Using Mutual Information to design class-specific phone recognizers

Patricia Scanlon[1,2], Daniel P.W. Ellis[1], Richard Reilly[2]

[1]LabROSA, Dept. of Elec. Eng., Columbia Univ., New York USA
[2]DSP Group, Dept. of Elec. Eng., Univ. College Dublin, Ireland
{patricia,dpwe}@ee.columbia.edu, richard.reilly@ucd.ie

Abstract

Information concerning the identity of subword units such as phones cannot easily be pinpointed because it is broadly distributed in time and frequency. Continuing earlier work, we use Mutual Information as measure of the usefulness of individual time-frequency cells for various speech classification tasks, using the hand-annotations of the TIMIT database as our ground truth. Since different broad phonetic classes such as vowels and stops have such different temporal characteristics, we examine mutual information separately for each class, revealing structure that was not uncovered in earlier work; further structure is revealed by aligning the time-frequency displays of each phone at the center of their hand-marked segments, rather than averaging across all possible alignments within each segment. Based on these results, we evaluate a range of vowel classifiers over the TIMIT test set and show that selecting input features according to the mutual information criteria can provides a significant increase in classification accuracy.

1. Introduction

The fundamental task of the acoustic model in a speech recognizer is to estimate the correct subword or phonetic class label for each segment of the acoustic signal. This task is complicated by the fact that the information relevant to this classification is spread out in time – due to mechanical limits of vocal articulators, other coarticulation effects, and phonotactic constraints – and may be unevenly distributed in frequency. One response to this situation is to collect information from a very large temporal context window as input to the classifier (perhaps as much as one second [1]), but this implies a large number of parameters in the classifier, and hence requires very large training sets, as well as frustrating the classifier training with redundant and irrelevant information.

In this work we investigate using Mutual Information (MI) as a basis for selecting particular cells in time-frequency (TF) to optimize the choice of features used as inputs to a classifier. MI is defined as the as the measure of the amount of information one random variable contains about another. Specifically, we have investigated the MI between TF features and phonetic and speaker labels within the phonetically-labeled TIMIT continuous speech corpus.

In related work, Morris et al. [2] investigated the distribution of information across the on/off aligned auditory spectrogram for a corpus of vowel-plosive-vowel (VPV) utterances. The MI was estimated between one time frequency feature and the VPV labels, also the joint mutual information (JMI) between two time frequency features and the VPV labels. The goal was to use MI and JMI to determine the distribution of vowel and plosive information in the time frequency plane. Features with high MI and JMI were used to train a classifier to recognize plosives in the VPV utterances.

Bilmes [3] used the Expectation Maximization (EM) algorithm to compute the MI between pairs of features in the TF plane. He verified these results in overlay plots and speech recognition word error rates. Ellis and Bilmes [4] used the same techniques of MI estimation to predict how and when to combine entire feature streams.

Yang et al. [5] used methods similar to [2] to perform phone and speaker/channel classification. Simple classifiers with one or two inputs demonstrated the value of individual TF cells with high MI and pairs with high JMI.

The idea of using class-specific feature subsets for classification was investigated in the BMMs of [6]. There, MI between pairs of feature elements, conditioned on class label, was used to build unique distribution models for every state in an HMM. Here, we propose just one unique TF pattern per broad class based on MI between features and the class label, then use discriminative classifiers to distinguish within that broad class.

Our work extends [5] in several ways. In addition to calculating MI over all phones, we look at subsets composed of broad phonetic classes, such as the MI between specific vowel labels across only the vowel tokens. Because different phonetic classes can have very different temporal and spectral structure, calculating the MI across all labels can 'wash out' interesting details that are specific to a particular class.

A second way in which detail can be washed out is by using every possible temporal alignment within each phone segment. Instead, we have taken just one time-frequency 'snapshot' for each labeled phone segment within the training data, and calculated MI over a *registered* TF grid, such that the center of the phone (or the burst initiation in the case of stops) is aligned across all the examples. This is in contrast to previous work which uses N translated versions of the TF features from a phone segment of duration N frames, thereby blurring out any discernible temporal structure.

