Using Broad Phonetic Group Experts for Improved Speech Recognition

Patricia Scanlon, Daniel P. W. Ellis, *Senior Member, IEEE*, and Richard B. Reilly, *Senior Member, IEEE*

***Abstract*—In phoneme recognition experiments, it was found that approximately 75% of misclassified frames were assigned labels within the same broad phonetic group (BPG). While the phoneme can be described as the smallest distinguishable unit of speech, phonemes within BPGs contain very similar characteristics and can be easily confused. However, different BPGs, such as vowels and stops, possess very different spectral and temporal characteristics. In order to accommodate the full range of phonemes, acoustic models of speech recognition systems calculate input features from all frequencies over a large temporal context window. A new phoneme classifier is proposed consisting of a modular arrangement of experts, with one expert assigned to each BPG and focused on discriminating between phonemes within that BPG. Due to the different temporal and spectral structure of each BPG, novel feature sets are extracted using mutual information, to select a relevant time-frequency (TF) feature set for each expert. To construct a phone recognition system, the output of each expert is combined with a baseline classifier under the guidance of a separate BPG detector. Considering phoneme recognition experiments using the TIMIT continuous speech corpus, the proposed architecture afforded significant error rate reductions up to 5% relative.**

***Index Terms*—Automatic speech recognition, broad phonetic groups (BPGs), mixture of experts, mutual information (MI).**

I. Introduction

THE fundamental task of the acoustic model in a speech recognizer is to estimate the correct subword or phonetic class label for each frame of the acoustic signal. The phoneme can be defined as the smallest phonetic unit in a language that is capable of conveying a distinction in meaning; however, phonemes that may be within the same broad phonetic group (BPG) contain very similar temporal characteristics and can be easily confused. In phoneme recognition experiments on the TIMIT database, reported in [5], it was observed that almost 80% of all misclassified frames are identified as phonemes within the same BPG as the correct target. The BPGs in these experiments were vowels, stops, weak fricatives, strong fricatives, and nasals.

Manuscript received August 2, 2005; revised April 17, 2006. This work was supported by Enterprise Ireland under the ATRP Program in Informatics, and by the Defense Advanced Research Projects Agency (DARPA) ubder the EARS Novel Approaches Program. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ananth Sankar.

P. Scanlon was with University College Dublin, Dublin 4, Ireland. She is now with Alcatel-Lucent, Bell Laboratories, Ireland, Dublin 15, Ireland (e-mail: scanlon@alcatel-lucent.com).

D.P.W. Ellis is with the Electrical Engineering Department, Columbia University, New York, NY 10027 USA (e-mail: dpwe@ee.columbia.edu).

R.B. Reilly is with the School of Electrical, Electronic and Mechanical Engineering, University College Dublin, Dublin 4, Ireland (e-mail: richard.reilly@ucd.ie).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2006.885907

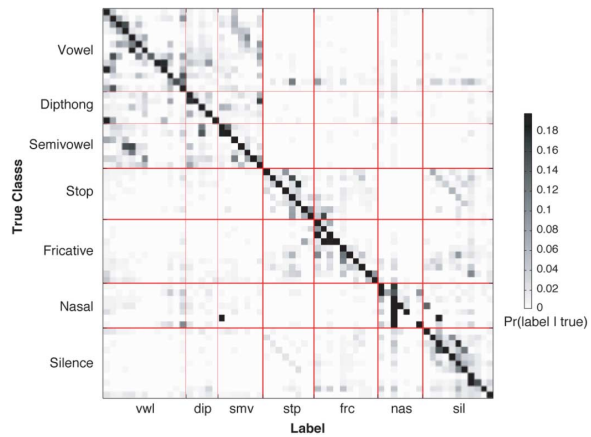


Fig. 1. Confusion matrix with phonemes grouped into BPGs. Rows are normalized to give conditional probabilities, and values larger than 20% (including most of the leading diagonal) are clipped to that level. Forty-eight percent of confusions fall into the same group, rising to 74% if vowels, dipthongs, and semivowels are merged into a single group.

Similar results to those reported in [5] are illustrated in Fig. 1, where almost 75% of misclassified frames were given labels within the same BPG. In Fig. 1, phonemes are divided into the BPGs of vowels, stops, fricatives, and nasals, where the vowel group contains all phonemes that may be labeled as vowels, semivowels, or dipthongs. Distinguishing between these three vowel-like groups, it is observed that almost 50% of confusions still lie within the same group as the true label. However, since the vowel-like sounds are especially confusable, they are placed in a single group. The confusion matrix of phonemes is given in Fig. 1, with the phonemes ordered in groups i.e., the first 25 are vowel or vowel-like phonemes, the next eight are stops, then ten fricatives, seven nasals, and finally 11 silence/pause/stop-closures.

The task of speech recognition is complicated by the fact that the information relevant to phoneme classification is spread out in both frequency and time—due to mechanical limits of vocal articulators, other coarticulation effects, and phonotactic constraints. As a result, it is generally advantageous to base classification on information from all frequencies and across a large temporal context window. This generalized feature window results in a large number of parameters as input to the classifier, and hence requires very large training sets, as well as frustrating the classifier training with redundant and irrelevant information. Using such a large, general-purpose feature space can lead to confusion between phonemes of the same BPG as seen in Fig. 1.

In this paper, a new modular architecture for speech recognition is proposed in which an expert is assigned to each BPG.

1558-7916/$25.00 © 2007 IEEE

Roku Exhibit 1017
Roku Inc. v. Universal Elecs. Inc.
IPR2022-01289
Page 00001

These experts focus discrimination capabilities of the classifier on the sometimes subtle differences between phonemes belonging to the same BPG, rather than between all phonemes in all BPGs. Since separate classifiers are used for each group, it is proposed that different feature sets be used for each expert that better support discrimination between the phonemes of that group. In this paper, mutual information (MI) is used as a basis for selecting particular cells in the TF plane to optimize the choice of features used as inputs to each BPG classifier.

MI-based feature selection for speech recognition has been investigated previously in the literature. Morris *et al.* [7] examined the distribution of information across a time-aligned auditory spectrogram for a corpus of vowel-plosive-vowel (VPV) utterances. The MI was estimated between each TF cell and the VPV labels, as was the joint mutual information (JMI) between pairs of TF cells and the VPV labels. The goal was to use MI and JMI to determine the distribution of vowel and plosive information in the TF plane. Features with high MI and JMI were used to train a classifier to recognize plosives in the VPV utterances. Bilmes [1] used the expectation-maximization (EM) algorithm to compute the MI between pairs of features in the TF plane. He verified these results in overlay plots and speech recognition word error rates. Yang *et al.* [13] used methods similar to [7] but their focus was on phone and speaker/channel classification. Multilayer perceptrons with one or two inputs were used to demonstrate the value for phone classification of individual TF cells with high MI and pairs with high JMI. In Scanlon *et al.* [10], in addition to calculating MI over all phonetic classes, the MI is examined for subsets formed by BPGs, such as the MI between specific vowel labels across only the vowel tokens, etc. The hypothesis that high MI features provide good discrimination was verified in [10] where a range of vowel classifiers are evaluated over the TIMIT test set and show that selecting input features according to the MI criteria can provide a significant increase in classification accuracy.

