BUILDING SEGMENTS WITH CHUNKS

Harald Höge

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

Abstract: In HMM-technology state tying is an approved method to achieve reliable estimation of model parameters. This method is based on clustering sub-phonetic units, derived from context dependent phones. Due to the theory of HMMs the clustering algorithm assumes that the feature vectors are statistic independent within a cluster. We derive the sub-phonetic units from tri-phones and call the resulting clusters 'HMM-segments'. In this paper we develop a new clustering algorithm, which is based on the theory of **H**idden **C**hunk **M**odels (HCM). The algorithm takes into account the statistic dependencies of the feature vectors realizing the sub-phonetic units. We call the resulting segments 'HCM-segments'. Both kinds of segments are modeled with HCMs. With these acoustic models we build two classification systems for context independent phonemes. Using a large Spanish speech database we compare the phoneme error rates achieved with the two kinds of segments. The HCM-segments showed higher performance.

1 Introduction

In speech recognition research is focused on achieving low error rates for words or utterances. The error rate is influenced mostly by the features and by the acoustic model used. We focus on improving acoustic models by taking into account the statistic dependencies of features. The lowest error rate is achieved, when the principle of maximum likelihood classification [1] is applied. But when recognizing whole utterances for this approach the conditional density function (cdf) $p(\vec{X}|utterance)$ must be known ($\vec{X}$ denotes the sequence of feature vectors realizing the utterance). As the statistic bindings of the complete sequence $\vec{X}$ must be treated, this cdf is too complex to be modeled as a statistical unit. In most LVCSR system, utterances are represented by sequences $\overrightarrow{PU}$ of phonetic units (*PU*). It is assumed that the sequences of feature vectors representing a different $PU \in \overrightarrow{PU}$ are statistic independent. Further it is assumed that a suited parametric model $p(\vec{X}|PU_i, \Theta_i)$; $i = 1, \ldots, N_{PU}$ for the distributions $p(\vec{X}|PU_i)$ can be provided. The parameters Θ_i have to be estimated reliably from a suited speech database. This approach leads to the following approximation for $p(\vec{X}|utterance)$:

$$\tilde{p}(\vec{X}|utterance) \equiv p(\vec{X}|\overrightarrow{PU}) \equiv \prod_n p(\vec{X}^n|PU_{i(n)}, \Theta_i) \quad (1)$$

(In the following $\tilde{p}(\cdot)$ denotes an approximation of a function $p(\cdot)$. The warping function *i(n)* maps the time index *n* of the sequences $\vec{X}$ to the index *i* of the corresponding $PU \in \overrightarrow{PU}$). To implement (1) three questions have to be answered

- which PU should be chosen
- how should the PU be modeled
- how should the parameters Θ be estimated

In [3] segmental models are proposed, which treat the PUs as statistical units i.e. the distribution $p(\vec{X}|PU_i, \Theta_i)$ must model all the statistical bindings of the sequence of feature vectors $\vec{X}$. If the PUs are large enough i.e. cover large portions of the utterance, minimum error rates should be achievable, when the maximum likelihood classification method is applied. To the author's knowledge, no segmental model for large PUs has been found, leading to competitive error rates. Many HMM based systems as [2] use as sub-phonetic units

PUs derived from context dependent phones (e.g. tri-phones) . Each phone is split into 3 parts, where for each phone each part is a specific **'sound' *S*.** The set of sounds is very large and most sounds have very low probabilities of appearance. Even when using very large speech databases, it turns out, that for many sounds too few samples are available to estimate reliably the parameters of $p(\vec{X}|S,\Theta)$. To solve this estimation problem, sounds are clustered. In HMM terminology this clustering scheme is called 'state-tying', which was introduced in [4]. Each cluster defines an PU. We call those PUs '**segments**' denoted by $Q_i, i = 1, \dots N_Q$. The theory of HMMs assumes that the feature vectors within a segment are i.i.d. distributed (i.e. the feature vectors are statistic **i**ndependent, and within a segment all feature vectors are **i**dentical distributed) leading to the segment model

$$p(\vec{X}_l|Q_i,\Theta_i) = \prod_{v=1}^{l} p(X_v|\,Q_i,\Theta_i) \qquad (2)$$

This model is not a segmental model [3], as the feature vectors X_v are statistic independent and the distributions $p(X_v|Q_i,\Theta_i)$ called the emission probabilities are identical for all X_v.

