AN INVESTIGATION OF PHONOLOGICAL FEATURE SYSTEMS USED IN DETECTION-BASED ASR

I-Fan Chen and Hsin-Min Wang

Institute of Information Science, Academia Sinica, Taipei
{ifanchen, whm}@iis.sinica.edu.tw

ABSTRACT

In this paper, we study the effect of using different phonological feature sets for detection-based automatic speech recognition in phone recognition tasks. Three phonological feature sets derived from different underlying phonological theories are investigated. Our experiments were conducted on the TIMIT database. By comparing the oracle phone recognition results achieved by assuming that all the phonological features are correctly detected based on each feature set, we show that selecting an appropriate phonological feature set is crucial to the performance of detection-based ASR. The highly accurate oracle phone recognition results show that the performance of the CRF-based backend, which is commonly used in detection-based ASR, is very satisfactory. Comparison of the oracle phone recognition results and the real phone recognition results indicates that investigation of high-accuracy front-end detectors is a key issue in improving the performance of detection-based ASR.

Index Terms— Detection-based ASR, phonological feature system, result fusion, speech recognition

1. INTRODUCTION

Currently, detection-based automatic speech recognition (ASR) is a popular research topic in fields related to ASR. Because human beings often understand speech by integrating multiple knowledge sources from the bottom up, detection-based ASR systems attempt to reduce the gap between human speech recognition (HSR) and ASR by simulating the HSR mechanism. Conceptually, the framework of detection-based ASR can be divided into two key components: a front-end knowledge attribute detection process and a backend knowledge integration process [1]. The front-end process collects a wide variety of knowledge attributes related to speech to form the knowledge sources; and the backend process integrates the attributes into higher-level speech units, such as phones, syllables, words, and sentences.

Among the various speech knowledge sources, phonological features are used most frequently by detection-based ASR research groups [2][3][4][5]. However, although there are several phonological feature sets based on different linguistic theories, it is not clear whether any of them could be considered as the best design for the detection-based ASR task [6]. Therefore, instead of building a detection-based ASR system by randomly selecting one of the phonological feature sets, it would be better to consider each of them before constructing a detection-based ASR system.

In this paper, we investigate the use of the three phonological feature sets described in [7] for detection-based phone recognition, namely the Sound Pattern of English (SPE) feature set [8], the multi-valued (MV) feature set [7], and the Government Phonology (GP) feature set [9]. To build our detection-based ASR system, we choose time delay recurrent neural networks for front-end phonological feature detection and Conditional Random Fields (CRFs) [10][11] for backend phone recognition [3][4]. Our experiments were conducted on the TIMIT database. For each feature set, we compare the upper bounds of the phone recognition performance, which are derived by assuming that all the phonological features are correctly detected, and show that selecting an appropriate phonological feature set is crucial to the performance detection-based ASR. In addition, we compare confused phone pairs induced by using the three feature sets and find that they tend to complement each other. Therefore, we propose improving the recognition accuracy by using CRFs to combine the recognition results of different ASR systems. We believe that our experiment results provide further insight into how to design a high-performance detection-based ASR system.

2. FRONT-END PHONOLOGICAL FEATURE DETECTION

2.1. Phonological features

The Sound Pattern of English (SPE) phonological feature system proposed by Chomsky and Halle [8] is based on speech production. The system demonstrates the power of phonological features to transform complex phonological rules represented by phones into concise forms. The feature set contains 13 binary features, which represent speech production characteristics, e.g., anterior, nasal, round, strident, and voice characteristics. All the SPE features with an additional silence feature are shown in Table 1.

The multi-valued (MV) phonological feature system [7] is also production-based. The feature set contains six speech production features that are commonly used in phonological analysis, e.g., manner, place, and phonation. Although the MV feature system contains fewer features than the SPE feature system, each feature can take one of between 2 and 10 possible values. For example, the phonation feature can be either *voiced* or *unvoiced*, while the attributes of the place feature include *low*, *mid*, *high*, *labial*, *coronal*, *palatal*, *corono-dental*, *labio-dental*, *velar*, and *glottal*. Table 2 shows the feature list of the MV feature set.

