MAXIMUM MUTUAL INFORMATION BASED REDUCTION STRATEGIES FOR CROSS-CORRELATION BASED JOINT DISTRIBUTIONAL MODELING

Jeff A. Bilmes
<bilmes@cs.berkeley.edu>

International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704, USA

CS Division, Department of EECS
University of California at Berkeley
Berkeley, CA 94720, USA

ABSTRACT

In maximum-likelihood based speech recognition systems, it is important to accurately estimate the joint distribution of feature vectors given a particular acoustic model. In previous work, we showed we can boost accuracy in this task by modeling the joint distribution of time-localized feature vectors along with statistics relating those feature vectors to their surrounding context. In this work, we evaluate information preserving reduction strategies for those statistics. We claim that those statistics corresponding to spectro-temporal loci in speech with relatively large mutual information are most useful in estimating the information contained in the feature-vector joint distribution. Furthermore, we claim that such statistics are most likely to generalize. Using an EM algorithm to compute mutual information between pairs of points in the time-frequency grid, we verify these hypotheses using both overlap plots and speech recognition word error results.

1. INTRODUCTION

A primary goal in maximum-likelihood based speech recognition is to accurately estimate the joint distribution of acoustic feature vectors for a given statistical model. That is, we wish to estimate the *feature-vector joint distribution* $P(X_1^T|M)$ where $\{X_1^T\} = \{X_1, \ldots, X_T\}$ and X_t is a time-localized feature vector. Improving the accuracy of $P(X_1^T|M)$ can often improve discriminability between different models and therefore reduce word error rate.

Hidden Markov models (HMMs) are the most commonly assumed underlying statistical model under which $P(X_1^T|M) = \sum_Q \prod_t P(X_t|Q_t, M)P(Q_t|Q_{t-1}, M)$, or the sum is replaced by a max under the Viterbi approximation. Q corresponds to variables comprising the hidden Markov chain.

Under the HMM assumptions, X_t is conditionally independent of its past given the current hidden Markov state variable Q_t. An important sub-goal, therefore, is the estimation of $P(X_t|Q_t)$. Because natural signals are not actually governed by HMMs, it is unrealistic to assume that the burden for determining the context-dependent distribution of X_t can be placed completely on Q_t. Representing the context as $\neg X_t = \{X_1, \ldots, X_{t-1}, X_{t+1}, \ldots, X_T\}$, this essentially is a requirement that $I(\neg X_t; Q_t) \geq I(X_t; \neg X_t)$ and $I(Q_t; X_t) \geq I(X_t; \neg X_t)$ where $I(Y; Z)$ is the mutual information between variables Y and Z. A necessary condition, therefore, is that the number of states have the lower bound $|Q_t| \geq 2^{I(X_t;\neg X_t)}$. Even with this condition satisfied, however, there is no guarantee training algorithms can find the correct mapping between hidden state Q_t and acoustic distribution X_t. This is the conditional independence problem associated with HMMs.

There have been many attempts to model the feature-vector joint distribution more accurately. Some involve estimating the short-time joint distribution of feature vectors [5] and others add to that a conditional dependence on additional variables along with a distribution of those variables [7].

In [1], we showed that we can model information contained in the feature-vector joint distribution by modeling the joint distribution of time-localized feature vectors along with statistics relating those feature vectors to their surrounding context, i.e.:

$$P(X_t, \frac{d}{dt}X_t, M(\bigcup_{l \in \mathcal{C}_t} E_s[X_t X_l^T])|Q_t), \qquad (1)$$

where $\mathcal{C}_t$ is a context around time point t, $E_s[]$ is the short-time expected value over a duration of length s, and $M(\cdot)$ is an information preserving reduction strategy. The delta ($\frac{d}{dt}X_t$) and the cross-correlational ($\bigcup_{l \in \mathcal{C}_t} E_s[X_t X_l^T]$) features determine short-time statistics relating the base-feature vector X_t to its surrounding context. It can be argued, therefore, that this captures information about the underlying feature-vector joint distribution.