The work described in this paper extends this work by extracting the relevant feature sets for each BPG. Specifically, the use of MI as measure of the usefulness of individual TF cells for each of the BPGs has been investigated, using the phonetically labeled TIMIT continuous speech corpus as the ground truth.

Modular or hierarchically organized networks as opposed to monolithic networks have been studied extensively in the literature. The speech recognition task is divided among several smaller networks or experts and the output of these experts are combined in some hierarchical way yielding an overall output.

A hierarchical mixture of experts (HME) was applied to speech recognition in [14], where the principle of divide-and-conquer was used. The training data was divided into overlapping regions which are trained separately with experts. Gating networks are trained to choose the right expert for each input. In the HME architecture the combining process is done recursively. The outputs from the experts are blended by the gating networks and proceed up the tree to yield the final output. In HME the decomposition is data driven and each expert has the same feature set as input.

The Boosting algorithm constructs a composite classifier by iteratively training classifiers while placing greater emphasis on certain patterns. Specifically, hard-to-classify examples are given increasing dominance in the training of subsequent classifiers. The hybrid NN/HMM speech recognizer in [11] shows it is difficult to take advantage of very large speech corpora, and that adding more training data does necessarily improve performance. The AdaBoost algorithm can be used to improve performance by focusing training on the difficult and more informative examples. In this paper log RelAtive SpecTrAl Perceptual Linear Predictive (log-RASTA-PLP) features, modulation-spectrogram-based features, and the combination of these feature sets are compared. It was shown that Boosting achieves the same low error rates as these systems using only one feature representation.

Previous research into using BPG experts in a modular architecture has been carried out in [5], which also includes the idea of using different feature sets for each of the BPG experts. These feature sets were varied in dimension and in time resolution and empirical measures were employed to determine the best feature set for each expert. BPG feature sets varied greatly using different feature vector dimensions, resolution, and including a variation of other features such as duration and average pitch for vowel and semivowel classes, zero-crossing rate, total energy of the segment, and time derivative of the low-frequency energy for the fricative class. In [5], no variation of the network parameters was made for each of the BPG experts. A maximum *a posterior* (MAP) framework was used for overall phoneme classification. This framework combines posterior probabilities from all BPG experts outputs with the posterior probability of its group.

Another approach to modular architecture for speech recognition was investigated in [9]. This architecture decomposes the task of acoustic modeling by phone. In the first layer, one or more classifiers or primary detectors are trained to discriminate each phone, and in the second layer, the outputs from the first layer are combined into posterior probabilities by a subsequent classifier. It is shown that the primary detectors trained on different front-ends can be profitably combined due to independent information provided by different front-ends. As different feature sets have individual advantages and disadvantages, the use of different feature sets such as mel-frequency cepstral coefficients (MFCCs), PLP, and linear predictive coding (LPC) feature sets and combinations of these feature sets were compared. In these experiments, the feature set combination that maximized the entire system was used. Another primary detector was incorporated into the framework to detect the presence of BPGs over a large context window, to combine with previous outputs to further improve performance.

Chang *et al.* [3] proposed that a hierarchical classifier based on phonetic features, i.e., one classifier for manner, then a conditional classifier for place given manner (which together distinguish all consonants), could significantly outperform traditional nonhierarchical classification based on experiments using the assumption of perfect recognition of the conditioning manner class. However, recent work [8] disproves this proposal by implementing a similar system where the conditioning manner class is automatically detected and showed that gains suggested in [3] were minimized.

In Sivadas and Hermansky [12], a hierarchical approach to feature extraction is proposed under the tandem acoustic modeling framework. This was implemented as hierarchies of MLPs such as speech/silence, voiced/unvoiced, voiced classes, and unvoiced classes. The output from the hierarchy of MLPs was subsequently used as feature set in a Gaussian mixture modeling (GMM) recognizer after some nonlinear transformation.

It was observed that the hierarchical tandem system performed better than the monolithic-based classifier using context-dependent models for recognition and worse when context-independent models were used. It was suggested that a more structured approach to the design of the classification tree would improve performance.

Modular approaches to speech recognition in the literature typically extract homogeneous feature vectors to represent the acoustic information required to discriminate between all phones [3], [8], [9], [11], [14]. While the performance of different feature sets and combinations of these sets has been compared in [9] and [11], homogeneous feature vectors are used as input to the entire system. The use of heterogeneous feature sets for modular-based ASR system has also been expored. A heuristic approach is used in [5] where empirical results are used to chose the feature set for each BPG (or phone-class). These feature sets vary greatly in dimensionality, inclusion of temporal features and inclusion of other features such as zero-crossing rate, energy, and pitch. In [12], the output from a hierarchy of MLP networks is used as the feature input to a GMM-based speech recognizer. In this paper, the use of MI criterion is proposed to select the most relevant features based on speech class information. In this way, just one unique TF pattern per BPG is selected and discriminative classifiers are used to distinguish within that group.

Our proposed approach combines modular network of BPG experts with a scheme to select only features relevant to each expert. Using a development set the size on the expert network's input layer, number of hidden nodes is chosen to maximize the performance of the BPG experts. Our implementation of this architecture assigns each frame to a BPG or the silence group. Each candidate frame is assigned to one of the BPGs or a silence group. In order to easily incorporate the proposed modular architecture into our existing baseline framework, the output from the set of experts is combined or "patched" into the baseline monolithic classifier posterior estimates.

The remainder of this paper is organized as follows. Section II describes the basic approach of decomposing acoustic classification into a set of subtasks, and then Section III provides the background for MI, its computation, and the subtask-dependent feature selection algorithm. In Section IV, the proposed classifier architecture is described. Details of the baseline system and the BPG experts and the BPG detector and integration methods are given. Section V discusses the benefits of the proposed feature selection method and provides experimental demonstration of the architecture.

II. Mixtures of Experts

Central to the system presented is the idea of decomposing the phone classification problem into a number of subtasks (i.e., our within-BPG classification) and building expert classifiers specific to each of those domains. This ensemble of experts is used as a (partial) replacement for a single classifier deciding among the entire set of phones, but in order to make these alternatives directly interchangeable, it is necessary to decide how to combine each of the experts into a single decision.

