BOOSTING ATTRIBUTE AND PHONE ESTIMATION ACCURACIES WITH DEEP NEURAL NETWORKS FOR DETECTION-BASED SPEECH RECOGNITION

Dong Yu[1], Sabato Marco Siniscalchi[2], Li Deng[1], and Chin-Hui Lee[3]

[1]Speech Research Group, Microsoft Research, Redmond, WA. USA
[2]Department of Telematics, Kore University of Enna, Enna, Italy
[3]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA. USA
dongyu@microsoft.com, siniscalchi77.19@gmail.com, deng@microsoft.com, chl@ece.gatech.edu

ABSTRACT

Generation of high-precision sub-phonetic attribute (also known as phonological features) and phone lattices is a key frontend component for detection-based bottom-up speech recognition. In this paper we employ deep neural networks (DNNs) to improve detection accuracy over conventional shallow MLPs (multi-layer perceptrons) with one hidden layer. A range of DNN architectures with five to seven hidden layers and up to 2048 hidden units per layer have been explored. Training on the SI84 and testing on the Nov92 WSJ data, the proposed DNNs achieve significant improvements over the shallow MLPs, producing greater than 90% frame-level attribute estimation accuracies for all 21 attributes tested for the full system. On the phone detection task, we also obtain excellent frame-level accuracy of 86.6%. With this level of high-precision detection of basic speech units we have opened the door to a new family of flexible speech recognition system design for both top-down and bottom-up, lattice-based search strategies and knowledge integration.

Index Terms — automatic speech attribute transcription, deep neural networks, detection-based ASR, phonological features, attribute detection, phone recognition

1. INTRODUCTION

State-of-the-art automatic speech recognition (ASR) systems are often based on a pattern matching framework that is motivated by expressing spoken utterances as sequences of stochastic patterns [1]. A single probabilistic finite state network (FSN), composed of acoustic hidden Markov model (HMM) states, phones, lexicon, grammar nodes, and their connecting arcs [2], is constructed in a top-down manner to represent all ASR task constraints. For a given input utterance the FSN is searched for the most likely sequence of words as the recognized sentence with maximum a posteriori (MAP) decoding [1], known as the integrated search strategy. On the other hand, the need for alternative ASR paradigms had attracted some research attention in recent years. Automatic speech attribute transcription (ASAT) [3][4] is a recently proposed framework based on bottom-up attribute detection of a collection of speech cues followed by knowledge integration of such cues to make linguistic validations. ASAT makes use of the articulatory-based phonological features studied earlier [6][7][8][9] in a new detection-based framework, and extends them to a number of tasks including rescoring of word lattices generated by state-of-the-art HMM systems [3], continuous phone recognition [5], cross-language attribute detection and phone recognition [10] and spoken language recognition [11].

In a recent attempt an LVCSR (large vocabulary continuous speech recognition) system was realized in a bottom-up, decoupled search fashion using weighted finite state machines (WFSMs) [12]. It was found that high-precision lattice generation at every stage of the knowledge integration phase and low-error lattice pruning with limited memory requirements are two critical research issues to warrant good performances for such a bottom-up, decoupled ASR search strategy. In this study we explore the first key challenge of generating high-precision attribute and phone lattices. We extend the conventional shallow MLPs used in [12] to deep neural networks (DNNs) [13], which has been shown to have very good theoretical properties [14] and demonstrated superior performances for both phone [16][17] and word recognition [15][13][18][19].

In this paper, we explore a wide range of DNN architectures by extending the conventional single hidden layer MLPs to five and seven layers. We also expand the number of hidden units in each layer from the original 800 and 1500 units [12] to 2048 units, and show that the DNNs lead to better attribute and phone recognition performance. The significantly boosted quality in attribute and phone estimation makes it highly promising to advance bottom-up LVCSR with DNNs and with new ways of incorporating the key asynchrony properties of the attributes. This also opens doors to new flexibility in combining top-down and bottom-up ASR.

2. ATTRIBUTES AND PHONES

An example detection-based frontend is shown in Figure 1 which was used in [12]. It consists of two main blocks: a bank of *speech attribute detectors* and an *evidence merger*. For English, which is what we evaluate in this paper, an *attribute detector* is built for each of the 21 phonological features listed in Table 1. Each *attribute detector* analyzes an expanded frame of the input speech signal and produces the posterior probability that pertains to some acoustic-phonetic attribute. Feed-forward multi-layer perceptrons (MLPs) can be used to build detectors. The input to each detector can be any speech features. Here in this paper we adopt the conventional 39-dim MFCC+Δ+ΔΔ vector. The number of outputs for each attribute detector is two: attribute present and absent. The long-term dependencies among attributes are taken into account in the *append & expand module*, which stacks together a window of eleven frames around the frame to be classified and generates a super-vector. The *merger module* is then fed with this super-vector. The merger can also be implemented with a frame-based MLP to discriminate among 40 phone classes.

