EVALUATION OF INFORMATION CONTAINED IN SPECTRAL FEATURES

Harald Höge[1] and Panji Setiawan[2]

[1]SVOX Deutschland GmbH, [2]Siemens Enterprise Communications GmbH & Co. KG
harald.hoege@svox.com, panji.setiawan@siemens-enterprise.com

Abstract: In this paper we estimate the information contained in features to recognize phones represented by HMMs states. The features investigated are derivatives of MFCCs. The information is defined as the 'Phone Entropy', which is Shannon's conditional entropy applied to phones. Related to the entropy are the bounds concerning the minimum phone error rate, which are achieved by a recognizer based on the Bayes principle. We use the Fano and Golić bounds as the lower and upper bounds, respectively. This paper is focused on estimating the Phone Entropy and the underlying probability functions. An approach to overcome the first order Markov model, as used in the state of the art HMM technology, is investigated. Experimental results are presented using the AURORA framework, where we determine the Phone Entropy and phone error rates for different noise levels. Phone error rates are compared to the bounds given by Fano and Golić.

1 Introduction

Research in speech recognition is focused on the reduction of word error rates. Improvements are achieved by using new features or using better statistical models. In this paper we regard features derived from the mel-frequency cepstral coefficients (MFCCs). For estimating the information contained in the features we use Shannon's conditional entropy *H(Q|X)* [1], where X denotes the features and Q the units to be recognized. *H(Q|X)* determines the number of bits needed to decode the units Q without error, i.e., if the condition *H(Q|X) = 0* holds, the units Q can be recognized without errors. For *H(Q|X) > 0* there exists a lower bound - the Fano bound [2] - for the minimal error rate achievable. The Fano bound is a function of *H(Q|X)*. This paper is focused on the estimation of *H(Q|X)* for different speech corpora which have different noise levels. As units Q we use 'phones' modelled by the HMMs states.

The paper is organized as follows. Chapter 2 provides the statistical framework to relate error rates to *H(Q|X)*. Chapter 3 is devoted to provide methods to estimate *H(Q|X)*. Finally we present experimental results using the AURORA framework ([9], [10]).

2 The Statistical Framework

According to the theory of pattern recognition [8], the Bayes classifier achieves minimum error rates. Given an uttered sequence $\vec{W}$ of words, and given a sequence of feature vectors $\vec{X}_t = (X_1,...,X_t)$ the Bayes estimate for $\vec{W}$ is given by

$$\vec{W} = \arg\max_{\vec{W}} p(\vec{W} \mid \vec{X}_t) = \arg\max_{\vec{W}} p(\vec{X}_t \mid \vec{W})\, P(\vec{W}). \qquad (2.1)$$

The minimum error rate can be achieved only if $P(\vec{W})$ - the language model - and $p(\vec{X} \mid \vec{W})$ - the acoustic model are known. The main obstacle to estimate the entropy $H(\vec{W} \mid \vec{X})$ is the high and variable dimensionality of the acoustic model $p(\vec{X} \mid \vec{W})$. In HMM technology each word W is modelled by a sequence of states $\vec{Q}_W$ and the acoustic model for $p(\vec{X} \mid \vec{W})$ is given by a state model $p(\vec{X} \mid \vec{Q}_{\vec{W}})$ based on a first order Markov process [7]. Using (2.1) the recognition of words is transformed to the recognition of a sequence of states given by[1]

[1] In the following we denote by $\tilde{p}(Z)$ an approximation of the real distribution *p(Z)* for any variable *Z*.

$$p(\vec{X} \mid \vec{Q}_{\vec{W}}) \approx \tilde{p}(\vec{X} \mid \vec{Q}_{\vec{W}}) \equiv \prod_m a_{j(m-1),j(m)} p(X_m \mid Q_{j(m)}), \qquad m = 1,...,t \tag{2.2}$$

$$\vec{Q}_{\vec{W}} = \arg\max_{\vec{Q}_{\vec{W}}} (\tilde{p}(\vec{X}_t \mid \vec{Q}_{\vec{W}}) \tilde{P}(\vec{Q}_{\vec{W}})) = \arg\max_{j(m)} (\tilde{P}(\vec{Q}_{\vec{W}}) \prod_m a_{j(m-1),j(m)} p(X_m \mid Q_{j(m)})). \tag{2.3}$$

