DYNAMIC SELECTION OF FEATURE SPACES FOR ROBUST SPEECH RECOGNITION

Bhuvana Ramabhadran, Yuqing Gao, Michael Picheny

Human Language Technologies IBM T. J. Watson Research Center
Yorktown Heights, NY, 10598.
(bhuvana,yuqing,picheny)@watson.ibm.com

ABSTRACT

Selection of acoustic features for robust speech recognition has been the subject of research for several years. In the past, algorithms that use feature vectors from multiple frequency bands [9], or employ techniques to switch between multiple feature streams [10] have been reported in the literature to handle robustness under different acoustic conditions. Acoustic models built out of different feature sets produce different kinds of recognition errors. In this paper, we propose a likelihood-based scheme to combine the acoustic feature vectors from multiple signal processing schemes within the decoding framework, in order to extract maximum benefit from these different acoustic feature vectors and models. The proposed technique is general enough to be applied to other pattern recognition fields, such as, OCR, handwriting recognition, etc. The fundamental idea behind this approach is to pick the set of features that classifies a frame of speech accurately with no apriori information about the phonetic class or acoustic channel that this speech comes from. Two methods of merging any set of acoustic features, such as, formant-based features, cepstral feature vectors, PLP features, LDA features etc., are presented here:

- Use of a weighted set of likelihoods obtained from these several alternative feature sets and
- Selection of the feature space that ranks the best when used in a rank-based recognizer.

These merging algorithms provide an impressive reduction in error rate between 8% to 15% relative across a wide variety of wide-band, clean and noisy large vocabulary continuous speech recognition tasks. Much of this gain is from the reduced insertion and substitution errors. Using the approach presented in this paper, we have achieved better improved acoustic modeling without increasing the number of parameters, i.e. two 40K Gaussian systems, when merged perform better than a 180K Gaussian system trained on the better of the two feature spaces.

1. INTRODUCTION

Spectral based feature vectors such as mel-frequency cepstral vectors, perceptual linear predictive coefficients (PLP), maximum-likelihood based linearly transformed features, formant-based features, etc. form the basis of most speech recognition systems. The acoustic models used to model these different feature spaces produce very different types of decoding errors. Depending on the feature space that these acoustic models operate on, their accuracy for classifying vowels, fricatives and other consonants vary. Also, the signal processing scheme used (LDA, PLP, cepstra, factor analysis, transformed features, etc) determines the robustness of these models to varying noise conditions. The use of information content in features extracted from Bark-spaced multiple critical frequency bands of speech have been proposed in [6, 9] in order to increase speech recognition accuracy. Typically, most of these feature streams contain complementary information and an efficient combination of these streams, would not only result in increased recognition accuracy, but will also serve as a means to select the feature stream that best represents the acoustics at the given time frame or segment. The overall performance of the final acoustic model, which is a combination of acoustic models based on several feature spaces, depends on how well the error patterns from these streams complement one another and how much redundant information they possess [1]. In some cases, even when the performance of one of the streams is not so robust or far worse than the best system, it may contain hidden characteristic information that becomes more valuable when the two streams are merged. The algorithm presented here, implicitly switches acoustic models dynamically during decoding, at the frame, phone, or syllable level, such that the feature space that best models the acoustics at any given time frame is always used.

In the past, multi-scale systems have been explored [2], where each stream operates on different time windows. In [7], an MLP was trained to map the feature spaces to the log-likelihoods of phonemes and the combination scheme involved the averaging of the features prior to orthogonalization. In the NIST-based ROVER scheme, a voting mechanism is used after an initial decoding pass to combine the best output from each model. In [5] a hierarchical architecture for combining classifiers for speech recognition is presented.

In this paper, we propose two maximum-likelihood based combination schemes, namely, a weighted, normalized likelihood combination and rank-based combination scheme. The selection algorithm proposed here does not increase the computational load during decoding. When several feature streams are decoded separately and combined thereafter (multi-pass decoding strategy), the decoding time increases linearly with the number of input streams. It is therefore possible to considerably reduce computational requirements while maintaining good robustness to changing acoustic conditions and improving recognition accuracy. The proposed strategy provides relative improvements over the best performing system in all of the test sets used. Maximum-likelihood is used as the criterion to select the best model. The feature selection method proposed here improves the overall recognition accuracy by 15% relative in large vocabulary continuous speech recognition tasks. Much of this gain is from the reduced insertion and substitution errors. As can be seen,
it is possible to increase clean speech performance by optimally selecting the feature that best represents a frame of speech, without increasing the number of system parameters.