Consider our basic classification problem of estimating, for each time frame, a phone label Q (which can take on one of a discrete set of labels $\{q_i\}$, based on an acoustic feature vector X. A monolithic classifier, such as a single MLP neural network, can be trained to make direct estimates of the posterior probability of each phone label $Pr(Q = q_i|X)$. If, however, a classifier is trained only to discriminate among the limited set of phones in a particular BPG, this new classifier is estimating posterior probabilities conditioned on the true BPG of the current frame C taking on a specific value (also drawn from a discrete set $\{c_j\}$). Thus, each expert classifier estimates $Pr(Q = q_i|C = c_j, X)$ for a different BPG class c_j. These can be combined into a full set of posteriors across all phones with

$$Pr(Q = q_i|X) = \sum_{c_j} Pr(Q = q_i|C = c_j, X)Pr(C = c_j|X) \tag{1}$$

i.e., as a weighted average of the experts, weighted by some estimate $P(C = c_j|X)$ of which expert is in fact best suited to the job—this process is called "patching in," since at different times the merged output stream consists of "patches" coming from different individual experts. The weights could constitute a "hard" selection (i.e., 1 for a particular c_j and 0 for all others), or they could be constants smaller than 1 (allowing some small proportion of different classifiers to come through at all times), or they could also be dynamic, varying in proportion to some kind of confidence estimate for the class estimation.

The BPG weights $Pr(C = c_j|X)$ need to be obtained somehow, most obviously through training a further classifier simply to identify the appropriate BPG. However, this expert-selection classifier will surely make some mistakes, and so the overall benefit of this two-stage classification (BPG, then phone given BPG) is a tension between the benefits of discrimination only within a narrow set of phones (as performed by the expert) and the degradation caused by imperfect estimation of BPG labels. Such systems can be "tuned" to be more conservative simply by making it less likely that a frame will be marked as relevant to one of the experts, assuming that the baseline classifier is used when none of the experts is selected, so that in the limit the system backs off to the simple baseline system.

With ideal classifiers, decomposing the problem this way should make no difference. However, since actual classifier performance is a complex function of classifier algorithms and available training data, the decomposition can have benefits. In particular, because each of the experts is looking at a distinct, homogeneous problem (discriminating phones within a single class), the "structural" discrimination of using different feature vectors for each expert can be incorporated, thereby reducing the number of parameters in the experts compared to the baseline classifier, and possibly improving their ability to exploit the finite training data. In Section III, how MI is used to select these distinct per-expert feature sets is discussed.

III. Mutual Information

A. Background

The entropy of a random variable is a measure of its unpredictability [4]. Specifically, if a variable X can take on one of a

set of discrete values $\{x_i\}$ with a probability $\Pr(X = x_i)$ then its entropy is given by

$$H(X) = -\sum_{x \in \{x_i\}} \Pr(X = x) \log \Pr(X = x). \quad (2)$$

If a second random variable C is observed, knowing its value will in general alter the distribution of possible values for X to a conditional distribution $p(x|C = c)$.

Because knowing the value of C can, on average, only reduce our uncertainty about X, the conditional entropy $H(X|C)$ is always less than or equal to the unconditional entropy $H(X)$. The difference between them is a measure of how much knowing C reduces our uncertainty about X, and is known as the MI between C and X

$$I(X;C) = H(X) - H(X|C) = H(C) - H(C|X). \quad (3)$$

Note that $I(X;C) = I(C;X)$; this symmetry emerges naturally from the expectations being taken over both variables, and leads to the intuitive result that the amount of information that C tells us about X is the same as the amount of information that knowing X would tell us about C. Further, $0 \leq I(X;C) \leq \min\{H(X), H(C)\}$, and $I(X;C) = 0$, if and only if X and C are independent.

B. Selection Algorithm and its Implementation

Putting aside for the moment the issue of computing (3), the MI-based algorithm for feature selection within the candidate pool of TF features can be expressed as

$$X_i = \underset{X \in \mathcal{X} \setminus \mathcal{X}_{i-1}}{\operatorname{argmax}} \{I(X;C)\} \text{ and } \mathcal{X}_i = \mathcal{X}_{i-1} \cup X_i \quad (4)$$

for $i = 1, 2, \ldots, d$, with $\mathcal{X}_o = \emptyset$, where d is the desired dimensionality of the selected feature vector. Note that this approach represents a simple sorting of all mutual information values and it results in a nested selected feature set $\mathcal{X}_1 \subset \cdots \subset \mathcal{X}_d \subset \mathcal{X}$. Note also, however, that this greedy strategy does not find the optimal set of d points since there may be information "overlap" between the successively chosen X points. In the worst case, two TF points that always had identical values would have equal $I(X;C)$ (and would thus be neighbors in the sorted list), but including the second would not add any additional information about C over that provided by the first.

To obtain estimates of the MI values needed in (4), the *histogram approach* was used to approximate the density functions required in (3), as in [13]. The histogram approach requires choosing the number of bins to be used and their bin widths. In order to exclude outliers (that can result in empty or sparsely filled bins), the range over which the histogram is computed, and hence the bin width, is determined by setting the lower bound equal to the mean of the samples minus three standard deviations; the maximum is similarly obtained.

Following [13], Doane's rule, $K = \log_2 n + 1 + \log_2(1 + \widehat{k}\sqrt{n/6})$ is used to determine the number of bins to estimate $p(X|C)$ and $p(X)$. In this rule, $\widehat{k}$ is the estimate of the kurtosis of the TF components (i.e., of random variable X), and n is the total number of training samples. In our experiments, $n \approx 10^5$, and, on the average, 30 bins are derived for each TF component. Note that the kurtosis estimates indicate that the TF components are non-Gaussian.

Given the number of bins, equally spaced intervals are formed b_k, $k = 1, 2, \ldots, K$, between the upper and lower bounds, as described above, computed for each X. Then $p(x) \approx n_k/n$, iff $x \in b_k$, is approximated where n_k denotes the number of observations $x \in b_k$. Assuming that class labels $c \in \{c_i\}$ are available for the training samples, the n_c and $n_{k,c}$ counts can similarly be obtained, thus estimating $p(c) = n_c/n$ and approximating $p(x|c) \approx n_{k,c}/n_c$, for all $x \in b_k$, $k = 1, 2, \ldots, K$, and $c \in \{c_i\}$.

Based on these estimates of the density functions, the computation of (3) becomes feasible.

TABLE I
PHONETIC BROAD CLASS GROUPS

Group	Phonemes
Vowels	iy ih eh ae aa ah ao uh uw ux ax ax-h ix
Dipthongs	ey aw ay oy ow
Semi-vowels	l el r w y er axr
Stops	b d g p t k jh ch
Fricatives	s sh z zh f th v dh hh hv
Nasals	m em n nx ng eng en
Silence	dx bcl dcl gcl pcl tcl kcl h pau epi q

C. Mutual Information for Broad Phonetic Groups

The phonemes are divided into phonetic broad classes as in Table I based on the distribution on confused phonemes in the confusability matrix in Fig. 1.

The MI was computed between the phonetic labels and the individual cells across the TF plane. The baseline features were perceptual linear predictive (PLP) cepstral coefficients [6] calculated over a 25-ms window with 10-ms advance between adjacent frames. For the TIMIT dataset, which is sampled at 16 kHz, 12-order PLP models were used.