The correlation information is estimated using the *modcrossgram* (MCG), a way of computing the cross-correlation between feature channels:

$$R_{i,j}(t, \ell) = \sum_{k=0}^{N} x_i(t+k)x_j(t+k+\ell)w_k,$$

where x_i is the i^{th} feature channel, t is the starting offset within the signals, ℓ is the correlation lag, $N+1$ is the number of points used to compute the correlation (corresponding to s in $E_s[]$ of Equation 1), and w_k are windowing coefficients. ℓ ranges over a time-span corresponding to the range of $\mathcal{C}_t$ in Equation 1. Using compressed envelopes as feature channels, we demonstrated a significant word error rate reduction in a noisy test condition with such a model.

In [1], we chose a data-independent reduction strategy $M(\cdot)$. In this paper, we argue that we can obtain a better reduction strategy by retaining MCG coefficients corresponding only to pairs of time-feature points in training data with a *strong* statistical dependence (i.e., large mutual information). For a given number of system parameters, we believe such a strategy will more accurately represent information contained in the feature-vector joint distribution and generalize better than a reduction strategy using weaker statistical dependencies.

Section 2 introduces our hypotheses. Section 3 describes our method for computing mutual information. Section 4 demonstrates how strong statistics in speech generalize better than weaker ones. Section 5 augments our evidence with word error rate improvements. Section 6 attempts to gain intuition via an information density plot. And Section 7 concludes and describes future work.

2. STRONG STATISTICS ARE ACCURATE AND GENERALIZE

The dependencies between a collection of random variables $\mathbf{y}_1, \ldots, \mathbf{y}_N$ (governed by $P(\mathbf{y}_1, \ldots, \mathbf{y}_N)$) can be described as

ICASSP'98
IEEE International Conference on Acoustics, Speech, and Signal Processing, Seattle, 1998

a graph where each node represents a variable and each edge represents a dependence relation between its two variables[8]. Chow showed that the best tree-dependent approximation (one whose graph is a tree) of such a distribution, in terms of least Kullback-Leibler distance, can be constructed by finding a maximum weight spanning tree of the original graph with edge weights set as the mutual information between the corresponding two variables [2].

Assuming, as we argue in [1], that the joint distribution of X_t and the statistics $\bigcup_{l \in \mathcal{C}_t} E_s[X_t X_l^T]$ captures important information about the feature-vector joint distribution, what fixed size subset of the statistics can best represent this distribution? With motivation from Chow's results, we argue that a subset containing *strong* statistics are better than a subset containing *weak* ones, where we define a strong and weak statistic according to the relative magnitude of mutual information [3] between the corresponding features elements in a training set. Intuitively, if we could choose only one of the two random variables Y and Z as an information source about variable X (i.e., we could either model $P(X|Y)$ or $P(X|Z)$), w.l.o.g., we would choose Y if $I(X;Y) > I(X;Z)$. Similarly, a variable Z that is informative about $X_{t,i}$ (the i^{th} element of feature vector X_t) should have relatively large $I(Z; X_{t,i})$. In other words, if we define the quantity $I(i,j,\ell) = I(X_{t,i}; X_{t-\ell,j})$ (where variation over t defines the sampling ensemble), then the set of K feature pairs with strongest statistics are defined as:

$$\{(i_k, j_k, \ell_k) : 1 \le k \le K\} = \underset{i,\,j,\,\ell}{\operatorname{arg}K\!\max}\ I(i,j,\ell).$$

Assuming K is large enough so all elements of X_t are included, these pairs should provide significant information about X_t. This, therefore, defines a simple MCG reduction method: choose the (size $2K$ because of symmetry) set:

$$\{R_{i_k,j_k}(t,\ell_k), R_{j_k,i_k}(t,-\ell_k) : 1 \le k \le K\}.$$

The use of this set, therefore, should be more informative about X_t than an equal sized weaker set.

Another source exists, however, providing information about $X_{t,i}$, namely Q_t. We must choose variables therefore that not only have information about $X_{t,i}$, but also do not include information already provided by Q_t about $X_{t,i}$. In other words, we must choose a variable $Z \in \{X_{t-\ell,j} : \forall \ell, j\}$ such that $I(X_{t,i}; Z|Q_t)$ is large. We could estimate $I(X_{t,i}; Z|Q_t)$ directly and choose Z accordingly. Alternatively, one heuristic is to chose Z to maximize the upper bound $I(X_{t,i}; Z|Q_t) \le I(X_{t,i}; Z) - I(Q_t; Z)$[3]. That is, choose Z such that $I(X_{t,i}; Z)$ is large and $I(Z; Q_t)$ is small. $I(Z; Q_t)$ will more likely be small if Z is chosen from a time frame other than t. But because $I(X_{t,i}; Z) - I(X_{t,i}; Q_t) \le I(X_{t,i}; Z|Q_t) \le I(X_{t,i}; Z) + I(X_{t,i}; Q_t|Z)$ [3], choosing Z with a large $I(X_{t,i}; Z)$ can produce a range of larger possible values for $I(X_{t,i}; Z|Q_t)$ than if we chose Z otherwise. An even simpler heuristic, therefore, is to choose Z with large $I(X_{t,i}; Z)$. This, as suggested above, is what we investigate in this paper.

