RC 22212 (99149) October 18, 2001 **Computer Science**

IBM Research Report

Maximizing Information Content of Feature Extraction for Improved Classification with Applications in Speech Recognition

M. Padmanabhan
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Maximizing Information Content of Feature Extraction for Improved Classification with Applications in Speech Recognition

M. Padmanabhan
IBM T. J. Watson Research Center
P. O. Box 218, Yorktown Heights, NY 10598
mukund@us.ibm.com

Abstract—**In this paper we consider the problem of extracting maximally informative feature vectors for classification, and explore the connection between the information content in the feature vector (quantified by the mutual information between the feature vector and the classes predicted by the classifier) and the classification accuracy. We explore these ideas in the context of a speech recognition system, where the classification problem is one of predicting the phonetic class, given an observed acoustic feature vector. The connection between information content and classification accuracy is first explored in the context of adding features to a baseline cepstral feature extraction scheme that is very commonly used in speech recognition applications. The features we chose to study relate to the spectral location and energy value of the spectral peaks in the speech signal (similar to formant frequencies). We first quantify the amount of incremental information that the spectral peak features can provide over and above cepstral feature vectors, and subsequently show a connection between the incremental information and speech recognition accuracy. Subsequently, the idea of optimizing mutual information to improve recognition accuracy is generalized to a linear transformation of the underlying features. We show that several prior methods to compute linear transformations (such as linear/heteroschedastic discriminant analysis) can be interepreted in this general framework of maximizing the mutual information. Finally, experimental results are provided that show that designing the feature space to maximize the mutual information can lead to improvements in the accuracy.**

I. Introduction

In this paper, we focus on the problem of designing a feature extraction method for classifiers. We develop the ideas in the context of a speech recognition system, however, we believe that the concepts should generalize to other classification problems as well. In speech recognition, the classification problem corresponds to predicting a phonetic class given an observed acoustic feature vector. There are several methods that are used today to extract acoustic feature vectors - most of them comprise of the step of extracting a multi-dimensional feature vector from the sampled speech signal at a uniform frame rate (typically every 10 ms). Quite often the procedure for extracting this feature vector is motivated by the workings of the human auditory system. As the human auditory system simulates a log-spaced filter-bank, with the sensitivity to the energy in each channel following a logarithmic relationship, most feature extraction schemes (examples are Mel cepstra and PLP cepstra) are based on mimicing these steps. However, though these feature extraction procedures are perceptually motivated, they do not explicitly take into account the aspect of discriminating between the different phonetic classes. As the ultimate goal of speech recognition systems is to identify the phonetic classes corresponding to a sequence of feature vectors, one possibility is to design the feature vector in such a way that the classes can be unambiguously predicted by observing the feature vector. In other words, the feature vector should have the power to discriminate between different classes, and our hope is that this will lead to better speech recognition accuracy.

One question that arises at this stage is how to quantify the discriminative power of a particular feature extraction scheme, and use it to select a scheme that gives the best speech recognition accuracy. A natural choice for quantifying this discriminative power is the mutual information between the feature vector and the phonetic classes. This is complicated by the fact that the feature vectors we are dealing with are multi-dimensional continuous-valued quantities, consequently the computation of the mutual information is not trivial - in the remainder of the paper, we will describe two different schemes to approximate this computation, and how to use it to select or design a feature extraction scheme.

Though the cepstral feature extraction procedure mentioned earlier is primarily perceptually motivated, these feature vectors still do a remarkably good job of discriminating between the classes, and several speech recognition systems use this feature extraction scheme. Consequently, in our search for an optimal feature extraction procedure, we will assume the Mel cepstra as our "baseline" feature vector, and try to improve on top of it.

We first describe experiments where the baseline Mel cepstra are augmented by additional quantities and examine the amount of incremental information that is added by these quantities. The quantities that we experimented with correspond to the location and energy of the spectral peaks in the speech signal [1]. This is motivated by the fact that spectral peak trajectories clearly provide valuable clues to identifying the underlying phonetic classes, and this information is not directly captured by the Mel cepstra. In order to compute these spectral peak quantities, we also propose a method based on bandpass filtering the speech signal using several filters with different pass-

bands (corresponding approximately to formant ranges), and using an adaptive IIR filter to track the locations of the spectral peaks in each bandpass output. This method has the advantage that along with the estimate of the spectral peak frequency, it also provides the energy at the spectral peaks (a feature that turns out to be quite informative). Experimental results are then presented that quantify the additional information that is provided by these spectral peak quantities and their effect on the recognition accuracy.

We next provide an alternative formulation for computing the mutual information between the acoustic feature vector and the phonetic classes, subject to making certain assumptions about the probability density function for the different phonetic classes, and generalize this formulation to incorporate an arbitrary linear transformation of the underlying feature vector. This formulation then provides a framework wherein the arbitrary linear transformation can be optimized in order to maximize the mutual information. The feature extraction scheme now comprises of computing the baseline acoustic feature vector and applying a linear transformation on it. Experimental results are provided that quantify both the mutual information between the transformed feature vector and the phonetic classes and the recognition accuracy corresponding to these feature vectors.

