Towards automatic speech recognition based on cochlear traveling wave delay trajectories

Tamás Harczos[1,2], Gero Szepannek[3], and Frank Klefenz[1]

[1] *Fraunhofer Institute for Digital Media Technology (Fraunhofer IDMT), 98693 Ilmenau, Germany; and Faculty of Information Technology, Péter Pázmány Catholic University, Budapest, Hungary*

[2] *Department of Statistics, University Dortmund, 44227 Dortmund, Germany*

[3] *Fraunhofer IDMT*

The evolution of automatic speech recognition (ASR) points out that employing principles having counterparts in the human auditory system may lead to better performance. Mel- or bark-warping of the frequencies, masking, compression and adaptation are some of these techniques. Hearing has already been modeled up to the cochlear nucleus (CN) to some degree. However, only few people question, whether one of the very first steps, namely the modeling of the basilar membrane delay trajectories, has been modeled and utilized sufficiently fair. To find the answer, we use an extraordinarily precise auditory model, and try to extract the excitation-dependent shapes of the delay trajectories. We use these features without any other spectral information to carry out speech recognition tasks under different noise conditions on the TIMIT database. We found that the shapes of the cochlear delay trajectories carry precious information, which can be extracted even in the presence of noise. This finding may play an important role in next generation cochlear implants.

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. Conventional speech recognition engines employ popular feature extraction methods like perceptual linear prediction (PLP) or mel-frequency cepstral coefficients (MFCC). Already a number of remarkable extensions, e.g. relative spectral transform for PLP (RASTA-PLP), see Hermansky and Morgan (1994), or auditory processing motivated adaptation for MFCC as in Holmberg *et al.* (2006), had been proposed, with which a tiny performance growth could always be achieved. Notwithstanding, there has been a very little improvement in the past few years and ASR seems to have reached a performance status-quo. Therefore, it seems reasonable to consider if there is a completely different way of sound representation, which could enable novel recognition techniques.

We chose to investigate the very first steps of any neurobiologically motivated sound processing, namely the frequency decomposition, and concluded to use a very precise auditory model for this task. The system we employ has an active basilar membrane (BM) model that faithfully represents outer hair cell (OHC) amplification, masking,

Auditory signal processing in hearing-impaired listeners. 1st International Symposium on Auditory and Audiological Research (ISAAR 2007). T. Dau, J. M. Buchholz, J. M. Harte, T. U. Christiansen (Eds.). ISBN: 87-990013-1-4. Print: Centertryk A/S.

and, the most relevant in our case, the cochlear delay trajectories (CDTs), which originate from the traveling waves on the BM. It is important to recall that the shape of each trajectory has a correlation to the momentary timbre of the sound, see Greenberg *et al.* (1998). Pitch, on the other hand, can be associated with the length, i.e. the place of decay, of each trajectory. In this paper we only use the timbral information, and we discard pitch and any direct spectral feature.

Motivation behind this idea was that the shapes of cochlear traveling wave delay trajectories are natural features. They form patterns, and if they do so, the auditory system surely makes use of it. On the other hand, the authors are not aware of any public recognition system, or, more important, any cochlear implant (CI) system, which makes use of these features. The fact, that vowels can be classified based on CDTs have been demonstrated by Harczos *et al.* (2006), yet, it has not been investigated if a larger set of phones, including consonants, could be classified that way. Therefore, an experimental speech recognition framework had been built to find an answer.

More details on the used auditory model are introduced in the following section. For the feature extraction an artificial Hubel-Wiesel network (HW-ANN) is employed as in Brückmann *et al.* (2004). First, the architecture and way of processing, then details on the automated speech recognition, and last, but not least, results, arising questions, and conclusions will be discussed.

