



On the Nature of the Features Generated in the Human Auditory Pathway for Phone Recognition

Harald Höge

Universität der Bundeswehr München, Munich, Germany

harald.hoege@unibw.de

Abstract

The features used by human phone recognition are generated along the auditory pathway by several transformations. In the first stages 'modulation features' are generated in lamina of neurons building a 3 dimensional strongly quantized structure, where each point of the structure corresponds to a feature component. One dimension concerns the different critical bands originated by bundles of inner hair cells. The second dimensions correspond to different locations of the acoustic field around the head. The third dimension is the modulation depth for different spectral and temporal modulation frequencies for each critical band and location. This structure is repeated in the auditory cortex, where a transformation to 'phone features' occurs. Due to insufficient neurophysiologic knowledge of these features, we conclude indirectly on their nature based on measurements of the accuracy of perceiving phones. We conclude that the phone features are statistic independent across adjacent phones. This implies that no acoustic context of neighbored phones is used to perceive phones. From these findings we speculate, that a phone oriented segment model is implemented in the auditory cortex. This model has the potential to model correctly the statistic dependencies of all phone features constituting an utterance.

Index Terms: human speech recognition, statistic independent features, segment model

1. Introduction

Performance of human speech recognition (HSR) is far superior compared to state of the art automatic speech recognition (ASR). 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 [4] and need no further fundamental improvement. Progress in acoustic modeling and feature extraction is still the most challenging issue. In ASR and HSR the primary input for feature extraction is some kind short term spectrum sampled linearly along the mel scale. The properties of the spectrum extracted in the cochlea is determined by the function of the inner hair-cells scanning the vibrations of the basilar membrane [10]. In HSR feature extraction is performed along the auditory pathway [5]. In recent years considerable progress was achieved in understanding the processing of auditory information along the auditory pathway. A short overview of the neuroanatomic and neurophysiological properties of the auditory pathway is given in chapter 2. First processing steps are done in the medulla and the midbrain by the cochlear nucleus and the central nucleus of the inferior colliculus. The resulting 'modulation features' describe the spectral and temporal modulation depth for each

critical band [6]. These features have a finer resolution of the spectrum as given by the MFCCs. Recently models of the modulation features have been used successfully in ASR [7,8]. Given the modulation features, phone features are extracted in the auditory cortex. To the author's knowledge the function of feature processing done in the auditory cortex is still unknown. To close this gap, in this paper perceptive relations on phone perception are used (see chapter 3). From these relations we conclude that the phone features are statistic independent for adjacent phones. The proof of these property is given in chapter 4. The statistic independence imply further that on the **acoustic** level no acoustic context of the neighboring phones is used. Given these findings we speculate on the kind of the acoustic model implemented in the auditory cortex. We hypothesize, that the acoustic model is a segment model [1], where the segments are phones. As described in chapter 4 this phone oriented segment model has the potential to model correctly all statistic dependencies of the phone features extracted from a stream of speech. Further no language model is involved, when perceiving phones in the auditory cortex. This is done in other areas of the cortex (Wernicke Area).

2. Neuroanatomical and Neurophysiological Findings

This chapter gives a short overview of the neuronal processing along the auditory pathway, which starts from the inner ear and ends in the auditory cortex. The pathway is build up by three subsystems: the cochlea located in the inner ear, the feature extraction system located in the medulla, in the midbrain and in the primary auditory cortex, and the phone perception system located in the secondary auditory cortex. Along the auditory pathway a complex processing of the information provided by the cochlea is performed. The information is coded by the number of spikes generated by a neuron per time unit, and by the temporal instant of the occurrence of a spike.

