ELSEVIER

Speech Communication 31 (2000) 35–50

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Relevance of time–frequency features for phonetic and speaker-channel classification

Howard Hua Yang *, Sarel Van Vuuren, Sangita Sharma, Hynek Hermansky

Department of Electrical and Computer Engineering, Oregon Graduate Institute of Science and Technology, 20000 NW Walker Road, Beaverton, OR 97006-8921, USA

Received 30 October 1998; received in revised form 30 August 1999; accepted 2 December 1999

Abstract

The mutual information concept is used to study the distribution of speech information in frequency and in time. The main focus is on the information that is relevant for phonetic classification. A large database of hand-labeled fluent speech is used to (a) compute the mutual information (MI) between a phonetic classification variable and one spectral feature variable in the time–frequency plane, and (b) compute the joint mutual information (JMI) between the phonetic classification variable and two feature variables in the time–frequency plane. The MI and the JMI of the feature variables are used as relevance measures to select inputs for phonetic classifiers. Multi-layer perceptron (MLP) classifiers with one or two inputs are trained to recognize phonemes to examine the effectiveness of the input selection method based on the MI and the JMI. To analyze the non-linguistic sources of variability, we use speaker-channel labels to represent different speakers and different telephone channels and estimate the MI between the speaker-channel variable and one or two feature variables. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Mutual information; Sources of variability; Spectral feature; Input selection; Phonetic classification; Multi-layer perceptron

1. Introduction

The speech research community has at its disposal rather large speech databases which are mainly used for training and testing of automatic speech recognition (ASR) systems. There has been little effort to date to use such databases for deriving reusable knowledge about speech and speech communication processes which could be used to improve ASR technology. In this paper we describe approaches for studying a large hand-labeled database of fluent speech using mutual information, an information-theoretic concept, to learn about the structure of the speech signal.

It has been known since the 1950s that the information about phonemes is extended in time and that there are no hard boundaries between phonemes. Over the past five years we have been advocating the use of ASR systems which selectively use relatively large temporal segments of speech signals. Our motivation was that the information about a phoneme is not localized to the region of that phoneme only, but instead is spread over a substantial (about one syllable long) segment of the signal (Hermansky, 1998). Knowledge of the spread of speech information is important since the basic task of a speech recognizer is to recognize

* Corresponding author. Tel.: +1-503-690-1331; fax: +1-503-690-1548.

E-mail address: hyang@cse.ogi.deu (H.H. Yang).

0167-6393/00/$ - see front matter © 2000 Elsevier Science B.V. All rights reserved.
PII: S0167-6393(00)00007-8

phonemes or phoneme-like units from information in the acoustic speech signal.

Mutual information, which measures the dependence between random variables, can be used to measure the spread of information. For speech data, we use this concept to deal with the question of how various elements of an information stream relate to each other. Some of our results were reported in (Yang et al., 1999). Previous works in this direction include (Morris et al., 1993; Bilmes, 1998).

Morris et al. (1993) analyzed 3000 vowel-plosive-vowel (VPV) utterances and estimated the mutual information (MI) between one acoustic feature and the aligned VPV labels and the joint mutual information (JMI) between two acoustic features and the VPV labels. Their goal was to use the MI and the JMI results to characterize the distribution of vowel and plosive information in VPV utterances in the time–frequency plane. They selected features with high MI and JMI for a Gaussian classifier and a multi-layer perceptron (MLP) to recognize plosives in the VPV utterances. To avoid the risk of assuming a wrong distribution for the feature variable, they used histograms to estimate the MI and the JMI. Although we use the same approach to estimate the MI and the JMI, our work is different from Morris et al.'s work in several ways. We consider the task of *phonetic* classification of telephone speech. We evaluate features in different contexts in the time–frequency plane. Our data set is larger to obtain more reliable MI and JMI estimates. Instead of showing the MI and JMI in grey levels in the time–frequency plane, we plot the levels of the MI and the JMI so that it is easier to visualize the maximum MI and JMI. We explicitly are trying to learn the structure of the information, hence we care to present the results in a more meaningful manner.

Bilmes (1998) showed recently that the salient information of speech appears to be spread over relatively long temporal spans. He estimated the MI between the spectral energy features and used it to optimize recognition models. We are interested in how the phonetic information is distributed in time and frequency. We have a data set in which every frame is hand-labeled by a phonetic label. We use this data set to estimate the MI and the JMI for one and two feature variables of energy observations in the time–frequency plane. This allows us to probe the distribution of phonetic information in time and frequency.

We represent the information in time and frequency by the short-term critical-band logarithmic energy $X(f_k,t)$. This is a feature representation commonly used in phonetic classification. A frame at time t is associated with a phoneme label Y_t. The problem is to determine the relevance of $X(f_k,t+d)$ across all frequencies f_k and a context window $-D \leqslant d \leqslant D$ for the classification of a phoneme at time t. We study the MI and the JMI of features for phonetic classification and for the classification of speaker-channels. Our motivation for using the MI and the JMI to analyze speech data is two-fold: (1) to study the distributions of phonetic information or speaker-channel information in the feature space for different target variables (phonetic label or speaker-channel label); (2) to select features for classifiers for different tasks: phoneme classification or speaker-channel classification.

It should be noted that feature selection based on the MI and the JMI is independent of the classification models. Our experimental results show that MLPs using the high MI or high JMI features as their inputs give high accuracies for phoneme classification.

2. Data set

Results are based on about 3 hours of phonetically labeled telephone speech from the English portion (Stories) of the OGI multi-lingual database (Cole et al., 1994). This represents approximately 50 s of extemporaneous speech from each of 210 different speakers. In this database, the average phoneme duration is about 65 ms and the average number of instances of a phoneme is 3440.

For the experiments in the paper we select a *subset* of the full phoneme set which consists of 19 phonemes that commonly occur in connected digits. These phonemes are denoted using the variable Y and labeled 1–19 (refer to Table 1). The speech segments corresponding to other phonemes and garbage sounds are removed. Since the 19

Table 1
A table of labels for 19 phonemes

Label	Phoneme	Example
1	w	[w]eed
2	∧	sev[e]n
3	n	[n]ine
4	uc	(unvoiced closure)
5	th	[t]wo
6	u	tw[o]
7	T	[th]ree
8	9r	fou[r]
9	i:	thr[ee]
10	f	[f]ive
11	aI	n[i]ne
12	v	se[v]en
13	s	[s]even
14	I	s[i]x
15	kh	[c]at
16	E	s[e]ven
17	ei	[ei]ght
18	z	[z]ero
19	oU	zer[o]

phonemes cover the major types of phoneme categories, it is believed that the results and conclusions obtained based on this smaller phoneme set may be generalized to the full phoneme set.

Acoustic features $X(f_k, t)$ are computed for 15 critical bands as follows. First, power spectra are computed from a short-time Fourier transform (STFT) analysis of the speech signal with a 20 ms Hamming window advanced in $\Delta t = 10$ ms steps, i.e., the frame rate is 100 Hz. Each step corresponds to a time frame. Then, logarithmic energies are computed by applying critical-band spaced (log-like in the frequency variable) weighting functions (Fig. 1) to the power spectra. This is done in a manner similar to that of the computation of Perceptual Linear Prediction Coefficients (Hermansky, 1990) by multiplying the power spectra by each weighting function in turn and integrating the result. The center frequencies of the critical bands may be expressed on a Hertz or Bark scale. The mapping from k in Bark to f in Hertz is given by the formula

$$f = 600 * \sinh(\mathrm{k}/6)$$

or by Table 2.

The total number of frames is about 500 000. The 50-speaker and 100-speaker subsets are used to obtain the MI and the JMI results in this paper.

