Entropy-Based Feature Analysis for Speech Recognition

Panji Setiawan1, Harald Höge2, and Tim Fingscheidt3

1Siemens Enterprise Communications GmbH & Co. KG, Munich, Germany
2SVOX Deutschland GmbH, Munich, Germany
3Technische Universität Braunschweig, Braunschweig, Germany

Abstract

Based on the concept of entropy, a new approach to analyse the quality of features as used in speech recognition is proposed. We regard the relation between the hidden Markov model (HMM) states and the corresponding frame based feature vectors as a coding problem, where the states are seen through a noisy recognition channel and received as feature vectors. Using the relation between Shannon’s conditional entropy and the error rate on state level, we estimate how much information is contained in the feature vectors to recognize the states. Thus, the conditional entropy is a measure for the quality of the features. Finally, we show how noise reduces the information contained in the features.

Index Terms: speech recognition, coding, entropy

1. Introduction

Bayes decision rule has been widely used as the fundamental approach to solve the problems in speech recognition which theoretically leads to minimum error rate. Given the observed sequence of feature vectors $x^M = \{x(\ell)\}_{\ell=0}^{M-1} = \{x(0), \ldots, x(M-1)\}$ of length $M$, where each $x(\ell)$ is a $d$-dimensional instance of a multivariate random variable $X = [X_1, \ldots, X_d]^T$, the Bayes estimator for the word is

$$
\hat{W} = \arg \max_W p(x^M|W)P(W),
$$

where the a posteriori probability is replaced by the conditional probability density $p(x^M|W)$ and the a priori probability $P(W)$. This leads to the well-known state-of-the-art maximum likelihood method\(^1\). Based on the state-of-the-art, $p(x^M|W)$ is approximated by $M$-HMMs and $P(W)$ by n-grams. Using (1) the quality of the features $x(\ell)$ is assessed by comparing recognition rates derived from different designed feature sets. This approach has been adopted in the Aurora evaluation campaigns\(^1\) to assess the quality of different feature extraction methods.

Our aim is to find a direct relationship between the statistical properties of the features and an achievable recognition rate using the concept of entropy. Due to the high dimensionality of the vector $x^M$, the probability function $p(x^M|W)$ in (1) can not be analysed directly. In this paper we regard as recognition units, the HMM states of words constructed by phonemes or whole word models. We assume that the recognition rate on the state level is highly correlated with that on the word level. The paper is organized as follows: In Section 2 the proposed concept of entropy for feature analysis is presented followed by a brief description of the entropy bounds of the Bayes probability of error in Section 3. Approximations to the entropy related terms are given in Section 4 and finally, its application for feature analysis is demonstrated in Section 5 followed by the experimental results in Section 6.

2. The Concept of Entropy

A sequence of features $x^M$ describing a speech utterance corresponds to a sequence of states $\psi^M = \{\psi(0), \ldots, \psi(M-1)\}$ where $\psi(\ell)$ denotes the state occupied at frame $\ell$. We regard as recognition task the mapping of the feature vectors to their hidden states $Q_j, j = 1 \cdots N_Q$. In order to analyse the statistical relationship between the states and the corresponding feature vectors experimentally, we have to align the feature vectors to the correct states. This is performed via a forced Viterbi algorithm alignment procedure as done in the HMM training. An example of the alignments of the same word being uttered several times is shown in Figure 1 assuming a left-to-right Bakis topology where a skip is allowed.

The mapping task can be formulated as a coding problem. A source transmits a chain of symbols $\psi^M$ (the input alphabet) through a recognition channel. The channel emits the features $x^M$ (the output alphabet). The concept of entropy\(^2\) allows to determine how much information is contained in the sequence of features $x^M$ in order to reconstruct the transmitted chain $\psi^M$. In order to decode the states without errors the features have to contain the information

$$
H(Q) = - \sum_{j=1}^{N_Q} P(Q_j) \log_2 P(Q_j),
$$

where $H(Q)$ denotes the entropy of the states. To proceed with this concept we need to define some entropy terms related to the feature random variable $X$. Shannon’s conditional entropy\(^2\) or equivocation which is defined as

$$
H(Q|X) = - \sum_{j=1}^{N_Q} P(Q_j|X) \log_2 P(Q_j|X),
$$

Figure 1: Example of different feature alignments to the HMM states for the same word $W$. Example of different feature alignments to the HMM states for the same word $W$. Figure 1: Example of different feature alignments to the HMM states for the same word $W$. Figure 1: Example of different feature alignments to the HMM states for the same word $W$. Figure 1: Example of different feature alignments to the HMM states for the same word $W$. Figure 1: Example of different feature alignments to the HMM states for the same word $W$. Figure 1: Example of different feature alignments to the HMM states for the same word $W$.
describes the information missing to reconstruct the state from the feature. As shown later in Chapter 3, \( H(Q|X) \) is closely related to the error rate and defines the quality of the features. If \( H(Q|X) = 0 \), we have an error free recognition channel. The maximum value of \( H(Q|X) \) equals \( H(Q) \). In this case the features contain no information to decode (recognize) the states.