Temporal regression features (or first derivative features) were computed over a context window of nine frames along with acceleration (or second derivative) features over the same window. These temporal features were appended to the feature vector, resulting in 39 PLP features. A temporal window of ± 15 frames around the labeled target frame (i.e., 31 time frames total) was used as the domain over which MI was computed. These features undergo a per-utterance mean and variance normalization prior to MI calculation providing a degree of invariance against variations in channel characteristic (microphone positioning etc.).

An MI plot consisting of 39×31 cells was calculated for each BPG. The MI calculation was performed for each individual time-quefrency cell, for PLP cepstra, against the phonetic labels within each BPG. An MI plot was generated for each of the groups as shown in Fig. 2. To take advantage of the MI plots, an MI feature selection mask is created by selecting N TF cells with the largest MI values. This results in an irregularly shaped pattern in the TF plane (MI-IRREG) consisting of all the cells with values above some threshold. The threshold was varied to extract different feature vector dimensionalities.

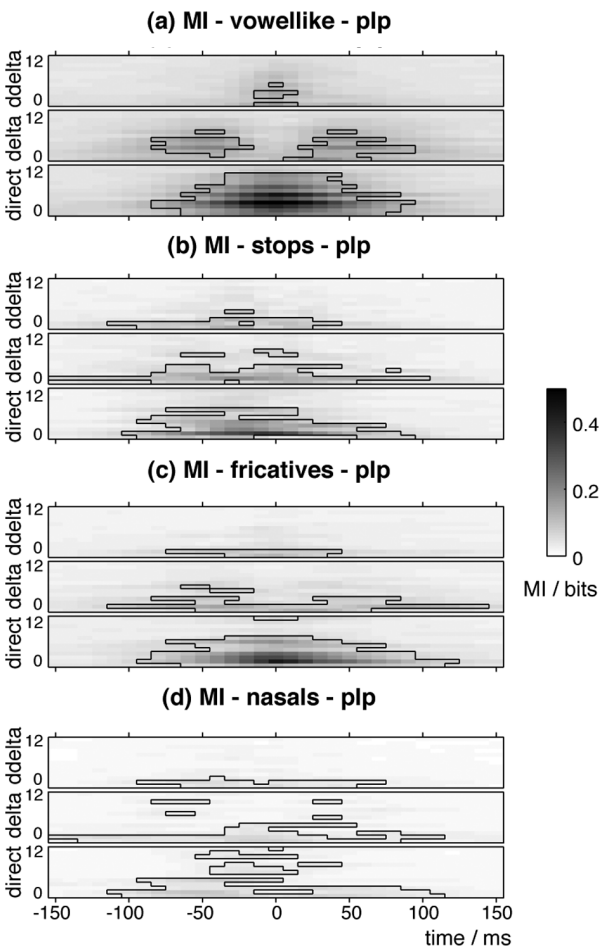


Fig. 2. Information distribution using MI for BPGs. (a) Vowels. (b) Stops. (c) Fricatives. (d) Nasals. The irregular outlines contain the top 200 cells in each case. Each block has three panes corresponding (from bottom to top) to static, and first and second derivatives, respectively. The PLP static features are 13 PLP coefficients.

As an example, Fig. 2 shows MI masks used to select 200 features as outlines. Standard feature vectors corresponding to rectangular regions in the TF plane (RECT) are also extracted in the experiments, where all spectral components, e.g., 13 PLP, plus first and second derivative features across a temporal window of nine successive 10-ms frames, are used.

It can be seen from Fig. 2 that the BPGs contain very different spectral and temporal characteristics.

It can be seen that information for discriminating between all the vowel-like phonemes is concentrated mainly in the static features. The information is spread out ± 50 ms and concentrated mainly in the third, fourth, sixth, and eighth coefficients. For Stops, information is spread out over the static and first derivative features. For the PLP features the most significant information exists in the second coefficient (spectral tilt) from -70 to 30 ms, with some less relevant information in the third, fourth, and fifth coefficients over a shorter time span. The MI between the TF cells and the fricative BPG phonemes is mainly concentrated in the static features. The greatest information exists in the second, third, and fourth coefficients from -30 to 50 ms. The nasal MI plots show only weak information, spread out over static and first and second derivative features. There appears to be a minimum of MI at the center of the window and information is concentrated in the second and fifth coefficients from -90 to -10 ms and in the first and third coefficients from 20 to 50 ms.

Due to the steady-state nature of vowels, most of the important information for discrimination between vowels exists in the static TF cells. Fricatives and nasals show an increasing trend of information shifting to the derivative features, with stops showing the greatest information in dynamic features. All this is consistent with our preconceptions concerning these BPGs.

Note that in the MI investigation above, the MI was computed for each TF cell in isolation and the relative MI for all cells is shown in Fig. 2. For steady-state phonemes such as vowel-like phonemes it is assumed that correlation is high along the time axis. This suggests that the immediate neighbors of a TF cell along the time axis may be omitted from the classifier without a significant loss of information. Therefore, conditional MI between the BPG phone labels and two feature variables in the TF plane was applied to measure the relevance of the feature cells before and after the current time frame.

Fig. 3 shows the MI between each cell on the TF plane and the phone label within each BPG (as before), additionally conditioned on the value of the TF cell centered on the labeling instant for that frequency band, i.e., the additional information provided by knowing a second cell's value. Thus, the values are zero for the 0-ms column, since this is the value already present in the conditioning. Note that the MI scale is much smaller compared to Fig. 2. Also note that each row of each spectrogram corresponds to a different experiment, since the conditioning value moves with the frequency band being measured. It can be seen in Fig. 3 that for the vowel BPG, the immediate neighboring features in time provide the lowest conditional MI with the current frame for all coefficients. However, for the fricative, stop, and nasal BPGs, the immediate neighboring coefficients in time do not always provide the lowest conditional MI.

To compute the conditional MI for more than two features, multivariate density estimation is required which is difficult to reliably obtain without an inordinate amount of data and computation time. Therefore, in order to approximate the N-way joint maximally informative set for the steady state vowel BPG, the selection masks are multiplied by a vertically striped pattern which reduces the inclusion of possibly redundant neighboring TF cells. An advantage of this method of "striping" the MI masks to reduce redundancy is that, for a given dimensionality, using the striped feature mask includes features spread out further in time when compared to nonstriped feature masks with the same dimensionality.

IV. Classifier Architecture

The proposed system first detects which BPG each frame belongs to. Once identified, the output for that frame is extracted from the corresponding BPG expert classifier and "patched" into the baseline classifier output to reduce the number of misclassifications that occur between phonemes within the same BPG. This system architecture is illustrated in Fig. 4. In this section, the implementation of the baseline classifier, the BPG experts, and the proposed modular architecture are described.

