RELEVANCY OF TIME-FREQUENCY FEATURES FOR PHONETIC CLASSIFICATION MEASURED BY MUTUAL INFORMATION

Howard Yang[1] *and Sarel van Vuuren*[2] *and Hynek Hermansky*[1,2]

Department of Computer Science and Engineering[1], Department of Electrical and Computer Engineering[2]
Oregon Graduate Institute, 20000 NW, Walker Rd., Beaverton, OR 97006-8921
hyang@cse.ogi.edu, sarelv@ece.ogi.edu, hynek@ece.ogi.edu

ABSTRACT

In this paper we use mutual information to study the distribution in time and frequency of information relevant for phonetic classification. A large database of hand-labeled fluent speech is used to (a) compute the mutual information between phoneme labels and a point of logarithmic energy in the time-frequency plane and (b) compute the joint mutual information between phoneme labels and two points of logarithmic energy in the time-frequency plane.

1. INTRODUCTION

The speech research community has at its disposal rather large speech databases which are mainly used for training and testing ASR systems. There has been relatively few efforts to date to use such databases for deriving reuseable knowledge about speech and speech communication processes which could be used for improvements of ASR technology. In this paper we describe some initial 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.

Over the past five years we have been advocating a move towards speech analysis techniques which would selectively use relatively large temporal segments of speech signal, believing that the information about a phoneme is not localized to the region of that phoneme only, but rather that it is spread over a substantial (about one syllable long) segment of the signal [1]. Thus, any evidence of the way in which the information about the underlying linguistic process is distributed in the signal is of importance in our quest.

Mutual information deals with the question of how various elements of an information stream relate to each other. Morris [2] used mutual information to find the critical points of information for classifying CVC utterances. Recent work of Bilmes [3] showing that the information appears to be spread over relatively long temporal spans spurred our interest in the technique. While Bilmes used mutual information between two variables on non-labeled data to reveal the mutual dependencies between the components of the spectral energies in time and frequency, we decided to focus on joint mutual information between the phonetic labels of a hand-labeled database and logarithmic energies at two points in the time-frequency plane. We use this concept to gain insight into how information about phonemes is distributed in time and frequency.

We represent the information in time and frequency by short-term critical-band logarithmic energy $X(f_k, t)$. This is a feature representation commonly used in phonetic classification. In particular, the goal is to determine the relevancy of $X(f_k, t-d)$ across all frequencies f_k and in a context window $-D \leq d \leq +D$ for classification of a phoneme labeled Y centered at time t.

2. DATA

Results are based on about 3 hours of phonetically labeled telephone speech from the English portion (Stories) of the OGI multi-lingual database [4]. This represents approximately 50 seconds of extemporaneous speech from each of 210 different speakers. The speech data is labeled by a variable Y taking 19 values from a set of commonly occurring phonemes. The average phoneme duration is about 65 ms and the average number of phoneme instances is 3440 for a grand total of 65421 phoneme instances.

Acoustic features X for the experiments are derived from a short-time analysis of the speech signal with a 20 ms analysis window (Hamming) advanced in 10 ms steps. The logarithmic energy at a frequency f_k is computed from the squared magnitude FFT using a critical-band spaced (log-like in the frequency variable) weighting function in a manner similar to that of the computation of Perceptual Linear Prediction coefficients [5]. In particular, the 5-th, 8-th and 12-th bands are centered around 0.5, 1 and 2 kHz respectively.

3. MUTUAL INFORMATION

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

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

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 $p(x)p(y)$

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

The MI measures the statistical dependency 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 mutual information can be computed analytically. However, this assumption is usually not true in practice. To estimate mutual information, one needs to approximate the probability density function $p(x, y)$. The typical density approximation methods for estimating mutual information are histogram, kernel function and EM algorithm [6, 3, 7]. Using the histogram for estimating MI is equivalent to computing the quantized version of the MI (see [8]).

The mutual information $I(X;Y)$ was used in [9, 10] for feature selection. It was also used in [2] to find those spectro-temporal areas in speech data which are most useful for classification and in [3] to estimate this information in the feature-vector joint distribution.

The correlation coefficient is often used to probe dependencies between variables. The correlation coefficient and the mutual information have three major differences. First, the correlation coefficient measures linear dependencies or second order statistics between random variables, whereas the MI measures the non-linear statistical dependencies between random variables. Second, the correlation coefficient is only invariant to component-wise linear transforms while the mutual information is invariant to component-wise monotonic transforms, $I(f(X), g(Y)) = I(X, Y)$ if the two functions $f(x)$ and $g(x)$ are monotonic and differentiable. Third, and most importantly, the mutual information works for categorical data while the correlation coefficient does not. For the phonetic classification problem, we encounter a classification variable for labeling each data frame. The dependencies between this variable and the feature variables can be probed by the mutual information but not by the correlation coefficient.

