Automatic Speech Recognition with an Adaptation Model Motivated by Auditory Processing

Marcus Holmberg, David Gelbart, *Student Member, IEEE,* and Werner Hemmert, *Member, IEEE*

***Abstract*— The Mel-Frequency Cepstral Coefficient (MFCC) or Perceptual Linear Prediction (PLP) feature extraction typically used for automatic speech recognition (ASR) employ several principles which have known counterparts in the cochlea and auditory nerve: frequency decomposition, mel- or bark-warping of the frequency axis, and compression of amplitudes. It seems natural to ask if one can profitably employ a counterpart of the next physiological processing step, synaptic adaptation. We therefore incorporated a simplified model of short-term adaptation into MFCC feature extraction. We evaluated the resulting ASR performance on the AURORA 2 and AURORA 3 tasks, in comparison to ordinary MFCCs, MFCCs processed by RASTA, and MFCCs processed by cepstral mean subtraction (CMS), and both in comparison to and in combination with Wiener filtering. The results suggest that our approach offers a simple, causal robustness strategy which is competitive with RASTA, CMS and Wiener filtering and performs well in combination with Wiener filtering. Compared to the structurally related RASTA, our adaptation model provides superior performance on AURORA 2 and, if Wiener filtering is used prior to both approaches, on AURORA 3 as well.**

***Index Terms*— Neural adaptation, speech recognition, noise robustness.**

EDICS Category: 1-RECO

Werner Hemmert*, Infineon Technologies AG Corporate Research CPR ST, Building 10-562, Otto-Hahn-Ring 6, 81730 Munich, Germany, Tel.: +49 (89) 234-53055, Fax: +49 (89) 234-9557068, werner.hemmert@infineon.com

Marcus Holmberg, Infineon Technologies AG, Corporate Research CPR ST, Building 10-562, Otto-Hahn-Ring 6, 81730 Munich, Germany, Tel.: +49 (89) 234-48682, Fax: +49 (89) 234-9554115, marcus.holmberg@infineon.com

David Gelbart, ICSI, Berkeley, USA International Computer Science Institute, 1947 Center Street, Suite 600 Berkeley, CA 94704-1198, Tel.: +1 (604) 737-9898, Fax: +1 (604) 221-7250, gelbart@icsi.berkeley.edu

I. Introduction

THE accuracy of human speech recognition motivates the application of information processing strategies found in the human auditory system to automatic speech recognition (ASR) [1], [2]. The most popular feature extraction methods for ASR, Mel-Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP), already employ several principles which have known counterparts in the cochlea and auditory nerve: frequency decomposition, mel- or bark-warping of the frequency axis, and compression of amplitudes. It therefore seems natural to consider the next processing step in the auditory periphery – synaptic adaptation in the auditory nerve. Adaptation (also known as synaptic depression) is a principal mechanism of neuronal information processing and is ubiquitous in the brain [3], [4], [5]. It accentuates signal onsets by following a high initial firing rate with a lower sustained rate. Adaptation is strong in the auditory nerve, as has been described in a number of measurements [3], [6], [7], [8].

Models of adaptation, or techniques resembling adaptation, have successfully been used in ASR. Adaptation has apparent similarities with the popular RASTA [9] technique. RASTA processing of speech is a bandpass modulation filtering, operating on the logarithmic spectrum. But whereas RASTA processing completely suppresses DC modulation, the auditory nerve shows a sustained firing rate to continuous stimuli. Recovery from adaptation might be partially responsible for temporal (forward) masking observed in psychoacoustic experiments [10]. Strope and Alwan [11] developed a model replicating psychoacoustic masking experiments with which they demonstrated ASR performance improvements, especially in noisy conditions. Seneff [12] included adaptation in her model of the auditory periphery, which was found to perform better in additive noise than a mel filter bank in [13]. Perdigão and Sá [14] found the Seneff model to be susceptible to noise (in contrast to the finding in [13]), but found that a simplified model of synaptic adaptation generally improved recognition scores. Tchorz and Kollmeier [15] used an auditory model to evaluate various adaptation parameters on an ASR task. They reported higher recognition scores in additive noise for their model compared to mel-frequency cepstral coefficients (MFCC) and attributed that mainly to their joint adaptation/compression model.

Accumulated knowledge of the synaptic processes of inner hair cells (e.g. [7]) has led to the evolution of fairly precise models [16], [17]. In this work, we first review the physiological facts and illustrate the effects of synaptic adaptation using a detailed model of auditory processing in the inner ear (Section II). We next derive a simplified model of adaptation and integrate it into conventional mel-frequency cepstral coefficient (MFCC) feature extraction (Section III). We evaluate the resulting ASR performance using the AURORA 2 and AURORA 3 speech recognition tasks (Section IV and Section V), in comparison to ordinary MFCCs, MFCCs processed by RASTA, and MFCCs processed by cepstral mean subtraction (CMS), and both in comparison to and in combination with Wiener filtering.