A. Speech Recognition Experiments

A.1 Speech Recognition System Overview

The speech recognition system uses a phonetic representation of the words in the vocabulary. Each phone is modelled with a 3-state left-to-right HMM. Further, we identify the variants of each state that are acoustically dissimilar by asking questions about the phonetic context in which the state occurs. The questions are arranged heirarchically in the form of a decision tree, and its leaves correspond to the basic acoustic units that we model. An acoustic feature vector is extracted every 10 ms, and we model the pdf of the feature vector for each leaf of the decision tree, with a mixture of gaussians (our system had 2063 leaves and 68.5k gaussians). The maximum-likelihood decoder hypothesises a sequence of words from the vocabulary and computes the probablity of this word sequence using a language model (LM) and an acoustic model (AM). The LM essentially models the probablity of a sequence words without considering the acoustic observations, and the AM uses the pdf's mentioned earlier to compute the probablity of an observed acoustic feature vector given a word sequence. An envelope search procedure [2] is used to limit the number of hypotheses that the decoder has to consider. The feature extraction procedure extracts Mel cepstra per frame and augments them with their first and second temporal derivatives. Subsequently, this baseline feature vector is augmented with additional spectral peak related features in the experiments in Section II The augmented features are further transformed by a linear transform in the experiments in Section III

A.2 Task Description

The speech recognition task that we experimented on is a Voicemail transcription task [7], which represents *large vocabulary spontaneous telephone speech*, and is consequently quite a difficult task. The size of the vocabulary is 11k words, and the perplexity of the trigram LM that was used is approximately 100. The amount of acoustic training data available is approximately 53 hours (6501 voicemail messages). We present results on a combined test set comprising of 105 voicemail messages (approximately 1 hour of speech) [1].

The remainder of the paper is organized as follows: Section II describes the procedure used to compute the spectral peak quantities, the mutual information computation associated with these new acoustic feature vectors, and the results of speech recognition experiments with these new features. Section. III describes an alternative formulation to compute the mutual information between the feature vectors and the phonetic classes and formulates the problem of computing a linear transform to maximize mutual information, and describes the speech recognition results associated with the new feature vectors. Section. IV presents the concluding remarks.

II. Spectral Peak Features

Though Mel cepstral features capture information related to the local spectral characteristics of the speech signal, they do not explicitly capture information about the trajectory of spectral peaks. However, it is well known that these trajectories contain the information that can help identify the underlying phonetic class. As these spectral peaks generally correspond to the formant frequencies in the speech production mechanism, one possible method of finding them is to use a formant tracker such as Xwaves. However, the limitation with this approach is that the formant frequencies are only defined for voiced sounds; consequently, the estimates provided for unvoiced regions of speech are more or less random. Further, we are interested in computing features that are not just related to the location of the spectral peaks but also to the energy.

A. Estimation of Spectral Peaks

A simple solution to computing these features is to use an adaptive filter that tracks the spectral peaks in the input signal [2]. This method provides both the estimate of

[1] A portion of this data is available publicly through the LDC in the form of the Voicemail Corpus Part I. This comprises 1806 voicemail messages in the training set and 42 messages from the above test set. We expect to be able to release an additional corpus in the near future which will comprise of the remainder of the training data used in these experiments (4695 voicemail messages) and an additional 50 messages from the evaluation test set.

[2] The general idea in using adaptive filters for spectral estimation/tracking is to design a notch filter whose notch frequency is parametrized by some of the filter coefficients, and then adapt the filter coeffcients (using some variant of gradient descent) to minimize the power of the notch filter output. The filter may be either finite impulse response (FIR) or infinite impulse response (IIR). The advantage with the former is that the filter poles are always inside the unit circle, however the disadvantage is that the spectral estimates provided by the filter tend to be biased. IIR adaptive filters are both

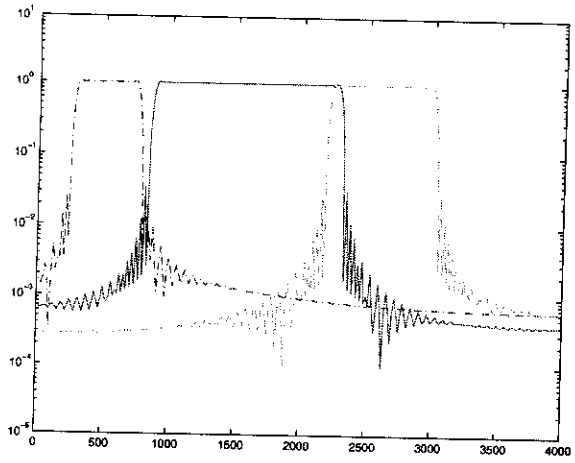

Fig. 1. Frequency response of bandpass filters

the spectral peaks as well as the energy at these peaks. For the voiced regions, this correlates roughly with the formant frequencies and energies, however, for the unvoiced regions, the adaptive filter essentially converges to the location of the spectral peaks in these regions. Further, as the adaptive filter is not allowed to change very abruptly, it also enforces a relatively smooth transition in the spectral peak locations over time.

