THE IMPACT OF CORRELATED FEATURES IN SPEECH RECOGNITION

Harald Höge

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

Abstract: Hidden Markov Models assume that adjacent feature vectors are statistically independent. Yet the use Δ and ΔΔ operations, super-vectors and LDA leads to highly correlated feature vectors contradicting the independence assumptions. Experiments [1] have shown, that more sophistical acoustical models lead to no substantial decrease in error rate. In order to investigate these findings we use simulated feature vectors having probability distributions similar as derived from real speech data. The used distribution model exactly the statistical properties of adjacent feature vectors. We made recognition experiments on 607 segments derived from tri-phones, which were realized by two feature vectors. The experiments confirm that the use of second order statistics does not improve the recognition rate substantially. Further we can show with simulated features that the error rate decrease with decreasing degree of correlation.

1 Introduction

Hidden Markov Models assume that adjacent feature vectors are statistically independent. Yet the use Δ and ΔΔ operations, super-vectors and LDA leads to highly correlated feature vectors contradicting the independence assumptions. Further Hidden Markov Models assume that within a state the features are distributed identically. These drawbacks are called the i.i.d (**i**ndependently **d**istributed, **i**dentically **d**istributed) problem. To overcome the i.i.d. problem sophisticated segment models [2] and recently trajectory models [1] inspired from speech synthesis [3] have been investigated. The results shown in [1] lead to the conclusion that acoustical models using different distributions within a state improve recognition performance significantly, but no significant improvements could be achieved by using acoustical models modeling statistical dependency. In order to understand the latter finding, in this paper experiments with simulated features generated by the Monte Carlo Method [4] were performed. The Monte Carlo Method allows to generate features with different degree of correlation between feature vectors given by predefined distribution function of the features. We experiment with distribution functions, which are derived from the emission probabilities of a HMM based recognition system trained on a large US-English database. In order to study the impact of correlation we evaluate the error rate on segments derived from tri-phones. The segments are modeled by chunks as described in [6] (see also chapter 2). The paper is organized as follows:

- Chapter 2 presents the statistical framework describing the concept of chunks, the used recognizer, the evaluation based on error rates and Shannon's entropy, and the application of the Monte Carlo Method .
- Chapter 3 describes, how the feature vectors are simulated with predefined distributions.
- Chapter 4 shows the experimental result, which are concluded in chapter 5.

2 The Statistical Framework

The concept of chunks [6] is derived from HMMs, where tri-phones are used as the basic phonetic units. The tri-phones are build up by three **segments**, which can be interpreted as the onset, middle and offset of a phone. In HMM technology each segment is modeled by a state. The acoustical model of a segment Q_i is defined by

$$p^{HMM}(\vec{X}_{n,l}|Q_i) = a_{ii}^l \prod_{v=1}^l p(X_{n+v-1}|Q_i);\ \vec{X}_{n,l} \equiv (X_n, \dots, X_{n-v+1}, \dots, X_{n-l+1}) \quad (1)$$

The duration of a segment is modeled by l and a_{ii}, whereas $p(X|Q_i)$ - the emission probabilities - describe the statistic properties of a segment independent of its duration.

In the following we define as a chunk a sequence of feature vectors $\vec{X}_{n,l}$ as defined in (1) assigned to a segment Q_i. The index l denotes the **length (**duration) of a chunk. The index ν denotes the **position** of a feature vector within a chunk. Given these notations we define an acoustic model for segments:

$$p_l(\vec{X}_{n,l}|Q_i) = p_{ll}(X_{n-l+1}|Q_i) \prod_{v=1}^{l-1} p_{vl}(X_{n-v+1}\,|X_{n-v}, \dots, X_{n-l+1}, Q_i)$$