The joint mutual information (JMI) $I(X_i, X_j; Y)$ between a random vector $\boldsymbol{X} = (X_i, X_j)$ and Y is defined by Eqn. 1 if X is replaced by $\boldsymbol{X}$. In [7] it has been used for input variable selection for radar pulse classification. In this paper, we shall apply both the MI and the JMI to identify those frames and frequency bands most relevant for phonetic classification.

4. MUTUAL INFORMATION IN SPEECH

To study the dependency structure in the speech data we need the mutual information, because as we show next, the data is strongly non-Gaussian.

4.1. A simple non-Gaussianity test of speech data

Using a histogram 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 bin number. Given a data set $\{x_t, t = 1, \cdots, T\}$, for Gaussian distributions, one may choose $\log_2 T + 1$ as the number of bins by Sturges's rule. For non-Gaussian distributions, one should 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 [11]).

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

$$S = \frac{1}{\sqrt{6T}\hat{\sigma}^3} \sum_{t=1}^{T} (x_t - \hat{\mu})^3,$$

$$K = \frac{1}{\sqrt{24T}\hat{\sigma}^4} \sum_{t=1}^{T} (x_t - \hat{\mu})^4 - \sqrt{\frac{3T}{8}},$$

where $\hat{\mu}$ and $\hat{\sigma}^2$ are 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 is Gaussian, both S and K are standard Gaussian asymptotically (see Vol.1 in [12]).

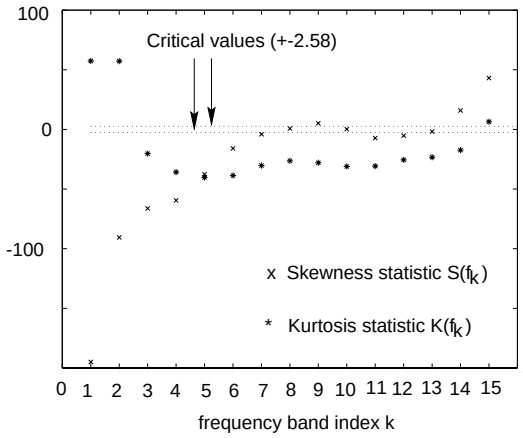

Figure 1: The statistics S and K in each frequency band of the spectrum of 50 speakers.

For each logarithmic spectral energy $X(f_k)$ we compute the statistics $S(f_k)$ and $K(f_k)$. The results in Fig. 1 are based on a 50 speaker subset of the speech data. Note that at the significance level $\alpha = 0.01$ the critical values for the standard normal distribution are ± 2.58. It is shown in Fig. 1 that the logarithmic spectral energies are strongly non-Gaussian. All absolute values of the K statistics exceed the critical value. 12 out of 15 absolute values of the S statistics exceed the critical value. From the K statistics, we know that most of the signals are sub-Gaussian with negative kurtosis except the signals in the two lowest frequency bands and the highest frequency band which are super-Gaussian[1]. Since the logarithmic spectral energies are non-Gaussian, instead of using the correlation coefficient, we use the MI to probe the dependencies between features and classification variable.

4.2. MI between a feature and phonetic label

In our speech data set each frame is assigned a phonetic label. Let Y be the target variable for the phonetic classi-

[1] From Fig. 1 it is seen that the first, second and 15-th band have rather different statistics. It is worthwhile to note that these bands fall mostly outside the telephone bandwidth (300-3400Hz) and may be noisy or less reliable for phonetic classification.

fication. Based on the data set

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

we estimate the mutual information as a function of frequency between a feature from the current frame and the target variable Y:

$$I(X(f_k, t-d); Y),\ k = 1, \cdots, 15,\ d = -D \cdots + D.$$

Fig.2 shows this mutual information as a function of frequency, with $d = 0$, for two sets of speech data with 100 speakers in each set. Fig. 2 reveals that along the frequency

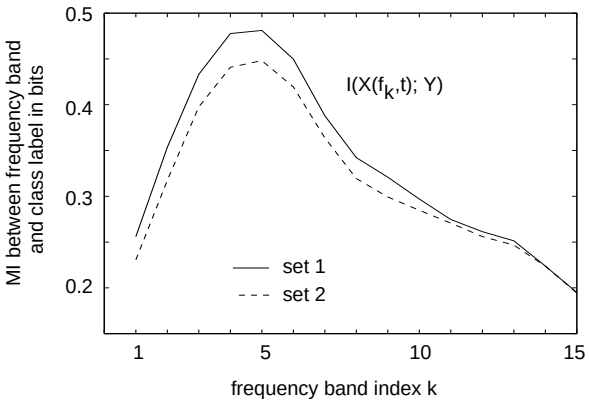

Figure 2: Mutual information as a function of frequency and the classification target variable for two disjunct sets of 100 speakers each. The two speaker sets exhibit similar MI patterns across frequency.

axis, all frequency components carry information about the underlying phoneme label, with dominant information in the 5-th frequency band (i.e. around 2-8 Bark or 350-1300 Hz).

