EXPLOITING COMPLEMENTARY ASPECTS OF PHONOLOGICAL FEATURES IN AUTOMATIC SPEECH RECOGNITION

Parya Momayyez, James Waterhouse, and Richard Rose

Department of Electrical and Computer Engineering
McGill University, Montréal, Québec, Canada

ABSTRACT

This paper presents techniques for exploiting complementary information contained in multiple definitions of phonological feature systems. Three different feature systems, differing in their structure and in the acoustic phonetic features they represent, are considered. A two stage process involving a mechanism for frame level phonological feature detection and a mechanism for decoding phoneme sequences from features is implemented for each phonological feature system. Two methods are investigated for integrating these features with MFCC based ASR systems. First, phonological feature and MFCC based systems are combined in a lattice re-scoring paradigm. Second, confusion network based system combination (CNC) is used to combine phone networks derived from phonological distinctive feature (PDF) and MFCC based systems. It is shown, using both methods, that phone error rates can be reduced by as much as 15% relative to the phone error rates obtained for any individual feature stream.

Index Terms— Speech Recognition, Acoustic Modeling, Phonological Features

1. INTRODUCTION

There has been a great deal of research on the use of phonological distinctive feature systems arising from multiple phonological theories in automatic speech recognition (ASR). This is motivated in part by well known linguistic and statistical arguments stating that speech recognition models should benefit from *some* independent, non-redundant underlying feature representation as an alternative to relying on the phoneme as a fundamental unit [1]. The particular issue investigated in this paper is whether complementary information that may be represented by ASR systems defined over multiple phonological feature systems can be exploited using system integration methodologies. The methodologies explored here include lattice re-scoring based methods like those reported in [2] and system combination methods like those described in [3, 4].

The techniques presented in this paper attempt to build on previous work in phonology, feature detection, phoneme decoding from phonological features, and system integration. First, many phonological feature systems have been defined in the linguistics community where the feature definitions themselves are based on different theories of speech production and acoustic phonetics [1, 5, 6]. Second, there has been a great deal of work by speech technologists attempting to extract acoustic parameters that are correlated with these features and using discriminative probabilistic methods for asynchronous detection of the occurrence of the features in the speech signal [7, 8]. Third, techniques have been proposed for combining information from phonological feature detectors for recognizing phone strings in ASR [9, 2]. Finally, there has been a large amount of work in the large vocabulary speech recognition (LVCSR) community devoted to recognizer output voting error reduction (ROVER) and confusion network combination (CNC) schemes for combining the results of ASR systems whose errors are assumed to be complementary in some way [3, 4].

King and Taylor implemented frame level phonological distinctive feature detectors based on feature definitions that were representative of several different phonological theories [1]. It was shown that reasonable frame level speaker independent phone classification accuracy could be obtained for each of these feature representations by using neural network (NN) based feature detectors. These detectors were trained from canonical feature labels obtained from phonemically labeled speech utterances [10]. It is not clear, however, that any particular linguistic theory or any particular definition of distinctive features could ever be considered "optimal" for this purpose. The premise of this paper is that it may be possible to exploit complementary aspects of different phonological feature systems for decoding a sequence of phone labels from speech.

Phonological feature based phone recognition was performed in [1] using an HMM based phoneme recognizer whose distributions were defined over estimates of the posterior probabilities of phonological distinctive features. This approach was motivated by the well known complex and non-linear relationship that exists between phones and direct acoustic observations. It was shown in [2] that this phonological distinctive feature based HMM system could significantly improve phone recognition performance when used as a mechanism for integrating phonological distinctive features with "traditional" MFCC based ASR through a lattice re-scoring paradigm.

This paper extends the work in [1] and [2] by investigating whether multiple definitions of phonological distinctive features can be shown to be complementary in their effect on ASR performance. Two methods are evaluated here for investigating this issue. Both methods rely on a two stage process involving a mechanism for frame level feature detection and a mechanism for decoding phoneme sequences from detected phonological features for each feature stream. The components of this process are described in Section 3. The first method, described in Section 4, involves integrating phone strings derived from phonological distinctive features with the solution space of a conventional mel frequency cepstrum coefficient (MFCC) based HMM speech recognizer. The second method, described in Section 5, attempts to exploit the potential diversity among the phone strings that are derived from the separate feature streams. If the phonetic networks generated by the different feature representations produce similar phone recognition accuracy and the errors generated by the different features are in some way complementary, then integrating the solution space of these networks through ROVER or CNC

This work was performed in collaboration with the DIVINES FP6 project and has been supported under NSERC Program Number 307188-2004

based system combination can result in significant improvement in phone accuracy.