We call this model a trajectory model, because along the trajectory of a segment given by a chunk of length l for each position v and each length l a specific distribution $p_{vl}(X_{n-v+1}|*,Q_i)$ is assigned. Compared to (1) these distributions generalize the emission probabilities and are called 'Trajectory Emission Probabilities - TEPs'. The TEPs are approximated by First Order (FO) distributions $p_{vl}(X_{n-v+1}|Q_i)$ and Second Order (SO) distributions $p_{vl}(X_{n-v+1}|X_{n-v}, Q_i)$ leading to the FO and SO trajectory model:

$$p_l^{FO}(\vec{X}_{n,l}|Q_i) \equiv \prod_{v=1}^l p_{vl}(X_{n-v+1}\,|Q_i) \quad (2)$$

$$p_l^{SO}(\vec{X}_{n,l}|Q_i) \equiv p_{ll}(X_{n-l+1}|Q_i) \prod_{v=1}^{l-1} p_{vl}(X_{n-v+1}\,|X_{n-v}, Q_i) \quad (3)$$

The evaluation of the two trajectory models is performed in recognizing segments Q_i. In section 2.1 the recognizer and the segment error rate are defined. Section 2.2 defines Shannon's Entropy, which is linked to the error rate by bounds. Finally in section 2.3 the application of the Monte Carlo method is sketched.

2.1 Recognizer and Segment Error Rates

To determine segment error rates (SER) we describe in the following recognizers based on first order trajectory model (2). For second order models (3) the recognizers are constructed equivalently. We define MAP recognizers operating on a complete chunk $\vec{X}_{n,l}$. The segment - recognizer is defined by

$$\hat{Q}_l^{FO}(\vec{X}_{n,l}) = argmax_j p_l^{FO}(\vec{X}_{n,l}|Q_j)P_l(Q_j)\,; \vec{X}_{n,l} \in \vec{X}_l(Q_i) \quad (4)$$

operating on chunks $\vec{X}_{n,l}$ of the set $\vec{X}_l(Q_i)$. This set contains all chunks of length l assigned to Q_i. The probability $P_l(Q_i)$ is defined of the occurrence probability of the segment Q_i represented by chunks of length l. If the output of the recognizer (4) is not Q_i the recognizer makes an error. The segment error rate of the recognizer using first order trajectory model is defined by

$$SER_l^{FO} = \int (1 - max_j(\,p_l^{FO}(\vec{X}_{n,l}|Q_i)P_l(Q_j))\,)\ d\vec{X}_{n,l} \quad (5)$$

In chapter 4 expression (5) is evaluated by the relation

$$SER_l^{FO} = \sum_{i=1}^{N_Q} P_l(Q_i)P(\hat{Q}_l^{FO}(\vec{X}_{n,l}) \neq Q_i)$$

2.2 Shannon's Conditional Entropy

A recognizer can be seen as a decoder of information sent via a noisy channel. In our case the decoder receives chunks and decodes the chunks to the index i of a segment Q_i. Shannon's conditional entropy $H_l(Q|\vec{X}_l)$ [7] determines the number of bits missing to decode the chunks without errors. $H_l(Q|\vec{X}_l)$ is defined by

$$H_l(Q|\vec{X}_l) \equiv -\sum_i \int p(Q_i, \vec{X}_l) ld\, p(Q_i|\vec{X}_l) d\vec{X}_l \ \ ;\ \vec{X}_l \in \vec{X}_l(Q_i) \quad (6)$$

The application of Shannon's entropy to chunks is described in [6]. Given Shannon's entropy, the lower and upper bounds of the segment error rates are given by the Fano [12] and the Golić [13] bound. In chapter 4 plots with these bounds are presented.

2.3 The Monte Carlo Method

We use the Monte Carlo Method [9] to evaluate the segment error rates (5) and Shannon's entropy (6). The error rates and the entropies are defined as an integral $\int f(Z)p(Z)dZ$, where *Z* denotes a random variable with distribution *p(Z)* and *f(Z)* is an operator on *Z*. Usually this integral cannot be evaluated analytically due to the high dimension of *Z*, which is in our case in the order of 50-100. The Monte Carlo Method uses the relation

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} f(Z_i)}{n} = E\big(f(Z)\big) = \int f(Z)p(Z)dZ \qquad (7)$$

Thus the expression $\frac{\sum_{i=1}^{n} f(Z_i)}{n}$ approximates the integral. The approximation error decreases with increasing n [4]. In our experiments n is chosen by observing the convergence with increasing *n*.

