INTEGRATION OF MULTIPLE FEATURE SETS FOR REDUCING AMBIGUITY IN ASR

Richard Rose and Parya Momayyez

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

ABSTRACT

The main goal of this paper is to investigate the feasibility of exploiting the invariance properties associated with articulatory based acoustic features to reduce ambiguity in ASR search. A multivalued phonological feature set defined by King and Taylor is used along with a time delay neural network implementation of phonological feature detectors to produce eight independent phonological feature streams [1]. Hidden Markov models (HMMs) defined over these phonological feature streams are combined with HMMs defined over spectral energy based mel frequency cepstrum coefficient (MFCC) acoustic features through a lattice re-scoring procedure. It is shown that significant improvements in phone recognition accuracy are obtained for this combined system relative to phone accuracy obtained for MFCC based HMMs alone. A study is also performed to analyze the effects of uncertainty in phonological feature detection.

Index Terms— Speech Recognition, Acoustic Modeling, Phonological Features

1. INTRODUCTION

The notion of an acoustic event detection and evidence combination paradigm for automatic speech recognition has been pursued by many researchers [2, 3, 4, 5]. In the most general scenario for this framework, multiple phonological feature detectors would generate a posteriori probabilities for every possible feature and these many asynchronous events would be integrated with other lexical and linguistic knowledge sources to obtain a decoded result. This is a compelling paradigm for ASR which has been dominated for many years by systems relying on statistical models defined over spectral energy based observations. This paper, however, is concerned with the more modest goal of exploiting this additional acoustic evidence for reducing ambiguity in existing ASR decoders and developing an understanding of the level of useful information that can be derived from features that are more closely associated with speech production. A lattice re-scoring approach for integrating phonological features obtained from a set of neural network based feature classifiers with a MFCC based ASR decoder is presented.

The idea that phonological or distinctive features may exhibit less overall variability than spectral energy based features derived directly from the acoustic waveform has been motivated primarily by their association with articulatory events in speech production [1, 6, 7]. An articulatory event that is considered "critical" to the production of a particular phoneme exhibits less variability during the production of that phoneme than non-critical articulatory events [6]. The use of phonological features has the potential to exploit this phenomenon of articulatory invariance. There are, of course, many challenges. These include the difficulty of obtaining reliable estimates of phonological features that serve as observations in this articulatory space. Challenges also include defining phonological features whose statistical properties can be efficiently modeled and incorporated with other knowledge sources in ASR search.

Many researchers have investigated practical means for exploiting phonological distinctive features and articulatory knowledge in general for ASR. This includes the use of neural networks to predict phonological features for MFCC observations [4, 1], the use of HMM's to model articulatory phenomena [8], and the use of HMM's defined directly over phonological features [5, 3]. Niyogi and Ramesh investigated a framework where the events corresponded to acoustic distinctive features [2]. This involved separate detectors for each distinctive feature category and a mechanism for merging the evidence obtained from the distinctive feature detectors with N-best string candidates produced by a HMM based continuous speech recognizer. While only a single distinctive feature detector, voice onset time, was implemented in this scenario, it was shown to have the potential for reducing uncertainty in traditional ASR search by disambiguating very specific acoustic confusions. Li et al also investigated the use of phonological features for re-scoring n-best lists generated from an MFCC based ASR system [4].

The main goal of this paper is to investigate whether the invariance properties that are generally attributed to phonological features can really be exploited to reduce ambiguity in ASR search. A simple generative model is presented in Section 2 where the acoustic speech waveform is assumed to be generated from a word sequence both indirectly through independent streams of phonological features and directly as a sequence of MFCC observations. Phonological features are generated by time delay neural network based feature detectors following the approach of King and Taylor [1]. A system for integrating these phonological feature streams with MFCC based ASR according to the generative model of Section 2 is described in Section 3. Section 4 presents ASR results obtained for this approach and discussion of the relationship of this approach to other feature based approaches is provided in Section 5.

2. MODEL FOR FEATURE INTEGRATION

This section describes a simple model for decoding an optimum phone string from a sequence of independent phonological feature vectors. The model is used to motivate the lattice re-scoring based feature integration strategy described in Section 3. A generative model for speech recognition is assumed where a phone string, F, generates a continually varying sequence of articulatory states. It is assumed that these articulatory states give rise to a set of N phonological feature streams, X^i, $i = 1, \ldots, N$. Each feature stream consists of a sequence of vectors, $X^i = \{\vec{x}_1^i, \ldots, \vec{x}_T^i\}$ that are updated at fixed 10 msec. time intervals. The definition of the phonological classes that each feature, $\vec{x}^i$, represents is given in Section 3.1.

