COMPARISON OF HMMS AND HCMS

Harald Höge

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

Abstract: We investigate acoustic models for segments, which are sub-phonetic units derived from clustered tri-phones. For each segment Q_i we evaluate acoustic models approximating the conditional density function (cdf) $p(\vec{X}_l|Q_i)$ of sequences $\vec{X}_l$ of features vectors aligned to a segment Q_i. We name those aligned sequences $\vec{X}_l$ 'chunks'. The quality of the acoustic models is evaluated by segment error rates (*SER*) and Shannon's Conditional Entropy $H(Q|\vec{X}_l)$. Further we develop a new method to answer the question, how close an acoustic models approximates the cdf $p(\vec{X}_l|Q_i)$. The method is based on the simulation of model generated chunks (MgCs), which have a cdf as given by the acoustic model. We evaluate Hidden Markov Models (HMMs) and Hidden Chunk Models (HCM) realized by GMMs with tied covariance matrices. Comparing the *SER* and $H(Q|\vec{X}_l)$ for the same amount of modes for HMMs and HCMs we see that HCMs perform in general better than HMMs. Evaluation experiments with MgCs show, that both acoustical models are still far away from the real distribution $p(\vec{X}_l|Q_i)$. It is still an open question, weather GMMs with tied covariance matrices are good candidates to approximate the $p(\vec{X}_l|Q_i)$ or weather any explored mixture show with increasing number of modes slow convergence to $p(\vec{X}_l|Q_i)$.

1 Introduction

This paper is focused on evaluating acoustic models for sub-phonetic units called segments, which are derived from clustered tri-phones. The clusters are constructed by a CART as used in HMM technology. Each clustered tri-phones is modelled by 3 segments, where each segment can be interpreted as the acoustic representation of either the onset or the middle or the offset of a tri-phone cluster. In total we regard about 600 segments yielding about 1500 tri-phone cluster. The duration of the segments is rather short. As explored by 3 large speech databases [12] recorded for 3 languages most segments have a duration of about 30ms. Using frame shifts of 10-15ms we found, that more than 92% of the segments are realized by 1-5 feature vectors. Due to this small amount of feature vectors aligned to a segment it seems to be feasible to find an acoustic model, which comes close to the exact distribution of **all** features building a segment. If the acoustic model approaches the exact distribution, the lowest segment error rate possible could be reached. Segment models [1,2] have the potential to achieve this goal. According to the segment model approach, the conditional density function (cdf) $p_l(\vec{X}_l|Q_i)$ of the complete **sequence** $\vec{X}_l = [X_l, \ldots, X_v, \ldots X_1]$ of feature vectors aligned to a segment Q_i is regarded. We use a specific segment model called Hidden Chunk Model (HCM) [3], where each aligned sequences $\vec{X}_l$ is called a 'chunk'. For each length l of a chunk a specific HCM depending on l has to be trained [12]. We use Gaussian Mixture Models (GMMs) with tied covariance matrices to construct HCMs. For chunks of length l, the HCM is given by:

$$\tilde{p}_l(\vec{X}_l|Q_i) = \sum_{k=1}^{K_{il}} c_{ikl} N(\vec{X}_l; \vec{\mu}_{ikl}, \vec{V}_l) \quad (1)$$

The tying approach of the covariance matrices $\vec{V}_l$ is motivated by the use of LDA transformed features, which normalize to a certain extend the covariance matrices of the individual

segments and which lead to good performance in HMM based LVCSR systems [14]. Each mean vector $\vec{\mu}_{ikl}$ is a composition of l mean vectors $\vec{\mu}_{ikl} = [\mu_{ikl1}, \ldots., \mu_{ikll}]$. A $\vec{\mu}_{ikl}$ can be interpreted as a trajectory or a template of feature vectors in the acoustic space of an 'exemplar sound'. To interpret $\vec{V}_l$ and $\vec{\mu}_{ikl}$ in (1) more intuitively, we assume that each trajectory $\vec{\mu}_{ikl}$ corresponds to a specific articulator gesture. We postulate that gestures of the same segment Q_i produced with different speed - given by l - lead to significant different acoustic realizations. Thus different speed l should lead to different means μ_{iklv}. Variations of the gestures are expressed by the covariance matrix $\vec{V}_l$ and by the modes k.

