THE USE OF CONDITIONAL GAUSSIANS FOR HIDDEN CHUNK MODELS

Harald Höge

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

Abstract: Hidden Chunk Models (HCMs) are duration dependent trajectory models, which model the statistic dependencies of all feature vectors assigned to a segment. The segments are derived from clustered tri-phones as used by HMMs, where each tri-phone is composed by three segments. A sequence $\vec{X}_l$ of l feature vectors assigned to a segment is called chunks of length l. For each segment Q_i the pdf $p_l(\vec{X}_l|Q_i)$ constituting the HCMs is modeled by a GMM where the mode k is given by a Gaussians $N(\vec{X}_l;\vec{\mu}_{ikl},\vec{V}_l)$. The sequence $\vec{\mu}_{ikl}$ of l mean vectors represents a trajectory of an 'exemplar' segment realized by a chunk of length l. Investigating speech from 3 languages shows, that over 92% of the chunks have a length ranging from l=1 to 5. From those chunks the means and covariance matrices of $N(\vec{X}_l;\vec{\mu}_{ikl},\vec{V}_l)$ can be trained. Using the properties of multivariate Gaussians, the $N(\vec{X}_l;\vec{\mu}_{ikl},\vec{V}_l)$ can be decomposed into conditional Gaussians allowing a frame by frame processing. Due to the problem of sparse training data and numerical inaccuracies, the parameters $\vec{\mu}_{ikl},\vec{V}_l$ cannot be estimated for $l > 5$. For this problem we present an extension approach. The properties of the conditional Gaussians and their extensions are evaluated by classification experiments of segments and phonemes. Further Shannon's entropy is evaluated showing the quality of the models used.

1 Introduction

Hidden Chunk Models (HCMs) [3] are specific segment models [1] similar to trajectory models [2]. The segments are derived from clustered tri-phones, where each tri-phone is composed by three segments. In the framework of HMMs [4] such segments are modeled by tied states with identical emission probabilities. HCMs are defined by duration dependent pdfs $p_l(\vec{X}_l|Q_i)$, where $\vec{X}_l = [X_l, \ldots X_1]^T$ denotes a sequence of l feature vectors X_ν assigned to a segment Q_i. The sequences $\vec{X}_l$ are called 'chunks' of 'length' l. $p_l(\vec{X}_l|Q_i)$ can be interpreted as an extended emission probability, which models not the distribution of a single feature vector X_ν within a state (HMM-approach) but the distribution of a complete chunk $\vec{X}_l$ within a segment (segment model approach).For each Q_i and given chunk $\vec{X}_l$ the HCM is defined by

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

The means and covariance matrices $\vec{\mu}_{ikl},\vec{V}_l$ model the statistical bindings of the feature vectors within a chunk. For a given length l all covariance matrices are tied to a single matrix $\vec{V}_l$. This tying approach can be applied as LDA transformed features are used [13]. Each mean vector $\vec{\mu}_{ikl} = [\mu_{ikll}, \ldots., \mu_{ikl1}]$ is composed by a sequence of l mean vectors $\mu_{iklv}, \nu = 1, \ldots, l$, which can be interpreted as an exemplar trajectory representing an 'exemplar' segment. Thus - in contrast to HMMs - the means $\mu_{iklv}, \nu = 1, \ldots, l$ depend in addition to the indices i,k on the indices ν and l, which describe the 'position' ν of each feature vector within a trajectory of length l. We investigate the length of segments in 3 languages. Using a frame shifts of 10-15ms we found, that more than 92% of the chunks have a length ranging from l=1 to l=5. In this range the parameters of the Gaussians $N(\vec{X}_l;\vec{\mu}_{ikl},\vec{V}_l)$ can be trained reliable. As

described in chapter 2 the Gaussians $N(\vec{X}_l;\vec{\mu}_{ikl},\vec{V}_l)$ can be decomposed by conditional Gaussians $N_{lv}, v=1,\dots,l$:

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

The conditional Gaussians N_{lv} are called Trajectory Emission Probabilities (TEPs) describing the distribution of the feature vectors along a trajectory. Due to limitations of training material and numerical problems in determining $\vec{V}_l$, the Gaussians $N(\vec{X}_l;\vec{\mu}_{ikl},\vec{V}_l)$ can be trained no more reliable for large *l*. In chapter 2 we present a solution for this problem.

In order to evaluate the properties of the HCMs we make classification experiments of segments and phonemes combined with the evaluation of Shannon's entropy. The evaluation methodology is treated in chapter 3. Finally in chapter 4 we present experimental results using speech data of 3 languages US-English, Spanish and French.

