ATTRIBUTE BASED LATTICE RESCORING IN SPONTANEOUS SPEECH RECOGNITION

I-Fan Chen[1], *Sabato Marco Siniscalchi*[1,2], *and Chin-Hui Lee*[1]

[1]School of ECE, Georgia Institute of Technology, USA
[2]Faculty of Architecture and Engineering, University of Enna "Kore", Italy

ABSTRACT

In this paper we extend attribute-based lattice rescoring to spontaneous speech recognition. This technique is based on two key features: (i) an attribute-based frontend, which consists of a bank of speech attribute detectors followed up by an evidence merger that generates confidence scores (e.g., sub-word posterior probabilities), and (ii) a rescoring module that integrates information generated by the frontend into an existing ASR engine through lattice rescoring. The speech attributes used in this work are phonetic features, such as frication and palatalization. Experimental results on the Switchboard part of the NIST 2000 Hub5 data set demonstrate that the proposed approach outperforms LVCSR systems based on Gaussian mixture model/ hidden Markov model (GMM/HMM) that does not use attribute related information. Furthermore, a small yet promising improvement is also observed when rescoring word-lattices generated by a state-of-the-art ASR system using deep neural networks. Different frontend configuration are investigated and tested.

Index Terms— Lattice Rescoring; Artificial Neural Networks, Phonetic Features, Automatic Speech Recognition.

1. INTRODUCTION

The top-down automatic speech recognition (ASR) approach, based on hidden Markov models (HMMs) (e.g., [1]), has enjoyed more than 30 years of technology advances (e.g., [2, 3, 4]), and it has thus been the leading paradigm to tackle the speech-to-text problem. Recently, its performance has been further enhanced by modeling frames of coefficients that represent the acoustic input with deep neural networks (DNNs) [5]. Indeed, a remarkable performance has been attained in many large vocabulary continuous speech recognition (LVCSR) tasks [6, 7, 8, 9] with the DNN/HMM solution. As an example, in the early 90's, a high error rate of over 40% was reported on the Switchboard task [10], but this error was reduced to the order of 20% (see [6], for example) using discriminative training (e.g., boosted maximum-mutual-information [11]), and feature space adaptation techniques (e.g., feature space maximum likelihood linear regression (fMLLR) [12]) when the acoustic observations are modeled with Gaussian mixture models (GMMs). In 2011, this error was further reduced down to 16% using a DNN/HMM acoustic model [6]. A 12% error rate was recently reported by employing DNN-derived acoustic features in GMM/HMM based acoustic models [9] and using 2000-hour training data. Although given these positive results, LVCSR errors are still rather high when compared with the recognition performance on read speech. In spontaneous speech, ill-formed utterances are often observed that cannot be completely characterized, even if a large amount of training speech data is collected to build language models.

Several speech researchers have tried to employ knowledge sources in speech production (e.g., [13, 14]) and auditory processing and perception (e.g., [15, 16, 17] to mitigate some of the ASR limitations. Many of these sources are not easily integrated into the conventional top-down ASR systems, and alternative multi-stage (a.k.a. stage-by-stage) ASR paradigms, which could ease this integration, have been explored. For example, a good performance was demonstrated in a speech understanding task by using key phrase detection followed by utterance verification [18]. New theories of nonlinear phonology, articulatory phonology, and landmark-based speech perception were employed in [19] to design a segment-based, multi-stage recognizer. In [20], the authors argued that speech features and lexical words are inherently correlated in natural language and demonstrated that better recognition accuracies can be obtained by jointly optimizing acoustic and linguistic parameters according to the maximum entropy principle. In the automatic speech attribute transcription (ASAT) framework [21], ASR is seen from a bottom-up and "divide-and-conquer" perspective. ASAT aims at identifying acoustic and linguistic information not fully exploited by the current top-down ASR paradigm. Within the ASAT paradigm a new family of lattice-based speech recognition systems grounded on accurate detection of speech attributes was implemented. It was demonstrated that high-accuracy phoneme recognition can be built within this framework [22]. Moreover, the lattice rescoring approach can be extended to LVCSR [23, 24] with good results on read speech.