In the Government Phonology (GP) feature set [9], unlike the above two production-based feature systems, sounds are destructed into a set of primes. Since the GP feature system is derived by examining the spectral properties of sounds, the phonological phenomena can be represented by fusing the primes structurally. For example, we can create /e/ by fusing **A** and **I**. In addition to simple fusion, we can use the most important prime in the fusion set as the head of the expression of the sound, which makes the GP feature system more expressive. However, the heavily structured representation demonstrates that

merely detecting the primes is not enough to identify sounds. To overcome this problem, King and Taylor added three primes as the head, namely a, i, and u [7], which are shown in Table 3 with the other primes used in the GP feature system.

Table 1. *The SPE feature set and the associated detection accuracy.*

Feature	Frame Acc	Feature	Frame Acc
Anterior	90 %	Nasal	97 %
Back	88 %	Round	94 %
Consonantal	90 %	Silence	98 %
Continuant	93 %	Strident	97 %
Coronal	89 %	Tense	90 %
High	88 %	Vocalic	87 %
Low	93 %	Voice	92 %

Table 2. *The MV feature set and the associated detection accuracy.*

Feature	Frame Acc	Feature	Frame Acc
Centrality	84 %	Phonation	91 %
Front back	82 %	Place	71 %
Manner	85 %	Roundness	91 %

Table 3. *The GP feature set and the associated detection accuracy.*

Feature	Frame Acc	Feature	Frame Acc
A	85 %	H	93 %
I	90 %	N	97 %
U	86 %	a	96 %
E	86 %	i	94 %
S	91 %	u	95 %
h	95 %		

2.2. Phonological feature detection

Following the work in [7], we use time delay recurrent neural networks [12] for phonological feature detection. For the SPE and GP feature sets, a single neural network with multiple outputs is trained to extract all features simultaneously; however, for the MV feature set, each feature is detected using an individual multi-output neural network with different numbers of hidden nodes. The inputs of all the neural networks are 12 Mel-frequency cepstral coefficients (MFCCs) plus energy, which are extracted by a 25-ms Hamming-windowed frame with 10-ms frame shifts. The outputs of the neural networks are real values ranging from 0 to 1, which can be treated as the posterior probabilities of the features. The phonological feature detection results are then obtained by applying a hard decision process to the neural networks' outputs[1]. The detection accuracies of the three phonological feature sets are shown in Tables 1, 2, and 3, respectively.

3. BACKEND PHONE RECOGNITION USING CONDITIONAL RANDOM FIELDS

Conditional Random Fields (CRFs) are undirected graphical models [10][11]. In a CRF, given an input sequence, the conditional probability of an output sequence is proportional to the product of the potential functions on cliques in the graph according to the Hammersley-Clifford theorem. Since the CRF framework directly models the posterior probability of a target sequence given a sequence of observations, it is also a kind of discriminative model. The flexibility to handle a wide variety of arbitrary and non-independent features makes the framework a powerful tool for the sequence to sequence assignment task. For example, CRFs are now widely used in natural language processing tasks, such as POS tagging and text chunking. Among the family of CRFs, the most widely used model is the first-order chain CRF model, which is expressed as

$$p(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_i \left(\sum_j \lambda_j s_j(y_i, \mathbf{x}) + \sum_k \mu_k t_k(y_{i-1}, y_i, \mathbf{x}) \right) \right), \quad (1)$$

where x and y are the observation and output sequences, respectively; Z(x) is the normalization term; and i is the index of the current position of the output sequence. In Eq. (1), there are two kinds of feature functions: the state feature function $s_j(\ )$ and the transition feature function $t_k(\ )$; λ_j and μ_k are weighting factors, which are the parameters of the model to be trained.

The state feature function only has a non-zero value when the current label y_i and the observation sequence **x** match some specific evidence. For example, the state feature function for a phone /ix/ might be defined as:

$$s(y_i, \mathbf{x}) = \begin{cases} 1, & \text{if } y_i = \text{/ix/}, \text{voice}(x_i) = true, \text{and vocal}(x_{i-1}) = false \\ 0, & \text{otherwise} \end{cases}. \quad (2)$$

The value of the function is 1 when the current observation contains a "*voice*" attribute and the preceding observation does not contain a "*vocal*" attribute; otherwise, it is 0. Note that the evidence of feature functions in CRFs is not restricted to the current observation, as the whole observation sequence can be considered. This capability of using longer observations over time is one advantage of CRFs over HMMs.

The transition feature function is very similar to the state feature function; however, it considers the preceding label and the current label simultaneously.

