Extraction of Relevant Speech Features using the Information Bottleneck Method

Ron M Hecht[1] *and* *Naftali Tishby*[2]

[1] Department of Computer Science, Tel-Aviv University
[2] School of Engineering and Computer Science, Hebrew University, Jerusalem, Israel
hechtr@post.tau.ac.il tishby@cs.huji.ac.il

Abstract

We propose a novel approach to the design of efficient representations of speech for various recognition tasks. Using a principled information theoretic framework – the Information Bottleneck method – which enables quantization that preserves relevant information, we demonstrate that significantly smaller representations of the signal can be obtained that still capture most of the relevant information about phonemes or speakers. The significant implications for building more efficient speech and speaker recognition systems are discussed.

10.21437/Interspeech.2005-182

1. Introduction

The extraction of the *relevant* features of speech signals is a long- standing challenge in signal processing. Feature extraction should be guided by the task, such as phoneme recognition, or speaker identification. Most speech recognition algorithms, however, do not take this ultimate goal into account at the front-end feature extraction level. Here we present a novel approach for automatic detection and selection of the most relevant features of speech signals using the information bottleneck method.

The information bottleneck method [1] aims at selecting a compact set of features $\hat{X}$ representing a much larger data variable X, which preserves high mutual information about a target variable Y. This method has been successfully employed in several applications such as document categorization [2], image classification [3], analysis of neural codes [4], and others.

Here we apply the method to a commonly used representation of speech signals – the Mel-cepstrum[9]. We perform a vector quantization of the mel-cepstrum feature set, and use the resulting quantization as our initial data-space. We then extract compact representations of the data-space that preserve information about our target variables via clusters obtained through the agglomerative information bottleneck procedure (AIB) [5]. This procedure takes into account the ultimate goal of recognition by iteratively merging cluster pairs which lead to the minimal reduction of mutual information with the target variable, phonemes in one case and speaker's identity in the other. We show that this procedure efficiently preserves the relevant information for either phoneme recognition or speaker recognition, and makes it possible to use a much smaller representation without reducing recognition relevant information. We show that starting with a higher number of codebook vectors and reducing them using the AIB preserves significantly more relevant information than starting with a quantization of the same size. This has obvious implications for designing efficient speech recognition systems. While we limit ourselves in this paper to discrete models (vector quantization), the method can be extended to continuous (GMM) models as well.

2. Methods

In this section we describe the information theoretic principle, the feature extraction procedure and the AIB algorithm applied, and the database we used.

2.1. Relevant Speech Quantization

Speech is a complex signal with a high entropy rate. It contains ample information about the various components of its acoustic structure, such as the spoken language, specific utterances, identity of the speaker, his/her physical conditions, mood, etc. Yet most speech processing algorithms employ standard front-end processing that eliminates much of the entropy of the signal, but do it in a universal – task independent – way. Such a representation is bound to contain irrelevant information which thus reduces the efficiency and performance of the recognizer that follows. It is therefore a fundamental problem in speech technology to filter out only (or mostly) the task-relevant components of the signal. This is a difficult problem, as the relevant acoustic distortion measure is unknown and involves both complex perceptual and linguistic variables. Our approach to this problem is to utilize the available tagging of the signal (phonemes or speakers) to guide the selection of its representation.

We begin with the joint representation of the speech signal, denoted by *X,* and its relevant labeling signal – whether phonemes, speaker identity, or other attributes – denoted here by *Y*. We then estimate their joint

distribution, *P(x,y)*. The amount of relevant information in X about Y is determined by Shannon's mutual information between the variables, defined as:

$$I(X;Y) = \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \qquad (1)$$

For continuous variables the sum should be replaced by the appropriate integral. Our goal is to find a compact representation of X, denoted by $\hat{X}$, that on one hand compresses it by minimizing the mutual information between them, $I(X;\hat{X})$, and on the other preserves as much as possible the information about Y, $I(\hat{X};Y)$. To this end we minimize the Information-Bottleneck variational functional,

$$L[p(\hat{x}\,|\,x)] = I(X;\hat{X}) - \beta I(\hat{X};Y) \qquad (2)$$

with respect to the (stochastic) mapping $p(\hat{x}\,|\,x)$, where β is a positive Lagrange multiplier . This is similar to the procedure of Rate-Distortion Theory – but without using an explicit distortion function. The solution to this variational problem yields an iterative converging algorithm for the mapping – thus the reduced representation – given *P(x,y)*, for any value of the parameter β. This procedure generates the optimal relevant quantization of the variable *X* with respect to *Y*, where the complexity of the quantization is controlled through the variable β (see [1]).

