Automatic Speech Recognition with a Cochlear Implant Front-End

Waldo Nogueira[1], *Tamás Harczos*[2], *Bernd Edler*[1], *Joern Ostermann*[3], *Andreas Buechner*[4]

[1]Information Technology Laboratory, Leibniz University of Hannover, Germany
[2]Fraunhofer Institute for Digital Media Technology, Ilmenau, Germany
[3] Institut fuer Informationsverarbeitung, Leibniz University of Hannover, Germany
[4]Hannover Hoerzentrum, Hannover, Germany

Abstract

Today, cochlear implants (CIs) are the treatment of choice in patients with profound hearing loss. However speech intelligibility with these devices is still limited. A factor that determines hearing performance is the processing method used in CIs. Therefore research is focused on designing different speech processing methods. The evaluation of these strategies is subject to variability as it is usually performed with cochlear implant recipients. Therefore an objective method for the evaluation would give more robustness compared to the tests performed with CI patients.

This paper proposes a method to evaluate signal processing strategies for CIs based on a hidden markov model speech recognizer.

Two signal processing strategies for CIs, the Advanced Combinational Encoder (ACE) and the Psychoacoustic Advanced Combinational Encoder (PACE), have been compared in a phoneme recognition task using the system mentioned above. Results show that PACE obtained higher recognition scores than ACE.

Index Terms: cochlear implant, speech recognition, HMM

1. Introduction

Cochlear implants significantly improve the auditory receptive abilities of people with profound hearing loss [1]. These devices consist of a microphone, a speech processor, a transmitter, a receiver and an electrode array which is positioned inside the cochlea. The electrode array carries a number of electrode contacts that can emit small electrical currents to evoke neural action potentials on the auditory nerve.

Speech processing strategies for cochlear implants determine the excitation patterns within the cochlea and subsequently have a strong influence on speech perception. Therefore research is focused on designing new advanced speech processing methods. In general, the speech processor decomposes the audio signal into different frequency bands and delivers a stimulation pattern to the implanted electrode determined by the speech processing strategy. The two main speech processing concepts are the CIS (Continuous Interleaved Sampling) and NofM strategies. NofM strategies such as Advanced Combinational Encoder (ACE) [4], separate speech signals into M subbands and derive envelope information from each band signal. N bands with the largest amplitude are then selected for stimulation (N out of M) in each time window. CIS could be considered as a special case of NofM with N=M, meaning that all bans are being selected for stimulation regardless of their envelope information.

Based on the general structure of the ACE strategy but incorporating a psychoacoustic masking model, a new approach has been designed in order to select the N bands in NofM strategies. The idea behind that was to neglect information that is inaudible to normal hearing persons and to concentrate only onto the signal components that are perceived by the normal hearing auditory system. It was anticipated to achieve improved speech recognition with this advanced speech coding strategy compared to he simple NofM type maxima selection of the ACE strategy. The new strategy was termed Psychoacoustic Advanced Combinational Encoder (PACE). The PACE strategy was evaluated in a pilot study conducted with eight cochlear implant recipients. Speech intelligibility tests, comparing the ACE and the PACE strategy, showed a superior speech performance for the PACE [4]. However, these results are generally subject to inter- and intra- subject variability. Results obtained from an objective method to measure speech intelligibility with both strategies would give more robustness to the study mentioned before.

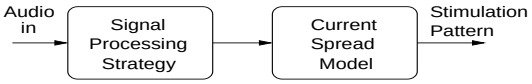

Figure 1: *Cochlear Implant Front-End.*

Automatic speech recognition systems based on neural networks and hidden markov models have been used to evaluate speech processors for cochlear implants [2], [3]. This paper also proposes a hidden markov model speech recognizer in order to compare the ACE and the PACE strategies. The speech recognizer uses as input the stimulation patterns obtained from a cochlear implant processor.

Section 2 presents the cochlear implant front-end. Section 3 outlines the structure of the hidden markov model speech recognizer . In section 4 the methods for testing both signal processing strategies are given. Finally, section 5 shows the results obtained and section 6 gives some conclusions.

2. The Cochlear Implant Front-End

Figure 1 presents the block diagram of the cochlear implant front-end. A speech signal is processed using a cochlear implant strategy. The output of this stage are electrical amplitudes. Afterwards a simple model of current spread has been used to estimate the stimulation pattern produced in the cochlea. The following subsections present each stage of the cochlear implant front-end in more detail.