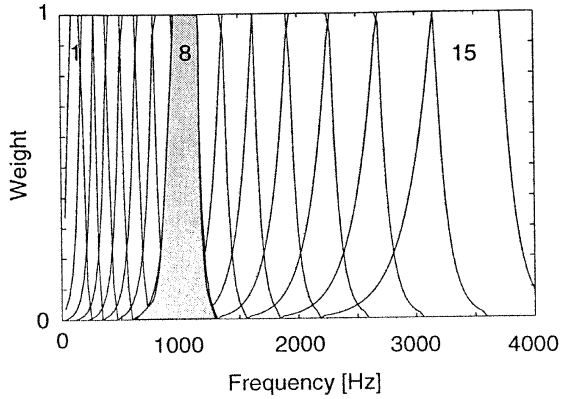


Fig. 1. The 15 weighting functions used for computing 15 samples from the power spectra of the speech signal. The weighting function for the eighth band is shown shaded to show the shape of the function.

Table 2
Center frequencies in Bark and Hertz

Center frequency in Bark	Center frequency in Hertz
1	100
2	204
3	313
4	430
5	560
6	705
7	870
8	1059
9	1278
10	1532
11	1828
12	2176
13	2584
14	3065
15	3630

The 50-speaker subset is a subset of the 100-speaker one. Unless specified, the MI and the JMI results are based on the 50-speaker data set. Half of the data set is not used at all in estimating the MI and the JMI. However, the whole data set is used in the phonetic classification experiments to verify the MI and JMI results.

3. MI and its properties

The MI between two random variables X and Y is defined by the entropies $H(X)$, $H(Y)$ and $H(X, Y)$:

$$I(X;Y) = H(X) + H(Y) - H(X,Y). \quad (1)$$

If X and Y have a joint density function $p(x,y)$, the MI is equal to the Kullback–Leibler divergence between $p(x,y)$ and $p(x)p(y)$,

$$I(X;Y) = \int p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \mathrm{d}x\,\mathrm{d}y. \quad (2)$$

The MI measures the statistical dependence between two random variables. It is zero when the two random variables are independent.

When X and Y are assumed to be jointly Gaussian, the MI can be computed analytically by the following formula:

$$I(X;Y) = -\log(1-\rho^2),$$

where ρ is the correlation coefficient between X and Y. However, the Gaussian assumption is usually not true for speech data. We shall give a non-Gaussianity test later in Section 4.1 which shows that energy observations are strongly non-Gaussian.

To estimate the MI, one needs to approximate the probability density function $p(x,y)$. Some typical density approximation methods are based on histogram, kernel function, or EM algorithm (Bonnlander, 1996; Bilmes, 1998; Yang and Moody, 1999, 2000). Like the approach in (Morris et al., 1993), we use the histogram method to approximate the density functions used for estimating the MI and the JMI. This is equivalent to computing the quantized version of the MI (Cover and Thomas, 1991).

The JMI $I(X_1,\ldots,X_n;Y)$ is defined by

$$I(X_1,\ldots,X_n;Y) = H(X_1,\ldots,X_n) + H(Y) - H(X_1,\ldots,X_n,Y). \quad (3)$$

To estimate the JMI, we need to estimate the joint probability $p(x_1,\ldots,x_n,y)$. This may suffer from the curse of dimensionality depending on the feature dimension n and the data size.

The JMI, $I(X_1,\ldots,X_n;Y) \geqslant 0$, measures the dependence between $(X_1,\ldots,X_n)$ and Y. It can also be written as

$$I(X_1,\ldots,X_n;Y) = H(X_1,\ldots,X_n) - H(X_1,\ldots,X_n|Y) \quad (4)$$

or

$$I(X_1,\ldots,X_n;Y) = H(Y) - H(Y|X_1,\ldots,X_n). \quad (5)$$

The entropy $H(X_1,\ldots,X_n)$ represents the uncertainty about the random vector $(X_1,\ldots,X_n)$. Conditional on Y, the uncertainty about this random vector is decreased. The JMI, $I(X_1,\ldots,X_n;Y)$, is the reduction in uncertainty about $(X_1,\ldots,X_n)$ when Y is observed. For example, if $(X_1,\ldots,X_n)$ is a feature vector and Y is a target variable representing phonemes or speaker-channels, then Eq. (5) gives a more appealing interpretation. $I(X_1,\ldots,X_n;Y)$ is the reduction in uncertainty about the target variable Y when the features or inputs $(X_1,\ldots,X_n)$ are used to predict Y.

For the phonetic classification problem, we encounter a classification variable that assigns a phonetic label to each frame. This variable is called a target variable for phonetic classification. The dependence between the target variable and the feature variables can be probed by the MI but not fully by a correlation coefficient since the energy observations are not Gaussian (See Section 4.1).

In practice, the correlation coefficient is often used to probe dependences between variables. But, it is only optimal for Gaussian variables. Also correlation is useless when one of the variables (phoneme category) is not even numerical, let alone not Gaussian. The correlation coefficient and MI have two major differences. First, the correlation coefficient measures linear dependences between random variables, whereas the MI measures the nonlinear statistical dependences between random variables. Second, the correlation coefficient is only invariant to component-wise linear transforms while the MI is invariant to component-wise monotonic transforms which may be nonlinear, i.e., $I(f(X),g(Y)) = I(X,Y)$ if the two functions $f(x)$ and $g(x)$ are monotonic and differentiable. The proof for the invariance property is given in Appendix A.

The MI has been used previously for feature selection in (Bonnlander, 1996; Barrows and Sciortino, 1996; Battiti, 1994). This method only selects individually optimal features. The JMI has been used in (Yang and Moody, 1999) to select jointly optimal features for radar pulse classification.

In this paper, we shall apply both the MI and the JMI to identify the features most relevant for phonetic classification and speaker-channel classification. For the features within one frequency band at different time shifts, we shall apply the JMI to measure the relevance of the features before and after the current time frame.

4. MI for phonetic classification

To study the dependence structure in the speech data we need the MI instead of the correlation coefficient, because as we show next, the energy observations are strongly non-Gaussian even after taking the logarithm.

4.1. Non-Gaussianity of speech data

We use the following statistics to test whether the distribution of the data is Gaussian:

$$\text{normalized skewness } S = \frac{1}{\sqrt{6T}s^3}\sum_{t=1}^{T}(x_t-\bar{x})^3$$

and

normalized kurtosis

$$K = \frac{1}{\sqrt{24T}s^4}\sum_{t=1}^{T}(x_t-\bar{x})^4 - \sqrt{\frac{3T}{8}},$$

where $\bar{x}$ and s^2 are the sample mean and sample variance of x_t. S and K are used to test the skewness and kurtosis of the data set. Under the null hypothesis that the distribution of the observation is Gaussian, asymptotically, both the normalized skewness S and the normalized kurtosis K follow the standard Gaussian distribution $N(0,1)$ (see Vol. 1 in (Stuart and Ord, 1994)).

