A PHONE-ORIENTED SEGMENT MODEL OF THE HUMAN CORTEX A HYPOTHESIS

Harald Höge

Universität der Bundeswehr München
harald.hoege@t-online.de

Abstract: This paper brings together perceptive knowledge of phone classification, biological knowledge of the auditory system (cochlea nuclei, inferior colliculus, primary and secondary auditory cortex), and statistic knowledge of acoustic models. Due to the finding we postulate:

– for perception the auditory system generates features extracted from critical bands. These features are statistic independent for adjacent phones[1]
– the perception of phones occurs in the secondary cortex based on a segment model, where all statistic dependencies of the features are modeled correctly for whole utterances

1 Introduction

Performance of human speech recognition (HSR) is far superior compared to state of the art automatic speech recognition (ASR)-systems. Performance in ASR depends on the choice of the acoustic features, the acoustic model, and the language model. The statistic bindings between words as given by the language model are well studied [5] and need no further fundamental improvement. Progress in acoustic modeling and feature extraction is still the most challenging issue. The primary input for feature extraction is very similar in ASR and HSR. The features are power spectra sampled along a mel scale. In most ASR systems the spectra are transformed to **M**el-**F**requency **C**epstral **C**oefficients (MFCCs). In HSR feature extraction is performed along the auditory pathway [6] (see chapter 3). First processing steps are done in the cochlear nucleus and the inferior colliculus. Final processing is done in the primary and secondary auditory cortex. The acoustic modeling done in the primary and secondary cortex is still unknown. From the perceptive experiments done by Fletcher [3] we speculate on the functionality of the acoustic model used by humans (chapter 4). We hypothesize, that the human acoustic model is a segment model [1], where the segments are phones. The features used are statistic independent for adjacent phones. This segment model is able to model all statistic dependencies of the features of an utterance.

2 Fletchers Experiments on Phone Recognition

In the years 1918-1950 Harvey Fletcher [1] studied the performance of human speech recognition (HSR) at Bell Labs. Recognition experiments were done using nonsense consonant-vowel-consonant (CVC)[2] syllables not existing in the English language. He called those CVCs 'syllables with no context'. In the experiments 'listeners' had to recognize CVCs presented by 'talkers' over telephone channels under various noise and bandwidth conditions. The CVCs were presented in carrier sentences like '*The first group is'* and this carrier sentence is finalized by 3 different choices of CVCs as *na'v, po't'h, kŏb.* The experiments were performed by many listener crews, which were trained to this task. Their performance was monitored over years and crew performance was found to stabilize within a few months. For each CVC an 'observer' noted the count of correct recognized CVCs and the count for correct recognition of the phones constituting the CVCs. Fletcher called the resulting accuracies (recognition rates) of syllables and phones with no context syllable and phone '**articulation**'. Due to the use of crews over years we assume that the voices of the speakers were known to

[1] phones are acoustic realizations of phonemes
[2] also CV and VC experiments were performed

the listeners. Further we assume, that the listeners knew the carrier sentences. Thus they had to perform a **classification** task of CVCs, as the beginning and the end of the CVCs was known. Finally we assume that the listeners knew that they had to listen to CVCs and had to identify the initial consonant, the vowels and the final consonants. Thus the listeners perform speaker dependent classification, where the number and kind of initial consonants, vowels, final consonants is known. The number of phones on the different position within a CVCs is about 20. Summarizing, the listeners had to perform a speaker dependent classification task of nonsense CVC-syllables with known structure.

The experimental settings are defined by a set of parameter α. α is defined by bandwidth and signal noise ratio (SNR) of a telephone channel. For many settings α Fletcher determined a syllable articulation S(α), a consonant articulation c(α) and an vowel articulation v(α). As shown in figure 1 he found with high accuracy the relation:

) (1)

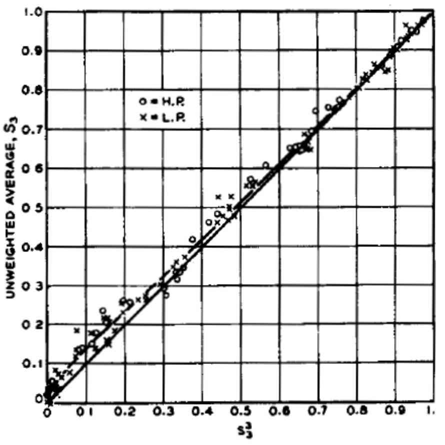

Figure 1 - relation between S and s^3 for different α realized by different high-pass (H.R.) and low-pass (L.R.) channels [3,9]