3 Simulation of Feature Vectors

Our simulations are restricted to segments realized by 2 feature vectors i.e. by chunks of length *l*=2 denoted as $Z_n = \begin{pmatrix} X_n \\ X_{n-1} \end{pmatrix}$. We choose for Z_n a mono-modal Gaussian distribution, which is defined in section 3.1. In section 3.2 an efficient method to generate the chunks is presented leading also to an explicit expressions for the TEPs used in (2) and (3).

3.1 Distribution of the Chunks

We assume that the chunks emitted by the segments Q_i have a mono-modal multivariate Gaussian distribution

$$p(Z|Q_i) = N(Z; \mu_i, V)\ ;\ i=1,\dots,N_Q \qquad (8)$$

We first define the covariance V_x of a single feature vector *X* of dimension *D*. We assume that V_x is the same for all segments and that the vector components within each feature vector *X* are statistically independent leading to a diagonal covariance matrix V_X

$$V_X \equiv \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma^2 \end{pmatrix} \equiv \sigma^2\, I;\ \sigma^2 = 1 \qquad (9)$$

The variances σ^2 (normalized to 1) are the same for each component (globally pooled variance). This approach is inspired from the use of the LDA [5]. The statistic dependency of adjacent feature vectors X_{n-1}, X_n is specified by the correlation matrix

$$Cor_{X^{n-1},X^{n}} = \begin{pmatrix} \rho_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \rho_D \end{pmatrix}; \qquad (10)$$

The correlation coefficients ρ_d (d=1,...D) are defined by

$$\langle (x_{d,n} - \mu_d) \cdot (X_{d',n} - \mu_{d'}) \rangle = \delta_{d,d'}\rho_d\ ; d, d' = 1, \dots, D$$

This structure of the correlation matrix assumes that adjacent feature vectors components $x_{d,n}, x_{d',n-1}$ are statistic independent except for the components which have the same position *d* within the vector (*d=d'*). Using the specifications (9), (10) the structure of *V* is given by:

$$V = \begin{pmatrix} V_X & Cor_{X_{n-1},X_n} \\ Cor_{X_{n-1},X_n} & V_X \end{pmatrix} \qquad (11)$$

To model the means μ_i of the emission probabilities $N(Z;\mu_i,V)$ we investigate several cases.

In a first study the means were set to achieve different error rates. In the following the means of the Gaussians are notated as

$$\mu_i \equiv \begin{pmatrix} \mu_{i,X_n} \\ \mu_{i,X_{n-1}} \end{pmatrix} \tag{12}$$

3.2 Generation of Chunks

To generate the chunks we assume we have a generator of statistic independent Gaussian vectors. Using the results of [8,chapter 45] we can generate efficiently chunks with the distribution (8). In the following we sketch shortly these results:

The inverse covariance matrix of V is defined as

$$V^{-1} \equiv A \equiv \begin{pmatrix} A_{22} & A_{21} \\ A_{12} & A_{11} \end{pmatrix};\ A_{21} = A_{12}; \dim A_{uv} = D \cdot D\ ;\ u, v = 1{,}2 \tag{13}$$

Given (13) we generate samples y_n and y_{n-1} ($\dim(y_n) = \dim(y_{n-1}) = D$), which have to be statistic independent and are distributed by the mono-modal Gaussians

$$p(y_n) = N(0, A_{22}^{-1})\ ; \qquad p(y_{n-1}) = N(0, B^{-1}), B \equiv A_{11} - A_{21}A_{22}^{-1}A_{21}\ . \tag{14}$$

The samples $Y \equiv \begin{pmatrix} y_n \\ y_{n-1} \end{pmatrix}$ can be generated using a standard Gaussian sample generator.

