Mutual Information and the Speech Signal
Mattias Nilsson and Bastiaan Kleijn
Sound and Image Processing Laboratory
KTH School of Electrical Engineering, Stockholm, Sweden
mattias.nilsson@skype.net, bastiaan.kleijn@ee.kth.se

Abstract
Mutual information is commonly used in speech processing in the context of statistical mapping. Examples are the optimization of speech or speaker recognition algorithms, the computation of performance bounds on such algorithms, and bandwidth extension of narrow-band speech signals. It is generally ignored that speech-signal derived data usually have an intrinsic dimensionality that is lower than the dimensionality of the observation vectors (the dimensionality of the embedding space). In this paper, we show that such reduced dimensionality can affect the accuracy of the mutual information estimate significantly. We introduce a new method that removes the effects of singular probability density functions. The method does not require prior knowledge of the intrinsic dimensionality of the data. It is shown that the method is appropriate for speech-derived data.

Index Terms: dimensionality, classification, speech

1. Introduction
Statistical mappings from speech-derived data to a discrete set of classes or to a second set of continuous parameters are commonplace in speech processing. Straightforward examples include phone recognition, e.g., [1], and bandwidth extension, e.g., [2]. Mutual information forms a powerful tool in the context of statistical mappings. The accuracy of the mappings can be characterized by mutual information, e.g., [3, 4] or the performance of a statistical mapping can be optimized by directly maximizing mutual information, e.g., [1, 5]. The probability of a classification error can be bound by the Fano bound [6].

The usage of mutual information in statistical mappings implies the need for its estimation from observed data. This is straightforward if all variables involved have a discrete alphabet. However, in speech processing the acoustic features generally have continuous alphabets. If all or some of the variables have a continuous alphabet, the straightforward estimation of mutual information generally involves the estimation of differential entropies. Consider as an example the computation of the mutual information between a random discrete class index $Y$ and a $k$-dimensional random speech-derived data vector $X^k$.

This mutual information is most conveniently computed as

$$I(Y; X^k) = h(X^k) - h(X^k|Y),$$  \hspace{1cm} (1)

where $h(X^k)$ is the differential entropy of $X^k$ and $h(X^k|Y)$ is the class-conditional differential entropy of $X^k$. Useful for classification is also the Fano bound [6] on probability of classification error, $P_e$, which for continuous $X^k$ is most conveniently expressed as

$$H_{P_e} + P_e \log((1/P_e) - 1) \geq H(Y) - I(Y; X^k),$$  \hspace{1cm} (2)

where $H_{P_e} = -P_e \log(P_e) - (1 - P_e) \log(1 - P_e)$.

Recent work has shown that the estimation of differential entropies from observed data is not trivial [7, 8, 9]. The estimation must account for the dimensionality of the manifold that the data lie on, a fact that has been ignored. It is well known that speech-derived data (as is common for experimental data) often lie on a manifold within the embedding space defined by the dimensionality of the observations, e.g., [10, 11, 12]. The situation becomes particularly complicated when the intrinsic dimensionality of the data varies over the embedding space. The goal of this paper is to address the estimation of mutual information under these difficult conditions that commonly occur for speech-derived data.

The difficulty in estimating differential entropy is apparent from its definition. The differential entropy of $X^k$ is defined as

$$h(X^k) = -\int_{R^k} f_{X^k}(x^k) \log(f_{X^k}(x^k)) dx^k,$$  \hspace{1cm} (3)

where $f_X(x)$ is the density. If $f_{X^k}$ is singular, then the differential entropy becomes negative infinity. A solution is to define the differential entropy on the manifold that the data lie on. This is difficult if the manifold is not known and not possible if a multitude of manifolds with different dimensions is involved.

Whereas the definition of differential entropy often is problematic for practical data (including speech-derived data), this is not the case for mutual information. If we quantize the continuous-alphabet variables then, with decreasing quantizer step-size, any mutual information measure involving the quantized variable converges to that of the corresponding continuous variable with decreasing step size. The effect of quantization is qualitatively similar to that of additive noise, $W^k$, with a nonsingular probability density and that the mutual information $I(Y; X^k + W^k)$ converges to $I(Y; X^k)$ with decreasing noise variance. Since the differential entropies of $X^k + W^k$ are defined this means that (1) can be used to estimate the mutual information. Our approach selects the minimum noise variance required for reliable estimation of the differential entropies.