3.1. PHONE RECOGNITION USING CRFS

In the phone recognition task, the observation sequence x is comprised of frame-based phonological features detected by the front-end neural networks, and the target label sequence y is a frame-based phone sequence. For each of the three phonological feature sets studied in this paper, we build a CRF model to map a frame-based phonological feature sequence to its associated frame-based phone sequence. Then, the frame-based phone sequence is merged into a phone sequence as the final phone recognition output. The state feature functions in CRFs are set to $s(y_i, x_{i-1}, x_i, x_{i+1})$, in which the preceding, current, and subsequent phonological feature observations are considered simultaneously. The transition feature functions are set to label bi-grams, i.e., $t(y_{i-1}, y_i)$. We use the CRF++ toolkit [13] to implement CRFs.

3.2. RESULT FUSION USING CRFS

Since the CRF framework assigns a label sequence to an observation sequence, it is reasonable to use it to combine the recognition results of different ASR systems. In a practical implementation, the observation sequence of the CRFs used to fuse the results is comprised of the frame-based phone sequences generated by a set of ASR systems, while the target label sequence is the final frame-based phone sequence. The state feature functions of these fusion CRFs only relate to the observation sequence and label of the current frame, i.e., $s(y_i, x_i)$, and the transition feature functions are implemented as label bi-grams, i.e., $t(y_{i-1}, y_i)$.

[1] Because the CRF++ toolkit used in this study to build the backend CRF-based phone recognition system does not support real valued features, the neural network output is discretized to 0 or 1.

To train the fusion CRFs, we combine multiple systems' recognition results for the development set with the reference labels to compile the training dataset. Unlike other system combination methods, such as ROVER [14] or CNC [15], which use aligned word/phone slots for fusion, we perform phone fusion at the frame level for simplicity. The fused result, which is a sequence of frames with tagged phone labels, is merged into a phone sequence as the final output

4. EXPERIMENTS

Our experiments were conducted on the TIMIT acoustic-phonetic continuous speech corpus, but the dialect utterances (SA1 and SA2) were not used. The database is divided into three parts: the training set (3296 utterances), the development set (400 utterances), and the test set (1344 utterances). The training and development sets are subsets of the TIMIT suggested training set. In the experiments, all the models are trained by the training set, and the configurations and parameters associated with the models are empirically assigned based on the development set. The set of 61 TIMIT phones are used as the recognition units. However, for the performance evaluation, the recognized TIMIT 61-phone results are mapped to the CMU/MIT 39-phone set [16].

In the first experiment, we want to determine whether the three phonological feature sets (SPE, MV, and GP) have the potential for detection-based ASR. We assume that all the phonological features are correctly detected; thus, the phone recognition results are in fact the upper bounds of using the three phonological feature sets in detection-based ASR[2]. Table 4 shows the oracle phone recognition results obtained in this way, where Corr (correction rate) and Acc (accuracy) are obtained by HTK's HResults tool and Acc = Corr – insertion rate. From Table 4, we observe that the GP feature set demonstrates higher potential for detection-based ASR than the other two feature sets, even though it only contains 11 distinct features. In contrast, the MV feature set performs the worst despite the fact that the total number of possible values associated with its 6 features is 24. The results show that a well defined phonological feature set is crucial to the performance of detection-based ASR. However, they do not indicate that the GP feature set is more suitable for detection-based ASR, since its features might be much more difficult to detect compared to those of the other feature sets.

In the second experiment, the inputs of the CRFs are the phonological features detected automatically by the neural network detectors. We used the conventional context-independent HMM-based phone recognizer, with 3 states per HMM and 16 Gaussian mixture components per state designed by HTK, as the baseline. The acoustic feature vector for the baseline system contains 39 components: 12 MFCCs plus energy and their first and second time derivatives. The results are shown in Table 5. First of all, the CRFs trained in the first experiment were used. These models are denoted as OT, the oracle-data trained CRFs, since they were trained with the oracle phonological features converted from the manual phone labels of the training data. Once again, the GP feature set is superior to the other two feature sets. However, although the phone correction rate of the GP-based recognizer is as good as that of the baseline HMM-based recognizer, it suffers from high insertion errors. This problem might be due to mis-classification by the front-end phonological feature detection process and the lack of features' asynchronous information in the oracle training data. Since our current backend CRFs can only accept simple binary-valued state and transition feature functions, the hard decision process, which maps the posterior probabilities to 0 or 1 according to a pre-defined threshold, could amplify the errors generated by the neural network detectors. Using CRFs that accept real-valued feature functions fed with the posterior probabilities generated by the front-end detectors might solve this problem to some extent. Another way to solve this problem is to train the CRFs directly on the front-end detection results of the training data, so that the CRFs can learn the errors of the front-end detector and reduce the mismatch between training and testing. In [3] and [4], the CRFs were also trained in this way. The results in the rows denoted as DT in Table 5 show that the accuracies are substantially improved because the insertion rates are significantly reduced whereas the correction rates only drop a little bit. Although the detection-based ASR systems still perform worse than the HMM-based recognizer, the former might become comparable to the latter when real-valued CRFs are used.