The first transformation of sound is performed by the inner hair cells sampling the vibrations of the basilar membrane [10]. The function of each hair cell can be described by a band pass filter, where the output of each filter is rectified and smoothed. The resulting information of the hair cells can be interpreted as a kind of short term spectrum $y(f,t)$ - the auditory signal - sampled equally on the mel-scale. For each ear the auditory signal is transported via the cochlea nerve to its cochlea nucleus and processed in different mono-aural streams. Each stream transports the information given by a bundle of inner hair cells corresponding to a critical band equally spaced on the mel-scale [10]. Each stream is processed by lamina of neurons handling different spectral and temporal aspects. In accordance to the tonotopic structure of the critical

10.21437/Interspeech.2015-361

bands on the cochlea, the lamina of neurons handling different streams are neuroanatomical ordered in the same way.

Additionally to the spectral information $y(f,t)$, spatial information of the acoustic field around the head is extracted by binaural processing. The delay of the spikes and the difference of spectral intensities originating by corresponding spikes of neurons from each ear are measured by specific neurons. In a first step the auditory signal is combined with the spatial information within the olivary complex. The most complex processing of the spectral and the spatial information is performed in the central nucleus of the inferior colliculus (ICC) [5]. The combination of the spatial information and the spectral information leads to a representation of the spatial short term spectrum $y(f, \vec{r}, t)$, where \vec{r} denotes the spatial locations. The information of $y(f, \vec{r}, t)$ is transported to lamina of disc-like structured complexes of dendrites, where each complex processes a specific aspect of $y(f, \vec{r}, t)$. Each complex of dendrites connected to a neuron with specific weights build a **spectral-temporal receptive field (STRF)**. Each STRF can be interpreted as a convolution kernel generating the depth of a specific aspect of the temporal and spectral modulation. The output of each neuron corresponds to a point in a strongly quantized three dimensional feature space. The three dimensions of the points are:

1. the different critical bands (quantization of mel-scale); in this dimension the tonotopic order is maintained neuroanatomical
2. the different location in the auditory space (quantization of spatial location)
3. the depth of modulation for different modulation frequencies implemented by a STRF for a specific critical band and for a specific location (quantization of the modulation frequencies).

This three dimensional structure of the ICC is repeated in the primary auditory cortex (see figure 1).

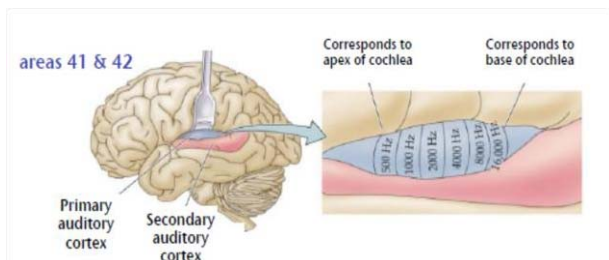


Figure 1: 3-dimensional structure of primary auditory cortex (blue, area 41); the tonotopic structure of the basilar membrane is maintained (in the picture on the right the horizontal direction marks the frequency direction)

Within the primary auditory cortex the modulation features are transformed to 'phone features'. As described in chapter 4, we conclude that the phone features are statistic independent for adjacent phones. How this property is implemented in the auditory cortex is an open question (see discussion) as stated in [5, 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'**.

3. Phone Perception

To the author's knowledge the most profound experiments to measure the performance on human phone perception was performed by Harvey Fletcher [3]. In the years 1918-1950 the experiments were done at Bell Labs using nonsense consonant-vowel-consonant (CVC)¹ syllables not existing in the English language. In the studies '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, kōb*. 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-articulation** and **phone-articulation**. In the following we use instead the wording syllable-accuracy and phone-accuracy. Due to the use of crews over years we assume that the voices of the speakers were known to the listeners. Further we assume, that the listeners knew the carrier sentences and could concentrate on the instant the unknown CVC starts. Thus the listeners had to perform a **classification** task of 'segmented' CVCs. Finally we assume that the listeners knew the list of potential consonants and vowels constituting the CVCs. Summarizing the listeners had to perform a speaker dependent classification task, 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. The phone-accuracy achieved under the best condition (no noise, no bandwidth constriction) was 98,5% (PER=1,5%). In ASR a benchmark for phone recognition is given by the performance achieved on the TIMIT database. This task is harder than Fletcher's set up, as about 39 phones have to distinguished. Further it is a speaker independent task and insertion/deletion errors can occur. In [7] a phone error rate of PER=15.7% is achieved on the TIMIT database.