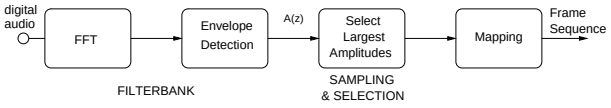

Figure 2: *ACE strategy block diagram.*

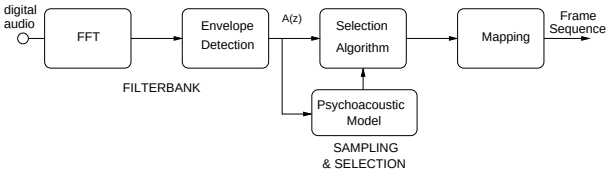

Figure 3: *PACE strategy block diagram.*

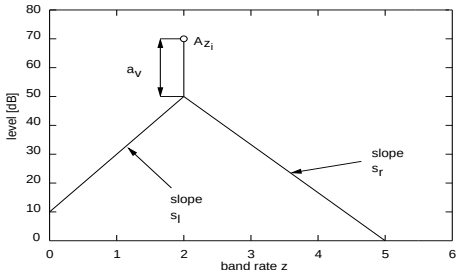

Figure 4: *Spreading Function.*

2.1. Signal Processing Strategy

The signal processing algorithms implemented are the Advanced Combination Encoder and the Psychoacoustic Advanced Combinational.

Both ACE (Figure 2) and PACE (Figure 3) are NofM-type strategies that can both be used with the Nucleus implant. In these strategies a digital signal sampled at 16 kHz is sent through a filterbank. The filterbank is implemented with an FFT (Fast Fourier Transform). The block update rate of the FFT is adapted to the rate of stimulation on a channel i.e. the Channel Stimulation Rate (CSR). The FFT is performed on windowed input blocks of 128 samples (8 ms at 16 kHz) of the audio signal using Hann window.

The uniformly-spaced FFT bins are combined by summing the powers to provide the required number of frequency bands. The bandwidths of these bands are approximately equal to the critical bands, where low- frequency bands have higher frequency resolution than high- frequency bands. The envelope in each spectral band $a(z)(z = 1, ..., M)$ is obtained as follows. The real part of the jth FFT bin is denoted with $x(j)$, and the imaginary part $y(j)$. The power of the bin is

$$r^2(j) = x^2(j) + y^2(j), j = 0, ..., L - 1. \quad (1)$$

The power of the envelope of a filter band z is calculated as a weighted sum of FFT bin powers

$$a^2(z) = \sum_{j=0}^{L/2} g_z(j) r^2(j), z = 1, ..., M, \quad (2)$$

where $g_z(j)$ are gains. The exact value of these gains can be obtained from [4].

The envelope of the filter bands z is $a(z)$.

In the ACE "sampling and selection" block, a subset of N (N $<$ M) filter bank envelopes $a(z_i)$ with the largest amplitude are selected for stimulation.

In the PACE "sampling and selection" block, a psychoacoustic-masking model is used to select the N bands. Consequently, the bands selected by this approach are not necessarily those with largest amplitudes (as is the case in the ACE strategy) but the ones that are, in terms of hearing perception, most important to the auditory system of normal-hearing people. The psychoacoustic masking model is configured by a so-called spreading function. This function models the masking effect of each band upon the others. The spreading function is defined using three parameters, the attenuation parameter a_v, the left slope s_l and the right slope r_l. In [4], speech tests with CI recipients were performed using two different spreading functions. These two configurations were termed PACE1 and PACE2, and they differed in the steepnes of the mentioned function. The PACE2 used a steeper function than the PACE2. More information on these two configurations of PACE can be found in [4].