M. Holmberg and W. Hemmert are with Infineon Technologies, Munich, Germany.

D. Gelbart is with ICSI, Berkeley, USA.

II. ADAPTATION PHYSIOLOGY AND DETAILED MODELING

A. Synaptic adaptation

The receptor cells in the inner ear, known as inner hair cells (see Fig. 1), transduce displacements of their hair bundles into electrical potentials. Voltage-sensitive Ca^{2+} channels located close to the synapses at the basal part of these cells open upon depolarisation of the cell membrane. Ca^{2+} mediates the fusion of neurotransmitter-filled synaptic vesicles with the cell membrane. The neurotransmitter diffuses across the synaptic cleft and binds to receptors on the post-synaptic membrane, which depolarizes and triggers an action potential, which propagates along the auditory nerve to the brain.

In the vicinity of each synapse is the "readily releaseable pool" (RRP) of synaptic vesicles. If enough time has elapsed since any stimulus at the corresponding basilar membrane location, this pool will be filled. At the beginning of an acoustic stimulus plenty of vesicles are available to fuse, causing a strong initial auditory nerve response. As the RRP is refilled at a lower rate than the initial vesicle fusion rate, it depletes. Auditory nerve activity is thus depressed shortly after stimulus onset during sustained stimuli. Adaptation in the auditory nerve can be described by a proportional part and two decaying exponentials [7] which are referred to as "rapid adaptation" and "short-term adaptation". Time constants[1] measured in the Mongolian gerbil were a few milliseconds for rapid adaptation and roughly 40-60 ms for short-term adaptation [7]. In cats, however, Chimento and Schreiner [18] found time constants of short term adaptation were level-dependent and decreased from 116 ms at 10 dB (above hearing threshold) to 73.5 ms at 30 dB. Spoor and Eggermont [19] estimated that human time constants of recovery from adaptation are about a factor of four longer compared to gerbil data; we therefore assumed a short-term adaptation time constant of 240 ms. This time constant is also motivated by the notion that higher levels in the auditory pathway – with presumably longer adaptation time constants than the auditory nerve – contribute to the temporal processing of speech. The 240 ms time constant is consistent with other auditory models that have been used as ASR front ends [15] and with the high-pass corner frequency of RASTA [9]. The corresponding high-pass corner frequency (0.66 Hz) is below but close to the maximum of the modulation spectrum of speech, which is between 2 to 8 Hz [1].

B. The effect of adaption in a detailed physiological model

In this section we use a detailed, physiologically motivated model of auditory processing in the inner ear (described in more detail in [22]) to show the effects of synaptic adaptation in Fig. 2. This model consists of a nonlinear model of the human cochlea [22] combined with an inner hair cell/synapse model adopted from [16]. Fig. 2 illustrates the dynamics of synaptic processing for a pure tone with added pink noise (upper panels) and for the spoken letter "p" (lower panels; this is a male speaker with pitch frequency approximately 125 Hz). Responses were derived from a channel with a characteristic frequency slightly higher than 3 kHz.

[1]All adaptation time constants τ_A mentioned here are defined as exponential decay constants. The corner frequency f_c of a corresponding high-pass filter can be derived from equation 2.

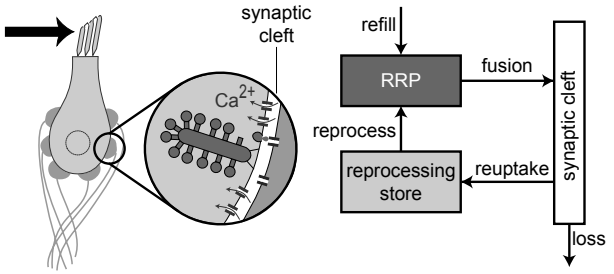


Fig. 1. Inner hair cell with synapse (left) and a model of vesicle pool dynamics adopted from [16]. The model consists of three pools (rectangular boxes). Vesicle traffic is indicated with arrows. Only vesicles in the RRP can fuse with the cell membrane and release their neurotransmitter into the synaptic cleft. Fusion rate depends on Ca^{2+} concentration. Transmitter in the cleft is partly recycled and partly lost. Transmitter which is recycled (reuptake) has to be reprocessed (reprocessing store) and then is added (reprocess) to the RRP. Additionally, there is refill by newly manufactured vesicles which replaces lost transmitter.

The inner hair cell receptor potentials (first column) code fine-grained temporal properties of the signal as well as the signal envelope; the voicing of the speech signal is preserved. To determine the poststimulus time histogram of a high spontaneous rate (HSR) nerve fiber (second column), action potentials from 1000 stimulus repetitions were counted in 1 ms wide bins. Spontaneous transmitter release into the synaptic cleft even with no stimulus present is responsible for the spontaneous rate, which was 30 spikes/s for HSR fibers. The response to the pure tone exhibits a large peak and then decays to the sustained rate of about 200 spikes/s. The two time constants of adaptation are clearly visible in response to the pure tone, where the rapid adaptation dominates the first 10 ms and the short-term adaptation and sustained rate dominate the subsequent response. The size of the short-term adaptation component compared to the steady-state response is approximately 1:1 [7]. The rapid component is larger and level dependent. In the speech sound, the rapid adaptation component appears to primarily enhance voicing. After the end of the signal, the ANF response is depressed and slowly recovers.

III. A SIMPLIFIED ADAPTATION MODEL FOR ASR

For application in speech recognition we developed a simplified adaptation model which we inserted into MFCC feature extraction just after the calculation of logarithmic mel spectra. Note that adaptation is a ubiquitous property of synapses; the model used here is to motivate a simplified adaptation stage that in fact may describe aspects of adaptation not only in the auditory nerve, but also of processing stages at higher levels along the auditory pathway.