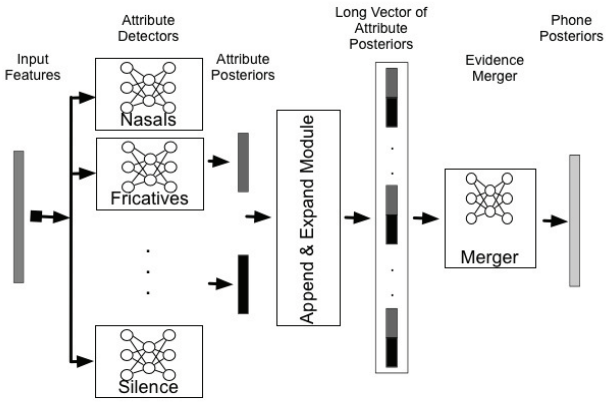

Figure 1: *Front-end module of detection-based ASR*

Table 1. Phonological features (attributes) and their associated phones used in this study.

	Attribute	Phonemes
manner	Vowel	iy ih eh ey ae aa aw ay ah ao oy ow uh uw er
	Fricative	jh ch s sh z zh f th v dh hh
	Nasal	m n ng
	Stop	b d g p t k
	Approximant	w y l r
place	Coronal	d l n s t z
	High	ch ih iy jh sh uh uw y ow g k ng
	Dental	dh th
	Glottal	hh
	Labial	b f m p v w
	Low	aa ae aw ay oy
	Mid	ah eh ey ow
	Retroflex	er r
	Velar	g k ng
others	Anterior	b d dh f l m n p s t th v z w
	Back	ay aa ah ao aw ow oy uh uw g k
	Continuant	aa ae ah ao aw ay dh eh er r ey l f ih iy oy ow s sh th uh uw v w y z
	Round	aw ow uw ao uh v y oy r w
	Tense	aa ae ao aw ay ey iy ow oy uw ch s sh f th p t k hh
	Voiced	aa ae ah aw ay ao b d dh eh er ey g ih iy jh l m n ng ow oy r uh uw v w y z
	Silence	sil

3. DEEP NEURAL NETWORKS

A DNN is a multi-layer perceptron (MLP) with many hidden layers. Due to the deep structure and the complicated nonlinear surface introduced by the large number of hidden layers it is important to employ effective training strategies. A popular trick is to initialize the parameters of each layer greedily and generatively by treating each pair of layers in DNNs as a restricted Boltzmann machine (RBM) before doing a joint optimization of all the layers [14]. This learning strategy enables discriminative training to start from well initialized weights and is used in this study.

3.1 Restricted Boltzmann Machines

An RBM can be represented as a bipartite graph with a visible layer and a hidden layer. The stochastic units in the visible layer only connect to the stochastic units in the hidden layer. The units in the visible layer are typically represented by Bernoulli or Gaussian distributions and the units in the hidden layer are typically represented with Bernoulli distributions. Gaussian-Bernoulli RBMs can be used to convert real-valued stochastic variables (such as MFCCs) to binary stochastic variables which can then be further processed using the Bernoulli-Bernoulli RBMs.

Given the model parameters θ, the joint distribution $\mathrm{p}(\mathbf{v}, \mathbf{h}; \theta)$ over the visible units $\mathbf{v}$ and hidden units $\mathbf{h}$ in the RBMs can be defined as

$$\mathrm{p}(\mathbf{v}, \mathbf{h}; \theta) = \frac{exp\big(-\mathrm{E}(\mathbf{v}, \mathbf{h}; \theta)\big)}{Z}, \quad (1)$$

where $\mathrm{E}(\mathbf{v}, \mathbf{h}; \theta)$ is an energy function and $Z = \sum_v \sum_h exp\big(-\mathrm{E}(\mathbf{v}, \mathbf{h}; \theta)\big)$ is the partition function. The marginal probability that the model assigns to a visible vector $\mathbf{v}$ is

$$\mathrm{p}(\mathbf{v}; \theta) = \frac{\sum_{\boldsymbol{h}} exp\big(-\mathrm{E}(\mathbf{v}, \mathbf{h}; \theta)\big)}{Z}. \quad (2)$$

The parameters in RBMs can be optimized to maximize log likelihood $\log p(\mathbf{v}; \theta)$ and can be updated as

$$\Delta w_{ij} = \langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model}, \quad (3)$$

where $\langle v_i h_j \rangle_{data}$ is the expectation that v_i and h_j occur together in the training set and $\langle v_i h_j \rangle_{model}$ is that same expectation under the distribution defined by the model. Because $\langle v_i h_j \rangle_{model}$ is extremely expensive to compute exactly, the contrastive divergence (CD) approximation to the gradient is used, where $\langle v_i h_j \rangle_{model}$ is replaced by running the Gibbs sampler initialized at the data for one full step [14].

