Neurocomputing 106 (2013) 148–157

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Exploiting deep neural networks for detection-based speech recognition

Sabato Marco Siniscalchi [a,b,*], Dong Yu [c], Li Deng [c], Chin-Hui Lee [b]

[a] Faculty of Engineering and Architecture, Kore University of Enna, Cittadella Universitaria, Enna, Sicily, Italy
[b] School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
[c] Speech Research Group, Microsoft Research, Redmond, WA, USA

ARTICLE INFO

Article history:
Received 29 May 2012
Received in revised form
21 September 2012
Accepted 5 November 2012
Communicated by R. Capobianco Guido
Available online 23 November 2012

Keywords:
Deep neural networks
Multi-layer perceptrons
Articulatory recognition
Speech recognition
Lattice rescoring

ABSTRACT

In recent years deep neural networks (DNNs) – multilayer perceptrons (MLPs) with many hidden layers – have been successfully applied to several speech tasks, i.e., phoneme recognition, out of vocabulary word detection, confidence measure, etc. In this paper, we show that DNNs can be used to boost the classification accuracy of basic speech units, such as phonetic attributes (phonological features) and phonemes. This boosting leads to higher flexibility and has the potential to integrate both top-down and bottom-up knowledge into the Automatic Speech Attribute Transcription (ASAT) framework. ASAT is a new family of lattice-based speech recognition systems grounded on accurate detection of speech attributes. In this paper we compare DNNs and shallow MLPs within the ASAT framework to classify phonetic attributes and phonemes. Several DNN architectures ranging from five to seven hidden layers and up to 2048 hidden units per hidden layer will be presented and evaluated. Experimental evidence on the speaker-independent Wall Street Journal corpus clearly demonstrates that DNNs can achieve significant improvements over the shallow MLPs with a single hidden layer, producing greater than 90% frame-level attribute estimation accuracies for all 21 phonetic features tested. Similar improvement is also observed on the phoneme classification task with excellent frame-level accuracy of 86.6% by using DNNs. This improved phoneme prediction accuracy, when integrated into a standard large vocabulary continuous speech recognition (LVCSR) system through a word lattice rescoring framework, results in improved word recognition accuracy, which is better than previously reported word lattice rescoring results.

1. Introduction

State-of-the-art automatic speech recognition (ASR) systems often rely on a pattern matching framework that expresses spoken utterances as sequences of stochastic patterns [1]. Top-down approaches are usually adopted to represent all constraints in a single, compact probabilistic finite state network (FSN), composed of acoustic hidden Markov model (HMM) states with emission probabilities generated by Gaussian mixture models (GMMs), phones, lexicon, grammar nodes, and their connecting arcs [2]. For a given input utterance, the maximum a posteriori decoding [1] procedure is used to find the most possible sequence of words embedded in the FSN as the recognized sentence. This search technique, known as the top-down integrated search strategy, has attained remarkable results in many ASR tasks. Nonetheless, recognition error rates for difficult tasks, such as spontaneous and unconstrained speech recognition, are still unacceptably high. In contrast, there is evidence to show that bottom-up, stage-by-stage ASR paradigm may do better under some spontaneous speech phenomena [3]. Automatic speech attribute transcription (ASAT) [4], a promising alternative ASR paradigm, is a bottom-up framework that first detects a collection of speech attribute cues and then integrates such cues to make linguistic validations. A typical ASAT system uses the articulatory-based phonological features studied earlier [5–9] in a new detection-based framework. ASAT has been extended and applied to a number of tasks including rescoring of word lattices generated by state-of-the-art HMM systems [10], continuous phoneme recognition [11], cross-language attribute detection and phoneme recognition [12] and spoken language recognition [13]. The speech cues detected in ASAT are referred to as *speech attributes*. The terms phonological features and speech attributes will be used interchangeably in this work.

In recent years there has also been a considerable resurgence of interest in neural network approaches to speech recognition. Neural networks are powerful pattern recognition tools that have been used for several real world applications [14], and different

* Corresponding author at: Faculty of Engineering and Architecture, Kore University of Enna, Cittadella Universitaria, Enna, Sicily, Italy.
Tel.: +39 3472913375.
E-mail addresses: marco.siniscalchi@unikore.it (S.M. Siniscalchi), dongyu@microsoft.com (D. Yu), deng@mircosoft.com (L. Deng), chl@ece.gatech.edu (C.-H. Lee).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.11.008

successful techniques have been developed around them since the early '80s in the speech community. For example, in connectionist speech recognition systems [15], neural networks are used to estimate the state emission probabilities of a HMM. In the TANDEM approach [16], a neural network extracts discriminative speech features that are fed into conventional GMM-HMM-based speech recognizers. In detection-based ASR paradigms (e.g., [11]), a set of neural networks learns the mapping from a spectral-based feature space to a phonetic feature space. Neural networks have also been used to model state and transition features in conditional random field (CRF) based ASR systems (e.g., [17]), in beam search pruning [18] and confidence measure estimation [19,20]. Although several architectures have been proposed to tackle different speech recognition tasks, such as recurrent neural networks (e.g., [21,22]) and time-delay neural network [23], the stylistic characteristics of the MLPs is by far the most popular due to the compromise realized between recognition rate, recognition speed, and memory resources. Furthermore, it has been shown that feed-forward neural architectures can approximate any function defined on compact sets in $\mathbf{R}^n$ [24], that is, they are *universal approximators* [14].

More recently, a major advance has been made in training densely connected, generative deep belief nets (DBNs) with many hidden layers. The core idea of the DBN training algorithm suggested in [25] is to first initialize the weights of each layer greedily in a purely unsupervised way by treating each pair of the layers as a restricted Boltzmann machine (RBM) and then fine-tune all the weights jointly to further improve the likelihood. The resulting DBN can be considered as a hierarchy of nonlinear feature detectors that can capture complex statistical patterns in data. For classification tasks, the same DBN pre-training algorithm can be used to initialize the weights in deep neural networks (DNNs) – MLPs with many hidden layers. The weights in the entire DNN can then be fine-tuned using labeled data. DNNs have been proven to be effective in a number of applications, including coding and classification of speech, audio, text, and image data [26–30]. These advances triggered interest in developing acoustic models based on DNNs and other deep learning techniques for ASR. For example, the context-independent DNN-HMM hybrid architectures have recently been proposed for phoneme recognition [31,32] and have achieved very competitive performance. A novel acoustic model, the context-dependent (CD)-DNN-HMM proposed in [33] has been successfully applied to large vocabulary speech recognition tasks and can cut word error rate by up to one third on the challenging conversational speech transcription tasks compared to the discriminatively trained conventional CD-GMM-HMM systems [34].

In this study[1], elements of both of these two research directions, namely ASAT and DNN, are merged together, and the conventional shallow MLPs used in [36] are replaced with DNNs, which has been shown to have very good theoretical properties [37] and demonstrated superior performances for both phoneme [31,32] and word recognition [33,34,38,39]. Following the ASAT paradigm, a *bank of speech attribute detectors* that assign probabilistic scores to manner and place of articulation events is built using DNNs. Then a DNN is designed to (1) combine together the output of these detectors and (2) generate phoneme posterior probability. A wide range of DNN architectures will be built by extending the conventional single hidden layer MLPs to five and seven layers. Experimental evidence on the speaker independent Wall Street Journal dataset [40] demonstrates that the proposed solution outperforms conventional shallow MLPs in both attribute and phoneme classification. Furthermore, by re-scoring the set of most likely hypotheses embedded in the word lattices generated by a conventional HMM-based LVCSR system using the DNN phoneme posterior probabilities, a two-stage LVCSR recognizer gave relative word error rate (WER) reductions ranging from 8.7% to 13.0% over the initial result and improves over previous studies on word lattice rescoring [10].

