Hermitian Polynomial for Speaker Adaptation of Connectionist Speech Recognition Systems

Sabato Marco Siniscalchi, *Member IEEE*, Jinyu Li, *Member IEEE*, and Chin-Hui Lee, *Fellow IEEE*

***Abstract*—Model adaptation techniques are an efficient way to reduce the mismatch that typically occurs between the training and test condition of any automatic speech recognition (ASR) system. This work addresses the problem of increased degradation in performance when moving from speaker-dependent (SD) to speaker-independent (SI) conditions for connectionist (or hybrid) hidden Markov model/artificial neural network (HMM/ANN) systems in the context of large vocabulary continuous speech recognition (LVCSR). Adapting hybrid HMM/ANN systems on a small amount of adaptation data has been proven to be a difficult task, and has been a limiting factor in the widespread deployment of hybrid techniques in operational ASR systems. Addressing the crucial issue of speaker adaptation (SA) for hybrid HMM/ANN system can thereby have a great impact on the connectionist paradigm, which will play a major role in the design of next-generation LVCSR considering the great success reported by deep neural networks – ANNs with many hidden layers that adopts the pre-training technique – on many speech tasks. Current adaptation techniques for ANNs based on injecting an adaptable linear transformation network connected to either the input, or the output layer are not effective especially with a small amount of adaptation data, e.g., a single adaptation utterance. In this paper, a novel solution is proposed to overcome those limits and make it robust to scarce adaptation resources. The key idea is to adapt the hidden activation functions rather than the network weights. The adoption of Hermitian activation functions makes this possible. Experimental results on an LVCSR task demonstrate the effectiveness of the proposed approach.**

***Index Terms*—Artificial Neural Networks, Model Adaptation, Speech Processing.**

I. Introduction

Despite the tremendous advances in automatic speech recognition (ASR) technology [1], [2], it has been reported that system performance is often degraded when there is a mismatch between the training and testing environments. A degradation of the recognition accuracy is typically observed when moving from a speaker-dependent (SD) to a speaker-independent (SI) condition due to inter-speaker variability [3]. For a successful deployment of ASR applications, the discrepancies between the training and testing environments must be addressed. Different approaches have been developed to reduce the mismatch between training and testing environments. For example, robust and invariant speech features are proposed in [4]. Bayesian adaptation or compensation techniques aiming to modify the recognition parameters or speech features are proposed in [5]. New robust decision strategies are devised in [6]. This paper focuses on acoustic model adaptation algorithms that try to automatically "tune" the ASR system parameters to a new test environment using a limited, but representative, set of new data, commonly referred to as *adaptation data*. In particular, this work addresses the batch, supervised speaker adaptation problem of hybrid hidden Markov model/artificial neural network (HMM/ANN) systems, where the ANN is implemented using a multi-layer perceptron (MLP) architecture. Addressing the crucial problem of speaker adaptation for ANNs can have a great impact on the connectionist ASR paradigm, which may play a major role in the design of next-generation LVCSR considering the great success of deep neural networks (e.g., [7], [8]).

In the hybrid HMM/ANN framework, the most effective speaker adaptation techniques consists in learning the parameters of a linear transformation network connected to either the input [9], [10] or the output layer [11]. The weights of this additional linear layer are estimated during the adaptation phase while all of the other weights are held constant. This approach has several disadvantages that will be described in Section II-B. In contrast, the proposed solution in this paper is to adapt the shape of the hidden activation functions [12] in the HMM/ANN system. A weighted sum of R orthonormal Hermite functions, which has already proven successful in speech classification and recognition tasks [12], is used as a non-linearity in the hidden nodes in order to obtain this adaptation capability. In the training phase, the proposed hidden activation function is automatically learned from the training data, but its shape is modified during the adaptation phase while all other neural parameters (i.e., weights) are kept frozen. Experimental results on the Wall Street Journal (WSJ) Nov92 task [13] not only demonstrate the effectiveness of our adaptation approach for LVCSR tasks but also show that the proposed solution outperforms conventional adaptation techniques for hybrid ASR systems as the amount of adaptation data decreases. The latter outcome implies that adapting the activation function to a target speaker confers robustness to scarcity of adaptation data. It is also observed that adaptation of the bias and slope of the sigmoidal activation function is not effective. This paper reorganizes, expands, and completes the study reported in [14]. In particular, a deeper performance comparison with MLPs adopting sigmoid activation functions has been carried out in the present paper. Furthermore, an analysis of the influence of