Section 2 explains the two selection schemes and the mathematical formulation. Section 3 discusses the implementation issues. In Section 4, the databases, feature spaces and the recognition systems used are described. Section 5 tabulates the results obtained. Section 6 contains a discussion of the results and directions for future work.

2. FEATURE SELECTION/COMBINATION ALGORITHM

The IBM Continuous speech recognition system used here uses a set of phonetic baseforms and context-dependent models. These models are re-defined by constructing decision tree networks that query the phonetic context to arrive at the appropriate models for the given context. Each decision tree is constructed for every sub-phonetic unit. Each terminal node (leaf) of the tree represents a set of phonetic contexts, such that the feature vectors observed in these contexts were close together as defined by how well they fit a diagonal Gaussian model. The feature vectors at each terminal node are modeled using a mixture of Gaussians with each Gaussian having a diagonal covariance matrix. The leaves of the decision tree correspond to context-dependent HMM states.

The IBM system also uses a rank-based decoding scheme [4]. The rank \( r(s,t) \) of a state \( s \) at time \( t \) is the rank order of the likelihood given the mixture model of this state in the sorted list of likelihoods computed using all the models of all the states in the system and sorting them in descending order. In such a system, the output distributions on the state transitions of the model are expressed in terms of the rank of the state. Each state transition has a probability distribution on ranks which typically has a peak at rank one and rapidly falls off to low probabilities for higher ranks.

This representation overcomes problems associated with the large fluctuations in the likelihood scores seen in continuous parameter HMMs. Also, if the observations for a few time frames are outliers for the correct mixture model, the penalty is not so high that the correct word comes out with a very low likelihood. This makes the rank-based system more robust compared to continuous parameter HMMs.

When there are several feature vectors from different feature spaces representing an observation, a different set of Gaussian mixture models is trained for each of these feature spaces, while keeping the context-dependent states the same. During decoding, the likelihoods computed from the Gaussian mixture models are used to rank order the states to form an ordered list of states. When there is more than one feature stream, there is one ordered list of states for each feature stream. The problem of combining/selecting feature spaces now becomes the problem of combining the ordered lists in a manner that best represents the acoustic at any given time frame.

Let \( x(t) = (x_1, x_2, \ldots, x_N) \) be the \( N \) observed feature vectors at time \( t \) and \( N \) is the number of feature spaces used to represent the speech signal. \( s_j \) is the \( j \)th state of the HMM. In order for the correct leaf to occupy the top rank positions, the probability given by

\[
p(s_j | x_1, \ldots, x_N) = \frac{p(x_1, \ldots, x_N | s_j)}{p(x_1, \ldots, x_N)}
\]

has to be maximized. It has been observed that recognition accuracy gets better when the correct leaves occupy the top rank positions more number of times.

The term \( p(x_1, \ldots, x_N | s_j) \) represents the combined feature space. Assuming that the feature vectors from the \( N \) streams are independent, Eqn 1 becomes

\[
p(s_j | x_1, \ldots, x_N) = \frac{p(x_1 | s_j)p(x_2 | s_j) \ldots p(x_N | s_j)}{p(x_1, \ldots, x_N)}
\]

in practice, we maximize \( p(x_1 | s_j)p(x_2 | s_j) \ldots p(x_N | s_j) \), which can be viewed as a function of a set of weights, \( \{w_n\} \), operating on \( p(x_n | s_j) \), i.e.

\[
f(w_1 p(x_1 | s_j), w_2 p(x_2 | s_j), \ldots w_N p(x_N | s_j))
\]

with the constraint,

\[
\sum_{n=1}^{N} w_n = 1, \forall n
\]