This result along with the significantly boosted quality in attribute and phoneme estimation makes it highly promising to advance *bottom-up* LVCSR with DNNs and with new ways of incorporating the key asynchrony properties of the articulatory-motivated phonetic attributes. This also opens doors to new flexibility in combining top-down and bottom-up ASR. Furthermore, it should be noted that modeling of articulatory-based phonetic features and phoneme is an active research filed in automatic speech recognition. Therefore, the current investigation can also impact research areas beyond the ASAT framework. For instance, several researchers have argued that better results can be achieved by modeling the underlying processes of co-articulation and assimilation rather than simply describing their effects on the speech signal (e.g., [7]). It is also believed that by integrating articulatory-motivated information into the speech recognition engine most of the problems of the current technology can be addressed. Finally, phoneme estimation also plays a very important role in many speech processing applications, such as out-of-vocabulary detection [41] and language identification (e.g., [42]).

The remainder of the paper is organized as follows. A brief survey on the ASAT paradigm for speech recognition is given in Section 2. Section 3 gives a light overview of related works on articulatory-motivated phonological attributes and phoneme estimation. The DNN architecture and training scheme are discussed in Section 4. The word lattice rescoring procedure adopted in this study is outlined in Section 5. Next, the experimental setup is given in Section 6 in which experimental results on attributes and phoneme classification, and word lattice rescoring are presented and discussed. Finally, we discuss our findings and conclude our work in Section 7.

2. ASAT in a nutshell

It is well known that the speech signal contains a rich set of information that facilitates human auditory perception and communication, beyond a simple linguistic interpretation of the spoken input. In order to bridge the performance gap between ASR systems and human speech recognition (HSR), the narrow notion of speech-to-text in ASR has to be expanded to incorporate all related information "embedded" in speech utterances. This collection of information includes a set of fundamental speech sounds with their linguistic interpretations, a speaker profile encompassing gender, accent, emotional state and other speaker characteristics, the speaking environment, etc. Collectively, we call this superset of speech information the attributes of speech. They are not only critical for high performance speech recognition, but also useful for many other applications, such as speaker recognition, language identification, speech perception, speech synthesis, etc. ASAT therefore promises to be knowledge-rich and capable of incorporating multiple levels of information in the knowledge hierarchy into attribute detection, evidence verification and integration, i.e., all modules in the ASAT system [4]. Since speech processing in ASAT is highly parallel, a collaborative community effort can be built around a common sharable platform to enable a "divide-and-conquer" ASR paradigm that facilitates tight coupling of interdisciplinary studies of speech science and speech processing [4]. A block diagram of the ASAT approach to ASR is shown in Fig. 1. The top panel shows the general ASAT front end that performs a collection of speech analyses geared to

[1] This work re-organizes, expands, and completes our study reported in [35].

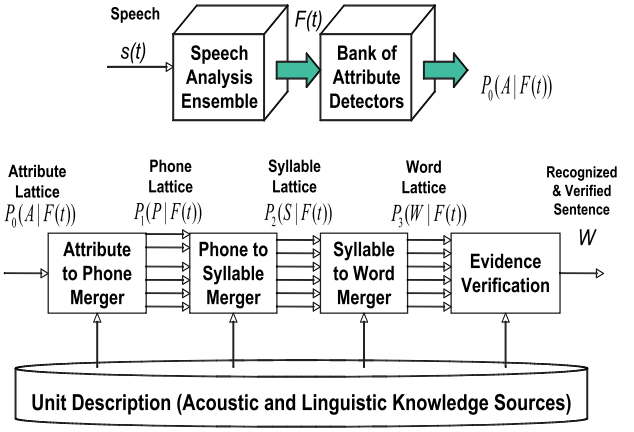


Fig. 1. A block diagram of the ASAT approach to ASR. On the top panel is the general ASAT front-end that contains a collection of speech analyzers each of which is geared to generate discriminative parameters. The bottom panel shows the ASAT backend knowledge integration model that produces all the intermediate sources of information in the speech knowledge hierarchy. This figure has been adapted from [4].

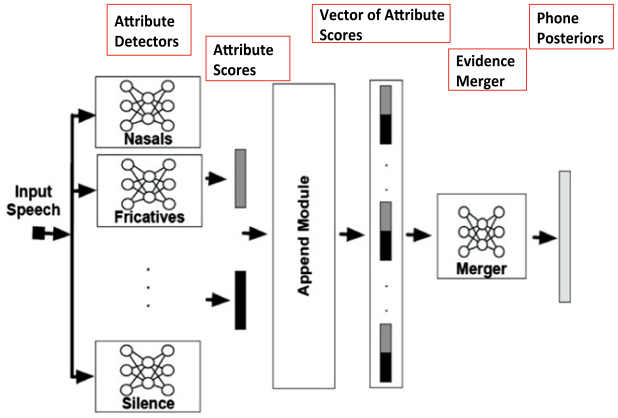


Fig. 2. The ASAT detection-based front-end. Each attribute detector analyzes any given input frame and produces a posterior probability score. The Append module stacks together attribute posteriors. The merger delivers phoneme posterior probabilities.

generate discriminative parameters. This ensemble of speech parameters are then used for further processing, including the design of a bank of speech attribute detectors to produce probabilistic scores for each of the key acoustic cues needed to generate evidences for making linguistic decisions. The bottom panel of Fig. 1 shows the ASAT backend knowledge integration that produces all the intermediate sources of information in the speech knowledge hierarchy.

The ASAT system is still under development, and the interested reader is referred to [4] for more details. In this study, we focus on the front-end of the ASAT framework, namely the bank of attribute detectors and the attribute-to-phoneme merger. The ASAT detection-based front-end as implemented in this paper is shown in Fig. 2 and consists of two main blocks: (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 "Append module" block, shown in Fig. 2, stacks together the outputs delivered by the attribute detectors for a given input and generates a supervector of attribute detection scores. This supervector is fed into the merger. Overall, the system shown in Fig. 2 maps acoustic features (e.g., short-time spectral features, or temporal pattern features) into phoneme 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 21 phonological features listed in Table 1. The merger discriminates among 40 phoneme classes shown in the third column of Table 1.

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

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

In Section 6, we show that such an architecture can achieve high-performance attribute and phoneme classification results and will be proven useful in word lattice rescoring implemented as described in [10].

3. Attribute and phoneme models in speech recognition

In the next two sections we survey research areas that can be impacted by the present study.

3.1. Attribute modeling

Broadly speaking, phonological-based approaches to speech recognition can be divided into two main categories depending on what they try to model, namely *abstract representation of articulation*, or *physical articulators*. The majority of these approaches fall into the first category. These studies are concerned with how to extract phonetic features and use them in the standard ASR framework (e.g., [11,12,43–46]). One of the main supporting arguments of these studies is that ASR engines can be improved by using more linguistically motivated features in addition to the standard frequency-based features. Articulatory features have

proven to have several nice properties, such as robustness to noise and cross-speaker variation [44], portability across different languages [12], and explicit modeling of linguistic information that makes it easier to deal with non-native and hyper-articulated speech [45]. These phonetic features can be extracted with data-driven techniques, and different approaches have been developed. For instance, a bank of artificial neural networks (ANNs) are used in [44] to score speech attributes. The attribute posterior probabilities generated by each ANN are concatenated and integrated by a higher-level ANN trained to produce phoneme posterior probabilities. A stream architecture to augment acoustic models based on context-dependent sub-word units with articulatory motivated acoustic units is proposed in [45]. Frame-level classification of a set of articulatory features inspired by the vocal tract variables of articulatory phonology is studied in [46]. In [43], it is shown that combining the recognition hypotheses resulting from the different articulatory specialized memories leads to significant phoneme recognition improvements. In [10], it is demonstrated that articulator information captured through a bank of phonetic feature detectors can be effectively used in a lattice rescoring process to correct utterances with errors in large vocabulary continuous speech recognition.