In order to simplify the task of the adaptive filter, we first isolate spectral regions of the input signal by passing it through a bank of bandpass filters, such that each region contains only one dominant spectral component. In initial experiments, we based the choice of bandpass filters on physio-acoustical studies that indicate that the spectral peaks correspond to formant frequencies, and further, the first three formant frequencies lie in the range 280-710 Hz, 870-2250 Hz, 2250-2890 Hz [3]. Consequently, the speech signal is first filtered using a bank of three bandpass filters, with the passbands corresponding to these formant ranges, and subsequently, an adaptive IIR filter [4] is used to track the frequency at which the spectral energy is maximum within each passband. The frequency response of these bandpass filters is shown in Fig 1 - they are linear phase filters with all three filters having the same group delay.

A.1 Adaptive Filter Structure

Denoting the outputs of the three bandpass filters as $y_1(t)$, $y_2(t)$ and $y_3(t)$, the adaptive filter stage identifies the spectral peak in the bandlimited spectra of $y_i(t)$. This adaptive filter structure has previously been described in [4]. The basic idea is as follows: the adaptive filter is a multiple notch IIR filter with the notch frequencies being directly related to the filter coefficients (each notch frequency depends on exactly one filter coefficient). This multiple

computationally more efficient and able to reduce the bias in the estimate; however, they have the disadvantage that during the process of adaptation the filter poles may move outside the unit circle making the structure unstable. However, for certain specific filter structures, the stability issue can be solved by imposing very simple conditions on the filter coefficients. In our implementation, we use such an IIR filter structure [4].

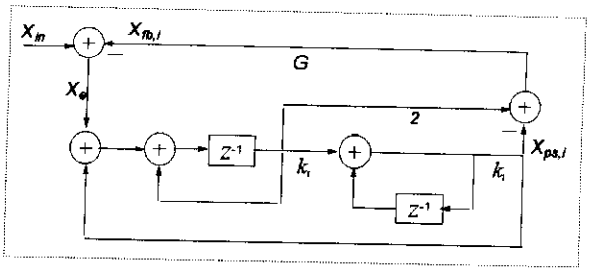

Fig. 2. Adaptive filter structure

notch transfer function is obtained by embedding several digital resonators in a feedback loop, with the notch frequencies corresponding to the resonator frequencies. The goal of the adaptive algorithm is to minimize the power of the output of the notch filter. For the case where there are N sinusoids in the input, and there are N notches in the filter transfer function, the optimal solution is when the notch frequencies are equal to the input sinusoidal frequencies. In [4] an adaptive algorithm was described that guarantees converegence under certain conditions, and has complexity that is linear in N.

For our case, we assume that each bandpass output, $y_i(t)$ has a single spectral peak and use an adaptive filter with a single notch to track each of the $y_i(t)$. This filter structure is shown in Fig 2. The transfer function from the input to various nodes in the filter structure are given below:

$$H_e(z) = \frac{x_e}{x_{in}} = \frac{1-(2-k_i^2)z^{-1}+z^{-2}}{1-(2-k_i^2)(1-G)z^{-1}+(1-2G)z^{-2}} \quad (1)$$

$$H_{fb,i}(z) = \frac{x_{fb,i}}{x_{in}} = \frac{(2-k_i^2)z^{-1}-2z^{-2}}{1-(2-k_i^2)(1-G)z^{-1}+(1-2G)z^{-2}} \quad (2)$$

$$H_{ps,i}(z) = \frac{x_{ps,i}}{x_{in}} = \frac{k_i^2 z^{-1}}{1-(2-k_i^2)(1-G)z^{-1}+(1-2G)z^{-2}} \quad (3)$$

The transfer function $H_e(z)$ represents a notch filter, with the notch frequency ω_i being related to the filter coefficient, k_i, through the following equation

$$k_i = 2\sin\left(\frac{\omega_i}{2}\right) \quad (4)$$

The algorithm for adapting the filter coefficient is given by

$$k_i(n+1) = k_i(n) - \mu \frac{x_e(n)x_{ps,i}(n)}{< x_{ps,i}(n)x_{ps,i}(n) > +\epsilon} \quad (5)$$

The term $x_e(n)x_{ps,i}(n)$ denotes the pseudo-gradient of the objective funtion (the coefficient is adapted in a direction opposite the pseudo-gradient), and the $< x_{ps,i}(n)x_{ps,i}(n) > +\epsilon$ denotes a power normalizing term that modfies the gradient direction to point in the Newton direction (i.e. the normalizing term approximates the inverse of the Hessian of the objective function). The spectral peak location can be inferred from the value of the filter coefficient k_i after it has converged using (4).

An additional feature of the filter structure is that the transfer function $H_{fb,i}$ is the complement of the notch transfer function, i.e., it represents a bandpass transfer function with the center frequency corresponding to the resonator frequency. Consequently, the power of the signal

at $x_{fb,i}$ represents the power of the input speech signal at this frequency. In the Appendix, we provide comparisons of the spectral estimates provided by the adaptive file with the formant estimates provided by Xwaves.