An experimental study is described in Section 6 that evaluates the effects of the above techniques on phone recognition accuracy (PAC). Performance obtained using different strategies for combining the outputs of the different phonological feature based and MFCC based systems is measured on the TIMIT speech corpus [10].

2. PHONOLOGICAL FEATURE SYSTEMS

Three different definitions of PDFs are investigated here and correspond to feature sets used in a previous study of frame level feature detection [1]. These feature sets are briefly introduced in this section in order to contrast the differing motivations that lead to their development, the differences in structure, and the differences in the level of articulatory and acoustic information that is characterized by each one.

Two of the phonological feature systems are based on properties of speech production. The first system was motivated by the distinctive features originally defined in Chomsky and Halle's, *The sound pattern of English* (SPE) [5]. The first feature set, referred to as SPE, consists of a set of thirteen binary features. These include binary values for the classes strident, nasal, continuant, voice, tense, round, coronal, anterior, low, back, high, consonantal, and vocalic. The second feature set consists of only eight features where each feature can assume anywhere from two to ten values. This is referred to as the multi-valued (MV) system. It defines phones in terms of well known linguistic terminology, such as manner and place, arranged into a hierarchy. The features include centrality, continuant, front-back, manner, phonation, place, roundness, and tenseness.

The third phonological feature system is based on multiple properties of the speech spectrum referred to as "primes" and is motivated by the theory Government Phonology (GP) [6]. This will be referred to as the GP phonological feature system. The GP feature system differs from SPE and MV feature systems primarily in that it is defined with respect to acoustic classes rather than the speech production classes of the SPE and MV. King and Taylor encoded the structure of this system into a set of modified features that could be detected from speech with reasonable accuracy [1].

The three feature sets, SPE, MV, and GP, differ both in the level of hierarchy that are embedded in the representations and also in the balance between acoustic and phonological features that are represented. It is natural to expect that using these different feature systems in phone recognition could potentially result in decoded phone strings whose errors are complementary. It will be shown in Section 6 how these differences result in a level of diversity that can be exploited in system combination.

3. PHONOLOGICAL DISTINCTIVE FEATURE BASED PHONE RECOGNITION

This section extends the work described in [2] where phonological distinctive feature (PDF) based phoneme recognition systems were developed as part of a larger formalism for integrating PDFs with MFCC based ASR. There are three components to these feature based phoneme recognition systems that will be described here and applied to all three of the feature sets described in Section 2. First, a set of neural network based classifiers that are used to extract the values for phonological distinctive features are presented. Second, the de-correlating and amplitude compression transformations applied to the neural network outputs are described. These transformations are necessary for providing a better match between the diagonal covariance Gaussian observation distributions in the HMMs and phonological feature vectors. Finally, separate HMM models defined over spectral energy (MFCC) based observations and multiple phonological feature based observations are trained and used for decoding phone sequences and generating phone lattices.

3.1. Neural Network Based Feature Detection

Following the work in [1], time delay neural networks (TDNNs) were used for phonological feature detection. For GP and SPE feature systems, all features were extracted simultaneously using a single network. For the multi-valued (MV) feature system, a separate TDNN was used for each of the eight features [1]. All the networks have a single hidden layer with varying number of hidden units. The input to each network is a vector of twelve MFCCs along with their first and second differences. The NICO toolkit was used for the back propagation training of all network parameters [11].

For the MV feature set, the output activations for each of the eight feature based TDNNs, 28 values in all, correspond to the binary values defined in the MV set. The frame classification accuracy of these TDNNs ranges between 73% for the 10 element "place" feature and 92% for the 2 element "phonation" feature. The frame accuracy for individual features for SPE set was between 88% and 98%. The frame accuracy was between 87% and 98% for the GP feature set.

3.2. Phonological Feature Transformations

The estimates of posterior probabilities generated at the output of the TDNN feature classifiers for the three different feature systems are transformed and applied as input to the phonological feature based phoneme recognizers described in Section 3.3. However, these feature based observation vectors suffer from both high dynamic range and from highly correlated vector components, as is the case for filterbank energies in MFCC analysis. To compensate for these issues, logarthmic amplitude compression and feature space rotation is applied to these vectors.