In order to boost the rank of the correct leaf, we need to boost the combined likelihood of all feature vectors from all the streams at any given time. One approach involves the averaging of the log posterior probabilities from the individual acoustic models trained on the separate feature streams. A second approach for combining the likelihoods of multiple feature streams uses the well-known sum and product rules. In [8], a discriminative model combination approach that optimally combines several acoustic and language models has been suggested. In this paper, we propose two methods for performing feature selection, which are detailed below. The combination of models/features can be implemented in both the rank domain or in the raw feature vector space. It should be emphasized here that while one of the methods serves as a means to merge feature streams, the other selects the best feature stream. Since both these methods operate at the frame level, no specific synchronization is needed at any sub-unit level, such as, the phone, word or syllable level.

2.1. Weighted Likelihood Combination

In a continuous-density HMM system, the context-dependent states of the HMM are derived from decision trees and modeled using Gaussian mixture densities. Maximum-likelihood is used as the criterion to select the best model from a set of \( n \) models according to the following equation:

\[
p(x_1 | s_j) = \exp \{K + [w_1 p(x_1 | s_j)^q + w_2 p(x_2 | s_j)^q \ldots w_N p(x_N | s_j)^q]^{1/q}\}
\]

where, \( w_1, w_2, \ldots, w_N \) are the weights, \( p(x_1 | s_j), \ldots, p(x_N | s_j) \) are the likelihoods from the multiple streams for the state \( s_j \) and \( K \) is a normalization constant.

The different values of \( q \) : 0, 1, \ldots, \infty represent different combination functions. The weights reflect the confidence in the quality of the states of the streams being merged. Hence, they could even be zero under certain noisy conditions, which would imply a form of feature selection. These weights can be tied across states and estimated using optimization techniques. We use this representation, because in the limit, when \( q \) goes to \( \infty \), this expression reduces to a \( \max \) operation and this is the theory behind the rank-based state selection scheme.
2.2. Rank-based state selection

As mentioned before, the rank \( r(s, t) \) of a state \( s \) at time \( t \) is the rank order of the likelihood given the mixture model of this leaf in the sorted list of likelihoods computed using all the models of all the leaves in the system and sorting them in descending order. The more number of times a correct leaf appears in the top rank positions, the better the recognition accuracy. In order to improve the rank of the correct state, its likelihood score has to be boosted up relative to other leaves. The emphasis here is on the selection of appropriate features that are robust to certain acoustic conditions and also model certain phonetic sounds better. Hence, we wish to pick the feature stream for which a state scores the topmost rank.

From Equation 5, it can be seen that when \( q \) tends to \( \infty \), this reduces to a max operation given by

\[
p(x(t)|s_j) = \exp\{K + \max\{p(x_1|s_j), p(x_2|s_j), \ldots, p(x_N|s_j)\}\}^{1/\gamma}
\]

This represents the choice of the feature stream as the one with the best rank in the ranking scheme described above.

3. IMPLEMENTATION

In the weighted likelihood based selection scheme, we used uniform weights as an initial choice for running our experiments and \( q = 1 \). Other choices of \( q \) did not yield any further improvements in recognition accuracy. Hence, for a two-stream input, the weights were chosen to be 0.5 and maintained constant throughout the test utterances. In the rank-based method, first, an n-best rank list for each stream is generated. For each observation vector the states are ordered based on the rank of the states. Next, the merged rank list from all the feature streams is generated by picking the state from the feature stream that yields the highest rank. This results in several states with the same rank in the n-best rank list. Although the correct state is ranked much lower in one feature space, an alternative feature space may be able to capture its characteristic information causing it to be ranked higher. This scheme picks the better feature stream based on the rank positions, thereby introducing discrimination between correct and incorrect states. The weighted-likelihood based method boosts the state likelihoods and provides robustness.

4. EXPERIMENTS

This section describes the experiments, the training and test data sets, the feature streams used and the results.

4.1. Training data

The training data used for all the systems built from the four different feature streams was the same. An in-house data base consisting of 100 hours of training data collected from 2000 speakers was used. The speaker adapted systems used to test read speech were adapted on 30 mins. of read speech from each speaker, while those used to test spontaneous speech were adapted using 30 mins. of spontaneous speech.