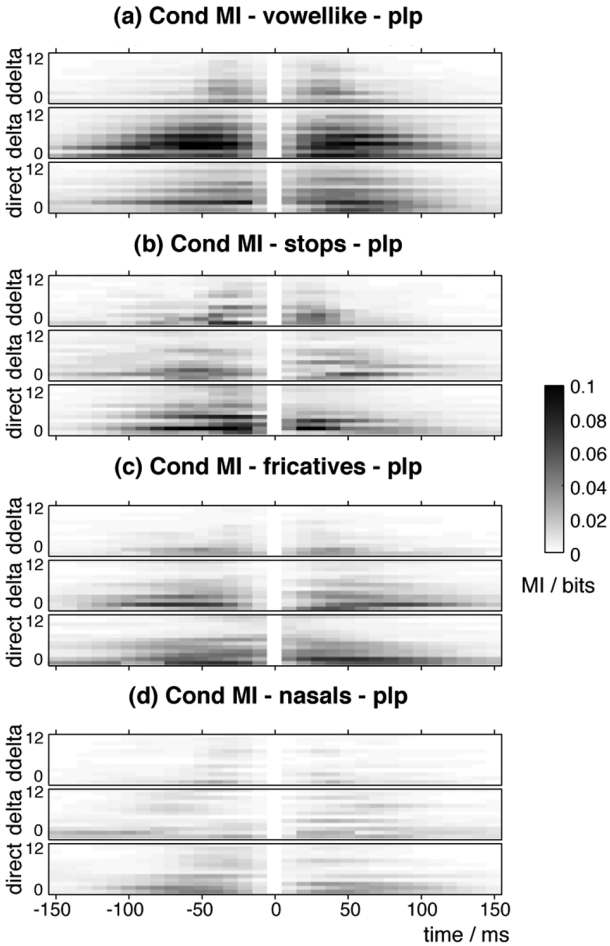


Fig. 3. Conditional MI distribution between each cell on the TF plane and the phone label within each BPGs. (a) Vowels, (b) Stops, (c) Fricatives, and (d) Nasals, additionally conditioned on the value of the time-frequency cell at time zero for that frequency band.

A. Baseline System

The hybrid ANN/HMM speech recognition framework described in [2] was used as our baseline system to estimate the 61 TIMIT phone posteriors. The neural network multilayer perceptron (MLP) classifier had an input layer of 351 units to accommodate the 39 PLP plus first and second derivative features as described in the previous section, over a context window of nine frames. The network also had a single hidden layer (whose size was varied in our experiments) and 61 output units, corresponding to each phone class. The network was trained to estimate the posterior probability for each of the 61 TIMIT phone classes for each frame of input by back-propagation of a minimum-cross-entropy error criterion against "one-hot" targets. The MLP was trained using all 468 speakers from the eight dialects of the TIMIT database—a total of 4680 utterances, of which 370 utterances were used for cross-validation. The cross validation set is used for adjusting the learning rate during MLP training and also for determining the early stopping point to prevent over-fitting.

These posteriors are scaled using phone priors, and the 61 phones were then mapped to a smaller set of 39 phones prior to being fed to an HMM decoder to find a single sequence of phone labels that best combines models and observations. This phone sequence is compared to the manual ground truth to produce a phone error rate (PER) that includes all substitutions, deletions, and insertions.

The 39 PLP plus first and second derivative features were computed for each frame in both the training and test sets. The mean and standard deviation was computed across all features in the training data for normalization. Each feature dimension in the training set is separately scaled and shifted to have zero mean and unit variance, which ensures the MLP input units are operating within their soft saturation limits. The same normalization is applied to the test sets.

The 168 test speakers were divided into two groups: 84 speakers were used in the development set to tune variables, and the other 84 were used in the final test set for evaluation of the proposed network.

B. Broad Phonetic Group Expert

The networks used for the BPG experts are similar to that of the baseline system but the output layers consist of a smaller number of units, e.g., 25, 8, 10, and 7 units for vowels, stops, fricatives, and nasals, respectively.

MI indicates which TF cell contain the most information for discriminating between each of the BPGs. A different feature set is extracted for each BPG to maximize discrimination capabilities of the expert, but the total number of input units is held constant across all experts.

C. BPG Detector

In order to determine whether to assign the candidate frame to the silence group or one of the BPG experts, two different methods were investigated. The first uses the baseline classifier output to determine which BPG or the silence group dominates the posterior distribution, by summing all the posteriors from each group and assigning the group with the greatest pooled posterior probability to the candidate frame. This is similar to the method described in [9]. Note if the silence group is assigned to the frame no expert is used, and the baseline posteriors are preserved in the final output stream.

The second method uses one classifier for each BPG and one for the silence group, each with a binary output (i.e., this group or not this group). The posterior probabilities from each of these detectors was combined to determine the inferred BPG or the silence group of the current frame.

Since these two mechanisms for estimating the current frame's group are different, they can give different results. A third method combines these two approaches and only assigns a candidate frame to a BPG or silence group once both methods agree. When the methods disagree the original baseline posteriors are maintained.

D. Integration

Given the outputs of several different classifiers (the baseline plus one or more experts), the question then arises of how to combine these differing values into a single set of posteriors to pass on to the decoder. One choice is to simply patch all the BPG phoneme posteriors in the baseline output with the posteriors of the BPG expert and set all other phoneme posteriors

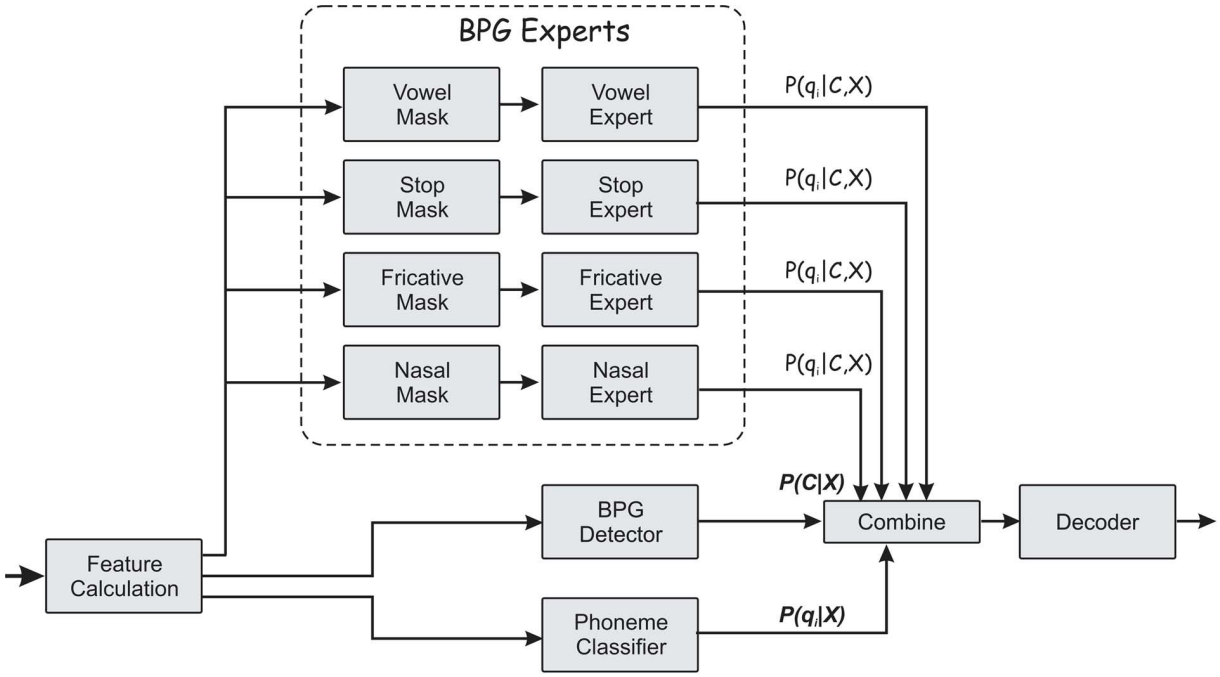


Fig. 4. Classifier architecture: Individual classifiers for each BPG are run on group-specific feature masks applied to the entire utterance, then combined with a general-purpose classifier at the posterior level according to the estimated current BPG.