Our method builds on existing differential entropy estimation methods, which were designed for nonsingular probability density functions. Classical methods for the estimation of differential entropy can be divided into two broad classes. In plug-in estimators, the density $f_{X^k}$ in (3) is replaced by an estimate of the density $f_{X^k}$. The estimated density can be based on histograms [13], kernels [14], and autoregressive models [15]. In contrast, the class of direct estimators does not require an estimate of the probability density. This class includes methods such as order-statistics [16], nearest-neighbor distances [17, 9], and entropic spanning graphs [18]. In this paper, we focus on the direct estimators, since they are not dependent on a particular choice of modeling of the probability density function. However, the basic principles of our approach can be carried over to plug-in estimators.
In the remainder of the paper, we first discuss the principles and practical implementation of mutual information estimation based on direct estimators for differential entropy. We then provide experimental results for artificial and speech source signals. The final section provides conclusions.

2. Mutual Information Estimation with Limited Resolution

In this section, we introduce the usage of additive noise to facilitate the estimation of the mutual information based on direct estimators for the differential entropy. We first study the effect of limiting the resolution of continuous-alphabet variables by adding noise and then provide a practical method to estimate mutual information.

2.1. The Impact of Added Noise

We consider classification of a random vector of speech-derived data, $X^k$, that lies with probability one on a $k$-dimensional smooth manifold $\mathcal{M}$ embedded in the $k$-dimensional Euclidean space ($1 \leq k < K$). Thus, the probability density functions $f_{X^k}()$ is singular with respect to the Lebesgue measure in $\mathbb{R}^k$.

The class indicator variable is denoted as $Y$.

To remove the singularity of the probability density functions, we add an independent random noise vector, $W^k$, with a probability density function that is not singular in the embedding space. We compute the mutual information between the noisy data vector $X^k + W^k$ and the class indicator $Y$:

$$I(Y; X^k + W^k) = h(X^k + W^k) - h(X^k + W^k | Y).$$  (4)

Let us use a subscript for the differential entropy to define the space it is defined in:

$$h_{\mathcal{M}}(X^k) = - \int_{\mathcal{M}} f_{X^k}(x^k) \log(f_{X^k}(x^k)) dx^k. \hspace{1cm} (5)$$

It is easily derived from (5) that

$$h_{\mathcal{M}}(X^k + W^k) = h_{\mathcal{M}}(X^k + W^k) + h_{\mathcal{M} - \mathcal{M}}(W^k). \hspace{1cm} (6)$$

An equivalent equation can be derived for the class-conditional differential entropy. This implies that the mutual information between the noisy data and the classes can be written as

$$I_{\mathcal{M}}(Y; X^k + W^k) = I_{\mathcal{M}}(Y; X^k + W^k),$$

independently of the noise component in $\mathbb{R}^k - \mathcal{M}$. If the manifold structure would be known, it would be possible to add noise that is orthogonal to the manifold yielding $I_{\mathcal{M}}(Y; X^k + W^k) = I_{\mathcal{M}}(Y; X^k)$. This is generally not the case.

Constraining the resolution of the data space by adding the noise $W^k$ facilitates the measurement of the differential entropy. The cost is an underestimate of the mutual information since $I_{\mathcal{M}}(Y; X^k + W^k) \leq I_{\mathcal{M}}(Y; X^k)$ from the data-processing inequality. With decreasing noise variance, the estimate converges to the true mutual information from below.

Usage of $I_{\mathcal{M}}(Y; X^k + W^k)$ instead of $I_{\mathcal{M}}(Y; X^k)$ in (2) invalidates the inequality. That is, we can only bound the classification error under the condition that additive noise is present.

Finally we note that our logic can be generalized to the case where the probability density consists of nonoverlapping components lying on manifolds of different dimensions. In the remainder of the paper, our objective is to constrain the resolution of the space such that the intrinsic dimensionality of the data space equals the dimensionality of the observed data vectors allowing measurement of the differential entropy.

2.2. Estimating the Mutual Information

We base our estimation of mutual information on the Kozachenko-Leonenko nearest-neighbor (NN) entropy estimator [17] in combination with adding noise to the observed data. The Kozachenko-Leonenko estimator is defined as

$$\hat{h}_{\text{NN}}(X) = \frac{1}{N} \sum_{n=1}^{N} \log_2 \left( \| e_n \| (N - 1) \gamma V_k \right), \hspace{1cm} (8)$$

where $\| e_n \|$ is the Euclidean distance between an observation $x_n$ and its nearest-neighbor, $k$ denotes the dimension, and $\gamma = \exp(C_E)$, where $C_E \approx 0.5772$ is the Euler constant. The factor $V_k$ in (8) is the volume of the $k$-ball of unit radius, $V_k = \left( \frac{\pi^{k/2}}{k!} \right)$.