2 Conditional Gaussians

2.1 Decomposition of Gaussians

Based on the properties of multivariate Gaussians [5] a Gaussian $N(Z;\mu_z,V_z)$ can be decomposed by $N(Z;\mu_z,V_z) \equiv p(Z_2)\cdot p(Z_1|Z_2)$ with Gaussians:

$$p(Z_2) = N(Z_2;\mu_{Z_2},V_{Z_2}); p(Z_1|Z_2) = N(Z_1;\mu_{Z_1|Z_2},V_{Z_1|Z_2});\ Z=[Z_1,Z_2]^T \quad (3)$$

The means and covariance matrices of the Gaussians (3) are given by

$$\left.\begin{aligned} V_z \equiv \begin{pmatrix} V_{Z_{11}} & V_{Z_{12}} \\ V_{Z_{21}} & V_{Z_{22}} \end{pmatrix};\ V_Z^{-1} \equiv A_Z \equiv \begin{pmatrix} A_{Z_{11}} & A_{Z_{12}} \\ A_{Z_{21}} & A_{Z_{22}} \end{pmatrix};\ \mu_z = \begin{pmatrix} \mu_{Z_1} \\ \mu_{Z_2} \end{pmatrix} \\ V_{Z_2}^{-1} = A_{Z_{22}} - A_{Z_{21}}A_{Z_{11}}^{-1}A_{Z_{12}} \\ \mu_{Z_1|Z_2} = \mu_{Z_1} - A_{Z_{11}}^{-1}A_{Z_{12}}(Z_2-\mu_{Z_2});\ \ {V_{Z_1|Z_2}}^{-1} = A_{Z_{11}} \end{aligned}\right\} \quad (4)$$

The dimension of the sub-means and sub-covariance matrices in (4) are related to the dimensions of the vectors Z_1 and Z_2. Now we apply (3), (4) to $N(Z;\mu_z,V_z) \equiv N(\vec{X}_l;\vec{\mu}_{ikl},\vec{V}_l)$.

$$\left.\begin{aligned} Z_1 \equiv X_l;\ Z_2 \equiv \vec{X}_{l,l-1} \equiv [X_{l-1},\dots,X_1]^T; \mu_z \equiv \vec{\mu}_{ikl};\ \vec{\mu}_{ikl} \equiv \vec{\mu}_{ikll} \equiv [\mu_{ikl1},\dots.,\mu_{ikll}]^T \\ V_Z \equiv \vec{V}_l; \vec{V}_l \equiv \vec{V}_{ll} \equiv \begin{pmatrix} V_l^{X_lX_l} & V_l^{X_l\vec{X}_{l,l-1}} \\ V_l^{\vec{X}_{l,l-1}X_l} & V_l^{\vec{X}_{l,l-1}\vec{X}_{l,l-1}} \end{pmatrix}, {\vec{V}_l}^{-1} \equiv \begin{pmatrix} A_l^{X_lX_l} & A_l^{X_l,\vec{X}_{l,l-1}} \\ A_l^{\vec{X}_{l,l-1},X_l} & A_l^{\vec{X}_l,\vec{X}_{l,l-1}} \end{pmatrix} \end{aligned}\right\} \quad (5)$$

This results in a decomposition

$$\left.\begin{aligned} N(\vec{X}_l;\vec{\mu}_{ikl},\vec{V}_l) = N_{ll}(X_l|\vec{X}_{l,l-1};\mu_{ikl|l},V_{l|l})\ N_{l-1}(\vec{X}_{l,l-1};\vec{\mu}_{ikl,l-1},\vec{V}_{l,l-1}) \\ \vec{\mu}_{ikl,l-1} \equiv [\mu_{ikl1},\dots,\mu_{ikl,l-1}]^T; \mu_{ikl|l} = \mu_{ikll} - A_l^{X_lX_l{}^{-1}}A_l^{X_l\vec{X}_{l-1}}(\vec{X}_{l,l-1}-\vec{\mu}_{ikl,l-1}) \\ {V_{l|l}}^{-1} = A_l^{X_lX_l};\ \vec{V}_{l,l-1}^{-1} = A_l^{\vec{X}_{l-1},\vec{X}_{l-1}} - A_l^{\vec{X}_{l-1},X_l}(A_l^{X_lX_l})^{-1}A_l^{X_l,\vec{X}_{l-1}} \end{aligned}\right\} \quad (6)$$

We continue the decomposing procedure of $N(\vec{X}_l;\vec{\mu}_{ikl},\vec{V}_l)$ by splitting $p(Z_2) \equiv N_{l-1}(\vec{X}_{l,l-1};\vec{\mu}_{ikl,l-1},\vec{V}_{l,l-1})$ again into two conditional Gaussians by setting $Z_1 \equiv X_{l-1}; Z_2 \equiv \vec{X}_{l,l-2}, Z=[Z_1,Z_2]^T; \mu_z = \vec{\mu}_{ikl,l-1}; V_z = \vec{V}_{l,l-1}$.The resulting Gaussians $p(Z_1|Z_2)$ and $p(Z_2)$ have the same structure as given by (6) but with different means and covariance matrices. The decomposing procedure can be continued by splitting the resulting $p(Z_2)$ again. This procedure can be repeated till $Z_2 \equiv X_1$. Thus this procedure leads to a recursive scheme to determine the means and covariance matrices of the conditioned Gaussians N_{lv} ; $v=l,\dots,1$. The scheme is performed in 3 steps: initialization, iteration and finalization.