The settings of Fletchers experiments is defined by a parameter α , which is defined by bandwidth and signal to noise ratio (SNR) of a telephone channel. For many settings α Fletcher determined a syllable accuracy $A(\alpha)$, a consonant accuracy $c(\alpha)$ and an vowel accuracy $v(\alpha)$. He found with high accuracy the relation:

$$A(\alpha) = a^3(\alpha); a^3(\alpha) \equiv c^2(\alpha) \cdot v(\alpha) \quad (1)$$

measured from high-pass and low-pass speech channels with various settings of α [3]. Equation (1) is the basis to conclude the existence of statistic independencies features for adjacent phones (see chapter 4).

Further Fletcher aimed to develop a relation, which characterize the 'information I contained in a given high-pass or low-pass channel. Later he extended I to a filter bank of band limited channels with no spectral overlap. For a given set of channels connected he called I the articulation index of those channels. I should be a function of the phone-accuracy a . The value of I of different channels should be equal, if the same accuracies are measured in each channel. The articulation index I of two connected channels should be the

¹ also CV and VC experiments were performed

sum of the articulation index of each single channel. Finally I should take the value one for unfiltered speech with no noise.

He found that $I(a)$ can be described by the relation:

$$I(a) = \frac{\log_{10} e}{\log_{10}(e_{min})};$$

$$e = 1 - a; e_{min} = 1 - a_{max}; a_{max} = 0.985 \quad (2)$$

where a_{max} is the highest accuracy for unfiltered speech without noise.

Given the phone accuracies a_I and a_{II} of two distinct band limited channels I and II with no overlap of the bands and given the accuracy a of the combined channel I&II, the additive property of $I(a)$ delivers

$$I(a) = I(a_I) + I(a_{II}) \quad (3)$$

Using (2) we get

$$\log_{10} e = \log_{10}(e_I) + \log_{10}(e_{II})$$

$$e = e_I e_{II} \quad (4)$$

From this result Fletcher concluded that speech is processed independently in bands which he called articulation bands. These bands are the smallest processing units. He experimented with a filter bank of $K=20$ filters (channels) spaced in articulation bands with bandwidth $b_k = [f_k, f_{k+1}]$, $k = 1, \dots, K$. The value of the bandwidths have been chosen in such a way, that the phone accuracy a_k of each channel has the same value for all K under no noise condition. For this filter bank (4) extends to the relation

$$e = \prod_{k=1}^K e_k \quad (5)$$

Fletcher proofed experimentally that (5) holds for various noise conditions. He developed a theory describing the influence of noise on phone accuracy. He showed that I_k of each articulation band has to be reduced by the amount of SNR present in this channel. This theory is motivated by the perceptive properties of critical bands called also 'Frequenzgruppen' [10]. As we know nowadays those bands are implemented neuroanatomic by lamina of neurons.

4. Segment Models

We regard a specific segment model from the class of segment models regarded in [1]. We restrict the segments to 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 of adjacent segments are statistic independent. This assumption is in not fulfilled by the features used in ASR. As the phone features processed by the auditory cortex are statistic independent for adjacent phones (see section 4.2), this phone 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 have been investigated in [11]. In section 4.2 we prove that statistical independences of features of adjacent phones and the use of the Bayes decision rule is consistent with (1). Finally we discuss in section 4.3 the nature of the features related to a single phone.