Now we discuss the estimation process of the parameters Θ_i of (2) in the context of clustering. For state of the art HMM-based speech recognition systems the MCE method is used for estimating the parameters Θ_i. This method delivers the minimal error rate for words or utterance given the wrong model assumption (1,2). The MCE method is a global optimization procedure, which minimize the error rate across words or utterances. To the author's knowledge, there exists no such a global optimization scheme for clustering sounds to segments. Instead the optimal clusters are generated by clustering 'similar' sounds using the principle of minimal increase of entropy [4]. We call the resulting segments 'HMM-segments'. In section 2.2 we show, that this approach is linked to achieve minimal error rate to classify HMM-segments with a single feature vector. It is well known, that parameters, estimated to deliver minimal error rates on segments, do not lead to lowest error rates for classification of larger phonetic units as words.

Recently we investigated segmental models called '**Hidden Chunk Model (HCM)**' [5, 6, 7], which realizes a specific segmental model [3]. In this approach the HCMs model HMM-segments. For generating the HMM-segments the clustering is performed by a top down approach using a CART [10]. In contrast to the segment model (2) the HCMs take into account the statistic dependencies of the feature vectors within the HMM-segments.

In this paper we evaluate a clustering scheme, which lead to different segments called HCM-segments. Similar to the HMM approach, we use for clustering the sounds S derived from tri-phones and we use the principle of minimal increase of entropy as in [4], but we take into account the statistic dependencies of the feature vectors within the sounds. Thus the clustering algorithm is consistent with the HCM-approach. In order to see the advantage of using the HCM-segment instead of the HMM-segments, we make phone classification experiments comparing phoneme error rates achieved for both kinds of segments. The paper is organized as follows. In chapter 2 we develop the principles of clustering using Shannon's conditional entropy [8] and describe the resulting algorithms for constructing HMM- and HCM-segments. Chapter 3 describes the experiments made.

2 Clustering

The section 2.1 defines more precisely the relation between sounds composing a segment and sets of feature vectors realizing segments. Section 2.2 concerns the basic principles for clustering based on Shannon's conditional entropy. Section 2.3 and 2.4 describes the algorithms to generate HMM-segments and HCM-segments.

2.1 Sounds, Segments and Related Feature Sets

We use as context dependent phones tri-phones Ph_{jc}. The context c is defined by the right and left phone of the 'central' phone Ph_j. Each tri-phone is composed by 3 sub-phonetic sounds $S_{jc}^{p.}; p = 1,2,3$, which can be interpreted as the onset, middle and offset of a phone. The sounds are clustered to segments.

The HMM-segments are notated by Q_i. They are defined by a mapping function $i = f(p,j,c)$, which maps all sounds $S_{jc}^{p.}$ to a segment Q_i according to

$$Q_i = \left\{ S_{jc}^{p.}, \forall p,j,c : \; i = f(p,j,c) \right\} \tag{3}$$

Thus the index i denote the phonetic properties p,j,c of all sounds S mapped to the HMM-segment Q_i.

The HCM-segments are denoted by $Q_{i.}^{l}$. They are defined by an extended mapping function $i = f(p,j,c,l)$ depending additionally on the length l of the chunks realizing the segment:

$$Q_{i.}^{l} = \left\{ S_{jc}^{p.}, \forall p,j,c,l : \; i = f(p,j,c,l) \right\} \tag{4}$$

The index i denotes the phonetic properties of the HCM-segments determined by the sounds clustered. Due to the dependency of i from l, the same index i can be related to different phonetic properties p,j,c. Further we restrict the HCM-segments $Q_{i.}^{l}$ to a maximal value of $l=m_0$. All sounds $S_{jc}^{p.}$ realized by chunks with $l > m_0$ are mapped to $Q_{i.}^{m_0}$.