$$\left.\begin{aligned} \textbf{\textit{Initialisation}}\text{: } Z \equiv \vec{X}_l \equiv [X_l,\dots,X_1]^T;\ Z \equiv [Z_1,Z_2]^T;\ Z_1 \equiv X_l;\ Z_2 \equiv \vec{X}_{l,l-1} \equiv [X_{l-1},\dots,X_1]^T \\ V_Z \equiv \vec{V}_l \equiv \vec{V}_{ll} \equiv \langle(\vec{X}_l-\vec{\mu}_{ikl})(\vec{X}_l-\vec{\mu}_{ikl})^T\rangle;\ \mu_z \equiv \vec{\mu}_{ikl} \equiv \vec{\mu}_{ikll} \equiv [\mu_{i,k,l,l},\dots,\mu_{i,k,l,1}]^T \end{aligned}\right\} \quad (7)$$

Iteration: for $\nu = l, \dots, 2$: $\mathrm{Z}_1 \equiv \mathrm{X}_\nu$; $\mathrm{Z}_2 \equiv \vec{X}_{l,\nu-1} \equiv [\mathrm{X}_{\nu-1}, \dots, \mathrm{X}_1]^{\mathrm{T}}$

$$\left.\begin{aligned} \vec{V}_{l\nu}{}^{-1} &\equiv \begin{pmatrix} A_l^{X_\nu X_\nu} & A_l^{X_\nu, \vec{X}_{l,\nu-1}} \\ A_l^{\vec{X}_{l,\nu-1}, X_\nu} & A_l^{\vec{X}_{l,\nu-1}\vec{X}_{l,\nu-1}} \end{pmatrix}; \left(\vec{V}_{l,\nu-1}\right)^{-1} \equiv A_l^{\vec{X}_{l,\nu-1}, \vec{X}_{l,\nu-1}} - A_l^{\vec{X}_{l,\nu-1}, X_\nu} \left(A_l^{X_\nu X_\nu}\right)^{-1} A_l^{X_\nu \vec{X}_{l,\nu-1}} \\ \mu_{ikl|\nu} &= \mu_{ikl\nu} - \left(A_l^{X_\nu X_\nu}\right)^{-1} A_l^{X_\nu \vec{X}_{l,\nu-1}} \left(\vec{X}_{l,\nu-1} - \vec{\mu}_{ikl,\nu-1}\right);\ \vec{\mu}_{ikl,\nu-1} \equiv [\mu_{ikl,\nu-1}, \dots., \mu_{ikl,1}]^T \\ p(Z_1|Z_2) &\equiv N_{l\nu}\left(X_\nu \middle| \vec{X}_{l,\nu-1};\ \mu_{ikl|\nu}, V_{l|\nu}\right); V_{l|\nu} = \left(A_l^{X_\nu X_\nu}\right)^{-1}; p(Z_2) \equiv N_\nu\left(\vec{X}_{l,\nu-1}; \vec{\mu}_{ikl,\nu-1}, \vec{V}_{l,\nu-1}\right) \end{aligned}\right\} \quad (8)$$

Finalization: for $\nu = 2$; $\mathrm{Z}_1 \equiv \mathrm{X}_2$; $\mathrm{Z}_2 = \vec{X}_{l,1} \equiv \mathrm{X}_1$; $\vec{\mu}_{ikl|1} \equiv \mu_{ikl1}$

$$\left.\begin{aligned} \vec{V}_{l2}{}^{-1} &\equiv \begin{pmatrix} A_l^{X_2 X_2} & A_l^{X_2, X_1} \\ A_l^{X_1, X_2} & A_l^{X_1, X_1} \end{pmatrix}; \left(\vec{V}_{l1}\right)^{-1} \equiv A_l^{X_1, X_1} - A_l^{X_1, X_2} \left(A_l^{X_2 X_2}\right)^{-1} A_l^{X_2 X_1} \\ p(Z_2) &\equiv N_{l1}\left(X_1; \mu_{ikl|1}, V_{l|1}\right);\ V_{l|1} \equiv \vec{V}_{l1} \end{aligned}\right\} \quad (9)$$