2.2. Agglomerative Information Bottleneck

A greedy approximation to the above optimization problem was introduced in [5]. This is done using an agglomerative greedy hierarchical clustering algorithm which merges *x* points that result in the smallest loss of mutual information about *Y*. The algorithm starts with a trivial partition of singleton clusters where each element of the data *X* is in its own cluster (we denote the cluster set by $\hat{X}$). At each step of the algorithm we merge two conditional distributions in the current partition into a single distribution in a way that locally minimizes the loss of mutual information on the relevant variable *Y*. The value of the information loss in merging the distributions $p(y\,|\,\hat{x}_i)$ and $p(y\,|\,\hat{x}_j)$ is given by the Jensen-Shannon divergence [6] between the distributions:

$$\begin{aligned} &JS_{\pi}\left[p(y\,|\,\hat{x}_i), p(y\,|\,\hat{x}_j)\right] = \\ &\pi_i D_{KL}\left[p(y\,|\,\hat{x}_i)\,|\,\pi_i p(y\,|\,\hat{x}_i) + \pi_j p(y\,|\,\hat{x}_j)\right] + \\ &\pi_j D_{KL}\left[p(y\,|\,\hat{x}_j)\,|\,\pi_i p(y\,|\,x_i) + \pi_j p(y\,|\,\hat{x}_j)\right] \end{aligned} \qquad (3)$$

where $D_{KL}\left[p\,|\,q\right]$ is the Kullback Leibler divergence:

$$D_{KL}\left[p\,|\,q\right] = \sum p(x) \log \frac{p(x)}{q(x)} \qquad (4)$$

and $\pi_i = p(\hat{x}_i)/(p(\hat{x}_i) + p(\hat{x}_j))$ for each *i*. This greedy merging yields hierarchical clusters that provide a simple approximation to the optimal solution of the IB problem in many cases. Throughout the merges we monitor the amount of information preserved by clusters $\hat{X}$ about Y, $I(\hat{X};Y)$, and compare it to the original value of *I(X;Y)*. This way we can stop the merging when we reach an acceptable level of relative information loss.

2.3. Database and feature extraction

Our database was taken from the OGI multi-language phonetically transcribed corpus [7]. We used files from the English story part. The database included over 100 PCM wave files of different speakers, each about one minute long. We used the standard Mel-cepstrum speech feature extraction (fig 1a) [9]. For each file the speech signal was divided into frames of 20 msec long. Each frame was multiplied by a Hamming window, and Fourier transformed. The power spectrum coefficients were mel-scaled. The Log-Mel-scaled coefficients passed through iDCT and cepstrum mean subtraction. The resulting features are a set of 16 coefficients of cepstrum values and their temporal derivative between adjacent frames. Each set of coefficients is tagged by the label phoneme (Y_1) and speaker (Y_2).

As the mutual information for the full continuous cepstral space is difficult to estimate, we performed vector quantization for discretizng the space. We used a training set taken from the database at different stages of the feature extraction process to produce codebooks of different sizes, N, which henceforth serve as our compressed variable *X*. We then calculated the empirical mutual information between the tagged phoneme Y_1 and the discretized speech features *X*.

3. Results

Our first interesting observation is that the mutual information of *X* and Y_1 increases at each stage of the standard feature extraction procedure (as shown in figure 1b), despite the fact that each stage reduces entropy and thus discards information. This can be explained by the facts that first, the information discarded is less relevant for phoneme identification – as it should be – and second, after quantization we are left with more efficient representation. This suggests the code-book obtained from the final processing stage, G, as the best choice of data-space for our relevant compression procedure.