We assume that the speech database is labeled; i.e. for each sound $S_{jc}^{p.}$ the corresponding sequence $\vec{X}_l$ is known ('aligned'). Applying (3) and (4) the alignment of each segment to a sequences $\vec{X}_l$ is given. We call sequences $\vec{X}_l$ aligned to segments '**chunks**'. Now we define sets of feature vectors and sets of chunks. The set of all feature vectors found in a speech database, which realize a HMM-segment Q_i, is notated by

$$C_i = \{X^n \in Q_i\} \tag{5}$$

(n denotes the time/frame-index) All chunks $\vec{X}_{l.}^{n}$ found in the speech database, which realize a HCM-segment $Q_{i.}^{l}$ is denoted by

$$C_{i.}^{l} = \left\{\vec{X}_{l.}^{n} \in Q_{i.}^{l}\right\} \tag{6}$$

The sets C_i are used to estimate the parameters Θ_i of the emission probabilities $p(X|Q_i,\Theta_i)$ of the feature vectors X. The sets $C_{i.}^{l}$ are used to estimate the parameters $\Theta_{i.}^{l}$ of the distributions $p\left(\vec{X}_l \middle| Q_{i.}^{l}, \Theta_{i.}^{l}\right)$. As described at the end of the following section, the clustering is done separately for each central phonemes $Ph_j j=1 \ldots N_{Ph}$ using sets:

$$\left.\begin{aligned} C_{j.}^{l} \equiv \left\{\vec{X}_{l.}^{n} \in Q_{i.}^{l};\; \forall p,c: i = f(p,j,c,l)\,; j = 1,\ldots,N_{Ph}; l = 1,\ldots,m_0\right\} \\ C^{l} \equiv \bigcup_{j=1.}^{N_{Ph.}} C_{j.}^{l} \end{aligned}\right\} \tag{7}$$

C^l is the set of all chunks $\vec{X}_{l.}^{n}$ of length l.

2.2 Basic Principles of Clustering

In this subsection we discuss the clustering approach in the framework of HCMs. This approach is very similar to the clustering approach in the framework of HMMs. In subsection 2.2.3 we sketch the HMM-approach.

The number of sounds S derived from tri-phones is notated by N_S. We want to cluster these sounds into $N_{Q.}^{m_o}$ segments under the condition $N_{Q.}^{m_o} \ll N_S$. $N_{Q.}^{m_o}$ has to be chosen in such a way that each set $C_{i.}^{l} = \left\{\vec{X}_{l.}^{n} \in Q_{i.}^{l}\right\}$ (6) contains enough chunks to estimate the parameters $\Theta_{i.}^{l}$

of the distribution $p\left(\vec{X}\middle|Q_i^l,\Theta_i^l\right)$ reliably[1]. Now the basic approach for clustering is formulated as follows: given the number $N_Q^{m_o}$ of target segments we want to construct $N_Q^{m_o}$ clusters of sounds *S*, which minimize the segment error rate (*SER*). This MCE-approach is motivated by the segmental approach which lead to low error rates for words or utterance if large PUs would be used (see discussion in the introduction). To the authors knowledge there exist no efficient clustering algorithm based on the MCE-criterion. Further our segments are very short *PU*s. Despite of these shortcomings, we stick to this basic approach but we use an approximation to minimize the *SER*. We use the fact that the *SER* is linked to Shannon's conditional entropy [8] $H(Q|Z)$ (Z denotes chunks $\vec{X}_l$ for all *l*). Within certain bounds [9] the *SER* decreases with decreasing $H(Q|Z)$. Thus minimizing $H(Q|Z)$ is linked to minimizing the *SER* within certain bounds. $H(Q|Z)$ is defined by

$$\left.\begin{aligned} H(Q|Z) &= \textstyle\sum_{l=1}^{\infty} P(C^l)H(Q^l|\vec{X}_l)\,;\; H\left(Q^l\middle|\vec{X}_l\right) = H(Q^l) - H(C^l) + H(\vec{X}_l|Q^l) \\ &= -\textstyle\sum_{l=1}^{\infty} P(C^l)H(C^l) + \sum_{l=1}^{\infty} P(C^l)\,(H(Q^l) + H(\vec{X}_l|Q^l)) \\ H(Q^l) &\equiv -\textstyle\sum_i P(Q_i^l)\; log\; P(Q_i^l)\;,\; H\left(\vec{X}_l\middle|Q^l\right) \equiv -\sum_i P(Q_i^l)\int p\left(\vec{X}_l\middle|Q_i^l\right) log\; p\left(\vec{X}_l\middle|Q_i^l\right) d\vec{X}_l \end{aligned}\right\} \quad (8)$$