For the trajectory at k Bark, the normalized skewness and the normalized kurtosis are denoted by $S(f_k)$ and $K(f_k)$, respectively. The results in Fig. 2 are based on a 50-speaker subset of the speech data. At the significance level $\alpha = 0.01$, the critical values for the standard normal distribution are ± 2.58. It is shown in Fig. 2 that the logarithmic spectral energy observations are strongly non-Gaussian. All absolute values of the normalized

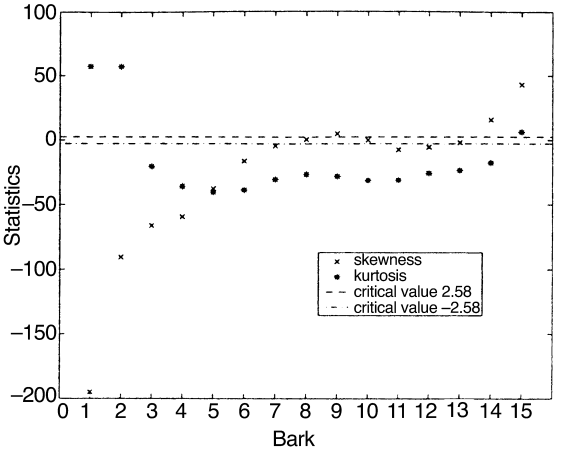


Fig. 2. The statistics S and K in different bands.

kurtosis statistics exceed the critical value, while 12 out of 15 absolute values of the normalized skewness statistics exceed the critical value. The normalized kurtosis statistics show that most of the bands are platykurtic or sub-Gaussian with negative kurtosis except the two lowest frequency bands and the highest frequency band which are leptokurtic or super-Gaussian. From Fig. 2 it is seen that the first, second and 15th band have rather different statistics. It is worthwhile to note that these bands fall mostly outside the telephone bandwidth (300–3400 Hz) and may be noisy or less reliable for phonetic classification.

The reason of testing non-Gaussianity is two-fold: (1) to justify why we use MI rather than correlation; (2) to determine the number of bins for a histogram to approximate a density function.

Since the logarithmic spectral observations are non-Gaussian, instead of using the correlation coefficient, we use MI to probe the dependence between features and phonetic labels. Using histograms to approximate a probability density function, one needs to choose the number of the bins to separate data points. There are several rules to choose the number of bins. Given a data set $\{x_t,\ t = 1,\ldots,T\}$, for Gaussian distributions, one may choose $\log_2 T + 1$ as the number of bins by Sturges's rule; but, for non-Gaussian distributions, one is better off to choose $\log_2 T + 1 + \log_2(1 + \hat{\kappa}\sqrt{T/6})$ as the number of bins by Doane's rule, where $\hat{\kappa}$ is the estimated kurtosis of x_t

(see (Venables and Ripley, 1994) for the two rules). We use Doane's rule to choose the number of bins based on our finding that the data is strongly non-Gaussian. For our data sets, the number of bins is between 20 and 30 depending on whether we use the whole data set or a subset to compute the number of bins.

4.2. MI between one feature and phonetic label

As a first step we obtain the MI between phonetic labels and a *single* feature in the time–frequency plane.

4.2.1. Distribution of information in frequency

While the features can be anywhere in the time–frequency plane, we first evaluate the MI for the features in the same frame as the phonetic label. Thus, the MI result will represent the distribution of information in the spectral feature vector which is aligned with the phonetic label.

Our speech data set is denoted by

$$D_T = \{(X(f_k,t), Y_t) : k = 1,\ldots,15,\ t = 1,\ldots,T\}.$$

Each feature vector $(X(f_1,t),\ldots,X(f_{15},t))$ is assigned a phonetic label Y_t that is used as a target variable for phonetic classification. It is also called a phonetic label variable. Based on the data set D_T, for each k, we estimate the MI between a feature variable $X(f_k,t)$ for the current time frame and the target variable Y_t,

$$I(X(f_k,t); Y_t), \quad k = 1,\ldots,15.$$

This MI function indicates the degree of relevance of each frequency band for phonetic classification. Fig. 3 shows two plots of MI as a function of frequency band, one plot for a data set with 50 speakers and one for a data set with 100 speakers.

It is revealed in Fig. 3 that along the frequency axis, all frequency bands carry information about the underlying phoneme label, with the dominant information around 5 Bark or 560 Hz. In the rest of this paper, the MI and JMI are estimated based on the 50-speaker data set.

4.2.2. Distribution of information in time

In the speech data, adjacent frames usually contain similar information about the same phoneme. The same concept as used above for finding the distribution of information along the frequency axis is used to examine features at the same frequency but misaligned in time. The result would indicate how much the information about the phoneme is spread out in time.

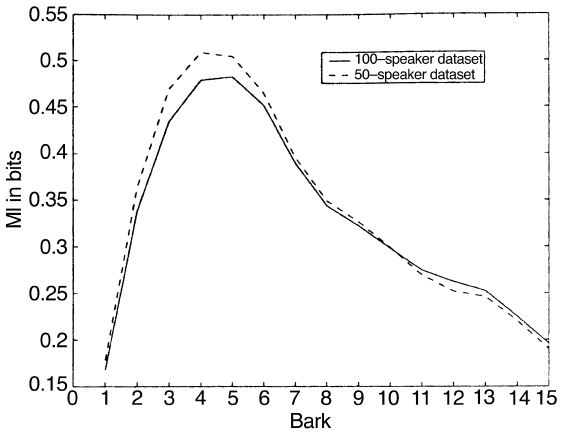


Fig. 3. MI between the phonetic label variable and one feature point in each frequency band for the 100-speaker and 50-speaker data sets. The two speaker sets exhibit similar patterns across different bands.

Define a set of frequency indices

$$Z_f = \{1,2,\ldots,15\},$$

and a set of time shifts

$$S(L,\Delta) = \{l\Delta : l = -L,\ldots,L\}.$$

The MI between the phonetic label variable and the features in different bands at different time shifts is

$$I(X(f_k,t+d); Y_t), \quad k \in Z_f,\ d \in S(L,\Delta).$$

The MI is shown in Fig. 4 as a function of k and d, for $L = 20$ frames and $\Delta = 10$ ms.

As a function of a time shift d, the MI function $I(X(f_k,t+d); Y_t)$ indicates the relevance of the features in the frames before and after the current frame at t. For $f_k = f_5$, this MI function is shown in Fig. 5.

To explain Fig. 5, let us consider two points $(X(f_k,t), Y_t)$ and $(X(f_k,t+d), Y_t)$. On average $X(f_k,t+d)$ contains little information on Y_t when the absolute time shift is greater than 100 ms.

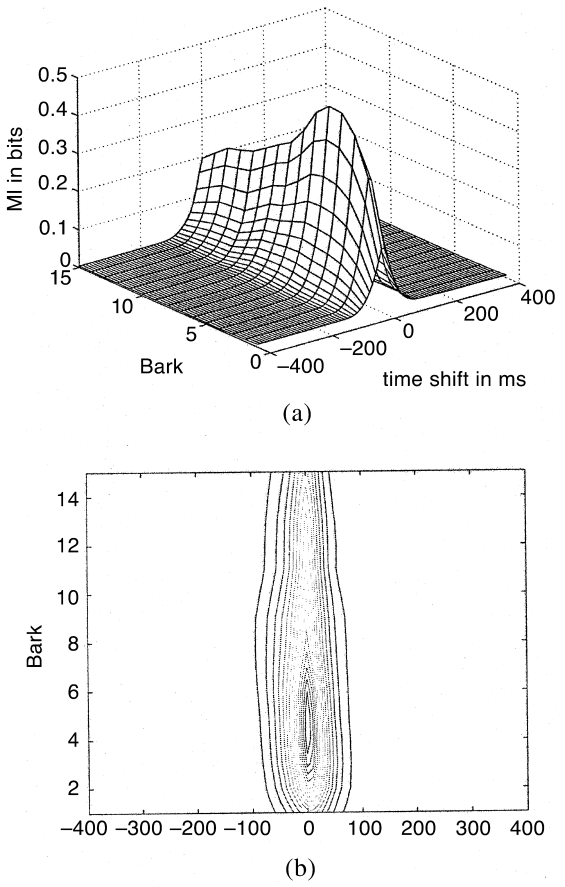


Fig. 4. (a) 3D plot of the MI, $I(X(f_k, t+d); Y_t)$, between the phonetic label variable and the features in different frequency bands with different time shifts based on the 100-speaker data set. (b) Contour plot of the MI function.