Let us also define the minimal probability of error \( P_e \) in estimating the state \( Q \) given an observed feature \( X = x \):

\[
P_e(Q|x) = 1 - \hat{Q}_{MAP},
\]

where \( \hat{Q}_{MAP} \) is the maximum a posteriori (MAP) estimator. The Bayes probability of error or Bayes risk \( P_B \) is shown as

\[
P_B = E_X \left[ P_e(Q|x) \right] = \int_X \left[ 1 - \arg \max_{d_j} p(Q|d_j|x) \right] p(x) \, dx,
\]

where \( X \) denotes the set of all possible outcomes of \( X \).

### 3. The Uncertainty Bounds

It is known that there is no one-to-one relation between the Bayes probability of error and the conditional entropy. The relation exists in the formulation of upper and lower bounds of the probability of error in terms of the entropy. While the Fano bound [3] is shown as a tight lower bound of the probability of error in terms of the entropy, the Golič bound was formulated in a general form to accommodate not only the conditional entropy as in (3), but also other concave information measures. It is shown as

\[
H(Q|X) \leq H(P_B) + P_B \log_2 (N_Q - 1),
\]

where \( H(P_B) = -P_B \log_2 P_B - (1 - P_B) \log_2 (1 - P_B) \) is the binary entropy function.

The Golič bound was formulated in a general form to accommodate not only the conditional entropy as in (3), but also other concave information measures. It is shown as

\[
H(Q|X) \geq j(j+1) \cdot \left( P_B - \frac{j-1}{j} \right) \log_2 \left( \frac{j+1}{j} \right) + \log_2 j,
\]

for \( (j-1)/P_B \leq j/(j+1) \) and \( 1 \leq j \leq N_Q - 1 \). This bound has also been derived in [5] and is valid for \( P_B \in [0, \frac{1}{N_Q}] \) and \( N_Q \geq 2 \). The error bounds are depicted in Figure 2 showing the case \( N_Q = 264 \) as used in the experimental part in Section 6. It shows that the lower bound is nearly proportional to the \( \log_2 P_B \). Given these relations we can see that the conditional entropy \( H(Q|X) \) is closely related to \( P_B \) and is suited for measuring the quality of the features.

### 4. Approximation of \( H(Q|X) \)

To evaluate \( H(Q|X) \) as given in (3) we use the relation\(^2\)

\[
H(Q|X) = H(Q) - \left[ H(X) - H(X|Q) \right],
\]

where the terms:

\[
H(X) = -\int_X p(x) \log_2 p(x) \, dx,
\]

\[
H(X|Q) = \sum_{j=1}^{N_Q} P(Q_j) H(X|Q_j),
\]

\[
H(X|Q_j) = -\int_X p(x|Q_j) \log_2 p(x|Q_j) \, dx,
\]

\(^2\) \( H(X) - H(X|Q) \) is known as the mutual information.

have to be determined. Only in special cases analytical solutions can be found. For the general case, approximations either by discrete distributions [6] or by continuous distributions such as Gaussian mixtures [7] have to be applied. To handle the distributions \( p(x|Q_j) \) and \( p(x) \) we use Gaussian approximations assuming statistical independence of the feature occurrences.

We assume that the \( d \)-dimensional multivariate conditional density \( p(x_1, \ldots, x_d|Q_j) \) is monomodal Gaussian:

\[
p(x|Q_j) = p(x_1, \ldots, x_d|Q_j) \sim N(\mu_{X|Q_j}, \Sigma_{X|Q_j}). \tag{11}
\]

Following the multivariate derivation shown in [8] we get:

\[
H(X|Q_j) = \frac{d}{2} \ln (2\pi e) + \frac{1}{2} \ln |\Sigma_{X|Q_j}|, \tag{12}
\]

with \( | \cdot | \) being the matrix determinant.

Given the approximation in (11) the distribution \( p(x) \) is a Gaussian mixture and (8) cannot be solved analytically. Given

\[
p(x) = p(x_1, \ldots, x_d) = p(x_d) \prod_{i=1}^{d-1} p(x_i | x_{i+1}, \ldots, x_d),
\]

leading to

\[
H(X) = -\int_X p(x) \log_2 \left[ p(x) \prod_{i=1}^{d-1} p(x_i | x_{i+1}, \ldots, x_d) \right] dx
\]

\[
= H(X_d) + \sum_{i=1}^{d-1} H(X_i | X_{i+1}, \ldots, X_d). \tag{13}
\]

The \( n \)-gram approximation

\[
p(x_i | x_{i+1}, \ldots, x_d) \approx p(x_i | x_{i+1}, \ldots, x_{i+n}), \tag{14}
\]

can be used. In our experiments we regard a mono- and bigram approximations \((n = 0, 1)\).