B. *Mutual information between acoustic feature and phonetic class*

The usefulness of an acoustic feature may be measured by the amount of information it provides in discriminating between phonetic classes. This can be quantified by the mutual information between the feature vector and the phonetic class. Let c denote the phonetic class, and Z denote the acoustic feature vector. The mutual information between Z and c is defined by

$$I(Z;c) = \sum_c P_c \int_Z p(Z/c) \log \left[\frac{p(Z/c)}{p(Z)} \right] dZ \qquad (6)$$

where P_c denotes the prior probability of class c. Though (6) cannot be expressed in closed form, by vector quantizing Z and approximating the integral with a summation, it may be rewritten as

$$I(Z;c) = \sum_c P_c \sum_{Z_j} p(Z_j/c) \log \left[\frac{p(Z_j/c)}{p(Z_j)} \right] \qquad (7)$$

The baseline cepstral features carry a certain amount of information and we are essentially interested in measuring the amount of incremental information provided by augmenting the baseline cepstral feature vector with the new spectral peak features. If Z represents the cepstral feature vector, and z represents one of the new features, then an augmented feature vector $\hat{Z}$ can be formed by concatenating Z and z. By vector quantizing $\hat{Z}$ (such that the number of codewords is the same as for the codebook for Z), the mutual information between the augmented vector $\hat{Z}$ and c, $I(\hat{Z};c)$, can be computed from (8)

$$I(\hat{Z};c) = \sum_c P_c \sum_{\hat{Z}_j} p(\hat{Z}_j/c) \log \left[\frac{p(\hat{Z}_j/c)}{p(\hat{Z}_j)} \right] \qquad (8)$$

The amount of information added by z can now be computed very simply as the difference between the information contained in the two feature sets

$$\delta I(\hat{Z}, Z) = I(\hat{Z};c) - I(Z;c) \qquad (9)$$

B.1 The amount of incremental information in spectral peak features

In this subsection, we examine the amount of information added by the spectral peak features. The baseline feature vector Z_D represents d-dimensional Mel cepstra and its first and second temporal derivatives. To this Z_D, we added one of the following two pairs of quantities: (i) the spectral peak locations, $z = [f_1 f_2]$, and their first and second temporal derivatives, (ii) the spectral peak energies $z = [e_1 e_2]$, and their first and second temporal derivatives. We will use d to denote the dimensionality of the cepstra and D to denote the dimensionality of the final feature and Z_D^1 and Z_D^2 to refer to the two augmented feature sets

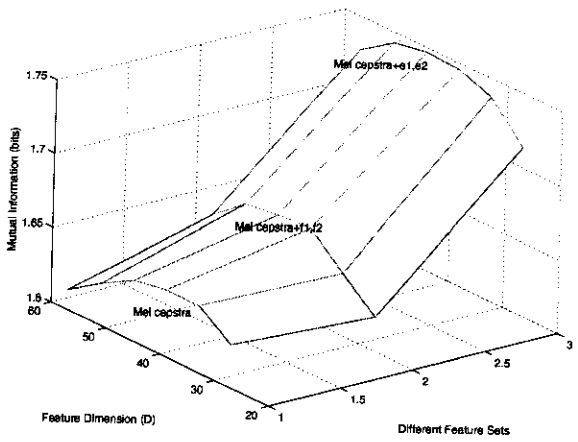

Fig. 3. Mutual information between features and phonetic classes

above. Hence, $D = 3d$ for the baseline cepstral system, and $D = 3d+6$ for each of the two augmented feature sets. We evaluated the $I(\hat{Z}, c)$ associated with these three feature sets for values of D varying from 27 to 57. The classes, c corresponded to 58 phonetic classes, and the number of feature vectors per class was limited to 200000. The feature vectors for each class were K-means clustered into 10 diagonal gaussians, leading to $\approx$ 570 gaussians in total. Next, a likelihood metric was used to label each feature vector with the label of the closest gaussian (from the set of 570 gaussians). Subsequently, (8) was used to compute the mutual information between the class and the quantized feature vectors.

The unconditional class distribution had an entropy of 5.6 bits, and the mutual information associated with 13-dimensional Mel cepstra ($Z_{39}, d = 13, D = 39$) was $I(Z_{39};c) = 1.63$ bits. The mutual information, $I(Z;c)$ (associated with the cepstra), and $I(\hat{Z};c)$ (associated with each of the two augmented features) are plotted in Fig 3 for various values of D.

Several observations may be made from Fig 3:
(i) the mutual information appears to be a concave function of the feature dimension i.e. it is maximum at a feature dimension, D, of 39 for the Mel cepstra and falls off as the dimensionality either increases or decreases. This indicates that the higher order cepstra after a certain point contain less *incremental* information than the lower order cepstra.
(ii) both the peak frequency and peak energy estimates appear to add to the information provided by the cepstra; however, the most informative features appear to be the energy at the peak frequencies, e_1, e_2. Further, the amount of additional information provided by these new features is around 0.1 bits, which is not negligible when compared to the 1.63 bits of information that are provided by the cepstra.

C. *Speech Recognition Experiments*

In this subsection, we describe the results of speech recognition experiments that used the three feature sets.

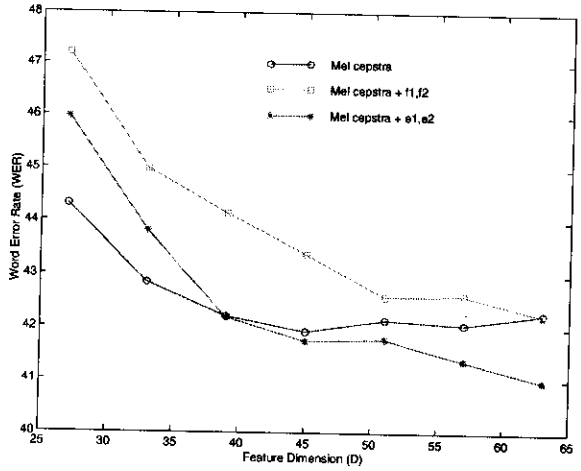

Fig. 4. WER vs feature dimension

The decision trees and number of gaussians was kept fixed for all systems, however the parameters of the gaussians were re-estimated in the appropriate feature space to construct the acoustic models for the system. We experimented with varying the dimension of the feature sets by varying the dimensionality of the underlying cepstra, and the results are shown in Fig. 4. The x-axis indicates the dimensionality of the feature vector, and the y-axis indcates the word error rate (WER) for the three feature sets.