A smaller number of investigations have been aiming at modeling the physical articulators directly. Speech researchers in this area argue that better recognition performance can be obtained by using articulatory motivated modes, since that allows to model the underlying processes of co-articulation and assimilation directly, rather than describing their effects on the speech signal [47]. For instance, an articulatory feature based HMM recognizer is studied in [48]. In this system, each state represents a particular articulatory configuration instead of representing an acoustic portion of a phoneme. Combination with conventional acoustic-based HMM is carried out by a weighted sum of the log-likelihood of the models. This recognition paradigm is quite flexible and allows the modeling of articulatory asynchrony, yet it suffers from data sparseness problem because the articulatory based HMMs require a large state space. In [5], it is shown that integration of high-quality global speech production models into the probabilistic analysis-by-synthesis strategy has the potential to close the gap between humans and machines, which is the ultimate goal of ASR. Finally, attempts to use articulatory knowledge for visual speech recognition have been also pursued (e.g., [49]).

3.2. Phoneme estimation

Phoneme estimation is an active research area in speech recognition as well, since it plays a very important role in many speech processing applications. For instance, the out-of-vocabulary (OOV) words detection problem is often tackled by representing those words as phoneme sequences (e.g., [41]). In automatic language identification (LID), it has been shown that language recognition results are highly correlated with phoneme recognition results [(e.g., [42]). Furthermore, phoneme recognition also plays a fundamental role in improving speaker recognition [50] and word recognition accuracy [15,51].

Designing a high-performance phoneme model is a challenge and several researchers have proposed various phoneme architectures in the recent years. For example, high-accuracy phoneme recognition results have been reported by using several MLPs arranged in a hierarchical structure (e.g., [11,52,53]). A remarkable performance has been achieved on the TIMIT task [54] using deep belief networks [31]. In [55], the authors propose a conditional augmented model as a way to incorporate generative models in a discriminative classifier, which leads to good phoneme classification results. In [56], the authors further extended this coupling between generative models and a discriminative classifier by using HMMs as regressors in a penalized logistic regression framework. Finally, (hidden) conditional random fields have also been proposed [17,57].

4. Deep neural networks

A DNN is a multi-layer perceptron with many hidden layers. The main challenge in learning DNNs is to devise efficient training strategies in order to escape poor local optimum of the complicated nonlinear error surface introduced by the large number of hidden layers . A common practice 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 performing a joint optimization of all the layers [37]. This learning strategy enables discriminative training to start from well initialized weights and is used in this study.

4.1. Restricted Boltzmann machines

A bipartite graph with a visible layer and a hidden layer can be used to represent an RBM. 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 commonly represented with Bernoulli distributions. Gaussian–Bernoulli RBMs can convert real-valued stochastic variables (such as short-term spectral features) to binary stochastic variables that can then be further processed using the Bernoulli–Bernoulli RBMs.

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

$$p(\boldsymbol{v},\boldsymbol{h};\theta)=\frac{\exp(-E(\boldsymbol{v},\boldsymbol{h};\theta))}{Z}, \tag{1}$$

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

$$p(\boldsymbol{v};\theta)=\frac{\sum_h\exp(-E(\boldsymbol{v},\boldsymbol{h};\theta))}{Z}. \tag{2}$$

For a Bernoulli (visible)–Bernoulli (hidden) RBM, the energy is

$$E(\boldsymbol{v},\boldsymbol{h};\theta)=-\sum_{i=1}^{V}\sum_{j=1}^{H}w_{ij}v_ih_j-\sum_{i=1}^{V}b_iv_i-\sum_{j=1}^{H}a_jh_j, \tag{3}$$

where w_{ij} represents the symmetric interaction between visible unit v_i and hidden unit h_j, b_i and a_j are the bias terms at the visible and hidden layers, respectively, and V and H are the numbers of visible and hidden units. The conditional probabilities can be efficiently calculated as

$$p(h_j=1\,|\,\boldsymbol{v};\theta)=\sigma\left(\sum_{i=1}^{V}w_{ij}v_i+a_j\right), \tag{4}$$

$$p(v_i=1\,|\,\boldsymbol{h};\theta)=\sigma\left(\sum_{j=1}^{H}w_{ij}h_j+b_i\right), \tag{5}$$

where $\sigma(x)=1/(1+\exp(-x))$.

Similarly, for a Gaussian–Bernoulli RBM, the energy is

$$E(\boldsymbol{v},\boldsymbol{h};\theta)=-\sum_{i=1}^{V}\sum_{j=1}^{H}w_{ij}v_ih_j+\frac{1}{2}\sum_{i=1}^{V}(v_i-b_i)^2-\sum_{j=1}^{H}a_jh_j. \tag{6}$$

The corresponding conditional probabilities become

$$p(h_j=1\,|\,\boldsymbol{v};\theta)=\sigma\left(\sum_{i=1}^{V} w_{ij}v_i+a_j\right), \tag{7}$$

$$p(v_i=1\,|\,\boldsymbol{h};\theta)=\mathcal{N}\left(\sum_{j=1}^{H} w_{ij}h_j+b_i,1\right), \tag{8}$$

where v_i takes real values and follows a conditional Gaussian distribution with mean $\sum_{j=1}^{H} w_{ij}h_j+b_i$ and variance one.

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

$$\Delta w_{ij}=\langle v_i h_j\rangle_{data}-\langle v_i h_j\rangle_{model}, \tag{9}$$

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 the 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 [37].

4.2. 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

$$p(l=k|\boldsymbol{h};\theta)=\frac{\exp(\sum_{i=1}^{H}\lambda_{ik}h_i+a_k)}{Z(\boldsymbol{h})}, \tag{10}$$

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, and 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, $E(h_j|\boldsymbol{v};\theta)=p(h_j=1|\boldsymbol{v};\theta)$ from the previous layer is used as the visible input vector based on the mean-field theory. This process continues until the last layer, where 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.

5. Word lattice rescoring

The final step in the present study is to verify whether DNN boosted accuracies can help improve the LVCSR performance, and we pursued this goal by integrating the information generated at the output of the DDN phoneme classifier, namely phoneme posterior probability, into an existing LVCSR system through the word lattice rescoring procedure outlined in the ASAT study reported in [10].

Note that a conventional ASR system can output either a single best hypothesis (that is, the decoded sentence) or a set of most likely sentence hypotheses for a given input utterance. In the latter case, the competing most likely hypotheses can be arranged in the form of a word lattice. In our investigation, we adopted the word lattice structure described in [58], which reflects the syntactic constraints of the grammar used during recognition. Thus, the work lattice is implemented as a direct, acyclic, and weighed graph, *G(N, A)*, with *N* nodes and *A* arcs. As shown in Fig. 3, the nodes of the lattice carry the timing information (i.e., temporal boundaries are given by the arcs's bounding nodes),

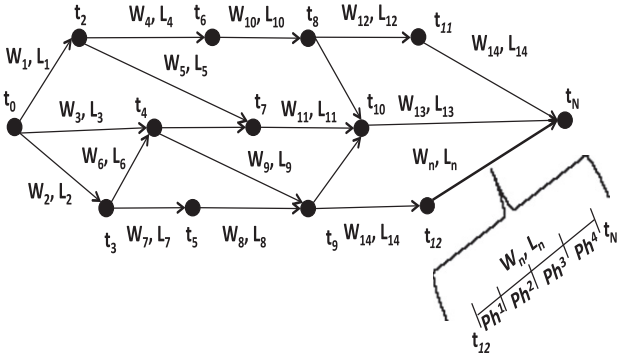


Fig. 3. An example of a word lattice implemented as a direct, acyclic, and weighted graph. The timing information is embedded in the nodes, whereas each arc carries out information about the word identity and its likelihood. The enlargement shows the generic *i*th word, W_i described as a sequence of phonemes.

whereas the arcs carry the symbol along with the score information. In particular, each arc corresponds to a recognized word. The rescoring algorithm aims at incorporating the scores generated by the ASAT detection-based front-end shown in Fig. 2 into speech word lattices. These word lattices are generated by the LVCSR baseline system to be presented in Section 6.1, and it is inspired by decoding scheme based on a generalized confidence score proposed in [59].