This work has been supported under NSERC Special Research Opportunities Program Number 307188-2004

A surface acoustic waveform, S, is generated from the sequence of articulatory states. An additional observation stream, X^0, is also defined to represent spectral energy based MFCC features resulting in a total of $N+1$ feature streams. The problem of decoding the optimum phone sequence, $\hat{F}$, corresponds to optimizing

$$\begin{aligned}\hat{F} &= \arg\max_F\{p(F|X^0,...,X^N,S)\} \quad (1)\\ &= \arg\max_F\{p(F,X^0,...,X^N,S)\}\\ &= \arg\max_F\{p(S|X^0,...,X^N,F)p(X^0,...,X^N,F)\}\\ &= \arg\max_F\{p(S|X^0,...,X^N)\\ &\qquad p(X^0|F),...,p(X^N|F)p(F)\} \quad (2)\end{aligned}$$

Equation 2 was obtained by assuming that the feature streams are conditionally independent given a phone sequence and the acoustic waveform, S, is not directly dependent on F. Each of the probabilities $p(X^i|F)$ represents the probability of the ith phonological feature stream for a given phone string. This is represented here by an acoustic HMM. While there are many different possibilities for integrating multiple feature streams, the initial model used here assumes that the phonological feature streams are simply concatenated. As a result, there are two HMMs. The first is defined over the MFCC observations, X^0, and the second is defined over eight concatenated observation streams, $X^1,...,X^8$ to be defined in Section 3.1. The probability, $p(F)$, represents the prior probability of generating the phoneme string. This is represented by a phone based bigram language model that will be used to constrain search within the MFCC based ASR system.

The probability, $p(S|X^1,...,X^N)$, in Equation 2 represents the probability of the acoustic waveform being generated from N feature streams. This provides a measure of the uncertainty associated with the estimation of the phonological feature vectors from speech. If $X^1,...,X^N$ are assumed to be independent, then

$$p(S|X^1,...,X^N) = \prod_{i=1}^{N} p(X^i|S)/p(X^i)p(S). \quad (3)$$

Each term, $p(X^i|S)/p(X^i)$, on the right side of Equation 3 corresponds to the posterior probability of feature stream i normalized by the prior for that stream. The probability of the MFCC observation stream, $p(X^0|S)$, can also be included in Equation 3. Non-trivial probability densities for spectral energy based observation vectors can be of particular interest in missing feature theory where models for noise corruption of spectral bands are incorporated directly into the observation distributions. However, it is assumed here to be uniformly distributed.

3. INTEGRATION BY LATTICE RE-SCORING

This section describes a system for integrating MFCC based and acoustic phonetic based features in a phone recognition task under the general modeling assumptions given in Section 2. It is made up of three parts. First, the phonological features used in the system are described. A set of neural network based classifiers that are used for computing the posterior probabilities given in Equation 3 are presented. Second, the use of HMM acoustic models for representing the likelihood for both spectral energy (MFCC) based features and the concatenated set of phonological features is discussed. Finally, the mechanism for integrating these features by re-scoring ASR lattices is described.

3.1. Phonological Features

The phonological features used here correspond to the multi-valued feature system used by King and Taylor [1]. Table 2 lists the eight features along with the possible values that can be assumed by each feature. Each of the eight features is modeled by a single layer time delay neural network (TDNN). The input to each network is a vector of twelve MFCC's along with their first and second differences. The outputs of each feature based TDNN correspond to the binary values given in Table 2. The NICO toolkit was used for all of the networks and back propagation training of all network parameters [9]. Neural network based classifiers for these features are used to generate the posterior probabilities given in Equation 3.

Phonological Feature Definition	
Feature Class	Values
centrality	central, full, undef.
continuant	continuant, non-continuant
front-back	back, front
manner	vowel, fricative, approximant, nasal
phonation	voiced, unvoiced
place	low, mid, high, palatal, corono-dental, labio-dental, labial, coronal, velar, glottal,
roundness	round, not-round
tenseness	lax, tense

Table 1. Definition of eight phonological features proposed by King and Taylor and the discrete values assumed by each feature [1].

