Auditory-based Automatic Speech Recognition

Marcus Holmberg[1], David Gelbart[2], Werner Hemmert[1]

[1] *Infineon Technologies AG, Corporate Research, 81730 München, Germany,*

Email: {marcus.holmberg,werner.hemmert}@infineon.com

[2] *International Computer Science Institute, Berkeley, CA 94704-1198, USA, Email: gelbart@icsi.berkeley.edu*

Introduction

We have developed a detailed model of the human inner ear, which replicates its spectral and temporal sound processing. In this paper we test our model's spectral coding quality on a standardized automatic speech recognition (ASR) task. Furthermore, we evaluate a simple model of neuronal processing known as adaptation with ASR. Adaptation suppresses the response to stationary signals over time and emphasizes "novel" information. Adaptation takes place at all levels of neuronal processing; it is already prominent at the level of the auditory nerve.

Speech recognition with auditory model

The auditory model we use for this part of the study was described in [3]. The model achieved filter shapes and dynamic compression of more than 60 dB in accordance with recent psychoacoustic measurements [3].

Fig. 1 (top row) shows the initial 50 ms of an artificial vowel [ɜ˞] as in *bird* and the modeled inner ear response. Fig. 1b shows basilar membrane (BM) displacement. The inner ear performs an almost logarithmic frequency decomposition, except at low frequencies where the scale becomes more linear. In the spectral domain, the model resolves the fundamental frequency (80 Hz; 33.8 mm) and the first two harmonic frequencies (31.5 mm and 29.7 mm); at higher frequencies it codes the vowel's spectral envelope. Fig. 1 (bottom row) illustrates the model's response as a function of time at three different output stages to the same stimulus. The frequency channel plotted responds best to 490 Hz (indicated by an arrow in Fig. 1b), which coincides with the first formant of the vowel. The traveling time in the inner ear causes the delay of the responses. BM displacement drives the model of the inner hair cells, the inner ear's sensory cells. These were modeled according to [5]. In the first stage, the mechanical vibrations are transduced into an electrical potential, the so-called receptor potential (RP). This step includes a soft half-wave rectification (compare Fig. 1d) and a saturating nonlinearity. At this stage, matching of inner ear compression and the sensory cell's dynamic range is crucial. The receptor potential drives synaptic mechanisms that lead to an action potential, or spike, in the auditory nerve (see [5] for details). Fig. 1e pictures transmitter release (TR) into the synaptic cleft. Transmitter is released via synaptic vesicles into the synaptic cleft; vesicle release rate is proportional to auditory nerve firing rate, except that refraction of the fiber is not taken into account. Refraction means that a neuron cannot fire twice within a short time interval (about 1 ms), but has to recover.

The coding of the glottis impulses (80 Hz) in the temporal domain is prominent at all model stages. The frequency channel also shows phase-locking to 480 Hz – the 6^{th} harmonic of the fundamental. Phase-locking diminishes above 1 kHz due to properties of the sensory cell. The coding of the fundamental frequency however remains prominent also in higher frequency channels.

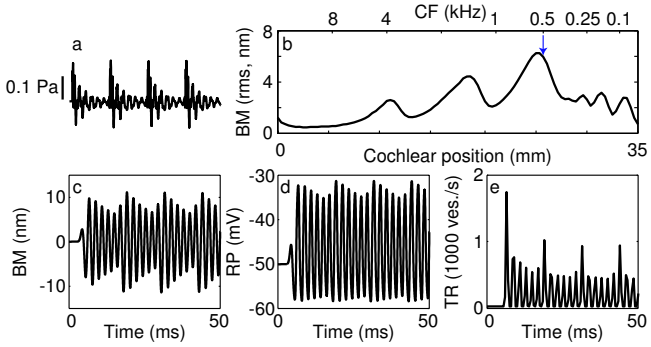

Figure 1: Human inner ear model response to an artificial vowel [ɜ˞] as in "*bird*". a) Stimulus. b) Basilar membrane (BM) displacement (rms-value over the whole vowel duration) as a function of location along the inner ear.The responses in the lower panels were calculated in a channel that responded best to 490 Hz (indicated by an arrow in panel b). c) BM displacement d) Receptor potential (RP) e) Transmitter release (TR). TR was plotted using 0.5 ms time-bins.

Our auditory model – like the human inner ear – works with a much higher resolution, both in frequency and time domain, than a normal ASR back-end can process. To make our model compatible with a standard speech recognizer (we use a back-end based on the HTK-recognizer), we had to reduce dimensionality of the features. In the time-domain, we averaged (or in the BM case we calculated rms-values) over a 25 ms Hanning window, which we advanced by 10 ms. As a result, the fine-grained temporal information such as the speaker's fundamental frequency, which is prominent in Fig.1, was lost. We thus rely just on the spectral filtering of the model, much like a standard ASR front-end. In the frequency domain, we integrated the frequency range from 5 kHz and downwards with another Hanning window (width 4.5 mm, advanced by 1.8 mm, compare Fig. 1b).