The rescoring formulation for word lattices is 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 phoneme scores, PS^i, of each phoneme composing that word (see [10] for more details). Thus *WS* is a linear combination of phoneme scores. In turn, PS^i is computed by summing up all of (log) posterior probabilities, optionally discounted by the prior probability, generated by DNN (or MLP) for that phoneme class. The weighted rescoring formula is defined as

$$S_n=w_1W_n+w_2L_n, \tag{11}$$

where W_n is defined as

$$W_n=\sum_{i=1}^{K} PS_n^i. \tag{12}$$

PS_n^i is the score of the i-th phoneme in the n-th arc, K is the number of phonemes 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. Both w_1 and w_2 are set to 0.5 in our experiments.

6. Experiments

In the following sections, the experimental setup is presented, and the results on attribute and phoneme classification are discussed. Word recognition results through lattice rescoring are also given. Nonetheless, before delving into the experimental part of the proposed work, the rational behind the use of neural architectures with more than one hidden layer for speech applications is now briefly discussed.

The nature of the speech signal is such that feature vectors extracted for different phonemic or phonetic classes greatly overlap in the input feature (hyper-)space. In [60], for example, the authors have found that there is a great overlap between formant frequencies for different vowel sounds by different talkers. More recently, it has also been demonstrated that Bhattacharyya distance distributions between 39-dimension MFCCs for the bilabial class and 39-dimension

MFCCs for the alveolar class is rather small [61]. These two pieces of experimental evidence imply that speech data lie on or near a nonlinear manifold, as pointed out in [62].

Shallow MLPs have the potential to learn good models of data that lie on or near a nonlinear manifold, but it has not been possible so far to seriously address the speech problem using a single hidden layer while avoiding over-fitting [62]. Deep neural networks are instead better suited to learn the kind of complicated functions that can represent high-level abstractions that are typical in speech problems [63]. Moreover, DNNs can implement arbitrary complex decision boundary with fewer overall hidden units than single hidden layer neural networks [14], and DNNs can learn more invariant and discriminative features at higher hidden layers and less likely to over-fit [62]. Finally, experimental evidence reported in [64] seems to suggest that DNNs are able to learn more appropriate features in their lower layers than shallow MLPs and are therefore better suited for speech applications.

6.1. Experimental setup

All experiments were conducted on the 5000-word speaker independent WSJ0 (5k-WSJ0) task [40]. The parameters of all classifiers presented in this study were estimated using training material from the SI-84 set (7077 utterances from 84 speakers, i.e., 15.3 h of speech). A cross-validation (cv) set was generated by extracting 200 sentences out of the SI-84 training set. The cv set accounts for about 3% of the SI-84 set and was used to terminate the training. The remaining 6877 SI-84 sentences were used as training material. Evaluation was carried out on the Nov92 evaluation data (330 utterances from 8 speakers).

Mel-frequency cepstrum coefficients (MFCCs) [65] were used as parametric representation of the audio signal. Spectral analysis to generate MFCCs was performed using a 23 channel Mel filter bank from 0 to 8 kHz. The cepstral analysis was carried out with a Hamming window of 25 ms and a frame shift of 10 ms. For each frame, 12 MFCC features plus the zeroth cepstral coefficient were computed. The first and second time derivatives of the cepstra were computed as well and concatenated to the static cepstra to yield a 39-dimensional feature-vector.

For the word lattice rescoring experiment, a gender independent LVCSR baseline systems was built. This system was designed with the HTK toolkit [66] and is based on tied-state cross-word triphone models and a trigram language model. The number of shared states is 2818, and these states were obtained with a phonetic decision tree and each state observation density was modeled by a GMM with 8 mixture components. The HMM parameters of the LVCSR baseline system were estimated using maximum mutual information (MMI) estimation procedure [67]. A language model within the 5k-WSJ0 vocabulary was used during decoding. The acoustic vector used to represent the speech signal contains 12 MFCCs, log energy, velocity, and acceleration coefficients.

6.2. Results on attribute classification

As stated, each detector estimates attribute posterior probabilities. Table 2 shows the classification accuracies at a frame level for the speech attributes used in this work. In this table, the prior probability of each attribute, P(attribute), 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. This information has been added to the Table 2 to demonstrate that the proposed solution can attain a classification result better than chance. The shallow MLP results obtained using MFCC input features were quoted from [35] and were obtained using a single hidden layer MLP with 800 hidden units. These results are referred to as shallow MLP in Table 2. The attribute accuracy obtained using a DNN is reported in the next-to-last column of Table 2. The DNN contains five hidden layers each with 2048 units following the previous work [33].

From this table we observe that higher attribute accuracies can be delivered using a DNN trained over short-time spectral features, that is, MFCCs. Furthermore, for many attributes, such as back, labial, and mid, the shallow MLP with a single hidden layer performs only slightly better than the Naïve approach. The DNN achieves a much higher accuracy with relative error rate reductions over the shallow MLP ranging from 40% to 90% for different attributes. Furthermore, the average relative error rate is reduced by 56% across all attributes over the shallow MLP. For instance, very good attribute classification accuracies can be obtained for several attributes, such as dental (99.0%) and voiced (95.4%).

6.3. Results on phoneme classification

Table 3 summarizes the average cross entropy (CE) and classification accuracies at the frame level for phonemes. The setup names are encoded as $\sharp$*_hidden_units* $\times$ $\sharp$*_hidden_layer input_feature*, where the input feature MFCC is extracted as described in Section 6.1, input features *Attr*1 and *Attr*2 refer to the attribute log posterior probability generated from the 800×1 MLP and 2048×5 DNN attribute detectors, respectively. All the four setups used 11 frames of features—5 frames looking ahead and 5 frames looking back.

From Table 3 we can make several observations. First, the shallow MLP based phoneme detector performs the worst even though it used the attribute detectors results as the input feature.

Table 2
Classification accuracies (in %) at a frame level for the speech attributes used in this work.

Attribute	P (attribute)	Naïve	Shallow MLP	DNN (2084×5)	STC
Anterior	36.2	63.8	85.6	92.5	93.2
Approximant	9.2	90.8	94.9	96.4	95.9
Back	19.6	80.4	87.6	93.1	92.9
Continuant	55.7	55.7	88.7	93.5	89.93
Coronal	25.5	74.5	87.9	92.4	93.1
Dental	1.4	98.9	98.9	99.0	99.1
Fricative	15.3	84.7	94.2	96.2	95.4
Glottal	0.8	99.2	99.3	99.7	99.7
High	16.7	83.3	90.7	95.0	94.9
Labial	11.0	89.0	92.5	96.9	92.5
Low	9.3	90.7	94.6	96.9	96.9
Mid	11.8	88.2	90.7	93.8	93.6
Nasal	8.7	91.3	95.9	97.7	97.1
Retroflex	6.2	93.8	97.6	98.5	98.4
Round	14.7	85.3	91.9	94.9	93.4
Stop	15.3	84.7	92.9	95.7	94.9
Tense	39.5	60.5	83.0	90.6	90.5
Velar	5.4	94.6	96.6	98.7	98.4
Voiced	59.9	59.9	92.1	95.3	95.4
Vowel	32.5	67.5	87.9	92.8	91.3

Table 3
Average cross entropy (CE) and phoneme classification accuracy at a frame level.