The next section describes how we calculated mutual information and presents MI as a function of time and frequency for a variety of data and label subsets. The hypothesis that high MI features provide good discrimination is verified in section 3 by performing vowel classification using the high MI points as input features to a multi-layer perceptron (MLP) classifier, and comparing these results to several baselines. In section 4, we discuss how the comparisons show a) the usefulness of only using the center frame in a vowel segment for training and b) the usefulness of high MI features for discrimination.

2. Mutual Information

Entropy of a discrete random variable X is defined as

$$H(X) = -\sum_{\forall x} p(x) \log(p(x)) \qquad (1)$$

where $0 \leq p(x) \leq 1$ implies $\log(p(x)) \leq 0$ and therefore $H(X) \geq 0$

Mutual Information is defined as a measure of the amount of information one random variable contains about another [7]. After observing the random variable Y, the uncertainty of X is reduced to $H(X|Y)$, which is the conditional entropy. Therefore the amount of information gained is the Mutual Information (MI):

$$I(X;Y) = H(X) - H(X|Y) \qquad (2)$$

In order to estimate The MI the probability density functions $p(x)$ and $p(x|y)$ need to be approximated. For each time-frequency feature the distribution of values was approximated with a diagonal-covariance Gaussian mixture model (GMM). Two Gaussian components were found to be adequate to capture the distributions we encountered. Once the GMMs has been determined, it is sampled on a dense grid over a fixed range. This range has its lower bound equal to the minimum of the mean minus three standard deviations across all the Gaussians; the maximum is similarly determined. This sampled data is used to compute the entropies $H(X(f,t))$ and $H(X(f,t)|Y)$, where $X(t,f)$ is the set of instances of the TF point $x(f,t)$ across the entire data set and Y is the set of vowel, phone or speaker class labels.

2.1. Mutual Information in Speech and Speaker Data

The MI was computed between labels and individual cells across the time-frequency plane. The base features were standard Bark-scaled log-spectral-energy coefficients calculated over a 25ms window with 10ms advance between adjacent frames. The TIMIT data, which is sampled at 16 kHz, gave 19 one-Bark frequency bands. A temporal window of $\pm$ 25 frames around the labeled target frame was used as the domain over which MI was computed. Within each utterance, the mean of each frequency band was set to zero, providing a kind of normalization against variations in channel characteristic (microphone positioning etc.).

Because our approach sought to calculate MI at a fixed temporal alignment of each phone segment, we initially had only one value for each TF cell from each segment in the training data. In order to increase the data available for training, and thus to improve the stability of our MI estimates, each TF value was augmented by its two immediate temporal neighbors, making the assumption that adjacent features in time are very similar and almost redundant. However, no further temporal shifting was performed in order to preserve fine time structure in the MI displays.

Each calculation thus returned an MI plot consisting of 19×49 cells. We performed these calculations against a variety of labels and over different subsets of the data. These results are shown in figure 1. The number of speakers from the TIMIT database used to calculate each pane was varied to ensure a sufficient number of examples were available.

The MI between TF features and all phone class labels is shown in the top pane. It can be seen that information useful for discriminating between phones exists in all frequency bands and that the information spreads out for about +/- 50ms. The highest MI can be seen at around 4-7 bark. Small differences between this plot and the corresponding figure in [5] are due to the fact that each window in our experiments is aligned on the center frame in a labeled segment, rather than using every frame carrying that label as a center-point.

The next four panes show the same calculation conducted over subsets of the data contributing to the top pane. Taking the
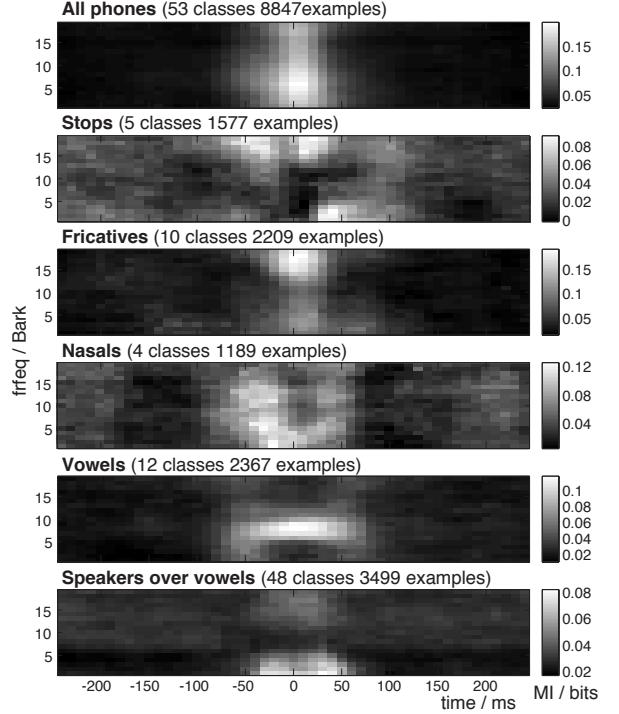

