A PHONETIC FEATURE BASED LATTICE RESCORING APPROACH TO LVCSR

Sabato Marco Siniscalchi[1], Torbjørn Svendsen[1], and Chin-Hui Lee[2]

[1]Department of Electronics and Telecommunications
Norwegian University of Science and Technology, Trondheim, Norway
[2]School of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, GA 30332 USA
{marco77, torbjorn}@iet.ntnu.no, chl@ece.gatech.edu

ABSTRACT

Large Vocabulary Continuous Speech Recognition (LVCSR) systems decode the input speech using diverse information sources, such as acoustic, lexical, and linguistic. Although most of the unreliable hypotheses are pruned during the recognition process, current state-of-the-art systems often make errors that are "unreasonable" for human listeners. Several studies have shown that a proper integration of acoustic-phonetic information can be beneficial to reducing such errors. We have previously shown that high-accuracy phone recognition can be achieved if a bank of speech attribute detectors is used to compute a confidence score describing attribute activation levels that the current frame exhibits. In those experiments, the phone recognition system did not rely on the language model to follow their word sequence constraints, and the vocabulary was small. In this work, we extend our approach to LVCSR by introducing a second recognition step during which additional information not directly used during conventional log-likelihood based decoding is introduced. Experimental results show promising performance.

Index Terms— Detectors, speech recognition, neural networks.

1. INTRODUCTION

The classical way of formulating automatic speech recognition (ASR) is as a statistical pattern classification problem [1]. In this approach, hidden Markov models (HMMs) have been the dominating technique for acoustic modeling, and knowledge about the language has usually been represented in terms of of N-gram language models. The parameters of these models are estimated from large amounts of task-specific speech and text examples; therefore, either a lack of data or a mismatch between training and testing conditions usually causes a severe reduction in accuracy, and the ASR system performs much worse than human speech recognition (HSR).

In the recent past, several speech scientists have advocated that a proper integration of some knowledge sources into standard ASR systems may be useful to bridge the gap between the performance of the ASR system and HSR, on the same task. Although different knowledge-based approaches have been proposed, most of them try to integrate phonetically motivated information into the design process. For example, speech knowledge represented by phonetically motivated acoustic parameters [2, 3] has been embedded into an HMM-based recognizer at the front-end level. In [4], and [5], articulatory knowledge is integrated at the HMM state level. In [6], a set of artificial neural networks (ANNs) is used to score articulatorily motivated features for manner and place of articulation. The posterior feature probability outputs from each network are fed into an ANN that generates phoneme probabilities. This ANN is used as an emission probability estimator in the HMM framework. In [7], the classifier outputs are used to directly train a standard HMM-based system producing an error pattern that differs from a conventional cepstra-based system, but no improvement is reported when system combination is carried out. In [8], three landmark-based speech recognizers are proposed and differ by the pronunciation model used. The leading idea in the designing process of the ASR systems is to implement the new theories of nonlinear phonology, articulatory phonology, and landmark-based speech perception in the form of an ASR system, along with the use of high-dimensional speech features.

We have previously shown that high-accuracy phone recognition can be achieved for different languages when articulatory motivated features are generated using a bank of speech attribute detectors [9, 10]. We have also introduced an ad-hoc rescoring technique to integrate articulatory based scores into word lattices [11], and promising results have been obtained for small vocabulary continuous speech recognition tasks. An aim of this paper is to extend our work on phone recognition [9] to LVCSR and show that direct integration of speech knowledge, such as articulatory based scores, can help remove some "unreasonable" errors which may be obvious for a human listener but difficult to correct with more sophisticated acoustic and language models. As in our previous work, the link between the observed acoustic evidence and linguistics is established by the design of a bank of phonetic feature detectors. Each detector computes a score describing an activation level of the specified speech attributes, such as frication, voicing, etc., that the current frame exhibits. These cues are combined by an event merger that provides some evidence about the presence of a higher level feature (e.g., phone) which is then used during a second pass decoding process (or rescoring) implemented as in [11] but where the language model contribution is now considered. Experimental results on the Wall Street Journal (WSJ0) corpus demonstrate the effectiveness of the proposed approach for several ASR baseline configurations which differ in the way the acoustic model is trained and in the language model adopted during the decoding step. While delivering the best performance on the WSJ0 corpus is not a key issue of the proposed study, it is a goal of this work to show that the proposed technique can correct errors in LVCSR tasks due to constraints imposed by the lexical and language models that sometimes produce recognition results that override the underlying acoustic phonetic constraints. Posteriogram plots of the speech detectors are also shown to support this claim.

