mean subtraction (CMS). The MFCC+RASTA column shows the effect when RASTA filtering was applied to the logarithmic mel-spectra instead of A2. MFCC+A2 outperforms MFCC+RASTA, especially with clean training. For the clean training condition MFCC+A2 outperforms MFCC+RASTA by more than 10% in the absolute recognition scores, which corresponds to a relative improvement of 31%. For the multi-condition training condition, differences are smaller, but recognition results of MFCC+A2 are still about 12% better than for MFCC+RASTA. The MFCC+CMS column shows the effect of cepstral mean subtraction, with means calculated over entire utterances. MFCC+A2 outperforms MFCC+CMS with clean training. With multicondition training, MFCC+CMS is superior; however, the extra delay resulting from mean calculation over an entire utterance is unattractive in some ASR applications.

Table 3: *Recognition accuracy for frame based adaptation processing (f is the corner frequency of highpass filter in A2).*

	MFCC	MFCC +A2 f=2.65Hz	MFCC +A2 f=1Hz	MFCC +RASTA	MFCC +CMS
clean	56.37	68.35	75.55	64.46	69.53
multi	89.34	90.44	91.04	89.86	92.55

8. Discussion

We have developed a physiologically motivated model of basilar membrane dynamics and a realistic model of the sensory cells in the human inner ear which is able to replicate a great amount of physiological and psycho-acoustical data. Due to the limited dynamic range of the sensory cells (<40 dB), dynamic compression of BM vibrations is essential for sound processing without major information loss. Our BM model achieved a dynamic range compression of more than 60 dB and accomplished speech recognition scores similar to classical MFCC features. The inner ear model features a natural mel-scaled frequency transformation, whereas with conventional FFT-based feature extraction (MFCC or PLP) a mel- or Bark-scale is achieved by weighted summing across FFT bins. This summing reduces the frequency resolution without improving the temporal resolution of the features. In the human auditory system, the weaker (absolute) frequency selectivity at higher frequencies is compensated by a higher temporal resolution. However, we reduced the temporal resolution of the output of our models when we integrated in time- and frequency domains to achieve 12 features each 10 ms. If we could better exploit the high temporal resolution of the auditory model, further improvements to speech recognition performance might be possible. Previous work on auditory modeling has involved the introduction of novel feature extraction techniques, for example Ghitza's ensemble interval histogram (EIH) [11] and

Sheikhzadeh and Deng's inter-peak interval histogram (IPIH). In some experiments, not included in this paper, we tried IPIH, but we did not observe a performance improvement.

Auditory models may improve speech recognition, but on the other hand, speech recognition can also be applied to test auditory models. In our model, recognition scores are not reduced by the harsh threshold- and saturating nonlinearities of the sensory cells, on the contrary, recognition scores even improve slightly. When the transmitter release is tested, a large drop of recognition scores is apparent. This step was modeled with a one-pool vesicle model, which does not correctly reproduce adaptation. We conclude that the physiological model of transmitter release we used requires improvements to satisfy both physiological measurements and high fidelity information processing requirements.

In an attempt to reduce the high computational requirements of physiologically motivated models, we also developed a simplified model, where we replaced dynamic compression by a simple logarithmic transformation. Table 2 suggests that the shallow high-frequency slope of a passive basilar membrane model does not provide features with sufficient frequency selectivity. This problem was solved by adding additional high pass filters. The ingenious design of the human auditory system becomes apparent again when we look at the conversion of vibrations into a receptor potential V_R. In the simplified model, this processing involves half-wave rectification and low-pass filtering, and it appears from the recognition scores that this is much better handled by the detailed model with its nonlinear dynamic compression and its soft rectification using a second-order Boltzmann function. On the other hand, the simplified model of adaptation (see section 6.3) improved recognition scores for the clean training condition. We also tested this scheme on features which were already integrated in frequency and time for dimensionality reduction (see Table 3). Recognition scores improved greatly, perhaps because half-wave rectification is not required in this case, as the RMS-values of BM vibration are processed.