AUDITORY MODEL

To model all the consecutive processing steps of the human ear, results of neurophysiological studies have to be incorporated. Basilar membrane is generally modeled by a gammatone filterbank, which seems to be a good compromise between accuracy and required computational power. Since we need very high time resolution for a detailed representation of CDTs, we use the system proposed by Baumgarte (2000) which is an extended inner ear model based on Zwicker's former work. Here, BM is divided into consecutive sections, which represent equal-width bands in the Bark scale. The movement of the interconnected sections, including the local amplification effect of the OHCs, is calculated via differential equations in the oversampled time domain.

Some masking is already contained in the BM-model by its nature. However, according to Holmberg *et al.* (2006), the presence of adaptation should further increase recognition capabilities. To introduce adaptation, we utilize two subsequent neural stages: inner hair cells (IHCs), and auditory nerve (AN) synapses. An inner hair cell model that represents the actual state of the art is developed by Sumner *et al.* (2002), and post-synaptic potential of the AN fibers (ANFs) can be modeled by the well-known equations by Hodgkin and Huxley (1952). To allow the synaptic adaptation effect to show up, IHC- and ANF-level calculation was repeated $m=10$ times and the results were averaged. The selection of the value m coincides with the fact that, in average, each human IHC has around 10 connecting afferent nerve fibers. The repetition of calculations makes sense also from the statistical point of view, since the variance $var(\underline{n})$ is proportional to $1/m$ with $\underline{n}$ being the average number of spikes during a short time interval Δt.

FEATURE EXTRACTION

The kind of information to look for and the way to extract it are common questions by designing any recognition or classification system. To make use of CDTs, some simple curve equations and an efficient curve extraction network will be introduced through the next paragraphs.

Cochlear delay trajectories

Greenberg *et al.* (1998) pointed out that the motion of the BM proceeds in an orderly fashion from the base to the point of maximum displacement, beyond which it damps out relatively quickly. The transit of the traveling wave is extremely fast at the base, but slowing dramatically for peak displacements at the apex of the cochlea. It has been showed, furthermore, that the delay trajectories can be efficiently modeled by the simple equation:

$$d_a = f_i^{-l} + k \qquad \text{(Eq. 1)}$$

Using (Eq 1), cochlear delay can be approximated by the given frequency f_i and delay constant k. The simple statement behind the equation above is that delay trajectories, in general, have characteristics similar to that of a reciprocal function. Since our data is digital, it is preferred that an array of matrices represent the possible trajectories. The following equation covers a range of N different curves, with an average inflexion of c.

$$T_i\left[y, \frac{c\cdot(i-1)\cdot(n-y-2)}{(c+y-1)\cdot(n-1)}\right]=1, \qquad \text{(Eq. 2)}$$

where $y=1,2,..,n$ for each $i=1,2,..,n$. T_i is therefore an n-element set of sparse matrices of size n-by-n. Each matrix of T_i has zero elements except for the positions given by (Eq 2). Two examples representing sets of possible curves with given parameters are shown in Fig 1.

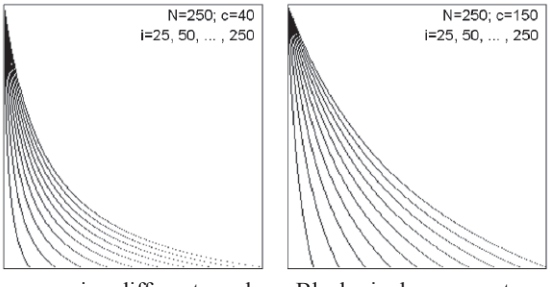

Fig. 1: Possible curves using different c values. Black pixels represent ones in the corresponding quadratic matrix T_i.