Thus we have derived a scheme to split $N\left(\vec{X}_l; \vec{\mu}_{ikl}, \vec{V}_l\right)$ into TEPs $N_{l\nu}\left(X_\nu \middle| \vec{X}_{l,\nu-1};\ \mu_{ikl|\nu}, V_{l|\nu}\right)$ for $\nu = 1, \dots, l$ leading to the decomposition (2). According to (8) the conditional means $\mu_{i,k,l|\nu}, \nu = 1, \dots, l$ depend on the feature vector $X_\nu, \dots, X_1$. Thus at frame ν the conditional means $\mu_{ikl|\nu}$ can be processed online without delay, as all feature vectors are known. This property allows a frame by frame processing as needed for online recognition system starting with the processing of the TEP $N_{l1}\left(X_1; \mu_{ikl|1}, V_{l|1}\right)$. The covariance matrices and their inverse (7-9) can be processed offline, as they depend from $\vec{V}_l$ only. Yet the values of the GMMs (1) have to be computed delayed at frame $\nu = l$, when all TEPs are available. The impact of delay together with different possible length of the chunks has to be implemented in the search process of an online recognizer.

2.2 Extension of Gaussians

We assume that the parameters of the Gaussians $N\left(\vec{X}_l; \vec{\mu}_{ikl}, \vec{V}_l\right)$ can be estimated reliable for a length $l \le m_0$. In this section we derive a scheme to extend the Gaussians for $l > m_0$.

The matrix $\vec{V}_l$ is composed by covariance elements $V_l^{X_\nu X_{\nu'}}$, which denotes the covariance matrix of the feature vectors $X_\nu, X_{\nu'}$ i.e. of feature vectors with different position ν within a chunk $\vec{X}_l$. $\vec{V}_l$ is composed by these elements in the form

$$\vec{V}_l = \begin{pmatrix} V_l^{X_l X_l} & V_l^{X_l X_{l-1}} \dots V_l^{X_l X_2} & V_l^{X_l X_1} \\ & \vdots & \\ V_l^{X_1 X_l} & V_l^{X_1 X_{l-1}} \dots V_l^{X_1 X_2} & V_l^{X_1 X_1} \end{pmatrix} \quad (10)$$

As the covariance elements $\mathrm{V}_\mathrm{l}^{\mathrm{X_\nu X_{\nu'}}}$ are known only for $\nu, \nu' \le m_0$, we have to find an approximation for the missing elements. Our approach is based on the idea that the correlation between feature vectors vanishes for increasing distance and that feature vectors with the same distance in time given by $|\nu - \nu'|$ have the same correlation as the feature vector X_{m_0}. This leads to the approximation:

$$\tilde{V}_l^{X_\nu X_{\nu'}} \equiv \left.\begin{cases} V_l^{X_\nu X_{\nu'}} & for\ \nu, \nu' \le m_0 \\ \left.\begin{cases} f_{|\nu-\nu'|} \cdot V_{m_0}^{X_{m_0} X_{m_0-|\nu-\nu'|}} & for\ \nu \ge \nu' \\ f_{|\nu-\nu'|} \cdot V_{m_0}^{X_{m_0-|\nu-\nu'|} X_{m_0}} & for\ \nu < \nu' \end{cases}\right\} for\ |\nu - \nu'| < m_0 \\ 0 \qquad for\ |\nu - \nu'| \ge m_0 \end{cases}\right\} for\ \nu\ or\ \nu' > m_0 \quad (11)$$

$$f_{|\nu-\nu'|} = \begin{cases} 1 & for\ \nu = \nu' \\ f_0 < 1 & for\ |\nu - \nu'| > 0 \end{cases}$$

In (11) an extension factor $f_{|\nu-\nu'|}$ is introduced. A value $f_{|\nu-\nu'|} = 1$ would be consistent with our approach. But it turned out that the value $f_{|\nu-\nu'|} = 1$ leads to matrices $\vec{\tilde{V}}_l$ with negative

determinant. Besides the extension of the matrix (10) we have to extend the means for $l>m_0$. Here we assume, that the trajectory gets stationary for $l>m_0$. Thus the sequence $\vec{\mu}_{i,k,l} = \left[\mu_{i,k,l,1}, \ldots, \mu_{i,k,l,l}\right]$ of the mean vectors $\mu_{i,k,\nu,l}$ is continued with the means estimated for $l=m_0$

$$\mu_{i,k,\nu,l} = \begin{pmatrix} \mu_{i,k,\nu,m_0} \; for \; \nu \leq m_0 \\ \mu_{i,k,m_0,m_0} \text{ for } \nu > m_0 \end{pmatrix} l > m_0 \tag{12}$$

When l exceed a certain value l_0- occurring often for 'non speech' segments – we follow a l_0-gram approach. Thus we limit the the statistic dependencies of the feature vectors till an order of l_0 leading to the approximation

$$N_{l\nu}(X_\nu|X_{\nu-1}, \ldots, X_1; \mu_{i,k,l|\nu-1}, V_{l|\nu}) \approx \left\{ \begin{matrix} N_{l_0\nu}(X_\nu|X_{\nu-1}, \ldots, X_1; \mu_{i,k,l_0|\nu-1}, V_{l_0|\nu}) \text{ for } \nu \leq l_0 \\ N_{l_0 l_0}(X_\nu|X_{\nu-1}, \ldots, X_{\nu+1-l_0}; \mu_{i,k,l_0|l_0}, V_{l_0|l_0} \text{ for } \nu > l_0 \end{matrix} \right\} l > l_0 \tag{13}$$