Will, however, the set of strong correlation pairs obtained from training data generalize to test data and noisy conditions? We argue that both the strongest and weakest statistics of a natural object (i.e., either a visual or an auditory object) are the most consistent with respect to samples of that object, with the strongest statistics being the most descriptively efficient. The strong statistical features of an object determine its defining characteristics, but the defining characteristics of an object are those traits that generalize best across all object instances. Furthermore, the strong statistical features will be those that remain most salient under distortions (i.e., noise). We explain with a simple analogy, descriptions of a *drinking cup*. "A cup has a concave cavity for holding liquid, is small enough to be held, has a relatively level base, etc." Such strong comparative statistics will be descriptively correct for almost all examples we encounter. "A cup has a handle, it is colored blue, it is translucent, etc." Such middle-weight statistics are true at times, false at others. "A cup looks like an elephant, acts like an inviscid compressible fluid, etc." Such weak statistics are almost always descriptively incorrect. Furthermore, we could describe a cup by listing all its unrepresentative properties, but greater descriptive efficiency is achieved by listing the archetypal features. We argue that the same phenomena holds for an auditory object, and in particular, descriptions thereof comprised of mutual information between pairs of auditory features. In Section 4 we provide empirical evidence for this phenomena in real speech data.

3. MUTUAL INFORMATION COMPUTATION USING EM

The mutual information between two continuous random variables X and Y is defined as [3]:

$$I(X;Y) = \int p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dx dy.$$

In practice, we have access only to samples $\{(x_i, y_i) : 1 \le i \le N\}$ drawn from the distributions governing X and Y. We therefore must use an estimation method.

One method computes a 2-dimensional histogram and then performs a discrete version of the above computation. Another method assumes that X and Y are jointly Gaussian distributed with correlation coefficient ρ_{xy}. If so, the quantity can be computed analytically [4] as $I_l(X;Y) = -\frac{1}{2}\log_2(1 - \rho_{xy}^2)$. Because ρ_{xy} captures the linear dependence between X and Y regardless of their joint distribution, we call this the linear mutual information. A third method fits a Gaussian mixture distribution to the sampled data using, say, EM. Unfortunately, we know of no analytical formula that, like in the single component case, maps from mixture parameters to mutual information. We can instead use numerical integration, or even simpler, sample the resulting distribution for the discrete computation.

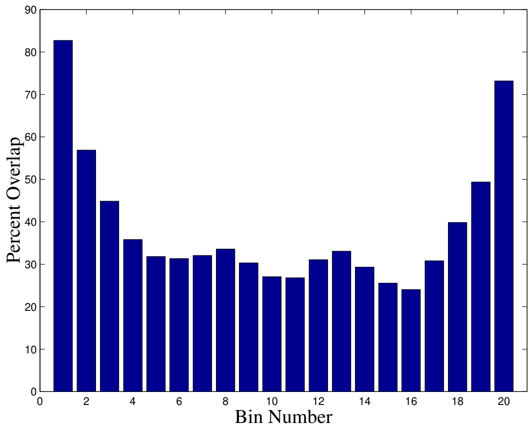

Figure 1: Statistical overlap between two independent 30 minute sections of digits+.

While the first non-parametric method is fairly simple, it suffers from several problems including the need for bias correction [6], a large number of histogram bins, and large amounts of "training" data. Because we need to compute thousands of mutual information values, this approach is not viable since thousands of simultaneously maintained 2-dimensional histograms would be required. Also, while the linear mutual information approximation is computationally simple, it does not capture potentially important non-linear statistical dependencies contained in distributions not well approximated by a single component Gaussian.