Within this paper we want to compare the performance of Hidden Markov Models (HMMs) [4] with HCMs for segments. We measure performance by evaluating segment error rates (*SER*) and Shannon's conditional Entropy as discussed in section 2.1. The acoustic model of HCMs is given by (1). The acoustic model for a segment using HMMs is given by

$$p^{HMM}(\vec{X}_l|Q_i) = \prod_{v=1}^{l} p_{S_i}(X_v|Q_i) \quad (2)$$

The emission probabilities $p_{S_i}(X_v|Q_i)$ is the probability, a state S_i - assigned to a segment - emits a feature vector. In analogy to (1) we model the emission probabilities by GMMs with a single tied covariance matrix. As seen on relation (2) HMMs assume, that the feature vectors are independently distributed and that all feature vectors of a segment are identically distributed. In HMM technology it is well known that the performance increase with increasing amount of modes of the GMMs. In order to make a fair comparison between HMMs and HCMs we evaluate error rates and Shannon's entropy for the same number of modes. As shown in chapter 3 the HCMs perform to a great extend better than the HMMs.

For any acoustic model it is an open question, weather a model is close to the exact distribution $p_l(\vec{X}_l|Q_i)$. To answer this question we developed a new paradigm to evaluate acoustic models, which is based on simulating specific chunks call **M**odel **g**enerated **C**hunks (MgCs). On the basis MgCs we define MgC-distances, which describe the distance of the model to the exact distribution. The new evaluation method is described in detail in section 2.2. Experimental results are shown in chapter 3.

2 Evaluation Methods

In speech recognition the most relevant method for measuring the performance of acoustic models is the error rate achieved on a given speech database for given phonetic unit. For this demand we evaluate segment error rates (*SER*s) for HMMs and HCMs in section 2.1. Further we evaluate on the basis of Shannon's conditional entropy the bounds of *SER*s in the same section. The *SER* achieved depends closely on the quality of approximation of the model to the exact distribution $p_l(\vec{X}_l|Q_i, l)$. A popular approximation measure is given by the Kullback-Leibler distance, which in our case has to be evaluated for GMMs. There exist a rich literature on finite mixture models [15], but to determine the KL-distance the exact distribution $p_l(\vec{X}_l|Q_i, l)$ must be known. As this is not given we describe in section 2.2 a new evaluation method, which allows to determine the distance between the real distribution and the model, even the cdf $p_l(\vec{X}_l|Q_i, l)$ is unknown.

2.1 Error Rates and Shannon's Conditional Entropy

We assume that a given speech database is labeled in such away, that each chunk aligned to the segments is given. Given the chunks we perform classification experiments using a maximum likelihood classifier

$$\hat{Q} = argmax_i\left(\tilde{p}_l(\vec{X}_l|Q_i, l)P(Q_i|l)\right) \quad (3)$$

$\tilde{p}_l(\vec{X}_l|Q_i,l)$ denotes the acoustical model for $p_l(\vec{X}_l|Q_i,l)$. We use as a-priori probability the duration dependent discrete distribution $P(Q_i|l)$. Based on (3) the segment error rate (*SER*) can be evaluated.

In order to find bounds for *SER*s we use Shannon's conditional Entropy [6,10]. Given segments Q_i ,$i=1,...,N_Q$ realized by chunks $\vec{X}_l$ of length l Shannon' conditional entropy $H_l(Q|\vec{X}_l)$ is defined by

$$\left.\begin{array}{c} H_l(Q|\vec{X}_l)\equiv H_l(Q)-I_l(\vec{X}_l;Q);\ I_l(\vec{X}_l;Q)\equiv H_l(\vec{X}_l)-H_l(\vec{X}_l|Q) \\ H_l(Q)\equiv-\sum_{i=1}^{N_Q}P(Q_i|l)\log\left(P(Q_i|l)\right);\ H_l(\vec{X}_l)\equiv-\int p(\vec{X}_l)log\ p(\vec{X}_l)\,d\vec{X}_l \\ p(\vec{X}_l)=\sum_{i=1}^{N_Q}P(Q_i|l)p_l(\vec{X}_l|Q_i) \\ H_l(\vec{X}_l|Q)\equiv-\sum_{i=1}^{N_Q}P(Q_i|l)\int p_l(\vec{X}_l|Q_i)\ log\ p_l(\vec{X}_l|Q_i)\,d\vec{X}_l \end{array}\right\} \quad (4)$$