Figure 1: *Mutual Information across time-frequency for various classes and data subsets. The number of classes and the total number of instances are given for each case.*

four broad phonetic classes of stops, fricatives, nasals and vowels, MI is calculated between the specific phones within each class, calculated only over the segments that are marked as belonging to that broad class. Although the top pane is, loosely speaking, a weighted average of these panes, we can see very different and specific patterns arising from the individual characteristics of each broad class. For stops (which were aligned at the starting edge, rather than the center, of the marked segment), we see a dip in MI just before the central frame, indicating the absence of information about the *particular* stop class in the silence immediately preceding the stop burst. However, for around 50ms before and after this dip, there is significant information in the frequency bands at Bark 14 and above, presumably indicating cues from incoming formant transitions and from the spectral characteristics of the burst. We also see a peak in MI in the low frequency (around 3 Bark) at about 40ms after the burst - presumably indicating the most useful place to look to distinguish voiced (b,d,g) and unvoiced (p,t,k) stops.

The third pane showing fricatives shows a fairly smooth distribution with most information again above Bark 14 – where most fricative energy exists – although again with some information below Bark 6 to distinguish voiced and unvoiced fricatives. Temporal information shows the same smooth central peak as seen in the first pane. Nasals, on the other hand, again show a minimum of MI at the very center of the window (at least above Bark 6), with most information coming from the middle frequency bands (Bark 5-15) between 20 and 60ms before the segment center, and somewhat less available in the symmetric region after the center. These off-center foci presumably cover the formant transitions in and out of the nasals, which are known to be the strongest cues to nasal identity.

The MI between TF features and the vowel classes is shown in fifth pane. It can be seen that the information is spread over a wider span than for all phones (first pane), specifically about ± 80ms, or roughly the duration of a short vowel. Also, the most relevant information for vowel discrimination exists between 6 and 11 bark which is somewhat higher than for all phones. An interesting aspect visible here is that while the lowest and highest frequency bands are not informative during the main vowel duration, outside the central region we see faint 'arms' above and below the central blob. These could again be capturing formant transitions in and out of the vowel; in the lower frequencies, they indicate the best place to look to distinguish between long- and short-duration vowels. Like the other fine temporal structures mentioned above, this is the kind of detail that only becomes visible when spectral frames are aligned on each label segment, rather than being averaged across all possible offsets within a segment.

The bottom pane is also evaluated over all segments falling into the broad class of vowels, but now the MI is to the *speaker* label rather than the specific vowel label. This plot was motivated by the idea that vowels are the most useful segments in identifying speakers. Interestingly, we see an image that is almost exactly the complement of the lexical (vowel-class) information from the pane above: relevant information for speaker exists above 12 bark and below 5 bark. Information is spread over a smaller time span of about ± 60ms. The high MI in the lowest frequency bands may include factors arising from the speaker channel, although our normalization should have removed the strongest effects.

3. Classification

To verify the hypothesis that high MI points correspond to relevant features for discrimination, several vowel classification experiments were performed. A multi-layer perceptron (MLP) classifier was using for vowel classification. The neural network (NN) was trained using the QuickNet software from ICSI Berkeley The network had a single hidden layer of 100 units, and an output layer of 12 units, one for each of the 12 vowel classes. Only the input layer geometry was varied in the experiments, to accommodate different numbers of input features.

The network was trained to estimate the posterior probability for each of the 12 vowel classes for each each frame of input by backpropagation of a minimum-cross-entropy error criterion against 'one-hot' targets. The NN was trained using all 468 speakers from the 8 dialects of the TIMIT database, a total of 4680 utterances of which 160 utterances were used for cross-validation. The 168 test speakers were used for testing.