3.2. HMM Models

Continuous diagonal mixture Gaussian observation density HMM models were used to build phoneme recognizers based on both MFCC observations and feature based observations. The MFCC based ASR system was based on thirteen component cepstrum vectors concatenated with the first and second difference cepstrum to obtain 39 dimensional observation vectors. A single phonological feature based ASR system was created using the estimates of the posterior probabilities obtained from the outputs of the feature based TDNNs described in Section 3.1.

There are several problems associated with using diagonal Gaussian HMMs to model observation vectors that are defined this way. A first problem is that the individual outputs are also potentially highly correlated. A second problem is dimensionality. There are 28 total components corresponding to the output values given in Table 1, resulting in an observation vector dimensionality of 84 when concatenated with first and second order difference vectors. Principal components analysis (PCA) was used to reduce the degree of correlation and to reduce the dimensionality of the phonological feature based observation vectors. The outputs of the 8 phonological feature TDNNs were concatenated to form a 28 dimensional vector. The covariance of the data was diagonalized using PCA, and the first 13 principle components were retained. This was then concatenated with first and second difference cepstrum to obtain a 39 component observation vector.

Both maximum likelihood and maximum mutual information training criteria were applied to training MFCC based and phonological feature based HMMs. It was originally thought that MMI training might hold a particular advantage for the phonological feature based observations due to the potential mismatch of these features to the structure of the HMMs. It will be shown in Section 4 that, while MMI training was able to improve performance for both systems

when only a small number of mixtures were used, a five mixture per state ML based system outperformed the equivalent MMI based system. As a result, all of the experiments described in Section 4 are based on ML trained HMMs. The prior phoneme probability, $P(F)$, was represented by a phone based bigram model estimated from the TIMIT training utterances.

3.3. Combining Models

The integration of MFCC and phonological feature based observations is performed by re-scoring lattices generated by the MFCC based HMM phone recognizer. This scenario provides the opportunity to observe the extent to which phonological feature based observations can be used to disambiguate errors in hypotheses generated by the more standard MFCC based ASR systems. The optimum phone string, $\hat{F}$, is obtained by optimizing the following score:

$$S = \lambda_1 \log p(X^1, ..., X^8|F) + \lambda_2 \log p(X^0|F) + \lambda_3 \log p(F) + \lambda_4 \sum_{i=1}^{8} \log \frac{p(X^i|S)}{p(X^i)}, \quad (4)$$

which is the log-linear combination of probabilities given in Equations 2 and 3. The first term in the Equation 4 corresponds to the acoustic log probability obtained from the HMM defined over the concatenated phonological feature based observations. The second term corresponds to the HMM log probability from the MFCC based HMM. The third term corresponds to the phonotactic language model. Finally, the last term in Equation 4 is derived from the TDNN based phonological feature detectors.

Each string hypothesis contained in the phone lattices produced by the MFCC based HMM has log probability $\lambda_2 \log p(X^0|F) + \lambda_3 \log p(F)$. These lattices are then re-scored using the phonological feature based HMM model to create a new lattice that incorporates the phonological feature log probabilities $\log p(X^1, ..., X^8|F)$. Finally, to incorporate the uncertainty associated with estimated phonological features, the arcs of these lattices are re-scored with the estimated log probabilities, $\log p(X^i|S)/p(X^i)$. The estimated posterior probability, $p(X^i|S)$, for each feature in Table 1 is obtained as the product of the estimated posterior probabilities that appear as the output values for the feature specific TDNN. The estimate of the prior probability, $p(X^i)$, is obtained simply from the relative frequency of the feature occurrence in the training corpus. All phone recognition results reported in Section 4 are obtained by ordering hypotheses using the scoring function in Equation 4.

4. EXPERIMENTAL STUDY

This section describes the experimental study performed to evaluate the lattice re-scoring approach to phonological feature integration described in Section 3. The Section consists of three parts. First, baseline ASR performance is presented on separate feature sets over a range of HMM model complexities and different model training criterion. Second, ASR results are presented for systems that optimize the criterion presented in Section 3.3 through a process of lattice re-scoring. The section concludes with a discussion of how improvements in the ability to detect phonological features may impact overall system performance in this framework.