$H_l(Q)$ is the information needed to recognize the segments Q_i aligned to a chunks of length l without error. The mutual information $I_l(\vec{X}_l;Q)$ is the information gained from the chunks $\vec{X}_l$. Whenever the relation $H_l(Q)>I_l(\vec{X}_l;Q)$ holds, errors occur. Given $H_l(Q|\vec{X}_l)$ upper and lower bounds for the *SERs* are known. We use as lower bound the Fano bound [7] and as upper bound the Golic bound [8]. These bounds are functions of the entropy $H_l(Q|\vec{X}_l)$. To evaluate the bounds we need the distributions $P(Q_i|l)$ and $p_l(\vec{X}_t|Q_i)$ as given by (4). Whereas the discrete distribution $P(Q_i|l)$ can be estimated with high accuracy on large databases, the cdf $p_l(\vec{X}_t|Q_i)$ is unknown. We approximate the cdf $p_l(\vec{X}_t|Q_i)$ by our model (1) leading to pdfs $\tilde{p}_l(\vec{X}_l|Q_i)$. The entities defined in (4) are approximated in 2 steps. In the first step expressions as $-\int p_l(\vec{X}_l|\,Q_i)\ log\ p_l(\vec{X}_l|Q_i)\,d\vec{X}_l$ are approximated by $-\int p_l(\vec{X}_l|\,Q_i)\ log\ (\tilde{p}_l(\vec{X}_l|Q_i)\,d\vec{X}_t$. In the second step we apply the Monte Carlo Method [9]:

$$-\int p_l(\vec{X}_l|\,Q_i)\ log\ (\tilde{p}_l(\vec{X}_l|Q_i)\,d\vec{X}_t \approx -\frac{1}{N_{S(i,l)}}\sum_{n=1}^{N_{S(i,l)}}log\ \tilde{p}_l(\vec{X}_l^n|Q_i) \quad (5)$$

using the $N_{S(i,l)}$ samples of chunks $\vec{X}_l^n\ n=1,\dots,N_{S(i,l)}$ of length l assigned to the segment Q_i. The quality of the approximation (5) increases with increasing number of samples. Taking into account the general inequality $-\int p(Z)logp(Z)dZ\le-\int p(Z)logf(Z)dZ$, which holds for any distributions $p(Z)$, $f(Z)$, we can relate $H_l(Q|\vec{X}_l)$ to its approximation $\widetilde{H}_l(Q|\vec{X}_l)$:

$$H_l(Q|\vec{X}_l)=-\sum_{i=1}^{N_Q}P(Q_i|l)\int p_l(Q_i|\vec{X}_l)\log p_l(Q_i|\vec{X}_l)\,d\vec{X}_l\le\widetilde{H}_l(Q|\vec{X}_l) \quad (6)$$

$$\widetilde{H}_l(Q|\vec{X}_l)\equiv-\sum_{i=1}^{N_Q}P(Q_i|l)\int p_l(Q_i|\vec{X}_l)\log\tilde{p}_l(Q_i|\vec{X}_l)\,d\vec{X}_l;\ \ \tilde{p}_l(Q_i|\vec{X}_l)\equiv\frac{\tilde{p}_l(\vec{X}_l|Q_i)P(Q_i|l)}{\sum_{i=1}^{N_Q}P(Q_i|l)\tilde{p}_l(\vec{X}_l|Q_i)}$$

Due to the approximations made, the bounds are approximations and depend on the quality of the acoustic model. Yet from (6) we conclude, that the Fano bound determined with $\widetilde{H}_l(Q|\vec{X}_l)$ is still an exact lower bound.