Setup	Train		cv		Test	
	avg CE	acc(%)	avg CE	acc(%)	avg CE	acc(%)
1500×1 *Attr*1	–	86.7	–	82.7	–	82.6
2048×5 MFCC	−0.26	91.6	−0.45	85.3	−0.46	85.1
2048×7 MFCC	−0.24	91.9	−0.45	85.5	−0.46	85.3
2048×5 *Attr*2	−0.28	90.2	−0.45	85.5	−0.48	85.0
2048×5 STC	−0.12	95.9	−0.37	88.8	–	88.3

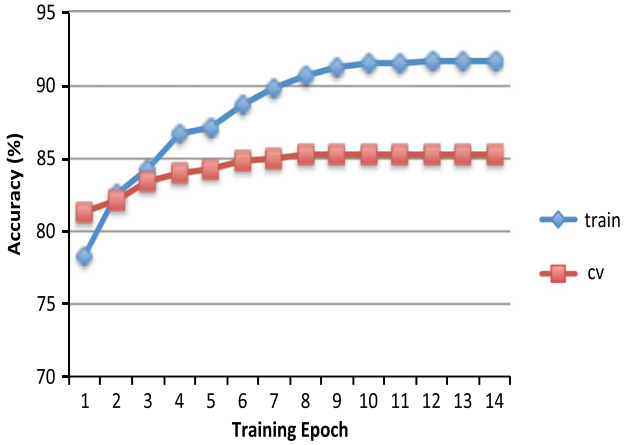


Fig. 4. DNN accuracy on the training and cv data using MFCCs.

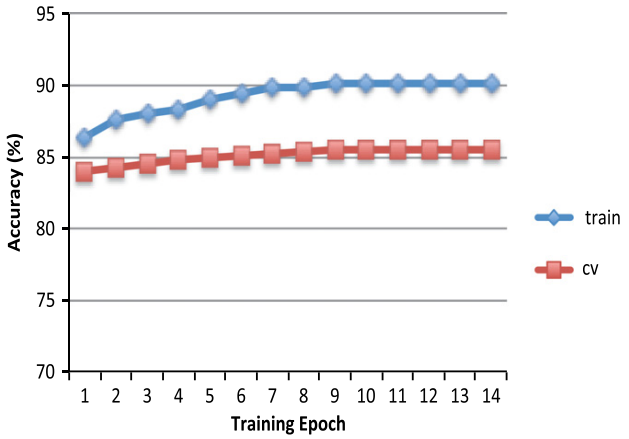


Fig. 5. DNN accuracy on the training and cv data using *Attr*2.

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. Finally, comparing cv and test set results we can see that the DNN results are robust.

Figs. 4 and 5 display the phoneme accuracies attained by a DNN with 5-hidden layer on the training and cv data at different training epochs. Specifically, the accuracies using MFCCs are reported in Fig. 4, whereas the phoneme accuracies using *Attr*2 are displayed in Fig. 5. A comparison between these two plots suggests that by breaking the phoneme detector into two stages—first to detect the attribute and then to estimate the phoneme identity based on the results of attribute detectors, higher phoneme accuracies can be achieved sooner. Nonetheless, this two-stage approach has not provided any gain over the direct approach that detect phonemes trained over MFCCs if DNN is used although the same two-stage detector did show advantages if shallow MLP or other shallow model (e.g. [68]) is used. This indicates that DNNs are powerful enough to capture useful discriminative information, and the performance on the test data reported in Table 3 confirms it. In this study, we have not exploited the temporal overlapping characteristics of speech attributes across different dimensions. Therefore, we believe that the potential of the deep layered structure may be proven useful in the future.

It has been demonstrated that better phoneme accuracies can be delivered by employing long-time temporal patterns of spectral energies in place of the conventional short-term spectral features (e.g., MFCCs) [53]. It would therefore be meaningful to verify whether further improvement on phoneme classification

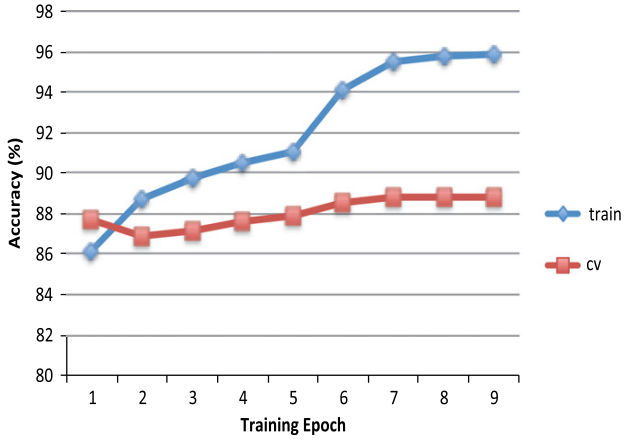


Fig. 6. DNN accuracy on the training and cv data using long-time temporal patterns of spectral energies.

can be gained by using long-time temporal based features. Each attribute classifier was thus trained using split temporal context features [52]. Specifically, spectral analysis was performed using a 23 channel Mel filter bank from 0 to 8 kHz. For each critical band a window of 310 ms centered around the frame being processed was considered and split in two halves: left-context and right-context . Two independent front-end as shallow MLPs (“lower nets”) were trained on those two halves and generated left- and right-context attribute posterior probabilities, respectively. The discrete cosine transform was applied to the input of these lower nets to reduce the dimensionality. The outputs of the two lower nets were then sent to a third MLP which acted as a merger and gave the attribute posterior probability of the target attribute. The attribute classification accuracy with STC features is reported in the last column of Table 2. It is worth noting that by arranging MLPs in a hierarchical structure and using long-time temporal information, results comparable to the DNN detectors with MFCC features can be attained. The output of these 21 attribute classifiers was combined into a supervector, as explained earlier. This supervector is fed into the merger shown in Fig. 2. We did not classify attributes using DNN and STC features since when DNN is used in the two-stage architecture with MFCC features, good phoneme recognition accuracy have been attained, as shown in Table 3.

Fig. 6 shows the framewise training and cv phoneme accuracy at different training epochs for a DNN when long temporal patterns of spectral energy are used. The final training and cv accuracies are reported in Table 3, 2048 $\times$ 5 STC, and these accuracies are equal to 95.9% and 88.8%, respectively. The framewise test accuracy is of 88.3%, as shown in the last column of Table 3. This result represents the best accuracy reported in this work.

In [69], it was shown that by using log filter bank features we can get much better results than using MFCC features if the same number of layers is used and sometimes even less layers are used, on real world voice search datasets. In our laboratories, evidence has also been gained to indicate that DNN is more robust than shallow MLP and GMM when noise exists. Specifically, we have got very remarkable results on the Aurora 4 task by simply plugging in the CD-DNN-HMM without using any noise robustness technique. Nonetheless, such experiments are out of the topic of this work and will be the subject of a future publication.

6.4. Results on word lattice rescoring

The performance of the MMI-based baseline system is reported in Table 4 in terms of word error rate (WER) on the Nov92 task. This result is comparable with the baseline result

Table 4
WER, in %, on the Nov92 task. Rescoring was applied to the MMI-based baseline systems, trained on the SI-84 material of the WSJ0 corpus.

System	WER (%)
MMI-based baseline	4.6
Shallow MLP-based rescoring	4.5
DNN-based rescoring	4.2
DNN-based rescoring with long-term energy trajectories	4.0

reported in [70,71]. Table 4 shows, in the second row, the performance of the rescored system when a shallow MLP is used to implement both the bank of attribute classifier and the merger. Eq. (11) with $w_1 = w_2 = 0.5$ is used to carry rescoring phase out. The same rescoring procedure is also carried out over the MMI baseline system using phoneme posterior probabilities generated using DNNs for both the bank of detectors and the merger, that is using the configuration 2048 *Attr*2 in Table 3. The interpolation weights in Eq. (12) are again clamped to 0.5. These results indicate that the rescored systems always achieve better performance than the conventional baseline system due to the system combination effect. Furthermore, DNN-based rescored system outperforms the MLP-based one, and a final WER of 4.2% is attained, which correspond to a relative improvement of 8.7%. The relative improvement represents the relative reduction of WER achieved by the DNN-based rescoring over the MMI baseline system, and it is computed as $(WER_{MMI} - WER_{RESCORED})/WER_{MMI}$. This performance also represents an improvement over our previously reported results on word lattice rescoring [10].