Given the matrix $C^T \equiv \begin{pmatrix} I & -A_{12}A_{22}^{-1} \\ 0 & I \end{pmatrix}$ we generate samples $Z_n \equiv \begin{pmatrix} X_n \\ X_{n-1} \end{pmatrix}$ emitted by a segment Q_i with the distribution (8) using the relation:

$$Z_n - \mu_i = C^T Y\ ;\ \begin{pmatrix} X_n - \mu_{\mu_{i,X_n}} \\ X_{n-1} - \mu_{\mu_{i,X_{n-1}}} \end{pmatrix} = \begin{pmatrix} y_n - A_{22}^{-1}A_{21}y_{n-1} \\ y_{n-1} \end{pmatrix};\ \mu_i = \begin{pmatrix} \mu_{\mu_{i,X_n}} \\ \mu_{\mu_{i,X_{n-1}}} \end{pmatrix} \tag{15}$$

Thus the samples Z_n can be generated very efficiently by using by for all segments the same samples of y_n and y_{n-1}. Given the relation

$$p(Z|Q_i) \equiv N(Z; \mu_i, V) = p(X_n|X_{n-1}, Q_i) \cdot p(X_{n-1}|Q_i)$$

the relations (14) and (15) leads to expressions of the TEPs used in (2) and (3):

$$p(X_n|X_{n-1}, Q_i) = N\left(X_n|X_{n-1}; \mu_{i,X_n|X_{n-1}}, A_{22}^{-1}\right) \tag{16}$$

$$p(X_{n-1}|Q_i) = N\left(X_{n-1}; \mu_{i,X_{n-1}}, V_X\right) \tag{17}$$

$$\mu_{\mu_{i,X_n|X_{n-1}}} \equiv \mu_{\mu_{i,X_n}} - A_{22}^{-1}A_{21}\left(X_{n-1} - \mu_{\mu_{i,X_{n-1}}}\right)\ ;\ \mu_i \equiv \begin{pmatrix} \mu_{i,X_n} \\ \mu_{i,X_{n-1}} \end{pmatrix}$$

(16) shows that the conditional distribution leads to a shift of the means and to a new covariance matrix.

4 Results

In the following we will present segment error rates and Shannon's entropy for different settings of the parameters defining the distribution (8). We compare the FO-segment recognizer (4) with the related SO-recognizer based on the 2 trajectory models (2) and (3) realized by the TEPs (16), (17). Thus we get segment error rates FO-SER and SO-SER. The entropy $H(Z|Q)$ (6) is evaluated by averaging the term $-ld\ p\left(Q_i|\vec{X}_l\right)$ using the Monte Carlo Method (7). The trajectory model $p\left(Q_i|\vec{X}_l\right)$ is given by the SO and FO TEPs (16) and (17). To simulate the feature vectors we use (15) by generating samples X_n for each segment Q_i. In order to investigate the impact of the correlation we vary the means μ_i, the correlation coefficients ρ_d and the number N_Q of the segments as defined in section 3.1. The series of experiments are guided by the strategy to change the parameters from artificial ones with easy to understand distributions to those related to the HMM recognizer and the speech database described in section 4.1.

4.1 The Speech Database and the Derived Parameters

For our investigation we use a set of SpeechDat-Car [9] like speech databases containing about 1800h of **read** speech uttered by 810 speakers. The databases were labeled automatically into N_Q=607 segments by forced Viterbi alignment using a state of the art tri-phone HMM recognizer designed with N_Q= 607 segments.

The features of the HMM recognizer are analysed at a frame rate of 15ms. Thus the duration of a chunk of the length l is l*15ms. To improve the recognition performance in noisy environments, noise reduction techniques have been applied [13]. From noise reduced spectra, MFCCs and their 'dynamics' are derived and extended to super vectors, which are transformed by an LDA [5] to the final feature vector with the dimension D=24.