2.2 The MgC-Distances

Our basic idea is to exchange the original chunks of the speech database by simulated chunks, which have exact the distribution $\tilde{p}_l(\vec{X}_l|Q_i,l)$ of the acoustic model. We call such chunks '**M**odel **g**enerated **C**hunks (MgCs)'. No we evaluate the *SER* and the entropy $H_l(Q|\vec{X}_l)$ for a MgCs generated database and for the original speech database. Thus we get 2 values for the *SER* and 2 values the entropy $H_l(Q|\vec{X}_l)$. We define as MgC-*SER*-distance the difference of the two *SER*s and the MgC-$H_l(Q|\vec{X}_l)$- distance the difference of the two entropies $H_l(Q|\vec{X}_l)$. We hypothesize, that if the MgC-distances are zero, the optimal acoustic model is found,

delivering the minimal *SER* and $H_l(Q|\vec{X}_l)$. Vice versa we hypothesize, that large distances hint for non optimal models.

Now we describe in detail, how we construct the MgC based database. We start with the labels of the speech database. We extract the sequence of segments together with the related length of chunks. Then for each segment the aligned chunk is exchanged by an MgC. To simulate an MgC of length l for a segment Q_i we use a mode generator and a multivariate Gaussian generator. According to (1) the mode generator has to generate a number k in the range [1,...K_{il}] with probabilities $c_{ikl.}$ Given the number k, a multivariate Gaussian generator generates a vector of the dimension of the chunk with the probability $N(\vec{X}_l; \vec{\mu}_{ikl}, \vec{V}_l)$. For both generators matlab® code is available. As the length of chunks can be very long, the resulting high dimension of such chunks may be no longer tractable in practice. To solve this problem we use the decomposition method described in [12], where a high dimensional multivariate Gaussian is decomposed by low dimensional Gaussians:

$$N(\vec{X}_l; \vec{\mu}_{ikl}, \vec{V}_l) = N_{l1}(X_1; \mu_{ikl|1}, V_{l|1}) \prod_{v=2}^{l} N_{lv}(X_v | X_{v-1}, \dots, X_1; \mu_{i,k,l|v-1}, V_{l|v}) \quad (7)$$

For long chunks, where no HCMs are trained, we use an n-gram approach [12]. According to (7) we have to generate only vectors from the low dimensional Gaussian distribution N_{lv} given by the conditional means $\mu_{i,k,l|v-1}$ and by the conditional covariance matrices $V_{l|v}$. By concatenating the l simulated vectors we construct a complete chunk, which is an MgC. The sequence of MgCs generated according to the labels of the speech database constructs the MgC based database, whose chunks have the distribution as given by the HCMs.

3 Experimental Evaluation of HCMs and HMMs

3.1 Experimental Set Up

The experimental set up is the same as described in [12]. We use the same Spanish and French speech databases from the QUAERO project [13] and the same CART to generate the labels for segments. We use 16 MFCCs per frame. 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. Using LDA the super vector is reduced to a 24 dimensional feature vector.

Speech database	#chunks for training	#chunks for test	length l of chunks & $P(l\|Q)in$ %					
			1	2	3	4	5	≥ 6
Spanish	7 721 815	570 492	26.3	62.5	6.4	1.9	0.8	2.2
French	27 164 564	4 095 502	26.8	31.6	20.6	9.4	4.1	7.6

Table 1 - amount of data and length distribution $P(l|Q)$ of the chunks

Speech database	# of segments N_Q	l & $H_l(Q)[bit]$					
		1	2	3	4	5	≥ 6
Spanish	604	8,80	8,92	8,82	8,53	8,06	3,21
French	598	8,77	8,75	8,72	8,62	8,54	7,63

Table 2 - number of segments and entropy $H_l(Q)$

The size of the databases and the length distribution of the chunks are shown in table 1. The French chunks are longer than those of Spanish. As shown in the next sections this difference has great impact on the properties of the acoustic models. In the following we use as *log*-function the base 2. Thus the entropies defined in (4) have as units *bit*. Table 2 shows the entropies $H_l(Q)$ evaluated by the distribution $P(Q_i|l)$. For equal distributed segments $H_l(Q)$

would take the value $log_2 N_Q$ (e.g. log_2 604=9.25 [*bit*]). The HCMs are trained on the speech databases using the unified EM algorithm [4] exchanging as samples the feature vectors by chunks. The HMMs are trained on the speech database and on the MgC based database.