S. M. Siniscalchi is with the Department of Computer Engineering, Kore University of Enna, Enna, Italy, and with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332.E-mail: marco.siniscalchi@unikore.it

J. Li is with Microsoft Corporation, Rendmond, WA, USA.E-mail: jinyli@microsoft.com

C.-H. Lee is with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, 30332.E-mail: chl@ece.gatech.edu

the amount of adaptation data has been investigated. The effect of the adaptation phase on the hidden activation function based on orthonormal Hermitian polynomial has also been reported showing that hidden activation functions assume different shapes after adaptation according to the specific target speaker.

The rest of the paper is organized as follows: related works are discussed in Section II. The Hermitian-based neural architecture is presented Section III. In Section IV the experimental environment is given, and the results are discussed. Concluding remarks are given in Section V.

II. Related Work

Model adaptation techniques can be divided into supervised adaptation and unsupervised adaptation. In supervised adaptation, each adaptation utterance is associated with a transcription. In unsupervised adaptation, it is not associated with a transcription. Furthermore, speaker adaptation can also be carried out off-line (batch) or on-line [15]. Batch adaptation is done after all the available adaptation utterances are collected whereas on-line adaptation is done each time one utterance is obtained. In the following two sections, a brief overview of existing acoustic model adaptation techniques is presented. First, top l adaptation approaches for conventional HMM/GMM systems are discussed in order to place our idea in the context of current ASR systems, although the proposed technique is tailored to hybrid HMM/ANN. Next, the underpinning of the principal speaker adaptation techniques for hybrid HMM/ANN, which is the focus of the paper, is presented, and the shortcomings of these techniques are highlighted.

A. Acoustic Model Adaptation of HMM/GMM Models

There exist two major adaptation approaches for HMM/GMM models, the transformation-based approach and the Bayesian approach. The best known example of transformation-based adaptation is the maximum likelihood linear regression (MLLR) framework [16], [17], in which an affine transformation is used to transform the mean and variance vectors of the Gaussian mixture densities in the original set of HMMs. The feature-space MLLR (fMLLR) [18], [19] has proven to be highly effective as a method for feature space adaptation. In Bayesian learning (e.g., [20]), prior densities are assumed, and MAP estimates are obtained for the HMM parameters. When the adaptation data size is limited, structural maximum a posteriori (SMAP) adaptation [21] improves the efficiency of MAP estimation. Bayesian estimation can also be applied to transformation parameters, e.g., MAPLR and joint MAP estimation of transformation and HMM parameters [22]. Correlated HMMs, in which HMM parameters are no longer assumed independent, have been shown to be advantageous over conventional HMMs, and when combined with online adaptation, they have also been shown to be both efficient and effective [23]. A hierarchical structure is leveraged by the proximity information of the mixture Gaussian densities, and it is therefore more efficient and effective than MAP in speaker adaptation, especially when the adaptation data set is very limited. For example, even with only one adaptation utterance in a resource management test, the word accuracy went from below 50% for MAP to over 70% for SMAP [21] Discriminative acoustic model adaptation methods have also been proposed over the years, e.g., [24], [25], [26]. These techniques alter the acoustic model such that a discriminative criterion is optimized.

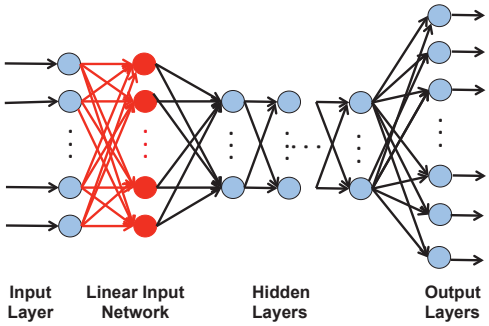

Fig. 1. *Basic neural architecture for adaptation of HMM/ANN models based on LIN. The red links are clamped to 1 during the training phase. In adaptation mode, the parameters (weights) associated to the red links are estimated using the adaptation utterances while all other weights are kept fixed. The activation function of each LIN neuron (red node) is a linear function.*