To determine $H(Q^l)$ the values of the discrete distribution $P(Q_i^l)$ are estimated from the size (counts) of the sets C_i^l (6). For estimating $H\left(\vec{X}_l\middle|Q\right)$ we approximate $p\left(\vec{X}_l\middle|Q_i^l\right)$ by a mono-modal Gaussian:

$$p\left(\vec{X}_l\middle|Q_i^l\right) \approx p_l\left(\vec{X}_l\middle|Q_i^l,\Theta_i^l\right) = \mathrm{N}(\vec{X}_l;\vec{\mu}_{il},\vec{V}_{il}),\Theta_i^l = \{\vec{\mu}_{il},\vec{V}_{il}\} \quad (9)$$

This approach is very crude, but few samples of chunks are needed to estimate the parameters Θ_i^l from the sets C_i^l. Further leads this approach to an computational feasible algorithm (see (12)).Now we have to find $N_Q^{m_o}$ clusters Q_i^l from the N_S sounds, which minimizes $H(Q|Z)$. The term $\sum_{l=1}^{\infty} P(C^l)H(C^l)$ is independent from the choice of the segments. Thus the minimum of $H(Q|Z)$ is given by the minimum of $R(Q)$:

$$min_Q R(Q);\; R(Q) \equiv \textstyle\sum_{l=1}^{m_0} P(C^l)\left(H(Q^l) + H(\vec{X}_l|Q^l)\right);\; \left|\{Q_i^l\}\right| = N_Q^{m_o}$$

Using (8, 9) $R(Q)$ is approximated by the log-likelihood function $L(Q)$ using the entropy $H\left(\mathrm{N}(\vec{X}_l;\vec{\mu}_{ikl},\vec{V}_{il})\right)$ of a mono-modal Gaussian:

$$\left.\begin{aligned} R(Q) \approx L(Q) &\equiv \textstyle\sum_{l=1}^{m_0} L(Q^l)\,;\; L(Q^l) \equiv P(C^l)\left(H(Q^l) + \sum_{i=1}^{N_Q^l} H(\mathrm{N}(\vec{X}_l;\vec{\mu}_{ikl},\vec{V}_{il}))\right) \\ H\left(\mathrm{N}(\vec{X}_l;\vec{\mu}_{ikl},\vec{V}_{il})\right) &= \left(\tfrac{lD}{2}\ln(2\pi e) + \tfrac{1}{2}ln\left|\vec{V}_{il}\right|\right)\;; H\left(Q^l\right) = -\textstyle\sum_{i=1}^{N_Q^l} P(Q_i^l)\; log\; P(Q_i^l) \end{aligned}\right\} \quad (10)$$

(*D* denotes the dimension of a feature vector *X*. N_Q^l denotes the number of segments Q_i^l with given *l*). We minimize $L(Q)$ separately for each component $L(Q^l)$:

$$min_{Q_i^l} L\left(Q^l\right) = min_{Q_i^l}\left(\textstyle\sum_{i=1}^{N_Q^l} P(Q_i^l)\left\{log\; P(Q_i^l) + \left(\tfrac{lD}{2}\ln(2\pi e) + \tfrac{1}{2}ln\left|\vec{V}_{il}\right|\right)\right\}\right); l = 1,\dots,m_0 \quad (11)$$

The *min* operator (11) has exponential complexity with increasing number N_Q^l. We use as approximation a top down clustering method for generating HMM-segments [10] and a bottom up method for generating HCM-segments (see section 2.4). Applying these methods the change of the likelihood $L(Q^l)$ has to be determined, when two segments clusters $Q_i^l, Q_{i'}^l$ are merged or split. The change of $L(Q^l)$ is given by $\Delta\left(Q_{ii'}^l\right) \equiv L\left(Q_{ii'}^l\right) - L(Q^l)$, where $L\left(Q_{ii'}^l\right)$ denotes the log-likelihood function when in (11) the two segments $Q_i^l, Q_{i'}^l$ are substituted by the segment $Q_{ii'}^l = Q_i^l \cup Q_{i'}^l$. Using (10) we get