Hough-transform via HW-ANN

Emerging auditory nerve activity patterns consist of similar curves to that shown in Fig 1. Now, a robust tool is needed to extract and classify the curvature information. An obvious choice for shape extraction would be the Hough-transform, however, it is known to be computationally intensive, see Van Swaaij *et al.* (1990). Hubel *et al.* (1978) demonstrated the natural orientation columns in the macaque monkey brain, which are believed to perform a kind of parallel Hough-transform, serving the orientation of the monkey by extracting features from the seen image in real-time. One can easily come to the idea of trying to model this naturally brilliant architecture, hoping that the same speed-up can be achieved. Epstein *et al.* (2002) designed a parallel Hough-transform engine, where, in reducing the n-dimensional feature space to two dimensions, the coordinate transform can be executed by a systolic array consisting of time-delay processing elements and adders. Brückmann *et al.* (2004) showed that not only video signals as bars of different slopes, but also audio signals as sinusoids are self-learned by feed-forward timing neural networks. It has also been shown that the training of HW-ANNs can be done in a straight and very fast way if several rules regarding the shapes to be trained are satisfied; see Harczos *et al.* (2006) for details. The previously presented curve equation (Eq 2) satisfies all these rules, and, can therefore be used for quick training of a HW-ANN.

Hough-transformed CDTs

Processing in the auditory model covers the whole audible frequency range. On the other hand, training and test data have an original sampling rate of 16000 Hz. For this reason we first apply cropping on the auditory images, and only the remaining part (around 150 channels) will be Hough-transformed. During the calculations for this paper we used $N=144$ and $c=40$ parameter values to build up the T_i matrices for the parallel Hough-transform.

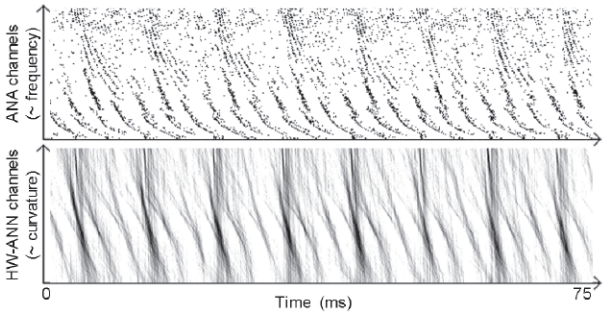

Fig. 2: A highly coherent 75 ms long auditory nerve activity (ANA) pattern induced by phone /eh/ (**top**), and its Hough-transformed image using parameter values of $N=144$ and $c=40$.

Since we are not looking for a rate of activity, but for well-defined shapes in a picture, we expect the Hough-transformed CDT (HCDT) representation to be largely noise-

robust (see findings in Section 5). We can refer to any point of a HCDT image by introducing *H(t,y),* where *t* denotes time (x axis) and y stands for the channel (y axis). For example, a local maximum at *H(τ,n)* would let us conclude that there was a curve with curvature *n* (defined by T_n with given *N* and *c*) at time τ on the input image.

Feature extraction from HCDTs

The HCDT representation (shown in Fig. 2, bottom) contains a great deal of interesting information, but it is largely over-detailed. The fine temporal information generated by the auditory model contains short-time features on time scales smaller than a millisecond. Since short-time features can be speaker- and situation-dependent, furthermore, because a temporally over-detailed sequence is unsuitable for a hidden Markov model (HMM) based ASR, we apply windowing onto the HCDT. We chose to use *l=0.01* second (10 ms) window length to stay comparable to other systems. The channels of the HCDT image will also be grouped into *b=12* bands, i.e., the used windows become 2D windows. The window height is non-equal, and non-linearly distributed, because with *c=40* the most CDTs project to the medial channels of the HCDT. The number of elements in each re-grouped band, in other words, the height of the windows, is given by (Eq. 3).

$$h_i = \left\lfloor 7 \cdot \sin\left(\pi + \pi \cdot \frac{i-1}{b-1}\right)^3 + 15 \right\rfloor, \; i = 1, 2, .., b \qquad \text{(Eq. 3)}$$