It may be seen that the systems that use e_1, e_2 start improving over the corresponding cepstral baseline as the dimensionality improves. The increase in dimensionality is achieved by adding either two additional cepstra, or e_1, e_2 or f_1, f_2 to an existing baseline system. For a low dimensionality baseline, the additional cepstra carry a good deal of additional information, consequently the WER improves by increasing the dimensionality of the baseline. However, as the dimensionality of the baseline increases, the additional cepstra are dominated by noise - consequently, there is not much information to be gained by adding higher order cepstra to the baseline. On the other hand, the spectral peak features e_1, e_2 contain complementary information and are thus able to help the performance further.

The WER of the best system (Z^2_{63})is 40.98% compared to a WER of 42.21% for a 39-dimensional cepstral baseline (Z_{39}), or a WER of 42.27% for a 63-dimensional cepstral baseline, which has the same number of parameters. This represents an absolute improvement of 1.2% (3% relative) in the WER.

III. Linear Transformations to maximize the Mutual Information

Earlier, we defined the mutual information between te feature vector and the phonetic classes as

$$I(Z;c) = \sum_c p(c) \int_Z p(Z/c) \log \left[\frac{p(Z/c)}{p(Z)}\right] dZ \qquad (10)$$

however, the mutual information can also be expressed as

$$I(Z;c) = H(Z) - H(Z/c) \qquad (11)$$

If we make the following assumptions:

Assumptions

(i) the c^{th} class is modeled by a single full covariance gaussian, in the Z space, the mean and covariance of this gaussian is given by μ_c, Σ_c, i.e.,

$$p(Z/c) = \frac{1}{(2\pi|\Sigma_c|)^{D/2}} e^{-\frac{1}{2}[Z-\mu_c]^T \Sigma_c^{-1} [Z-\mu_c]} \qquad (12)$$

(ii) rather than model the complete data as $p(Z) = \sum_c P_c\, p(Z/c)$, the complete data is modeled with a single full covariance gaussian with mean and covariance μ, Σ

$$p(Z) = \frac{1}{(2\pi|\Sigma|)^{D/2}} e^{-\frac{1}{2}[Z-\mu]^T \Sigma^{-1} [Z-\mu]} \qquad (13)$$

the quantities $H(Z)$ and $H(Z/c)$ are given by

$$\begin{aligned} H(z) &= -\int_Z p(Z) \log p(Z)\, dZ \\ &= \frac{D}{2} log(2\pi) + \frac{1}{2} log|\Sigma| + D \qquad (14) \\ H(Z/c) &= \sum_c P_c \left[\frac{D}{2} log(2\pi) + \frac{1}{2} log|\Sigma_c| + D\right] \qquad (15) \end{aligned}$$

Consequently the expression for $I(Z;c)$ can be written as

$$I(Z;c) = log|\Sigma| - \sum_c P_c\, log|\Sigma_c| \qquad (16)$$

and (16) can be used to select the most informative feature space from a discrete set of possible feature sets, Z.

Let us now consider the case where Z is linearly projected into the feature vector X, i.e.

$$X = A\, Z \qquad (17)$$

where the dimensionality of X is D' and A is a $D'xD$ matrix. Now the models for the complete data and the different classes in the X space are also gaussians, with their means and covariances being given by $A\mu, A\Sigma A^T$ for the complete data, and $A\mu_c, A\Sigma_c A^T$ for the c^{th} class. Consequently, the mutual information between the projected feature and the classes is given by

$$\begin{aligned} I(X;c) &= H(X) - H(X/c) \\ &= log|A\, \Sigma\, A^T| - \sum_c P_c\, log|A\, \Sigma_c\, A^T| \qquad (18) \end{aligned}$$

We can make several observations from (18):
(i) if A represents a full rank transformation rather than a projection, then $I(X;c) = I(Z;c)$ and the mutual information is independent of the transformation A
(ii) if A represents a projection where the dimensionality of X is less than that of Z, then the linear projection A may be computed in order to maximize $I(Z;c)$. This objective function is almost exactly the objective function for heteroschedastic discriminant analysis (HDA) as described in [8]. *Consequently, we can make the claim that HDA is*

actually a projection that maximizes the mutual information between the projected feature and the phonetic classes (subject to the assumptions made earlier). Similarly, if all the classes are assumed to have the same average within class covariance, this formulation leads to linear discriminant analysis (LDA).