Conversely, $X(f_k, t+d)$ does contain information on Y_t for absolute time shifts less than 100 ms. This suggests that one may want to use contextual information in a window of about 100 ms to either side of the frame being classified.

Along the time axis, features further than 100 ms from the labeled frame are basically irrelevant for the classification of that label. But, this is only true when the feature $X(f_k, t+d)$ is used *alone* to predict the phoneme label Y_t. We shall show in Section 4.4.2 that the feature $X(f_k, t+d)$ with an absolute time shift d larger than 100 ms is still relevant for predicting Y_t when the feature $X(f_k, t+d)$ is used jointly with the feature $X(f_k, t)$.

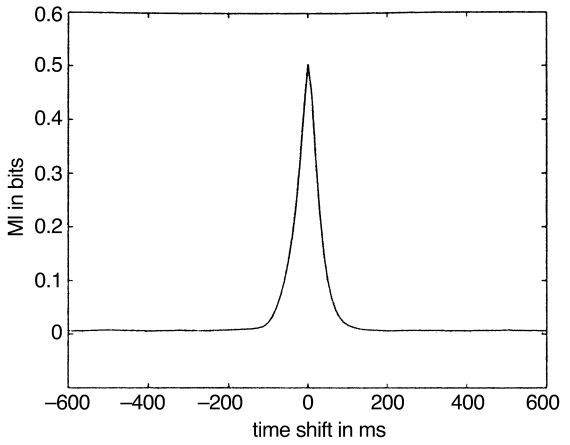


Fig. 5. Mutual information as a function of time shift (in frames) and the classification target variable for the fifth frequency band. The dotted line shows the lower bound (0.0028 bits) which is the MI between the fifth frequency band and scrambled phonetic labels.

4.3. Informationless MI levels

When a feature variable is not relevant to the target variable, the MI estimate is small. We use two methods to obtain informationless MI levels: (1) scramble the observations for the target variable; (2) use a random input to replace the observations for the feature variable. For the first method, the observed labels Y_t are treated as an array and indices of the array are randomly permuted to obtain the scrambled version of Y_t.

For the 50-speaker data set, in all bands, the MI between the feature and the scrambled phonetic labels is less than 0.003 bits. An informationless MI level of 0.0028 bits for the fifth band is shown by the dotted line in Fig. 5. For the 100-speaker data set, in all bands, the MI between the feature and the scrambled phonetic labels is less than 0.001 bits. Roughly speaking, the informationless MI level becomes smaller by half when the data set is doubled.

Replacing energy observations by a Gaussian random sequence of the same size, we obtain the informationless MI level of 0.0017 bits. Hereafter, any MI or JMI estimates which are less than 0.0028 bits will be considered as insignificant.

4.4. JMI between two features and the phonetic label

The MI between phonetic labels and a single feature gives a general indication which shows how linguistic information is distributed in the time–frequency plane. However, phonetic classification is seldom done using only a single feature. Rather, a whole vector of features is typically used to estimate the phonetic identity of the underlying linguistic event. How is the linguistic information distributed in a *combination* of features in the time–frequency plane? To address this problem, we use the concept of the JMI between the phonetic label and *two* features in the time–frequency plane which tells us how much more additional information (not contained in the first feature) is provided by the second feature. [1]

4.4.1. Additional feature at different frequency but the same time

Consider the relevance of two features in the same frame but in different frequency bands for phonetic classification. The JMI between these two features and the classification variable is

$$J_1(k,j) = I(X(f_k,t), X(f_j,t); Y_t),$$

where the two features are aligned with the target variable in time. Here, we define $J_1(k,k) = I(X(f_k,t); Y_t)$.

For the fifth band as the first feature, for example, the JMI as a function of frequency in Bark of the second feature band is shown in Fig. 6.

By the chain rule for MI (see Cover and Thomas, 1991), we have

$$J_1(5,k) = I(X(f_5,t); Y_t) + I(X(f_k,t); Y_t | X(f_5,t)).$$

The quantity $I(X(f_k,t); Y_t|X(f_5,t))$ is the MI between $X(f_k,t)$ and Y_t conditional on $X(f_5,t)$ and is called the *conditional* MI. It is the information gain due to the second feature $X(f_k,t)$ given the first feature $X(f_5,t)$. It is shown in Fig. 6 that the conditional MI reaches a maximum at 9 Bark (from about 0.5 bits for a single feature to as much as 0.85 bits for an additional feature at 9 Bark). In general, we find that the inclusion of a second feature always provides information in addition to that provided by a single feature.

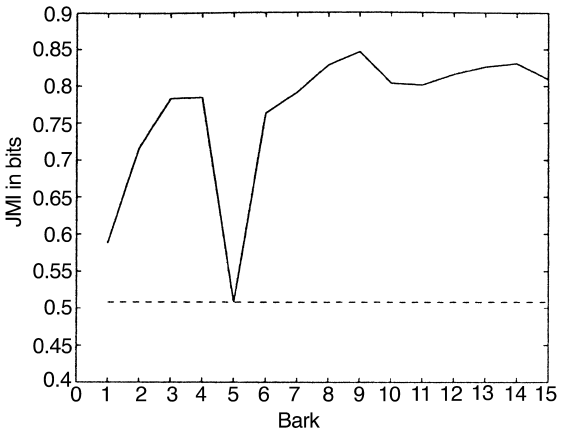


Fig. 6. The MI of one feature at 5 Bark, the horizontal line. The JMI of two features, one at 5 Bark and another one at k Bark, the solid line.

The JMI between the phonetic label variable and two features in the same frame but in different bands are shown in Fig. 7. The kth row in the matrix $[J_1(k,j)]$ is plotted in the kth panel in Fig. 7. Note that Fig. 6 is an enlargement of the fifth panel in Fig. 7.

For the two optimal features in the same frame but different frequency bands, we compute the maximum

$$J_1(4,9) = \max_{k,j \in Z_f} J_1(k,j) = 0.9 \text{ (bits)}. \tag{6}$$

This means that among all possible pairs $(X(f_k,t), X(f_j,t))$ in the same frame, the pair $(X(f_4,t), X(f_9,t))$ is the most relevant for phonetic classification.

Let $J_1(k,j_k^*) = \max_{j \in Z_f} J_1(k,j)$ be the maximum JMI in the kth band. The maximum JMI and the MI are compared in Fig. 8. Given the 1st feature at k Bark, the information gain is maximized at j_k^* Bark. The gap between the two curves in Fig. 8 represents the maximum information gain due to a

[1] The concept of JMI can be extended to more than two features. However, the amount of data needed for reliable estimates increases exponentially as the number of features increases.

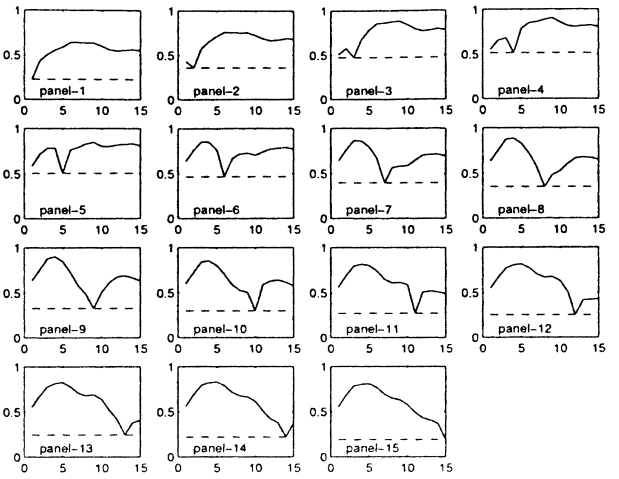


Fig. 7. For $k = 1, \ldots, 15$, the JMI, $\{J_1(k, j), j \in Z_f\}$, is plotted in the kth panel where $Z_f = \{1, 2, \ldots, 15\}$. The horizontal line in the kth panel represents the MI of one feature at k Bark. In each panel, excess values of the JMI above the horizontal line are the information gain due to a second feature.