During the processing of an L-second long input sound, the whole HCDT image will be split into *b* times *(L\l+1)* windows of data (where '\' stands for integer division). The set of windows is denoted by $W_{p,q}$, where *p=1,2,..,b,* and *q=1,..,L\l+1*. Feature vectors, later also referred to as HCDTf, will be built up by calculating the variance, and the difference between the two extremes of each window (also referred to as range). Additionally, two quasi-energy terms, and the first order difference of the variance and that of one energy term are added. For a formal description of the structure of feature vectors see (Eq. 4).

$$f_{p,q} = \begin{bmatrix} \lfloor \mathrm{var}(W_{p,q}) \rfloor \\ \left\lfloor \sum_{k=1}^{p} \mathrm{var}(W_{k,q}) \right\rfloor \\ \lfloor \max(W_{p,q}) - \min(W_{p,q}) \rfloor \\ \left\lfloor \sum_{k=1}^{p} \max(W_{k,q}) - \min(W_{k,q}) \right\rfloor \\ \lfloor diff(\mathrm{var}(W_{p,q})) \rfloor \\ \left\lfloor diff\left(\sum_{k=1}^{p} \mathrm{var}(W_{k,q})\right) \right\rfloor \end{bmatrix}, \quad \begin{matrix} p = 1, 2, .., b; \\ q = 1, .., L \setminus l + 1 \end{matrix} \qquad \text{(Eq. 4)}$$

DESIGN OF ASR EXPERIMENTS

We evaluated our recognition concept by developing two independent but identical continuous speech recognition systems, both based on the HTK Toolkit by Woodland and Young (1993). The only difference between the two systems is that they use differently generated front-end feature set. Three-state hidden Markov models are used for classification, where each HMM state initially consists of a Gaussian mixture with

diagonal covariance. Using the calculated features and the given original phonetic transcriptions, both systems carry out 23 rounds of re-estimation of the HMM parameters using an embedded training version of the Baum-Welch algorithm of HTK.

Used features

Both recognition systems process input sounds on a 10 ms window skip basis. The first system performs a conventional pre-emphasis on any 25 ms piece of windowed input sound, and then calculates a 20 channel cepstrum from which the mel-frequency cepstrum coefficients (MFCC) of the first 12 channels, along with an energy term, will form the base features. Additionally, the first and second order derivatives are appended to the feature vector. The second system only employs the HCDT-based features with the same number of feature vector elements as that of the first system.

Training and test data set

Recognition experiments were conducted on the TIMIT database, see Zue *et al.* (1990), using the whole corpus of 6300 sentences. Sound files were normalized in amplitude to -6 dB, but were left unfiltered in any other meaning. It is very important to emphasize that we do not use any kind of grammar, bi-gram or triphone models. Recognition of phones is therefore completely based on the actual feature vectors. A recognized word is stated to be correct, if both its consisting phones and their sequence are recognized correctly. The recognition experiments on both systems were carried out three times using three different set of phones:

- the *full set* of TIMIT phones as in Zue *et al.* (1990),
- a *reduced set* of 38 phones introducing phone-level equalities (this set includes *aa, ae, ah, aw, ay, b, ch, d, dh, dx, eh, er, ey, f, g, hh, ih, iy, jh, k, l, m, n, ng, ow, oy, p, r, s, sh, t, th, uh, uw, v, w, y, z*),
- and a *minimal set* of the 27 most common phones applying even more equalities (this set includes *aa, ah, b, ch, d, eh, er, f, g, hh, ih, iy, k, l, m, n, ow, p, r, s, sh, t, uw, v, w, y, z*).

Since the recognition variability decreases by employing the reduced or minimal phone set, the overall recognition rate is expected to increase. The process of calculating with different phone sets is intended to help highlight problematic phones, or, on the contrary, to point out possible weaknesses of the system.

Noise conditions

All recognition tasks were run five times, simulating different noise conditions. First, the original (clean) TIMIT sounds without added noise were used, and then signal to noise ratios (SNRs) of 15 dB, 10 dB, 7.5 dB, and 5 dB were introduced using white noise. In each situation, the whole training and test corpus was affected, and the HMMs of both recognition systems were re-trained.