Rapid adaptation enhances the temporal fine structure of speech signals, but as this fine structure is removed by MFCC feature extraction, we did not include rapid adaptation in our simplified model. Our model of adaptation is implemented by summing (with equal weights) a temporally high-pass filtered version of the logarithmic mel spectra with the original

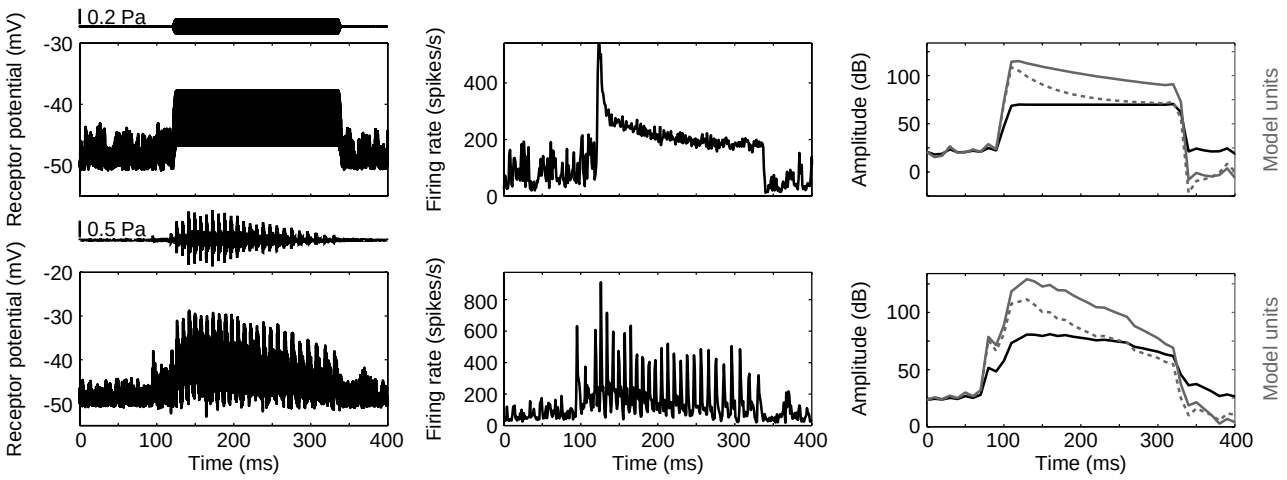


Fig. 2. Upper panel: Modeled excitation caused by a 3 kHz tone burst (sound pressure: 70 dB(A), raise time: 5 ms) with pink noise added (SNR: 39 dB(A)). Bottom panel: Response to the spoken letter "p" (ISOLET recording mtkm0-P2-t, sound pressure: 75 dB(A) (SNR: 41 dB(A)), fundamental frequency: 125 Hz, third formant at approximately 3 kHz). The sound stimulus is plotted on top of the leftmost panel. Columns from left to right: receptor potential (RP) in the detailed model, firing rate of HSR fibers (1000 repetitions, 1 ms time bins) in the detailed model, value of logarithmic mel-frequency channel without (black line, scaled in dB) and with simplified adaptation (gray lines, scaled in model units). Adaptation responses are shown for time constants of 60 ms (dotted line) and 240 ms (solid line). The mel-frequency channel was tuned to 3 kHz, the auditory model channel to a slightly higher frequency to avoid auditory nerve saturation.

logarithmic mel spectra. We performed the temporal filtering using a first-order IIR high pass filter which represents the decaying exponential effect of short-term adaptation. The transfer function of the high-pass filter is:

$$H(z) = \frac{2f_s\tau - 2f_s\tau z^{-1}}{1 + 2f_s\tau + (1 - 2f_s\tau)\, z^{-1}} \qquad (1)$$

where $f_s = 100$ Hz, the MFCC frame rate. We initialized the filter's memory to zero and, to avoid an initial transient, we subtracted the first frame of logarithmic mel spectra from all of the frames when applying this filter. Results for the 3 kHz mel-frequency channel with and without added adaptation are plotted in the rightmost column of Fig. 2. The figure illustrates that the the original logarithmic mel-frequency value closely resembles the envelope of the receptor potential, and the adaptation model replicates the effects of short-term adaptation. Particularly for the speech stimulus, the trace of the adaptation output replicates the envelope of the auditory nerve activity: signal onsets are enhanced and the response decays during stimulus duration.

Fig. 3 shows the effect of adaptation processing on the logarithmic mel-frequency spectrogram for an utterance from the AURORA 2 test set (spoken digit "two"; female speaker), without noise (left column) and with car noise (right column).

IV. DESIGN OF AUTOMATIC SPEECH RECOGNITION EXPERIMENTS

We evaluated the effect of our simplified adaptation model on ASR performance using the AURORA 2 and AURORA 3 speech recognition tasks, using plain MFCC features as a baseline. For comparison, we also evaluated the effect of other robustness methods: RASTA and cepstral mean subtraction, which can be viewed as related temporal filtering techniques, and Wiener filtering, which is designed specifically for additive noise.