To reduce the dynamic range of the phonological feature values, a logarithmic compression was applied to the posterior values at the output of TDNNs. For this to be effective, it was important to limit the minimum value for the non-linear compression to a value of approximately -10. One more important problem associated with all three sets of phonological feature vectors is the potential high correlation among individual outputs from the neural networks. Principal components analysis (PCA) was applied to estimating linear transformations that result in a transformed feature space where feature vectors are approximately uncorrelated. PCA was also used for dimensionality reduction for the multi-valued feature system. Since there are 28 total components corresponding to the output values for the MV feature set, the overall observation vector dimensionality would be 84 when these 28 components are concatenated with first and second order difference vectors.

The observation vectors associated with the SPE and GP feature systems have dimension 14 and 11 respectively. The outputs of the eight phonological feature TDNNs for the MV based feature system are concatenated to form a 28 dimensional vector. Only the first 13 principle components were retained after PCA analysis for this feature system. No dimensionality reduction was performed for the SPE or GP based observation vectors. Each of the resulting transformed vectors was then concatenated with first and second difference vectors to obtain 42, 33, and 39 components observation vectors for

SPE, GP and MV, respectively. These observation vectors are input to the HMM-based feature-to-phone mapping in Section 3.3.

3.3. HMM Based Feature-to-Phone Mapping

Continuous diagonal mixture Gaussian observation density HMM models were used to map from frame level phonological feature based observations derived from the discriminative networks described in Section 3.1 to phone sequences. This approach to defining HMM based acoustic models over observation vectors obtained from discriminative networks is very similar to the approach used in [12]. The three phonological feature based ASR systems were trained using the estimates of the posterior probabilities obtained from the feature based TDNNs described in Section 3.1 transformed as described in Section 3.2. The phone accuracy obtained for feature based and MFCC based systems is summarized in Section 6.

4. LATTICE RE-SCORING BASED FEATURE INTEGRATION

This section describes a lattice re-scoring method for integrating MFCC and phonological feature based models in a phone recognition task. It is made up of two parts. First, the lattice re-scoring strategy, as originally introduced in [2], is described as a general method for integration of multiple independent features with traditional MFCC based ASR. Second, a discriminative model combination algorithm is applied to estimating the relative weighting of multiple knowledge sources in the lattice re-scoring approach [13]. This method is used as an alternative to empirical estimation of these weights as was previously done in [2].

4.1. Feature Integration

In previous work, a strategy for integrating phonological feature based models with traditional MFCC based ASR was investigated. A simple model for decoding an optimum phone string from multiple independent phonological feature vectors was presented in [2]. The problem of decoding the optimum phone sequence, $\hat{F}_m$, for the mth utterance corresponds to optimizing

$$\hat{F}_m = \arg\max_{F_m} \left\{ p\left(F_m | X_m^0, \ldots, X_m^N, S_m\right) \right\}. \quad (1)$$

In Equation 1, F_m is assumed to be a phone string generating a continually varying sequence of articulatory states which gives rise to N phonological feature streams, $X_m^i, i = 1, \ldots, N$, where X_m^i corresponds to feature vectors generated for the mth utterance and the ith distinctive feature system. Each feature stream consists of a sequence of T vectors, $X_m^i = \{\vec{x}_m^i(1), \ldots, \vec{x}_m^i(T)\}$. X_m^0 is also defined to represent spectral energy based MFCC features and S_m is the surface acoustic waveform.

It was also shown that optimizing Equation 1 leads to finding the phone string that optimizes the following log linear combination

$$\log p_{\{\Lambda\}}\left(\boldsymbol{X_m}|F_m\right) = \lambda_0 \log p\left(X_m^0|F_m\right) + \lambda_1 \log p\left(X_m^1|F_m\right) + \ldots + \lambda_N \log p\left(X_m^N|F_m\right) + \lambda_{N+1} \log p\left(F_m\right). \quad (2)$$

Equation 2 is the log-linear combination of probabilities associated with multiple feature streams and the language model probability $p(F_m)$. $\boldsymbol{X_m} = \{X_m^0, X_m^1, \ldots, X_m^N\}$ represents the set of $N+1$ feature streams for utterance m. Although the weight values $\Lambda = \{\lambda_0, \lambda_1, \ldots, \lambda_{N+1}\}$ can be estimated empirically, a discriminative model combination approach is used here to estimate these weights automatically.