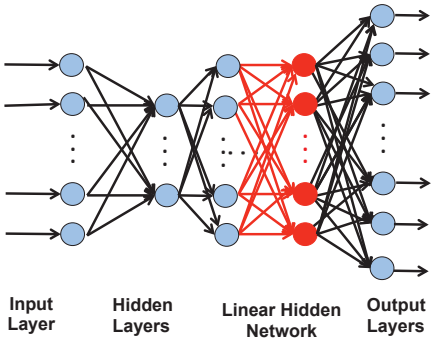

Fig. 2. *Basic neural architecture for adaptation of HMM/ANN models based on LHN. The red links are clamped to 1 during the training phase. In adaptation mode, the parameters (weights) associated to the red links are estimated using the adaptation utterances while all other weights are kept fixed. The activation function of each LHN neuron (red node) is a linear function.*

B. Acoustic Model Adaptation of hybrid HMM/ANN Models

Adaptation techniques based on a linear transformation network added to either to the input or the output layer of the ANN have proven beneficial for speaker adaptation and represent the most successful adaptation solution in the connectionist paradigm. In Figure 1, a linear input network (LIN) is added to the input layer of the SI MLP to map SD input vectors to the SI ASR system [9], [10]. The linear transformation, which is trained to minimize the error at the output of the neural architecture while keeping all other MLP weights frozen, rotates the input space to reduce the discrepancy between target and training conditions. In [11], the authors propose to add a linear transformation network before the output layer, referred to as a linear hidden layer (LHN) (see Figure 2). The rationale behind LHN is that the added linear layer generates discriminative features of the input pattern suitable for the classification performed at the

output of the MLP. In these linear network techniques, the number of adaptation parameters cannot be set according to the amount of the available adaptation data, since it is bound to the number of inputs, or outputs of the SI MLP. Furthermore, these linear transformation network approaches do not perform well when only a few adaptation utterances are available [11], in which case the SA ASR performance drops below the SI ASR one. These techniques are also prone to severe performance degradation when the adaptation data do not contain examples for a subset of the output classes [11].

Other techniques for adapting connectionist ASR systems have been proposed in the past years. For instance, *regularized adaptation of discriminative classifiers* described in [27] shows that adaptation can be successfully accomplished using a small amount of speaker-specific material by an L2 regularization that penalizes large deviations from the original, speaker-independent weights. The parallel hidden network approach in [9] is another example of adaptation techniques that uses two independent parallel hidden layers: one hidden layer is tuned on SI data whereas the weights of the other hidden layer are trained using adaptation data while keeping all other parameters fixed. In the GAMMA approach, a gamma filter is used to map the speaker-dependent input vectors into the SI system. Nonetheless, to the best of the authors' knowledge, these techniques have never been applied to LVCSR systems. The interested reader is referred to [9] for a complete review of these techniques.

III. Connectionist ASR Systems with Hermitian-based ANNs

It is instructive to briefly review the statistical pattern matching approach to ASR and highlight the key components of a conventional ASR system before delving into the details of the proposed acoustic model adaptation technique.

A. Speech Decoding in a Nutshell

The goal of an ASR system is to "recognize" the word sequence, W, given a sequence $\mathbf{X} = (x_1, \ldots, x_T)$ of T feature vectors, or frames, extracted from a speech signal; that is, x_t is the acoustic observation at time t. This can also be considered as a *decision problem*; i.e., based on the information in $\mathbf{X}$ and the other relevant aspects of the problem, we attempt to make the best inference, in some sense, about W that is embedded in $\mathbf{X}$. This problem, under the statistical approach to ASR, is solved by decomposing the joint distribution, $p(W, \mathbf{X})$, into two components, $p(\mathbf{X}|W)$ and $P(W)$, known as an acoustic model (AM) and a language model (LM), respectively. The forms of $p(\mathbf{X}|W)$ and $P(W)$ are assumed *parametric* probability density functions (PDFs), i.e., $p_\Lambda(\mathbf{X}|W)$ and $P_\Gamma(W)$, respectively. The parameters, Λ and Γ, are estimated from some *training data*. The decoded word sequence $\hat{W}$ is then determined using the well-known *plug-in MAP* (maximum *a posteriori*) *decision rule* (e.g., [28], [29], [30]),

$$\hat{W} = \arg\max_{W \in \Omega} P(W|\mathbf{X}) = \arg\max_{W \in \Omega} p_{\hat{\Lambda}}(\mathbf{X}|W) \cdot P^{\alpha_L}_{\hat{\Gamma}}(W) \quad (1)$$