A. Feature extraction

1) Baseline MFCC features: We used the ETSI Distributed Speech Recognition reference source code for MFCC (available from www.etsi.org) to calculate 13 cepstral coefficients including C0. Delta and double-delta features were calculated within HTK using two frames of past context and two frames of future context. All other techniques for feature extraction used in our experiments, including our simplified adaptation model, were incorporated as additions onto this baseline.

2) RASTA: For RASTA filtering, we used a publicly available Matlab implementation of RASTA [23]. We inserted the RASTA filter into the MFCC calculation just after the calculation of logarithmic mel spectra.

The logarithmic domain in which our adaptation model and RASTA operate is a natural domain for attempts to compensate for convolutional distortions (since, ignoring the effects of framing and mel filtering, convolution in time becomes multiplication in frequency, and this multiplication becomes an addition when a logarithm is taken). For distortion due to additive noise, however, compensatory signal processing may be easier in a linear domain. This perspective motivated the development of the J-RASTA [9] variant of RASTA, in which the domain is closer to logarithmic when levels of background noise are low and closer to linear when levels of background noise are high. Rather than evaluate J-RASTA, we chose to evaluate RASTA preceded by Wiener filtering, as the performance of this can be compared to the performance of cepstral mean subtraction and our adaptation model themselves preceded by Wiener filtering.

3) Cepstral mean subtraction: We also tried cepstral mean subtraction (CMS), in which the mean of the MFCC features across frames is calculated and then subtracted from the features. CMS, like RASTA, suppresses DC, but unlike RASTA it suppresses only DC and it is noncausal. The noncausality of the CMS technique can be a drawback in interactive ASR applications (there are causal and nearly-causal variants but

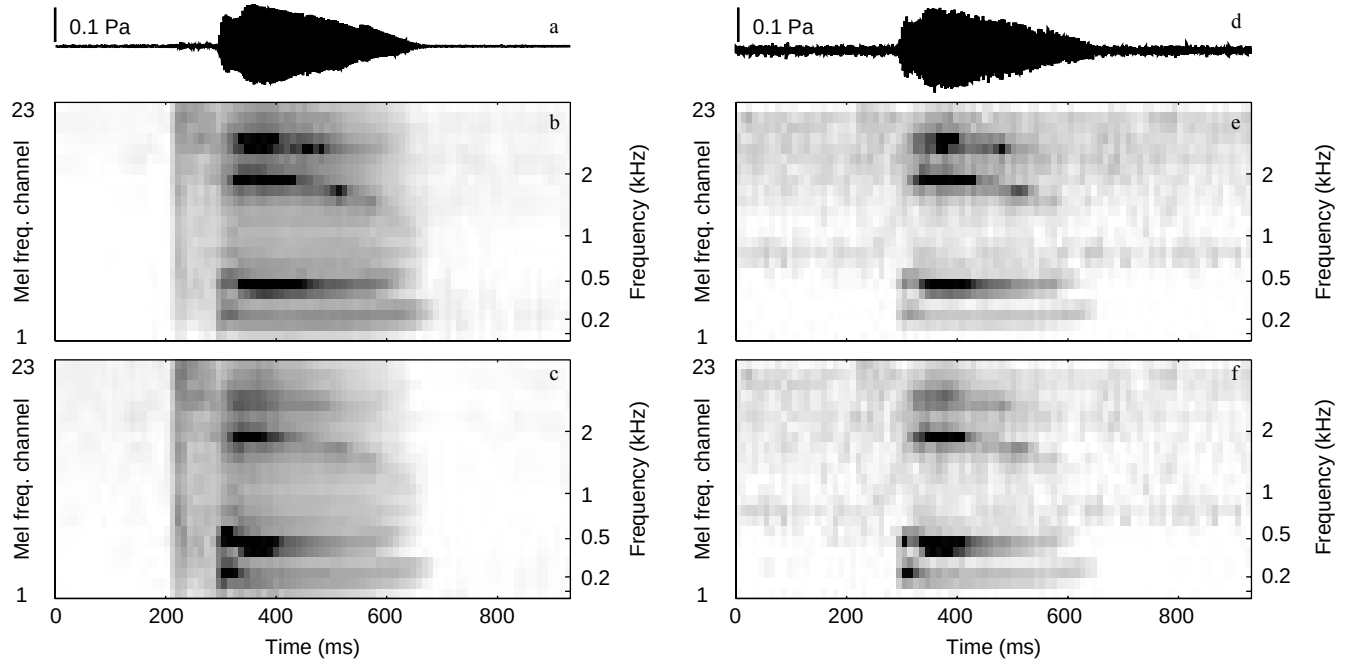


Fig. 3. Comparison of mel-spectrograms (b+e) and mel-spectrograms with adaptation (c+f) in response to the spoken digit "two" (AURORA 2, utterance FCJ_2A) in clean condition (left column), and with 15 dB SNR car noise (right column).

we will not consider them here). On the other hand, RASTA and our adaptation filter (described below) are left-context-dependent, which could be a disadvantage in systems which, unlike in this work, use context-independent acoustic modeling [9], [24]. CMS was of interest to us because it is a common ASR robustness technique and because it and RASTA are sometimes viewed as competing alternatives (although they may also be combined with each other, as in [25]). Further discussion of CMS, along with an examination of improving CMS performance through incorporating speech detection, can be found in [26].