Given the segmented speech database, chunks of length l=2 were extracted and the sets $\vec{X}_{n,2} \in \vec{X}_2(Q_i); i = 1, \dots, 606$ were constructed. From these sets the 'database adapted' parameters of the distribution (8) were determined (see section 4.3). For each set the means μ_i were determined. The correlation coefficients ρ_d are determined by averaging over all segments.

4.2 A 2 Class Experiment

The first experiment is done with an easy to understand distribution (8) and were the number of segments N_Q is restricted by 2. Further we use following segment probabilities, means and correlation coefficients.:

- p denotes the probability $P_2(Q_1)$ of the segments Q_1 ($P_2(Q_2)$=1-p) (see (4))
- ρ_0 denotes the correlation parameter: $\rho_d = \rho_0; d = 1, \dots, D$ (see (10))
- the means are defined as follows: $\begin{pmatrix}\mu_{1,X_n} \\ \mu_{1,X_{n-1}}\end{pmatrix} = \begin{pmatrix}\mu_0 I_D \\ \mu_0 I_D\end{pmatrix}; \begin{pmatrix}\mu_{2,X_n} \\ \mu_{2,X_{n-1}}\end{pmatrix} = \begin{pmatrix}-\mu_0 I_D \\ -\mu_0 I_D\end{pmatrix}; D = 24$

To define the means we use the notation (12) and use the unity vector I_D defined by $I_D = (1,...,1)^T$, which has D components. The means of the segments were chosen in such a way that the 2 distributions $p(Z|Q_i)$, i=1,2 are symmetric around 0. The value of μ_0 determines the overlap of the two distributions and thus the segment error rates (SER). The number n of simulated features was set to 50 000 samples. Higher number of n did not change the results significantly. Initial an experiments with p=0.5 was performed leading to no change in error rates between the SO and the FO case. This is caused by the symmetric structure of the distributions, where the decision boundaries are the same for the SO and the FO case. Thus further experiments are made with p=0.3 ($H(Q)$=0.9).Table 1 shows the results for different values of ρ_0 and the same value of the means.

	SER	H(Q|Z)
SO	0.084	0.30
FO	0.084	0.30

$\rho_0 = 0$

	SER	H(Q|Z)
SO	0.192	0.51
FO	0.193	0.55

$\rho_0 = 0.5$

	SER	H(Q|Z)
SO	0.155	0.51
FO	0.161	0.61

$\rho_0 = 0.95$

Table 1 - parameter for the mean $\mu_0 = 0.19$

For the case $\rho_0 = 0$ the feature vectors are statistic independent leading to the same results for FO and SO case. With increasing correlation parameter ρ_0 the error rate increases. This is consistent with the increase of Shannon's entropy in both cases as the information contained in the chunk decreases. In the SO case the error rate is less than in the FO case. But this effect is very small. Bigger differences can be seen in Shannon's entropy.

In fig. 1 scatter plots of *SER* - $H(Q|Z)$ are shown. The different points are gained by simulations with varying the parameter μ_0. The error rate and Shannon's entropy drops with increasing μ_0. Fig.1 shows the case ρ_0=0.5 and ρ_0=0.95. With increasing ρ_0 the SER increases as less information is in the feature vectors. Also the *SER* and the related Shannon's entropy show higher values for the FO-case as in the SO-values showing the impact of the correlation. The differences in error rates are small compared to differences in Shannon's entropy as already shown in table 1. The case $\rho_0 = 0.95$ demonstrates that the points *H*(*Q*|*Z*)-*SER* tend to leave the upper Fano bound for the FO case.