4.2. Discriminative Model Combination

One approach pursued in automatic speech recognition for optimum integration of multiple acoustic and language models is *discriminative model combination* (DMC). This approach has been applied in situations where the cost or likelihood of a system incorporating multiple models can be represented as one general log-linear posterior probability distribution [13].

In discriminative model combination, the coefficients Λ are optimized based on the decision error rate of the following discriminant function:

$$g\left(F_m, \hat{F}_m\right) = \sum_{j=0}^{N} \lambda_j \left(\log p\left(X_m^j|F_m\right) - \log p\left(X_m^j|\hat{F}_m\right)\right) + \lambda_{N+1} \log p\left(F_m\right) - \lambda_{N+1} \log p\left(\hat{F}_m\right). \quad (3)$$

Let $f\left(\mathscr{L}\left(k_{mr}, k_{m0}\right)\right)$ represent an "ideal" discriminant function, where $\mathscr{L}\left(k_{mr}, k_{m0}\right)$ is the *Levenshtein-distance* between the correct observation string k_{m0} and the competing strings $k_{mr}, r = 1, \ldots, K$, and $f()$ is a sigmoid function. The objective of the discriminative method is to minimize the mean distance between the discriminant function of the log-linear posterior probability distribution in Equation 3 and the ideal discriminant function [13]. This criterion can be optimized with respect to the log linear model coefficients, Λ, using training samples. This mean squared distance is given as

$$D(\Lambda) = \frac{1}{KM} \sum_{m=1}^{M} \sum_{r=1}^{K} \left(\log \frac{p_{\{\Lambda\}}\left(k_{m0}|\boldsymbol{X_m}\right)}{p_{\{\Lambda\}}\left(k_{mr}|\boldsymbol{X_m}\right)} - f\left(\mathscr{L}\left(k_{mr}, k_{m0}\right)\right) \right)^2, \quad (4)$$

where $p_{\{\Lambda\}}\left(k_{mr}|\boldsymbol{X_m}\right) \propto p_{\{\Lambda\}}\left(\boldsymbol{X_m}|k_{mr}\right) p\left(k_{mr}\right)$ and $p_{\{\Lambda\}}$ corresponds to the probability given in Equation 2. The first summation in the above equation is over all M utterances in the training set. The second summation is over the K most likely phone ystrings produced by the MFCC based ASR system.

In this paper, the discriminative model combination paradigm has been used for combining each of the HMMs defined over phonological feature based observations with the HMM model defined over MFCC observations. A set of K-best strings for $K = 10$ were generated from MFCC-based HMMs and used afterward in Equation 4 to compute the weight values Λ for each model combination.

In Section 6, the performance of these combined models are summarized and compared to the first-level HMMs defined in Section 3.3. The optimum phone sequences and lattices decoded by these new combined models are also used as another set of inputs to the system combination paradigm described in Section 5.

5. SYSTEM COMBINATION

System combination techniques have been widely used in LVCSR for combining output word strings or word lattices obtained from multiple ASR systems [4]. Significant reduction in composite word error rate (WER) is generally obtained relative to the individual ASR

system WER when the average individual system error rates are similar but the nature of the errors from the individual systems are different. The experimental study in Section 6 investigates the effect of applying this same class of system combination techniques as a means for integrating the multiple phonological feature based phone recognition systems described in Section 3. Two such techniques which are used for phoneme level system combination are ROVER [3] and confusion network combination (CNC) [4].

ROVER, as originally introduced by Fiscus [3], is performed in two stages. In the first stage, the decoded output strings produced by the individual speech recognition systems are aligned using dynamic programming. This is a sequential procedure that is initiated by first picking a reference string. Each remaining string is then aligned to this reference string, one at a time, until a single word transition network (WTN) is obtained. All arcs leaving a node, or alignment point, in this WTN have the same destination node. There is no optimum ordering of strings in this alignment process that is guaranteed to give the best results. The second stage consists of selecting the best scoring label at each alignment point through a voting process. While this voting process can be wighted by assigning confidence scores to each arc on the WTN, this was not done in the experiments described in Section 6.