[1] To the authors knowledge there exist no criterion for reliability. Large differences in error rates on the training database and the test database (mismatch) hint to poor estimation

$$\Delta\left(Q_{ii'}^{l}\right) = P\left(Q_{ii'}^{l}\right)\left\{log\ P(Q_{ii'}^{l}) + \frac{1}{2}ln\left|\vec{V}_{ii'l}\right|\right\} - P\left(Q_{i}^{l}\right)\left\{log\ P(Q_{i}^{l}) + \frac{1}{2}ln\left|\vec{V}_{il}\right|\right\} - P\left(Q_{i'}^{l}\right)\left\{log\ P(Q_{i}^{l}) + \frac{1}{2}ln\left|\vec{V}_{i'l}\right|\right\} \quad (12)$$

According to [11, page 544] $\Delta\left(Q_{ii'}^{l}\right)$ is always positive. We modify further (11) by introducing some restrictions, which sounds can be clustered. These restrictions are applied for HMM-segments and HCM-segments: We do not merge sounds belonging to different central phonemes PH_j. This is done for two reasons. First we want to speed up computation as done in [4]; second we want to avoid that sequences of segments representing a phone of a central phoneme are equal for different central phones. Further we have to consider that each central phoneme Ph_j is represented by sequences of sounds $\{S_{jc}^{p}\}$; p = 1,2,3. We do not cluster sounds belonging to different position p. This restriction eases the search algorithm needed for phoneme recognition, when skipping of segments is allowed (for some languages as for French this happens often in the QUAERO databases). Given all these restrictions, the minimum of (11) has to be evaluated only for restricted sets of sounds realized by the sets C_j^l (see (7)). In this case the total number N_j^l of sounds realized by C_j^l is no more determined by the minimization procedure (11). We use following heuristic rule:

$$N_j^l = P\left(C_j^l\right)N_Q^{m_0};\ N_Q^{m_0} = \sum_{j=1}^{N_{Ph}}\sum_{l=1}^{m_0} N_j^l \quad (13)$$

2.3 HMM-Clustering

For generating HMMs we have to estimate the parameters of the emission probabilities $p(X|Q_i, \Theta_i)$ (see (2)). In analogy to (9) we use a monomodal Gaussian for clustering:

$$p(X|Q_i, \Theta_i) = \mathrm{N}(\mathrm{X}; \mu_\mathrm{i}, \mathrm{V_i}), \Theta_i = \{\mu_\mathrm{i}, \mathrm{V_i}\} \quad (14)$$

In analogy to (10, 11) the log-likelihood function $L(Q)$ is given by

$$\left.\begin{aligned} L(Q) &\equiv H(Q) + H\left(\mathrm{N}(\mathrm{X}; \mu_\mathrm{i}, \mathrm{V_i})\right);\ H\left(\mathrm{N}(\mathrm{X}; \mu_\mathrm{i}, \mathrm{V_i})\right) = \left(\frac{D}{2}\ln(2\pi e) + \frac{1}{2}ln|V_i|\right) \\ H(Q) &\equiv -\sum_i P(Q_i)\log P(Q_i)\ ;\ H(X|Q) \equiv -\sum_i P(Q_i)\int p(X|Q_i) log\ p(X|Q_i)dX \end{aligned}\right\} \quad (15)$$

and we have to minimize

$$min_{Q_i}L(Q) = min_{Q_i}\left(\sum_{i=1}^{N_Q} P(Q_i)\left\{log\ P(Q_i) + \left(\frac{D}{2}\ln(2\pi e) + \frac{1}{2}ln|V_i|\right)\right\}\right) \quad (16)$$

To ease the calculation of the covariance matrices V_i we assume that they are diagonal with tied diagonal elements σ_i^2 leading to $|V_i| = \sigma_i^{2D}$ This approach is motivated from the use of LDA-transformed feature vectors, which produces globally a covariance matrix with this structure. Equivalent to (12) we get

$$\Delta(Q_{ii'}) = P(Q_{ii'})\left\{log\ P(Q_{ii'}) + \frac{1}{2}ln|V_{ii'}|\right\} - P(Q_i)\left\{log\ P(Q_i) + \frac{1}{2}ln|V_i|\right\} - P(Q_{i'})\left\{log\ P(Q_{i'}) + \frac{1}{2}ln|V_{i'}|\right\} \quad (17)$$

To approximate (17) approximately, we use a top down clustering algorithm based on phonetic decision trees (CART) as described by [10]. We start with a single cluster, which contains all sounds. Then the cluster is split into two clusters in such a way that $\Delta(Q_{ii'})$ for the two clusters is maximal. Than the resulting clusters are split further till the chosen value N_Q is reached. The phonetic decision tree restricts the merging of sounds according to the rules encoded in the tree.