A. Bringing in the assumption of diagonal covariances

One of the constraints that most speech recognition systems operate under is the assumption of diagonal covariance gaussians to model the probablity density function for the different phonetic classes. This constraint is primarily imposed due to computational considerations, and can play a significant role in determining the system performance. For instance, if optimizing the objective function in (18) yields a feature space where the dimensions of X are heavily correlated, then the inaccuracies associated with modeling this feature space with diagonal gaussians will nullify any benefit that we obtained in applying the transformation A in the first place. Consequently, it is necessary to take the diagonal gaussian assumption into account when designing the linear transformation, in order to see the benefit associated with designing the transformation to maximize the mutual information.

Let $\Sigma_{c,diag,X}$ denote a diagonal matrix formed by taking the diagonal entries of $A\Sigma_c A^T$. For the case where the pdf of the classes are assumed to be full covariance gaussians, the mutual information between the projected feature and the classes is given by (18). For the case where the pdf of the classes is assumed to be a diagonal covariance gaussian, we obtain

$$H(Z/c) = \sum_c P_c \left[\frac{D'}{2} log(2\pi) + \frac{1}{2} log|\Sigma_{c,diag,X}| + D \right]$$
$$= \sum_c P_c \left[\frac{D'}{2} log(2\pi) + \frac{1}{2} \sum_{i=1}^{D'} \left[\log(a_i \Sigma_c a_i^T) \right] + D' \right] \quad (19)$$

where a_i is the i^{th} row of A. Consequently, the mutual information between the projected feature and the classes, under the assumption of diagonal gaussian pdf's for the classes is

$$I_{diag}(X;c) = log|A\ \Sigma\ A^T| - \sum_c P_c \sum_{i=1}^{D'} log\left[a_i\ \Sigma_c\ a_i^T\right] \quad (20)$$

A.1 Maximizing mutual information under diagonal assumption

The objective function to be optimized under the diagonal gaussian assumption is now given by (20)

$$f(A) = \stackrel{argmax}{A}\ I_{diag}(X;c)$$
$$= \stackrel{argmin}{A} \sum_c P_c \left[\sum_{i=1}^{D} \log(a_i \Sigma_c a_i^T) \right] - \log(|A\Sigma A^T|) \quad (21)$$

The solution to this optimization problem may be obtained using iterative methods that pursue some form of gradient descent. From [9], the derivative of this objective function is given by

$$\frac{df(A)}{dA} = 2 \sum_c P_c \left[S_c\ A\ \Sigma_c\right] - \left[\left(A\Sigma A^T\right)^{-1}\right]^T A\Sigma^T \quad (22)$$

where S_c is a diagonal matrix whose i^{th} diagonal entry is given by $\frac{P_c}{a_i\ \Sigma_c\ a_i^T}$ [3].

A.2 Minimizing loss in mutual information due to diagonal assumption

Rather than maximize the mutual information under the assumption of a diagonal gaussian model as in Section III-A.1, we could also consider selecting the linear projection A such that the mutual information under the diagonal gaussian assumption is as close to the mutual information for the full covariance gaussian case. This is equivalent to minimizing the difference between $I(X;c)$ in (18) and $I_{diag}(X;c)$ in (20). Consequently, our new objective function becomes

$$f(A) = \stackrel{argmin}{A}\ I(X;c) - I_{diag}(X;c)$$
$$= \stackrel{argmin}{A} \sum_c P_c \left[\sum_{i=1}^{D} \log(a_i \Sigma_c a_i^T) - \log(|A\Sigma_c A^T|) \right] \quad (24)$$

It is interesting to note that this objective function is very similar to that proposed in [10], [11], [12], where a linear transformation of the feature space was computed in order to maximize the likelihood of the data, given diagonal gaussian models [4]. *Consequently, we can make the claim that the linear transformation computed in [10], [11], [12] is actually the transformation that minimizes the difference in mutual information between the transformed feature and the class assuming full-covariance-gaussian class models and diagonal-covariance-gaussian class models.*

The solution to this optimization problem may be obtained using iterative methods that pursue some form of gradient descent. From [9], the derivative of this objective function is given by

$$\frac{df(A)}{dA} = 2 \sum_c P_c \left[S_c\ A\ \Sigma_c - \left[\left(A\Sigma_c A^T\right)^{-1}\right]^T A\Sigma_c^T \right] \quad (25)$$

where S_c is a diagonal matrix whose i^{th} diagonal entry is given by $\frac{P_c}{a_i\ \Sigma_c\ a_i^T}$ [5].

[3] For the case where A is assumed to be a transformation rather than a projection, this reduces to

$$\frac{df(A)}{dA} = 2 \sum_c P_c \left[S_c\ A\ \Sigma_c - A^{-T}\right] \quad (23)$$

[4] This also follows from the fact that $I(X;c) = H(X) - H(X/c)$ and $H(X/c)$ is the loglikelihood of the training data. In computing $I(X;c) - I_{diag}(X;c)$, the $H(X)$ terms cancels out, consequently, minimizing the difference between $I(X;c)$ and $I_{diag}(X;c)$ is equivalent to minimizing the difference in loglikelihood of the training data under the assumption of full-covariance or diagonal-covariance gaussian class models.