4) Wiener filter: Since the ASR tasks we used contain large amounts of additive noise, we were interested in trying a robustness method designed specifically for that distortion. We used a Wiener filter implementation originally developed as part of the Qualcomm-ICSI-OGI ASR front end described in [25][2]. This performs Wiener filtering with engineering modifications such as a noise over-estimation factor, smoothing of the filter response, and a spectral floor. The noise power spectrum estimate is initialized using the beginning frames of the input waveform before the start of speech, and updated using later frames which are judged as non-speech according to a frame energy test. We ran the Wiener filter prior to all other feature extraction, as a pre-processing stage which created noise-reduced waveforms.

B. Recognition tasks and recognizer back end

We used the AURORA 2 and AURORA 3 speech recognition tasks. AURORA 2 [27] is based on the TIDIGITS database (connected digits spoken by both male and female speakers), bandpass filtered to telephone bandwidth, with eight different noise types artificially added at SNRs from -5 dB to 20 dB. This task has a standardized ASR back end: a hidden Markov model (HMM) recognizer implemented with Cambridge's HTK tools. The acoustic modeling uses word-level digit models with 16 states per word, a three-state pause model, and a one-state short pause model which is identical to the middle state of the pause model. Each state is described by a diagonal-covariance Gaussian mixture model. We used the "complex" version of the back end, defined by Asela Gunawardana, in which there are 20 Gaussians per word state and 36 Gaussians per pause state. There are two training conditions in AURORA 2: clean training where no noise was added and multi-condition training with four different noise

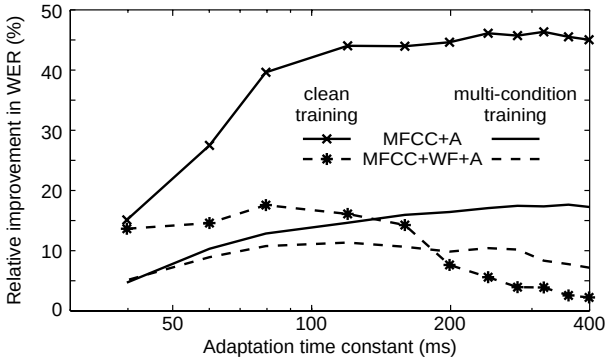


Fig. 4. ASR performance of MFCC in combination with the adaptation model (MFCC+A), and with Wiener filtering and adaptation (MFCC+WF+A) as a function of adaptation time constant. Results are given as relative improvement of word error rates (relative to plain MFCC for MFCC+A and to Wiener filtered MFCC features for MFCC+WF+A).

[2]An archive containing additional description of the Wiener filter along with software for it can be downloaded at http://www.icsi.berkeley.edu/Speech/papers/gelbart-ms/pointers

types added at various SNR levels. There are three test sets: set A uses the same noise types that are used in the multi-condition training, set B uses noise types not found in training, and set C uses noise types not found in training and is also filtered to simulate a channel mismatch. We report our performance separately for each training condition as word recognition accuracies (in percent) averaged over the three test sets and SNRs from 0 dB to 20 dB.

In order to evaluate performance on real-world data, and to judge the generalization of our results, we also used the AURORA 3 task, previously used by [25] and [28] among others. AURORA 3 involves recognition of connected digits strings recorded in in-car environments under various driving conditions, using both close-talking and hands-free microphones. The recordings are a subset of SpeechDat-Car [29][3]. A standardized HTK back end configuration is used, using word-level digit models with 16 states per word and three Gaussians per state, a three-state pause model with six Gaussians per state, and a one-state short pause model which is identical to the middle state of the pause model. There are three different training/testing conditions for AURORA 3. In the well-matched condition, 70% of the data (including both microphone types and all driving conditions) are used as training data, and the remaining 30% are used as test data. In the medium-mismatched condition, only the hands-free recordings are used, with the less noisy driving conditions used for training data and the remainder used for test data. In the highly-mismatched case, the close-talking recordings from all driving conditions are used as training data, and the hands-free recordings from all but the quietest condition are used for the test set.

We present recognition scores averaged over four languages (German, Danish, Finnish and Spanish). When averaging across the different mismatch conditions, we followed the usual convention for AURORA 3 and weighted the high-mismatch, medium-mismatch, and well-matched conditions by 0.25, 0.35, and 0.4 respectively.[4]

The speech data and back end configuration files for the AURORA 2 and AURORA 3 tasks are available from ELDA[5].

V. ASR RESULTS AND DISCUSSION

A. The effect of the adaptation time constant

As discussed in section II-A, adaptation time constant in humans might differ considerably from values measured in animals. Also, adaptation at higher processing levels in the human auditory system with longer time constants might be important for speech processing, and indeed the parameters of adaptation in the human auditory system are not necessarily optimal for ASR. We have therefore evaluated a range of adaptation time constants for ASR. Figure 4 summarizes AURORA 2 speech recognition results (expressed as relative word error rates compared to a front end without the adaptation processing) as a function of adaptation time constants. Adaptation time constants in the range found in gerbils and guinea pigs (40-60 ms) improve recognition scores, but longer time constants provide larger benefits. For the case of clean training with Wiener filtering, the figure shows that the best time constant is 80 ms, but for the other three cases the best time constant lies between 200 ms and 300 ms. We therefore chose an adaptation time constant of 240 ms for all other ASR experiments. This value is four times as large as in gerbils, just as estimated for humans by [19]. Our AURORA 3 experiments provide a generalization test for this choice of time constant.