4.1. Baseline System

All HMM acoustic models and TDNN based phonological feature detectors were trained from the TIMIT training utterances. A small development set was held out for empirical estimation of the weights, $\lambda_1, ..., \lambda_4$ in Equation 4. All phone recognition results are reported on the 1344 utterance TIMIT test corpus. The baseline phone accuracies (PAC) measured on the TIMIT test set for MFCC based and phonological feature (FEAT) based HMM ASR systems are summarized in Table 2. Varying levels of system complexity are represented in the table ranging from single mixture per state monophone context subword models to five mixture per state phonetic context clustered triphone models. The performance of systems configured using HMMs trained from maximum likelihood (ML) and maximum mutual information (MMI) based training criteria are also given in Table 2. The baseline phone accuracy (PAC) for a five mixture per state HMM model defined over MFCC observation vectors is shown in the table to be 69.1%. This is below the state of the art performance obtained for TIMIT, but is reasonable for an HMM system that has not been optimized for a phone recognition task.

There are several observations that can be made from Table 2. First, it is clear that the PAC for the best FEAT based ASR system is below that of the best MFCC system. However, the lower complexity FEAT system obtains far better PAC than the MFCC system with equivalent model complexity. This suggests that the use of phonological features can potentially provide an advantage over MFCC's, but these features are not at all well modeled by mixtures of diagonal Gaussians. Second, comparing the second and third rows of Table 2, it is clear that discriminative training of HMM parameters leads to a small but significant improvement in PAC for monophone models. There is no improvement, however, for triphone models. It appears that MMI training does have the potential for reducing the effects of mismatch between the data distribution and model structure in this context, but there are insufficient utterances in the TIMIT corpus for MMI training for all but the simplest HMM models.

Baseline Phone Accuracy (PAC)			
Obs.	Mono-1Mix	Tri-1Mix	Tri-5Mix
ML-MFCC	51.7%	63.8%	69.1%
ML-FEAT	59.4%	62.4%	64.1%
MMI-FEAT	60.9%	62.5%	64.5%

Table 2. PAC measured on the TIMIT phone set for ML and MMI trained HMMs with varying model complexity trained from MFCC coefficients and phonological feature (FEAT) based observations.

4.2. Combined System Performance

Table 3 displays the phone accuracy for systems implemented using a lattice re-scoring scenario to optimize the criterion given by Equation 4. The interpolation weights in Equation 4 are currently estimated empirically from a small 100 utterance development set.

The PAC in first row of Table 3 repeats the PAC for the five mixture per state context clustered triphone HMM system based on MFCC observations (M-HMM) shown in the first row of Table 2. The PAC displayed in the last row of Table 3 (M-HMM-LatMat) represents the best alignment of the reference string for each utterance with the phone lattice generated for the utterance. Therefore, this figure represents the best possible PAC obtainable through re-scoring of lattices generated by the MFCC based HMM recognizer.

The second row of Table 3 displays the PAC when the lattices from system M-HMM are re-scored using the FEAT observation based F-HMM system. A 3.1% absolute improvement in PAC is obtained. This is a significant performance improvement relative to the relative richness of the lattices as characterized by the LatMat performance. The third row of the table displays the performance ob-

tained when these lattices, after having been re-scored by the FEAT based HMMs, are then re-scored again to incorporate the probabilities obtained from the phonological feature detectors (D-RESC). No significant increase in performance was obtained using by including this information. This suggests that posteriors corresponding to feature detector activations carry little information about the underlying uncertainty associated with feature detection.

PAC for Lattice Re-scoring	
System	PAC
M-HMM	69.1%
M-HMM / F-RESC	72.2%
M-HMM / F-RESC / D-RESC	72.2%
M-HMM - LatMat	87.6%

Table 3. Phone accuracy for MFCC based HMM baseline ASR system (M-HMM), combined system with FEAT based HMM re-scoring (F-RESC), combined system re-scored with feature detector scores (D-RESC), and the lattice inclusion rate (LatMat)

5. DISCUSSION

The performance reported in Table 3 for the combined MFCC/FEAT systems is limited by many issues. A first limitation is the richness of the set of the hypothesized phone strings contained in the MFCC-based ASR lattices. This is characterized by the LatMat performance in Table 3 which shows that the best possible PAC obtainable by lattice re-scoring is 87.6%. A second limiting issue is the choice of the phone rather than, for example, the syllable as the fundamental acoustic unit for ASR.