Once the 19 spectral coefficients are extracted for every frame in the training set, the mean and standard deviation are computed of all the features for normalization. Each feature dimension in the training set is separately scaled and shifted to have zero mean and unit variance. The same normalization is applied to the test set.

The MI indicates which TF features contain the most information for discriminating the vowel classes. Therefore, in order to show that these high MI features provide good discrimination, three vowel classification experiments were carried out, using three different methods to choose which TF cells are used as input to the classifier. First, a baseline system used a rectangular block (RECT) of all 19 frequency bands across a temporal window of n successive 10ms frames, where n is 7,5,3,1 .i.e. $\pm 3,2,1,0$ frames. One training pattern block was extracted from each vowel segment in the training data, aligned with the
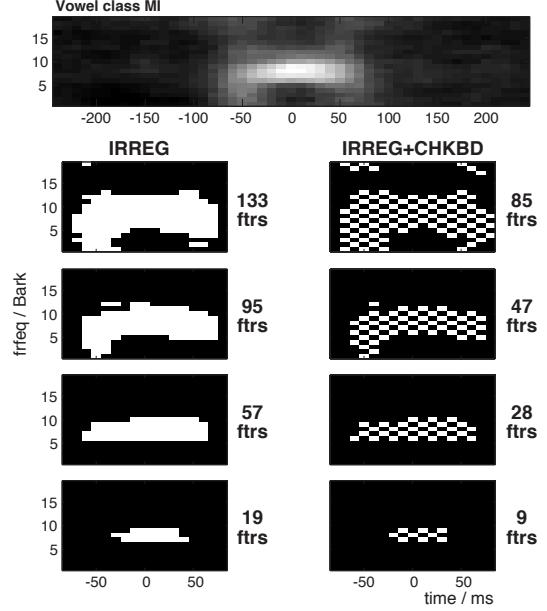

Figure 2: *Mutual Information thresholded masks (IRREG) for features lengths 133,95,57,19 without checkerboarding (left) and features lengths 85,47,28,9 with checkerboarding (CHKBD, right).*

center frame of the segment. (In the case of an odd number of frames in a vowel sequence, two windows are used for training, each centered on one of the center frames.) The features in the temporal window are concatenated together after normalization to create one feature vector that is presented to the input layer of the NN.

To take advantage of the MI results, the second system took as input the TF cells with the largest MI values to the vowel label. This results in an irregularly-shaped pattern in the TF plane (IRREG) consisting of all the cells from the 5th pane of figure 1 with values above some threshold. The threshold was varied to match the total number of inputs in the baseline systems with different temporal context widths i.e. 19, 57, 95 and 133 features. The mask was used to select TF values from each training segment, and all selected values were concatenated and presented as input to the MLP classifier. The mask always lay within a block of 19x17 cells i.e. 19 frequency channels by ± 8 frames around the center. The thresholded masks used in the experiments to extract the various number of features for this method, are shown in the left-hand column of figure 2.

Note that in the MI investigation above, MI was computed for each TF cell in isolation. While the relative MI can be seen in the MI plots, the JMI between each MI point and all its neighbors is not so easily obtained. We can however assume that every spectral dimension is highly correlated with its immediate neighbors, and correlation is high along the time axis too: This suggests that the immediate neighbors of a cell used in classification could be omitted from the classifier without a significant loss of information. This can be implemented by multiplying the selection masks of both the RECT and IRREG systems by a checkerboard pattern (CHKBD), so no cells in the resulting mask can be immediate neighbors. The right column of figure 2 illustrates the checkerboard MI masks used for the IRREG case. Note that applying the checkerboard changes the number
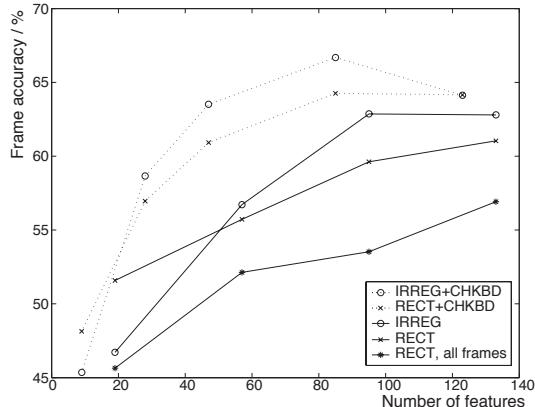
Figure 3: *Variation of classification accuracy with the number of TF cells used in the classifier input across all three systems and with and without the checkerboard feature masking.*

of input units available.