The adaptation attenuates low frequencies in the modulation spectrum. A given adaptation time constant τ_A corresponds to a high-pass filter with a corner frequency f_c of:

$$f_c = \frac{1}{2\pi\tau_A} \tag{2}$$

Arai et al. [21] found that the 1–16 Hz portion of the modulation spectrum is important for human speech intelligibility. Thus, our time constant of 240 ms, corresponding to a corner frequency of 0.66 Hz, is consistent with their results. Notably, the adaptation time constant is also in accordance with psychoacoustic time constants. As already pointed out by Zwicker, forward masking can last up to 200 ms and the time constant effective in simultaneous masking is also 200 ms [20].

TABLE I
RECOGNITION ACCURACIES (%) FOR THE AURORA 2 TASK.

	Training cond.	MFCC	MFCC + A	MFCC + RASTA	MFCC + CMS
Unprocessed speech	**clean**	56.4	76.5	64.5	69.5
	multi	89.3	91.2	89.9	92.5
Wiener filtered speech	**clean**	78.9	80.1	72.6	76.7
	multi	91.4	92.3	90.8	93.2

A indicates adaptation filtering and CMS cepstral mean subtraction.

B. AURORA 2 feature extraction performance comparison

The performance of the simple adaptation model is plotted in Figure 5 as a function of SNR, while results averaged over SNR are summarized in Table I. The largest improvement occurred when changing from clean to multi-condition training; for plain MFCC the relative improvement in word error rate (WER) was 75.6%. Wiener filtering gave 51.6% / 18.9% relative WER improvement over plain MFCC for clean- and multi-condition training, respectively. Adding the adaptation model (indicated with +A) improves performance by a lesser but still significant amount: 46.1% / 17.0% relative WER improvement over plain MFCC. Given the much lower computational cost of our adaptation model compared to Wiener filtering, this improvement is impressive. Of special interest is that combining adaptation processing with Wiener filtering provides a 5.6% / 10.5% relative improvement in WER over Wiener filtering alone. With and without Wiener

[3]More information can be found online at http://www.speechdat.org

[4]More detailed AURORA 2 and AURORA 3 performance breakdowns than included here can be found online at http://www.icsi.berkeley.edu/Speech/papers/TSAP-Adaptation. We may use the same location in the future to inform our readers about new developments in our work.

[5]The Evaluations and Languages resources Distribution Agency, www.elda.org

TABLE II
RECOGNITION ACCURACIES (%) FOR THE AURORA 3 TASK.

	Mismatch case	MFCC	MFCC + A	MFCC + RASTA	MFCC + CMS
Unprocessed speech	**Well**	91.0	92.5	89.5	91.3
	Medium	70.2	73.9	76.3	79.5
	High	48.4	63.7	67.4	59.2
	Average	**73.1**	**78.8**	**79.3**	**79.2**
Wiener filtered speech	**Well**	92.3	93.4	90.3	93.6
	Medium	68.1	78.6	78.2	82.6
	High	63.0	77.2	76.9	74.2
	Average	**76.5**	**84.2**	**82.7**	**84.9**

filtering, the adaptation model outperforms RASTA and CMS for clean training, while for multicondition training it outperforms RASTA and is outperformed by CMS.

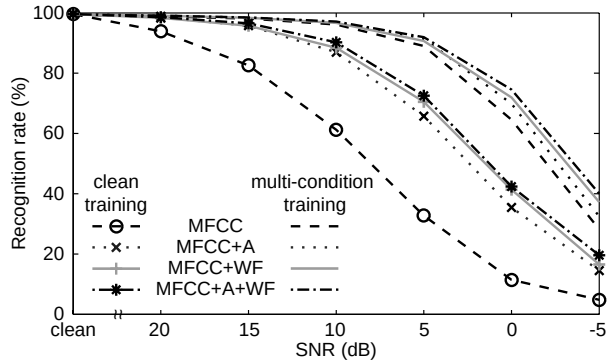


Fig. 5. Recognition accuracy in noise using MFCC based features only, MFCC in combination with our adaptation model (MFCC+A), MFCC with Wiener filtering (MFCC+WF) and MFCC with Wiener filtering and adaptation (MFCC+A+WF).

C. AURORA 3 feature extraction performance comparison

Table II summarizes the results obtained on the AURORA 3 task. Without Wiener filtering, adaptation processing (indicated with +A) improved average recognition accuracy from 73.1% to 78.8%, a relative improvement of 21.2% in the word error rate (WER). RASTA and CMS processing both slightly outperformed adaptation filtering in this case, providing relative improvements of 23.2% and 22.6% respectively. For comparison, introducing Wiener filtering only yielded 12.8% relative WER improvement compared to plain MFCC feature extraction. In the Finnish medium mismatch case, there was, surprisingly, a 60% drop in relative WER from the introduction of Wiener filtering (we have been told we are not the only ones to observe this sort of behavior when applying noise reduction to the Finnish task; Hans-Günter Hirsch, personal communication. Without taking Finnish into account, the relative WER improvement for Wiener filtering was 19.9%.) Adding adaptation processing to Wiener filtering resulted in a relative improvement of 32.7% (41.3% relative to plain MFCC features). This was better than adding RASTA processing, but worse than adding CMS to Wiener filtering.