Table 4. *The oracle phone recognition results derived by using different phonological feature sets.*

	Corr (%)	Acc (%)
SPE	93.28	93.20
MV	88.75	88.56
GP	98.39	98.36

Table 5. *The real phone recognition results derived by different recognizers, where OT means using oracle-data trained CRFs and DT means using detected-data trained CRFs.*

		Corr (%)	Acc (%)
HMM-based		69.02	63.45
OT detection-based	SPE	66.19	29.68
	MV	59.24	30.33
	GP	69.03	31.38
DT detection-based	SPE	56.56	55.27
	MV	51.84	50.68
	GP	55.74	54.53

Following the results of the first experiment, we analyze the characteristics of the three phonological feature sets in terms of their phonetic confusion pairs. The phonetic confusion property can be considered as a measure of an ASR system's resolution in the phonetic space. Table 6 shows the confusion sets associated with the three phonological feature sets identified from the oracle phone recognition results of the first experiment. It is clear that the GP feature set achieves the best resolution in the phonetic space. Moreover, it is noteworthy that since the confusion sets of the three feature sets are quite different, they may be complementary. Based on this assumption, we can expect that combining the three ASR systems will achieve better results. Therefore, in the last experiment, we use CRFs to combine the recognition outputs of different detection-based ASR systems in Table 5. The results are shown in Table 7. Although the DT CRF models outperform the OT CRF models in the single system case, combining DT detection-based systems does not improve the performance. In contrast, the combination of the OT detection-based systems substantially improves the recognition performance over that of each detection-based ASR system using only a single phonological feature set. Moreover, if we combine the output of the HMM-based recognizer with outputs of the three OT detection-based

[2] However, we should note that since we generate our oracle phonological features from TIMIT phone labels with their time boundaries, it lacks the asynchronous phenomenon which usually happens between phonological features.

recognizers (although this is not our goal here), we observe an absolute 0.86% improvement in accuracy over that of the HMM-based recognizer. Our results might indicate that the OT CRF models are better than the DT CRF models in extracting specific information contained in different phonological feature sets, and thus are more suitable for system combination.

Table 6. *Confusion pairs identified from the oracle phone recognition results by using different feature sets.*

Feature Set	#pair	Top 5 most confused pairs and their frequency counts
SPE	38	(iy,dh):1809 (z,aw):1236 (p,ey):956 (m,en):939 (f,v):911
MV	59	(iy,ih):995 (s,sh):395 (er,ah):394 (ey,iy):371 (ae,ah):315
GP	14	(el,sil):163 (uh,ah):126 (w,uw):64 (y,ih):39 (ah,sil):6

Table 7. *The real phone recognition results of the combined recognizers*.

Method	#sys	Corr (%)	Acc (%)
Baseline HMM	1	69.02	63.45
OT: SPE+MV+GP	3	61.97	**60.65**
DT: SPE+MV+GP	3	52.90	52.06
OT+DT: SPE+MV+GP	6	60.81	59.20
OT: SPE+MV+GP plus HMM	4	65.53	**64.31**
DT: SPE+MV+GP plus HMM	4	59.57	58.64
OT+DT: SPE+MV+GP plus HMM	7	64.22	62.59