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) $p(\vec{X}_t | utt)$. \vec{X}_t denotes the sequences $\vec{X}_t = [X_1, \dots, X_n, \dots, X_t]$ of feature vectors X_n realizing the utterance 'utt'. n denotes the frame index. Because all the statistic bindings of the complete sequence \vec{X}_{utt} 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 $\vec{P}\vec{U}$ of small phonetic units (PU) taken from a set $\{PU_i, i = 1, \dots, N_{PU}\}$. We use as phonetic units context independent phones ph . Thus an utterance utt - consisting of T phones - is given by a sequence $ph_{i(m)}, m = 1, \dots, T$. Each of the phones $ph_{i(m)}$ is realized by a sequence $\vec{X}_{l(m)}^m$ of $l(m)$ feature vectors. We call a sequence $\vec{X}_{l(m)}^m$ assigned to phones $ph_{i(m)}$ a **chunk**. The number l of feature vectors building a chunk $\vec{X}_{l(m)}^m$ is called its **length l** . The sum of the lengths $l(m)$ must span the whole utterance given by the condition $t = \sum_{m=1}^T l(m)$. The set of lengths $l(m)$ defines a segmentation S of the frame indices $n = 1, \dots, t$. Consequently the sequence \vec{X}_t of feature vectors is segmented in a sequence of chunks $\vec{X}_{l(m)}^m, m = 1, \dots, T$ with lengths $l(1), \dots, l(T)$:

$$\vec{X}_t = [\vec{X}_{l(1)}^1, \dots, \vec{X}_{l(m)}^m, \dots, \vec{X}_{l(T)}^T], S \equiv [l(1), \dots, l(T)]$$

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

$$utt_{rec} = \underset{p}{\operatorname{argmax}} P(utt | \vec{X}_t) =$$

$$\underset{p}{\operatorname{argmax}}_{\vec{p}\vec{h}} \left(p(\vec{X}_t | \vec{p}\vec{h}) P(\vec{p}\vec{h}) \right)$$

$$= \underset{p}{\operatorname{argmax}}_{\vec{p}\vec{h}, S} \left((\sum_S p(\vec{X}_t | \vec{p}\vec{h}, S) P(S | \vec{p}\vec{h})) P(\vec{p}\vec{h}) \right) \quad (6)$$

To evaluate (6) three distributions $p(\vec{X}_t | \vec{p}\vec{h}, S), P(S | \vec{p}\vec{h}), P(\vec{p}\vec{h})$ have to be approximated by models. $P(\vec{p}\vec{h})$ is approximated by a language model. The relation $P(S | \vec{p}\vec{h})$ is given by a model of the duration of groups of phones. For the segment models [1] the acoustic model $p(\vec{X}_t | \vec{p}\vec{h}, S)$ is approximated by

$$\tilde{p}(\vec{X}_t | \vec{p}\vec{h}, S) = \prod_{m=1}^T p_{l(m)}(\vec{X}_{l(m)}^m | ph_{i(m)}, \theta_i^m);$$

$$S = [l(1), \dots, l(T)] \quad (7)$$

(7) reflects the assumption, that the chunks of adjacent phones are statistic independent. The pdfs $p_{l(m)}(\vec{X}_{l(m)}^m | ph_{i(m)}, \theta_i^m)$ are models of the distribution of the chunks of length $l(m)$ for phones of given length. As shown in [11] the pdfs could be represented by multimodal Gaussians with parameters θ_i^m .

4.2. The Impact of Statistic Independent Chunks

Given 2 adjacent phones ph^1, ph^2 we denote by $a(ph^1), a(ph^2)$ their accuracies and denote by $a(ph^1, ph^2)$ the accuracies of the concatenated phones ph^1, ph^2 . Now we prove, that the relation

$$a(ph^1, ph^2) = a(ph^1) a(ph^2) \quad (8)$$

holds, if the chunks \vec{X}^1, \vec{X}^2 representing the phones ph^1, ph^2 are statistic independent and if the phones are context independent.

Using Bayes decision rule the accuracy $a(ph)$ for a set ph of phones $ph_i, i=1, \dots, N_{ph}$ is given by

$$a(ph) = \int_{\vec{X}=-\infty}^{+\infty} \max_i P(ph_i | \vec{X}) p(\vec{X}) d\vec{X} \quad (9)$$

For simplification we do not regard the different dimensions of the chunks \vec{X} given by the length l of a chunk. For detailed treatment of (9) we have to introduce distributions $P_l(ph_i | \vec{X}_l)$ depending on l as done in (7). As shown in [11] (9) can be easily extended handling different lengths l .