### 5. Influence of Noise on the Features

In this section we are going to use the concept of entropy to analyse the influence of noise on the features. We denote the clean speech feature random variable with \( X \) and the noisy

![Figure 2: A lower bound (LB) and upper bound (UB) of Bayes probability of error \( P_B \) with \( N_Q = 264 \).](image-url)
speech feature random variable with \( Y \). The distortion feature random variable \( E \) includes the distortions caused by the noise on the speech signal and by the artifacts of denoising algorithms. We assume that the distortions are additive and stationary:

\[
y = x + e.
\]

(15)

Furthermore, we assume that the distortion feature \( E \) is statistically independent from \( X \), i.e., \( p(x, e) = p(x) \cdot p(e) \).

Denoting the mean vectors and covariance matrices of the noisy speech and clean speech vectors with \( \mu_Y, \Sigma_Y \) and \( \mu_X, \Sigma_X \), respectively, we get the relation

\[
\mu_Y = \mu_X + \mu_E \quad \text{and} \quad \Sigma_Y = \Sigma_X + \Sigma_E.
\]

(16)

As the distortion vector \( E \) is also present on \( Y \) in each state \( Q_j \), (16) also holds on the state level:

\[
\mu_{Y|Q_j} = \mu_{X|Q_j} + \mu_E \quad \text{and} \quad \Sigma_{Y|Q_j} = \Sigma_{X|Q_j} + \Sigma_E.
\]

(17)

We also assume that the density functions \( p(e) \) and \( p(y|Q_j) \) are monomodal Gaussians leading to a multimodal Gaussian

\[
E \sim \sum_{j=1}^{N_Q} p(Q_j) \cdot p(y|Q_j).
\]

In the following, however, we also assume a monomodal Gaussian of \( p(y) \) for simplicity.

Following the Gaussian formulation in (12), we get

\[
H(Q|Y) = H(Q) - \frac{1}{2} \sum_{j=1}^{N_Q} P(Q_j) \ln \left( \frac{\sigma_{X|Q_j}^2 + \sigma_{E|Q_j}^2}{\sigma_{X|Q_j}^2 + \sigma_{E|Q_j}^2} \right),
\]

(18)

For diagonal covariance matrices we can decompose the mutual information \( \Delta = H(Y) - H(Y|Q) \) in terms of each feature dimension \( \gamma \):

\[
\Delta = \sum_{i=1}^{d} \Delta_i, \quad \Delta_i = H(Y_i) - H(Y_i|Q).
\]

(19)

where

\[
H(Y_i) - H(Y_i|Q) = -\frac{1}{2} \sum_{j=1}^{N_Q} P(Q_j) \ln \left( \frac{\sigma_{X_i|Q_j}^2 + \sigma_{E_i|Q_j}^2}{\sigma_{X_i|Q_j}^2 + \sigma_{E_i|Q_j}^2} \right).
\]

(20)

with

\[
\gamma_{i,j} = \frac{\sigma_{X_i|Q_j}^2}{\sigma_{X_i|Q_j}^2} \quad \text{SNR}_i = \frac{\sigma_{X_i}^2}{\sigma_{E_i}^2}.
\]

As shown in (21), where it employs the monogram approximation, the distortion influences \( H(Q|Y) \) via the SNR. If the SNR approaches 0 (high noise case), \( H(Q|Y) \) approaches \( H(Q) \), i.e., no information is contained in the features.

### 6. Experimental Results

#### 6.1. Setup

To investigate the conditional entropy, we have used the well matched training set of the Aurora 3 German digits database [1]. This database contains utterances which were recorded simultaneously via a noisy speech hands-free channel and a clean speech close talk channel. This set up allows the extraction of the features \( X = x \) from the clean speech channel and the features \( Y = y \) from the noisy speech channel\(^3\). According to

\(\footnote{The two channels have a time delay. The features \( x, y \) are based on the synchronized signals.} \)

(15), given \( Y \) and \( X \) the noise features \( E = e \) can be reconstructed. In this way the estimates for \( H(Q|X) \) and \( H(Q|Y) \) can be evaluated. As features we used MFCCs \((d = 39)\) combined with a least-squares based denoising method [9] to obtain from the noisy speech channel the denoised speech features \( \hat{X} = \hat{x} \). The amount of feature vectors available is \( N \approx 400000 \) and the total amount of states derived from whole word models of digits is \( N_Q = 264 \). Note that the entropies are calculated on the training set since the testing set yields insufficient data.