ASR RESULTS AND DISCUSSION

We evaluated average correct phone and word recognition rates using three different phone sets. We have kept track of both under different noise conditions. A performance comparison is presented through Table 1. We introduced relative scores to confront the tendencies of both methods. Results show on one hand that MFCC outperforms HCDTf under clean condition, but, on the other hand, performance of MFCC also decreases much more rapidly in the case of increasing noise. In other words, by using HCDT instead of MFCC features, relative degradation decreases with noise.

	Phone set	clean	15 dB SNR	10 dB SNR	7.5 dB SNR	5 dB SNR
MFCC	Full	66.8	43.5	29.4	21.1	13.3
HCDTf		21.5	20.4	19.5	17.2	13.4
Relative		**32.2**	**47.1**	**66.4**	**81.5**	**101.2**
MFCC	Reduced	73.1	48.7	36.3	25.7	16.8
HCDTf		24.6	24.3	21.4	20.2	16.5
Relative		**33.7**	**49.9**	**58.9**	**78.7**	**97.9**
MFCC	Minimal	76.3	50.5	41.0	30.8	21.1
HCDTf		30.5	29.1	27.0	25.3	23.0
Relative		**39.9**	**56.7**	**65.8**	**82.1**	**108.8**

Table 1: Average correct phone recognition rates (%) with MFCC and HCDTf features on the three phone sets. Relative score indicates HCDTf performance compared to that of MFCC.

Noise robustness is also preserved if phone frequencies (numbers of occurrences) are reckoned in the recognition performance, but, by studying the resulting confusion matrices generated by HTK's statistics tool we concluded that the two systems (MFCC-based vs. HCDTf-based) recognize different set of phones most properly.

MFCC identifies the phones /aa/, /ao/, and /ey/ most accurately, which are more frequent than the best-recognized phones of the HCDT-based system, /ay/, /ow/, and /ux/. We also found that in the presence of noise, phone /f/ was always least accurately recognized by the MFCC features, while the same phone was always among the three most accurately identified ones (so was phone /t/, too) in the HCDT-based system. This confirms the possibility of advantageously combining the two feature sets.

CONCLUSION

A large number of different spectral analysis methods for automated speech recognition have already been published; see e.g. Papandreou-Suppappola (2003) for an overview. There have been great efforts to improve some of these by employing principles revealed from the human auditory system as in Moore (1995). At the same time, some of these spectral analysis techniques have been modified, or even simplified, to be utilizable in CI systems, see Zeng *et al.* (2004). It, therefore, seems natural to ask, if biologically reasonable natural features exist, which might be used for both purposes.

By employing a very precise inner ear model we experimented with cochlear delay trajectories originating from the traveling waves on the BM. We found that the vary-

ing shapes of these delay curves, without any additional spectral information, can be used as feature for speech recognition. Average recognition rates do not reach that of the MFCC features (except for very low SNRs), but the system behaves very stable under different noise conditions. We found furthermore that the two feature sets could possibly be combined for a better overall performance.

The presented novel method along with the results confirms that cochlear delay trajectories contain important clues, which can be used for speech recognition purposes. We believe that trajectory shapes would be important features in CI systems to enable increased understanding of speech, and should not be neglected in future CI strategies.

OUTLOOK

The aim of the study was to investigate two aspects of the new HCDT features: noise robustness and information content as compared to standard MFCC features. During our studies we found that the MFCC- vs. HCDTf-based systems recognize different set of phones most properly, which indicates the possibility to combine these features to achieve a better overall performance.

We already initiated a pilot study to investigate this potential. Until now, we combined MFCCs with HCDTfs only for clean speech and performed LDA for extraction of uncorrelated features as in Haeb-Umbach *et al.* (1993). Results were compared to those using only MFCCs. The experiment showed that for several feature vector dimensionalities the recognition rates of the standard MFCC front ends can be increased using both features. We conclude again, that the HCDT features contain additional information that is not included in the MFCCs alone.