The rest of the paper is organized as follows. Section 2 describes the overall systems in more detail. In particular, the articulatory fea-

tures used are presented, information about the detection modeling and process is given, and the event merger is described. The rescoring procedure is outlined in Section 3. Section 4 describes the experimental results, and Section 5 concludes the paper.

2. SYSTEM OVERVIEW

The overall recognition system consists of two main parts: (1) a word recognizer that provides word lattices, and (2) a module that provides phone posterior probabilities needed during the rescoring step. The former is a LVCSR system designed using the HTK toolkit[1]. The latter is the combination of a bank of speech phonetic feature detectors and non-linear discriminant functions (ANN). The overall system is a two-stage speech decoder (Figure 1). The set of phonetic features and a more detailed description of the speech detectors and the event merger are presented in the following sections.

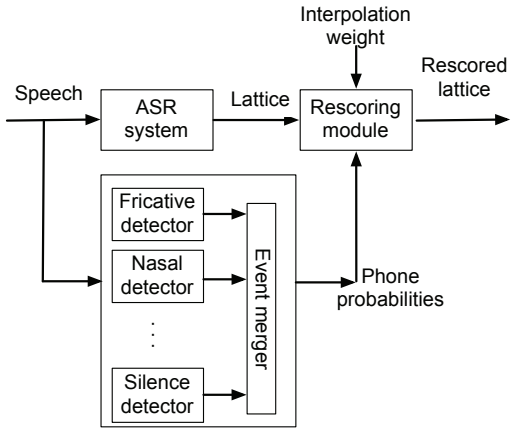

Fig. 1. Two-stage LVCSR decoder.

2.1. Phonetic Features

Table 1 lists the set of phonetic features used in our experiments. This set of 22 features is adopted from [9], and it is a combination of the Sound Pattern of English (SPE) features defined by Chomsky and Halle [12], and the manner and place of articulation features used in [11].

List of phonetic features
fricative glide liquid nasal stop vowel coronal dental glottal high labial low mid retroflex velar anterior back continuant round tense voiced silence

Table 1. *The set of speech features to be detected. This set is a combination of manner and place of articulation and SPE features.*

It is well known that vowels and consonants cannot be mapped into a common linguistic space because place of articulation has been differently defined for them. To circumvent this issue, we follow [13] and force vowels and diphthongs to be organized into the same place classes as the consonants. Also, we consider all articulatory features as binary during the training phase, although some of them take on non-binary discrete and continuous values.

[1]HTK toolkit, http://htk.eng.cam.ac.uk/

2.2. Speech Event Detectors & Event Merger

Each speech detector analyzes the input speech signal and produces a posterior probability that some acoustic-phonetic attribute is present during the frame being processed. Each detector is implemented with 3 feed-forward ANNs with one hidden layer and 500 hidden nodes organized as in [9]. The softmax activation function is used in the output layer. Energy trajectories in mel-frequency bands that are organized in split-temporal context [14] are used as parametric representations of the speech signal. To learn the parameter of each detector, the training data is separated into feature present and feature absent regions for every articulatory event of interest using the available phonetic transcription.

The event merger combines the outputs of the event detectors using different weights, and it delivers evidences at a phone level. Therefore, a single feed-forward ANN with one hidden layer and 800 hidden nodes is used. The activation function at the output layer is the softmax function.

The parameters of both the bank of speech detectors and the event merger are estimated on the training data from the TIMIT corpus [15]. Actually, these two blocks were trained at the time of our studies on high-accuracy phone recognition [9], and they have not been retrained for our experiments on the WSJ0 corpus. This avoids the cumbersome phase typical of state-of-the-art ASR systems, which need to be trained from scratch or modified through adaptation of the acoustic models every time that the speech task changes.