Although most of those alternative ASR paradigms have overcome ASR limitations only on specific speech tasks, we believe that bottom-up, stage-by-stage paradigms will be proven to be useful as more knowledge sources will become available. This paper is thus our first attempt to extend the

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

ASAT lattice rescoring technique from read to spontaneous speech. Specifically, the lattice rescoring approach is evaluated on the Switchboard task. With respect to our previous implementation, several modifications have been introduced and evaluated, namely DNNs have been used to implement the set of ASAT detectors, and senone output classes have also been evaluated at the merger level. Furthermore, rescoring was applied to word lattices generated by either GMM/HMM or DNN/HMM ASR systems.

The remainder of the paper is organized as follows. Section 2 provides a brief survey of the ASAT framework. Section 3 describes the ASAT lattice recognition algorithm. Next, the experimental setup is reported in Section 4. The experimental results are given and discussed in Section 5. Finally we summarize our findings in Section 6.

2. A GLIMPSE INSIDE ASAT

The ASAT detection-based front-end consists of two key elements: (a) a bank of attribute detectors that can produce detection results together with confidence scores, and (b) an evidence merger that combines low level events (attribute scores) into higher level evidence, such as phoneme posteriors. The outputs delivered by the attribute detectors can be stacked together for a given input in order to generate a supervector of attribute detection scores. This supervector is fed into the evidence merger. In practice, this front-end maps acoustic features into posterior probabilities. An intermediate transformation is accomplished by a bank of speech attribute detectors that scores events embedded into the speech signal. For English, which is what we evaluate in this paper, an attribute detector is built for each of the following phonetic features: *fricative, nasal, stop, approximant, coronal, dental, glottal, labial, low, mid, retroflex, velar, anterior, back, continuant, tense, voiced*. The merger discriminates among either context-independent phoneme classes or context-dependent phoneme (senone) classes.

Attribute detectors and event merger used through all experiments in this paper are implemented using a feed-forward multi-layer perceptron (MLP) networks with either a single hidden layer, or multiple hidden layers. In both cases, MLPs are designed for estimating class posterior probabilities in a discriminative way. The conditional probability of a class label y given an input vector x is estimated using a nonlinear model of the form

$$\hat{p}_k = \hat{p}(y = k|x) = \frac{\exp g_k}{\sum_{i=1}^{N} \exp g_i}, \quad (1)$$

where g_k is the linear activation function of the kth output.

The sigmoidal activation function is used as non-linearity in hidden neurons. The training protocol follows the classical stochastic back-propagation algorithm used to train the neural networks [25]. Furthermore, the cross-entropy criterion, which measures a "distance" between probability distributions, is adopted as the criterion function for training phase. To avoid over-fitting during the training process, the reduction in classification error on the development set was adopted as the stopping criterion.

3. ASAT LATTICE RESCORING

ASAT word lattices rescoring [23] is implemented as follows: each arc in a lattice corresponds to a word in a string hypothesis. A score at the end of each word, a *word-level score*, WS, is obtained by summing up the scores, PS^i, of each phoneme/senone composing that word. Thus WS is a linear combination of phoneme/senone scores. In turn, PS^i is computed by summing up all of (log) posterior probabilities, optionally discounted by the prior probability, generated by the MLP for that class. The weighted rescoring formula is defined as

$$S_n = w_1 W_n + w_2 L_n, \quad (2)$$

where W_n is defined as $W_n = \sum_{i=1}^{K} PS_n^i$. PS_n^i is the score of the *i*-th phoneme/senone in the *n*-th arc, K is the number of phonemes/senones in the word associated with the *n*-th arc, w_2 is the interpolation weight of the log-likelihood score computed by the LVCSR baseline system, L_n,and w_1 is the interpolation weight of the word-level score.

4. EXPERIMENTAL SETUP

4.1. Corpora

We evaluate the effectiveness of the ASAT lattice rescoring algorithm on the task of conversational telephone speech-to-text transcription using the 309-hour Switchboard- I Release 2 training set [10] together with the Mississippi State transcripts2 and the 30K-word lexicon released with those transcripts. The lexicon contains pronunciations for all words and word fragments in the training data. In this first attempt to extend lattice rescoring to spontaneous speech, the primary test set adopted is the 1831-segment SWB part of the NIST 2000 Hub5. That is, the Callhome subset of the NIST 2000 Hub5 task is excluded. This allows us to have ASR baseline systems comparable with those reported in [6].