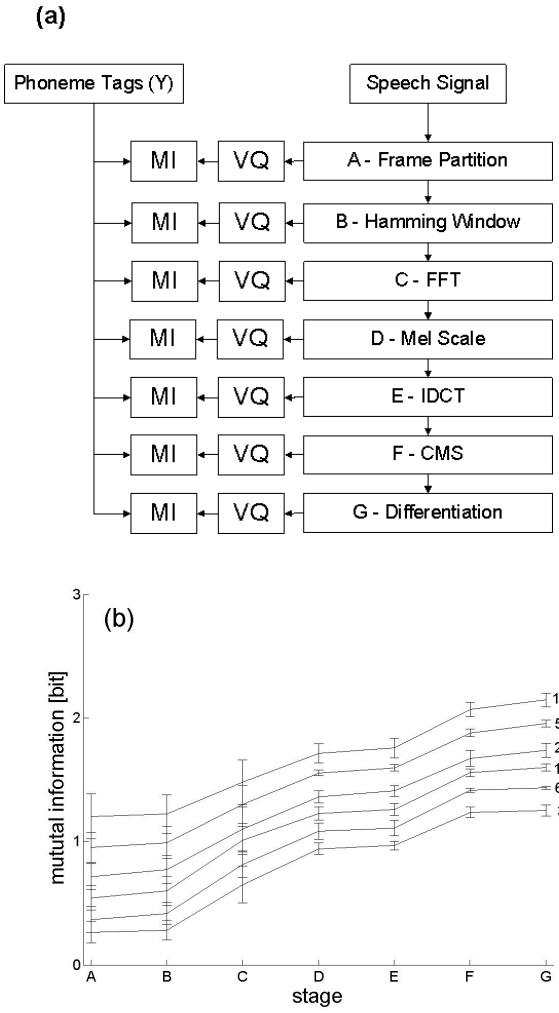

Figure 1: (a) The feature selection process. (b) The mutual information between the VQ codebook and the phonemes. Note the monotonic increase of the relevant information along the stages as well as with the VQ size. Error bars are for 7 cross-validation batches.

The application of the agglomerative information bottleneck procedure leads to the desired result. For a given number of clusters, the amount of mutual information between the clusters and the target variable Y is much higher when starting from a larger codebook and then reducing its size using the AIB algorithm (Fig. 2). As an example, the mutual information between a code book of size N=256 extracted directly from the vector quantization stage described in section 2.2 is lower than the mutual information between the target Y and a codebook of size 256 obtained by AIB applied to an initial data-space of 512 vectors.

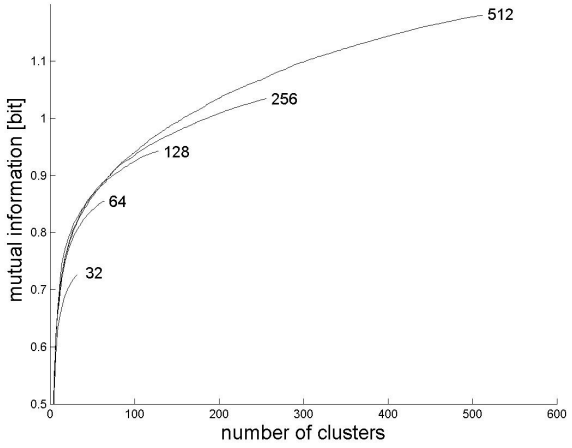

Figure 2 : Mutual information between the clusters $(\hat{X})$ and the phonemes (Y_1) at different stages of the AIB algorithm. Shown are results for N=32, 64, 128, 256, and 512.

An important advantage of the IB algorithm is flexibility in choosing the relevant target. When performing the IB algorithm for purposes of maximizing the mutual information with the tagged phoneme Y_1 we observe a large range in which the number of clusters decreases without a considerable decrease in the mutual information about Y_1 (Fig. 3). On the other hand, the mutual information with the speaker target (Y_2) drops rapidly (see fig 3). When applying the IB algorithm with the goal of maximizing the mutual information about the speaker's identity Y_2, we observe a larger range where $I(\hat{X};Y_2)$ barely drops, while $I(\hat{X};Y_1)$ decreases rapidly (Fig. 3). Interestingly, this is more noticeable when starting from a smaller data-space size N. The IB indeed captures the relevant information about the target, whether phonemes or speaker, as it is designed to do. The difference between the phoneme-cluster information and the speaker-cluster information is about 50% of the mutual information, which (roughly) predicts a similar improvement in recognition performance for each task (respectively). Moreover, the analysis of the resulting clusters can identify the speaker vs. phoneme dependent components of the Mel-cepstra, identifying them as clearly different components of the signal.