3. LATTICE RESCORING ALGORITHM

Lattice Rescoring is used as a mechanism to integrate articulatorily motivated knowledge into the ASR system. First, a speech decoder generates a collection of competing speech hypotheses. It is then followed by a rescoring algorithm to re-rank these hypotheses by incorporating additional information not directly used in the decoding process.

The lattice structure adopted in the proposed work reflects the syntactic constraints of the grammar used during recognition and is implemented as a direct, acyclic, and weighed graph, *G(N, A)*, with *N* nodes and *A* arcs. The timing information is embedded in the nodes (i.e., temporal boundaries are given by the arcs's bounding nodes); whereas the arcs carry the symbol along with the score information. In particular, each arc corresponds to a recognized word.

The rescoring algorithm proposed for acoustic evidence in this work incorporates scores generated by the evidence merger into the speech lattice, and it is inspired by the decoding scheme based on a generalized confidence score proposed in [16]. Generally speaking, combining independent sources of information can be carried out by

$$p(o_t|\Lambda) = C \prod_{i=1}^{N} p(o_t|\Lambda^i)^{\alpha^i}, \quad (1)$$

where Λ^i represents the set of acoustic parameters for the *i*-th system, Λ is the set of acoustic parameters of the combined system, $p(o_t|\Lambda^i)$ are different independent sources, C is a normalization constant, and α^i is the i-th interpolation weight. In the log-space, the above multiplication of exponentially weighted terms becomes a weighted sum.

In our experiment, we assume C equal to 1. Moreover, the weighted sum is carried out on an arc-by-arc basis using the segmentation generated by the word decoder as a linear combination of the log-likelihood acoustic score of each arc and the logarithm of

the event merger associated output that corresponds to the arc phone label.

3.1. Rescoring Formulation for Word Lattices

Each arc in a lattice corresponds to a word in a string hypothesis. A word (arc) score, W_n, is computed as follows,

$$W_n = \sum_{i=1}^{K} PS_n^i. \quad (2)$$

where PS_n^i is the sum of the logarithm of the frame-level phone probabilities generated by the event merger in correspondence to the *i*-th phone in the *n*-th arc, and K is the number of phones in the word associated with the *n*-th arc.

The weighted rescoring formula is finally defined as

$$S_n = w_w\, W_n + w_l\, L_n, \quad (3)$$

where w_l is the interpolation weights of the log-likelihood score, and and w_w is the interpolation weights of the word-level score W_n. During the search of the best path in the lattice, the new S_n acoustic score is combined with the language model score. Finally, it may be worth pointing out that no sequential information of phonemes in a word is used to generate the PS_n^i.

4. EXPERIMENTS

In the following sections, we present the experimental setup and discuss the results.

4.1. Experimental Setup

All experiments were conducted on the 5,000-word speaker independent Wall Street Journal (5k-WSJ0) corpus. The acoustic model parameters were estimated using training material from the SI-84 set (7077 utterances from 84 speakers, i.e., 15.3 hours of speech material). The testing material was the Nov92 evaluation set (330 utterances from 8 speakers). As already stated in Section 2.2, the set of speech detectors and the merger are those used for our phone recognition experiments [9]. Therefore, these two blocks were trained on the training material provided with the TIMIT corpus, and they were not further trained for this study.

For our experiments, four different gender independent LVCSR baseline systems using different acoustic and language models were built. All four systems were designed with the HTK toolkit. The first system was based on tied-state cross-word triphone models trained by Maximum Likelihood Estimation (MLE) and a bigram language model. The second system employed tied-state cross-word triphone models trained by MLE and a trigram language model. The third system used tied-state cross-word triphone models trained by Maximum Mutual Information (MMI) and a bigram language model. The fourth system was based on tied-state cross-word triphone models trained by MMI and a trigram language model. Closed vocabulary language models for the 5k-WSJ0 vocabulary were used during decoding. In all of the above systems each HMM has 3 states with 8 Gaussian mixture components per state. Finally the acoustic vector contains 12 MFCCs, log energy, velocity, and acceleration coefficients.

4.2. Experimental Results

The performance of all four systems are summarized in Table 2, and they are comparable with the results reported in [17]. These baselines are also comparable in terms of Word Error Rate (WER) with more recently reported results (e.g., [18], and [19]). Other studies show better results than the proposed baseline systems by using different setups (e.g. [20]). Those baseline systems were not available to us; therefore, the attempt was to improve over available baseline systems.