If we connect two phones ph^1, ph^2 realized by chunks \vec{X}^1, \vec{X}^2 the accuracy of the connected phones is given by

$$a(ph^1, ph^2) =$$

$$\int_{\vec{X}^1, \vec{X}^2=-\infty}^{+\infty} \max_{i,j} P(ph_i^1, ph_j^2 | \vec{X}^1, \vec{X}^2) p(\vec{X}^1, \vec{X}^2) d\vec{X}^1, d\vec{X}^2$$

Assuming that the chunks \vec{X}_i, \vec{X}_j are statistically independent we get

$$P(ph_i^1, ph_j^2 | \vec{X}^1, \vec{X}^2) = \frac{p(\vec{X}^1, \vec{X}^2 | ph_i^1, ph_j^2) P(ph_i, ph_j)}{p(\vec{X}^1, \vec{X}^2)} = \frac{p(\vec{X}^1 | ph_i^1) p(\vec{X}^2 | ph_j^2) P(ph_i, ph_j)}{p(\vec{X}^1) p(\vec{X}^2)}$$

leading to

$$a(ph^1, ph^2) = \frac{P(ph^1, ph^2)}{P(ph^1)P(ph^2)} \cdot \left(\int_{\vec{X}^1, \vec{X}^2 = -\infty}^{+\infty} \max_{ij} \{ P(ph_i^1 | \vec{X}^1) P(ph_j^2 | \vec{X}^2) \} p(\vec{X}^1, \vec{X}^2) d\vec{X}^1 d\vec{X}^2 \right) = \frac{P(ph^1, ph^2)}{P(ph^1)P(ph^2)} \int_{\vec{X}^1 = -\infty}^{+\infty} \max_i \{ P(ph_i^1 | \vec{X}^1) \} p(\vec{X}^1) d\vec{X}^1 \cdot \int_{\vec{X}^2 = -\infty}^{+\infty} \max_j \{ P(ph_j^2 | \vec{X}^2) \} p(\vec{X}^2) d\vec{X}^2 = \frac{P(ph^1, ph^2)}{P(ph^1)P(ph^2)} a(ph^1) a(ph^2) \quad (10)$$

Assuming that the phones ph^1, ph^2 are context independent, we get (8). Equation (10) can be extended easily to 3 phones leading to (1) for nonsense CVCs.

4.3. Within Phone Features

As described in chapter 2 the modulation features and the phone features are processed independently for each critical band B_k . Thus each feature vector X can be split in K groups $X_{B_k}, k = 1, \dots, K$ of feature components, where X_{B_k} are the features processed in a critical band. In this way the chunks \vec{X} can be split in sub-chunks $\vec{X}_{B_k}, k = 1, \dots, K$, where each \vec{X}_{B_k} contains the feature groups X_{B_k} leading to $\vec{X} = [\vec{X}_{B_1}, \dots, \vec{X}_{B_K}]$. In chapter 3 the perception error e_k is introduced. e_k denotes the phone error from speech limited to the critical band B_k . Accordingly we can define accuracies $a_k(ph)$. These accuracies can be calculated with (9), where the chunks are reduced to sub-chunks. Due to (5) we have the relation

$$1 - a(ph) = \prod_{k=1}^K (1 - a_k(ph)) \quad (11)$$

From this relation we conclude, that there must be a specific statistic relation between the sub-chunks. In (3) the information $I(a_k)$ could be interpreted as the mutual information between the phones and the sub-chunks \vec{X}_{B_k} . Under the condition, that the sub-chunks are statistic independent, the mutual information [12] adds as given by (3). There is no evidence that the sub-chunks are statistic independent within a phone, nor that the mutual information is linked to the error rate as given by (5). Thus the statistic interpretation of (5) is still an open issue.