Looking at the mismatch conditions individually, we see that in the well-matched case adaptation filtering improved recognition scores (16.9% / 14.4% relative improvement, without and with Wiener filtering respectively), while RASTA worsened them. Under high mismatch, adaptation filtering improved recognition scores (29.6% / 38.4% relative improvement), but was outperformed by RASTA processing without noise reduction (36.8% / 37.5%). Since RASTA completely suppresses DC modulation, one might have expected RASTA to outperform adaptation in the high-mismatch condition, but this is not the case if Wiener filtering is used first. In the medium mismatched case, CMS clearly outperformed both adaptation and RASTA processing.

It is notable that adaptation processing alone outperformed Wiener filtering alone in all mismatch conditions and that the combination of adaptation and Wiener filtering always yielded better recognition scores than either method by itself.

VI. CONCLUSION

A number of auditory-based approaches for noise-robust ASR have been proposed by different groups. They include models of particular psychoacoustic or physiological effects (e.g., temporal masking [11]), more comprehensive models of "effective" sound processing [15], [30], and complex, physiologically-based inner ear models [31], [14]. We chose to model a single effect, synaptic adaptation. Compared to past work we chose an approach of greater simplicity, modeling adaptation with a single first-order IIR filter stage which is an incremental addition to conventional MFCC feature extraction. This simplicity reduces computational cost and reduces the engineering effort required for integration into existing ASR systems. This follows the tradition of employing auditory principles in simple or stylized form represented in the very popular MFCC and PLP feature extraction methods. (Given the superb performance of human speech recognition, we do also see great merit in research on detailed auditory modeling for ASR, and indeed are pursuing it ourselves [32].)

Using the AURORA 2 recognition task, we experimented with the effect of the short-term adaptation time constant in our model, and found a time constant of 240 ms appropriate, which is consistent with estimates for adaptation in the human auditory nerve [19]. With that time constant, incorporating our model into MFCC calculation reduced AURORA 2 word error rate by 46% relative for clean training and by 17% relative for multicondition training. This was less than the improvement from introducing Wiener filtering (51.6% / 18.9%), but still notable considering the lower computational cost and implementation complexity of our adaptation model compared to the Wiener filter. Furthermore, when combined with Wiener filtering adaptation provided further WER reductions relative to Wiener filtering alone.

On the AURORA 3 task, of the three methods evaluated, adaptation was the only one to outperform Wiener filtering in all mismatch conditions, and like CMS it consistently improved recognition scores when combined with Wiener filtering. Thus we see that our adaptation model is an effective strategy for feature robustness. Considering this, and its useful

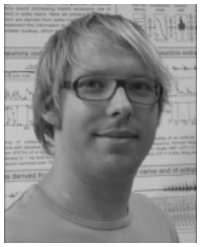
properties of simplicity and causality, it seems its employment in ASR systems deserves serious consideration.

ACKNOWLEDGMENT

We want to thank Nelson Morgan, Hynek Hermansky, Adam Janin, Carmen Pelaez, Marc Ferras and Qifeng Zhu for their advice and the reviewers for their valuable comments. This work was funded in part by the German Federal Ministry of Education and Research (reference number 01GQ0443).