Table 2. Word Error Rate For Nov92 task

	Bigram	Trigram
Baseline (MLE)	7.32%	5.06%
Rescored	7.01%	4.86%
Baseline (MMI)	6.64%	4.60%
Rescored	6.20%	4.39%

The recognition WERs after rescoring are also reported in Table 2 and indicate that the rescored system always outperforms the conventional decoding scheme. In all experiments, the interpolating weights in Eq. 3 were estimated empirically. The acoustic model and language model weights are identical for the baseline and rescored systems. The incorporation of scores generated from the knowledge module improves the baseline systems in all of the experiments.

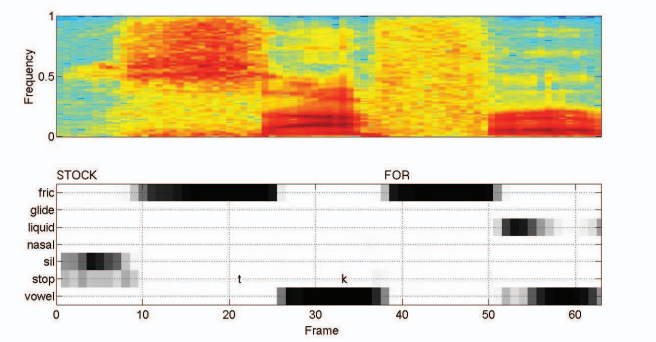

Fig. 2. Spectrogram (upper panel) and posteriogram (bottom panel) for the utterance numbered 446c0210 in the region of the recognition error.

Some examples illustrate the effect of the rescoring. The correct sequence of words for the utterance numbered 441c020t in the Nov92 set is: *Has exposure really been reduced*. The force of the trigram language model leads the decoding process to generate the sentence *Has exposure are really been reduced*. When rescoring is applied, the correct sentence is restored, and the insertion error (e.g., the word "are") is removed. A better understanding of the rescoring effect can be gained by using the *posteriogram* plots. Towards this end, we consider the utterance numbered 446c0210. The correct sequence of words is *The company said its European banking affiliate Safra republic plans to raise more than four hundred fifty million dollars through an international offering*, but the MMI-based system produced *stock for* instead of *Safra* when a trigram language model was used. The bottom panel of Figure 2 shows the posteriogram plot, which is the time evolution of the speech detector output, in

the location around the error. The posterior evolution is shown only for some of the detectors, namely (top to bottom), fricative, glide, liquid, nasal, silence, stop, and vowel, to avoid cluttered plots. The spectrogram of the word *Safra* is shown in the upper panel. In the posteriogram, the decoded (erroneous) word transcription is shown on the upper side. The positions of the stop sounds "t" and "k", which belong to the recognized word "stock", are superimposed at the bottom. The posteriogram shows that the output of the stop detector is null, and the spectrogram confirms the lack of stop events. During the second decoding step, such information can be useful to correct the LVCSR errors, as in this example. Obviously, the detectors are not perfect, some errors cannot be recovered and others can be introduced during rescoring, yet an overall improvement was observed in all of our experiments.

5. CONCLUSION

This paper studies the utility of incorporating acoustic phonetic information not directly utilized by the conventional ASR systems during the decoding phase with particular attention to the LVCSR task. Such articulatory information was obtained through a data-driven approach based on a bank of speech detectors and an event merger. To show that portability and generality are two key properties of the proposed framework the set of detectors and merger used during the rescoring phase are not retrained for the current task. Experimental evidence clearly demonstrates that the proposed approach improves the performance of ASR systems in several continuous speech recognition applications. Moreover, these experiments show that the proposed approach is particularly effective in dealing with errors that do not observe strict acoustic phonetic constraints.

The set of detectors and the merger can be further improved by introducing additional phonetic features and using more sophisticated data-driven approaches. Nonetheless, the goal of this study was to extend our detector based approach to the LVCSR task, and verify whether our approach can help remove some of the decoding errors caused by linguistic constraints. Our ongoing research includes the refinement of the bank of speech event detectors using more sophisticated data-driven techniques.