3.2 Error Rates and Shannon's Conditional Entropy

The segment error rates are evaluated using (3). Shannon's conditional entropy is evaluated using approximations as described in section 2.1. In the following we denote by $H(Q|\vec{X}_l)$ the approximated version. Due to the data available (see table 1) HCMs for Spanish are trained till l_0=3 and for French till l_0=5. To compare HMMs with HCMs we compare models with the same number of modes (NoM). For HCMs the NoM is defined by the sum of all modes of the HCMs trained. Thus the French HCMs have for the same value of NoM less modes per HCM than the Spanish ones. Further the HCMs are trained in such a manner, that the number of modes of each HCM for given l is equal i.e. each HCM(l) has NoM/l modes. The HMMs evaluated on the MgC test database are trained on the training part of the MgC database.

	HCMs						**HMMs**					
	length *l* of chunks						**length *l* of chunks**					
	1	2	3	1	2	3	1	2	3	1	2	3
NoM	**SER**			$\boldsymbol{H(Q\|\vec{X}_l)}$			**SER**			$\boldsymbol{H(Q\|\vec{X}_l)}$		
1 812	82.9	73.1	62.6	6.72	6.13	5.91	82.5	72.3	69.4	7.64	7.59	9.50
3 624	79.8	71.4	60.5	6,71	5.89	5.51	81.5	71.0	68.3	7.32	6.95	8.59
10 872	77.5	70.5	59.6	6.23	5.75	5.35	79.9	69.5	67.6	6.89	6.21	7.44
21 744	76.2	70.0	59.4	5.99	5.77	5.34	79.2	68.7	66.7	6.66	5.92	7.10
43 488	75.6	69.9	59.0	5.83	5.73	5.33	78.7	68.6	66.9	6.50	5.70	6.90

Table 3 - SER and $H(Q|\vec{X}_l)$ for HCMs and HMMs for the **Spanish** test database

NoM	**HCM**				**length *l* of Chunks**					
	1	2	3	4	5	1	2	3	4	5
	SER					$\boldsymbol{H(Q\|\vec{X}_l)}$				
2 990	69.6	57.0	53.0	52.1	52.9	5.50	4.38	4.18	4.14	4.25
5 980	66.6	55.2	52.0	51.0	52.3	5.38	4.31	4.24	4.25	4.27
20 000	63.0	54.3	51.6	51.0	52.1	4.77	4.15	4.12	4.13	4.31

NoM	**HMM**				**length *l* of Chunks**					
	1	2	3	4	5	1	2	3	4	5
	SER					$\boldsymbol{H(Q\|\vec{X}_l)}$				
2 990	70.5	61.0	61.7	65.1	68.7	6.01	5.90	7.30	9.33	11.97
5 980	67.0	57.2	59.1	64.5	70.9	5.36	5.07	6.60	9.35	13.33
20 000	65.1	55.8	58.4	64.5	71.1	4.90	4.56	6.03	8.62	12.33

Table 4 - SER and $H(Q|\vec{X}_l)$ for HCMs and HMMs for **French** test database

Table 3 (Spanish) and 4 (French) show the *SER*s and approximated $H(Q|\vec{X}_l)$ for HCMs and HMMs achieved for different NoM. In general the values of *SER*s and $H(Q|\vec{X}_l)$ drop with increasing l. Further for l>2 the HCMs perform always better than the HMMs. The good performance of the HMMs for l=2 for Spanish can be explained that most chunks have the length l=2 and the HMM concentrates most of its modes to these chunks. This is not the case for French.

3.3 MgC-Distances

For simulating model generated chunks (MgCs) we have to distinguish between two models characterized by their number of modes (NoMs). The **first** model concerns the model used to generate the MgCs. The modes from this model are called chunk modes. The **second** model concerns the model used to perform maximum likelihood classification (3). The modes from this model are called model modes. The first model is always a HCM, as this model delivers the lowest values for *SER* and $H(Q|\vec{X}_l)$. The second model is either a HCM or a HMM. To evaluate the impact of the NoMs we investigate two cases: the 'case of match', where the number of NoMs of the chunk modes and number of NoMs of the model modes are equal; The 'case of mismatch', where the values of the two kinds of NoMs are different.