4.2. Acoustic Features

The Kaldi toolkit [26] is used to generate the acoustic features needed to train the acoustic models (both GMM/HMM and DNN/HMM) of the ASR baseline systems, and the attribute detectors. The 40-dimensional acoustic vector is generated as follows: 13-dimensional Mel-frequency cepstral coefficient [27] features are spliced in time taking a context size

of 9 frames (i.e., 4), followed by de-correlation and dimensionality reduction to 40 using linear discriminant analysis. maximum likelihood linear transform is used to further decorrelate the acoustic features. This is followed by speaker normalization using fMLLR. The fMLLR has 40×41 parameters and is estimated using the GMM-based system applying speaker adaptive training.

4.3. Baseline Systems

Three independent LVCSR baseline systems were built:

ML GMM/HMM: The HTK [28] toolkit was used to design this baseline GMM/HMM system trained on the acoustic features described above. The models trained on the full training data 40 mixture components per observation state. The GMM/HMM models were trained with maximum likelihood (ML). Using more than 40 Gaussians did not improve the ML result.

BMMI GMM/HMM: The KALDI recipe was used to design this baseline GMM/HMM system trained on the acoustic features described above. The models trained on the full training data contain 5230 tied triphone states and 300K Gaussians. These models were trained using boosted maximum mutual information (BMMI) [11] with a 0.1 boosting factor.

DNN/HMM: The KALDI toolkit was used to design this baseline DNN/HMM system trained on the acoustic features described above. These features are globally normalized to have zero mean and unit variance. The DNN trained on the full training data has 6 hidden layers, where each hidden layer has 2048 neurons. The output layer has 8857 output units corresponding to senone classes. The input to the network is an 11 frame (5 frames on each side of the current frame) context window of the 40 dimensional features. A greedy layer-wise pre-training [29] is performed to initialize each hidden layer. The interested reader is referred to [30] for additional details on the design and training of the KALDI DNN.

4.4. Attribute Detectors and Evidence Merger

Each attribute detector was independently trained on the acoustic features described above, except a global normalization transformation was applied to them in order to have to zero mean and unit variance. A development set was created by using 10% of the Switchboard-I Release 2 data, and it was used for stopping the training phase. The remaining 90% of the data were used for the training phase. The actual input to each attribute detector is constructed using $2n + 1$ frames of speech features, and n is the number of look-forward and look-backward frames. In the following experiments, n was set to 4. The neural architecture evaluated in the work is: a deep MLP with four (4) hidden layer having 1500 neurons with sigmoidal activation function with no pre-training. The softmax function is used in order to obtain attribute posterior probabilities. Each attribute detector classifies an input speech frame into one of the following classes: target class, non-target class, voiced noise, noise, laugh, silence. In our previous studies [23], each detector classified the input frame into only two classes: target, and non-target. Yet, SWB data contain non-speech events (such as noise, and laugh) that may harm the detector performance if assigned to the non-target class. In order to be consistent with the ASAT terminology [21], we decided to use the term attribute detectors to refer to these attribute classifiers.

The merger is implemented using an MLP, and it combines the evidence generated at output of each attribute detectors and generates class posterior probabilities. The actual input to the merger is constructed using the present frame along with four look-forward and four look-backward frames. Hence, the input dimension is 1080. Different neural configurations have been evaluated: 1) MLP with single hidden layer having 1500 nodes, and 46 phone-based output classes ($s-$Merger), 2) deep MLP with five hidden layers having each 2048 nodes, and 46 phone-based output classes ($d-$Merger). In the deep configuration, senone classes were also used. That resulted in a deep MLP configuration with 8957 output classes. In all cases, the softmax function is used as output non-linearity. In the deep MLP configuration ($d-$Merger$-sen$). In this work, the possible advantages of pre-training was not investigated.

5. RESULTS

5.1. Results on Attribute and Phone Classification

Table 1 shows the classification accuracies at a frame level for the speech attributes used in this work. From this table we observe that high attribute accuracies can be delivered using a deep MLP trained over short-time spectral features. Furthermore, for some attributes, such as nasal, and retroflex the classification accuracy is over 90%. In general, the attribute classification accuracy is in the range between 80-90%. The lowest accuracy is observed for the tense class. It is also interesting to notice that the classification accuracy is only 93.7% which is much lower than that attained using high-quality read speech [21].