Finally, the effectiveness of using only one temporal alignment per labeled segment was tested with a contrast system that used the same input feature geometries as the RECT system, but was trained on TF windows centered on *every* frame labeled with a particular class ("all frames") rather than only the single alignment at the center of the segment. This classifier thus saw a much larger number of training patterns, but the patterns were less consistently structured since they were composed of all possible temporal offsets for each class. This is also the closest equivalent to the classifier used in conventional connectionist speech recognition, although it is discriminating only among vowels, not among all phone classes.

Every system was tested in the same way: For every vowel in the test data (a total of 14641 cases), a single test pattern aligned to the center of the vowel segment is collected, and subjected to the appropriate mask. These patterns are passed to the MLP classifier, and the output unit with the greatest activation is taken as the class. The 'frame accuracy' value in figure 3 is the proportion of such patterns correctly classified. (A difference of around 1% between any of these results is significant at the 5% level.)

4. Discussion and Conclusions

Figure 3 shows that for a given number of input features, the MI-based IRREG classifier (circles) outperforms the more conventional RECT classifier (crosses) for all except the very smallest input sizes (at which point IRREG is looking at just 3 of the 19 frequency channels). We also see that checkerboarding the input mask to reduce input dimensionality (dotted lines) significantly improves accuracy in all cases except the very narrowest, even though the number of inputs is approximately halved. Our best performance of 66.7% correct is achieved by checkerboarding the large IRREG MI region to include 85 features.

We can also see that, for this task, training only on the center alignment of each segment affords very considerable improvement over the conventional approach of training on every frame labeled with a given class (asterisks). Of course, this is an unfair comparison, because the "all frames" classifier will presumably do a much better job of classifying frames closer to the edge of segments, i.e. patterns on which the other classifiers

are not trained.

This points out the main weakness with these classifiers, that they rely on the hand-marked phone boundaries to locate the segment centers at which the classification is performed. In a real speech recognition task, no such boundaries would be available, and some additional scheme would be needed to decide which frames should be passed to the classifier. However, the bewilderment at misaligned frames might be helpful here: we plan to investigate using the entropy of the classifier output as a way of detecting the phone centers in continuous speech.

It is worth noting that TIMIT is a small corpus by today's standards. This is useful to the point we are trying to make, since the task is strongly data-limited. It is imperative to minimize the number of input features being employed, since there is not enough data to learn large numbers of parameters. This is apparent in the local maxima in figure 3, indicating the point at which adding further features (and hence model parameters) actually hurts performance.

We have shown that MI is a practical and interesting tool for locating the information in the speech signal. MI confirms and quantifies the way in which different phone classes rely on different timescales and frequency regions for discrimination, suggesting a scheme of phone-class-specific classifiers each looking at different regions in time-frequency. An investigation into a vowel classifier confirmed the value of MI-based feature selection, as well showing that training only on one, center-aligned pattern per training segment gave much greater accuracy for an equivalent test than the conventional approach of training on every frame offset with the segment.

5. Acknowledgments

6. References

[1] H. Hermansky and S. Sharma, "Temporal patterns (TRAPS) in ASR of noisy speech," in *ICASSP'99*, March 1999.

[2] A. Morris, J.L. Schwartz, and P. Escudier, "An information theoretical investigation into the distribution of phonetic information across the auditory spectrogram," *Computer Speech and Language*, vol. 7, no. 2, pp. 121–136, 1993.

[3] J. Bilmes, "Maximum mutual information based reduction strategies for cross-correlation based joint distribution modelling," in *ICASSP'98*, April 1998, pp. 469–472.

[4] D. Ellis and J. Bilmes, "Using mutual information to design feature combinations," in *Int. Conf. on Spoken Language Processing*, 2000, pp. 79–82.

[5] H. H. Yang, S. Sharma, S. van Vuuren, and H. Hermansky, "Relevance of time-frequency features for phonetic and speaker-channel classification," *Speech Communication*, Aug. 2000.

[6] J. Bilmes, "Buried markov models for speech recognition," in *ICASSP'99*, March 1999.

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