19th INTERNATIONAL CONGRESS ON ACOUSTICS
MADRID, 2-7 SEPTEMBER 2007

COMPARATIVE EVALUATION OF SUCCESSIVE COCHLEAR MODELING STAGES AS POSSIBLE FRONT-ENDS FOR AUTOMATIC SPEECH RECOGNITION

PACS: 43.66.Ba

Harczos, Tamás[1]; Nogueira, Waldo[2]; Szepannek, Gero[3]; and Klefenz, Frank[4]

[1] Fraunhofer Institute for Digital Media Technology (Fraunhofer IDMT), Ehrenbergstrasse 29, 98693 Ilmenau, Germany; **and** Faculty of Information Technology, Péter Pázmány Catholic University, Práter u. 50/a, 1083 Budapest, Hungary; hzs@idmt.fraunhofer.de

[2] Information Technology Laboratory, Leibniz Universität Hannover, Schneiderberg 32, 30167 Hannover, Germany; nogueira@tnt.uni-hannover.de

[3] Department of Statistics, University Dortmund, Vogelpothsweg 87, 44227 Dortmund, Germany; szepannek@statistik.uni-dortmund.de

[4] Fraunhofer IDMT; klz@idmt.fraunhofer.de

ABSTRACT

The ear, our natural spoken language interface, has always been a fascinating research object. During the last few decades a large number of both physiological and computational models of the human auditory system were developed, while only some of them became widely accepted. The evolution of automatic speech recognition (ASR) systems also demonstrated that employing principles having counterparts in the human ear might increase performance. Logarithmic warping of the spectrum, masking, compression and adaptation are some of these techniques, and they rely on different stages of cochlear processing. Since several attempts have already been made to implement ASR on data directly obtained from cochlear models of different accuracy, we decided to create a layered cochlear model, and to evaluate successive modeling stages as possible front-ends for ASR. The distinct stages are basilar membrane motion, vesicle release probability of the inner hair cell, neurotransmitter concentration in the synaptic cleft, auditory nerve activity, and mean activity of auditory nerve populations; the system consists of up-to-date physiological models. Speech recognition tasks are carried out on a significant part of the TIMIT database, including various noise conditions, and the results are discussed.

INTRODUCTION

Even though automatic speech recognition (ASR) has been a much-discussed research topic during the last decades, there is still no technical solution which would be superior to human performance. Moreover, machine recognition error rates exponentially escalate in the presence of increasing noise. Many biologically motivated approaches have been proposed to enhance ASR performance [1] [2], but, unfortunately, most of them can hardly be compared. We have constructed a computer program that effectively simulates the successive cochlear processing stages, which will be employed as front-ends for ASR. Training and testing procedures as well as noise conditions will be varied in a controlled manner so that recognition results remain comparable. We hope that this practical comparison may not only play an important role for future ASR systems, but may help understand what features are used by the human auditory system to recognize speech, which would be an essential step towards better cochlear implants (CIs).

AUDITORY MODEL

Natural audio signal processing of the ear consists of a chain of consecutive steps. The outer ear (auricle and ear canal) and the middle ear (auditory ossicles) work like transducers towards the inner ear (cochlea) and amplify frequencies in the range of 0.2-5 kHz. Via the stapes, hammer and anvil, the incoming sound waves are transmitted to the cochlea through the oval window into propagating fluid waves around the basilar membrane (BM). Due to varying

properties of the BM along the cochlea its movement at any certain position is dominated by a small range of frequencies. This phenomenon is called tonotopic bandpass-filtering. Along the human BM three rows of outer hair cells (OHCs) and one row of inner hair cells (IHCs) are located. The OHCs are assumed to enhance perception, while the inner hair cells are transducers of the mechanical waves into electrical potentials that generate action potentials (APs or spikes) at the auditory nerve (AN) fibres. The signal-transmission inside the IHCs is well described in [3]. On top of any hair cell, three rows of stereocilia follow the movement of the BM. Stereocilia deflection results in opening and closing of $[K^+]$-ion-channels. Influx of $[K^+]$ leads to depolarization (or inversely, hyper-polarization) of the IHC resulting in half-way rectification of the bandpass-filtered sound wave. As a function of the IHC-membrane potential $[Ca^{++}]$-ions enter the IHC and evoke the release of neurotransmitter at its pre-synaptic end. The diffusion of the neurotransmitter across the synaptic cleft causes post-synaptic depolarization of the auditory nerve fibres. If the post-synaptic membrane potential reaches a specific threshold, an AP is generated, after which the nerve fibre underlies some refractory period, where the threshold for firing is largely increased and thus firing is less probable. In average, 10 synaptic end bulbs of auditory nerves are coupled per IHC. The information carried by them will start to be integrated at the next processing stage, the cochlear nucleus.