CNC is an extension of ROVER where, instead of aligning the strings produced by the individual ASR systems in the first stage, confusion networks are aligned [4]. Confusion networks are a compact representation of lattices which are in turn graph representations of the search space of the recognizer containing the most likely word hypotheses. They maintain the ordering of the original contents of the lattices but have the structure of the WTN described above. CNC involves three steps. First, the lattices generated by the individual ASR systems are converted to confusion networks. Second, these networks are aligned in sequential order, using the confusion network of one of ASR systems as a reference. This results in a WTN corresponding to the combination of all confusion networks produced by the individual systems. Finally, this combined network is used for obtaining the highest scoring string using a modified voting procedure. This involves adding the posterior probabilities or, if available, confidence scores on the labels at each alignment point.

6. EXPERIMENTAL STUDY

An experimental study is presented which evaluates the performance of the two different strategies described in Sections 4 and 5 for integrating phonological feature based representations with MFCC based ASR. First, Section 6.1 begins by introducing the phoneme recognition task domain used for the experiments along with the phone recognition accuracies associated with the phonological distinctive feature based phone recognizers described in Section 3. Second, Section 6.2 presents the performance obtained by combining these feature based phone recognizers with a "traditional" MFCC based phone recognizer using the system combination techniques from Section 5. Finally, the performance of the lattice re-scoring approach is described in Section 6.3 where discriminative model combination (DMC) is used to estimate the weights of the log linear model for feature integration described in Section 4.

6.1. Feature Based Phoneme Recognition

The experiments described in this section were performed using the TIMIT speech corpus [10]. HMM acoustic models and TDNN based phonological feature detectors were trained from 3572 utterances taken from the TIMIT training set with a small 124 utterance development set held out for empirical estimation of the log linear weights, Λ, given in Section 4.1. All results is this section are reported as phone recognition accuracy (PAC) evaluated on the 1344 utterance TIMIT test corpus using the reduced phone set described in [14].

The phonological distinctive feature (PDF) based phone recognizers described in Section 3 consist of TDNN based feature detectors, log/PCA based transformation of feature detector outputs, and HMM based feature-to-phone mapping. Table 1 displays the phone accuracies obtained for the MV, GP, and SPE features. The HMM based component of all three systems includes context dependent tri-phone three state HMM phone models with continuous diagonal covariance Gaussian mixture observation densities containing 5 mixtures per state. HMM models in all three systems were trained using a maximum likelihood (ML) criterion. All of the PACs displayed in Table 1 can be compared to a PAC of 69.1% obtained for a similarly configured MFCC based HMM phone recognizer. The two rows in the table correspond to the case where PCA transformation alone is applied to the feature detector outputs and both log and PCA based transformations are used.

There are several observations that can be made from Table 1. First, the nonlinear amplitude compression performed by the log transformation results in significant improvement in phone accuracy when compared with systems implemented with the PCA transformation alone. This improvement only occurs when the log is performed with amplitude compression implemented to constrain the allowable range of feature detector output values to 10 dB. Second, while the GP based system obtains the best PAC at 68.1%, all three of the feature based phone recognizers obtain relatively similar performance. Third, the difference in phone accuracy between the best performing feature based system and the MFCC based system was only 1% absolute. This similarity in performance for the MV, GP, SPE, and MFCC based systems is important when combining these systems using the techniques described in Section 6.2.

Feature Based Phone Recognition Accuracy			
Feature Trans.	MV	GP	SPE
PCA	64.1%	66.1%	64.2%
log+PCA	66.9%	68.1%	66.5%

Table 1. PAC measured for HMM based phone recognition using three different feature sets with PCA based and PCA+log based transformations applied to feature detector outputs.

6.2. System Combination Performance

Given the similarity of the phone accuracies obtained from the MFCC and feature based phone recognizers shown in Table 1, it is likely that performance could be further improved through some combination of these systems. This is true, of course, only if the errors produced by these different systems are complementary in some way. By observing the level of improvements obtained using the ROVER and CNC based system combination techniques described in Section 5, it is possible to get an indication of the degree to which the different feature representations convey complementary information.

Table 2 shows the phone accuracies obtained for ROVER and CNC based system combination of the MFCC based ASR system with the GP, MV, and SPE based systems. For the ROVER based system combination, a single phone string was produced by each

system and the phone strings were successively aligned in the order shown in the table with the phone string produced by the MFCC system used as the reference string in the alignment process. The best phone label at each alignment point was chosen by a weighted voting procedure. For the CNC based system combination, a phone lattice was produced by each system, confusion networks were created from each lattice, and the confusion networks were successively aligned in the order shown in the table with the confusion network produced by the MFCC system used as the reference in the alignment process. The best phone label at each alignment point was chosen by summing the posterior probabilities for arcs containing a given phone label.