3.2 Deep Neural Network Training Process

The last layer of a DNN transforms a number of Bernoulli distributed units into a multinomial distribution using the softmax operation

$$\mathrm{p}(l = k|\mathbf{h}; \theta) = \frac{exp\big(\sum_{i=1}^{H} \lambda_{ik} h_i + a_k\big)}{Z(\boldsymbol{h})}, \quad (4)$$

where $l = k$ denotes the input been classified into the kth class, and λ_{ik} is the weight between hidden unit h_i at the last layer and class label k.

To learn the DNNs, we first train a Gaussian-Bernoulli RBM generatively in which the visible layer is the continuous input vector constructed from $2n + 1$ frames of speech features, in which n is the number of look-forward and look-backward frames. We then use Bernoulli-Bernoulli RBMs for the remaining layers. When pre-training the next layer, $E\big(h_j|\mathbf{v}; \theta\big) = p\big(h_j = 1|\mathbf{v}; \theta\big)$ from the previous layer is used as the visible input vector based on the mean-field theory. This process continues until the last layer at which time error back-propagation (BP) is used to fine-tune all the parameters jointly by maximizing the frame-level cross-entropy between the true and the predicted probability distributions over class labels.

4. EXPERIMENTS

In this section we report results on using DNNs to train attribute and phone detectors and show that the detectors trained using DNNs significantly outperform those trained using shallow MLPs used in [12].

4.1. Experimental Setup

Table 2. Average cross entropy (CE) and classification accuracies at a frame level for the speech attributes.

Attribute	Prior	Naïve	800x1	2048x5 (DNN)					
	train	test	test	train		cv		test	
	%	Acc (%)	Acc (%)	Avg CE	Acc (%)	Avg CE	Acc (%)	Avg CE	Acc (%)
anterior	36.2	**63.8**	**85.6**	-0.14	94.4	-0.21	91.8	-0.19	**92.5**
approximant	9.2	**90.8**	**94.9**	-0.06	97.5	-0.09	96.5	-0.09	**96.4**
back	19.6	**80.4**	**87.6**	-0.14	94.4	-0.18	92.6	-0.17	**93.1**
continuant	55.7	**55.7**	**88.7**	-0.15	94.3	-0.18	93.0	-0.16	**93.5**
coronal	25.5	**74.5**	**87.9**	-0.14	94.3	-0.19	92.6	-0.20	**92.4**
dental	1.4	**98.6**	**98.9**	-0.02	99.4	-0.02	99.2	-0.02	**99.0**
fricative	15.3	**84.7**	**94.2**	-0.08	96.9	-0.10	96.1	-0.10	**96.2**
glottal	0.8	**99.2**	**99.3**	-0.01	99.7	-0.01	99.6	-0.01	**99.7**
high	16.7	**83.3**	**90.7**	-0.09	96.8	-0.13	95.2	-0.13	**95.0**
labial	11.0	**89**	**92.5**	-0.07	97.3	-0.10	96.4	-0.08	**96.9**
low	9.3	**90.7**	**94.6**	-0.05	98.1	-0.09	96.7	-0.09	**96.9**
mid	11.8	**88.2**	**90.7**	-0.13	94.9	-0.15	93.9	-0.15	**93.8**
nasal	8.7	**91.3**	**95.9**	-0.04	98.5	-0.07	97.4	-0.07	**97.7**
retroflex	6.2	**93.8**	**97.6**	-0.02	99.1	-0.04	98.5	-0.04	**98.5**
round	14.7	**85.3**	**91.9**	-0.07	97.4	-0.13	95.3	-0.14	**94.9**
silence	19.0	**81**	**97.6**	-0.04	98.5	-0.04	98.2	-0.03	**98.7**
stop	15.3	**84.7**	**92.9**	-0.10	96.2	-0.13	95.3	-0.12	**95.7**
tense	39.5	**60.5**	**83.0**	-0.16	93.8	-0.23	90.7	-0.24	**90.6**
velar	5.4	**94.6**	**96.6**	-0.03	99.1	-0.05	98.4	-0.04	**98.7**
voiced	59.9	**59.9**	**92.1**	-0.11	95.9	-0.13	95.0	-0.12	**95.3**
vowel	32.5	**67.5**	**87.9**	-0.13	94.6	-0.19	92.6	-0.18	**92.8**

All experiments were conducted on the 5,000-word speaker independent WSJ0 (5k-WSJ0) task [20]. The training material from the SI84 set (7077 utterances, or 15.3 hours of speech from 84 speakers) is separated into a 6877-utterance training set and a 200-sentence cross-validation (CV) set. Evaluation was carried out on the Nov92 evaluation data with 330 utterances from 8 speakers. In all the studies, 13 MFCCs+Δ+ΔΔ were chosen as the short-time spectral representation of the speech signal.