There are many competing models of each stage of cochlear processing. For the BM part some examples could be gammatone or dual resonance nonlinear (DRNL) filter banks. We chose a less common but more detailed BM model that is based i.a. on Steele's, Taber's [4], Zwicker's and Peisl's work [5]. It has been augmented by Watts [6] in the early nineties, and later further extended and re-implemented for PCs by means of wave digital filters by Baumgarte [7]. We will refer to this model as the extended Zwicker model (EZB). It includes the non-linear behaviour of OHCs and is hence an active BM model. In our case it is configured to have 251 channels with linearly increasing centre frequency on the Bark scale, covering the whole human hearing range. The time step was set to 1/44100 sec, which fits EZB's requirements for our task.

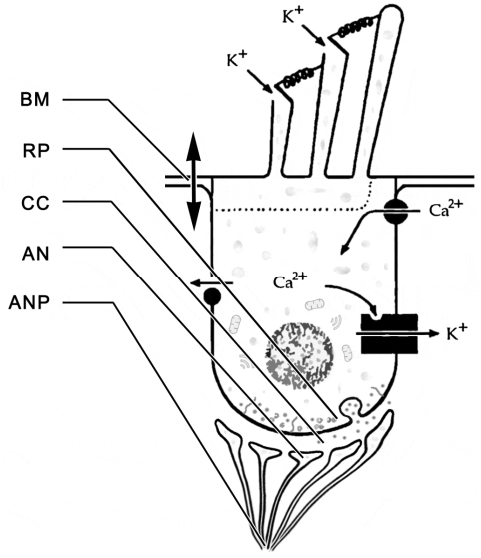

Figure 1.- Schematics of an inner hair cell, and overview of cochlear processing stages

For the calculation of IHC- and AN-dynamics the models of Sumner et al. [3] are used. (Preliminary tests showed that with the much more computationally intensive Hodgkin-Huxley equations [8] better AN results could not be achieved from the aspect of automatic speech recognition.) We have also implemented parts of former work of Meddis [9] to be able to calculate the neurotransmitter concentration in the synaptic cleft in a non-stochastic manner.

Figure 1 gives an overview of the cochlear processing stages discussed during this paper. These are the basilar membrane motion (BM, should be imagined as the vertical displacement of the whole cell), vesicle release probability of the inner hair cell (RP), neurotransmitter concentration in the synaptic cleft (CC), non-stochastic CC (CCns, not indicated in the figure), auditory nerve activity (AN), and mean activity of auditory nerve populations (ANP). Please note that ANP is the mean activity of 12 ANs, which consist of 6 HSR, 4 MSR, and 2 LSR (high/medium/low spontaneous rate, respectively) ANs. For details on differently tuned nerve fibres please see [3].

SPEECH CORPUS

Recognition experiments were conducted on a subset of the DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus [10], which we will refer to as the TIMIT CORE database. It has been built respecting the following rules:

- the subset should be small enough to considerably decrease calculation time, but large enough to allow convergence of the results;

- the ratio of training and test sentences of the full TIMIT database should be maintained;
- no speaker should appear in both the training and testing portions;
- all dialect regions should be represented in both subsets, with at least one female and one male speaker from each dialect;
- the amount of overlap of text material in the two subsets should be minimized;
- all the phonemes should be covered in the test material, each of them should preferably occur multiple times in different contexts.

TIMIT CORE inheritably includes the *Core Test Set*, originally recommended by the makers of TIMIT. It consists of 192 sentences, each having a distinct text prompt, spoken by 24 speakers. Another 576 utterances for training, selected by the rules above, make TIMIT CORE complete. Its 768 utterances have a total duration of over 38 minutes, which allows each phone to occur at least 100 times in average.