The ROVER performance in the first row of Table 2 displays the baseline MFCC PAC. The performance displayed for CNC in the first row of the table represents the PAC obtained using a modified word error rate optimization criterion [4]. It is clear from rows two through four of Table 2 that combining all three feature based systems with the MFCC based phone recognizer using ROVER results in a small but significant improvement in PAC. However, combining a single feature based system with the MFCC system using CNC results in a far greater improvement, with the MFCC+GP combination resulting in a 3.9% absolute improvement. Finally, the last two rows of Table 2 show that significant additional improvements in PAC are obtained as the additional feature based systems are combined using both ROVER and CNC.

Combination of Feature and MFCC Based Systems		
Combined Systems	PAC	
	ROVER	CNC
MFCC	69.1%	69.6%
MFCC + GP	70.0%	73.0%
MFCC + MV	69.9%	72.4%
MFCC + SPE	69.8%	72.2%
MFCC + GP + MV	72.9%	73.9%
MFCC + GP + MV + SPE	73.6%	**74.3%**

Table 2. Phone accuracies obtained for combination of three feature based phone recognition systems with the MFCC based system using both ROVER and CNC based techniques. Systems are combined in the order shown.

6.3. System Integration using Lattice Re-scoring

Table 3 displays the phone accuracy for systems implemented using the lattice re-scoring scenario originally introduced in [2] to optimize the weighted log-linear combination criterion given in Equation 2. Rows two through four in Table 3 display the phone accuracies obtained by integrating the MFCC based phone recognizer with each of the three phonological distinctive feature based systems. The interpolation weights in Equation 2 are estimated using discriminative model combination. Rows five and six in Table 3 display the phone accuracies obtained by combining multiple MFCC and feature based systems by sequential lattice re-scoring. At each stage of this sequential process, a phone lattice is produced and this phone lattice is then re-scored by the following feature based HMM system. The relative weights that are applied to the log likelihoods of each these multiple systems are also estimated using DMC.

There are several observations that can be made from Table 3 and from comparing the performance of the systems shown in Tables 2 and 3. It is clear from rows two through four of Table 3 that feature integration of a single PDF with MFCC based ASR results in substantial improvement in PAC with respect to the MFCC baseline system. Comparison with rows two through four of Table 2 shows that these improvements are similar to those obtained using CNC based system combination. It is also clear from rows five and six in Table 3 that feature combination performed through successive lattice re-scoring results in performance improvements that are similar to those obtained through CNC based system combination. All of these results suggest that these feature based systems convey complementary information which can significantly reduce that ambiguity associated with MFCC based ASR.

Lattice Re-scoring / DMC Feature and MFCC Integration	
Combined Systems	PAC
MFCC (Baseline)	69.1%
MFCC + GP (DMC)	72.9%
MFCC + MV (DMC)	72.7%
MFCC + SPE (DMC)	72.6%
MFCC + MV + GP (DMC)	73.3%
MFCC + MV + GP + SPE (DMC)	**73.8%**

Table 3. Comparison of PACs measured for different feature representations re-scoring MFCC lattices using DMC

The use of discriminative model combination was extremely important for estimation of the log linear weights, Λ, given in Section 4.1. Without DMC, it would be nearly impossible to obtain empirical estimates of these weights by exhaustively tuning them on a development set for each different experimental scenario. In order to demonstrate that there is no significant performance degradation associated with using DMC in estimating Λ, the performance of PDF feature integration for both DMC and empirical estimation of Λ is shown in Table 4. The second and third rows in Table 4 display the the PACs obtained for MV based feature integration with MFCC based ASR where only PCA based transformation was applied to the feature detector outputs [2]. In the second row, the parameters were empirically estimated from the development set and in the third row, the parameters were estimated using DMC. It is clear from the table that the phone accuracy obtained using DMC based estimation of the log linear weights is not significantly different from the phone accuracy obtained by empirical estimation of these weights.