Finally, the "mapping block" determines the current level based on the envelope magnitude and the channel characteristics. This is done by using the Loudness Growth Function (LGF), which is a logarithmically-shaped function that maps the acoustic envelope amplitude $a(z_i)$ to an electrical magnitude.

$$p(z_i) = \begin{cases} \frac{log\left(1+\rho\left(\frac{a(z_i)-s}{m-s}\right)\right)}{log(1+\rho)} & s \leq a(z_i) \leq m \\ 0 & a(z_i) < s \\ 1 & a(z_i) \geq m \end{cases} \quad (3)$$

The magnitude $p(z_i)$ is a fraction in the range 0 to 1 that represents the proportion of the output range (from the Threshold *T* to the Comfort level *C*). An input at the base-level *s* is mapped to an output at Threshold level, and no output is produced for an input of lower amplitude. The parameter *m* is the input level at which the output saturates; inputs at this level or above result in stimuli at Comfort level. The parameter ρ controls the steepness of the LGF [5]

Finally, the channels z_i are stimulated with levels:

$$l_i = T + (C - T)p_i \quad (4)$$

The set of l_i $(i = 1..N)$ form the frame sequence. A frame is generated at a rate defined by the channel stimulation rate. This parameter is fixed for each patient and its typical value is around 1000 Hz.

2.2. Current Spread Model

A simple model of current spread was used to estimate the electrical excitation along the auditory nerve with a cochlear implant. The current density was modeled with an exponential decay function in K sections along the cochlea.

$$E_m(k) = e^{\frac{-|X_{elec}(m) - X_{sect}(k)|}{\lambda}}, \quad m = 1...M, \; k = 1...K \quad (5)$$

$X_{sect}(k)$ represents the position in [mm] along the cochlea for the section k. $X(m)$ is the position along the cochlea for the electrode m and λ represents the degree of spread of excitation.

The number of sections K was set to 211. The sections were chosen to be spaced 0.1 Bark in frequency, this resolution was chosen to obtain the same number of sections as used by other auditory models known from the literature [10]. By selecting

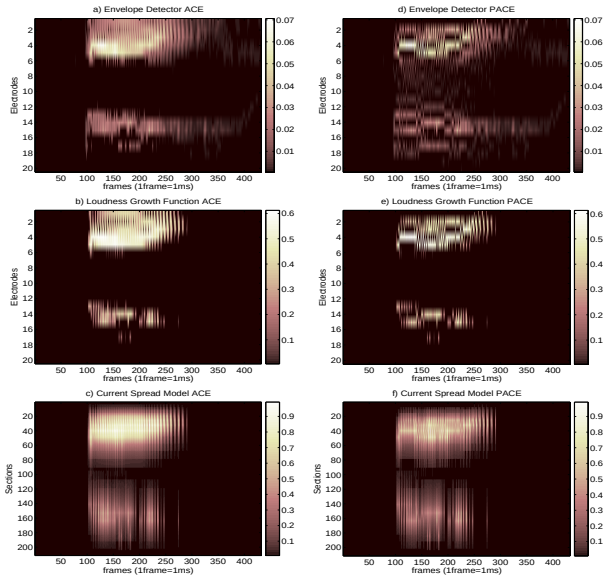

Figure 5: *a) Acoustic Envelope Response with ACE b) Loudness Growth Function Response with ACE. c) Electrical Field Response with ACE, d) Acoustic Envelope Response with PACE e) Loudness Growth Function Response with PACE,f) Electrical Field Response with PACE*

the same number of sections in both models, a direct comparison in the excitation patterns can be made. The position along the cochlea associated to each section was obtained by directly inverting equation 6 given by Greenwood [7] .

The number of electrodes M was set to 20. The position of each electrode $X_{elec}(m)$ was approximated by substituting the center frequencies of the ACE filterbank in equation **??**.

$$F = A(10^{aX} - k), \quad A = 165.4,\ a = 0.06 \qquad (6)$$

The value of λ was set to 1 mm, this value agrees with acoustic experiments comparing a cochlear implant vocoder and experiments with cochlear implant patients [6].

Finally, the excitation produced by each frame was obtained by adding the current spread produced by all electrodes stimulated in that frame.

Figure 5 presents the stimulation patterns of a speech token obtained at the different processing stages using the ACE and the PACE2 strategies. The token was a vowel 'a' uttered in english by a woman.