#### 6.2. Analysis

Firstly, an estimate of \( H(Q) \) is needed to obtain the conditional entropy as in (7). The state entropy \( H(Q) \) is a discrete entropy and is calculated as in (2). The discrete probabilities \( P(Q_j) \) are estimated from a finite set of \( N \) observed features:

\[
P(Q_j) \approx \frac{n_j}{N} \quad \text{and} \quad \tilde{H}(Q) = -\sum_{j=1}^{N_Q} \tilde{P}_j \log_2 \tilde{P}_j, \quad (22)
\]

where \( n_j \) is the number of occurrences of state \( Q_j \) and \( N = \sum_{j=1}^{N_Q} n_j \). We have obtained the value of \( \tilde{H}(Q) = 7.95 \) bits given the speech features which shows that the distribution of the state is nearly uniform.

Given the clean speech features \( x \), the entropies \( H(X|Q) \) and \( H(X) \) are calculated according to (12) and (14), respectively. Similar formulations have to be applied to the entropies related to the noisy speech \( y \) and the denoised speech \( \hat{x} \) features. \( H(X), H(Y), \) and \( H(\hat{X}) \) are calculated based on the monogram and bigram approximations and in addition to that, three cases are considered for the assumed underlying distribution, i.e., monomodal Gaussian, Gaussian mixture, and the measured distribution.

The conditional entropies are tabulated in Table 1 for the denoised speech features obtained from the combined hands-free and close talk utterances. The table shows that the monomodal assumption is generally not suitable since it tends to exceed \( H(Q) \), hence giving out negative values for the conditional entropy. The measured distribution and Gaussian mixture give reasonable estimates with the bigram approximation theoretically giving more accurate estimates since it contains more information regarding the dependencies between adjacent feature dimensions.

<table>
<thead>
<tr>
<th>( p(\hat{x}) )</th>
<th>Monogram [bits]</th>
<th>Bigram [bits]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monomodal Gaussian</td>
<td>-0.08</td>
<td>-0.08</td>
</tr>
<tr>
<td>Measured distribution</td>
<td>0.13</td>
<td>1.5</td>
</tr>
<tr>
<td>Gaussian mixture</td>
<td>0.03</td>
<td>0.18</td>
</tr>
</tbody>
</table>

Due to the different values of \( H(Q|\hat{X}) \) as shown in Table 1, we conclude that modeling the distributions is crucial. Figure 3 and 4 depict the joint probability distributions shown, we conclude, that the distributions \( p(\hat{x}|Q) \) are not monomodal Gaussians as assumed. But due to the lack of data, no reliable models for these distributions could be made.
Finally, to analyse the quality of the features the conditional entropies for \( X, \hat{X} \), and \( Y \) are evaluated\(^4\). \( H(X|Q) \) is determined following (12), where the monomodal Gaussian distributions \( p(x_i|Q_i) \) are assumed. \( H(X) \) is evaluated using the monogram Gaussian mixture assumption. In Figure 5 the terms \( \Delta_i \) decomposing the mutual information are shown. Based on the monogram assumption the decomposition is done in analogy to (20) for \( X \) and \( \hat{X} \). Compared to the results in Table 1, the denoised features are only obtained from the hands-free utterances.

The crude entropy estimation introduces inaccuracies, e.g., negative clean speech conditional entropy and a higher entropy for the denoised speech compared to the noisy speech where ideally, given the recognition rate improvement made by the denoising method as observed in [10] the following condition should hold:

\[
H(Q|\hat{X}) \leq H(Q|X) \leq H(Q|Y).
\]

(23)

Despite the inaccuracies, interesting observations can still be drawn from the figure. It shows that the clean speech features contain most information and that lower feature indices contain more information than the higher ones. These observations confirm the ability of the proposed concept to fulfill the expected behaviour. The figure also shows that a denoising method has improved the features on, for example, the first six feature dimension indices except the fourth feature index. This implies that the condition in (23) is satisfied for lower feature indices and it is valid given the assumed Gaussian mixture approximation.

Although there is no one-to-one relationship between the gained bits and the relative recognition rate improvement, it nevertheless gives us a hint that the denoising method has created a better representation of the features. Future denoising improvements can then be directed to obtain a much better improvement on all feature dimension indices.

\(^4\)In the following, entropies related to \( X \) also apply to those of \( \hat{X} \) and \( Y \).

7. Conclusions

We have derived a new method, to measure the quality of features given by the Shannon’s conditional entropy. Although crude approximations were made to estimate the entropy terms, the measure can be used to evaluate denoising methods. Better estimates of Shannon’s conditional entropy can be achieved by using more data and using more accurate models for the distributions.

8. References