3 Evaluation

3.1 Classification of Segments and Phonemes

We assume that the boundaries of each clustered tri-phone Ph_c including the boundaries of its segments Q_i are known. We denote the 3 segments composing Ph_c by $Q_{i(c,p)}, p = 1,2,3$ and the given boundary information by $\vec{S} \equiv \{l_p, p = 1,2,3\}$. For classification of phonemes Ph_j we need an acoustic model $p(\vec{X}_t|Ph_j, \vec{S})$. We first define an acoustic model $p_t(\vec{X}_t|Ph_c, t)$ for each clustered tri-phone Ph_c given by

$$p_t(\vec{X}_t|Ph_c) = \prod_{p=1}^{3} p_{l_p}\left(\vec{X}_{l_p}\middle|Q_{i(c,p)}\right); t = \sum_{p=1}^{3} l_p \; ; \; \vec{X}_t = [\vec{X}_{l_1}, \ldots, \vec{X}_{l_3}]^T \tag{14}$$

The model (14) assumes that the chunks $\vec{X}_{l_p}$ for different segments are statistic independent. This assumption is quite crude and corresponds to the statistic independence assumption of feature vectors on frame level of HMMs. Given a set of phonemes Ph_j, j=1,...,N_{Ph} we define sets $C(j)$ containing all indices of clustered tri-phones Ph_c having Ph_j as central phoneme. This leads to a context independent acoustic model for phonemes:

$$p(\vec{X}_t|Ph_j) = \sum_{j=1}^{N_{Ph}} \left[P(Ph_c|Ph_j, t) \prod_{p=1}^{3} p_{l_p}\left(\vec{X}_{l_p}\middle|Q_{i(c,p)}\right)\right]_{c \in C(j)} \tag{15}$$

For classification of phonetic units $PU \in \{Q, Ph\}$ we use a maximum likelihood classifier

$$\widehat{PU} = argmax_j\left[P(PU_j|t)p(\vec{X}_t|PU_j, \vec{S})\right] \tag{16}$$

We use as a-priori probability the duration dependent unit-monogram $P(PU_j|l)$, which can be estimated directly from databases. This approach reduces the dependencies of error rates from the syntactic and semantic restrictions given by a language. Thus we focus the evaluation on the quality of the acoustic models. Given the acoustic models (1) and (15) and the ML-classifier (16) we get

$$\hat{Q} = argmax_i\left[P(Q_i|l)p_l(\vec{X}_l|Q_i)\right] \tag{17}$$

$$\widehat{Ph} = argmax_j\left[P(Ph_j|t)p(\vec{X}_t|Ph_j)\right] \tag{18}$$

3.2 Entropy

Based on Shannon's theory on entropy [6] relations between error rates and Shannon's conditional Entropy can be derived [7,8]. Given phonetic units PU_j ,j=1,...,N_{PU} realized by chunks $\vec{X}_t$ of length t Shannon' conditional Entropy $H_t(PU|\vec{X}_t)$ is defined by

$$\left.\begin{matrix} H_t(PU|\vec{X}_t) \equiv H_t(PU|t) - I_t(\vec{X}_t; PU); \; I_t(\vec{X}_t; PU) \equiv H_t(\vec{X}_t) - H_t(\vec{X}_t|PU) \\ H_t(PU|t) \equiv -\sum_{j=1}^{N_{PU}} P(PU_j|t) \log(P(PU_i|t)) \; ; \; H_t(\vec{X}_t) \equiv -\int p(\vec{X}_t) log \; p(\vec{X}_t) \, d\vec{X}_t \\ H_t(\vec{X}_t|PU) \equiv -\sum_{j=1}^{N_{PU}} P(PU_j|t) \int p_t(\vec{X}_t|PU_j) \; log \; p_t(\vec{X}_t|PU_i) \, d\vec{X}_t \end{matrix}\right\} \tag{19}$$

$H_t(PU|t)$ is the information needed to recognize the PUs without error from chunks of length t. The mutual information $I_t(\vec{X}_t; PU)$ is the information gained from the chunks $\vec{X}_t$. Whenever

the relation $H_t(PU) > I_t(\vec{X}_t; PU)$ holds, errors occur. As the correct distribution $p_t(\vec{X}_t|PU_i)$ is unknown, we use approximations as $\tilde{p}_l(\vec{X}_l|Q_i), \tilde{p}_l(\vec{X}_l)$ given by the GMMs (1). We determine the entities defined in (19) by the Monte Carlo Method [9]. For example the entity $H_l(\vec{X}_l|Q)$ is evaluated by $N_{S(i,l)}$ samples of chunks $\vec{X}_l^n$ assigned to the segment Q_i using

$$-\int p_l(\vec{X}_l|\,Q_i)\; log(p_l(\vec{X}_l|Q_i)\; d\vec{X}_t \approx -\frac{1}{N_{S(i,l)}}\sum_{n=1}^{N_{S(i,l)}} log\; \tilde{p}_l(\vec{X}_l^n|Q_i) \qquad (20)$$