One major goal of our work with auditory models is to identify processing schemes which improve speech recognition. The principles underlying these may also be useful in conventional ASR implementations. When we added adaptation to MFCC processing (computational cost is less than 0.02 MOPS), we found large improvements of speech recognition accuracy especially with clean training. Here adaptation outperforms CMS (Table 3). Unfortunately, these improvements do not always translate to the multi-condition training, where CMS is superior. Adaptation is a fundamental principle of sensory- and neuronal processing, which suppresses equally distributed information and enhances changes. In auditory information processing, adaptation in the auditory nerve accentuates temporal onsets of signals. Maximum recognition accuracy was obtained for a high pass filter corner frequency of 1 Hz which matches the value used in RASTA. The corresponding time constant is slightly lower than the adaptation time constant observed in the auditory nerve. However, adaptation is apparent at all neuronal stages and time constants are thought to increase along the auditory pathway. This might explain why a lower filter corner frequency appears to be more appropriate in speech recognition experiments.

It is noteworthy that a model of neuronal adaptation outperforms RASTA processing. Like adaptation, RASTA enhances temporal changes of signal amplitudes but it is using a high-pass filter which completely suppresses stationary signal components. This processing scheme enhances speech recognition in background noise from 56% (MFCC) to 64% (MFCC+RASTA) in the clean training condition. However, sometimes stationary-seeming signals, for example vowels with a long duration, do carry linguistic information which might be destroyed by RASTA. The optimal temporal filtering for improving ASR performance can depend on the recognition task [1, 14]. At least for the Aurora task, adaptation seems to provide a good compromise: in the clean training condition adaptation outperformed RASTA (31% relative improvement in word error rate) and reached a recognition score of 75.55%.

9. Acknowledgements

Thanks to Nelson Morgan, Hynek Hermansky, Barry Chen, Arlo Faria, Pratibha Jain, Adam Janin, Carmen Pelaez and Qifeng Zhu for their advice and assistance.

10. References

[1] Chen, C., "Noise robustness in automatic speech recognition", *PhD thesis*, University of Washington, 2004, http://ssli.ee.washington.edu/papers/publications.html

[2] Greenwood, D.D., "A cochlear frequency-position function for several species – 29 years later", *JASA*, vol. 87, pp. 2592–2605, 1990.

[3] Hermansky, H. and Morgan, N., "RASTA processing of speech", *IEEE TSAP*, vol. 2, pp. 578–589, 1994.

[4] Hirsch H. and Pearce, D., "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions", *ISCA ITRW ASR*, pp. 181-188, 2000.

[5] Hunke, M., Hyun, M., Love, S. and Holton, T., "Improving the noise and spectral robustness of an isolated-word recognizer using an auditory-model front end", *ICSLP 1998*, Sydney, Australia.

[6] Jankowski, C., Vo, H. and Lippmann, R., "A comparison of signal processing front ends for automatic word recognition", *IEEE TSAP*, vol. 3, no. 4, 1995.

[7] Moser, T. and Beutner, D., "Kinetics of exocytosis and endocytosis at the cochlear inner hair cell afferent synapse of the mouse", *Proc. Natl. Acad. Sci. USA*, vol. 97, pp. 883–888, 2000.

[8] Mountain, D. C. and Cody, A. R., "Multiple modes of inner hair cell stimulation", *Hearing Research*, vol. 132, pp. 1–14, 1999.

[9] Robert, A. and Eriksson, J. L., "A composite model of the auditory periphery for simulating responses to complex sounds", *JASA*, vol. 106, pp. 1852–1864, 1999.

[10] Robles, L. and Ruggero, M. A., "Mechanics of the mammalian cochlea", *Physiological Reviews*, vol. 81, pp. 1305–1352, 2001.

[11] Sandhu, S., Ghitza, O. and Lee, C.-H., "A comparative study of mel cepstra and EIH for phone classification under adverse conditions", *ICASSP 1995*, Detroit, USA.

[12] Sheikhzadeh, H. and Deng, L., "Speech analysis and recognition using interval statistics generated from a composite auditory model", *IEEE TSAP*, vol. 6, no. 1, pp. 90–94, 1998.

[13] Strube, H. W., "A computionally efficient basilar-membrane model", *Acustica*, vol. 58, pp. 207–214, 1985.

[14] van Vuuren, S. and Hermansky, H., "Data-driven design of RASTA-like filters", *Eurospeech 1997*, Rhodes, Greece.

[15] Wang, K. and Shamma, S., "Self-normalization and noise robustness in early auditory representations", *IEEE TSAP*, vol. 3, pp. 382-395, 1994.

[16] Westerman, L.A. and Smith, R.L. "Rapid and short-term adaptation in auditory nerve responses", *Hearing Research*,vol. 15, pp. 249-260, 1984.