(2.3) defines a search task where an optimal alignment function *j(m)* has to be found, which assigns each frame *m* to a state $Q_{j(m)}$, which maximizes the probability $\tilde{p}(\vec{X} \mid \vec{Q}_{\vec{W}}) \tilde{P}(\vec{Q}_{\vec{W}})$. $a_{j(m-1),j(m)}$ denote the transition probabilities, which allow only the transitions on states compatible with the word specific sequences $\vec{Q}_W$. The optimal alignment *j(m)* assigns each feature vector X_m to a state $Q_{j(m)}$ defining a segmentation $S_X=(X_m, Q_{j(m)})_{m=1,...,t}$. Furthermore, this segmentation defines a state sequence $\vec{Q}_{\vec{W}} \equiv (Q_{j_Q(n)})_{n=1,T}$ given by the state assignment function $j_Q(n)$. We transform the segmentation S_X into a segmentation S_Q

$$S_Q \equiv (\vec{X}_{l_n}, Q_{j(n)})_{n=1,T}, \tag{2.4}$$

where $\vec{X}_{l_n} = X_{m_n},..., X_{m_n+l_n-1}$ denotes a sequence of feature vectors belonging to the same state $Q_{j_Q(n)}$. This segmentation can be improved further using a forced Viterbi algorithm, where the sequence of states is known. Using the improved segmentation S_Q we can define the phone recognition task, where the functions *j(m)* and $j_Q(m)$ are fixed, but the phones $Q_{j_Q(n)}, n=1,...,T$ have to be recognized[2]:

$$\vec{Q} = \arg\max_{\vec{Q}} p(\vec{X} \mid \vec{Q}, S_Q) P(\vec{Q}) \approx \arg\max_{\vec{Q}} \tilde{p}(\vec{X} \mid \vec{Q}, S_Q) \tilde{P}(\vec{Q}),$$

$$\tilde{p}(\vec{X} \mid \vec{Q}, S_Q) \equiv \prod_{\vec{m}_n} p_{l_n}(\vec{X}_{l_n} \mid Q_j, S), \qquad n = 1,...,T \tag{2.5}$$

$$\vec{Q} = \arg\max_{\vec{Q}} \tilde{p}(\vec{X} \mid \vec{Q}, S_Q) \tilde{P}(\vec{Q}) \approx \prod_{n=1}^{T} \arg\max_j p_{l_n}(\vec{X}_{l_n} \mid \vec{Q}, S_Q) \tilde{P}(Q_j).$$

The approximation made in (2.5) assumes that both, the feature vectors assigned to different phones and the phones themselves are statistically independent. We call the functions $p_l(\vec{X}_l \mid Q_j)$ the 'extended emission probabilities' (**EEPs**). They model the statistical dependencies of all feature vectors assigned to a phone; thus we have to model EEPs for different values of *l*. This is in contrast to the HMM approach given by (2.3), where all feature vectors are regarded as statistically independent (case *l*=1). (2.5) defines a phone recognition task with the maximum a posteriori solution

$$Q(\vec{X}_{l_n}) = \arg\max_j p_{l_n}(\vec{X}_{l_n} \mid Q_j, S) \tilde{P}(Q_j), \qquad j = 1,..., N_Q; n = 1,...,T$$

where N_Q denotes the number of phones. Given the following probabilities[3]:

$$p_l(\vec{X}_l \mid Q_j) = p_l(X_1,..., X_l \mid Q_j); p_l(\vec{X}_l) = \sum_{j=1}^{N_Q} P(Q_j) p_l(\vec{X}_l \mid Q_j), \qquad l = 1,..., L \tag{2.6}$$

where $L = max\ (l)$, the following entropies dependent on *l* can be defined:

$$H(\vec{X} \mid l) \equiv -\int_{\vec{X}_l} p_l(\vec{X}_l)\ ld\ p_l(\vec{X}_l)\ d\vec{X}_l;\ H(\vec{X} \mid l, Q_j) \equiv -\int_{\vec{X}} p_l(\vec{X}_l \mid Q_j) ld\ p_l(\vec{X}_l \mid Q_j) d\vec{X}_l. \tag{2.7}$$

Given the entropies defined in (2.7) we define the **Phone Entropy *PE(X)***

---

[2] It is well known in speech recognition technology, that the optimal sequence of states does not represent sequences corresponding to words.