2.4 HCM Clustering

We apply a bottom up clustering method. If we would use each individual sound S as an initial segment we would run in the problem of many very small clusters C_i^l leading to unreliable estimates for the covariance matrices for evaluating (12). Instead we use as initial segments HMM-segments. For this purpose we create a large number N_Q of HMM-segments

with $N_Q \gg N_Q^{m_0}$ and assume that most segments represent well the individual sounds S. We take the sets C_i given from the HMM-segments and split each set C_i to m_0 'extended' sets $C_i^l; l = 1, \dots, m_0$ (see (5, 6)). These extended sets represent 'extended HMM-segments'. Thus we have $N_Q^{m_0} \equiv m_0 \cdot N_Q$ extended HMM-segments Q_i^l. With the extended HMM-segments we perform bottom up clustering. We start calculating $\Delta(Q_{ii'}^l)$ of all Q_i^l, $Q_{i'}^l$ separately for each l and phoneme Ph_j . The pair of minimal increase of $\Delta(Q_{ii'}^l)$ is merged giving a new segment $Q_{ii'}^l$. This procedure is continued till the requested number $N_Q^l; l = 1, \dots, m_0$ of HCM-segments is reached. This procedure approximates (11). The merging process is started only for segments, where the extended sets C_i^l are large enough to estimate the parameters $\Theta_i^l = \{\vec{\mu}_{il}, \vec{V}_{il}\}$. Especially the condition $|\vec{V}_{il}| > 0$ must hold. Each 'small' segment is merged to that final HCM-segment with the smallest heuristic distance.

3 Evaluation of HCM- and HMM-Segments

The experimental set up is the same as described in [6], where a Spanish speech database from the QUAERO project [12] was used. As primary features we use 16 MFCCs per frame with a frame rate of 10ms. The feature vectors are constructed in two steps. First a super vector of dimension 144 is build concatenating the 4 right and left MFCC vectors inclusive the central MFCC vector. The super vector is transformed by an LDA, which reduces the dimension to a 24 dimensional feature vector. Table 1 shows the amount and probabilities of chunks. As the probabilities $P(C^l)$ of the chunks $\vec{X}_l$ for $l > 3$ are very small, we set m_0=3 i.e. segments Q_i^l for l=1,..., m_0 are constructed.

Language	# chunks training	# chunks test	length l of chunks & $P(C^l) in$ %					
			1	2	3	4	5	≥ 6
Spanish	7 721 815	570 492	26.3	62.5	6.4	1.9	0.8	2.2

Table 1 - amount of data and probability $P(C^l)$ of the chunks $\vec{X}_l^n$

We start generating 604 HMM-segments (N_Q=604) using the CART described in section 2.3. The CART is provided from the speech research group of Prof. Ney (Chair of Informatics, RWTH). The **HMM-segments** are transformed to extended HMM-segments as described in section 2.4. We take the $m_0 \cdot N_Q = 1812$ extended HMM-segments $Q_i^l; l = 1, \dots, m_0; i = 1, \dots, N_Q$. In this experimental set up the extended HMM-segments are used for **two purposes**. First they are used for making phoneme recognition experiments second they are used as starting segments for constructing HCM-segments. This experimental set up is not optimal because the amount of HMM-segments to start the clustering should be much larger (see chapter 'conclusion'). In this set up the 1812 extended HMM-segments are clustered down to the target number of $N_Q^{m_0} = 1220$ HCM-segments.

l	# extended HMM-segments	# HCM-segments	HMM $H(Q^l)$	HCM $H(Q^l)$
1	604	435	8.80	8.42
2	604	436	8.92	8.50
3	604	349	8.82	8.22
4	604	349	8.53	7.98
5	604	349	8.06	7.52