Fletcher developed a relation, which connects the phone articulation of high-pass and low-pass filtered speech with the phone articulation of unfiltered speech. This relation should be additive. This idea was in the spirit, that each band filtered speech with a bandwidth delivers a certain information . Adding the information from a low pass and a high pass filtered speech with the same cut-off frequency f_c should result in the information of the unfiltered speech. He found a transformation, which he called the **articulation index *A*,** with the property

(2)

should be one for unfiltered speech with no noise condition (maximal information). If the articulation is the same for speech for different bands, the articulation index should be the same.

For clean speech he found experimentally a density function *D*(*f*) tabulated in [4], which he called **articulation index density**, where the articulation index is given by

(3)

He found that A(s) can be described in the form:

(4)

and the inverse relation

. (5)

Combining (3) and (5) we get

leading to

(6)

From this result Fletcher concluded that speech is processed independently in frequency bands called articulation bands. These bands are equivalent to the critical bands [10] in the range of 700Hz-5000Hz. Zwicker called these bands 'Frequenzgruppen'. In chapter 3 we see that the processing in critical bands is implemented biologically.

He experimented with a filter bank of $K=20$ filters spaced in articulation bands with bandwidth . The partial phone articulations have the same value for all K under no noise condition. Extending (2) to the filter bank we get:

(7)

3 Biological Findings

In this chapter we shortly describe the stages of neuronal processing implemented by the human auditory system to perform phone recognition.

The auditor system can be divided in three subsystems: the cochlea located in the inner ear, the feature extraction system located in the medulla and in the midbrain, and the phone classification system located in the primary and secondary auditory cortex. The information transported from the inner ear to the primary auditory cortex in an **ascending** way is called the **auditory pathway** (see fig. 2). An overview of this pathway is given in [6, chapter 1]. There exist also descending feedback information flows interacting with the ascending information.

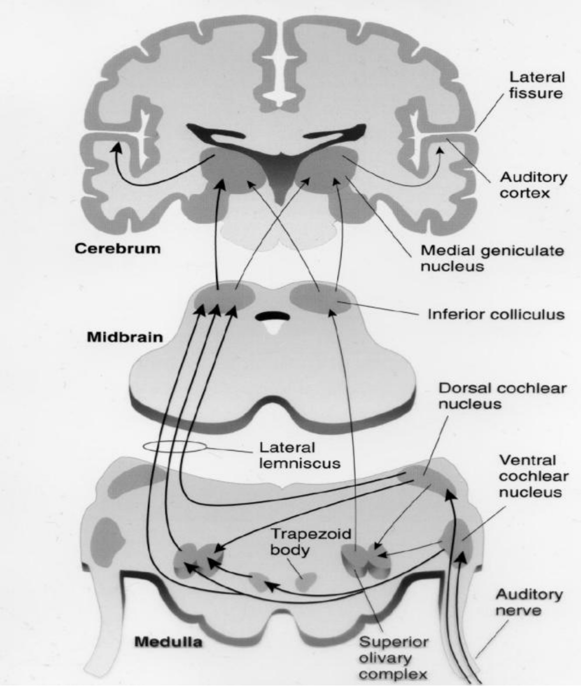

Figure 2 - auditory pathway

The information - transported by dendrites - is spatially organized by **streams** realized by lamina of neurons. Each stream handles a certain spectral-temporal aspect of the auditory signal. As stated in [6, chapter 1]: **'The particular contribution of most of these concurrent streams originating in the cochlear nucleus to the process of feature extraction, object**

recognition, and, ultimately, perception remain to be established' leads to the conclusion, that the neural based modeling of perceptive observation is stilled an open field for research.

The processing is performed by specific 'processors' called **nuclei**. They contain thousands of neurons dedicated to perform a specific task. There are special ensembles of dendrites, which form **perceptive fields**. The fields are spatial organized in a **tonotopic** structure, which reflects the spatial position of the hair cells in the cochlea. Further the perceptive fields are embedded in lamina, where each lamina handle a streams handling the information of hair cells of a critical band (see figure 3).