TABLE II
FRAME PHONE CLASSIFICATION ACCURACIES (%) FOR DIFFERENT METHODS OF FEATURE SELECTION: PLP RECT, PLP MI-IRREG FOR ALL BPGS USING 100 HIDDEN UNITS. THREE-HUNDRED FIFTY-ONE PLP FEATURES ARE USED

BPG	Vowels	Stops	Fricatives	Nasals
RECT	55.2	78.9	77.9	73.9
MI-IRREG	55.6	79.8	79.7	72.9

to zero—i.e., fully replacing the outputs of the baseline classifier for frames detected as belonging to a particular BPG. However, if the BPG classification is in error, this may result in irreparable damage to the posterior stream. Another approach is to mix the phoneme posteriors in the baseline output with the posteriors of the BPG expert using fixed mixing weights, so that even when a particular BPG class has been chosen, the posteriors remain a mixture of both expert and baseline classifiers. It would also be possible to make variable interpolations between the two sets of posteriors based, e.g., on the degree of confidence of the current BPG label, but in preliminary experiments a variable-mixing-weight rule that showed any advantage over hard decisions was not found.

V. CLASSIFIER EXPERIMENTS

A. BPG Experts

Table II illustrates that high-MI feature selection leads to improved performance. The table compares the accuracies for frame-level phone classification of each expert individually for both baseline RECT and MI-IRREG features using 351 features. In all cases, the expert MLP classifiers had 100 hidden units. It can be seen from Table II that the performance of the MI-IRREG features are significantly better than the baseline RECT features for all BPGs except Nasals. Similarly based on these results, MI-IRREG features are used for Vowel, Stop, and Fricative experts, and RECT features are used for Nasal experts. Significance at the 5% level is 0.4%, 0.9%, 0.6%, and 0.9% for Vowel, Stop, Fricative, and Nasal frame accuracies, respectively; note that the improvements due to MI-IRREG are at the lower limit of significance in most cases.

Since each BPG has different characteristics, with different feature selections made according to the MI criteria, it is worth investigating the variation of accuracy with the size of the feature vector independently for each expert: it is expected that increasing the amount of information available for each classifier will improve performance up to a point, beyond which the burden of the added complexity fails to outweigh the added information, and performance actually declines due to overtraining. Fig. 5 shows the frame accuracy across 195, 273, 351, and 429 features. Fig. 5 also examines the effect of omitting adjacent feature vectors in time to avoid any possible correlation of the features. A feature vector dimensionality of 273 was found to maximize frame accuracy for the Vowel, Stop, and Nasal experts while 351 maximized performance for the Fricative expert. Experts for Stops, Fricatives, and Nasals maximized performance using all features, whereas the Vowel expert performed best when the "striped" MI-IRREG mask was used for feature selection. Again, the variables which performed best for each BPG were used for the remainder of the experiments.

As the expert networks have fewer outputs than the baseline classifier (i.e., 7 to 25 versus 61 in the baseline), the BPG expert units can afford to have larger hidden layers without increasing the total complexity of the classifiers. The results of varying the hidden layer sizes to 100, 500, 1000, 2000, 3000, 4000, and 5000 units are shown in Table III. Although the gains due to the much larger networks are sometimes quite small, for the Vowels and Stops experts, 4000 hidden units provided

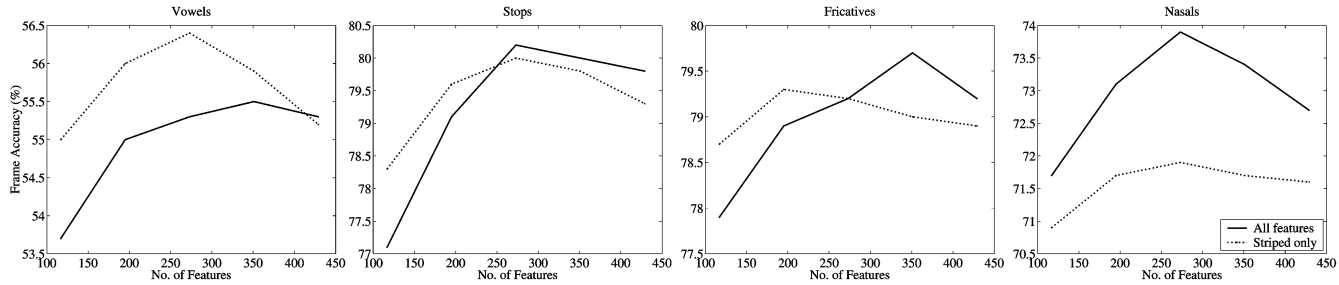


Fig. 5. Frame accuracy for different feature vector dimensions using all features or "striping."

TABLE III
PHONE CLASSIFICATION FRAME-LEVEL ACCURACIES (%) FOR DIFFERENT NETWORK HIDDEN UNITS FOR ALL BROAD PHONETIC GROUPS

Hidden Units	100	500	1000	2000	3000	4000	5000
Vowels	56.4	58.8	59.6	59.2	59.8	**60.5**	60.2
Stops	80.2	82.3	82.9	83.0	83.1	**83.3**	83.3
Fricatives	79.7	80.8	80.9	81.0	**81.4**	81.4	80.7
Nasals	73.9	74.9	75.1	75.3	**75.9**	75.8	75.7

TABLE IV
BPG DETECTOR FRAME ACCURACIES (%) FOR DIFFERENT FEATURE VECTOR DIMENSIONS (MLP INPUT UNITS). HIDDEN UNITS ARE HELD CONSTANT AT 100

Features	117	195	273	351	429
BPG Detector	90.7	91.0	91.0	91.1	91.0

TABLE V
BPG DETECTOR FRAME ACCURACIES (%) FOR DIFFERENT NUMBER OF MLP HIDDEN UNITS. THERE ARE 351 INPUT UNITS IN EACH CASE

Hidden Units	100	500	1000	2000	3000
BPG Detector	91.1	91.3	91.4	91.4	90.8

maximum frame accuracy, while for both Fricatives and Nasals, 3000 hidden units maximized performance.