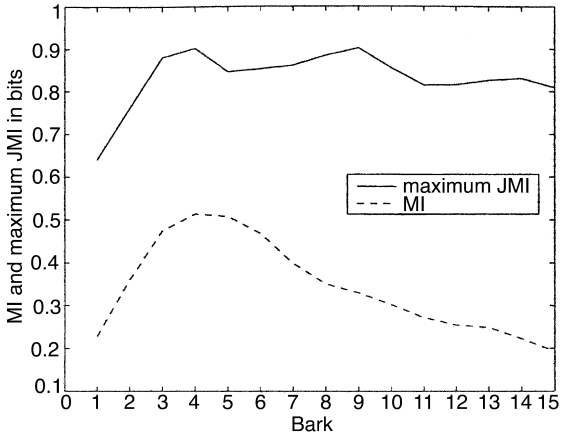


Fig. 8. The maximum JMI, $J(k, j_k^*)$, versus the MI in the kth band. The gap between two curves is equal to $J_1(k, j_k^*) - I(X(f_k, t); Y_t)$ which is the information increase due to one contextual feature at the current time frame but in a different frequency band. The average information increase across 15 bands is 0.49 bits.

second feature given the first feature in each of the frequency bands. The band indices $\{j_k^*\}$ are listed in Table 3.

4.4.2. Additional feature at the same frequency but at a different time

To examine the relevance of two features in the same frequency band but at different times in predicting the phoneme, we define the JMI $J_2(k, d)$ as follows:

$$J_2(k, d) = I(X(f_k, t), X(f_k, t + d); Y_t), \quad k \in Z_f, \quad d \in S(L, \Delta).$$

Here, $L = 60$ frames and $\Delta = 10$ ms and we define $J_2(k, 0) = I(X(f_k, t); Y_t)$. When $k = 5$, the JMI of the fifth frequency band $J_2(5, d)$ as a function of time shift is shown in Fig. 9.

Information is gained by using a feature located at a different time (points labeled by B, C, D or E

Table 3
The frequency bands of optimal second features

k (Bark)	j_k^* (Bark)
1	6
2	6
3	9
4	9
5	9
6	3
7	3
8	4
9	4
10	4
11	4
12	5
13	5
14	5
15	5

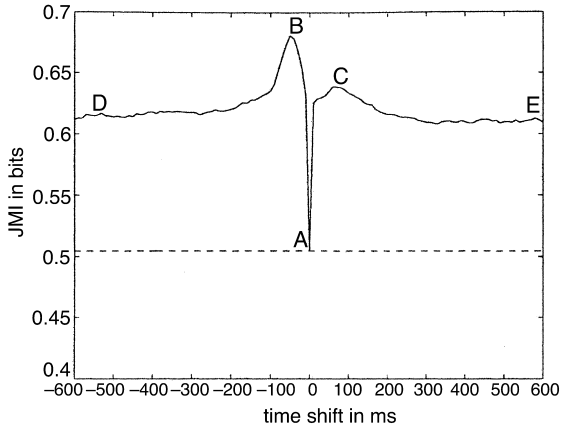


Fig. 9. The JMI of the fifth frequency band as a function of time shift in milliseconds. Information is gained by using a feature located at a different time (e.g., B, C, D or E) in addition to the feature from the current time frame (located at point A).

in Fig. 9) in addition to the feature from the current time (labeled by point A). Point A represents the MI, $J_2(5,0) = I(X(f_5,t);Y_t)$, which is also shown by the maximum value in Fig. 5. The asymptotic level of the JMI is 0.61 bits even when the absolute time shifts are larger than 600 ms. This suggests that the second feature, even if it is far apart from the first one, will increase the JMI by removing a constant bias in the data. The maximum of the JMI is achieved at point B corresponding to a feature 50 ms before the current time. Based on the data set we used to estimate the JMI, the average phoneme duration is about 65 ms. This may be the reason why the frame at -50 ms gives the maximum additional information for phonetic classification. Note that the current frame is at time zero and so negative time is in the past while the positive time is in the future.

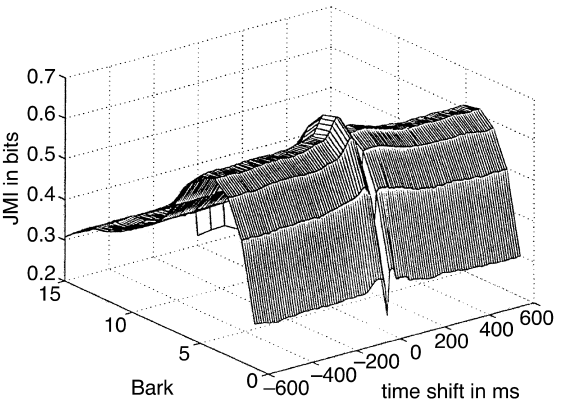
The spread of the additional information is asymmetric in time with most of the supporting information found in the intervals $[-200,-20]$ and $[20,200]$ ms. Around the point B in Fig. 9, the mutual information increases from 0.5 bits for a single feature to around 0.68 bits for two features. This may indicate possible asymmetries in the coarticulation pattern with a weaker anticipatory coarticulation. The asymmetry may also indicate a causal direction in time, i.e., the label at the current time is more predictable from features in the past rather than from features in the future.

For the 50-speaker data set,

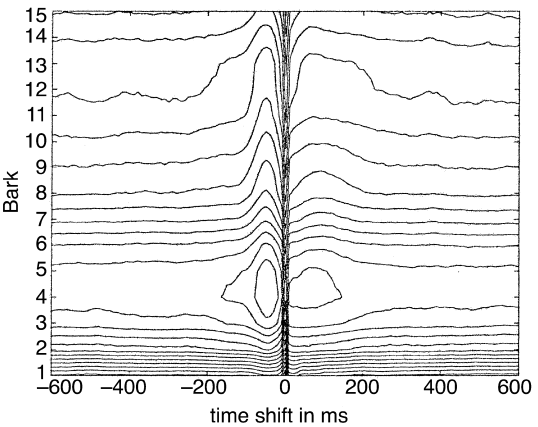
$$J_2(4,50) = \max_{k\in Z_f, d\in S(60,10)} J_2(k,d) = 0.69 \text{ (bits)}. \quad (7)$$

The JMI patterns are quite similar in all frequency bands. Fig. 10 shows the JMI, $J_2(k,d)$, as a function of the Bark index k and the time shift d.

The same JMI function is shown in Fig. 11 by 15 panels. The kth panel shows the JMI of two features at k Bark, $I(X(f_k,t),X(f_k,t+d);Y_t)$ at different time shifts. The JMI at 5 Bark shown in



(a)



(b)

Fig. 10. The JMI as a function of the frequency band index k and the time shift d: (a) 3D plot; (b) contour plot.