Therefore, we chose the third parametric method which we call I_{mg}. We use the EM algorithm to fit a 5 component full covariance

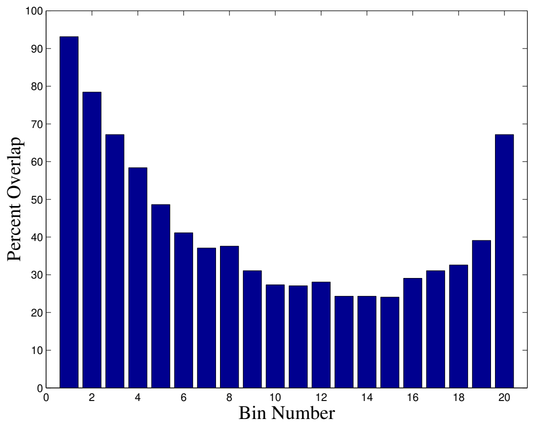

Figure 2: Statistical overlap between two independent 64 minute sections of Switchboard.

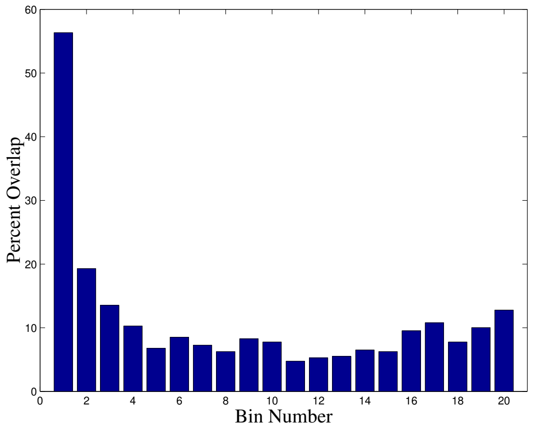

Figure 3: Statistical overlap between a 64 minute Switchboard and 30 minute digits+ section.

Gaussian mixture to each data set (for our data sets, more than 5 components showed no appreciable benefit). The resulting density is sampled at points on a 250x250 grid (again, greater values did not appreciably effect results). In each dimension $d = 1, 2$, the grid spans the range $[\min_i(\mu_{i,d} - 3\sigma_{i,d}), \max_j(\mu_{j,d} + 3\sigma_{j,d})]$ where $\mu_{i,d}$ and $\sigma_{i,d}$ are the mean and standard deviation of component i for dimension d. This surface is normalized and used in the discrete computation. With one mixture component, the method produces results almost identical to linear mutual information. For a larger number of components, the resulting values almost always get larger (i.e., we empirically find $I_{mg} \geq I_l$) indicating the addition of non-linear ingredients of the true mutual information.

4. OVERLAP OF STRONG STATISTICS IN SPEECH

In this section, we verify the hypothesis that the strongest statistics in speech generalize better than weaker ones. We look at mutual information between pairs of points in the time-frequency plane – we find those pairs with large value in one speech corpus are also large in an independent corpus.

Each corpus has telephone bandwidth and is processed by a 22 channel quarter-octave FIR filter bank. Each channel is then rectified, band-pass filtered (restricting the modulation energy to a range between roughly 1 and 35Hz), downsampled to 80Hz, and cube-root compressed. The resulting envelopes define the time-frequency grid X_t. We compute $I_{mg}(i, j, \ell) = I_{mg}(X_{t,i}; X_{t-\ell,j})$ for various i, j and for $\ell \in 0 \ldots 17$ (0 to 200ms into the past). We compute $K = 22\times22\times17 - 22\times23/2 = 7975$ mutual information values (the 253 upper triangular values for which $\ell = 0$ and $j \leq i$ are not needed for obvious reasons).

For each corpus $\mathcal{C}$, we obtain the sorted set $I_{\mathcal{C}}(k) = \{(i_k, j_k, \ell_k) : 0 \leq k < K\}$ such that $I_{mg}(i_k, j_k, \ell_k) \geq I_{mg}(i_m, j_m, \ell_m)$ for $k < m$. We divide the index range $0 \leq k < K$ into N equal sized subsets $R_n : 1 \leq n \leq N$ such that $R_n = \{k : (n-1)K/N \leq k < nK/N\}$. For two corpora $\mathcal{C}$ and $\mathcal{D}$, we define the percentile-region overlap ratio as:

$$O_{\mathcal{C},\mathcal{D}}(n) = \frac{|I_{\mathcal{C}}(R_n) \cap I_{\mathcal{D}}(R_n)|}{|R_n|},$$

where $1 \leq n \leq N$ is the bin number, and $I_{\mathcal{C}}(R_n)$ are the set of time-frequency pairs contained in the n^{th} bin. For example, if $N = 20$, $O_{\mathcal{C},\mathcal{D}}(1)$ indicates the percentage of the strongest 5% of time-frequency pairs in corpus $\mathcal{C}$ that are also contained in the strongest 5% in corpus $\mathcal{D}$, $O_{\mathcal{C},\mathcal{D}}(2)$ indicates the percentage of overlap in the non-intersecting next strongest 5%, etc.