Table 2 summarizes the classification accuracies at the frame level for phoneme and senone classes. We observe a 1% absolute improvement going from shallow to deep MLPs although pre-training was not applied to initialize the deep network architecture. The latter makes us believe that further improvement can be attained by applying a layer-wise pre-training of the neuron parameters. In the last column, the senone-based deep MLP accuracies is reported. It is the first time we use senones at the output of the merger, and the frame classification accuracy attained is equal to 51%.

Table 1. *Classification accuracies (in %) at a frame level on the development set for the speech attributes used in this work.*

Attribute	Frame Accuracy
anterior	82.8
approximant	89.1
back	82.4
continuant	84.3
coronal	83.9
dental	92.4
fricative	88.2
glottal	93.5
high	86.2
labial	89.3
low	88.5
mid	85.2
nasal	90.3
retroflex	92.0
stop	87.2
tense	79.9
velar	91.7
voiced	86.8
vowel	81.4
silence	93.7

Table 2. Merger Accuracy Results (in %) on the Development Data.

setup	$s-$MLP	$d-$MLP	$d-$MLP$-sen$
Accuracy	69.0%	71.2%	51.1%

5.2. Lattice Rescoring Results

Table 3 shows system performance for the three ASR systems evaluated in this work along with the lattice rescoring accuracies in terms of word error rate (WER). The performance of the three ASR baseline systems shown in the first row of Table 3 are comparable with those reported in [6] on the Switchboard part of the Hub5 2000 data with similar architecture solutions. Hence lattice rescoring is applied to top ASR systems for the task at hand. We first discuss the results using a standard GMM/HMM baseline system, and we therefore focus on the first two columns of Table 3. When a shallow MLP ($s-$Merger) is used to implement the merger (second row of the table), the rescored systems always achieves better performance than the conventional baseline system due to the system combination effect. In particular, the WER is reduced from the initial 24.2% down to 22.8% when the ML GMM/HMM baseline system is used. The same rescoring procedure carried out over the BMMI GMM/HMM system produces a final WER of 19.2% starting form the initial 19.5%. A bigger improvement is observed when the shallow merger is replaced by a deep merger ($d-$Merger), as shown in the third row of the table. A final WER of 21.8% is attained by rescoring lattices generated with ML GMM/HMM system, and a WER of 18.6% is instead observed rescoring over lattices generated with the BMMI GMM/HMM baseline system.

Table 3. Lattice Rescoring Results in terms of WER (in %).

setup	ML	BMMI	DNN
Baseline	24.2	19.5	15.7
$s-$Merger	22.8	19.2	no-improv.
$d-$Merger	21.8	18.6	no-improv.
$d-$Merger$-sen$	–	–	15.5

It should be pointed out that if we were to use the output of the merger as state probability density function of a conventional hybrid ANN/HMM system, the WER would be equal to 24.9% and 28.8% for the $s-$Merger, and the $d-$Merger case, respectively. These WERs are worse than the results attained through lattice rescoring; therefore, beneficial complementary information has been injected into the baseline systems.

DNNs have boosted ASR system performance, as abovementioned, and revitalized the "hybrid" frame-work. Hence, porting our ASAT rescoring technique within this ASR architecture is an important step of our studies. The last column of Table 3 shows results related to DNN/HMM baseline system. No improvement is observed when using phoneme classes at the output of the MLP-based merger. The WER is reduced from 15.7% to 15.5% instead if a deep merger with context-dependent phone classes is employed in the rescoring phase. Although the small improvement, we believe it is quite promising for several reasons: 1) attributed detectors can be improved using attribute-specific acoustic features, 2) the attribute detectors can be improved using context-dependent attributes and pre-training, and 3) attribute-based approaches provide a more flexible framework than DNN/HMM baseline systems for capturing pronunciation details, and 4) a more sophisticated rescoring scheme can be developed.

6. SUMMARY

We have presented our first attempt at poring the ASAT lattice rescoring technique to spontaneous speech. Several ASAT frontend configurations were developed and evaluated, and it was demonstrated that our rescoring technique reduced the WER of standard GMM/HMM systems built employing state-of-the-art techniques, such as speaker adaptive training, and discriminative training. In particular, the WER was reduced from 24.2% down to 21.8% when the acoustic model is trained using maximum likelihood estimation, and the ASAT merger is a deep MLP with no pre-training. Further, the WER is reduced to 18.6% by rescoring lattices generated with a BMMI GMM/HMM system trained over fMLLR acoustic features. A first attempt at rescoring lattices generated with a DNN/HMM system was also carried out, and a small improvement was observed. In future work, more sophisticated rescoring combination technique will be devised, and MLP pre-training will be investigated.