Results obtained on the AURORA-test [2] are shown in

Fig. 2 (see caption for details). BM and RP features achieved about the same recognition rates as standard mel-frequency cepstrum coefficients (MFCC, solid line with circles), whereas the TR-feature performed considerably worse.

A frame-based adaptation model

Many properties of standard ASR features have counterparts in early processing stages of the ear: the mel-spaced frequency channels have their direct counterpart in the inner-ear frequency decomposition; the logarithmic scaling of feature amplitudes corresponds to the inner-ear compression. The strong accentuation of signal onsets seen in the auditory nerve however, has no direct counterpart in ASR feature extraction. We therefore built a simple model of adaptation and evaluated it using the AURORA test. We applied our adaptation model on standard MFCC-feature calculation as an extra step between the logarithmic mel-spectrum calculation and the cepstrum calculation.

The response of the auditory nerve to a steady-state stimulus (compare the modeled response in Fig. 1e) can be described as a sum of a steady-state term and two exponentially decaying terms (time constants in the order of 60 ms and 3 ms, respectively [6]). Since the so-called rapid term has a time-constant (3 ms) much shorter than the window used to calculate the mel-frequency vectors (25 ms), we omitted the corresponding term. The output of our adaptation model was the sum of a high-pass filtered version of the input signal (first order high-pass filter with corner frequency $f_c = 1$ Hz) and the signal itself. The corner frequency corresponds to a time constant of 160 ms, much longer than the 60 ms found in physiological measurements on animals [6]. However, adaptation time constants seem to be considerably longer in humans than in animals [4]. Our adaptation model has common traits with the widely used RASTA-technique [1]. However, RASTA lacks the proportional term of the adaptation model, and thus its response to a constant input will decay to zero over time.

Speech recognition results with adaptation are shown in Fig. 2 (solid lines). The adaptation filtering outperforms both plain and RASTA-processed MFCC-features. On average, the relative improvement in word error rate is 41% compared to MFCC and 31% compared to RASTA (clean training condition).

Discussion

With our inner ear model features, we achieve speech recognition scores comparable to those of MFCC features. This is quite remarkable considering the many nonlinear transformation steps included in the model – steps needed to preserve the temporal fine-structure of sound signals in the auditory nerve response. The transmitter release (TR) feature performs worse than the other features in the clean training condition shown in Fig. 2. However, for multi condition training, where the recognizer is presented with noisy utterances also in the training phase, TR features performed slightly better than RP (data not shown).

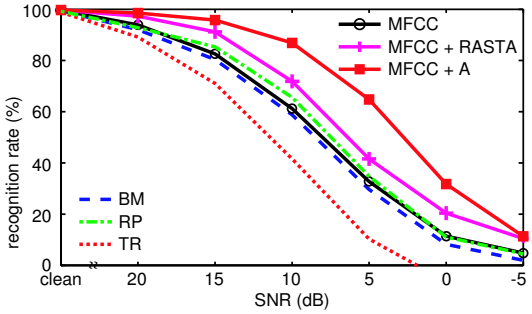

Figure 2: Recognition rates for the features derived from our inner ear model and from MFCC's. As test-bed we used the AURORA test [2]. Recognition results are averages over all three AURORA test sets for the clean training condition. The inner-ear model features (BM, RP and TR) perform about as good as standard MFCC-features, except TR which performs substantially worse. Adaptation features (MFCC+A) outperform both plain MFCC-features and RASTA-processed ones.

In this paper we only used the spectral coding (known as rate-place coding) of our inner ear model; all temporal information on a shorter time-scale than 25 ms was lost when we extracted features for the speech recognizer. We hypothesize that evaluation of the temporal information, which is present in inner ear processing but not in MFCCs, is crucial for robust speech recognition in noisy environments. This will be the subject for further studies.

One goal of our work is to use knowledge of human sound processing to improve conventional ASR systems. The large improvement in recognition scores with our adaptation model is a successful example of such a strategy.

Acknowledgements

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

References

[1] Hermansky H., Morgan N., *IEEE TSAP*, vol. 2, pp. 578–589, 1994.

[2] Hirsch H., Pearce D., in *ISCA ITRW ASR*, Paris, France, 2000, pp. 181–188.

[3] Holmberg M., Hemmert W., in *CFA/DAGA'04*, Strasbourg, France, 2004, pp. 773–774.

[4] Spoor A., Eggermont J. J., Odenthal D. W., in *Electrocochleography*, Baltimore, MD: University Park, 1976, pp. 183–198.

[5] Sumner C. J., Lopez-Poveda E. A., O'Mard L. P., Meddis R., *JASA*, vol. 111, pp. 2178–88, 2002.

[6] Westerman L. A., Smith R. L., *Hear. Res.*, vol. 15, pp. 249–260, 1984.