We assume that $N$ is sufficiently large, that we have access to the class label $c_n$ of each observation $x_n$, and that the data of a particular class have a particular intrinsic dimension. We denote the cardinality of $Y$ by $|\gamma Y|$. Starting with the set of data that belong to the class $y$, we create $M$ subsets by randomly selecting data vectors from the set. The size of the subsets should range from some $L_y$ to the maximum number of observations available for the class $L_y$, where $L_y$ is sufficiently large to satisfy the high-rate assumptions implied by (8).

If the dimensionality $k$ is not correct, then the differential entropy estimates from (8) form an affine map (linear function plus offset) of the logarithm of the number of observations (cf. [8, 9]). The slope $s$ is related to the difference between the actual intrinsic data dimensionality and the dimensionality used. If the data density is not singular, then the differential entropy estimates using (8) are independent of the number of observations. We estimate the differential entropies for each of the $M$ data subsets of class $y$ and use linear regression to fit a line to the pairs of entropy estimates and the logarithm of corresponding subset sizes. The least-squares solution to the estimate of the slope $s$ and the offset $b$ of the line is

$$\hat{\theta} = (\hat{L}^T \hat{L})^{-1} \hat{L}^T \hat{h}, \hspace{1cm} (9)$$

where $\hat{\theta} = (\hat{s}, \hat{b})^T$, and where

$$\hat{h} = [\hat{h}_{\text{NN}}(X|Y = y; L_y), ..., \hat{h}_{\text{NN}}(X|Y = y; L_y)]^T, \hspace{1cm} (10)$$

$$\hat{L} = \left[ \log_2 \left( \frac{L_y}{1} \right), ..., \log_2 \left( \frac{L_y}{1} \right) \right]^T. \hspace{1cm} (11)$$

In (10), $\hat{h}_{\text{NN}}(X|Y = y; L_y)$ denotes the entropy estimate using a subset of size $L_y$ from class $y$.

We initially perform the dimension estimation with no added noise. If the estimated slope $\hat{s} < -\epsilon$ (where $\epsilon$ represents some numerical tolerance) we add noise $W$ to the data and perform the regression of (9) using the resulting noisy data. We search for the lowest noise variance that results in $\hat{s} \geq -\epsilon$. An effective search procedure is the binary search.

When all classes have been processed as described above, the unconditioned differential entropy $\hat{h}_{\text{NN}}(X^k + W^k)$ is estimated using (8) from all observed data with the added class-conditioned noise. Given the unconditioned differential entropy and the class-conditioned differential entropies we can estimate the mutual information between the observed data and the classes from (4).

The procedure is summarized in Table 1, Table 2 and Table 3. Note that when all intrinsic dimensionalities equal the dimensionality of the embedding space the mutual information estimate is the conventional estimate, since no noise is added in that case.
Table 1: Estimation of mutual information.

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Set class counter $y = 1$ and the class-conditional noise level $\sigma_{W</td>
</tr>
<tr>
<td>2.</td>
<td>Select $M$ subsets of the class-conditional observations ranging in size from $L_y$ to $L_y$ (the maximum number of observations available for class $y$).</td>
</tr>
<tr>
<td>3.</td>
<td>Estimate $\hat{h}_{\text{NN}}(X+W</td>
</tr>
<tr>
<td>4.</td>
<td>Use (9) to find $\hat{s}$.</td>
</tr>
<tr>
<td>5.</td>
<td>If $\hat{s} &lt; -\epsilon$ increase $\sigma_{W</td>
</tr>
<tr>
<td>6.</td>
<td>If $y &lt; \Omega_Y$ set $y = y + 1$, $\sigma_{W</td>
</tr>
<tr>
<td>7.</td>
<td>Estimate $\hat{h}_{\text{NN}}(X+W)$ using (8) from all observations added with the class-conditional noise (obtained from the previous steps).</td>
</tr>
<tr>
<td>8.</td>
<td>Estimate $I(X + W; Y)$ using (4).</td>
</tr>
</tbody>
</table>

Table 2: Results for the artificial experiment described in Section 3.1. The true mutual information accounting for the resolution constraint ($\Delta = 0.09$ for the class $A$ data) is 0.93 bits.

<table>
<thead>
<tr>
<th>Number of observations $N$</th>
<th>True MI (bits)</th>
<th>Estimated MI (constr. res.)</th>
<th>Estimated MI (conventional)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.92 ± 0.01</td>
<td>1.19 ± 0.01</td>
<td></td>
</tr>
</tbody>
</table>

The results of the reference experiment for 20,000 observations are shown in Table 2. Whereas straight application of (8) results in a mutual information estimate of 1.19, the proposed method results in a mutual information estimate of 0.92. We note that the value 1.19 is physically impossible since we have only two classes, corresponding to 1 bit. More importantly, the results of the proposed method converge towards the correct answer with increasing data size, whereas straight application of (8) does not. This is illustrated in Fig. 2, where it is evident that for the noise level used ($\Delta = 0.1$), the estimates of the differential entropies have converged to values that correspond to $I(X^3; Y) = 0.92$. This is consistent with the theoretical result at this noise level, which is 0.93 (we used uniformly distributed noise to facilitate this comparison with theory). More accurate answers can be obtained by simply increasing the number of observed data. In contrast, the conventional approach using straight application of (8) does not result in convergence.