[3] In the following we do not distinguish between $P(Q_j)$ and $\tilde{P}(Q_j)$, because these values can be conveniently estimated.

$$PE(X) \equiv H(Q \mid \vec{X}) = H(Q) - I(\vec{X};Q); \qquad I(\vec{X};Q) \equiv H(\vec{X}) - H(\vec{X} \mid Q),$$
$$H(\vec{X} \mid Q) \equiv -\sum_{j=1}^{N_Q} P(Q_j) \sum_{l=1}^{L} P(l) H(\vec{X} \mid l, Q_j); \qquad H(\vec{X}) \equiv -\sum_{l=1}^{L} P(l) H(\vec{X} \mid l). \tag{2.8}$$

Thus *PE(X)* is determined [3] by *H(Q)* and the mutual information $I(\vec{X};Q)$. *H(Q)* is the number of bits needed to recognize the phones. $I(\vec{X};Q)$ is the number of bits extracted from the feature vectors *X*. Consequently, *PE(X)* represents the missing information measured in bits to recognize the phones without errors $(PE(X) \geq 0)$. Missing information leads to errors. The minimal achievable phone error rate is bounded by the Fano and the Golić bounds. These bounds are a function of *PE(X)* (see Section 6.1). The bounds together with the experimental measured values of *PE(X)* and the related phone error rates are presented in Section 4.3.

3 Mutual Information of Features

In this chapter we aim to estimate $I(\vec{X};Q)$ as needed to evaluate the Phone Entropy (2.8). The main issue is to develop a model for the extended emission probabilities (EEPs) as defined in (2.6). This model depends on the statistical properties of the features used.

State of the art recognizers generally use MFCCs features [20], which are processed further to increase the temporal context [15] and to remove long term statistical dependencies, which are not modelled by the first order Markov assumption. Channel compensation [14], vocal tract length normalization [13], and noise reduction [16] are typical methods to remove the long term statistical dependencies. The increase of temporal context results in a higher feature vector dimension. This dimension is decreased by using a linear discriminative analysis LDA [11]. The LDA plays an important role on the statistical properties of the EEPs (2.6), because it de-correlates the components x_k of a feature vector $X=(x_1,...,x_d)^T$ and equalizes the variances of the components [6].

The information contained in the feature vectors depends on the quality of the speech recorded. In Chapter 4 we regard speech, which is distorted by environmental noise. These environmental distortions influence the phone error rates and the Phone Entropy.

In order to achieve some analytical solutions, we assume that the feature components $x_{k,j}$, $k=1,...,d$ of the d-dimensional feature vector X_j assigned to a phone Q_j are statistically independent with respect to k and Gaussian distributed. Furthermore, we assume that the components $x_{k,j}^m,...,x_{k,j}^{m+l_n}$ emitted by the same phone Q_j are generated by an ARMA process

$$y_{k,j}^q = \begin{cases} \xi_{k,j}^q, & for \quad l_n = 1 \\ \sum_{i=1}^{l_n} (-a_{k,j,l}^i y_{k,j}^{q-i}) \; + \; \xi_{k,j}^q, & for\, l_n > 1 \end{cases} \qquad k=1,...,d; \quad j=1,...,N_Q \tag{3.1}$$
$$x_{k,j}^{m_n+q-1} = y_{k,j}^q + \mu_{k,j}; \; y_{k,j}^{q-i} = 0, \qquad for \; q-i<1 \; ; \; q = 1,...,l_n$$

driven by a Gaussian process $\xi_{k,j}^q$ with the distribution $N(\xi_{k,j}, 0, \sigma_{k,j}^2)$. Whenever a feature vector 'jumps' into a new state, the ARMA process starts with a value $\xi_{k,j}^1$ of the random variable $\zeta_{k,j}$. For each l the distribution of the EEPs is given by

$$p_l(\vec{X}_l \mid Q_j) = \prod_{k=1}^{d} p_1(x_{1,k,},...,x_{l,k} \mid Q_j),$$
$$p_l(x_{1,k},...,x_{l,k} \mid Q_j) = \begin{cases} p_1(x_{1,k} \mid Q_j), & for\, l = 1 \\ p_1(x_{1,k} \mid Q_j) \prod_{i=l}^{2} p_i(x_{i,k} \mid x_{i-1,k},...,x_{1,k}, Q_j). & for\, l > 1 \end{cases} \tag{3.2}$$