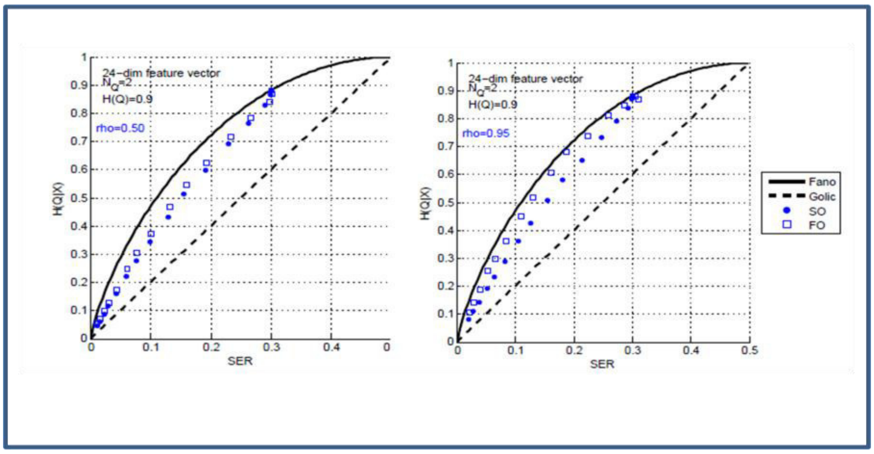

Figure 1a and 1b- left figure 1a: ρ_0=0.5 ; right figure 1b: ρ_0=0.95

4.3 Experiments with Database Adapted Parameters

In the following we describe 3 experiments adapting step by step the parameters to those derived from the speech database as described in section 4.1. For all experiments the number of segments is set to N_Q=606, where we omitted the silence class. The probabilities of the segments $P_l(Q_i)$ needed for the recognizer (4) were set to the values as found in the speech database. This probabilities lead to an entropy of *H(Q)*=8.3.

4.3.1 Experiment I

The means $\begin{pmatrix} \mu_{i,X_n} \\ \mu_{i,X_{n-1}} \end{pmatrix} = \begin{pmatrix} \mu_{r,i} I_D \\ \mu_{r,i} I_D \end{pmatrix}$ are generated by a random number generator for the random values $\mu_{r,i}$ with the distribution $N(\mu_{r,i}; 0, \sigma^2_{\mu_r} I)$. Depending on $\sigma^2_{\mu_r}$ different error rates were achieved. The correlation coefficients are set as described in section 4.2.

	SER	H(Q|Z)
SO	0.31	1.72
FO	0.31	1.72

$\rho_0 = 0.$

	SER	H(Q|Z)
SO	0.55	4.98
FO	0.56	5.73

$\rho_0 = 0.7$

	SER	H(Q|Z)
SO	0.62	4.01
FO	0.64	5.16

$\rho_0 = 0.95$

Table 2- $\sigma_{\mu_r} = 0.56$

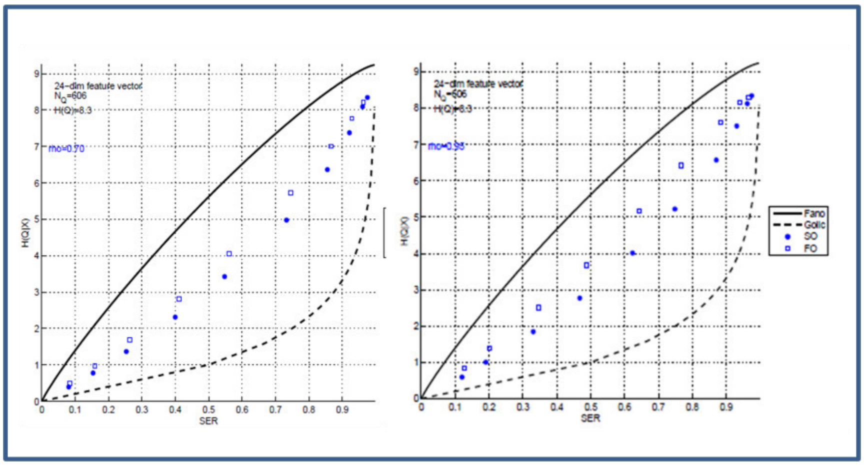

Figure 2 - simulations with ρ_0=0.5 and ρ_0=0.95 and varying σ_{μ_r}

In table 2 we fix σ_{μ_r} and vary the correlation coefficients as in table 1. The results are similar as described in section 4.2.