It has been shown that the dimension of TIMIT CORE is sufficient for the hidden Markov model (HMM) based training to converge. Random tests verified that, for our purposes, there is no statistically relevant difference in recognition rates between TIMIT-based and TIMIT CORE-based training and testing.

FEATURE EXTRACTION

We tried different feature extraction methods on each distinct type of auditory images (AIs), i.e. on the data of distinct cochlear stages, to see whether and how they affect performance.

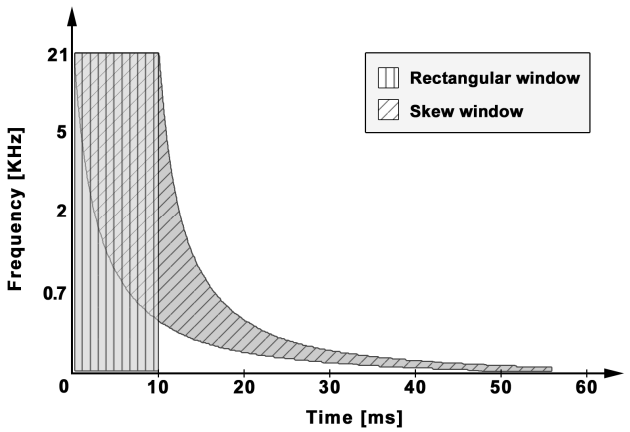

Figure 2.-Rectangular vs. non-rectangular (skew) time windowing

Firstly, AIs with their original 251 channels, are time-windowed, using two different kinds of window-shapes, a rectangular one, and a skewed one, see Figure 2. Both windows have equal width (10 ms) over the channels, i.e. over frequencies, and the weighting factors inside the windows are equally one. Smoothing techniques like rounding at the edges or overlapping along adjacent time-windows are not employed, because of the nature of data to be processed.

Skew windowing tries to make use of the fact that the impulse response (IR) of any stage of the auditory model results in a bunch of delay trajectories of reciprocal fashion [11]. This would effect smearing of frequency components in the case of rectangular time-windowing.

Since each channel includes lots of data along any time-window, a norm will also be used to compress the information. Three types of norms have been tested: absolute average (later referred to as A^2 norm), L^2, and L^∞ norm, which are defined through Table I.

Table I.- Types of norms employed during feature extraction

A^2 norm (Eq. 1)	**L^2 norm** (Eq. 2)	**L^∞ norm** (Eq. 3)
$\|x\|_{A^2} = \frac{1}{n} \cdot \sum_{k=1}^{n} \|x_k\|$	$\|x\|_{L^2} = \sqrt{\sum_{k=1}^{n} \|x_k\|^2}$	$\|x\|_{L^\infty} = \max_{k=1}^{n} \|x_k\|$

Next, we warp the 251-channel frequency information, expressed by the norms of the respective time-windows. A triangular, overlapping filter bank has been designed that has 20 output bands, of which the centre frequencies are equally spaced along the mel-scale. The implementation of the frequency warping is very similar to that of HTK (see [12] for details),

except for the input of the filter bank, which here is Bark-scaled instead of being linearly spaced (like the discrete Fourier transform output).

We also wanted to see if the discrete cosine transform (DCT) manages to increase feature quality by decorrelation, so a transformed version of the vectors will also be tested. At this point, we have the original 20-dimensional filter bank feature vectors and the first 12 cepstral coefficients (completed with the energy term) based 13-dimensional DCT feature vectors.

Since the performance of a speech recognition system can greatly be enhanced by adding time derivatives to the basic feature vectors, as a final step, we add first and second order regression coefficients, which results in 60 (filter bank) or 39 (DCT) dimensions for the final feature vectors. Both kinds of vectors will be tested as exclusive input for speech recognition.