The entropy of this Gaussian ARMA process is given by [3]

$$H_l(x_k \mid Q_j) = \tfrac{1}{2}\ln(2\pi e\sigma^2_{k,j}) + \tfrac{1}{4\pi}\int_{\omega=-\pi}^{\pi} \log g_{j,k}(\omega)\, d\omega; \quad g_{j,k}(\omega) = \left|\sum_{i=0}^{l} a^i_{k,j,l} e^{i\omega l}\right|^{-2}; a^0_{k,j} = 1. \quad (3.3)$$

The coefficients $a^i_{k,j}$ can be determined by the Durbin-Levinson Algorithm [3] using the autocorrelation function $\varphi_{k,j}(x_{k,}|Q_j)$. Furthermore, $g_{j,k}(\omega)$ is given by the Fourier transform of $\varphi_{k,j}(x_{k,}|Q_j)$. In the following two sections we investigate the case l=1 and l=2, which allows to estimate the Phone Entropy $PE(X)$ for L<3 (L=max(l); see Table 4.1).

3.1 Case l=1

This case is described in [6] in more detail. As the statistics of the feature vector X are modelled by the EEPs with l=1, $p_1(x_{1,k}|Q_j)$ is equivalent to the emission probabilities $p(x_k|Q_j)$ used in HMM technology (see (3.2))

$$p_1(X \mid Q_j) = \prod_{k=1}^{d} p_1(x_k \mid Q_j); \quad p_1(x_k \big| Q_j) = N(x_k, \mu_{x_k|Q_j}, \sigma^2_{x_k|Q_j}), \qquad X = (x_1,...,x_d)^T \quad (3.4)$$

leading to the entropy

$$H_1(X \mid Q) = \sum_{j=1}^{N_Q} P(Q_j) H_1(X \mid Q_j); \qquad H_1(X \mid Q_j) = \sum_{k=1}^{d} (\tfrac{1}{2}\ln(2\pi e) + \tfrac{1}{2}\ln(\sigma^2_{x_k|Q_j}). \quad (3.5)$$

Using (2.6) and (3.4) we can determine $H_1(X)$

$$H_1(X) = -\int p_1(X)\, ld\ p_1(X)\, dX; \qquad p_1(X) = \sum_{j=1}^{N_Q} P(Q_j) p_1(X \mid Q_j). \quad (3.6)$$

In [6] several methods have been investigated to estimate $H_1(X)$. If we assume $p_1(X)$ is approximated by a monomodal Gaussian distribution with variances and means as given by the multimodal distribution (see (3.6)), we get for the mutual information the following nice formulation:

$$I_{1,M}(X;Q) = \sum_{k=1}^{d} I_M(x_k;Q); \qquad I_{1,M}(x_k;Q) = \tfrac{1}{2}\sum_{j=1}^{N_Q} P(Q_j) \ln \frac{\sigma^2_{x_k}}{\sigma^2_{x_k|Q_j}}, \quad (3.7)$$

$$\sigma^2_{x_k} = \sum_{j=1}^{N_Q} P(Q_j)(\sigma^2_{x_k|Q_j} + (\mu_{x_k} - \mu_{x_k|Q_j})^2); \qquad \mu_{x_k} = \sum_{j=1}^{N_Q} P(Q_j)\mu_{x_k|Q_j}, \qquad k = 1,...,d$$

where the index M denotes a monomodal approximation. This approximation shows that the mutual information depends on the variances and means of the related distributions. Based on the distributions $p_1(X|Q_j)$ (3.4) and $p_1(X)$ (3.6), alternatively, the entropies $H_1(X)$ and $H_1(X|Q)$ can be estimated by the Monte Carlo method as described in the Appendix.

3.2 Case l=2

According to (3.2) we have to treat the distributions

$$p_2(\vec{X}_2 \mid Q_j) = \prod_{k=1}^{d} p_1(x_{1,k}, x_{2,k} \mid Q_j),$$

$$p_l(x_{1,k}, x_{2,k} \mid Q_j) = p_1(x_{1,k} \mid Q_j) p_2(x_{2,k} \mid x_{1,k}, Q_j),$$

which are approximated by using the properties of the ARMA process defined by (3.1)

$$p_1(x_{1,k} \mid Q_j) \equiv N(x_{1,k}, \mu_{x_k|Q_j}, \sigma^2_{x_k|Q_j}),$$

$$p(x_{2,k} \mid x_{1,k}, Q_j) \equiv N_{\xi_{k,j}}(x_{2,k} - a_{k,j}x_{1,k} - \mu_{x_k|Q_j}(1-a_{k,j}), 0, \sigma^2_{\xi_{k,j}}); \qquad a_{k,j} \equiv a^1_{k,j,2}, \quad (3.8)$$

$$a_{k,j} = \frac{E(x_{2,k}x_{1,k} \mid Q_j)}{\sigma^2_{x_{1,k}|Q_j}}; \qquad \sigma^2_{\xi_{k,j}} = \sigma^2_{x_k|Q_j}(1-(a_{k,j})^2).$$