Fig. 3 shows the mutual information $I(X(f_k, t-d); Y)$ as a function of time shift d (in frames) with $f_k = f_5$. In interpreting Fig. 3 consider two data points $\{X(f_k, t), Y_t\}$ and $\{X(f_k, t-d), Y_{t-d}\}$. Then in average $X(f_k, t-d)$ can be said to contain little information on $Y(t)$ when the absolute time shift is greater than about 100ms. Conversely, $X(f_k, t-d)$ does contain information on $Y(t)$ for absolute time shifts less than 100ms. This suggests that one may want to use contextual information in a window of about 100ms to either side of the frame that is to be classified.

4.3. JMI between two features and phonetic label

The MIs displayed in Figs. 2-3 show the relevancy of each individual logarithmic spectral energy. In practice, we use spectral energies jointly for phonetic classification. If we use two feature points in the same frame but in different frequency bands, the joint MI between these two points and the classification variable is $I(X(f_i, t), X(f_k, t); Y)$. For the 5-th band for example, define

$$J_1(k) = I(X(f_5, t), X(f_k, t); Y),\ k = 1, \cdots, 15$$

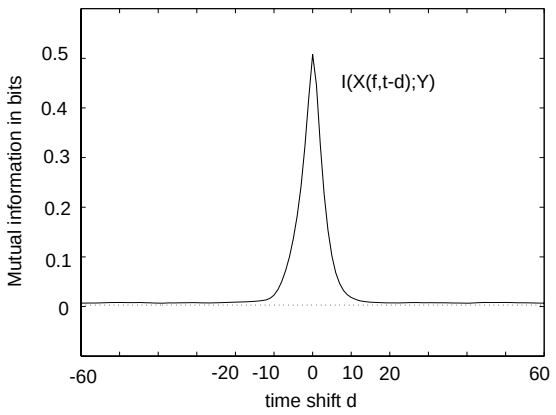

Figure 3: Mutual information as a function of time shift (in frames) and the classification target variable for the 5-th frequency band. The dotted line shows the lower bound (0.0028 bits) which is the mutual information between the 5-th frequency band and scrambled phonetics labels.

where $J_1(5) = I(X(f_5, t); Y)$. The joint MI as a function of frequency is shown in Fig.4. By the chain rule for informa-

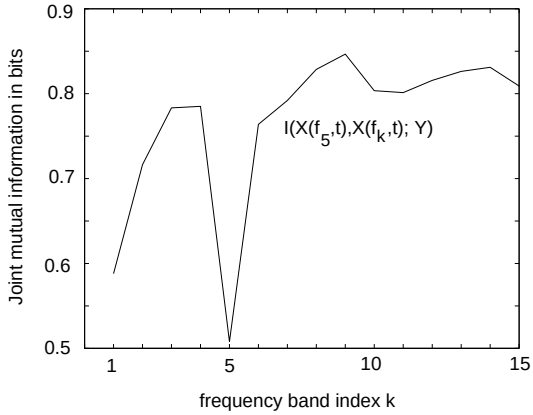

Figure 4: JMI as a function of frequency for a feature point in the 5-th band and a feature point in the k-th band.

tion [8], we have

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

We call $I(X(f_k, t); Y | X(f_5, t))$ the relative mutual information between $X(f_k, t)$ and Y conditional on $X(f_5, t)$. It is shown by Fig. 4 that this relative mutual information reaches a maximum in frequency band 9 (from about 0.5 bits for a single measurement to as much as 0.85 bits for an additional measurement at 9 Bark). In general we find that the inclusion of a second point always provides information in addition to that provided by a single point.