To our knowledge there exists no theory, which predicts exactly the error rates given $H_l(PU|\vec{X}_t)$. Yet there exist bounds for the error rates, which depend only on the number N_{PU} of units to be recognized. We regard as upper bound the Fano bound [10] and we regard as lower bound the Golic bound [11]. In the next chapter we present Fano-Golic plots, which relate the error rates to Shannon's entropy. Bad approximation of an acoustical model is indicated when the measured points are close or outside the Fano-Golic bounds.

3.3 Variance Analysis of TEPs

The covariance matrices $V_{l|\nu}, \nu = 1, \dots, l$ iare a measure for the variation of the feature vectors along the trajectory. The conditional Gaussians can be written in the form

$$N_{l\nu}(X_\nu|\vec{X}_{l,\nu-1};\, \mu_{ikl|\nu}, V_{l|\nu}) \equiv \frac{1}{(2\,\pi\Delta_{l\nu})^{\frac{D}{2}}} e^{-\frac{1}{2}(X_\nu-\mu)^T A(X_\nu-\mu)};\; A \equiv {V_{l|\nu}}^{-1};\; \mu \equiv \mu_{ikl|\nu} \text{ with } \Delta_{l\nu} \equiv \det(V_{l|\nu})^{\frac{1}{D}} \qquad (21)$$

(D denotes the dimension of the feature vector X_ν). The entity $\Delta_{l\nu}$ can be interpreted as the average variance of the vector components of X_ν. (For example for a diagonal matrix $V_{l|\nu} = \sigma^2 I$ with dim(I)=$D \cdot D$ we get $\Delta_{l\nu} = \sigma^2$). For correlated feature vectors, $\Delta_{l\nu}$ should decrease with ν, as the statistical dependency increases with increasing length of vector $\vec{X}_{l,\nu}$.

4 Experimental Results

4.1 Experimental Set-Up

Our experiments are performed with Spanish and French speech databases of broadcast news, conversations and podcast downloaded from various internet sources. The databases were developed during the QUAERO project and used during for ASR evaluation [12]. The labeling into segments was performed by the HMM training system [13] using 3 or 6 state right to left HMMs. The segments are constructed by clustering tri-phones using CART under the condition that the central phoneme is not tied. We also use results [3] derived from an American English Database of in car recorded speech according to the SpeechDat recommendations [14]. To compare results from all 3 databases the number of segments is chosen to be in the order of 600 for all 3 languages. The feature vectors are derived from LDA transformed augmented MFCCs. For labeling the Spanish database a 3-2-HMM with tied emission probabilities was used, where each of the 3 segments of the clustered tri-phones has to be assigned at least to one feature vector. For the French database a 3-1-HMM was used allowing skipping of segments. For training the GMMs (1) we collect from the labeled data for each labeled segments Q_i all the aligned chunks $\vec{X}_l$ of length l. Given those sets of chunks the HCMs were trained using the EM-algorithm in its unified form [4, chapter 7] without changing the boundaries of the segments. The size of the databases and the length distribution shown in tab.1 allows to train HCMs till l=3 for Spanish and to train HCMs till l=6 for French. Yet for French for the HCM for l=6 the determinant of the covariance matrix $\vec{V}_6$ became negative due to limited numerical accuracy given by MATLAB-R2010b. Thus for French, extensions (11) for $l > m_0$=5 has to be applied. As shown in tab. 1 most chunks have the length l=2. More than 92% of the segments have a length l less than 6. Compared to US-English and Spanish, the French database has more longer segments (about 20% for l>3).

In the following we use as *log*-function the base 2. Thus the entropies have as units *bit*. In table 2 the distribution $P(Q_i|l)$, derived from counting the chunks of length l assigned to Q_i, is

characterized by the entropies $H_l(Q)$. For equal distributed segments $H_l(Q)$ would take the value $log_2\ N_Q$ (e.g. log_2 607=9.25 [bit]). Table 2 shows smaller values than for equal distribution $P(Q_i|l)$.