Given the distributions derived in (3.8) and the result in (3.5) we get

$$H_2(\vec{X}_2 \mid Q_j) = H_2(X \mid Q_j) + H_1(X \mid Q_j),$$

$$H_2(X \mid Q_j) \equiv -\int_{\vec{X}_2} p_2(\vec{X}_2 \mid Q_j)\, ld\ p_2(X_2 \mid X_1, Q_j)\, d\vec{X}_2, \qquad \vec{X}_2 \equiv (X_1, X_2)^T$$

$$= \sum_{k=1}^{d} (\tfrac{1}{2}\ln(2\pi e) + \tfrac{1}{2}\ln(\sigma^2_{x_k|Q_j}(1-(a^j_{k,2})^2))) = H_1(X \mid Q_j) - \tfrac{1}{2}\sum_{k=1}^{d} \ln\frac{1}{1-(a_{k,j})^2},$$

$$H_2(\vec{X}_2 \mid Q_j) = 2H_1(X \mid Q_j) - \tfrac{1}{2}\sum_{k=1}^{d} \ln\frac{1}{1-(a_{k,j})^2} \le 2H_1(X \mid Q_j). \tag{3.9}$$

In the case, where the feature vectors X are statistically independent, the entropy adds equally with the value $H_1(X|Q_j)$ for each frame, whereas for a second order ARMA process, the entropy is reduced depending on the parameters $a_{k,j}$. If these parameters are 0 (no correlation), there is no loss in entropy. If the parameters approach 1 the entropy goes to $-\infty$.

4 Experiments

In this chapter we investigate the Phone Entropy and error rates for phones represented by the states of whole word HMM models from the German digits having N_Q = 264 phones. As speech corpus we use the AURORA-3 German digits database [10], which is a subset of the German SpeechDat-Car (SDC) database containing digits utterances [17]. The dataset used has a total number t = 407 978 feature vectors (without pauses).

4.1 Environmental Distortions

The utterances were recorded under different noise conditions. For each utterance an SNR value was assigned using the method described in [18]. This method measures the SNR values separately for each mel-filterbank channel summing up to a value of SNR_{mel} for each utterance. According to the noise level, the utterances are clustered into different datasets covering a range of SNR_{mel} values (see Table 4.2).

4.2 Analysis of ARMA Processes

Applying the segmentation as given by (2.4) to the AURORA-3 database, we have the following distribution of the length l of feature vectors assigned to the same phone:

l	1	2	3	4	5	6	7
P(l)	0.409	0.128	0.041	0.017	0.008	0.005	0.003

Table 4.1 – Distribution of P(l).

Table 4.1 shows that 59% of the phones could be handled by an ARMA process with $l > 1$. Furthermore, in Figure 4.1 the normalized autocorrelation coefficients for different components x_k are shown for those cases, where at least 2 feature vectors are assigned to a state. The figure shows that the correlation decreases with increasing index k. This is in contrast to the property of the LDA to order the components with increasing index to have less discriminative power [6].

4.3 Phone Entropy and Phone Error Rates

In this section we estimate phone entropies as investigated in Chapter 3 for the case l=1, which applies for 41% of all the phones observed. As shown in [6] the Phone Entropy $PE(X)$ is estimated with the Monte-Carlo method leading to estimates $\tilde{H}^N_{MC}(Q \mid X)$ (see Section 6.2).