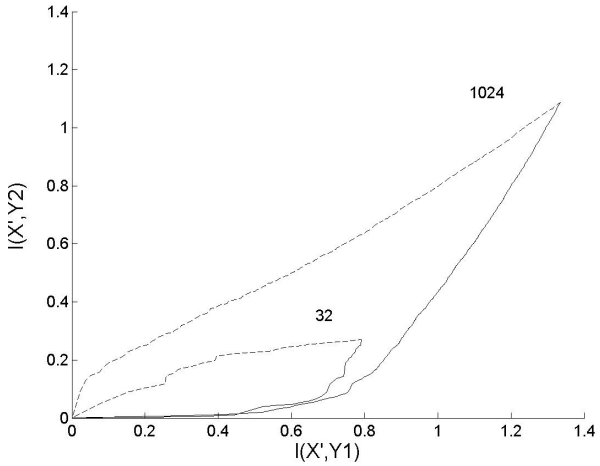

Figure 3: Mutual information between clusters and phonemes, $I(\hat{X};Y_1)$ vs. the mutual information between clusters and speakers $I(\hat{X};Y_2)$. Results shown are for initial data-spaces of sizes N=1024 and N=32. The results shown are for the AIB applied to the phonemes (dashed) and the AIB applied to the speakers (solid).

	Entropy	**VQ 1024 MI**	**MI after AIB to speakers**	**MI after AIB to Phonemes**
Phonemes	3.58	1.33	0.35	0.82
Speaker	4.0	1.08	0.35	0.15

Table 1: Entropies and Mutual Information (bits) for the phonemes and speakers for the initial size N=1024 VQ and after applying the AIB algorithms for phonemes and speakers with reduced N=32 clusters.

4. Discussion

Current speech recognition algorithms commonly use a universal set of features that are applied uniformly to classify a wide variety of attributes of the original speech signal. These attributes, or recognition targets, include the spoken phoneme, the speaker's identity, gender, language, prosody etc.. Such a uniform front-end approach does not utilize the clearly different characteristics of the signal that affect each of these attributes. It would be much more productive to use a different-task-specific-front-end approach that specifically extracts the relevant features of the speech signal. The information bottleneck method provides natural framework to tackle this challenge.

Here we present a novel approach for the selection of a reduced set of features for different speech recognition tasks using the information bottleneck method, by applying the simple agglomerative proxy algorithm (AIB). By starting with a standard vector quantization of mel-cepstrum features, we iteratively reduce the feature set size in a way that minimizes the loss of mutual information between the reduced feature set and the relevant target variable. We show that this method preserves the relevant information while considerably reducing the feature set dimensionality. The method produces a feature set that is much more appropriate for the task than a feature set of the same size obtained with a standard vector quantization of the mel-cepstrum. This method is designed to capture the relevant information of the target, and is not suited for the recognition of other targets.

The information bottleneck approach opens up many other directions to pursue in speech processing. The method is suited to the recognition of other speech characteristics, and to combined speech characteristics such as words. In addition the AIB can be used to analyze the correlation between frames.

5. References

[1] Tishby, N., Pereira, F., and Bialek W., ``The Information Bottleneck Method", *The 37th annual Allerton Conference on Communication, Control, and Computing,* 1999.

[2] Slonim, N and Tishby, N, "Document Clustering using Word Clusters via the Information Bottleneck Method", *proceedings of "SIGIR"* 2000.

[3] Seldin, Y. Starik, S. and Werman M., "Unsupervised Classification of Images using their Joint Segmentation." *In the proceedings of the 3rd International Workshop on Statistical and Computational Theories of Vision (*SCTV) 2003.

[4] Schneidman E, Slonim N, Tishby N, de Ruyter van Steveninck R, and Bialek W, "Analyzing neural codes using the information bottleneck method", (NIPS-13) 2002.

[5] Slonim, N and Tishby, N, "Agglomerative Information Bottleneck", *Advances in Neural Information Processing Systems* (NIPS-12) 2000.

[6] Lin, J., IEEE Trans. on IT,**37**: 145-150, 1991.

[7] Muthusamy, Y. K., Cole, R.A., Oshika, B.T. "The OGI Multi-Language Telephone Speech Corpus", *ICSLP,* 1992.

[8] Yang, H. H. and Hermansky, H., "Search for Information Bearing Components in Speech", *Advances in Neural Information Processing Systems* (NIPS-12) 2000.

[9] Shtein, Y. J., *Digital Signal Processing A Computer Science Perspective,* John Wiley & Sons, NY, 2000.