[5] For the case where A is assumed to be a transformation rather

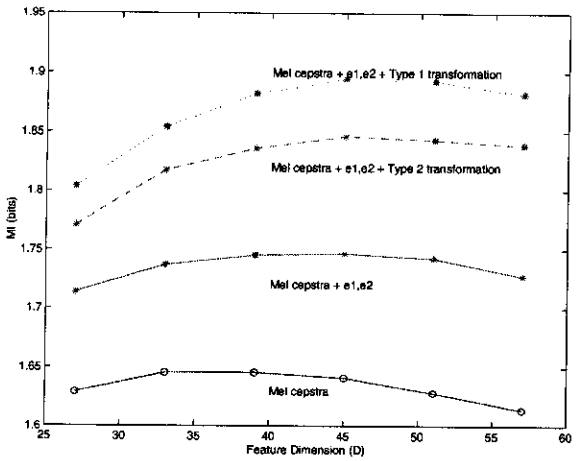

Fig. 5. Mutual Information vs feature dimension

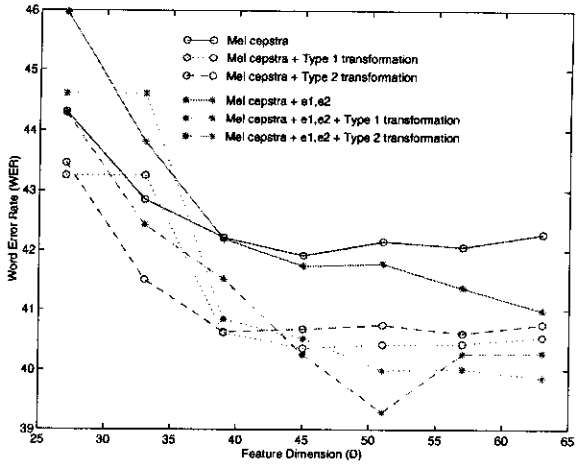

Fig. 6. WER vs feature dimension for several feature sets

D	Z_D	Z_D^2	$X_D^{'}$	$X_D^{2'}$	$X_D^{''}$	$X_D^{2''}$
27	44.31	45.97	43.25	44.61	43.46	44.29
33	42.85	43.81	41.64	42.24	41.51	42.42
39	42.21	42.19	40.62	40.84	40.63	41.52
45	41.92	41.74	40.36	40.52	40.68	40.25
51	42.15	41.78	40.42	39.99	40.75	39.30
57	42.06	41.37	40.43	40.01	40.61	40.26
63	42.27	40.98	40.54	39.86	40.76	40.28

TABLE I
WER VS FEATURE DIMENSION

A.3 Mutual information in the transformed feature space

To verify the ideas presented above, we computed a full rank linear transformation for the feature set Z_D (cepstra) and Z_D^2 (cepstra+e_1, e_2) to maximize the objective functions (21) and (24). We will refer to these transformations as Type 1 ($A^{'}$ that optimizes (21) and Type 2 ($A^{''}$ that optimizes (24), and the associated feature spaces as $X_D^{'} = A^{'} Z_D$, $X_D^{2'} = A^{2'} Z_D^2$, $X_D^{''} = A^{''} Z_D$, $X_D^{2''} = A^{2''} Z_D^2$. The mutual information between these sets of features and the phonetic classes was computed as in Section. II-B, and shown in Fig 5. The figure shows that both transformations consistently increase the mutual information.

B. Speech Recognition Experiments

In this subsection, we describe the results of speech recognition experiments that use the feature sets,
Cepstra: Z_D, $X_D^{'} = A^{'} Z_D$, $X_D^{''} = A^{''} Z_D$
Cepstra+e1,e2: Z_D^2, $X_{D'}^2 = A^{2'} Z_D^2$, $X_{D''}^2 = A^{2''} Z_D^2$
The results are shown in Fig. 6, and tabulated in Table I. Based on the results, we can make the following observations:
(i) both the Type 1 and Type 2 transformations consistently improve performance over the corresponding baseline
(ii) the transformed features for the "cepstra+e_1, e_2" systems are better than the corresponding transformed features for the "cepstra" baseline for feature dimensionalities greater than a threshold. This threshold varies from 45-51 depending on the system, indicating the utility of the spectral-peak related features over the higher order cepstra (which are known to contain less information)
(iii) the WER of the best system ($X_{51}^{2''}$) is 39.30% compared to a WER of 42.21% for the 39-dimensional cepstral baseline (Z_{39}), and a WER of 42.15% for a 51-dimensional cepstral baseline with the same number of parameters. This represents an absolute improvement of 2.9% (relative 6.8%) in the WER.

than a projection, this reduces to

$$\frac{df(A)}{dA} = 2 \sum_c P_c \left[S_c \, A \, \Sigma_c - A^{-T} \right] \qquad (26)$$

IV. CONCLUSION

In this paper we consider the problem of extracting feature vectors that contain the maximum amount of information about the predicted class. We first explore the connection between the information content in the feature vector (quantified by the mutual information between the feature vector and the predicted classes) and the classification accuracy. The connection is first explored in the context of adding features to a baseline feature extraction scheme - and using the mutual information between the augmented feature and the predicted class to select the final feature extraction scheme. This problem was considered in the context of a speech recognition system, where the baseline feature extraction scheme corresponds to the commonly used Mel cepstra. To this baseline feature, we added other features related to the spectral location and energy value of the spectral peaks in the speech signal (similar to formant frequencies). The new features were estimated by passing the speech signal through a bank of bandpass filters, and using an adaptive filter to track the location of the spectral peak in each band. We show that augmenting the baseline feature with these features can help increase the mutual information between the augmented feature and the predicted classes, which translates into an improvement in the word error rate. Subsequently, the idea of optimiz-