For all plots, $N = 20$ and each two corpora has **no** speaker overlap. Figure 1 shows the plot for two independent 30 minute sections of *digits+*, a telephone quality database of isolated digits and control words from Bellcore. The figure shows a relatively large overlap for both strong statistics (lower bins numbers) and weak statistics (higher bin numbers) and the middle-weight statistics show lower overlap. Figure 2 shows the plot for two independent 64 minute randomly-chosen sections of the Switchboard database. We see a trend similar to the previous figure. The next two plots show inter-corpora prediction. Figure 3 (resp. figure 4) shows the plot between a 64 minute section of switchboard and a 30 minute section of digits+ (resp. digits+ corrupted by 10db SNR additive car noise recorded over a cellular phone). In both plots, the strong statistics have a significantly greater overlap than the weaker ones. Therefore, it seems likely that if we select a set of strong statistics from one corpora, they are likely to generalize.

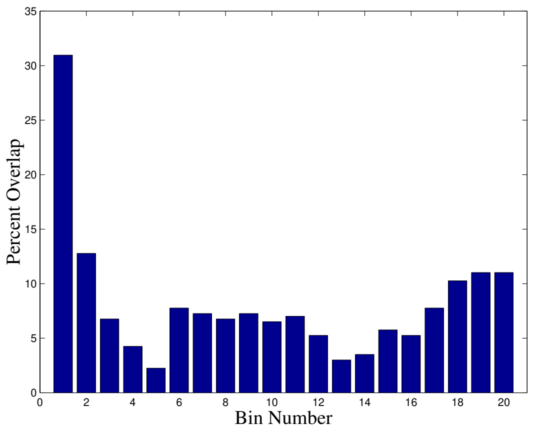

Figure 4: Statistical overlap between a 64 minute Switchboard and 30 minute 10dbSNR digits+ section.

5. SPEECH RECOGNITION RESULTS

To lend further evidence to our hypothesis, we evaluated both strong and weak MCG features in a hybrid artificial neural network/hidden Markov model (ANN/HMM) speech recognition system [5] using

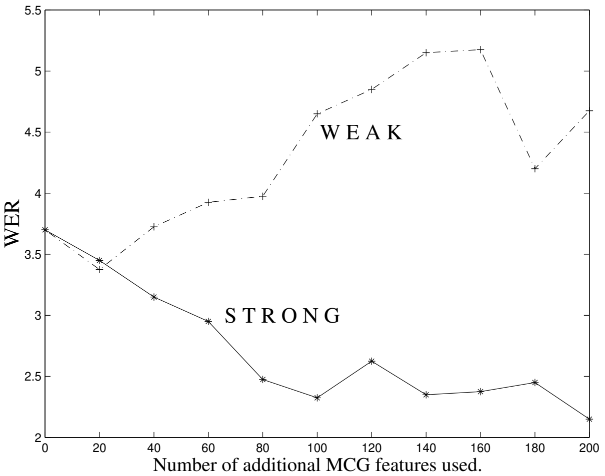

Figure 5: ASR word error rate (WER) results for digits+ comparing strong and weak MCG features used in addition to baseline.

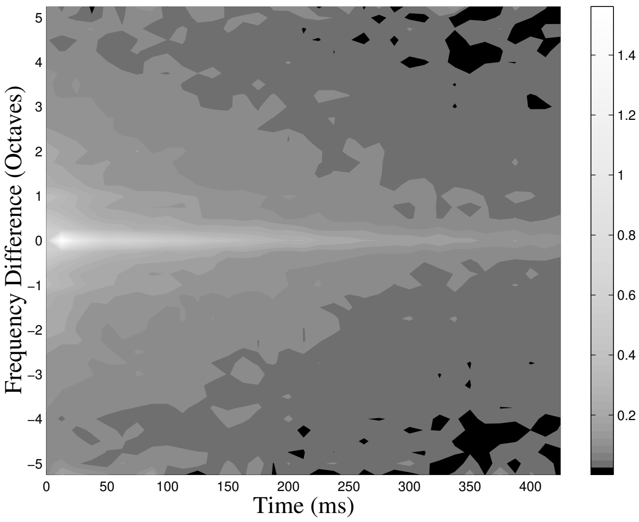

Figure 6: The information density of a randomly selected 2-hour section of Switchboard (in bits per unit area).