In Section 6.3, it was demonstrated that the use of long-temporal evolution of spectral energies in place of MFCCs allows us to obtain better phoneme classification accuracies. Therefore, word lattice rescoring was performed with those boosted phoneme posterior probabilities, and a final WER of 4.0% was observed (see last row in Table 4). This result corresponds to a relative performance improvement of 13.0%.

7. Discussion and conclusion

We have demonstrated in this work that we can achieve high accuracies for both phonological attribute detection and phoneme estimation using DNNs. Furthermore, DNN-based rescoring has proven useful in an LVCSR application. This opens up new opportunities to some old problems, such as speech recognition from a phoneme lattice [2] and from phonological parsing [72]. It also creates an exciting avenue to provide high-precision attribute and phoneme lattices for bottom-up, detection-based speech recognition where words can be directly specified in terms of attributes free from phonemes. For speech understanding, concepts may be also directly specified in terms of attributes free from words.

It is well known that speech utterances can be characterized by two types of acoustic/phonetic cues in spectrogram reading. The first kind of cues are relatively stable, e.g. voicing, places, and manners of articulation. On the other hand, there are many more cues that varies strongly with context, e.g. flapping and spectral manifestation of vowels. By combining these cues and other higher level information such as lexical constraints, researchers repetitively demonstrated that with some training, they can "interpret" the underlying sentence embedded in a speech utterance. The gist seems to lie in the relatively reliable "human visual detection" of such acoustic landmarks, or events, and "bottom-up knowledge integration" of these linguistic events. This process of integrating diverse knowledge sources as the basis for ASR could be accomplished by *phonological parsing* (e.g. [72]) and related techniques (e.g. [73]).

Furthermore, in spontaneous speech partial understanding is often needed because an integrated approach is not sufficient to properly capture the overall knowledge sources. Thus for ill-formed utterances in spontaneous speech it is expected that the proposed framework will be robust and will give a better performance than the standard HMM-based technology as demonstrated previously in the proposed key-phrase detection frameworks [3]. The main challenge here is robust connection between the target concepts in the understanding tasks and attribute specification and robust detection of such attributes. With our initial success reported in this paper, we intend to continue to explore the cross-fertilization of ASAT and DNNs for LVCSR and for potential speech understanding applications.

One clear limitation of the current framework in 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 phoneme-to-attribute mapping (Table 1), and it may account for why the use of attribute detectors has not led to superior phoneme classification as compared to DNN directly trained over MFCC features. Nonetheless, such asynchrony is central to modern phonological theory (see a review in [74]). Incorporation of asynchrony will significantly modify the attribute targets in running speech in a principled and parsimonious way, as demonstrated in [5,7,8,75]. For spontaneous speech that exhibits significantly more variation in pronunciation than read-style speech, such asynchrony plays a more important role. With the attribute targets modified in a phonologically meaningful manner, we hope that the DNN approach will further enhance the value of the attributes for making word recognition and spontaneous speech recognition more accurate within the detection-based ASAT framework.

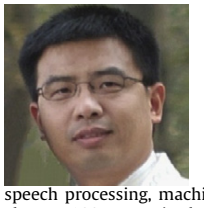
References

[1] C.-H. Lee, Q. Huo, On adaptive decision rules and decision parameter adaptation for automatic speech recognition, Proc. IEEE 88 (8) (2000) 1241–1269.
[2] S.E. Levinson, Structural methods in automatic speech recognition, Proc. IEEE 73 (1985) 1625–1650.
[3] T. Kawahara, C.-H. Lee, B.-H. Juang, Flexible speech understanding based on combined key-phrase detection and verification. IEEE Transactions on Speech and Audio Processing 6 (6) (1998) 558–568.
[4] C.-H. Lee, M.A. Clements, S. Dusan, E. Fosler-Lussier, K. Johnson, B.-H. Juang, L.R. Rabiner, An overview on automatic speech attribute transcription (ASAT), in: Proceedings of the Interspeech, Antwerp, Belgium, 2007, pp. 1825–1828.
[5] L. Deng, D. Sun, A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features, J. Acoust. Soc. Am. 85 (5) (1994) 2702–2719.
[6] K. Kirchhoff, Robust Speech Recognition Using Articulatory Information, Ph.D. Thesis, University of Bielefeld, 1999.
[7] S. King, P. Taylor, Detection of phonological features in continuous speech using neural networks, Comput. Speech Lang. 14 (4) (2000) 333–345.
[8] K. Livescu, Feature-Based Pronunciation Modeling for Automatic Speech Recognition. Ph.D. Thesis, MIT, September 2005.
[9] E. Eide, Distinctive Features for Use in an Automatic Speech Recognition System, in: Proceedings of the Eurospeech 2001, Aalborg, Denmark, pp. 1613–1616.
[10] S.M. Siniscalchi, C.-H. Lee, A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition, Speech Commun. 51 (11) (2009) 1139–1153.
[11] S.M. Siniscalchi, T. Svendsen, C.-H. Lee, Towards bottom-up continuous phone recognition, in: Proceedings of the ASRU, Kyoto, Japan, 2007, pp. 566–569.
[12] S.M. Siniscalchi, D.-C. Lyu, T. Svendsen, C.-H. Lee, Experiments on cross-language attribute detection and phone recognition with minimal target-specific training data, IEEE Trans. on Audio, Speech, and Language Processing, 20 (3) (2012) 875–887.
[13] S.M. Siniscalchi, J. Reed, T. Svendsen, C.-H. Lee, Exploring universal attribute characterization of spoken languages for spoken language recognition, in: Proceedings of the Interspeech, Brighton, UK, 2009, pp. 168–171.
[14] C.M. Bishop, Neural networks for pattern recognition, Oxford University Press, New York, USA, 1995.
[15] H. Bourlard, N. Morgan, Connectionist Speech Recognition: A Hybrid Approach, Kluwer, Boston, USA, 1994.