DESIGN OF ASR EXPERIMENTS

Auditory features have been evaluated by a continuous speech recognition system built with the Hidden Markov Model Toolkit (HTK) by Cambridge University [12]. For the hidden Markov modeling left-to-right HMMs are trained for each of the 61 monophones, consisting of 3 states each. States are modeled as Gaussians. Based on the calculated features and the given original phonetic transcriptions of the training material, the system carries out 19 rounds of re-estimation of the HMM parameters using HTK's embedded Baum-Welch algorithm. After the 6^{th}, 9^{th} and 12^{th} iterations of the Baum-Welch re-estimation the Gaussians are split into mixtures of two Gaussians, so that at the end of the training there are 8 Gaussian mixtures per state. State initialization is done by linearly segmenting the utterances (flat init). No advanced grammar model (such as bi-gram, or tri-phone) is used, recognition of phones is therefore completely based on the actual feature vectors.

All recognition tasks were run several times, simulating different noise conditions by adding different level pseudorandom white noise. Preliminary tests showed that the most interesting changes in performance can be obtained by evaluating the 96 dB (no added noise), 24 dB, 18 dB, 12 dB, 9 dB, 6 dB, and 0 dB (both signal and noise are normalized to the same level) signal to noise ratio (SNR) setups.

Each noisy setup (24 dB SNR and beyond) will be evaluated under three different training conditions. These are called *clean*, *mixed*, and *dirty* training. In the case of clean training the HMMs will only be trained on noiseless (96 dB SNR) data, while during dirty training the training material only includes the actual noisy data. In the mixed case training is carried out using 50% noisy and 50% clean data. Please note that the total duration of the training material is invariant among the three conditions. Independent of the actual training variant, recognition tests will always be accomplished exclusively on data corresponding to the actual SNR.

Tests have been made to use reduced monophone sets instead of the original 61-element set. By introducing phone-level equalities, we constructed a reduced set of 38 phones and a minimal set of 27 most common phones, and repeated each recognition experiment accordingly. Since the variability decreased by employing the restricted phone sets, overall recognition rates increased (by over 10% in average), as expected. Still, performance trends were not affected, and therefore these additional results will not be included in the final evaluation.

ASR RESULTS

An overview of ASR results is presented in Figure 3, where only the best results among different norms and time-windows have been included in the plot. Please keep in mind that a large phonetic set was used without grammar. An analysis of variance (ANOVA) has also been implemented [13] to linearly decompose the observed variation in experimental data (excluding results obtained with dirty training) into fixed effects of different modeling factors. Recognition result y_i (in terms of word level *correctness*, see [12]) for the i^{th} simulation is modeled as

$$y_i = \hat{y}_i + \varepsilon_i = y_0 + \sum_j a_j x_{j,i} + \sum_{k \neq l} a_{kl} x_{kl,i} + \varepsilon_i + n \qquad \text{(Eq. 4)}$$

where y_0 is the intercept term, $x_{j,i}$ denotes different factor levels (e.g. the type of time windowing for simulation i, being 1 if skewed window is used and 0 if not), and a_j represents the associated effects to the factors. A level is picked for each factor and its effect is set equal to zero for comparison with the effects of all other levels. Of special interest are a_{kl} standing for interactions between any two levels of different factors $x_{kl,i}=x_{k,i}x_{l,i}$. This way, effects of different model stages

can be detected under varying noisy conditions, for example. Term ε_i denotes unexplained errors and will be minimized by the model in the least squares sense.