To examine the relevancy of two feature points in the

same frequency band we use the joint mutual information

$$J_2(d) = I(X(f,t), X(f,t-d); Y), d = -D, \cdots, +D.$$

Here, we define $J_2(0) = I(X(f,t); Y)$. The joint MIs of the 5-th frequency band are shown in Fig.5. Information is

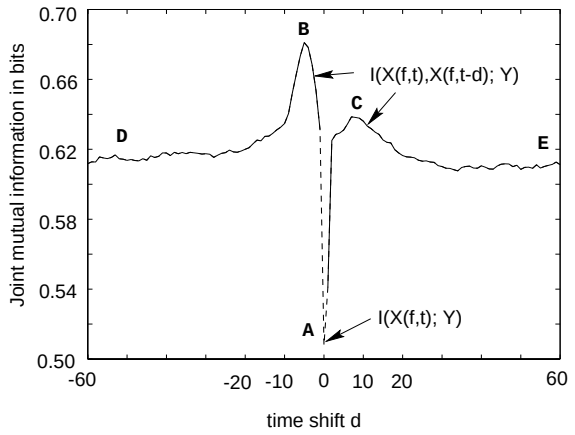

Figure 5: The JMI of the 5-th frequency band as a functin of time shift (in frames). Information is gained by using a feature located at a different time (e.g. B, C, D, or E) in addition to the feature from the current frame (located at point A). See the text for a discussion.

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 frame (located at point A). Point A represents the mutual information $J_2(0) = I(X(f,t); Y)$ which is also shown by the maximum value in Fig. 3. The asymptotic level of the JMI is 0.61 bits and indicates that a systematic bias may be corrected by using more than one feature in time[2]. The spread is asymmetric in time with most of supporting information found between 20 and 80 ms beyond the current frame (points B and C) where it can yield an increase in information from 0.5 bits for a single measurement to around 0.68 bits for two measurements). This indicates possible asymmetries in the coarticulation pattern with a weaker anticipatory coarticulation.

5. CONCLUSIONS

Analysis of distributions of critical band spectral energies showed that the distributions are non-Gaussian, thus requiring analysis of non-linear dependencies using the mutual information criteria.

Analysis of mutual information between phonetic labels and logarithmic energies at points in the time-frequency plane revealed that along the frequency axis, all frequency components carry information about the underlying phoneme label, with dominant information around 2-8 Bark (350-1300 Hz). Along the time axis, on average, components no further than about 100 ms outside the labeled segment are still relevant for the classification of a given phoneme.

The analysis of joint mutual information between a label and logarithmic energy at the current frame at two different points in frequency shows that the addition of a measurement at the second frequency considerably increases the information about the phonetic label.

It is the analysis of joint mutual information along the time axis which we find the most interesting. Even though the additional information from the second measurement is not as high as in the case of the second measurement at the same time and different frequency as discussed above, it indicates that significant information for phonetic classification is spread in time over at least 200 msec but possibly more. This spread is asymmetric in time with most supporting information found between 20 and 80 ms beyond a given time instant.

[2]Cepstral mean subtraction is a well-known technique used to correct for a long-term bias in a logarithmic spectral energy as introduced by a time-invariant transmission channel.

6. REFERENCES

[1] H. Hermansky, "Should recognizers have ears?, invited paper," *Speech Communication*, vol. 25, no. 1-3, pp. 3–27, 1998.

[2] A. Morris, J.-L. Schwartz, and P. Escudier, "An information theoretical investigation into the distribution of phonetic information across the auditory spectogram," *Computer Speech & Language*, vol. 7, pp. 121–136, April 1993.

[3] J. A. Bilmes, "Maximum mutual information based reduction strategies for cross-correlation based joint distribution modeling," in *ICASSP98* , pp. 469–472, April 1998.

[4] R. Cole, M. Fanty, M. Noel, and T. Lander, "Telephone speech corpus development at CSLU," in *ICSLP*, (Yokohama), pp. 1815–1818, Sept. 1994.

[5] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, pp. 1738–1752, April 1990.

[6] B. Bonnlander, "Nonparametric selection of input variables for connectionist learning," tech. rep., PhD Thesis. University of Colorado, 1996.

[7] H. H. Yang and J. Moody, "Input variable selection based on joint mutual information," tech. rep., Oregon Graduate Institute of Science and Technology, 1998.

[8] T. M. Cover and J. A. Thomas, *Information Theory*. John Wiley & Sons, Inc., 1991.

[9] G. Barrows and J. Sciortino, "A mutual information measure for feature selection with application to pulse classification," in *IEEE Intern. Symposium on Time-Frequency and Time-Scale Analysis* , pp. 249–253, 1996.

[10] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Trans. on Neural Networks*, vol. 5, pp. 537–550, July 1994.

[11] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S-Plus*. New York: Springer, 1994.

[12] A. Stuart and J. K. Ord, *Kendall's Advanced Theory of Statistics*. Edward Arnold, 1994.