Speech database	frame shift	#chunks for training	#chunks for test	$P(l\|Q)$ *in % of training data;* l:					
				1	2	3	4	5	≥ 6
US-English	15ms	33.879.857	-	29	46	25	n.a.*	n.a	n.a.
Spanish	10ms	4.529.470	3.760.799	26.3	62.5	6.4	1.9	0.8	2.2
French	10ms	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 (*n.a. = not available)

Speech database	# of segments N_Q	l / $H_l(Q)[bit]$					
		1	2	3	4	5	6
US-English	607	7.78	8.35	7.37	n.a.	n.a.	n.a.
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 $H_l(Q)$

4.2 Error Rates and Shannon's Conditional Entropy

The segment error rates are determined by evaluating (15). Due to table 3 the *SER*s drop with increasing l till l=3. For larger l the *SER*s tend to be stable till $l=m_0$ is achieved. Afterwards *SER* increases. This result is discussed in more detail in section 4.3.

Speech database	# of Modes per HCM	$SER;\ l$:						m_0	f_0
		1	2	3	4	5	6		
US-English	607	83.9	71.1	49.8	n.a.	n.a.	n.a.	3	-
Spanish	604	83.4	71.2	57.5	62.2	66.9	39.8	3	0.3
French-5	598	75.1	62.1	57.5	56.2	56.7	61.7	5	0
French-3	598	75.1	62.1	57.5	62.9	73.7	79.0	3	0

Table 3 - Segment Error Rate (SER) for mono-modal HCMs; for US-English a diagonal covariance matrix was chosen; For French and Spanish the HCMs were extended according to (14) till l_0=6.

Fig. 1 shows Fano-Golic plots for the segments ($PU_i = Q_i$), where $H(Q|\vec{X}_l)$ is determined or outside the Fano bound. For Spanish with m_0=3 for l=4 the point is on the border and for

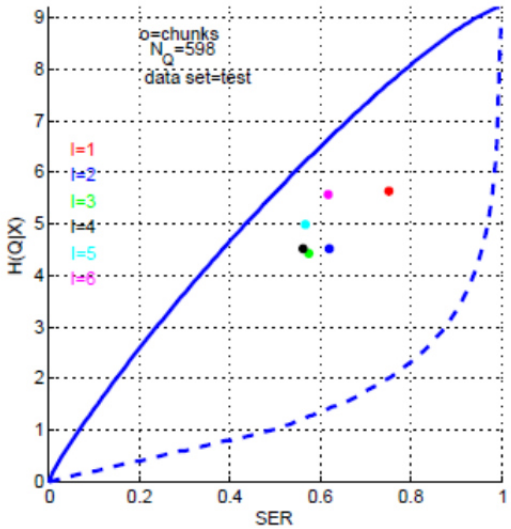

Figure 1.a - French: #modes= 46 101, m_0=5 , l_0=6

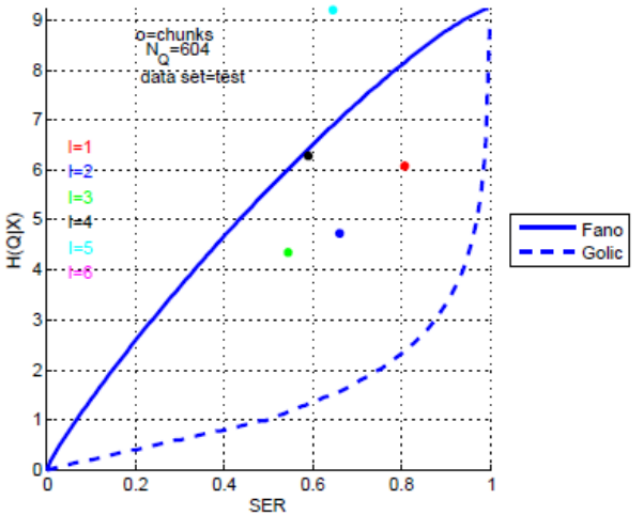

Figure 1.b - Spanish: #modes=50 000, m_0=3 , l_0=6

for l=6 outside the Fano bound. For French with m_0=5 for l=6 the point is outside of the plot. These results indicates the poor acoustic model of the extension approach (11). Table 4 shows the phoneme error rates and related mutual information for Spanish and French for different

number of modes. The mutual information for French is negative indicating a poor acoustic model (17).

Language\#modes	Mono-Modes	10000	50000	N_{Ph}	$H(Ph)$
ES-PER\I(Ph;X)	33.4/3.74	28.7/4.48	27.3/4.8	32	n.a.
FR-PER\I(Ph;X)	70.7/-10.32	n.a.	68.2/-11.6	40	4.72

Table 4 - Phoneme error rates (*PER*) and mutual information $I(Ph|X)$

4.3 Analysis of Decomposition

Table 5 and 6 show the values of $\Delta_{l\nu}$ (21) for the TEPs $N_{l\nu}$ and for $\vec{V}_l$ for different extension configurations for French. To avoid negative determinants the extension value f_0 (11) is set to 0. As expected $\Delta_{l\nu}$ of the TEPs decreases for increasing ν . This decrease is observed till $\nu \leq m_0$. For $\nu > m_0$, $\Delta_{l\nu}$ takes values as for $\nu = 1$, as for $f_0 = 0$ the feature vector X_ν is modeled to be statistic independent from the feature vectors $\vec{X}_{\nu,\nu-1}$. In analogy to (21) the values Δ_l is defined for the complete covariance matrix $\vec{V}_l$. $\vec{V}_l$ decrease with increasing l indicating, that trajectories with increasing l show less variations.