The auditory pathway with all his nuclei and its functions is similar to all mammals. The functionality of the human pathway is derived from measurement of neuronal activities of mammals, where the functionality of the brain of the cat has been investigated most deeply. The function of a neuron located in a nucleus is measured by the responses to acoustic stimuli. The output of neurons are electrical spikes. The number of spikes per second is the most relevant value to characterized the 'sensitivity' to a specific acoustic stimulus (e.g. a tone of specific frequency). The result of the measurements are **tuning curves,** where the sensitivity is measured by the values of a series of specific acoustic stimuli. The sensitivity can be interpreted as the probability that a certain stimulus is present.

Another source of information is the temporal position of individual spikes, which plays an important role in binaural processing. The delay of the spikes located in different ears is measured by specific neurons. These delays are the basis to build up a three-dimensional auditory field [6 chaper15].

The first neural processing of sound is performed in the hair cells sampling the vibrations of the basilar membrane [10]. Each hair cells realizes a band limited filter operating on the acoustic signal. The output of each filter is rectified and smoothed. The resulting information can be interpreted as short term power spectra *P(f,t)* - the auditory signal - sampled equally in the mel-domain. This information is transported via the acoustic nerve to the cochlea nucleus, were the auditory signal are processed in different streams delivering specific spectral and temporal aspects. A stream corresponds to the information derived from the hair cells handling a critical band. This is done separately for each ear. The monoaural streams are directed to the olivary complex, where binaural processing starts.[3] Next the streams are directed to lateral lemniscal nuclei, whose function is mostly unknown. The streams are processed further by the central nucleus of the inferior colliculus (ICC). In the ICC the one-dimensional tonotopic structure of the cochlea is transformed to a three dimensional structure. This structure is build up by lamina (see fig. 3)

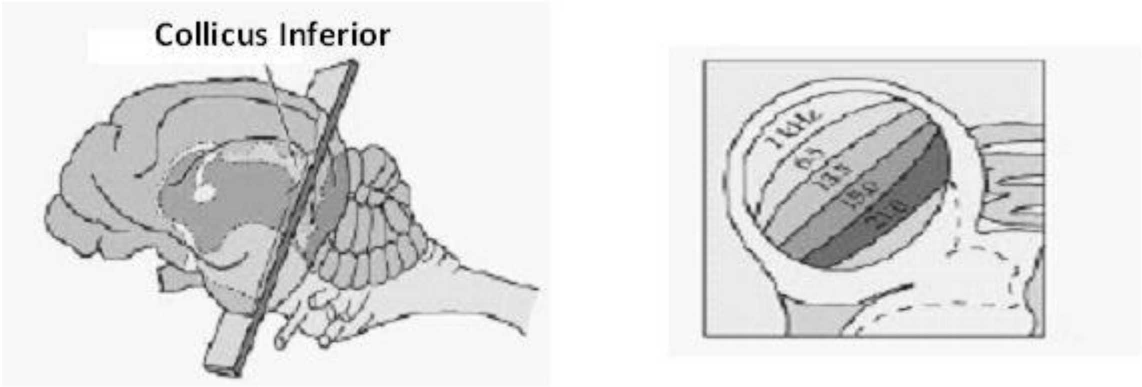

Figure 3 - lamina in the ICC ordered in critical bands for critical frequencies 7Khz, 6.5kHz,...

[3] binaural processing is located in the evolutional old medulla in contrast to binocular visual processing located . in the evolutional young cerebral cortex. Thus hearing is older than seeing.

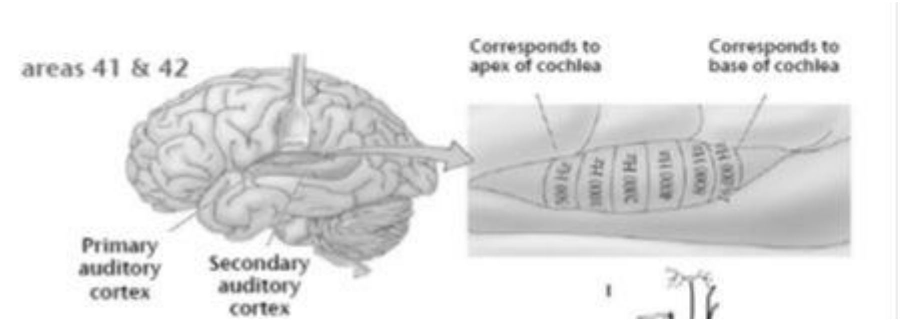
areas 41 & 42
Corresponds to apex of cochlea
Corresponds to base of cochlea
500 Hz
1000 Hz
2000 Hz
4000 Hz
8000 Hz
16,000 Hz
Primary auditory cortex
Secondary auditory cortex