Table 2 – number of extended HMM-segments and HCM-segments for l=1,...,5 and the related entropies (in Bit)

Given the probabilities $P(Q_i^l)$ of the extended HMM-segments and HCM-segments the related entropies $H(Q^l)$ (8) are calculated (see table 2). These entropies and all other entropies

have as units *bit,* i.e. they are calculated using as *log*-function the base 2. $H(Q^l)$is the number of bits needed from the information of the features to achieve a segment error rate (*SER*) of *SER*=0. As we have less HCM-segments than extended HMM-segments for each length the entropies $H(Q^l)$ are lower for HCM-segments.

Given the extended HMM-segments and the final HCM-segments, the HCM-classifiers are trained for both segments using as acoustic model:

$$p_l(\vec{X}_l|Q_i^l,\Theta_i^l)=\sum_{k=1}^{K_{il}} c_{ikl}N(\vec{X}_l;\vec{\mu}_{ikl},\vec{V}_l);\ \Theta_i^l=\{\vec{\mu}_{ikl},\vec{V}_l\} \tag{18}$$

In (18) we have to distinguish between the extended HMM-segments and HCM-segments, which are both notated as Q_i^l. The same holds for the related parameters Θ_i^l.The covariance matrices $\vec{V}_l$ of the multimodal Gaussians are tied as described in [6]. The HCM-parameters Θ_i^l are trained using the unified EM algorithm [13]. This algorithm has to be adapted for the HCM approach. The EM-algorithm uses instead of the feature vectors chunks. For classification of the two kinds of segments we use a maximum likelihood classifier $\hat{Q}=argmax_i\left(p_l(\vec{X}_l|Q_i^l,\Theta_i^l)P(Q_i^l)\right)$. For classifying a context independent phonemes Ph_j we have to regard all N_j sequences of segments $SQ_{n_j}, n_j=1,\dots,N_j$ building that phoneme. For a given test database we have to classify phonemes $Ph_j^{l_{Ph}}$ realized by a sequence of l_{Ph} feature vectors and realized by a sequence $SQ_{n_j}^{l_{Ph}}=[Q_{i_p}^{l_p},p=1,2,3]$ of 3 segments (this corresponds to the split of a phone into 3 sounds (see section2.1)). We assume, we know the length l_p of each segment $Q_{i_p}^{l_p},p=1,2,3$. Thus the classification task is to determine the indices $i_p,p=1,2,3$ The maximum likelihood classifier is given by

$$\left.\begin{aligned}\widehat{Ph}_j^{l_{Ph}}&=\text{argmax}_\text{j}\left(\text{p}(\vec{\text{X}}_{\text{l}_{\text{Ph}}}|Ph_j^{\text{l}_{\text{Ph}}},l_p,p=1,2,3)(Ph_j^{\text{l}_{\text{Ph}}})\right)\\ p(\vec{X}_{l_{Ph}}|Ph_\text{j}^{l_{Ph}},l_p,p=1,2,3)&=\sum_{n_j=1}^{N_j}\text{P}(Q_{n_j}^{l_{Ph}}|\text{Ph}_\text{j}^{l_{Ph}})\prod_{\text{p}=1}^{3}p_{l_p}\left(\vec{X}_{l_p}\middle|Q_{i(n_j,p)}^{l_p},\Theta_{i(n_j,p)}^{l_p};\right)\end{aligned}\right\}$$

In [14] we regarded additionally the case that the length l_p is not known. This knowledge leads to slightly higher *SER*s. Shannon's conditional entropies $H(Q^l|\vec{X}_l)=H(Q^l)-H(C^l)+H(\vec{X}_l|Q^l)$ (see (8)) is the information missing to classify the segments Q_i^l of given length l without error. We approximate $H(Q^l|\vec{X}_l)$ by the method of Monte Carlo [15] as done in [14]. As shown in table 3 the high values of $H(Q^l|\vec{X}_l)$ corresponds to the high values of *SER*. Table 3 shows that the SERs using the extended HMM-segments are lower than those of the HCM-segments. This is a strange result to be further explored, because the HCM-segments show much lower values for $H(Q^l|\vec{X}_l)$. Although less segments are used the phoneme error rate (*PER*) is lower for HMM-segments.