digits+. The MCG parameters are the same as in [1]. Each word error rate (WER) displayed is obtained using data from 200 speakers totaling 2600 examples from 4 jackknifed cuts – scores shown are the average of 4 tests in which 150 speakers were used for training and 50 different speakers used for testing. Before each ANN training, all weights were set to small random values.

The far left of Figure 5 shows the baseline score consisting of 1 frame of JRASTA [5] features plus deltas (17 total) and a 572 hidden unit ANN probability estimator. Moving to the right, each point shows the addition of MCG features, in increments of 20, starting from the strongest and moving down or the weakest and moving up in strength. The number of hidden units of the ANN is adjusted to equalize the number of free parameters in each case.

As can be seen, the addition of strong MCG features significantly reduces the error rate (at 200 strong MCG features, the WER is insignificantly different then our best result [1]). The addition of weak MCG features, however, significantly increases the error rate probably because the number of hidden units decreases while adding useless input features. This plot therefore demonstrates that we can boost recognition performance by adding strong MCG features, presumably due to more accurate estimation of information contained in the feature-vector joint distribution.

6. WHERE IS THE INFORMATION IN SPEECH?

Where in the time-frequency plane can we find information about $X_{t,i}$? Figure 6 shows the information density (i.e., $I_{mg}(d,\ell) = avg_{i-j=d} I_{mg}(i,j,\ell)$) in bits per unit area spanning 425ms into the past computed from a 2 hour random selection of Switchboard. As can be seen, significant information can be gained about $X_{t,i}$. Of course, the mutual exclusivity of this information as well as the quantity already provided by Q_t is not shown. The degree to which the HMM assumptions are invalid, however, should correspond to the degree to which the addition of information in the surrounding acoustics is useful about $X_{t,i}$. Our speech recognition results seem to indicate that strong information in the surrounding acoustic context is indeed useful.

7. CONCLUSIONS

We claim that strong statistics (i.e., correlation between pairs of time-frequency points with large relative mutual information) are better both for estimating the actual feature-vector joint distribution and for generalization than are weaker statistics. We provide evidence for these claims using both overlap plots and speech recognition word error rate results.

We are currently working on methods to determine which set of strong MCG features are best and how they should be represented (i.e., should further transformations be performed). Ultimately, we would like to select MCG features based on a better indication of their utility such as a large $I(X_{t,i};Z|Q_t)$ or a measure that also considers their informational mutual exclusivity (e.g., large $I(X_{t,i};Z_1|Q_t)$, $I(X_{t,i};Z_2|Q_t,Z_1)$, etc.).

This work has benefited from many useful discussions with Dan Ellis, Jeff Zweig, Nelson Morgan, Steve Greenberg, and other members of the speech group at ICSI. This work has been partially sponsored by JSEP contract F49620-94-C-0038 and ONR URI Grant N00014-92-J-1617.

8. REFERENCES

[1] Jeff A. Bilmes. Joint distributional modeling with cross-correlation based features. In *Proc. ASRU*, Santa Barbara, December 1997. IEEE.

[2] C.K. Chow and C.N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Trans. on Info. Theory*, 14:462–467, 1968.

[3] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, 1991.

[4] Solomon Kullback. *Information Theory And Statistics*. Dover, 1968.

[5] N. Morgan and H. Bourlard. Continuous speech recognition. *IEEE Signal Processing Magazine*, 12(3), May 1995.

[6] A.C. Morris. *An information-theoretical study of speech processing in the peripheral auditory system and cochlear nucleas*. PhD thesis, Institut National Polytechnique De Grenoble, 1992.

[7] M. Ostendorf, V. Digalakis, and O. Kimball. From HMM's to segment models: A unified view of stochastic modeling for speech recognition. *IEEE Trans. Speech and Audio Proc.*, 4(5):360–378, September 1996.

[8] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.