Figure 4 - topographic map of area 41

Each lamina are build by disc-like structured complexes of neurons, which represent **s**pectral-**t**emporal **r**eceptive **f**ields (STRFs) [6, chapter12]. For each critical band the neurons operating on the STRFs analyses the amplitude and frequency modulation of that band in parallel. Additionally this processing is done in parallel in different areas of the laminas covering different spatial orientations. Those 'modulation features' are not statistic independent, as they are a more or less linear convolution of the power spectra of the cochlea. The modulation features are processed further in the primary auditory cortex (area 41). The topographic structure of that area is similar to that of the ICC.

We speculate, that in area 41 statistic independent features are generated. This is done in parallel for each critical band and for different spatial orientation of the acoustic field. These features are processed further in the secondary cortex (area 42), which performs associative processing for perceiving phones.

4 Segment Models

We regard a specific segment models [1], where the segments are phones. In contrast to HMMs, segment models take into account the statistic bindings between all feature vectors covering a segment. The weak point of the segment model is given by the assumption, that the features across segments are statistic independent. This assumption is in not fulfilled by the features used in ASR. As we postulate, that the features processed by the human brain are statistic independent across phones, this phone based model is able to describe all statistic dependencies correctly for whole utterances.

In section 4.1 we describe a phone based segment model. Similar segment models, which use sub-phones as segments, have been investigated in [13]. In section 4.2 we prove that statistic independences of features across phones lead to the finding of Fletcher as described in equation (1). Equation (7) is discussed in section 4.3.

4.1 The Phone Based Segment Model

The lowest error rate can be achieved, when Bayes decision rule is applied [2]. For recognizing whole utterances, this rule needs the conditional density function (cdf) . denotes the sequences of feature vectors realizing the utterance '*utt*'. *n* denotes the frame index. Because all the statistic bindings of the complete sequence must be treated, this cdf is too complex to be modeled as a single statistical unit. In most LVCSR systems, utterances are represented by sequences of small phonetic units (*PU*) taken from a set . Thus an utterance *utt* - consisting of *T* phonetic units - is given by the sequence . Each of the is realized by a sequence of feature vectors. We call a sequence assigned to a **chunk**. The number *l* of feature vectors building a chunk is called its **length *l***. The sum of the lengths *l*(*m*) must span the whole utterance given by the condition . The set of lengths *l*(*m*) defines a segmentation *S* of the frame indices . Thus the sequence of feature vectors is segmented in a sequence of chunks of length :

,

Based on these definition we formulate the Bayes decision rule for the recognized utterance as follows:

(8)

To evaluate (8) three distributions have to be approximated by models. is approximated by a language model. The relation between words and is given by a model of the length of chains of phonetic units. The acoustic model is approximated by

(9)

(9) reflects the assumption of segment models, that the chunks of adjacent segments are statistic independent.

4.2 Classification with Statistic Independent Chunks

Given 2 adjacent phones we denote by their accuracies and denote by the accuracies of the concatenated phones . Now we prove, that the relation

(10)

holds, if the chunks representing the phones are statistic independent and if they are context independent.

Using Bayes decision rule the accuracy *a* for a set *ph* of phones ph_i, i=1,...,N_{ph} is given by

If we connect two phones the accuracy of the connected phones is given by

Assuming that the chunks are statistic independent we get

leading to

(11)

Assuming that the phones are statistic independent, we get (10). Equation (11) can be extended easily to 3 phones leading to (1).