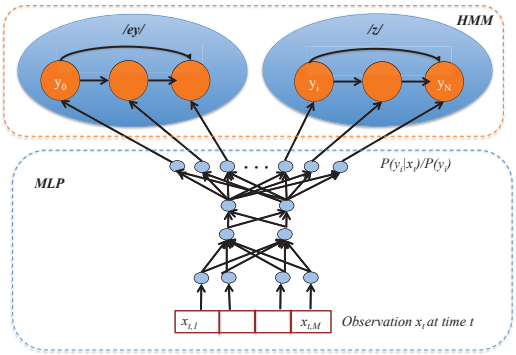

Fig. 3. *A block diagram of a generic hybrid HMM/ANN system. The HMM models the sequential property of the speech signal, and the MLP models the scaled observation likelihood of all the phone-state labels. The same MLP is replicated over different points in time.*

where $\hat{\Lambda}$ and $\hat{\Gamma}$ are the estimated parameters obtained during training, $\hat{W}$ is the recognized sentence from decoding, and Ω is the set of valid candidate word sequences to be searched during testing. This decision rule, derived from the optimal Bayes decision rule, is also widely used in many other pattern recognition applications. In the above equation, α_L, commonly known as a language model multiplier, is used to balance the AM and LM contributions to the overall probability due to unknown distributions and the use of a likelihood function $p_{\hat{\Lambda}}(\mathbf{X}|W)$ to compute the acoustic probability.

The acoustic model is implemented as a hidden Markov process that governs the transitions between states $Y = (y_1; \ldots; y_K)$. An HMM is completely specified given the initial state probability distribution $\pi = p(q_0 = y_i)$, where q_t is the state at time t, the transition probabilities $a_{ij} = p(q_t = y_j | q_{t-1} = y_i)$, and a model to estimate the observation probabilities $p(x_t|y_i)$. Thus, the acoustic probability can be computed as

$$\begin{aligned} p_{\hat{\Lambda}}(\mathbf{X}|W) &= \sum_q (p(\mathbf{X}, q|W) p(q|W) \\ &\approx \arg\max_\pi (q_0) \prod_{t=1}^{T} a^{(k)}_{q_{t-1}, q_t} \prod_{t=0}^{T} p(x_t|q_t) \end{aligned} \quad (2)$$

In conventional HMMs used for ASR, $p(x_t|q_t)$ is directly modeled using GMMs [31], [32]. However, this work is concerned with hybrid HMM/ANN models, where the MLP directly estimates the *a posterior* probability, $p(q_t|x_t)$, of the q_t state, given the speech observation x_t, as shown in Figure 3. Bayes' rule is used to compute the observation probability:

$$p(x_t|q_t) = \frac{p(q_t|x_t) p(x_t)}{p(q_t)}, \quad (3)$$

where $p(q_t)$ is the prior probability of each state estimated from the training set, and $p(x_t)$ is independent of the word sequence and thus can be ignored.

B. Hermitian-Based MLP for Speech Recognition

Feed-forward single-hidden-layer MLPs are employed in the present work. The MLP is designed for estimating class

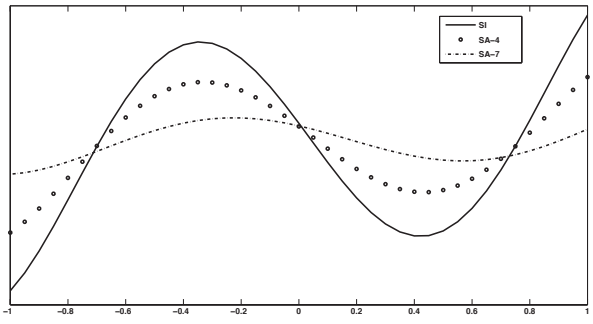

Fig. 4. *Hermitian hidden activation functions before and after adaptation for a randomly selected hidden neuron. The value of R is 10 for all example in the figure. It is interesting to notice that the shape of the activation function changes moving from speaker independent (solid curve) to speaker adapted (dashed, and dot curves) system. In the figure, SI stands for speaker independent, whereas SA is for speaker adapted. The speakers numbered 4 and 7 are considered in the figure, and the activation functions related to them are referred to SA-4, and SA-7, respectively. The activation function assumes a different shapes for different speakers, as shown by the dot and dashed curves.*