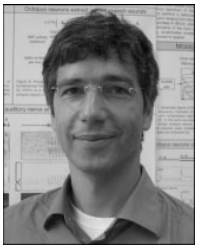
REFERENCES

[1] H. Hermansky, "Should recognizers have ears?" *Speech Commun.*, vol. 25, pp. 3–27, 1998.
[2] B. Kollmeier, "Auditory principles in speech processing: Do computers need silicon ears?" in *Proc. Eurospeech*, Geneva, Switzerland, 2003, pp. 5–8.
[3] 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.
[4] M. N. Kvale and C. E. Schreiner, "Short-term adaptation of auditory receptive fields to dynamic stimuli," *J. Neurophysiol.*, vol. 91, pp. 604–612, 2004.
[5] N. Ulanovsky, L. Las, and I. Nelken, "Processing of low probability sounds by cortical neurons," *Nature Neurosc.*, vol. 6, pp. 391–398, 2003.
[6] R. L. Smith, "Short-term adaptation in auditory-nerve fibers: Some poststimulatory effects," *J. Neurophysiol.*, vol. 40, pp. 1098–1112, 1977.
[7] L. A. Westerman and R. L. Smith, "Rapid and short-term adaptation in auditory nerve responses," *Hear. Res.*, vol. 15, pp. 249–260, 1984.
[8] W. S. Rhode and P. H. Smith, "Characteristics of tone-pip response patterns in relationship to spontaneous rate in cat auditory nerve fibers," *Hear. Res.*, vol. 18, pp. 159–168, 1985.
[9] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 578–589, 1994.
[10] A. J. Oxenham, "Forward masking: Adaptation or integration?" *J. Acoust. Soc. Am.*, vol. 109, pp. 732–741, 2001.
[11] B. Strope and A. Alwan, "A model of dynamic auditory perception and its application to robust word recognition," *IEEE Trans. Speech Audio Processing*, vol. 95, pp. 451–464, 1997.
[12] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," *J. Phonetics*, vol. 16, pp. 55–76, 1988.
[13] C. Jankowski, H. Vo, and R. Lippmann, "A comparison of signal processing front ends for automatic word recognition," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 286–293, 1995.
[14] F. Perdigão and L. Sá, "Auditory models as front-ends for speech recognition," in *Proc. NATO ASI on Computational Hearing*, Il Ciocco, Italy, 1998, pp. 179–184.
[15] J. Tchorz and B. Kollmeier, "A model of auditory perception as front end for automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 106, pp. 2040–2050, 1999.
[16] 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. Acoust. Soc. Am.*, vol. 111, pp. 2178–88, May 2002.
[17] ——, "Adaptation in a revised inner-hair cell model," *J. Acoust. Soc. Am.*, vol. 113, pp. 893–901, February 2003.
[18] T. C. Chimento and C. E. Schreiner, "Time course of adaptation and recovery from adaptation in the cat auditory-nerve neurophonic," *J. Acoust. Soc. Am.*, vol. 88, pp. 857–864, 1990.
[19] A. Spoor, J. J. Eggermont, and D. W. Odenthal, "Comparison of human and animal data concerning adaptation and masking of eighth nerve compound action potential," in *Electrocochleography*, J. Ruber, C. Elberling, and G. Solomon, Eds. Baltimore, MD: University Park, 1976, pp. 183–198.
[20] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, 2nd ed. New York: Springer-Verlag, 1990.
[21] T. Arai, H. Hermansky, M. Pavel, and C. Avendano, "Intelligibility of speech with filtered time trajectories of spectral envelopes," in *Proc. ICSLP'96*, Philadelphia, USA, 1996, pp. 2490–2493.
[22] 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.
[23] D. Ellis. (2004) PLP and RASTA in matlab. [Online]. Available: http://www.ee.columbia.edu/˜dpwe/resources/matlab/rastamat/
[24] J. de Veth and L. Boves, "Channel normalization techniques for automatic speech recognition over the telephone," *Speech Commun.*, vol. 25, pp. 149–164, 1998.
[25] A. Adami *et al.*, "Qualcomm-ICSI-OGI features for ASR," in *Proc. ICSLP'02*, Denver, Colorado, 2002, pp. 21–24.
[26] C. Mokbel, D. Jouvet, and J. Monné, "Deconvolution of telephone line effects for speech recognition," *Speech Commun.*, vol. 19, pp. 185–196, 1996.
[27] H. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," in *Proc. ISCA ITRW ASR*, Paris, France, 2000, pp. 181–188.
[28] D. Macho *et al.*, "Evaluation of a noise-robust DSR front-end on AURORA databases," in *Proc. ICSLP'02*, Denver, Colorado, 2002, pp. 17–21.
[29] A. Moreno *et al.*, "Speechdat-car: A large speech database for automotive environments," in *Proc. LREC*, Athens, Greece, 2000, pp. 895–900.
[30] M. Kleinschmidt, J. Tchorz, and B. Kollmeier, "Combining speech enhancement and auditory feature extraction for robust speech recognition," *Speech Commun.*, vol. 34, pp. 75–91, 2001.
[31] H. Sheikhzadeh and L. Deng, "Speech analysis and recognition using interval statistics generated from a composite auditory model," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 90–94, 1998.
[32] W. Hemmert, M. Holmberg, and D. Gelbart, "Auditory-based automatic speech recognition," in *ICSA SAPA'04*, Jeju Island, Korea, 2004, pp. 1–6.



Marcus Holmberg Marcus Holmberg was born in Villstad, Sweden. He received the M.Sc. degree in electrical engineering from the Chalmers University of Technology in Gothenburg, Sweden, in 2003. Since 2003, he has been pursuing a PhD from Darmstadt Technical University, Germany. He is working as a research engineer at Infineon Technologies, Corporate Research, in Munich. His research interests include auditory modeling, spiking neuronal networks and robust speech recognition.

David Gelbart David Gelbart was born in Vancouver, Canada, in 1977. He received a bachelor's degree in computer science and mathematics from the University of British Columbia in 1999. Since 2000 he has been pursuing a PhD in automatic speech recognition from the University of California Berkeley as a member of the Speech Group at the International Computer Science Institute.



Werner Hemmert Werner Hemmert was born in Moosburg, Germany, in 1964. He received the M.Sc. degree in electrical engineering and computer science from the Technical University Munich, Germany, in 1991. He then joined the Hearing Research Laboratories at the ENT hospital in Tübingen, where he worked on the micromechanics of the inner ear. For his three-dimensional vibration measurements he won the 1996 Helmholtz Prize of the PTB, Germany. In 1997, he received the Ph.D. degree from the Ruhr-University Bochum, Germany. From 1998 to 2000, he investigated technical and biological micromechanical systems at the Massachusetts Institute of Technology, Cambridge. At the IBM Research Laboratories, Rüschlikon/Switzerland (2000-2001), he developed a life-sized, hydrodynamical, micromechanical inner ear model. Since 2001, he is working on robust speech recognition at the Corporate Research Department of Infineon Technologies, Munich.