5. CONCLUSION AND FUTURE WORK

We have investigated the use of three phonological feature sets for detection-based ASR. The oracle phone recognition results show that the GP feature set outperforms the other two feature sets and the upper bound phone accuracy of using this feature set in detection-based ASR is 98.36%. However, the real phone recognition results show that the three compared feature sets perform comparably to each other. In our experiments, the CRFs for backend phone recognition were trained with the oracle phonological features converted from the manual phone labels of the training data (denoted as OT) or the automatically detected phonological features of the training data (denoted as DT). The OT detection-based ASR systems based on the three feature sets achieve similar correction rates compared to the HMM baseline system, but they suffer from serious insertion errors, while the DT detection-based ASR systems perform slightly worse than the HMM baseline system. This might be due to the binary state feature functions of our backend CRF models. We find that the three compared feature sets complement each other by comparing their phonetic confusion sets. Based on this observation, we proposed using CRFs to combine the detection-based ASR systems based on the three phonological feature sets. Although the current accuracies of various detection-based ASR systems are not satisfactory, our initial study indicates the potential for detection-based ASR.

In our future work, we will change the binary-valued state feature functions of the backend CRFs to real-valued ones. We expect that, to some extent, this will improve the performance of both the OT detection-based ASR systems and the DT detection-based ASR systems. We will also try to introduce the asynchronous information into our oracle articulatory feature labels of the training data to see the importance of this knowledge to the performance of a detection-based ASR system. In addition to the backend issues, comparison of the results in Table 4 and Table 5 shows that there is still a large gap between the ideal and real performance of our three detection-based ASR systems. Thus, how to improve the accuracy of front-end detectors is a major issue. A possible research direction could use a group of models, such as SVMs, HMMs, and MLPs, to detect phonological features simultaneously and fuse their detection results. We believe that, with appropriate feature sets and accurate detectors, detection-based ASR is a promising technique for improving the recognition accuracy and resolving the bottleneck problem of current ASR systems.

6. ACKNOWLEDGEMENTS

This work was supported in part by the National Science Council, Taiwan, under Grants: NSC 96-2221-E-001-003 and NSC 96-3113-H-001-012.

7. REFERENCES

[1] C.-H. Lee, M.A. Clements, S. Dusan, E. Fosler-Lussier, K. Johnson, B.-H. Juang and L.R. Rabiner, "An Overview on Automatic Speech Attribute Transcription (ASAT)", in Proc. InterSpeech2007.

[2] C.-Y. Lin and H.-C. Wang, "Attribute-based Mandarin Speech Recognition using Conditional Random Fields", in Proc. InterSpeech2007.

[3] J. Morris and E. Fosler-Lussier, "Combining Phonetic Attributes using Conditional Random Fields", in Proc. InterSpeech2006.

[4] J. Morris and E. Fosler-Lussier, "Further Experiments with Detector-Based Conditional Random Fields in Phonetic Recognition", in Proc. ICASSP2007

[5] I. Bromberg, Q. Fu, J. Hou, J. Li, C. Ma, B. Matthews, A. Moreno-Daniel, J. Morris, S.M. Siniscalchi, Y. Tsao and Y. Wang, "Detection-Based ASR in the Automatic Speech Attribute Transcription Project", in Proc. InterSpeech2007.

[6] P. Momayyez, J. Waterhouse and R. Rose, "Exploiting Complementary Aspects of Phonological Features in Automatic Speech Recognition", in Proc. IEEE Automatic Speech Recognition and Understanding Workshop, 2007

[7] S. King and P. Taylor, "Detection of Phonological Features in Continuous Speech using Neural Networks", Computer Speech and Language, vol. 14, pp. 333-353, 2000.

[8] N. Chomsky and M. Halle, The Sound Pattern of English. Harper & Row. New York: 1968.

[9] J. Harris, English Sound Structure, Blackwell, 1994.

[10] J. Lafferty, A. McCallum and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", in Proc. ICML2001.

[11] F. Sha and F. Pereira, "Shallow Parsing with Conditional Random Fields", in Proc. HLT-NAACL2003.

[12] N. Strom, "The NICO Artificial Neural Network Toolkit", http://nico.nikkostrom.com

[13] T. Kudo, "CRF++: Yet Another CRF Toolkit", http://crfpp.sourceforge.net

[14] J. Fiscus, "A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)", in Proc. IEEE Automatic Speech Recognition and Understanding Workshop, 1997.

[15] 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, pp. 373-400, 2000.

[16] K.F. Lee and H.W. Hon, "Speaker-independent Phone Recognition using Hidden Markov Models", IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. 37, no. 11, pp. 1641-1648, 1989.