posterior probabilities in a discriminative way. The MLP estimates the conditional probability of a class label y given an input vector $\mathbf{x}$ using a nonlinear model of the form

$$\hat{p}_k = \hat{p}(y = k|\mathbf{x}) = \frac{\exp g_k}{\sum_{i=1}^{N} \exp g_i}, \tag{4}$$

where g_k is the linear activation function of the kth output, and it is given by

$$g_k = \sum_{j=1}^{L} w_{kj}^{(2)} f_j \left(\sum_{i=1}^{M} w_{ji}^{(1)} x_i \right). \tag{5}$$

Here $w_{kj}^{(2)}$ and $w_{ji}^{(1)}$ denote weights in the second and first layer, respectively; f_j is the activation function of jth hidden neuron. L and M are the number of nodes in the hidden and input layer, respectively. The ith component of the input vector $\mathbf{x}$ is indicated with x_i. There exist several types of activation functions that can be used in hidden neuron. If the orthonormal Hermite polynomials are chosen, f_j is a linear combination of Hermite functions of the form

$$f_j(z) = \sum_{r=0}^{R} c_{jr} h_r(z), \tag{6}$$

where R is the degree of the Hermite polynomial, and $h_r(z)$ is the rth Hermite orthonormal function. The c_{jr} coefficients are learned during the training phase along with the $w_{kj}^{(2)}$ and $w_{ji}^{(1)}$ weights. The orthonormal Hermite polynomials will be described later along with their first-order derivatives. Figure 4 shows the Hermitian-based activation function for one of the hidden neurons (solid curve).

In this work, all of the hidden neurons employ a Hermite polynomial of the same degree, and the softmax activation function is employed at the output layer. Hermitian-based MLP have already proposed in the past, e.g. [33], [34], but several differences exist between the proposed implementation and other similar architectures. For example, in the constructive MLP system [34], the neural architecture grows as part of the training phase by adding a new hidden neuron till convergence is reached. Furthermore, the order of the Hermitian-based activation function increases by one each time a new hidden unit is added to the network. We believe that this architecture is not suitable for speech applications where the number of hidden neurons is on the order of hundreds or thousands, and it did not seem reasonable to use Hermite polynomials with such a high degree because it may lead to an unstable training phase. In our implementation, the number of hidden neurons is fixed at the beginning of the training phase. All hidden neurons employ a Hermite polynomial of the same degree. This results in a configuration similar to [33], but [33] uses linear activation functions instead of softmax activation functions at the output layer. Furthere, [33] uses the sum-of-squares as the error function to be minimized whereas the cross-entropy error function is chosen as the function to be minimized during the training phase in this study. Furthermore, the work presented in [33] is not related to ASR.

1) Hermite Regression Formula: The r-th orthogonal Hermite polynomial, $H_r(z)$, is defined over the interval $(-\infty, \infty)$. The orthonormal Hermite polynomial of order r can be then expressed in terms of $H_r(z)$:

$$h_r(z) = \alpha_r H_r(z) \phi(z), \tag{7}$$

where

$$\alpha_r = (r!)^{-\frac{1}{2}} \pi^{\frac{1}{4}} 2^{-\frac{r-1}{2}}, \tag{8}$$

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \tag{9}$$

$$\begin{aligned} H_r(z) &= (-1)^r e^{z^2} \frac{\partial^r}{\partial z^r} \left(e^{-z^2} \right) \\ &= 2z H_{r-1}(z) - 2(r-1) H_{r-2}(z), \\ &\quad r > 1, \quad H_0(z) = 1, \quad H_1(z) = 2z. \end{aligned} \tag{10}$$

The derivation of the first-order derivative of Eq. 6 is very simple due to the recursive nature of the orthonormal Hermite polynomials. As a consequence, the proposed Hermitian-based activation function can be easily plugged into the learning procedure based on gradient descent. The first-order derivative is

$$\begin{aligned} \frac{\partial}{\partial z} f(z) &= \sum_{r=1}^{R} c_{jr} \frac{\partial}{\partial z} \left(h_r(z) \right) \\ &= \sum_{r=0}^{R} c_{jr} \left[(2r)^{\frac{1}{2}} h_{(r-1)}(z) - z h_r(z) \right]. \end{aligned} \tag{11}$$