NoM		length *l* of Chunks					
chunk	model	1	2	3	1	2	3
modes	modes	SER [%]			$H(Q\|\vec{X}_l)$[*bit*]		
1 812	1 812	80.0	61.7	42.9	5.40	3.62	2.27
3 624	3 624	73.6	59.6	43.1	4.72	3.44	2.30
3 624	1 812	82.0	63.9	46.3	5.81	4.10	2.95
10 872	10 872	71.0	59.4	44.0	4.43	3.43	2.35
21 744	21 744	69.7	59.2	44.2	4.31	3.41	2.38
21 744	10 872	76.2	63.1	46.2	5.16	3.78	2.56
21 744	3 624	79.9	63.1	46.6	5.97	3.93	2.73
21 744	1 812	83.2	65.1	48.3	6.34	4.37	3.19
43 488	43 488	67.7	59.3	44.6	4.16	3.42	2.41

Table 5 - SER and $H(Q|\vec{X}_l)$ **Spanish** HCMs

NoM		length *l* of Chunks					
feature	model	1	2	3	1	2	3
modes	modes	SER [%]			$H(Q\|\vec{X}_l)$[*bit*]		
1 812	1 812	84.8	69.8	65.7	6.04	4.72	5.18
10 872	10 872	80.9	66.7	64.9	5.33	4.37	5.04
21 744	21 744	80.6	67.3	66.3	5.51	4.41	5.10

Table 6 - *SER* and $H(Q|\vec{X}_l)$ of MgC based **Spanish** test database**;**
HMMs trained on MgC based **Spanish** training database

NoM		length *l* of Chunks									
feature	model	1	2	3	4	5	1	2	3	4	5
modes	modes	SER [%]					$H(Q\|\vec{X}_l)$[*bit*]				
2 990	2 990	59.2	36.3	29.9	28.7	29.2	3.32	1.78	1.43	1.37	1.41
5 980	5 980	54.0	36.2	31.2	30.2	30.7	2.94	1.77	1.49	1.46	1.51
20 000	20 000	51.8	37.6	32.8	32.0	31.4	2.81	1.86	1.61	1.61	1.65

Table 7 - *SER* and $H(Q|\vec{X}_l)$ **French HCMs** for MgC based test database

We first regard the match case. As expected, the values of *SER* and $H(Q|\vec{X}_l)$ drop with increasing NoMs (see table 5-7). Comparing these tables with the tables 3 and 4 for the same NoM of model modes we can determine the MgC-distances. The MgC-distances are larger for $l>1$ than for $l=1$. This result can be explained, that for the case $l=1$, the chunk is a single feature vector and no statistic dependencies between feature vectors has to be modeled. This is in contrast to the case $l>1$, where statistic dependencies have to be modelled.

NoM		length *l* of Chunks									
feature modes	model modes	1	2	3	4	5	1	2	3	4	5
		MgC-SER [%]					MgC - $H(Q\|\vec{X}_l)$ [*bit*]				
2 990	2 990	10.4	20.7	23.1	23.4	23.7	2.18	2.60	2.75	2.77	2.84
5 980	5 980	12.6	19.0	20.8	20.8	21.6	2.44	2.54	2.75	2.79	2.76
20 000	20 000	11.2	16.7	18.8	19.0	20.7	1.96	2.29	2.51	2.52	2.66

Table 8 - MgC-distances for the French database

As seen in table 8, the MgC-distances drop with increasing NoMs in general. For *l*=1 we see that the MgC-distances increase first and drops for larger NoMs. This behavior can also be observed for Spanish for *l*=1,2. For larger NoM we expect, that the distances drop further continuously. If a convergence to zero can be achieved at all, is still an open question. Due to the values for SER and $H(Q|\vec{X}_l)$ for increasing NoM shown in table 8, the NoMs must be very large to achieve small distances. To yield reliable models very large databases are needed.

Regarding the miss match case we see on table 5 a large decrease in the values of *SER* and the $H(Q|\vec{X}_l)$ with increasing NoMs. Comparing this decrease of *SER* and the $H(Q|\vec{X}_l)$ of table 3 and 4 for increasing NoMs, we see that this descrease is much bigger for the MgCs. This difference in behavior indicates that the chunks have not a multimodal Gaussian distribution with tied covariance matrices.