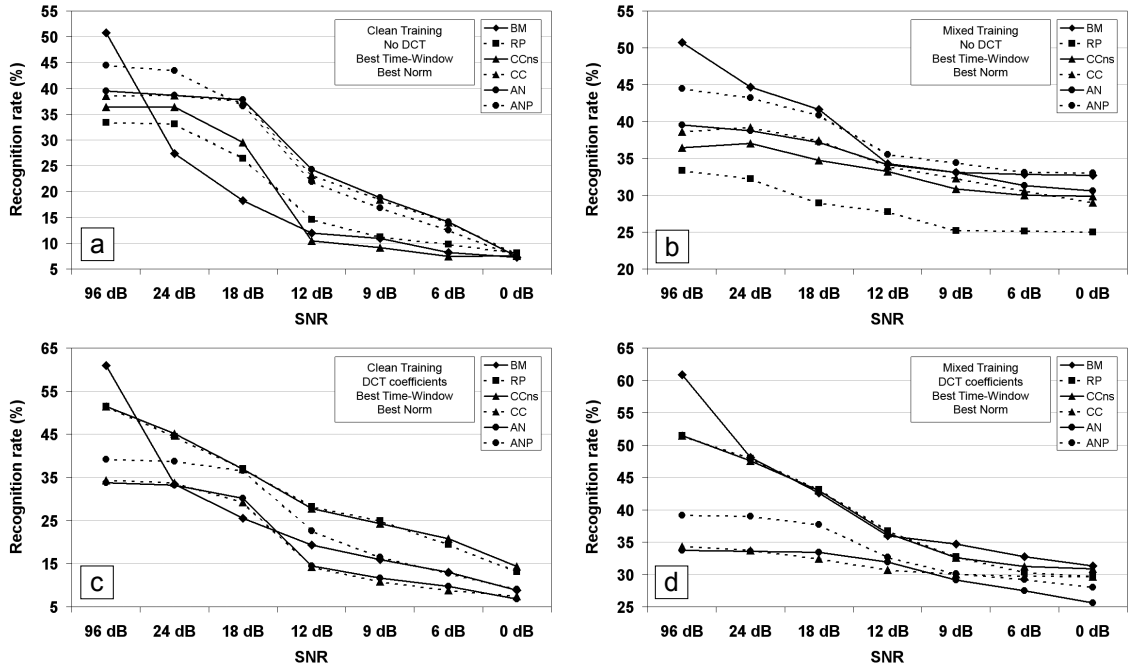

Figure 3.- Overview of recognition rates in terms of word level correctness

Since all experiments are conducted on the same speech data base observed results are far from independent. But even if this theoretical assumption is not met, results are quite well interpretable in a descriptive way and show hints towards a more effective feature extraction. The proportion of explained variability in the data, R^2, is 0.94. This lets us conclude that the investigated factors are meaningful ones.

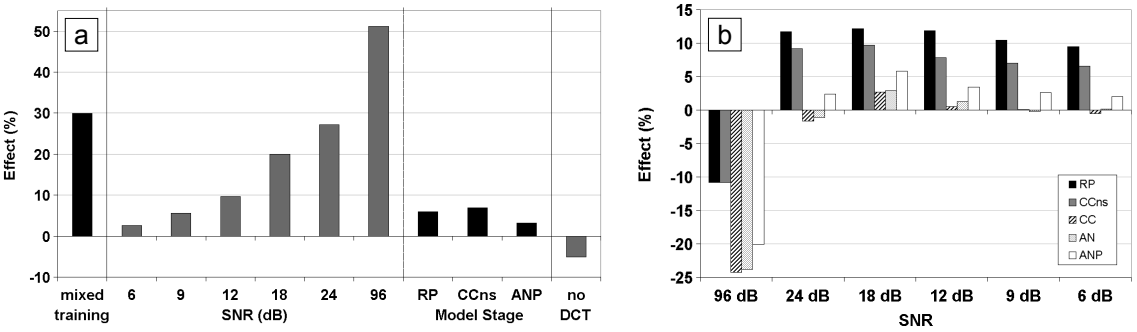

Figure 4.- (a) Main effects of different factors compared to “clean training, 0 dB SNR, BM features with A^2 norm and DCT” setup, only factors with an absolute effect of at least 1% are shown; (b) Main and 2nd order effects with reference to “BM features at 0 dB SNR”.

The strongest effect on the recognition performance results from SNR as well as the chosen training condition (see Figure 4a). In case of noisy speech recognition tasks higher level auditory modeling has an additional positive effect on the recognition rates as compared to directly using BM motion for feature extraction. Recognition can be improved up to more than 10% using a non-stochastic cleft model or the neurotransmitter release probability (see Figure 4b). For clean speech, directly using BM showed to be better. On the other hand, even if the interactions suggest using RP as best basis for feature extraction under noisy conditions there are additional negative effects as opposed to employ CCns under mixed training (~3.9%), and when no DCT is performed (~4.4%). All in all, the use of non-stochastic cleft model may be the best choice under noisy conditions, which is in harmony with the observed results from Figure 3.