Extended HMM-segments								HCM-segments							
	length *l* of chunks								length *l* of chunks						
#	1	2	3	1	2	3		#	1	2	3	1	2	3	
modes	*SER*			H(Q|$\vec{X}_l$)			*PER*	modes	*SER*			H(Q|$\vec{X}_l$)			*PER*
1 812	82.9	73.1	62.6	6.72	6.13	5.91	41.8	1 220	83.1	73.4	62.9	6.22	5.30	4.66	41.5
3 624	79.8	71.4	60.5	6.71	5.89	5.51	39.4	3 624	80.4	71.3	62.9	6.22	5.40	5.00	39.0
10 872	77.5	70.5	59.6	6.23	5.75	5.35	38.1	10 872	78.0	71.1	61.9	5.85	5.28	5.02	37.7

Table 3 - *SER*, *PER* and $H(Q|\vec{X}_l)$ for HCM-segments and HMM-segments

4 Acknowledgement

We would like to thank Muhammad Tamir, Christian Plahl, and Hermann Ney from the RWTH Aachen University, Germany for kindly providing the labeled QUAERO databases and the related CARTs.

5 Conclusion

We derived a theory for clustering sounds to segments, where the clustered sounds are modeled by HCMs. The theory is based on cluster-algorithms minimizing Shannon's entropy using the concept of HCMs. We presented preliminary experiments showing that HCM-segments lead to slightly lower phoneme error rates. This encouraging result is in contrast to higher SERs, which still has to be explored. Further in this experiment, the initial amount of HMM-segments is low. In future we will experiment with larger amount of HMM-segments.

References

[1] Fukunaga, K.: Introduction to Statistical Pattern Recognition. Second Edition, Academic Press, 1990

[2] Rybach, D., Gollan, C., Heigold, G., Hoffmeister, B., Lööf, J., Schlüter, R., and Ney, H.: The RWTH Aachen University open source speech recognition system. in Proc. Interspeech, Brighton, U.K.:2111–2114, 2000

[3] 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.

[4] Hwang, M.Y., and Huang, X.: Shared-distribution hidden Markov models for speech recognition. IEEE Trans. Speech Audio Processing Vol. I,pp.414-420:Oct. 1993

[5] Höge, H., Setiawan, P.: Improvements of Hidden Chunk Models. ESSV Berlin 2010

[6] Höge, H.: The Use of Conditional Gaussians for Hidden Chunk Models. In Proc. ESSV Cottbus 2012

[7] Höge, H.: Comparison of HMMs and HCMs. In Proc. ESSV Bielefeld 2013

[8] Shannon, C.E.: A Mathematical Theory of Communication. Bell System Technical Journal, Vol. 27: July and October 1948, pp. 379-423 and 623-656.

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

[10] Beulen, K. : Phonetische Entscheidungsbäume für die automatische Spracherkennung mit großen Wortschätzen. Dissertation RWTH 1999

[11] Papoulis, A.: Probability, Random Variables, and Stochastic Processes. Third Edition, McGraw-Hill Series in Electrical Engineering 1991

[12] Sundermeyer, M., Nußbaum-Thom, M., Wiesler, S., Plahl, C., El-Desoky Mousa, C.A., Hahn, S., Nolden, D., Schlüter, R., and H. Ney: The RWTH 2010 QUAERO ASR evaluation system for English, French, and German. In Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, Prague, Czech Republic: 2212–2215, 2011.

[13] Huang, X. D., Ariki, Y. and Jack, M. A.: Hidden Markov Models for Speech Recognition. Information Technology Series, Edinburg University Press, 1990

[14] Höge, H.: Modeling Statistic Dependencies of Feature Vectors within Phonemes. In: Mehnert, D., Kordon, U. und Wolff, M. (Ed.): Systemtheorie, Signalverarbeitung, Sprachtechnologie, Vol. 68: Studientexte zur Sprachkommunikation. TUDpress: Dresden 2013, pp. 36 – 64.

[15] Robert, C. and Casella, G.: Monte Carlo Statistical Methods. Second edition Springer Verlag: New York 2004