B. BPG Detector

In this section, three methods of assigning candidate frames to BPG experts are compared. The first method considered uses the baseline classifier's output to determine which BPG or silence group dominates the posterior distribution. This approach provides a frame-level BPG classification accuracy of 90.8%.

The second method uses a separate network for each BPG and a silence group with a binary output. The frame is labeled with the group corresponding to the network with the greatest confidence (largest posterior), given the silence group the baseline posteriors are maintained for that frame. Table IV provides the frame accuracies for a number of different feature vector sizes; best performance is achieved for 351 inputs. In these BPG detector networks, only two output units are required, and since the number of output units is so small more hidden units can be used without increasing complexity of the system. The results of varying the hidden units for 100, 500, 1000, 2000, and 3000 are shown in Table V. In these results, a difference of around 0.2% is significant at the 5% level.

C. Integration With Baseline System

The BPG phoneme posteriors in the baseline output are merged with the posteriors of the BPG expert using constant

TABLE VI
PERS (%) OBTAINED FROM PATCHING WEIGHTED BPG EXPERT OUTPUTS INTO BASELINE SYSTEM, USING 100 HIDDEN UNITS IN THE BASELINE SYSTEM, FOR DIFFERENT METHODS OF BPG DETECTION, AS A FUNCTION OF THE MIXING WEIGHT

Weight	1	0.9	0.7	0.5	0.3	0.1	0
BPG Detector	28.8	26.7	26.8	27.5	29.0	31.9	33.9
BPG Posteriors	28.7	26.7	26.9	27.8	28.9	32.2	33.9
Combined	26.9	**26.4**	27.0	27.8	29.3	32.2	33.9
Oracle	22.6	22.8	23.4	24.6	26.9	31.3	33.9

mixing proportions. The PERs in Table VI were obtained by varying the mixing weights then passing the merged posteriors to the HMM decoder to obtain a final inferred phoneme sequence; when the mixing weight is zero, the baseline classifier posteriors are unchanged regardless of the detected BPG, and the baseline PER is achieved.

Both basic methods of BPG detection ("BPG Detector" and "BPG Posteriors") perform similarly. The "Combined" method combines the results of the previous approaches and only assigns a candidate frame to a BPG once both methods agree; it can be seen that this provides improvement in performance—indicating that the two basic methods differ in their errors, and that combining them avoids some of these errors. The "oracle" results are obtained by using the the ground-truth BPG label to control the patching, i.e., using the labels of the database to assign each frame to the silence group or one of the BPG experts. This gives an idea of the upper bound achievable by the BPG experts given ideal BPG detection.

The results of Table VI were given using a baseline network with 100 hidden units. In Table VII, the number of hidden units in the baseline classifier was varied over 100, 500, 1000, and 2000 hidden units. When the mixing weight is zero, the PER corresponds to the baseline system without BPG experts. While baseline performance improves markedly for larger classifier networks, significant improvements can still be seen over baseline as the experts are patched in. Significance at the 5% level is achieved for a difference of 0.7% in these results

The results in the experiments were maximised for the development set. Given a baseline PER of 26.5%, using the proposed modular architecture reduces this error to 25.2%. Application to the omitted test set of speakers from dialects 4 to 8 in the TIMIT dataset gives a baseline PER of 27.3%, which is reduced to 26.3% using BPG experts. For both the development and test sets, 5% statistical significance is achieved for a difference of around 0.7%. Over the entire test set the baseline PER is reduced from 26.9% to 25.8% using the proposed architecture,

TABLE VII
PERS (%) OBTAINED FROM PATCHING IN WEIGHTED BPG EXPERT OUTPUTS INTO BASELINE SYSTEM FOR DIFFERENT NUMBERS OF HIDDEN UNITS, USING THE "COMBINED" METHOD OF BPG DETECTION FROM TABLE VI, AS A FUNCTION OF THE MIXING WEIGHT

Weight	**1**	**0.9**	**0.7**	**0.5**	**0.3**	**0.1**	**0**
100 hidden units	26.9	26.4	27.0	27.8	29.3	32.2	33.9
500 hidden units	26.2	25.6	25.6	25.8	26.5	27.1	27.7
1000 hidden units	26.1	25.3	**25.2**	25.3	25.5	26.0	26.5
2000 hidden units	26.2	25.4	25.4	25.4	25.5	26.3	26.8

for the entire test set 5% statistical significance is achieved for a difference of around 0.5%.

VI. DISCUSSION

The spread of relevant information for each of the BPGs was illustrated in the MI plots of Fig. 2. These observations reinforce received wisdom concerning different phone classes based purely on objective measurements. Of course, the great contrast shown between the BPGs reinforces the case that BPGs should benefit from distinct, expert classifiers, structurally adapted to obtain the most information from the front-end features.

The number of hidden nodes has a strong impact on the performance of a neural network classifier. The more hidden nodes it contains, the more complex the model it can capture. Good recognition performance, however, depends on the availability of sufficient training data.

Training an NN on limited data can lead to over fitting which is more likely to occur as more hidden nodes are introduced. To prevent overfitting, training is usually stopped early, using the performance of the network measured with a cross validation (CV) dataset held out from the main training data. In our learning schemes, training is typically stopped when the performance of the CV set increases by less than 0.5% after an entire back-propagation pass through the training set. When training the single, baseline classifier stopping criteria represents an average across all phonemes and may not be ideal for each BPG. In using the expert networks proposed in this paper, not only are the feature sets specific to each broad phonetic class of phonemes, but also the early stopping point can specifically prevent overfitting of this class.

Given the limited amount of training data available using the TIMIT database there is a limit to the number of hidden nodes that can be used to model the complexities of the data without overfitting the training set. As was seen in the experiments, performance ceases to improve, and in some cases decreases, past a certain number of hidden nodes. The baseline system performance is at maximum with 1000 hidden units, while the smaller expert system performance is maximized at 3000–4000 nodes. However, even in these cases, very little improvement is seen above 1000 units.

Current methods of computing MI and conditional MI use the histogram approach to obtain the density estimation between one or two features and the classes of interest, but ideally the joint MI between the entire feature set selected so far and each successive candidate could be computed. This approach would benefit from more sophisticated methods to obtain a multivariate probability density estimation between a complete set of features.

In Table VI, the oracle results illustrate the potential of the system given an ideal BPG detector. Therefore, crucial to the performance of the proposed system is the BPG detection. Based on the confusion matrix in Fig. 1, given division of phonemes into the BPGs: vowel, semivowels, dipthongs, stops, fricatives, and nasals, only 50% of misclassified frames fell within the same BPGs. However, grouping the similar vowel-like BPGs vowels, semivowels, and dipthongs increased this percentage to 75%. Therefore, the task of BPG detection is simplified and improved BPG feature extraction is achieved, by further increasing the number of misclassified frames that fall within the same BPG. For this reason, it is hypothesized that a more rigorous approach to grouping phonemes into BPGs would improve system performance.