In this study, the phone labels were derived from the forced alignments generated using a 2818 8-mixture tied-state cross-word tri-phone GMM-HMM LVCSR system trained with maximum likelihood criterion. The attribute labels were generated by mapping phone labels to attributes according to Table 1. No assimilation of attributes from one phone to the adjacent ones was represented and modeled in the results reported below.

4.2. Results on Attribute Detector

Table 2 compares the average cross entropy (CE) and classification accuracies at a frame level for the speech attributes used in this work. In this table, the prior $p(attr)$ is estimated from the training data. The naïve algorithm assigns each frame with the most probable label (true or false). That is, when the majority of the frames in the training set is true for an attribute, then we assign value "true" to that attribute for all frames. The shallow MLP results were quoted from [12] and were obtained using a single hidden layer MLP with 800 hidden units. The DNN contains 5 hidden layers each with 2048 units following previous work [13].

From this table we observe that the DNN significantly outperforms the shallow MLP, with relative error rate reductions ranging from 40% to 90% for different attributes. The average relative error rate is reduced by 56% across all attributes over the shallow MLP. In fact, for many attributes, such as *back*, *labial*, and *mid*, single hidden layer MLP performs only slightly better than the naïve approach while the DNN achieves a much higher accuracy.

4.3. Results on Phone Estimation

Table 3 summarizes the average cross entropy (CE) and classification accuracies at the frame level for phones. The setup names are encoded as "#_hidden_units x #_hidden_layer input_feature", where the input feature MFCC is the standard 39-dim MFCCs+Δ+ΔΔ feature, input features *Attr1* and *Attr2* refer to the attribute log posterior probability generated from the 800x1 MLP and 2048x5 DNN attribute detectors, respectively, and the input feature *phone* at the bottom row of Table 3 is the phoneme log posterior probability computed from the 2048x5 phone detector DNN with the MFCC input. All the five setups used 11 frames of features - 5 frames looking ahead and 5 frames looking back.

Table 3. Average cross entropy (CE) and phone classification accuracies at the frame level

Setup	train		cv		test	
	avg CE	acc(%)	avg CE	acc(%)	avg CE	acc(%)
1500x1 Attr1	-	86.7	-	82.7	-	**82.6**
2048x5 MFCC	-0.26	91.6	-0.45	85.3	-0.46	**85.1**
2048x7 MFCC	-0.24	91.9	-0.45	85.5	-0.46	**85.3**
2048x5 Attr2	-0.28	90.2	-0.45	85.5	-0.48	**85.0**
2048x2 phone	-0.22	92.3	-0.41	86.8	-0.43	**86.6**

From Table 3 we can make several observations. First, the shallow MLP based phone detector performs the worst even though it used the attribute detector's results as the input feature. For example, we can increase the test set accuracy by absolute 2.5% and 2.7% over the shallow MLP detector, respectively, using a 5-hidden layer and 7-hidden layer DNN. Second, breaking the phone detector into two stages - first to detect the attribute and then to estimate the phone identity based on the results of attribute detectors – has not provided any gain over the direct approach that detect phones using the MFCC features if DNN is used, although the same two-stage detector did show advantages if shallow MLP

or other shallow model (e.g. [24]) is used. This indicates DNNs are powerful enough to capture useful discriminative information. Third, we can obtain additional 1.5% absolute accuracy improvement over the "2048x5 MFCC" configuration by using its output (augmented by the adjacent frames) as the features into a new, higher-level phone detector. Finally, comparing cv and test set results we can see that the DNN results are robust.