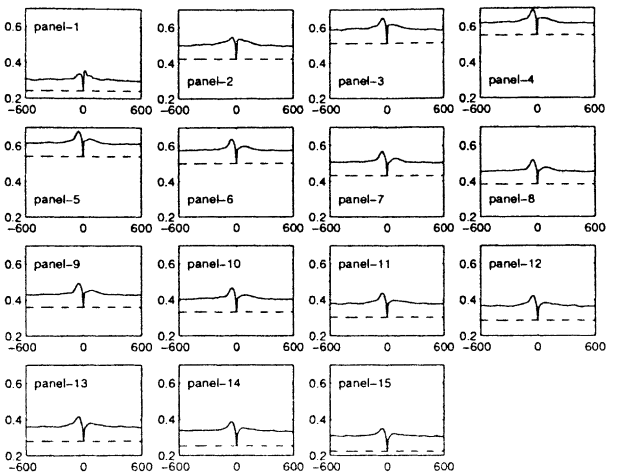


Fig. 11. The JMI between the phonetic label variable and two features in the same band. The kth panel (counting row-by-row) shows the JMI, $J_2(k,d)$ for $d \in S(60,10)$.

Fig. 9 is an enlargement of the fifth panel in Fig. 11.

The information gain due to a contextual feature in a different frame is clearly seen in Fig. 11. The maximum JMI in the kth band is

$$J_2(k,d_k^*) = \max_{d \in S(60,10)} J_2(k,d).$$

Given the first feature at time shift 0 in the kth band, the second feature in the same band achieves a maximum information gain at the time shift d_k^* ms. The optimal time shifts are given in Table 4.

Table 4
The optimal time shifts in different bands

k (Bark)	d_k^* (ms)
1	20
2	−40
3	−50
4	−50
5	−50
6	−50
7	−50
8	−50
9	−50
10	−50
11	−50
12	−50
13	−50
14	−50
15	−60

This table shows that adding a feature at 50 ms before the current time gives the maximum information gain in all bands except the very low and the very high frequency bands.

Figs. 8 and 12 show the information gain of the second feature given the first feature in each band. But in Fig. 8 the second feature is in the same time frame but in a different band as the

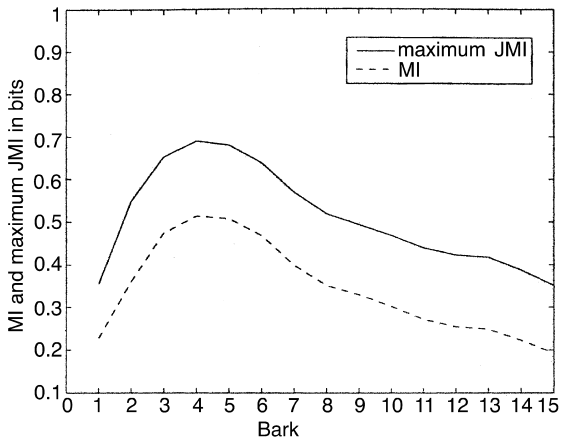


Fig. 12. The maximum JMI, $J_2(k,d_k^*)$, versus the MI in the kth band. The gap between the two curves at k Bark is $J_2(k,d_k^*) - I(X(f_k,t);Y_t)$ which is the information gain due to one contextual feature in the same band but at a different time frame. The average information increase across 15 bands is 0.17 bits.

first feature, while in Fig. 12 the second feature is in the same band but in a different time frame as the first one.

4.4.3. Additional feature in a neighborhood of the first one

We have so far estimated the JMI of two features either in the same current frame or in the same frequency band. The question still remains whether a pair of features in different frames and frequency bands would give a larger JMI.

When we maximized the JMI in (6) and (7), we considered the following two contexts of the feature $X(f_4,t)$:

$$C_f = \{X(f_j,t) : j \in Z_f\}, \tag{8}$$

$$C_T = \{X(f_4,t+d) : d \in S(60,10)\}. \tag{9}$$

In the context C_f, the feature $X(f_9,t)$ gives the maximum information gain while in the context C_T the feature $X(f_4,t-50)$ provides the maximum information gain. In a bigger context $C_f \cup C_T$, the feature $X(f_9,t)$ achieves the maximum information gain.

In general, it seems reasonable to conclude that an additional feature at the same time frame but at a different frequency band provides the most additional information. To check this, we consider an even bigger context of the feature $X(f_4,t)$:

$$C_{f,T} = \{X(f_j,t+d) : j \in Z_f,\ d \in S(60,10)\}. \tag{10}$$

For every feature in $C_{f,T}$, the conditional MI, $I(X(f_j,t+d);Y_t|X(f_4,t))$, is estimated and plotted in Fig. 13(a) and (b). The distribution of the information gain of a second feature in the neighborhood of the first feature is again asymmetric (see Fig. 13(b)). The result confirms that the second feature $X(f_9,t)$ at the current frame gives the maximum information gain in the larger context $C_{f,T}$.

4.4.4. Phonetic classification experiments

We use MI to measure the relevance of the features for phonetic classification based on the following hypothesis: the classifiers with features of high MI or high JMI perform better. To verify this, we conducted several phonetic classification experiments using MLP classifiers.

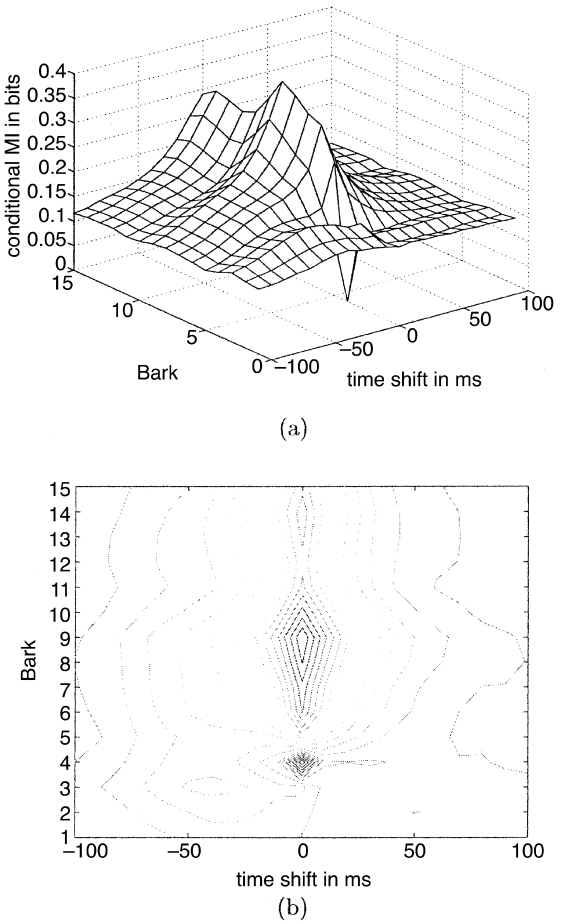


Fig. 13. The conditional MI $I(X(f_j,t+d);Y_t|X(f_4,t))$ in the context $C_{f,T}$: (a) 3D plot; (b) contour plot. The feature $X(f_9,t)$ gives the maximum information in this context.

Each MLP classifier used in the experiments is a fully connected feed-forward neural network with a single hidden layer. The units in the hidden layer use a sigmoid activation function while the output units employ a soft-max function defined as follows. Let $\boldsymbol{y} = (y_1,\ldots,y_n)'$ be the hidden layer outputs and $\boldsymbol{W} = (w_{ij})$ be a weight matrix of $m \times n$ connecting the hidden layer and the output layer. Let $\boldsymbol{u} = \boldsymbol{W}\boldsymbol{y}$ and

$$z_i = \frac{e^{u_i}}{\sum_{k=1}^{m} e^{u_k}}, \quad i = 1,\ldots,m,$$

where $(z_1,\ldots,z_m)$ are the soft-max outputs of the MLP.

The software was developed at the International Computer Science Institute, Berkeley, California. We used $n = 300$ units in the hidden layer and $m = 19$ units in the output layer, where each output unit corresponds to one of the 19 phonetic classes used in this article.