3. Experimental Results

In this section, we show that our algorithm is appropriate for the estimation of mutual information of speech-derived data. We first show with artificial data that the procedure gives the expected results. We then show that the intrinsic dimensionality of speech data can vary over the embedding space, which means that conventional methods for estimating the mutual information will fail. Finally we provide an example application of the method to a speech-data classification problem.

3.1. Reference Experiment

To verify the effectiveness of the method, we consider example data that are divided into two classes that we label $A$ and $B$. The random data vectors $X^3$ are embedded in a three-dimensional Euclidean space and the class variable $Y$ takes two discrete values. As is illustrated in Fig. 1, the data of class $A$ data are located on a two-dimensional planar manifold and the data of class $B$ are located on a three-dimensional manifold. As the overlap of the data of class $B$ with the plane vanishes, the actual mutual information between the data and the two classes is one bit, i.e., $I(X^3; Y) = 1$.

In the experiment we made up to 20,000 observations and we use $M = 5$ subsets and a slope threshold of $\epsilon = 0.1$. We added noise with a uniform distribution in each dimension. The optimal noise levels were found to be $\Delta_A = 0.09 \pm 0.015$ for the class $A$ data and $\Delta_B = 0.0$ for the class $B$ data.

Figure 1: Reference experiment.

Figure 2: Convergence of differential entropies in reference experiment. Circles represent the estimates with noise added (our algorithm) of $h(X + W)$ (solid), $h(X + W|Y = A)$ (dotted), and $h(X + W|Y = B)$ (dashed), respectively. The corresponding estimates without noise added (conventional) are marked with squares.

3.2. Relevance to Speech

To illustrate the applicability of the proposed procedure we consider the parameters of an order-10 autoregressive (AR) models, as determined by linear predictive (LP) analysis. Such data have previously been shown to lie on a manifold embedded in the embedding space [12]. The predictor coefficients are commonly used to characterize the short-term power-spectral enve- loppe of the speech signal. As parameters we use the line spectral frequencies (LSF) [19], which are well-behaved [20]. We first estimate the dimensionality for regions around random points as described in [7, 9]. We find that the intrinsic dimensionality varies from 7 to 10. Low intrinsic dimensionality corresponds to strongly colored power spectra and high intrinsic dimensionality corresponds to relatively flat power spectra.

The low intrinsic dimensionality strongly colored power spectra is likely the result of physiological constraints on the human vocal tract. The resulting variations in the intrinsic dimensionality of the LSF parameter set must be accounted for in mutual information estimates, thus motivating the usage of our methods for speech data.
3.3. Classification of Speech Vowels

We now apply our methods to the classification of speech vowels. We consider cepstral coefficients for the two vowel classes /aa/ and /ix/. The cepstra are computed from linear prediction coefficients estimated from 20 ms segments. Each class had about 50,000 data. The class probabilities are, therefore, equal and $H(Y) = 1$.

The estimation of the dimensionality yields 5 and 7 for respectively /aa/ and /ix/. Table 3 shows the results for the mutual information estimation. The mutual information between the classes can not be more than 1 bit since there are only 2 classes. Yet straight application of (8) yields a value of 1.44, whereas our procedure provides a reasonable 0.777. For only 2 classes. Yet straight application of (8) yields a value of 1 bit since there are equal and $H(Y) = 1$.

The estimation of the dimensionality yields 5 and 7 for respectively /aa/ and /ix/. Table 3 shows the results for the mutual information estimation. The mutual information between the classes can not be more than 1 bit since there are only 2 classes. Yet straight application of (8) yields a value of 1.44, whereas our procedure provides a reasonable 0.777. For only 2 classes. Yet straight application of (8) yields a value of 1 bit since there are equal and $H(Y) = 1$.

<table>
<thead>
<tr>
<th>Dimensionality</th>
<th>Number of hits</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 3: Results for the classification of /aa/ and /ix/.

4. Conclusions

We confirmed that commonly used speech-derived data have an intrinsic dimensionality that is lower than that of the embedding space. This means that conventional estimation procedures for the differential entropy are not accurate. Our method of controlled noise addition leads to a reliable estimation of mutual information for speech-derived data. When the number of observations is increased our method converges asymptotically to the true mutual information from below.

5. References