Different norms as well as skew windowing seem not to show any strong effect on the recognition rates, while performing an additional discrete cosine transform on the features improves results by about 10% in average. However, the latter effect vanishes for later stages of the auditory model (CC, AN, ANP). Here, the results get better by directly using filter bank outputs.

CONCLUSIONS

Humans still highly outperform ASR systems under non-ideal conditions [14]. Adaptation, which mainly takes place in the cochlear apparatus by means of neurotransmitter release and refractoriness of the auditory nerves, is assumed to be the origin of this phenomenon. This should also be the motivation for further steps in modeling the last stages of peripheral auditory processing with special respect to its contribution to speech recognition.

Our purpose was not to build and ultimate ASR, but to show effects of employing consecutive levels of auditory modeling. This way, we tried to give hints for both building more robust ASR systems and for possible improvements of cochlear implant processing strategies. Results are especially encouraging for current CIs, where very simple models are used to bypass the peripheral auditory system, and today, the performance of these devices is highly affected by environmental noise [15]. We analysed the effect of noise systematically and found that higher level auditory modeling can indeed be an advantage in terms of noise-robustness.

ACKNOWLEDGMENT

The authors would like to thank András Kátai and Stephan Werner for their faithful co-operation.

References:

[1] M. Holmberg, D. Gelbart, W. Hemmert: Automatic speech recognition with an adaptation model motivated by auditory processing. IEEE Trans. Speech Audio Processing **14, No.1** (2006) 43-49

[2] M. E. Munich, Q. Lin: Auditory Image Model features for Automatic Speech Recognition. **9th** European Conference on Speech Communication and Technology, Interspeech (2005) 3037-3040

[3] C. J. Sumner, L. P. O'Mard, E. A. Lopez-Poveda, R. Meddis: A revised model of the inner-hair cell and auditory nerve complex. Journal of the Acoustical Society of America **111, No.5** (2002) 2178-2189

[4] C. R. Steele, L. A. Taber: Comparison of WKB and finite difference calculations for a two-dimensional cochlear model. Journal of the Acoustical Society of America **65** (1979) 1001-1006

[5] E. Zwicker, W. Peisl: Cochlear preprocessing in analog models, in digital models, and in human inner ear. Hearing Research **44** (1990) 209-216

[6] L. Watts: Cochlear Mechanics: Analysis and Analog VLSI. California Institute of Technology, Pasadena, California (1993) PhD dissertation

[7] F. Baumgarte: Ein psychophysiologisches Gehörmodell zur Nachbildung von Wahrnehmungsschwellen für die Audiocodierung. Universität Hannover, Germany (2000) PhD dissertation

[8] A. L. Hodgkin, A. F. Huxley: A Quantitative Description of Membrane Current and its Application to Conduction and Excitation in Nerve. Journal of Physiology **117** (1952) 500-544

[9] R. Meddis: Simulation of mechanical to neural transduction in the auditory receptor. Journal of the Acoustical Society of America **79, No.3** (1986) 702-711

[10] J. Garofolo, L. Lamel, W. Fiesher, J. Fiscus, D. Palett, N. Dahlgren: DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus. NISTIR 4930, NIST, Gaithersburgh, MD (1993) Technical report

[11] S. Greenberg, D. Poeppel, T. Roberts: A space-time theory of pitch and timbre based on cortical expansion of the cochlear traveling wave delay. Psychophysical and Physiological Advances in Hearing, London (1998) 293-300

[12] S. Young, G. Everman, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland: The HTK Book (v 3.3). Cambridge University, Engineering Dept. (2005)

[13] J. R. Turner, J. F. Thayer: Introduction to analysis of variance: design, analysis & interpretation. Thousand Oaks, California, SAGE Publications (2001)

[14] A. V. Ivanov, A. A. Petrovsky: Anthropomorphic feature extraction algorithm for speech recognition in adverse environments. Speech and Computer, SPECOM (2004) 166-173

[15] J. J. Remus, L. M. Collins: The Effects of Noise on Speech Recognition in Cochlear Implant Subjects: Predictions and Analysis Using Acoustic Models. EURASIP Journal on Applied Signal Processing **18** (2005) 2979-2990