4.3.2 Experiment II

The means $\begin{pmatrix} \mu_{i,X_n} \\ \mu_{i,X_{n-1}} \end{pmatrix}$ of the chunks measured in the speech database were used for the simulation. The correlation coefficients are as set in section 4.2. As shown in table 3 the results are similar to those in table 2. This concludes that a Gaussian model for the means approximates good the behavior of the recognizer.

	SER	H(Q|Z)
SO	0.43	2.10
FO	0.43	2.10

$\rho_0 = 0$

	SER	H(Q|Z)
SO	0.57	3.06
FO	0.58	3.30

$\rho_0 = 0.5$

	SER	H(Q|Z)
SO	0.66	3.80
FO	0.68	4.69

$\rho_0 = 0.95$

Table 3 - simulation with means given by the database

4.3.3 Experiment III

In a final experiment we adapt the correlation coefficients ρ_d as defined in (10) to the values found in the speech database. ρ_1 has the greatest value of 0.87, whereas ρ_{24} has the smallest value of 0.29. To see the impact of correlation we modified these coefficients by the formula

$$\rho'_d = \varphi_0 \, \rho_d \, ; d = 1, \ldots, 24$$

For $\varphi_0 = 0$ we have no correlation, for $\varphi_0 = 1$ we have the correlation as in the database. In table 4 we see results for different values of φ_0.

	SER	H(Q|Z)
SO	0.43	2.10
FO	0.43	2.10

$\varphi_0 = 0$

	SER	H(Q|Z)
SO	0.54	2.86
FO	0.55	3.02

$\varphi_0 = 0.5$

	SER	H(Q|Z)
SO	0.62	3.51
FO	0.64	4.09

$\varphi_0 = 1.0$

Table 4 - simulation with modified correlation coefficients

Table shows, that differences in error rates observed for the FO and SO case are small. This is a similar result as found in table 3. The case $\varphi_0 = 1.0$ comes closest to the distribution of the features found in the database. For a multi modal acoustical models as presented in [6] the error rate is 0.49, which shows that more accurate FO-models achieve better results than simple SO-models.

5 Conclusion

Based on the simulations we conclude that the modeling of correlation of adjacent feature vectors does not improve the recognition rate of segments substantially, when using mono-modal distributions. We assume, that the decision surfaces do not change significantly leading to this finding. It is still an open question, if this finding holds also for multimodal distributions. Shannon's entropy shows significant differences. As this entropy is related to scores used for rejection, the use of models modeling correlation may lead to better rejection strategies.

Literature

[1] Höge, H.: Second-Order Models for Hidden Chunk Models. In Proc. ITG Fachtagung Speech Communication: October 2010.

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

[3] Tokuda, K., Zen, H., and Kitamura, T.: Trajectory modeling based on HMMs with the explicit relationship between static and dynamic features. In Proc. of Eurospeech2003: 2003, pp. 865-868

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

[5] Hauenstein, A. and Marschall, E.: Methods for Improved Speech Recognition over Telephone Lines. In Proc. ICASSP 95:1995,pp. 425-428

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

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

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

[9] Moreno, A., Lindberg, B., Draxler, C., Richard, G., Choukri, S. Euler, K. , and Allen, J.: SpeechDat-Car: A Large Database for Automotive Environments. In Proc. of LREC: May 2000.

[10] Setiawan, P. , Beaugeant, C., Fingscheidt, T., and Stan, S.: Least-Squares Weighting Rules Formulations in the Frequency Domain. In Proc. ESSP: September 2005

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

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

[13] Setiawan, P. , Beaugeant, C., Fingscheidt, T. and Stan, S.: Least-Squares Weighting Rules Formulations in the Frequency Domain. In Proc. ESSP: September 2005.