[16] H. Hermansky, D.P. Ellis, S. Sharma, Connectionist feature extraction for conventional HMM systems, in: Proceedings of the ICASSP, Instanbul, Turkey, 2000, pp. 1635–1638.
[17] J. Morris, E. Folser-Lussie, Combining phonetic attributes using conditional random fields, in: Proceedings of Interspeech, Pittsburgh, Pennsylvania, USA, 2006, pp. 597–600.
[18] S. Abdou, M.S. Scordillis, Beam search pruning in speech recognition using a posterior-based confidence measure, Speech Commun. 42 (3–4) (2004) 409–428.
[19] G. Bernardis, H. Bourlard, Improving posterior confidence measures in hybrid HMM/ANN speech recognition system, in: Proceedings of the ICSLP, Sydney, Australia, 1998, pp. 775–778.
[20] D. Yu, J. Li, L. Deng, Calibration of confidence measures in speech recognition, IEEE Trans. Audio Speech Lang. Process. (November) (2011) 2461–2473.
[21] J.S. Bridle, Alpha-nets: a recurrent "neural" network architecture with a hidden Markov model interpretation, Speech Commun. 9 (1) (1990) 83–92.
[22] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Networks 18 (2005) 602–610.
[23] K.J. Lang, A.H. Waibel, G.E. Hinton, A time-delay neural network architecture for isolated word recognition, Neural networks 3 (1) (1990) 24–43.
[24] T. Chen, H. Chen, R. Liu, Approximation capability in $C(R^n)$ by multilayer feedforward networks and related problems, IEEE Trans. Neural Networks 6 (1) (1995) 25–30.
[25] G.E. Hinton, S. Osindero, Y. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.
[26] V. Nair, G. Hinton, 3-d object recognition with deep belief nets, Adv. Neural Inf. Process. Syst. 22 (2009) 1339–1347.
[27] R. Salakhutdinov, G. Hinton, Semantic hashing, Int. J. Approximate Reasoning 50 (7) (2009) 969–978.
[28] R. Collobert, J. Weston, A unified architecture for natural language processing: deep neural networks with multitask learning, in: Proceedings of 25th international conference on Machine learning, ser. ICML, New York, USA, 2008, pp. 160–167.
[29] V. Mnih, G. Hinton, Learning to detect roads in high-resolution aerial images, in: Proceedings of ECCV, Crete, Greece, 2010, pp. 210–223.
[30] K. Jarrett, K. Kavukcuoglu, M. Ranzato, Y. LeCun, What is the best multi-stage architecture for object recognition? in: Proceedings of ICCV, Kyoto, Japan, 2009, pp. 2146–2153.
[31] A. Mohamed, G.E. Dahl, G.E. Hinton, Acoustic modeling using deep belief networks, IEEE Trans. Audio Speech Lang. Proc. 20 (1) (2012) 14–22.
[32] A. Mohamed, D. Yu, L. Deng, Investigation of full-sequence training of deep belief networks for speech recognition, in: Proceedings of Interspeech, Makuhari, Japan, 2010, pp. 1692–1695.
[33] G.E. Dahl, D. Yu, L. Deng, A. Acero, Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition, IEEE Trans. Audio Speech Lang. Proc. 20 (1) (2010) 30–42.
[34] F. Seide, G. Li, D. Yu, Conversational speech transcription using context-dependent deep neural networks, in: Proceedings of Interspeech, Florence, Italy, 2011, pp. 437–440.
[35] D. Yu, S.M. Siniscalchi, L. Deng, C.-H. Lee, Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition, in: Proceedings of ICASSP, Kyoto, Japan, 2012, pp. 4169–4172.
[36] S.M. Siniscalchi, T. Svendsen, C.-H. Lee, A bottom-up stepwise knowledge-integration approach to large vocabulary continuous speech recognition using weighted finite state machines, in: Proceedings of Interspeech, Florence, Italy, 2011, pp. 901–904.
[37] G.E. Hinton, Training products of experts by minimizing contrastive divergence, Neural Comput. 14 (2002) 1771–1800.
[38] D. Yu, M.L. Seltzer, Improved bottleneck features using pretrained deep neural networks, in: Proceedings of Interspeech, Florence, Italy, 2011, pp. 237–240.
[39] D. Yu, L. Deng, G. Dahl, Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition, in: Proceedings of NIPS Workshop on Deep Learning and Unsupervised Feature Learning, 2010.
[40] D.B. Paul, J.M. Baker, The design for the Wall Street Journal-based CSR corpus, in: Proceedings of ICSLP, 1992.
[41] I. Bazzi, J. Glass, Modeling out-of-vocabulary words for robust speech recognition, in: Proceedings of ICSLP, Beijing, China, 2000, pp. 401–404.
[42] P. Matějka, P. Schwarz, J. Černocký, P. Chytil, Phonotactic language identification using high quality phoneme recognition, in: Proceedings of Interspeech, Lisboa, Portugal, 2005, pp. 2237–2240.
[43] S. Demange, S. Ouni, Continuous episodic memory based speech recognition using articulatory dynamics, in: Proceedings of Interspeech, Florence, Italy, 2011, pp. 2305–2308.
[44] K. Kirchhoff, Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments, in: Proceedings of ISCLP, Sydney, Australia, 1998, pp. 891–894.
[45] F. Metze, Articulatory Features for Conversational Speech Recognition, Ph.D. Dissertation, University of Karlsruhe, Germany, 2005.
[46] J. Morris, E. Folser-Lussie, Combining phonetic attributes using conditional random fields, in: Proceedings of Interspeech, Pittsburgh, PA, USA, 2006, pp. 597–600.
[47] J. Frankel, M. Western, S. King, Articulatory feature recognition using dynamic Bayesian networks, Comput. Speech Lang. 21 (2007) 620–640.
[48] M. Richardson, J. Bilmes, C. Diorio, Hidden articulator Markov models for speech recognition, Speech Commun. 41 (2) (2003) 511–529.
[49] K. Saenko, K. Livescu, J. Glass, T. Darrell, Multistream articulatory feature-based models for visual speech recognition, IEEE Trans. Pattern Anal. Mach. Intell. 31 (9) (2009) 1700–1707.
[50] Q. Jin, T. Schultz, A. Waibel, Speaker identification using multilingual phone strings? in: Proceedings of ICASSP, Orlando, USA, 2002, pp. 145–148.
[51] B. Chen, Q. Zhu, N. Morgan, Learning long-term temporal features in LVCSR using neural networks, in: Proceedings of ICSLP, Jeju Island, Korea, 2004, pp. 612–615.
[52] P. Schwarz, P. Matějaka, J. Černocký, Hierarchical structures of neural networks for phoneme recognition, in: Proceedings of ICASSP, Toulouse, France, 2006, pp. 325–328.
[53] H. Hermansky, S. Sharma, Temporal Patterns (TRAPS) in ASR of Noisy Speech, in: Proceedings of ICASSP, Phoenix, AZ, USA, 1999, pp. 289–292.
[54] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgren, DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus, U.S. Dept. of Commerce, NIST, Gaithersburg, USA, 1993.
[55] M. Layton, M. Gales, Augmented statistical models for speech recognition, in: Proceedings of ICASSP, Toulouse, France, 2006, pp. 129–132.
[56] Ø. Birkenes, T. Matsui, K. Tanabe, S.M. Siniscalchi, T. Myrvoll, M.H. Johnsen, Penalized logistic regression with HMM log-likelihood regressors for speech recognition, IEEE Trans. Speech Audio Process. 18 (2010) 1440–1454.
[57] A. Gunawardana, M. Mahaja, A. Acero, Hidden conditional random fields for phone classification, in: Proceedings of Interspeech, Lisboa, Portugal, 2005, pp. 1117–1120.
[58] W. Macherey, L. Haferkamp, R. Schlüer, H. Ney, Investigations on error minimizing training criteria for discriminative training in automatic speech recognition, in: Proceedings of Interspeech, Lisboa, Portugal, Sept. 2005, pp. 2133–2136.
[59] M.-W. Koo, C.-H. Lee, B.-H. Juang, Speech recognition and utterance verification based on a generalized confidence score, IEEE Speech Audio Process. 9 (8) (2001) 821–831.
[60] G.E. Peterson, H.L. Barney, Control methods used in a study of the vowels, J. Acoust. Soc. Am. 24 (1952) 175–194.
[61] O. Scharenborg, V. Wan, R.K. Moore, Towards capturing fine phonetic variation in speech using articulatory features, Speech Commun. 49 (2007) 811–826.
[62] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition: the Shared Views of Four Research Groups, IEEE Signal Processing Magazine, vol. 29, no. 6, 2012, pp. 82–97.
[63] Y. Bengio, Learning deep architectures for AI, Found. Trends Mach. Learn. 2 (2009) 1–127.
[64] F. Seide, G. Li, X. Chen, D. Yu, Feature engineering in context-dependent deep neural networks for conversational speech transcription, in: Proceedings of IEEE ASR, HI, USA, December 2011, pp. 24–29.
[65] S.B. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. Acoustic Speech Signal Process. 28 (4) (1980) 357–366.
[66] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P.C. Woodland, The HTK Book (for HTK Version 3.3), Cambridge University Press, Cambridge, UK, 2005.
[67] L.R. Bahl, F. Jelinek, R.L. Mercer, A maximum likelihood approach to continuous speech recognition, IEEE Trans. Pattern Anal. Mach. Intell. 5 (2) (1983) 179–190.
[68] U.V. Chaudhari, M. Picheny, Articulatory feature detection with SVM for integration into ASR and Phone recognition, in: Proceedings of ASRU, Merano, Italy, 2009, pp. 93–98.
[69] J. Li, D. Yu, J.-T. Huang, Y. Gong, Improving wideband speech recognition using mixed-bandwidth training data in CD-DDD-HMM, in: Proceedings of Workshop on Spoken Language Technology, 2012.
[70] J. Li, M. Yuan, C.-H. Lee, Approximate test risk bound minimization through soft margin estimation, IEEE Trans. Audio Speech Lang. Process. 15 (8) (2007) 2393–2404.
[71] Q. Fu, Y. Zhao, B. Juang, Automatic speech recognition based on non-uniform error criteria, IEEE Trans. Audio Speech Lang. Process. 20 (3) (2012) 780–793.
[72] K. Church, Phonological Parsing in Speech Recognition, Ph.D. Thesis, MIT, 1986.
[73] J. Sun, L. Deng, An overlapping-feature based phonological model incorporating linguistic constraints: applications to speech recognition, J. Acoust. Soc. Am. 111 (2) (2002) 1086–1101.
[74] L. Deng, D. O'Shaughnessy, Speech Processing—A Dynamic and Optimization-Oriented Approach, Marcel Dekker Inc., 2003. (June).
[75] L. Deng, Articulatory features and associated production models in statistical speech recognition, Computational Models of Speech Pattern Processing, Springer, 1999, pp. 214–224.