The whole data set used in the experiments consists of 3 hours of speech from 210 speakers. It is divided into three subsets: 67% of the whole data set being a training set, 8% being a cross-validation set, and 25% being a test set. Each MLP is trained using an on-line error-back-propagation algorithm with relative entropy criterion and binary outputs. During the training process, the performance of the network on the cross-validation set is used to determine when to stop the training and how to set the learning rate after each training epoch. The training starts at a learning rate of 0.008 that is held constant until the performance on the cross-validation set does not improve. For every subsequent epoch the learning rate is divided by a factor of 2. The training is stopped when the performance on the cross-validation set shows no further improvement. The accuracy of the MLP is computed using the test set.

To speed up on-line training, the input–output patterns are randomly presented to the network. The input features are normalized to have zero mean and unit variance. The normalizing mean and variance values are computed only on the training data and applied to the test data during recognition.

We did three experiments to obtain frame accuracies of the MLPs with different inputs for phonetic classification.

In the first experiment, each MLP was trained and tested using a single feature in one of 15 critical bands. The frame accuracies of these single-input MLPs are shown in Fig. 14. The results are consistent with the MI shown in Fig. 3. The features from 3 to 8 Bark have relatively high MI. Consequently, the MLPs with a single feature in these bands (3–8 bands) achieve relatively high accuracies.

In the second experiment, each MLP had two inputs in two different critical bands. The results are shown in Fig. 15. The kth panel in Fig. 15 shows the accuracies of the MLPs with one input in the kth band and another input in other bands. The accuracies of the MLPs shown in Fig. 15 are compatible with the patterns of the JMI of two features shown in Fig. 7.

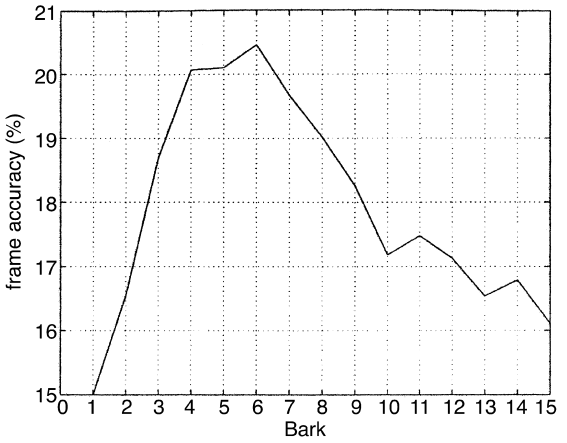


Fig. 14. Frame accuracies of the MLP with one frequency feature.

In the third experiment, each MLP has 30 inputs consisting of two vectors $\boldsymbol{x}_0$ and $\boldsymbol{x}_d$. Here,

$$\boldsymbol{x}_0 = (X(f_1, t), \ldots, X(f_{15}, t))$$

and

$$\boldsymbol{x}_d = (X(f_1, t + d), \ldots, X(f_{15}, t + d)).$$

The accuracies of the MLPs shown in Fig. 16 are in agreement with the JMI in Fig. 9. The frame accuracy is maximum when $d = -50$ ms.

Fig. 9 only depicts the JMI for the fifth band while Fig. 16 depicts the frame accuracy for those MLPs using inputs from all bands. Therefore, Fig. 16 may reflect a combined effect across all bands.

5. MI between spectral features and speaker-channel labels

Phonetic variability is not the only source of variability in speech. The other non-linguistic sources of variability are different speakers and different communication channels. The concept of mutual information can also be applied to non-linguistic information. To illustrate this, we show

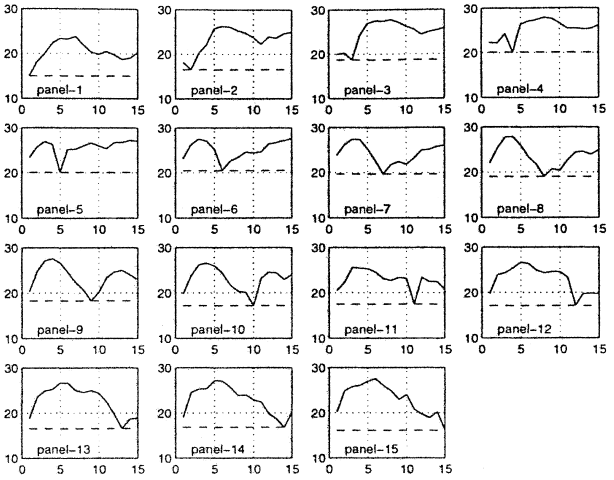


Fig. 15. Frame accuracies of the MLP with two frequency features.

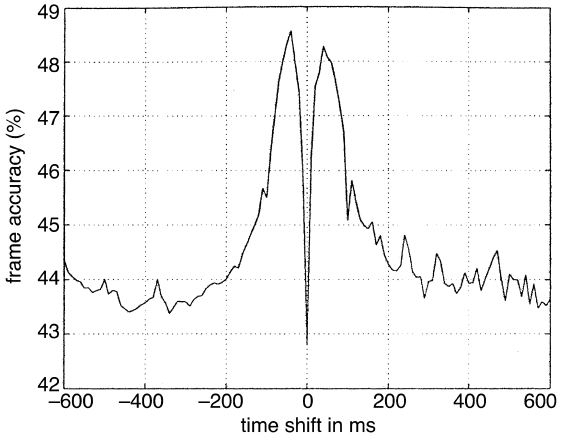


Fig. 16. Frame accuracies of the MLP with 30 inputs.

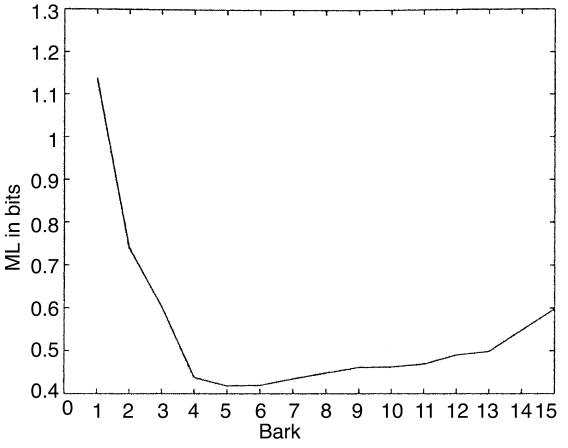


Fig. 17. The MI between the speaker-channel labels and one feature for speaker-channel classification. The 50-speaker data set is used to evaluate the MI.

how the non-linguistic information is distributed in different frequency bands.

In this case we use speaker-channel labels $\{Z_t\}$ to label the frames in the time–frequency plane. The MI estimates $I(X(f_k,t);Z_t)$ shown in Fig. 17 indicate that for speaker-channel classification the low and high frequency bands are most relevant and the fourth and fifth bands are the least relevant. This is in contrast to the MI for phonetic classification. The low frequency components in the first and second bands contain line noise. It is possible that features in the first and second bands give more information than those in other bands to differentiate speaker-channels.