4.3 Classification with Statistic Independent Sub-Chunks

We split a chunk given by a phone into 'sub-chunks' . Each sub-chunk represent the feature vectors extracted from the output of neurons of a stream originated from the output of the hair cells of a critical band. We assume, that the sub-chunks are statistic independent. Now we regard Shannon's conditional [11] to find a relation to (7). In [9] Allen postulates also that their exists such a relation.

is the number of *bits* missing to recognize the phones without error. In the case the error rate *e* is zero. Shannon's entropy can be expressed by the mutual information and the phone entropy :

(12)

The mutual information is the number of bits gained from the sub-chunks . The phone entropy is the number of bits needed to recognize the phone without error. To the author knowledge there exist no direct relation between the error rate *e* and , but there exist bounds of the error rates depending on . For given and given number of phones an upper bound e_b for the error rate is given by the Fano bound [12]:

(13)

e_b increases monotonically with fulfilling :

(14)

If we assume, that the feature vectors are statistic independent the mutual entropy can be expressed by [11]

(15)

represents the information gained from the sub-chunk originating from of a single critical band. We define a partial entropy

(16)

which gives bounds for the error rate e_k defined in (7). Analog to (14) e_k is bounded by :

(17)

Due to (12), (15) the bound decreases, when more sub-chunks contribute to the recognition. From (17) equation (7) cannot be derived directly. For getting insight into the (7) we assume, that the bounds follow a similar relation as given by (7). We define

(18)

Using (13) we get for small errors

(19)

Using (15), (16),(17) and approximation (19) we get

leading to

(20)

(20) is not the relation (7), but nevertheless it hints to (7).

5 Conclusion

We have shown, that the assumption of statistic independent chunks generated in the primary auditory cortex is consistent with the perceptive experiments leading to (1). Using the framework of Shannon's conditional entropy we have shown further, that the assumption of statistic independent sub-chunks hints to the multiplicative character of (7).

The independent processing of sub-chunks in the ICC and in the primary auditory cortex leads to the assumption, that statistic independent sub-chunks are generated in the primary cortex.

Still two main questions are open. First: How transforms the auditory cortex the features generated in the ICC into statistic independent chunks and sub-chunks. Second: Is the Bayes decision rule realized in the cortex, and if yes, how is the search process (8) implemented to perform the segmentation and to include the language model.

A special issue is the fact, that within the ICC a spatial acoustic field is introduced. It is highly probable that for different spatial orientation in parallel features for different orientations are generated. Thus different phones originating from different spatial orientation could be perceived.

6 References

[1] Ostendorf, M., Digalakis, V., and Kimball, O.: From HMMs to segment models: a unified view of stochastic modeling for speech recognition. IEEE Trans. on Speech and Audio Proc., 4(5): 360-378, 1996

[2] Duda, R. O., Hart, P.E. and Stork, G.S.: Pattern Classification Second Edition. John Wiley & Sons, Weinheim 2001.

[3] Allen, J.B.: Harvey Fletcher 1884-1981, The ASA-Reprint of Speech and Hearing Communication. Allen, J.B.,(Ed.): Acoustical Society of America, New York, 1994

[4] Fletcher, H.: Speech and Hearing in Communications. New York 1953, pp. 283-285

[5] Brown, P.F. et al.: An Estimate of an Upper Bound for the Entropy of English. Computer Linguistics, Vol. 18: 1992, pp.31-40

[6] Winer, J. A., Schreiner, C. E.: The Inferior Colliculus. Springer Verlag, 2005

[7] Abdel-Hamid, O., Abdel-Rahman, M., Jiang, H., Penn, G.: Applying convolutional neural networks concepts to hyprid-NN-HMM model for speech recognition. Proc. ICASSP 2012 pp.4277-4280

[8] Toth, L.: Combining time- and frequency-domain convolution in convolutional neural network-based phone recognition. ICASSP 2014 pp. 190-194

[9] Allen, J.B.: How Do Humans Process and Recognize Speech?. IEEE Trans. on Speech and Audio Processing. 1994, pp. 567-577

[10] Zwicker, E., Feldtkeller, R.: Das Ohr als Nachrichtenempfänger. Hirzel Verlag 1967

[11] Cover, T.M., Thomas, J.A.: Elements of Information Theory. 2nd edition A Wiley-Interscience publication 2006

[12] Fano, R.M.: Transmission of Information: A Statistical Theory of Communications. MIT Press and John Wiley & Sons, Inc., New York, third edition: 1991

[13] Höge, H.: Impact of Correlated Features in Speech Recognition. . In Proc. ESSV: Aachen 2011