VII. CONCLUSION

In this paper, using the observation that phone-level confusions fall most often into the same BPG as the true target, a phone recognition system was designed with separate experts trained to discriminate only within the broad classes of Vowels, Stops, Fricatives, and Nasals. Since the TF characteristics of these different speech sounds are so different, the experts were each given individual, distinct "perspectives" on the input signal by selecting subsets of the feature dimensions drawn from a wide time window and choosing the feature dimensions exhibiting the greatest MI with the class-conditional label. It was shown empirically that this feature selection gave a small but meaningful improvement in classification accuracy for three of the four broad classes.

To construct a complete phone recognition system, we needed to mix the judgments of the experts with the baseline classifier under the guidance of a separate broad-class detector. The method of simply pooling groups of posteriors from the baseline classifier was compared with an ensemble of separately trained detectors, one for each broad class. While both approaches performed similarly, combining them such as to detect a broad class only when both detectors agreed gave the best overall performance.

An elaborate classification scheme must of course prove itself superior to the simple approach of increasing the complexity of a single baseline classifier—in our case, adding more hidden units to the MLP neural network. For both baseline and experts, the hidden layer sizes were increased to the maximum supportable by the TIMIT training set used in the. Even with the rather large networks this implied, the expert-based system continued to afford significant error rate reductions; for smaller, more computationally efficient systems, the gains possible with the experts are even larger.

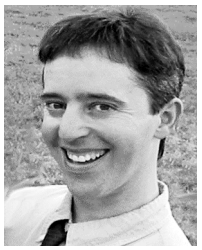
REFERENCES

[1] J. Bilmes, "Maximum mutual information based reduction strategies for cross-correlation based joint distribution modelling," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Seattle, WA, 1998, pp. 469–472.
[2] H. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach*. Norwell, MA: Kluwer, 1994.
[3] S. Chang, S. Greenberg, and M. Wester, "An elitist approach to articulatory-acoustic feature classification," in *Proc. Eurospeech*, 2001, pp. 1725–1728.
[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[5] A. Halberstadt and J. Glass, "Heterogeneous acoustic measurements for phonetic classification," in *Proc. Eurospeech*, 1997, pp. 401–404.
[6] H. Hermansky, "Perceptual linear predictive (PLP) analysis for speech," *J. Acoust. Soc. Amer.*, vol. 87, pp. 1738–1752, 1990.
[7] A. Morris, J. Schwartz, and P. Escudier, "An information theoretical investigation into the distribution of phonetic information across the auditory spectrogram," *Comput. Speech Lang.*, vol. 7, no. 2, pp. 121–136, 1993.
[8] M. Rajamanohar and E. Fosler-Lussier, "An evaluation of hierarchical articulatory feature detectors," in *IEEE Autom. Speech Recognition Understanding Workshop*, 2005, pp. 59–64.
[9] T. Reynolds and C. Antoniou, "Experiments in speech recognition using a modular mlp architecture for acoustic modelling," *Inf. Sci.*, vol. 156, pp. 39–54, 2003.
[10] P. Scanlon, D. P. W. Ellis, and R. Reilly, "Using mutual information to design class specific phone recognizers," in *Proc. Eurospeech*, 2003, pp. 857–860.
[11] H. Schwenk, "Using boosting to improve a hybrid HMM/neural network speech recognizer," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1999, pp. 1009–1012.
[12] S. Sivadas and H. Hermansky, "Hierarchical tandem feature extraction," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2002, pp. 809–812.
[13] H. Yang, S. V. Vuuren, S. Sharma, and H. Hermansky, "Relevance of time-frequency features for phonetic and speaker-channel classification," *Speech Commun.*, vol. 31, pp. 35–50, 2000.
[14] Y. Zhao, R. Schwartz, J. Sroka, and J. Makhoul, "Hierarchical mixtures of experts methodology applied to continuous speech recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1995, pp. 3443–3446.

Patricia Scanlon received the B.Sc.Eng. degree from Trinity College Dublin, Ireland, in 1997 and the Ph.D. degree from the National University of Ireland, in 2000, both in electrical and electronic engineering.

From 1997 to 2000, she worked in the telcommunications industry as a Software Engineer and Technical Team Leader. From 2002 to 2003, she was a Visiting Scholar at the Laboratory for Recognition and Organization of Speech and Audio (LabROSA), Columbia University, New York, NY. In 2004, she interned with the Audio Visual Speech Technologies Group at the IBM T. J. Watson Research Center, Yorktown Heights, NY. She currently holds a Postdoctoral position with Alcatel-Lucent, Bell Laboratories Ireland. Her research interests are in the area of feature extraction and pattern recognition for audio and video signals. Her application areas include audiovisual speech recognition, multimodal biometrics, and machine health monitoring.



Daniel P. W. Ellis (S'92–M'96–SM'04) received the Ph.D. degree in electrical engineering from the Massachusetts Institute of Technology (MIT), Cambridge.

He was a Research Assistant in the Media Laboratory at MIT while pursuing the Ph.D. degree. He spent several years as a Research Scientist at the International Computer Science Institute, Berkeley, CA. Currently, he is an Associate Professor with the Electrical Engineering Department, Columbia University, New York. His Laboratory for Recognition and Organization of Speech and Audio (LabROSA) is concerned with all aspects of extracting high-level information from audio, including speech recognition, music description, and computational auditory scene analysis. He also runs the AUDITORY e-mail list of 1700 worldwide researchers in perception and cognition of sound.

Richard B. Reilly (M'92–SM'04) received the B.E., M.Eng.Sc., and Ph.D. degrees in electronic engineering from the National University of Ireland, Dublin, U.K., in 1987, 1989, and 1992, respectively.

In 1988, he joined Space Technology Ireland and the Department de Recherche Spatiale (CNRS group), Paris, France, developing DSP-based on-board experimentation for the NASA satellite WIND. In 1990, he joined the National Rehabilitation Hospital and in 1992 became a Postdoctoral Research Fellow at University College Dublin, focusing on signal processing for speech and gesture recognition. Since 1996, he has been on the academic staff of the School of Electrical, Electronic, and Mechanical Engineering, University College, Dublin. He is currently Senior Lecturer and researches into neurological signal processing and multimodal signal processing.

Prof. Reilly was the 1999/2001 Silvanus P. Thompson International Lecturer for the IEE. In 2004, he was awarded a U.S. Fulbright Award for research collaboration into multisensory integration with the Nathan Kline Institute for Psychiatric Research, New York. He is a member of the IEEE Engineering in Medicine and Biology Society and Signal Processing Society. He is the Republic of Ireland representative on the Executive Committee of the IEEE U.K. and Republic of Ireland Section. He is an Associate Editor for IEEE TRANSACTIONS ON MULTIMEDIA and also a reviewer for IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, IEEE TRANSACTIONS ON NEURAL SYSTEMS AND REHABILITATION ENGINEERING, IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, *Journal of Applied Signal Processing*, *Signal Processing* and *IEE Proceedings Vision, Image and Signal Processing*.