For each $k = 1,\ldots,15$, a sequence of JMI estimates $\{I(X(f_k,t),X(f_j,t);Z_t)\}_{j=1}^{15}$ is plotted in the kth panel in Fig. 18. The first panel shows that if one feature is in the first band then the most relevant feature for speaker-channel classification is in the second band. Other panels show that if one feature is not in the first band then the most relevant feature is always in the first band. The maximum JMI is

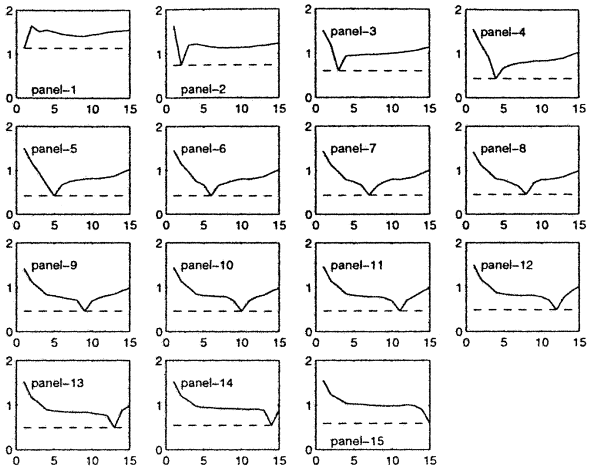


Fig. 18. The JMI between the speaker-channel labels and two features for speaker-channel classification. The kth panel (counting row-by-row) shows the JMI, $I(X(f_k,t),X(f_j,t);Z_t)$ for $j=1,\ldots,15$.

$$I(X(f_1,t),X(f_2,t);Z_t)=\max_{k,j} I(X(f_k,t),X(f_j,t);Z_t)$$
$$=1.65 \text{ (bits)}.$$

The MI and JMI for speaker-channel classification can be used to measure the sensitivity of the features to speaker-channel changes. Features in the low and high frequency bands that have high speaker-channel MI are sensitive to the speaker-channel variability and may not be good features for phonetic classification. This is consistent with the result shown in Fig. 3.

6. Discussion and conclusions

The question of how information is coded in the speech signal is of both practical and theoretical importance. For the past few years we have been advocating that established data analysis techniques could be applied to large speech corpora to gain more knowledge on this important topic. This work is an attempt in this direction in which we use a relatively large phoneme-labeled data set and the information-theoretic concept of mutual information to gain insight into the distribution of phonetically relevant information and speaker-channel variability.

Analysis of the distributions of critical-band spectral energy observations shows that these distributions are non-Gaussian. This motivates the analysis of non-linear dependence using the mutual information.

Analysis of the MI between phonetic labels and spectral energy observations at points in the time–frequency plane revealed that along the frequency axis, all frequency bands carry information about the underlying phoneme label, with dominant information around 3–8 Bark (313–1059 Hz).

Analysis of the JMI between the phonetic labels and spectral energy observations at the current frame at two different points in different frequency bands showed that the addition of a feature at a different frequency band considerably increases the information about the phonetic label.

Along the time axis, from Fig. 5, individual features more than 100 ms in the past or in the future are not relevant for the phonemes at the current time frame. However, from Fig. 9, jointly with a feature at the current time frame, the second features more than 100 ms in the past or the future are still relevant for predicting the phoneme.

It is the analysis of the JMI along the time axis which we find the most interesting. Even though the additional information from the second feature is not as high as in the case of the second feature at the same time frame and different frequency band as discussed above, it indicates that significant information for phonetic classification is spread in

time over at least 200 ms. This spread is asymmetric in time with relatively more information found in the interval $[-200,-20]$ ms and less found in the interval $[20,200]$ ms (see Fig. 9).

The MI and the JMI between the speaker-channel label and the features in the time–frequency plane show that the features most relevant for phonetic classification are least relevant for speaker-channel classification.

Acknowledgements

The work was supported by DoD (MDA904-98-1-0521, MDA904-97-1-0007) and by NSF (IRI-9712579) grants to the Anthropic Signal Processing Group at OGI. We thank David Broad, a SILCOM.COM scientist in Santa Barbara, for his insightful suggestions.

Appendix A. Proof of the invariance property

We want to show that if $f_i(y), i=1,\ldots,n$, and $g(y)$ are differentiable monotonic functions, then

$$I(f_1(X_1),\ldots,f_n(X_n);g(Y)) = I(X_1,\ldots,X_n;Y).$$

In fact, let $Z=(Z_1,\ldots,Z_n,Z_{n+1})=(f_1 \quad \text{trun}(X_1),\ldots,f_n(X_n),g(Y))$, then

$$p(z_{n+1}) = p(y)/|g'(y)|,$$
$$p(z_1,\ldots,z_n,z_{n+1}) = p(x_1,\ldots,x_n,y)/|\Delta_{n+1}|,$$
$$p(z_1,\ldots,z_n) = p(x_1,\ldots,x_n)/|\Delta_n|,$$

where $|\Delta_{n+1}| = |\Delta_n||g'(y)|$ and $|\Delta_n| = \prod_{i=1}^{n}|f_i'(y_i)|$. Hence,

$$\begin{aligned}
&I(Z_1,\ldots,Z_n;Z_{n+1})\\
&\quad = \int p(z_1,\ldots,z_n,z_{n+1})\log\frac{p(z_1,\ldots,z_n,z_{n+1})}{p(z_1,\ldots,z_n)p(z_{n+1})}\,\mathrm{d}z_1\cdots\mathrm{d}z_{n+1}\\
&\quad = \int \frac{p(x_1,\ldots,x_n,y)}{|\Delta_{n+1}|}\log\frac{p(x_1,\ldots,x_n,y)|\Delta_n||g'(y)|}{|\Delta_{n+1}|p(x_1,\ldots,x_n)p(y)}|\Delta_{n+1}|\mathrm{d}x_1\cdots\mathrm{d}x_n\,\mathrm{d}y\\
&\quad = \int p(x_1,\ldots,x_n)\log\frac{p(x_1,\ldots,x_n,y)}{p(x_1,\ldots,x_n)p(y)}\mathrm{d}x_1\cdots\mathrm{d}x_n\,\mathrm{d}y\\
&\quad = I(X_1,\ldots,X_n;Y).
\end{aligned}$$

References

Barrows, G., Sciortino, J., 1996. A mutual information measure for feature selection with application to pulse classification. In: IEEE International Symposium on Time–Frequency and Time-Scale Analysis, pp. 249–253.

Battiti, R., 1994. Using mutual information for selecting features in supervised neural net learning. IEEE Transactions on Neural Networks 5 (4), 537–550.

Bilmes, J., 1998. Maximum mutual information based reduction strategies for cross-correlation based joint distribution modeling. In: ICASSP98, pp. 469–472.

Bonnlander, B., 1996. Nonparametric selection of input variables for connectionist learning. Technical report. PhD Thesis. University of Colorado.

Cole, R., Fanty, M., Noel, M., Lander, T., 1994. Telephone speech corpus development at CSLU. In: ICSLP, Yokohama, pp. 1815–1818.

Cover, T.M., Thomas, J.A., 1991. Information Theory. Wiley, NewYork.

Hermansky, H., 1990. Perceptual linear predictive PLP analysis of speech. J. Acoust. Soc. Amer. 87 (4), 1738–1752.

Hermansky, H., 1998. Should recognizers have ears? Speech Communication 25, 3–27.

Morris, A., Schwartz, J.-L., Escudier, P., 1993. An information theoretical investigation into the distribution of phonetic information across the auditory spectrogram. Comput. Speech Language 7 (2), 121–136.

Stuart, A., Ord, J.K., 1994. Kendall's Advanced Theory of Statistics. Edward Arnold, Paris.

Venables, W.N., Ripley, B.D., 1994. Modern Applied Statistics with S-Plus. Springer, New York.

Yang, H., Moody, J., 1999. Feature selection based on joint mutual information. In: Advances in Intelligent Data Analysis (AIDA), Computational Intelligence Methods and Applications (CIMA), International Computer Science Conventions, Rochester, NY.

Yang, H., van Vuuren, S., Hermansky, H., 1999. Relevancy of time–frequency features for phonetic classification measured by mutual information. In: ICASSP'99, Phoenix, pp. I:225–228.

Yang, H., Moody, J., 2000. Data visualization and feature selection: New algorithms for nongaussian data. In: Advances in Neural Information Processing Systems, Vol. 12.