6. ACKNOWLEDGMENTS

This work was partly funded by (1) The Research Council of Norway through the SIRKUS project, and (2) The Italian Ministry of Education, University and Research (MIUR) through Prof. Sorbello's speech project (University of Palermo).

7. REFERENCES

[1] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, New Jersey, USA, 1993.

[2] N. N. Bitar and C. Y. Espy-Wilson, "Knowledge-based parameters for HMM speech recognition," in *Proc. of ICASSP*, Atlanta, USA, May 1996, pp. 29–32.

[3] E. Eide, "Distinctive features for use in an automatic speech recognition system," in *Proc. of the EuroSpeech*, Aalborg, Denmark, Sept. 2001, pp. 1613–1616.

[4] M. Richardson, J. Blimes, and C. Diorio, "Hidden articulator Markov models for speech recognition," *Speech Communication*, vol. 41, no. 2, pp. 511–529, 2003.

[5] F. Metze and A. Waibel, "A flexible stream architecture for ASR using articulatory features," in *Proc. of ICSLP*, Denver, USA, Sept. 2002, pp. 16–20.

[6] K. Kirchhoff, "Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments," in *Proc. of ISCLP*, Sydney, Australia, Nov./Dec. 1998, pp. 891–894.

[7] B. Launay, O. Siohan, A. C. Surendran, and C.-H. Lee, "Towards knowledge-based features for HMM based large vocabulary automatic speech recognition," in *Proc. of ICASSP*, Orlando, USA, May 2002, pp. 817–820.

[8] M. Hasegawa et al., "Landmark-based speech recognition: Report of the 2004 Johns Hopkins summer workshop," in *Proc. of ICASSP*, Philadelphia, USA, May 2005.

[9] S. M. Siniscalchi, T. Svendsen, and C.-H. Lee, "Towards bottom-up continuous phone recognition," in *Proc. of ASRU*, Dec. 2007, pp. 566–569.

[10] S. M. Siniscalchi, T. Svendsen, and C.-H. Lee, "Toward a detector-based universal phone recognizer," in *Proc. of ICASSP*, Mar./Apr. 2008, pp. 4261–4264.

[11] S. M. Siniscalchi, J. Li, and C.-H. Lee, "A study on lattice rescoring with knowledge scores for automatic speech recognition," in *Proc. of InterSpeech*, Pittsurgh, USA, Sept. 2006, pp. 517–520.

[12] N. Chomsky and M. Halle, *The Sound Pattern of English*, Harper and Row, New York, USA, 1968.

[13] M. Tang, S. Seneff, and V. W. Zue, "Modeling linguistic features in speech recognition," in *Proc. of Eurospeech*, Geneva, Switzerland, Sept. 2003.

[14] P. Schwarz, P. Matějka, and J. Černockỳ, "Hierarchical structures of neural networks for phoneme recognition," in *Proc. of ICASSP*, Toulouse, France, May 2006, pp. 325–328.

[15] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus," in *U.S. Dept. of Commerce, NIST*, Gaithersburg, USA, Feb. 1993.

[16] M.-W. Koo, C.-H. Lee, and B.-H. Juang, "Speech recognition and utterance verification based on a generalized confidence score," *IEEE Speech Audio Processing*, vol. 9, no. 8, pp. 821–831, 2001.

[17] P. C. Woodland, J. Odell, V. Valtchev, and S. Young, "Large vocabulary continuous speech recognition using htk," in *Proc. of ICASSP*, Adelaide, Australia, Apr. 1994, pp. 125–128.

[18] B. Mak, T.-C. Lai, and R. Hsiao, "Improving reference speaker weighting adaptation by the use of maximum-likelihood reference speakers," in *Proc. of ICASSP*, Toulouse, France, May 2006, pp. 229–232.

[19] J. Li, M. Yuan, and C.-H. Lee, "Approximate test risk bound minimization through soft margin estimation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2393–2404, 2007.

[20] W. Macherey, L. Haferkamp, R. Schlter, and H. Ney, "Investigations on error minimizing training criteria for discriminative training in automatic speech recognition," in *Proc. of Interspeech*, Lisboa, Portugal, Sept. 2005, pp. 2133–2136.