ACKNOWLEDGMENT

The authors would like to thank András Kátai, Stephan Werner, and Zoltán Fodróczi for their faithful co-operation, and Prof. Karlheinz Brandenburg for the valuable talks.

REFERENCES

Baumgarte, F. (**2000**). “Ein psychophysiologisches Gehörmodell zur Nachbildung von Wahrnehmungsschwellen für die Audiocodierung,” Ph.D. thesis, Uni. Hannover, Germany.

Brückmann, A., Klefenz, F., and Wünsche, A. (**2004**). “A neural net for 2D-slope and sinusoidal shape detection,” Int. Scient. J. of Computing, **3**/1, 21-26.

Epstein, A., Paul, G. U., Vettermann, B., Boulin, C., and Klefenz, F. (**2002**). “A Parallel Systolic Array ASIC for Real-Time Execution of the Hough Transform,” IEEE Trans. Nuclear Science, **49**/2, 339-346.

Greenberg, S., Poeppel, D., and Roberts, T. (**1998**). “A space-time theory of pitch and timbre based on cortical expansion of the cochlear traveling wave delay,” Psychophysical and Physiological Advances in Hearing Proceedings, London, UK,

293-300.
Haeb-Umbach, R., Geller, D., Ney H. (**1993**). "Improvements in connected digit recognition using linear discriminant analysis and mixture densities," IEEE ICASSP Proceedings, Adelaide, Australia, II / 239-242.
Harczos, T., Szepannek, G., Kátai, A., and Klefenz, F. (**2006**). "An auditory model based vowel classification," IEEE BioCAS Proceedings, London, UK, 69-72.
Harczos, T., Klefenz, F., and Kátai, A. (**2006**). "A neurobiologically inspired vowel recognizer using Hough-transform," VISAPP Proceedings, **1**, 251-256.
Hermansky, H., and Morgan, N. (**1994**). "Rasta processing of speech," IEEE Trans. Speech Audio Process., **2**, 578-589.
Hodgkin, A. L., and Huxley, A. F. (**1952**). "A Quantitative Description of Membrane Current and its Application to Conduction and Excitation in Nerve," J. of Physiology, **117**, 500-544.
Holmberg, M., Gelbart, D., and Hemmert, W. (**2006**). "Automatic Speech Recognition With an Adaptation Model Motivated by Auditory Processing," IEEE Trans. Speech and Language Process., **14**/1, 43-49.
Hubel, D. H., Wiesel, T. N., and Stryker, M. P. (**1978**). "Anatomical demonstration of orientation columns in macaque monkey," J. Comparative Neurology, **177**, 361-380.
Moore, B. C. J. (Editor). (**1995**). "Hearing," Academic Press, ISBN 0-1250-5626-5.
Papandreou-Suppappola, A. (Editor). (**2003**). "Applications in Time-Frequency Signal Processing," CRC Press, ISBN 0-8493-0065-7.
Sumner, C. J., O'Mard, L. P., Lopez-Poveda, E. A., and Meddis, R. (**2002**). "A revised model of the inner-hair cell and auditory nerve complex," J. Acoust. Soc. Am., **111**/5, 2178-2189.
Van Swaaij, M., Catthoor, F., and De Man, H. (**1990**). "Deriving ASIC architectures for the Hough transform," Parallel Computing, **16**, 113-121.
Woodland, P. C., and Young, S. J. (**1993**). "The HTK tied-state continuous speech recognizer," EuroSpeech Proceedings, 2207-2210.
Zeng, F.-G., Popper, A. N., and Fay, R. R. (Editors). (**2004**). "Cochlear Implants: Auditory Prostheses and Electric Hearing," Springer, ISBN 0-3874-0646-8, 2004.
Zue, V., Seneff, S., and Glass, J. (**1990**). "Speech database development at MIT: TIMIT and beyond," Speech Communication, **9**, 351-356.