5. DISCUSSION AND CONCLUSION

We have demonstrated in this work that we can achieve high accuracies for both phonological attribute detection and phone estimation using DNNs. This opens up new potentials to some old problems, such as speech recognition from a phone lattice [2] and from phonological parsing [22]. It also creates an exciting avenue to provide high-precision attribute and phone lattices for bottom-up, detection-based speech recognition where words can be directly specified in terms of attributes free from phones. For speech understanding, concepts may be also directly specified in terms of attributes free from words. More specifically, for ill-formed utterances, such as in spontaneous speech where partial understanding is often needed because an integrated approach is not sufficient to properly capture the overall knowledge sources, it is expected that the proposed framework will be robust and will give a better performance than the standard HMM-based technology as demonstrated in previously proposed key-phrase detection frameworks [23]. With our initial success reported in this paper, we intend to continue to explore the cross-fertilization of ASAT and DNNs for LVCSR and other applications.

One clear limitation of the current framework in the detection-based speech recognition is the lack of temporal overlapping (i.e., asynchrony) characteristics in the attributes across different dimensions. This limitation is reflected in the static phone-to-attribute mapping (Table 1), and may account for why the use of attributes has not achieved better phone estimation compared with no use of attributes. Yet such asynchrony is central to modern phonological theory. Incorporation of asynchrony will significantly modify the attribute targets in running speech in a principled and parsimonious way, as demonstrated in [6][8][9][21]. With the attribute targets modified in a phonologically meaningful manner, it is hopeful that the DNN approach will further enhance the value of the attributes for making word recognition more accurate in the detection-based framework.

6. REFERENCES

[1] C.-H. Lee and Q. Huo, "On adaptive decision rules and decision parameter adaptation for automatic speech recognition," *Proceedings of the IEEE*, Vol. 88, No. 8, pp. 1241-1269, August 2000.

[2] S. E. Levinson, "Structural methods in automatic speech recognition," *Proc. IEEE*, Vol. 73, pp. 1625-1650, 1985.

[3] 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.

[4] 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)," *Proc. Interspeech*, Antwerp, Belgium, August 2007.

[5] S. M. Siniscalchi, T. Svendsen, and C.-H. Lee, "Towards bottom-up continuous phone recognition," *Proc. ASRU*, 2007.

[6] L. Deng and D. Sun, "A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features," *J. Acoust. Soc. Am*, vol. 85, pp. 2702-2719, 1994.

[7] K. Kirchhoff. "Robust speech recognition using articulatory information," *Ph.D Thesis*, University of Bielefeld, 1999.

[8] S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," *Computer, Speech and Language 14(4)*, pp. 333-345, 2000.

[9] K. Livescu, "Feature-based pronunciation modeling for automatic speech recognition." *Ph.D Thesis*, MIT Sept. 2005.

[10] S. M. Siniscalchi, D.-C. Lyu, T. Svendsen, and C.-H. Lee, "Experiments on cross-language attribute detection and phone recognition with no target-specific training speech data," to appear in *IEEE Trans. Audio, Speech and Language Proc.* 2011.

[11] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, "Exploring universal attribute characterization of spoken languages for spoken language recognition," *Proc. Interspeech*, Brighton, UK, September 2009.

[12] S. M. Siniscalchi, T. Svendsen, and C.-H. Lee "A bottom-up stepwise knowledge-integration approach to large vocabulary continuous speech recognition using weighted finite state machines", *Proc. Interspeech 2011*, pp. 901-904.

[13] G.E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pretrained deep neural networks for large vocabulary speech recognition", *IEEE Trans. Audio, Speech, and Lang. Proc. Jan.* 2012.

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

[15] D. Yu, L. Deng, and G. Dahl, "Roles of pretraining and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition," *Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2010.

[16] A. Mohamed, G. E. Dahl, and G. E. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. on Audio, Speech, and Lang. Proc. Jan. 2012*.

[17] A. Mohamed, D. Yu, and L. Deng, "Investigation of full-sequence training of deep belief networks for speech recognition", in *Proc. Interspeech* 2010, pp. 1692-1695.

[18] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks", *Proc. Interspeech 2011*, pp. 237-240.

[19] F. Seide, G. Li and D. Yu, "Conversational speech transcription using context-dependent deep neural networks", Interspeech 2011, pp. 437-440.

[20] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," *Proc. ICSLP*, 1992.

[21] L. Deng, "Articulatory features and associated production models in statistical speech recognition," *Computational Models of Speech Pattern Processing*, pp. 214-224, Springer, 1999.

[22] K. Church, "Phonological parsing in speech recognition," *Ph.D Thesis*, MIT, 1986.

[23] T. Kawahara, C.-H. Lee and B.-H. Juang, "Key-phrase detection and verification for flexible speech understanding," *IEEE Trans. on Speech and Audio Proc.*, Vol. 6, No. 6, pp. 558-568, Nov. 1998.

[24] U. V. Chaudhari, M. Picheny, "Articulatory Feature Detection with SVM for Integration into ASR and Phone Recognition", *ASRU 2009*, pp. 93-98.