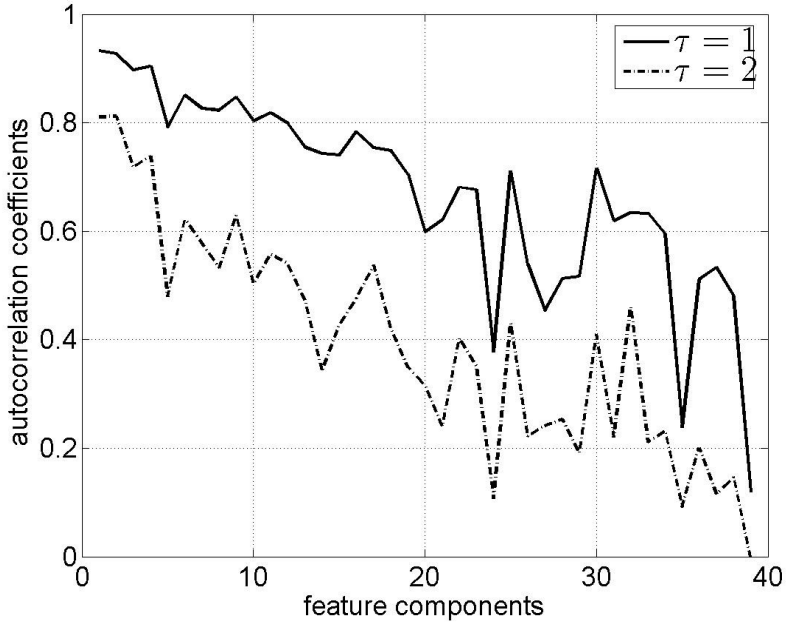

Figure 4.1 – Coefficients $a_{k,j}$ as defined in (3.8), but averaged over all phones Q_j.

range of SNR$_{mel}$	$\tilde{H}_{MC}^{N}(Q \mid X)$	e_{Ph} **[%]**
16-24 dB	3.77	64
8-16 dB	4.22	71
0-8 dB	4.85	76

Table 4.2 – Measured Phone Entropies $\tilde{H}_{MC}^{N}(Q \mid X)$ and error rates e_{ph}.

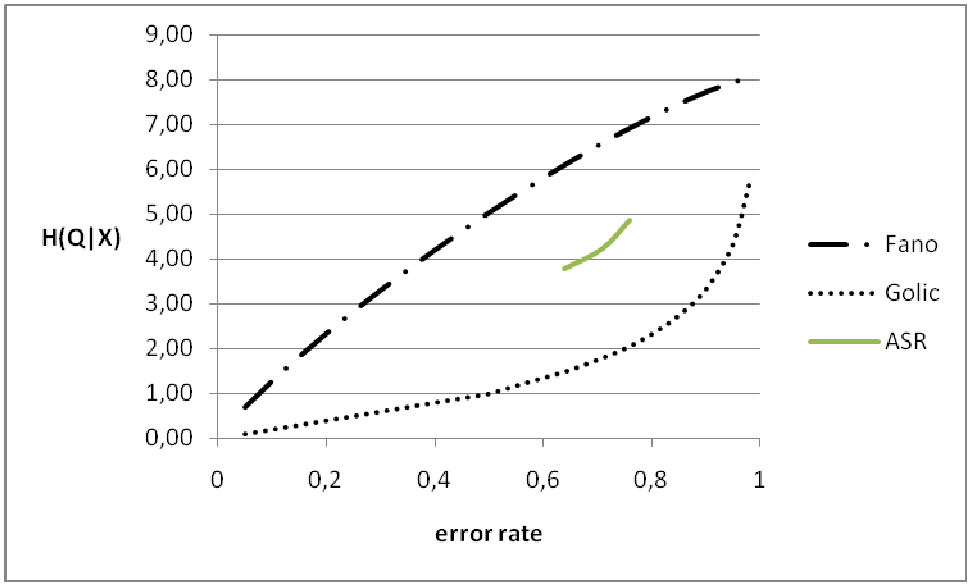

Figure 4.2 – Phone error rates e_{Ph} and $\tilde{H}_{MC}^{N}(Q \mid X)$ related to the upper and lower bounds.

Table 4.2 shows the error rate e_{Ph} for l=1 leading to the 'ASR' curve shown in Figure 4.2. Furthermore, Figure 4.2 shows the relation between both the measured error rates and phone entropies and the Fano and Golić bounds. Although the position of the curve denoted by 'ASR' lies between the lower and upper bounds as the theory predicts, yet we do not know the precision of the models used to estimate the phone entropy in making the final conclusions.

5 Conclusions and Future Work

To evaluate the information included in the feature vectors we have investigated a phone recognition task. We extended the statistical framework of HMMs by modeling the statistical dependencies of the feature vectors within a phone using an ARMA approach. Still this extension has to be explored experimentally. We defined a Phone Entropy *PE(X)*, which is related to the error rates, and compared the measurement of these values with the Fano and Golić bounds. Still an open issue is to model the statistical dependencies of feature vectors belonging to different phones.