A third limitation is the performance of the phonological feature detectors described in Section 3.1. To gain insight into the limitations posed by imperfect feature detectors, the combined M-HMM/F-RESC system in Table 3 was re-run assuming an ideal manner class feature stream. The performance of a TDNN based phonological feature detector is obtained by measuring the accuracy of the detector in classifying analysis frames against ideal feature labels obtained from human-labeled TIMIT transcriptions. This frame classification accuracy was found to range from 72.8% for the 10 element "place" feature to 92.2% for the 2 element "phonation" feature. An experiment was performed where the manner feature detector, which obtained a frame classification accuracy of 84.4%, was replaced by the ideal values taken from the human labeled training transcriptions. Lattice re-scoring was performed using the ideal values taken from the test transcriptions. When the F-HMMs were re-trained and lattice re-scoring was re-run using these ideal manner features, the PAC for the M-HMM/F-RESC system in Table 3 increased from 72.2% to 76.2%. This rather large increase in performance obtained from perfect knowledge of only one of the eight feature classes suggests that even incremental improvements in feature detection performance may have a significant impact on systems like the one descibed here.

6. SUMMARY AND CONCLUSIONS

A lattice re-scoring approach for integrating phonological feature streams obtained from TDNN based feature detectors with an MFCC based ASR decoder has been presented. An improvement in phone recognition accuracy of from 69.1% for an MFCC based system to 72.2% for the combined system was obtained on the TIMIT corpus. There are two major issues that have not been addressed in this work. A first issue arises from the fact that performance was strictly measured according to phone recognition accuracy. This does not allow for consideration of how incomplete phonological information associated with a given syllable may have different impact on word recognition accuracy depending on factors like the stress level associated with the syllable [10]. Another issue is that one has to be careful about making broad interpretations based on results measured on read speech corpora. This is especially true for the TIMIT corpus which was collected under controlled acoustic conditions and where human labeled phonetic boundaries are available for training phonological feature detectors. To address this issue, our goal is to extend this work to spontaneous speech corpora which have been derived from the conversational telephone speech Switchboard corpus [11].

7. ACKNOWLEDGMENTS

The authors would like to thank Simon King of the Centre for Technology Research at the University of Edinburgh for his helpful advice in building phonological feature classifiers.

8. 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] P. Niyogi and P. Ramesh, "Incorporating voice onset time to improve letter recognition accuracies," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 13–16, March 1998.

[3] Sebastian Struker, Florian Metze, Tanja Schultz, and Alex Waibel, "Integrating multilingual articulatroy features into speech recognition," *Proc. European Conf. on Speech Communications*, pp. 1033–1035, September 2003.

[4] Jinyu Li, Yu Tsao, and Chin-Hui Lee, "A study on knowledge source integration for candidate rescoring in automatic speech recognition," *Proc. Int. Conf. on Acoust., Speech, and Sig. Processing*, pp. 837–840, May 2005.

[5] 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.

[6] G. Papcun, J. Hochberg, T. R. Thomas, F. Laroche, J. Zacks, and S. Levy, "Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data," *J. Acoust. Soc. Am.*, vol. 92, no. 2, pp. 688–700, August 1992.

[7] R. C. Rose, J. Schroeter, and M. M. Sondhi, "Speech production models in automatic speech recognition," *J. Acoust. Soc. America*, vol. 99, no. 3, pp. 1699–1709, Apr. 1995.

[8] J. Sun, X. Jing, and L. Deng, "Data-driven model construction for continuous speech recognition using overlappling articulatory features," *Proc. Int. Conf. on Spoken Lang. Processing*, October 2000.

[9] N. Strom, "The NICO artificial neural network toolkit," http://www.speech.kth.se/NICO.

[10] Eric Fosler-Lussier, C. Anton Rytting, and Soundararajan Srinivasan, "Phonetic ignorance is bliss: Investigating the effects of phonetic information reduction on ASR performance," *Proc. European Conf. on Speech Communications*, pp. 1249–1252, September 2005.

[11] Simon King, Chris Bartels, and Jeff Bilmes, "Svitchboard 1: Small vocabulary tasks from Switchboard 1," *Proc. European Conf. on Speech Communications*, September 2005.