4.2. System Description

All systems had approximately 3000 context-dependent HMM states. The speech recognition system uses an alphabet of 52 phones. Each phone is modeled with a 3-state left-to-right HMM. Systems with 160K Gaussians and 245K Gaussians were built for comparing the recognition accuracies obtained using the feature selection scheme with models with a large number of parameters.

4.3. Feature spaces

The basic acoustic front-end uses a cepstral feature vector extracted every 10 ms, along with \( \Delta + \Delta \Delta \) and sentence based cepstral mean normalization. The LDA feature space comprised of 9 frames of cepstral vectors spliced together and the top 40 dimensions with the largest eigen values were selected. The centroid-based feature stream computes the centroids [9] in different frequency bands, based on the formula,

\[
C_m = \frac{\int_{l_m}^{h_m} f_{wm}(f) P_q(f) df}{\int_{l_m}^{h_m} w_{wm}(f) P_q(f) df}
\]

where, \( C_m \) is the centroid in the mth subband with \( l_m \) and \( h_m \) being the lower and higher frequencies of the mth subband, \( P(f) \) is the power spectrum and \( \gamma \) is a constant controlling the dynamic range of the power spectrum. These features are known to be robust to noise and similar to formant frequencies. As can be seen, from Table 3, these are supplementary to the traditional cepstral feature vectors. The enhanced likelihood computation scheme [11] which incorporates forward and backward prediction errors into the regular cepstral stream was the fourth stream used in this research. This feature space captures the correlation between adjacent vectors using regression. The regression predicts the neighboring frames of the current frame of speech. The incorporation of prediction error likelihoods into the overall likelihood computation improves the rank position of the correct leaf, without increasing the complexity of the HMMs.

4.4. Test Data

Different test data sets were used in this study. They can be broken down into read speech (BOS), higher perplexity read speech from both native and non-native speakers (NRR), spontaneous speech (SPO), read speech from a variety of domains (LMNEW) and spelling data (SPL). All the test sets consist of both native and non-native speakers. The BOS data set is an hour of speech from 10 speakers, the NRR data set consists of over 2 hours of speech from 15 speakers, and the SPO test set comprises of an hour of speech from each of the 10 speakers. The LMNEW test set contains 5 speakers and the SPL test set contains 10 speakers.

5. RESULTS

Experiments were performed using speaker adapted and speaker independent models. Table 1 summarizes the results using speaker dependent models using both methods of feature selection. The weights chosen for the likelihood-based scheme were 0.5 and constant. It can be seen that our proposed methods outperform ROVER's performance on these data sets. The extra computations resulting from the additional feature extraction and selection process is approximately an additional 15% of the decoding time. This is considerably less when compared to a combination scheme after several decoding passes. Table 2 tabulates the use of cepstral and LDA feature streams using speaker independent models. It also compares the performance of the two selection schemes with a system built with many more parameters (160K Gaussians) than the combined system (effectively 80K Gaussians). There was negligible
HMM states are selected based on two sets of criteria. These methods provide a reduction in error rate between 8% to 15% relative across a wide variety of LVCSR tasks, both in speaker adapted and speaker independent scenarios. Methods to optimally estimate the weights are also presented. With just the additional cost of computation required by the additional feature spaces, it is possible to achieve recognition accuracies that are higher than those attained by ASRs trained on those individual streams. Furthermore, this approach is as good as training ASRs with a large number of parameters and hence does not suffer from sparse training data issues. As our results demonstrated, when the feature streams contain complimentary information, it is possible to switch between them in a manner that boosts recognition accuracy. As part of this on-going work, some of the ideas that will be explored further are presented here. The weights used for merging different streams can be changed dynamically and tied to specific phones or syllables. They can also be estimated statistically from the training data based on an approach similar to the discriminative model described in [8]. The optimization function will now include the rank distribution function. Mutual information-based techniques for deciding a priori if a feature stream will contribute any additional information will also be explored. The use of this selection scheme in segmental models will also be explored. The proposed technique is general enough to be applied to other pattern recognition fields, such as, OCR, handwriting recognition, etc.

7. REFERENCES