Figures 5a) and d) present the acoustic envelopes selected by the ACE and the PACE respectively. It can be observed that the spectrum obtained with the ACE is concentrated in two areas coinciding with the two first formants of the token 'a'. These two areas contain high energy as ACE selects the bands with largest amplitude. Figures 5 b) and e) present the spectrum obtained after converting the acoustic amplitudes into electrical amplitudes using the Loudness Growth Function for the ACE and the PACE. It can be observed that the pattern obtained with PACE has less components than the one determined by ACE. This is, because PACE selects more bands that are below the base level of the LGF function than ACE. Figure 5 c) and f) present the stimulation pattern obtained with ACE and PACE after modeling the current spread produced in the cochlea. It can be seen that with ACE the two formants have been smeared

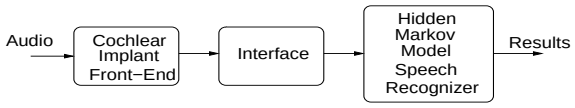

Figure 6: *Block Diagram of the Automatic Speech Recognizer using a Cochlear Implant Front-End.*

over the frequency. However, with PACE the formants can still be recognized.

3. Automatic Speech Recognition with a Cochlear Implant Front-End

Figure 6 presents the basic block diagram of the automatic speech recognition system using a cochlear implant front-end. The following subsections explain the speech recognizer and the interface between the cochlear implant front-end and the speech recognizer in detail.

3.1. Interfacing to the speech recognizer

As explained in previous sections, feature vectors of dimension K (K=211) are obtained at a rate determined by CSR. As described earlier, a typical value for the CSR is 1000 Hz. Therefore, the dimensionality of this data was reduced in order to make it more suitable for speech recognition with a Hidden-Markov-Model (HMM). For each section, the stimulation pattern was integrated every 10 ms. Afterwards, to further reduce the number of spectral features, a DCT was applied and the first 12 cepstral coefficients were stored. Finally, the feature vector was augmented by adding first- and second- order temporal derivatives. A similar dimensionality reduction was used in [5].

3.2. Hidden Markov Model Speech Recognizer

The recognition system was built with Cambridge's HTK Toolkit [9]. The experiments were performed using the TIMIT core database [8]. This database contains 192 sentences from 24 speakers. 576 sentences were used for training and 192 sentences were used for testing. The test set was not included in the training. The recognizer used a five-state HMM for each phoneme. Each state was modeled by a "Gaussian mixture". We did not use any kind of grammar, bi-gram or triphone model. Recognition of phones was therefore completely based on the actual feature vectors. Using Mel-frequency cepstral coefficient (MFCC) features, the system obtained 64.00% on phoneme recognition rate using the TIMIT core database.

4. Methods

The experiments consisted on comparing ACE and PACE strategies in a phoneme recognition task for different conditions. The different testing conditions were determined by varying the parameter N (number of selected bands N from the total number of filter bank bands M) in both signal processing strategies. All other parameters of the presented system were kept fixed.

The total number of bands M was set to 20, the CSR was set to 1000 Hz and the values that define the LGF function were set to ρ=20, s=4/256 and m=150/256. These values, are the standard values used by most of the users of the Nucleus-24 cochlear implant device. Finally, λ was set to 1 mm.

The PACE strategy was configured using two different spreading functions termed PACE1 and PACE2 as explained in

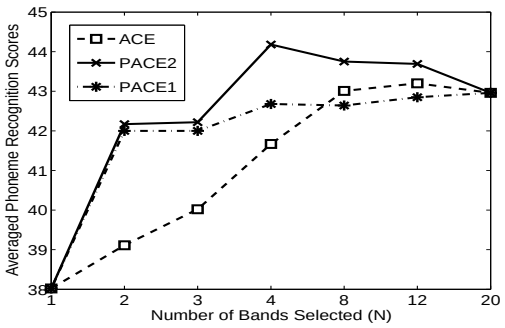

Figure 7: *Phoneme Recognition Results.*

section 2.1

5. Results

Figure 7 presents the recognition scores obtained by the automatic speech recognizer using ACE, PACE1 and PACE2 respectively.

For N=1 only one band is selected and for this condition there are no processing differences between ACE and PACE. Subsequently, the same scores were obtained in this condition. The same is true for N=20 where all the bands are selected in both ACE and PACE. Maximum performance is achieved by the PACE2 strategy, with N=4. Interestingly, PACE stimulating only 4 channels per cycle achieves better performance than ACE stimulating 20 channels per cycle.

These results are consistent with speech intelligibility tests obtained with cochlear implant patients at our center. Patient trials shown that PACE2 with N=4 performs better than ACE with N=8 [4].