PAC for Lattice Re-scoring	
System	PAC
MFCC (Baseline)	69.1%
MFCC + MV - PCA (Empirical)	72.3%
MFCC + MV - PCA (DMC)	72.4%

Table 4. PACs measured for systems integrating MFCC based and MV phonological feature based with empirical and DMC estimation of log-linear weights (No logarithmic amplitude compression was applied to MV feature values)

7. SUMMARY AND CONCLUSION

Two methods for integrating multiple phonological feature based phone recognizers with more "traditional" MFCC based phone recognition have been investigated. First, systems defined over separate feature representations were integrated with an MFCC based ASR

decoder through a lattice re-scoring approach. The system integration was based on weighted log linear combination of feature based and MFCC based likelihoods where the weights were estimated automatically using a discriminative model combination approach. Each of the three feature based systems provided similar improvements in phone accuracy of approximately 3.5% absolute when individually combined with MFCC based ASR through lattice re-scoring. When all three feature based systems were sequentially applied in a sequential lattice re-scoring scenario, a 4.7% increase in PAC was obtained. The second method proposed for integrating multiple phonological feature based systems with MFCC based phone recognition was based on a ROVER and CNC system combination paradigms. Integrating all feature sets simultaneously through CNC system combination resulted in an absolute increase of 5.2% in phone accuracy.

The performance improvements achieved by these two methods are significant for two reasons. First, it suggests that the phonological feature systems used here are not simply functionally equivalent representations leading to the same decoded phone sequences. In fact, based on the performance improvements obtained here, they indeed appear to be complementary in the information that they represent. Second, for both approaches, feature integration itself was performed without having to apply any domain knowledge or any structural considerations. Both the weight estimation in lattice re-scoring and confusion network construction and combination in CNC are entirely data driven.

8. ACKNOWLEDGMENTS

The authors would like to thank Simon King of the Center for Technology Research at the University of Edinburgh for his helpful advice in building phonological feature classifiers. All HMM training and recognition simulations were based on the HTK HMM Toolkit [15]. All CNC system combination experiments were performed with the help of the SRI LM Toolkit[16].

9. REFERENCES

[1] Simon King and Paul Taylor, "Detection of phonological features in continuous speech using neural networks," *Computer Speech and Language*, vol. 14, pp. 333–353, 2000.

[2] R.C. Rose and P. Momayyez, "Integration of multiple feature sets for reducing ambiguity in ASR," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Hawaii, April 2007, vol. IV, pp. 325–328.

[3] J. Fiscus, "A post-processing system to yield reduced word error rates: Recogniser output voting error reduction (ROVER)," *Proceedings 1997 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 347–352, 1997.

[4] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: Word error minimization and other applications of confusion networks," *Computer, Speech and Language*, vol. 14, no. 4, pp. 373–400, 2000.

[5] Noam Chomsky and Morris Halle, *The sound pattern of English*, Studies in language (New York, N.Y.). Harper & Row, New York, 1968.

[6] John Harris, *English sound structure*, Blackwell, Oxford, UK; Cambridge, Mass., 1994.

[7] P. Niyogi and P. Ramesh, "Incorporating voice onset time to improve letter recognition accuracies," *Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, vol. 1, pp. 13–16, March 1998.

[8] K. Kirchhoff, G. Fink, and G. Sagerer, "Conversational speech recognition using acoustic and articulatory input," *Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, pp. 1435–1438, June 2000.

[9] J. Morris and E. Fosler-Lussier, "Further experiments with detector-based conditional random fields in phonetic recognition," in *Proceedings of the International Conference on Acoustic, Speech, and Signal Processing*, Honolulu, Hawaii, 2007.

[10] W. Fisher, V. Zue, J. Bernstein, and D. Pallet, "An acoustic-phonetic data base," *Journal of the Acoustical Society of America*, vol. 81, pp. 92–93, 1987.

[11] N. Strom, "The NICO artificial neural network toolkit," http://nico.nikkostrom.com, 1996.

[12] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature stream extraction for conventional HMM systems," *Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, May 2000.

[13] P Beyerlein, "Discriminative model combination," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. ICASSP, May 1998, vol. I, pp. 481–484.

[14] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Transaction on Acoustics, Speech, and Signal Processing*, vol. 37, pp. 1641–1648, 1989.

[15] S.J. Young, "The HTK hidden Markov model toolkit: Design and philosophy," Tech. Rep., Cambridge University Engineering Department, Speech Group, Cambridge, 1993.

[16] Andreas Stolcke, "SRILM - an extensible language modeling toolkit," in *Proceedings of the International Conference on Spoken Language Processing*, 2002, pp. 901–904.