7. REFERENCES

[1] L. R. Rabiner, "A tutorial on hidden Markov models and selected application in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[2] J.-L. Gauvain and L. Lamel, "Large vocabulary continuous speech recognition: Advances and applications," *Proc. IEEE*, vol. 88, no. 8, pp. 1181–1200, 2000.

[3] C.-H. Lee and Q. Huo, "On adaptive decision rules and decision parameter adaptation for automatic speech recognition," *Proc. IEEE*, vol. 88, no. 8, pp. 1241–1269, 2000.

[4] H. Ney and S. Ortmanns, "Progresses in dynamic programming search for LVCSR," *Proc. IEEE*, vol. 88, no. 8, pp. 1224–1240, 2000.

[5] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, pp. 1771–1800, 2002.

[6] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Proc. Interspeech*, Florence, Italy, Aug. 2011, pp. 437–440.

[7] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 20, no. 1, pp. 30–42, 2012.

[8] S. M. Siniscalchi, D. Yu, L. Deng, and C.-H. Lee, "Speech recognition using long-span temporal patterns in a deep network model," *Signal Processing Letters*, vol. 20, no. 3, pp. 201–204, 2013.

[9] Z.-J. Yan, Q. Huo, and J. Xu, "A scalable approach to using DNN-derived features in GMM-HMM based acoustic modeling for LVCSR," in *Proc. INTERSPEECH*, Lyon, France, Aug. 2013, pp. 104–108.

[10] J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: telephone speech corpus for research and development," in *Proc. ICASSP*, San Francisco, CA, USA, Mar. 1992, pp. 517–520.

[11] D Povey, D.i Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature space discriminative training," in *Proc. ICASSP*, Las Vegas, NV, USA, Mar./Apr. 2008, pp. 4057–4060.

[12] Y. Li, H. Erdogan, T. Gao, and E. Marcheret, "Incremental on-line feature space MLLR adaptation for telephony speech recognition," in *Proc. ICSLP*, Denver, CO, USA, Sept.. 2002, pp. 1471–1420.

[13] L. Deng, "Computational models for speech production," in *Computational Models for Speech Pattern Processing*, NATO ASI. Springer-Verlag, 1999.

[14] L. Deng, "Articulatory features and associated production models in statistical speech recognition," in *Computational Models for Speech Pattern Processing*, NATO ASI. Springer-Verlag, 1999.

[15] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 1, pp. 115–132, 1994.

[16] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," *J. Phonetics*, vol. 16, pp. 55–76, 1988.

[17] L. Deng, "Computational models for auditory speech processing," in *Computational Models for Speech Pattern Processing*, NATO ASI. Springer-Verlag, 1999.

[18] T. Kawahara, C.-H. Lee, and B.-H. Juang, "Flexible speech understanding based on combined key-phrase detection and verification," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 6, pp. 558–568, 1998.

[19] M. Tang, S. Seneff, and V. W. Zue, "Modeling linguistic features in speech recognition," in *Proc. Eurospeech*, Geneva, Switzerland, Sept. 2003, pp. 2585–2588.

[20] J.-T. Chien and C.-H. Chueh, "Joint acoustic and language modeling for speech recognition," *Speech Communication*, vol. 52, no. 3, pp. 223–235, 2010.

[21] C.-H. Lee and S. M. Siniscalchi, "An information-extraction approach to speech processing: Analysis, detection, verification, and recognition," *Proc. IEEE*, vol. 101, no. 5, pp. 1089–1115, 2013.

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

[23] S. M. Siniscalchi and C.-H. Lee, "A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition," *Speech Communication*, vol. 51, pp. 1139–1153, 2009.

[24] S. M. Siniscalchi, D. Yu, L. Deng, and C.-H. Lee, "Exploiting deep neural networks for detection-based speech recognition," *Neurocomputing*, vol. 106, pp. 148–157, 2013.

[25] R.Õ. Duda and P.Ẽ. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, 1973.

[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *ASRU*, Big Island, Hawaii, US, Dec. 2011.

[27] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllable word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, 1980.

[28] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book (for HTK Version 3.3)*, Cambridge University Press, Cambridge, UK, 2005.

[29] G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.

[30] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proc. INTERSPEECH*, Lyon, France, Aug. 2013, pp. 2345–2349.