6. Conclusions

This paper has presented an automatic speech recognizer that uses stimulation patterns coming from a cochlear implant front-end. The goal of the system was to obtain an objective measure to determine which signal processing strategy performs better in terms of speech intelligibility for cochlear implant patients. Using such an objective measure it is possible to achieve more robustness to results obtained from cochlear implant patients directly.

In a pilot experiment with the TIMIT core database two NofM signal processing strategies for cochlear implants were compared for different settings. This experiment has shown that the PACE strategy achieves better recognition performance than the ACE strategy even with PACE stimulating less bands per stimulation cycle compared to ACE. The maximum phoneme recognition score using PACE was 44.18% stimulating only 4 channels per stimulation cycle. The maximum score obtained by ACE was 43.01% and it was necessary to stimulate 12 channels per stimulation cycle. This experiment was consistent with speech intelligibility results obtained with cochlear implant recipients. As PACE can achieve at least the same results as ACE, but with significantly lower stimuli per cycle, the new strategy saves power and can lead to the design of smaller devices with less batteries.

The speech recognition system proposed is comparable to existing models of the normal hearing auditory system. Auditory models together with HMM back-ends have shown to obtain at least similar recognition scores to that of mel-frequency-coefficients (MFCCs) based hidden markov model speech recognizers. Therefore, it is anticipated that an auditory model representing the limitations of hearing with cochlear implants performs worse than the models representing a fully functioning auditory system. Future work will comprise the comparison of the above mentioned systems. Furthermore, the repetition of the experiments in adverse listening conditions and the influence of the parameters that configure the cochlear implant front-end in speech recognition will be investigated. It will be also evaluated if these results are consistent with the results obtained from cochlear implant patients. On our long way to closer mimic the neural excitation patterns within the cochlea and the ever increasing complexity of speech coding strategies, the approach of evaluating speech processing algorithms with the help of auditory models will gain more and more importance in the future.

7. References

[1] B. S. Wilson, C. C. Finley, D. T. Lawson, R. D. Wolford, D. K. Eddington, and W. M. Rabinowitz, “Better speech recognition with cochlear implants,”, Nature, vol. 352, no. 6332, pp. 236-238, 1991.

[2] C-H Chang, G. T. Anderson, P. C. Loizou, “A Neural Network Model for Optimizing Vowel Recognition by Cochlear Implant Listeners”, IEEE Transactions on Neural Systems and Rehabilitation Engineering, Vol. 9., No. 1, March 2001.

[3] J. Yao, Y-T Zhang, “The Application of Bionic Wavelet Transform to Speech Processing in Cochlear Implants Using Neural Network Simulations”, IEEE Transactions on Biomedical Engineering, Vol. 49, No. 11, 2002.

[4] W. Nogueira, A. Buechner, Th. Lenarz, B. Edler, “ A Psychoacoustic “NofM” -Type Speech Coding Strategy for Cochlear Implants”, EURASIP Journal on Applied Signal Processing, vol. 2005, no. 18, pp. 3044-3059, 2005.

[5] “Nucleus MATLAB Toolbox 2.11,” Software User Manual, N95246 Issue 1, Cochlear Corporation, October, 2001.

[5] M. H. Holmberg, D. Gelbart, U. Ramacher, W. Hemmert, “Automatic Speech Recognition with Neural Spike Trains”, IEEE Interspeech 2005, Lisbon, Portugal, September 4-8, 2006.

[6] J. Laneau, M. Moonen, J. Wouters, “Factors affecting the use of noise-band vocoders as acoustic models for pitch perception in cochlear implants”, J. Acoust. Soc. Am. 119(1), 2006.

[7] D. D. Greenwood, “A cochlear frequency-position function for several species-9 years later,” J. Acoust. Soc. Am. 87, 2592-2605, 1990.

[8] V. Zuc, S. Seneff, and J. Glass, “Speech database development at MIT: TIMIT and beyond”, Speech Communication, vol. 9, pp. 351-356, 1990.

[9] P. C. Woodland and S. J. Young, “The HTK tied-state continuous speech recognizer,”, in Proc. Eurospeech, 1993.

[10] F. Baumgarte, “Ein psychophysiologisches Gehoermodell zur Nachbildung Wahrnemungsschwellen fuer die Audiocodierung”, Dissertation, Universitaet Hannover, 2000.