3.4 Bounds of Error Rates

Figure 1 shows a Fano-Golic plot for the bounds of the SERs for Spanish for speech derived chunks and MgCs. The MgCs are simulated for the case, that the NoMs for both models have the same value (NoM=43.488). The values of *SER* and $H(Q|\vec{X}_l)$ are taken from table 3 and 5. In the case of MgCs, Shannon's entropy $H(Q|\vec{X}_l)$ is exact, as the distribution $\tilde{p}_l(\vec{X}_l|Q_i, l)$ is known. In the case of speech derived chunks, $H(Q|\vec{X}_l)$ is approximated. Due to (6) the exact $H(Q|\vec{X}_l)$ is smaller. This means, that the points shown in figure 1 on the left side have to be shifted down to an unknown amount. Thus also for the approximated entropy the Fano bound is still a correct lower bound, but more conservative, whereas the upper bound is only an approximation. As the *SERs* of the MgCs are smaller, it seems that the MgC based error rates could be taken as a much tighter upper bound than the Golic bound. If this hypothesis is correct, has to be explored theoretically.

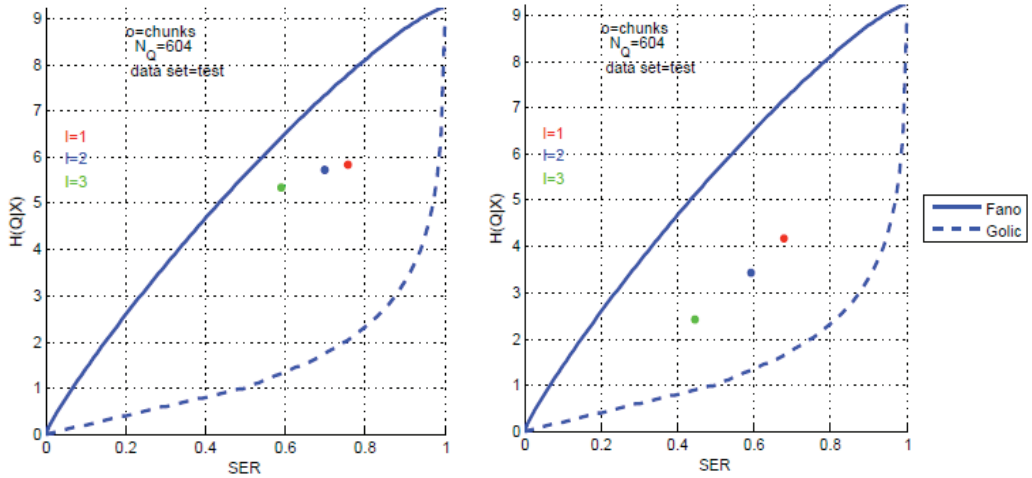

Figure 1 - Bounds of SER for Spanish with 43.488 modes; left plot: speech chunks - right plot: MgCs

4 Acknowledgement

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

5 Conclusion

Using GMM based HMMs and HCMs we have shown that in general HCMs perform better than HMMs for equal number of modes, especially for longer chunks. A new evaluation paradigm delivering a distance - the MgC- distance - has been presented for evaluating the quality of approximation between the acoustic model and the exact distribution. The experiments have shown that HCMs and HMMs show quite large distances. It seems that GMMs with very large number of modes are needed to achieve low distances.

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] Tokuda, K., Zen, H., and Kitamura, T., "Trajectory modeling based on HMMs with the explicit relationship between static and dynamic features“, Proc. of Eurospeech, 865-868, 2003.

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

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

[5] S. Kotz, N. Balakrishnan, N.L. Johnson, “Continuous Multivariate Distributions”, Vol 1: Models and Applications, John Wiley & Sons, Inc., 2000

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

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

[8] Golic, J.: On the Relationship between the Information Measures and the Bayes Probability of Error. IEEE Transactions on Information Theory, Vol. IT-33(5): 1987, pp. 681-693.

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

[10] Höge, H., Setiawan, P.: Shannon’s Conditional Entropy and Error Rates on Segment Level. Beiträge zur Signaltheorie, Signalverarbeitung, Sprachakustik und Elektroakustik - Dietrich Wolf zum 80. Geburtstag. Hrsg.: A. Lacroix, Studientexte zur Sprachkommunikation Band 52, TUD Press-Verlag, Dresden 2009

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

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

[13] 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.

[14] 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, 2009.

[15] Mclachlan, G. and Peel, D., "Finite Mixture Models". Wiley, New York. MR1789474, 2000.

Autorenverzeichnis