m_0=3 l_0=6 f_0=0

	$\Delta_{l\nu}$						Δ_l
l\nue	1	2	3	4	5	6	
1	1.00						1.00
2	0.97	0.28					0.52
3	0.93	0.26	0.16				0.34
4	0.93	0.26	0.16	0.94			0.33
5	0.93	0.26	0.16	0.94	0.94		0.32
6	0.93	0.26	0.16	0.94	0.94	0.94	0.32

Table 5a-$\Delta_{l\nu}$,Δ_l for mono-mode HCMs

m_0=3 l_0=6 f_0=0

	$\Delta_{l\nu}$						Δ_l
l\nue	1	2	3	4	5	6	
1	0.71						0.71
2	0.73	0.23					0.41
3	0.83	0.24	0.15				0.31
4	0.83	0.24	0.15	0.84			0.29
5	0.83	0.24	0.15	0.84	0.84		0.27
6	0.83	0.24	0.15	0.84	0.84	0.84	0.26

Table 5b - $\Delta_{l\nu}$,Δ_l for HCMs with 50 000 modes

m_0=5 l_0=6 f_0=0

	$\Delta_{l\nu}$						Δ_l
l\nue	1	2	3	4	5	6	
1	1.00						1.00
2	0.97	0.28					0.52
3	0.93	0.26	0.16				0.34
4	0.91	0.23	0.15	0.08			0.22
5	0.92	0.25	0.2	0.13	0.11		0.13
6	0.91	0.22	0.14	0.07	0.02	0.92	0.13

Table 6a-$\Delta_{l\nu}$, $\Delta_l$for mono-mode HCMs

m_0=5 l_0=6 f_0=0

	$\Delta_{l\nu}$						Δ_l
l\nue	1	2	3	4	5	6	
1	0.71						0.71
2	0.73	0.23					0.41
3	0.83	0.24	0.15				0.31
4	0.84	0.22	0.14	0.08			0.21
5	0.85	0.21	0.13	0.07	0.02		0.13
6	0.85	0.21	0.13	0.07	0.02	0.86	0.13

Table 6b - $\Delta_{l\nu}$,Δ_l for HCMs with 50 000 modes

# of Modes		$SER;\ l$:							m_0	f_0
	1	2	3	4	5	6	6 >	*PER*		
1.794	78,2	63,6	58,8	63,6	69,6	78,9	80,5	67,9	3	0
2.290	78,2	63,6	58,8	57,9	59,0	77,5	82,9	69,6	5	0
37.818	73,4	56,5	55,4	61,9	68,1	77,5	79,0	66,4	3	0
46.101	72,2	55,8	55,0	55,6	56,7	80,8	87,9	68,2	5	0

Table 7 - *SER* and *PER* for different extension configurations and modes

The related *SER* and *PER* values are shown in tab.7 . Increasing number of modes decreases *SER* for $\nu \leq m_0$. For $\nu > m_0$ *SER* gets worth caused by crude extension approach (11),(13).

5 Acknowledgements

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

6 Conclusion

We have presented the decomposition of chunk related Gaussians into conditional Gaussians allowing a frame based processing of the TEPs and allowing to handle arbitrary long chunks. Numerical problems in estimating covariance matrices of long chunks are observed. Perhaps the HCM training can done for conditional Gaussians using a similar decomposition approach as (8) avoiding the inversion of high dimensional matrices. The used extension strategy leads to poor TEPs as the extended covariance matrices are not realistic covariance matrices. Here a better extension model is needed. Perhaps an ARMA model for feature vectors of a chunk can be applied. The high phoneme error rates for French has to be investigated further. It seems that the used segments are not appropriate to model trajectories for phonemes.

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] Höge, H.: Estimating an upper Bound for the Error Rate for Speech Recognition using Entropy. Int. J. of Electronics and Communications Vol.53: 1999.

[8] Höge, H., Setiawan, P.: Shannon’s Conditional Entropy and Error Rates on Phone Level. In: Lacroix, A. (Ed.): Beiträge zur Signaltheorie, Signalverarbeitung, Sprachakustik und Elektroakustik - Dietrich Wolf zum 80. Geburtstag, Studientexte zur Sprachkommunikation, Vol. 52,: TUDpress-Verlag: Dresden 2009.

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

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

[11] 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

[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] 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.

[14] Moreno, A., B. Lindberg, C. Draxler, G. Richard, K. Choukri, J. Allen and S. Euler:'SpeechDat-Car: A Large Speech Database for Automotive Environments'. In Proc. International Conference on Language Resources and Evaluation (LREC), June 2000.