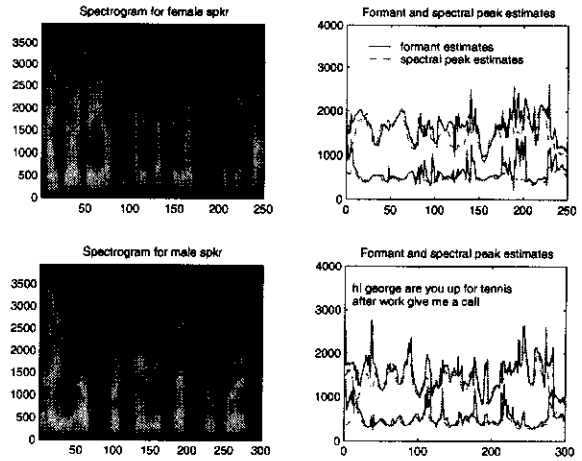

Fig. 7. Spectrogram and formant frequency estimate of speech for one sentence with male and female speakers

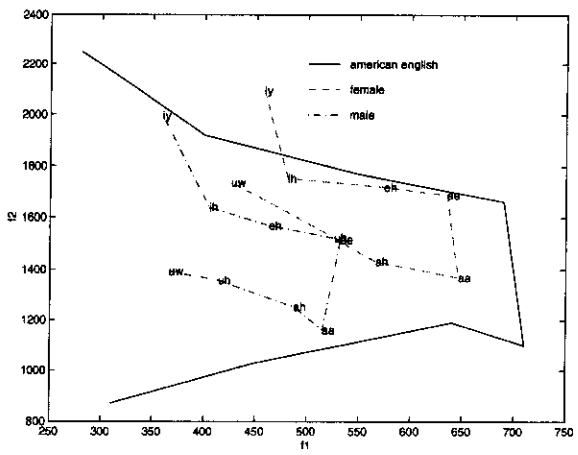

Fig. 8. Location of vowels for male, female and average speakers

ing mutual information to improve recognition accuracy is generalized to a linear transformation of the underlying features. We show that several prior methods to compute linear transformations can be interepreted in this general framework of maximizing the mutual information. We also report the results of experiments that show that further improvements in the word error rate can be achieved by applying a mutual information maximizing linear transformation on the augmented feature. By the use of these techniques, we show an overall reduction of 2.9% (6.8% relative) in the word error rate (WER), with respect to a 39-dimensional cepstral baseline system.

V. Appendix

A. Comparison to Xwaves Formant Estimates

The bandpass filtering that preceded the spectral peak tracking was motivated by trying to isolate different formants, consequently, it would be interesting to examine the correlation between the spectral peaks and the formant frequencies for vowels which have a well defined formant structure. However, the true formant estimates are not available for any of the data that we experimented with, consequently, we used Xwaves [6] to estimate the locations of the formants from the speech signal. In Fig 7, we show the spectrogram of the speech for one sentence from a male and female speaker, as well as the first two formant estimates obtained from Xwaves (x_1, x_2), and the first two spectral peaks obtained from the adaptive filter (s_1, s_2). It may be seen that both the formant estimates provided by Xwaves and the spectral peaks tracked by the adaptive filter look reasonable, To validate the formant estimates that we obtained from Xwaves, we computed the average f_1 and f_2 frequencies for several vowels for one male and one female speaker, and plot them in Fig 8. Also plotted in the same figure are the average vowel positions for average American English, as specified in [3].

References

[1] M. Padmanabhan, "Spectral Peak Tracking and its use in Speech Recognition", ICSLP 2000.
[2] P. S. Gopalakrishnan, L. R. Bahl and R. L. Mercer, "A Tree Search Strategy for Large Vocabulary Continuous Speech Recognition", Proceedings of ICASSP 1995.
[3] P. Ladefoged, "A Course in Phonetics", 1993, Harcourt Brace College Publishers, 301 Commerce Street, Suite 3700, Fort Worth, TX 76102.
[4] M. Padmanabhan and K. Martin, "Resonator-based filter-banks for frequency domain applications", IEEE Trans. Circuits and Systems, Oct 1991.
[5] S. M. Kay and S. L. Marple, "Spectral Analysis - A Modern Perspective", Proceedings of the IEEE, vol. 69, pp 1380-1419, 1981.
[6] Entropic Xwaves package, http://www.entropic.com
[7] M. Padmanabhan et al., "Speech Recognition Performance on a VoiceMail Transcription Task", Proceedings of the Eurospeech 1999.
[8] G. Saon, M. Padmanabhan, R. Gopinath, S. Chen, "Maximum Likelihood Discriminant Feature Spaces", ICASSP 2000.
[9] K. Fukunaga, "Introduction to Statistical Pattern Recognition", Academic Press, ISBN 0-12-269851-7.
[10] R. Gopinath, "Maximum Likelihood Linear Transformation", ICASSP 1998.
[11] M. J. F. Gales, "Semi-tied Covariance Matrices", ICASSP 1999.
[12] N. Kumar and A. G. Andreou, "Heteroschedastic discriminant analysis and reduced rank HMM's for improved speech recognition", Speech Communication, 26:283-297, 1998.