5. Discussion

In this chapter we discuss further the nature of the features generated by the articulatory pathway, but in a speculative way. The neurophysiologic properties of the modulation features are understood quite well. It should be emphasized that those features are extracted in the same way for all mammals. Thus they are not tuned to perceive phones, but to perceive the surrounding sounds carrying important cues for living. The neurophysiologic nature of the phone features are mostly unknown. Due to property of plasticity of the neurons in the cortex, the phone features can be learned and thus tuned to perceive phones of a given language. The process of learning lead to the properties described by (1) and (5), which

corresponds to statistic properties of the phone features as described in section 4.2 and 4.3. Thus there must be a relation between the learning process and the statistic properties of the phone features. This is an open issue for research. Further we assume, that the phone features are generated in two stages. In the first stage phone features related to each critical band are generated. This view is motivated by the neuroanatomic structure of the primary auditory cortex. Those partial phone features have already the properties (1), (5). In a second stage the partial phone features are combined to generate the phone features. It is highly probable that those features are generated for all locations around the head in parallel. Thus human can listen to different spatial voices in parallel (party effect). As the access to the Wernicke area seems to be limited, the parallel understanding of different voices is limited. Further it is unclear, what acoustic information besides the categorical phonetic aspect is contained in the phone features, which activate the perception of words in the Wernicke area.

6. Conclusions

We have combined biological, perceptive and statistical knowledge to get inside in the nature of the features used for phone perception. In this paper we showed that the gap in neurophysiologic knowledge can be closed partly by combining knowledge from different scientific fields. Applying this method we showed that statistic independent chunks are generated in the primary auditory cortex and that Bayes decision rule is applied for classification. Still two main questions are open. First: How transforms the auditory cortex the modulation features generated in the ICC into statistic independent chunks. Second: 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 spatial information is included in the feature given by the neuroanatomic structure of the STRFs. Using this aspect features could be developed leading to more robust ASR.

7. References

- [1] M. Ostendorf, V. Digalakis and O. Kimball, "From HMMs to segment models: a unified view of stochastic modeling for speech recognition," *IEEE Trans. on Speech and Audio Proc.*, 4(5), pp. 360-378, 1996
- [2] R. O. Duda, P.E. Hart and G.S. Stork, "Pattern Classification Second Edition," *John Wiley & Sons*, Weinheim 2001.
- [3] H. Fletcher, and R.H. Galt, "The perception of Speech and Its Relation to Telephony," *The Journal of the Acoustic Society of America*, Vol. 22, number 2, pp. 89-151, 1950
- [4] P.F. Brown et al., "An Estimate of an Upper Bound for the Entropy of English," *Computer Linguistics*, Vol. 18: pp.31-40, 1992
- [5] J. A. Winer, C. E. Schreiner, "The Inferior Colliculus," Springer Verlag, 2005
- [6] P.X. Joris, C.E. Schreiner, A. Rees, "Neural Processing of Amplitude-Modulated Sounds," *Physiol Rev* 84, pp.541-577, 2004
- [7] L. Toth, "Combining time- and frequency-domain convolution in convolutional neural network-based phone recognition," *ICASSP 2014*, pp. 190-194, 2014
- [8] S.K.Nemala, K.Patil, and M. Elhilali, "A Multistream Feature Framework Based on Bandpass Modulation Filtering for Robust Speech Recognition," *IEEE Transactions on Audio and Language Processing*, Vol.21, pp. 416-426, 2013
- [9] J.B.Allen, "How Do Humans Process and Recognize Speech?," *IEEE Trans. on Speech and Audio Processing*, pp. 567-577, 1994

- [10] E. Zwicker, H. Fastl, "Psychoacoustics," *Berlin: Springer* 1999
- [11] H. Höge, " The Use of Conditional Gaussians for Hidden
Chunk Models, " *Proc. ESSV Cottbus Germany*, 2012
- [12] T.M. Cover, J.A. Thomas, "Elements of Information Theory,"
second edition, Wiley 2006