Sabato Marco Siniscalchi is an Assistant Professor at the University of Enna "Kore," and affiliated with the Georgia Institute of Technology. He received his Laurea and Doctorate degrees in Computer Engineering from the University of Palermo, Palermo, Italy, in 2001 and 2006, respectively. In 2001, he was employed by STMicroelectronics where he designed optimization algorithms for processing digital image sequences on very long instruction word (VLIW) architectures. In 2002, he was an Adjunct Professor at the University of

Palermo and taught several undergraduate courses for Computer and Telecommunication Engineering. From 2005 to 2007, he was a Post Doctoral Fellow at the Center for Signal and Image Processing (CSIP), Georgia Institute of Technology, Atlanta, under the guidance of Prof. C.-H. Lee. From 2007 to 2009, he joined the Norwegian University of Science and Technology, Trondheim, Norway, as a Research Scientist in the Department of Electronics and Telecommunications under the guidance of Prof. T. Svendsen. In 2010, he was a Researcher Scientist in the Department of Computer Engineering, University of Palermo, Italy. His main research interests are in speech processing, in particular automatic speech and speaker recognition, and language identification.



Dong Yu joined Microsoft Corporation in 1998 and Microsoft Speech Research Group in 2002, where he is a researcher. He holds a Ph.D. degree in Computer Science from the University of Idaho, an MS degree in Computer Science from Indiana University at Bloomington, an MS degree in Electrical Engineering from Chinese Academy of Sciences, and a BS degree (with honor) in Electrical Engineering from Zhejiang University (China). His current research interests include speech processing, machine learning, and pattern recognition. He has published close to 100 papers in these areas and is the inventor/coinventor of more than 40 granted/pending patents. Most recently, he has been focusing on deep learning and its applications in speech processing. The context dependent deep neural network hidden Markov model (CD-DNN-HMM) that he co-proposed and developed has been seriously challenging the dominant position of the Gaussian mixture hidden Markov model framework for large vocabulary speech recognition.

Dong Yu is a Senior Member of IEEE. He is currently serving as an Associate Editor of IEEE Transactions on Audio, Speech, and Language Processing (2011-) and has served as an Associate Editor of IEEE Signal Processing Magazine (2008–2011) and the Lead Guest Editor of IEEE Transactions on Audio, Speech, and Language Processing – special issue on deep learning for speech and language processing (2010–2011).

Li Deng received the Bachelor degree from the University of Science and Technology of China and the Ph.D. degree from the University of Wisconsin Madison. He joined Department of Electrical and Computer Engineering, University of Waterloo, Ontario, Canada in 1989 as an Assistant Professor, where he became a Full Professor with tenure in 1996. From 1992 to 1993, he conducted sabbatical research at Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, Mass, and from 1997 to 1998, at ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan. In 1999, he joined Microsoft Research, Redmond, WA as a Senior Researcher, where he is currently a Principal Researcher. Since 2000, he has also been an Affiliate Professor in the Department of Electrical Engineering at the University of Washington, Seattle, teaching the graduate course of Computer Speech Processing. His current (and past) research activities include automatic speech and speaker recognition, spoken language identification and understanding, speech-to-speech translation, machine translation, language modeling, statistical methods and machine learning, neural information processing, deep-structured learning, machine intelligence, audio and acoustic signal processing, statistical signal processing and digital communication, human speech production and perception, acoustic phonetics, auditory speech processing, auditory physiology and modeling, noise robust speech processing, speech synthesis and enhancement, multimedia signal processing, and multimodal human–computer interaction. In these areas, he has published over 300 refereed papers in leading journals and conferences, 3 books, 15 book chapters, and has given keynotes, tutorials, and lectures worldwide. He is elected by ISCA (International Speech Communication Association) as its Distinguished Lecturer 2010–2011. He has been granted over 40 US or international patents in acoustics/audio, speech/language technology, and other fields of signal processing. He received awards/honors bestowed by IEEE, ISCA, ASA, Microsoft, and other organizations. He is a Fellow of the Acoustical Society of America, and a Fellow of the IEEE. He serves on the Board of Governors of the IEEE Signal Processing Society (2008–2010), and as Editor-in-Chief for the IEEE Signal Processing Magazine (SPM, 2009–2011), which ranks consistently among the top journals with the highest citation impact.

Chin-Hui Lee is a Professor at School of Electrical and Computer Engineering, Georgia Institute of Technology. Lee received the B.S. degree in Electrical Engineering from National Taiwan University, Taipei, in 1973, the M.S. degree in Engineering and Applied Science from Yale University, New Haven, in 1977, and the Ph.D. degree in Electrical Engineering with a minor in Statistics from University of Washington, Seattle, in 1981.

Lee started his professional career at Verbex Corporation, Bedford, MA, and was involved in research on connected word recognition. In 1984, he became affiliated with Digital Sound Corporation, Santa Barbara, where he engaged in research and product development in speech coding, speech synthesis, speech recognition and signal processing for the development of the DSC-2000 Voice Server. Between 1986 and 2001, he was with Bell Laboratories, Murray Hill, New Jersey, where he became a Distinguished Member of Technical Staff and Director of the Dialogue Systems Research Department. His research interests include multimedia communication, multimedia signal and information processing, speech and speaker recognition, speech and language modeling, spoken dialogue processing, adaptive and discriminative learning, biometric authentication, and information retrieval. From August 2001 to August 2002 he was a visiting professor at School of Computing, The National University of Singapore. In September 2002, he joined the Faculty Georgia Institute of Technology.

Lee has participated actively in professional societies. He is a member of the IEEE Signal Processing Society (SPS), Communication Society, and the International Speech Communication Association (ISCA). In 1991–1995, he was an Associate Editor for the IEEE Transactions on Signal Processing and Transactions on Speech and Audio Processing. During the same period, he served as a member of the ARPA Spoken Language Coordination Committee. In 1995–1998 he was a member of the Speech Processing Technical Committee and later became the chairman from 1997 to 1998. In 1996, he helped to promote the SPS Multimedia Signal Processing Technical Committee in which he is a founding member.

Lee is a Fellow of the IEEE, and has published more than 350 papers and 25 patents on the subject of automatic speech and speaker recognition. He received the SPS Senior Award in 1994 and the SPS Best Paper Award in 1997 and 1999, respectively. In 1997, he was awarded the prestigious Bell Labs President's Gold Award for his contributions to the Lucent Speech Processing Solutions product. Lee often gives seminal lectures to a wide international audience. In 2000, he was named one of the six Distinguished Lecturers by the IEEE Signal Processing Society. He was also named one of the two ISCA's inaugural Distinguished Lecturers in 2007–2008. Recently he won the SPS's 2006 Technical Achievement Award for "Exceptional Contributions to the Field of Automatic Speech Recognition".