6 Appendix

6.1 Bounds given by Fano and Golić

Given the phone error rate e_{Ph} from a maximum a posteriori classifier the Fano bound is given by

$$H(Q \mid \vec{X}) = H(e_{Ph}) + e_{Ph}\, ld(N_Q - 1); \qquad H(e_{Ph}) \equiv -e_{Ph} ld\, e_{Ph} - (1 - e_{Ph})\, ld(1 - e_{Ph}).$$

The Golić bound is given by

$$H(Q \mid \vec{X}) = j(j+1).(e_{Ph} - \frac{j-1}{j}) \log_2 (\frac{j+1}{j}) + \log_2 j,$$

for $\frac{j-1}{j} \le e_{Ph} \le \frac{j}{j+1}$ and $1 \le j \le N_Q - 1$.

6.2 Monte Carlo Approximation

Given an estimate $\tilde{p}(Z)$ for a distribution $p(Z)$ for an arbitrary random variable Z we define the approximated entropy

$$\tilde{H}(Z) \equiv E(-ld\ \tilde{p}(Z)) = -\int_X p(Z)\, ld\ \tilde{p}(Z)\, dX.$$

Given a set of samples Z_m from a speech database, the expectation $E(-ld\ \tilde{p}(Z))$ can be approximated by the 'Monte-Carlo Entropy'

$$H_{MC}^N(p(Z)) \equiv -\tfrac{1}{N} \sum_{m=1}^{N} ld(\tilde{p}(Z_m)) \approx \tilde{H}(Z).$$

Due to the law of large numbers the relation

$$E((H(Z) - H_{MC}^N(p(Z)))^2) \propto \frac{\sigma^2}{t},$$

holds, i.e., the Monte-Carlo entropy converges to *H(Z)* for infinite *t*.

References

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

[2] Fano, R.M.: Transmission of Information: A Statistical Theory of Communications. MIT Press and JohnWiley & Sons, Inc. New York: 1961.

[3] Papoulis, A.: Probability, Random Variables, and Stochastic Processes. Third Edition, McGraw-Hill, Inc.: 1991.

[4] Golić, 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.

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

[6] 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.

[7] Rabiner, I.R., Juang, B.-H.: Fundamentals of Speech Recognition. Englewood Cliffs, New Jersey, Prentice Hall: 1993.

[8] Duda, H., Hart, P.: Pattern Classification and Scene Analysis. John Wiley & Sons: 1973.

[9] Pearce, D.: Enabling New Speech Driven Services for Mobile Devices: An Overview of the ETSI Standards Activities for Distributed Speech Recognition Front-Ends. Proc. Applied Voice Input/Output Society Conference (AVIOS): May 2000.

[10] ETSI STQ-Aurora, AU/273/00 V1.1: Description and Baseline Results for the Subset of the Speechdat-Car German Database used for ETSI STQ Aurora WI008 Advanced DSR Front-end Evaluation: January 2001.

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

[12] Andrassy, B., Vlaj, D, Beaugeant, C.: Recognition Performance of the Siemens Front-End with and without Frame Dropping on the Aurora 2 Database. Proc. INTERSPEECH: Sept.2001, pp.193-196.

[13] Pitz, M., Molau, S., Ralf Schlüter, R., Ney, H.: Vocal Tract Normalization Equals Linear Transformation in Cepstral Space, Eurospeech 2001: 2001.

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

[15] Höge, H., Hohenner, S., Kunstmann, B., Schachtl,S.,Schönle, M., Setiawan, P.: Automotive Speech Recognition. In: Z.-H. Tan and B. Lindberg (Ed): Automotive Speech Recognition on Mobile Devices and over Communication Networks (Advances in Pattern Recognition): Springer-Verlag: 2008, pp. 347-374.

[16] Setiawan, P., Beaugeant, C., Fingscheidt, T., Stan, S.: Least-Squares Weighting Rules Formulations in the Frequency Domain. Proc. Electronic Speech Signal Processing Conference (ESSP): September 2005.

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

[18] Andrassy, B., Höge, H.: Human and machine recognition as a function of SNR. Proc. LREC 2006: 2006.

[19] Bauer, J.: Enhanced Control and estimation of parameters for a telephone based isolated digit recognizer. Proc. ICASSP97: 1997, pp.1531-1534.

[20] Davis, S.B., Mermelstein, P.: Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 28: 1980, pp. 357-366.