2) Training Phase: The training protocol is displayed in Algorithm 1, which is the classical stochastic back-propagation algorithm used to train the neural networks [35] with the difference that the c_{jr} are also adapted. The cross-entropy error criterion, J_{ce}, which measures a "distance" between probability distributions, is adopted as criterion function. For N training samples and C output classes, the cross-entropy error criterion is of the form:

$$J_{ce}(w_{ji}^{(1)}; w_{kj}^{(2)}; c_{jr}) = \sum_{m=1}^{N} \sum_{l=1}^{C} t_{ml} \log(t_{ml}/y_{ml}). \qquad (12)$$

Algorithm 1 Network Learning *Pseudocode*

Input:

N, $\underline{\theta}$, $\underline{\beta}$, $\underline{\eta}$ ◁ *values defined by the user*
$\underline{w}_{ji}^{(1)}$, $\underline{w}_{kj}^{(2)}$, $\underline{c}_{jr}$ ◁ *randomly chosen values*
$\underline{c}_{j1}, \alpha$ ◁ *randomly chosen values between 0 and 1*
$r \triangleleft 1$, $temp \triangleleft \underline{c}_{j1}$
repeat $r \leftarrow r + 1$
$\underline{c}_{jr} \triangleleft \alpha \times temp$
until $(r == R)$

Output:

$\hat{w}_{ji}^{(1)}$, $\hat{w}_{kj}^{(2)}$, $\hat{c}_{jr}$ ◁ *learned parameters*

Begin

Initialize $m \leftarrow 0$, $\theta \leftarrow \underline{\theta}$, $w_{ji}^{(1)} \leftarrow \underline{w}_{ji}^{(1)}$,
$w_{kj}^{(2)} \leftarrow \underline{w}_{kj}^{(2)}$, $c_{jr} \leftarrow \underline{c}_{jr}$
$n_{\mathcal{T}} = n_{\mathcal{V}} = n_{\mathcal{T}}^{*} = n_{\mathcal{V}}^{*} \leftarrow 0$
repeat $m \leftarrow m + 1$
$n_{\mathcal{T}} \leftarrow$ fraction of misclassified training samples
$n_{\mathcal{V}} \leftarrow$ fraction of misclassified validation samples
foreach $x \in \mathcal{T}$ **do**
feed-forward(x, $w_{ji}^{(1)}$,$w_{kj}^{(2)}$, c_{jr}) ◁ *The input x is fed into the network.*
$w_{ji}^{(1)} \leftarrow w_{ji}^{(1)} + \eta \delta_j x_i$
$w_{kj}^{(2)} \leftarrow w_{kj}^{(2)} + \eta \delta_k y_j$
$c_{jr} \leftarrow c_{jr} + \eta \delta_{jr} h_r(z)$
end foreach
$n_{\mathcal{V}}^{*} \leftarrow$ fraction of misclassified validation samples
$n_{\mathcal{T}}^{*} \leftarrow$ fraction of misclassified training samples
if $(n_{\mathcal{V}}^{*} - n_{\mathcal{V}}) < \underline{\beta}$ **then**
$\eta \leftarrow \eta \times 0.7$
else
$\hat{w}_{ji}^{(1)} \leftarrow w_{ji}^{(1)}$
$\hat{w}_{kj}^{(2)} \leftarrow w_{kj}^{(2)}$
$\hat{c}_{jr} \leftarrow c_{jr}$
end if
until $(n_{\mathcal{T}}^{*} - n_{\mathcal{T}} > \theta) \wedge (m \leq N)$

End

In Algorithm 1, $\mathcal{T}$ and $\mathcal{V}$ represent the training and validation data, respectively. If an *epoch* corresponds to a single presentation of all input patterns in the training set, N is the maximum number of epochs that can be performed. The number of misclassified training and validation input vectors are indicated with $n_{\mathcal{T}}$, and $n_{\mathcal{V}}$, respectively. The number of misclassified training and validation input vectors at the mth epoch are indicated with $n_{\mathcal{T}}^{*}$, and $n_{\mathcal{V}}^{*}$, respectively. The learning rate is indicated with η, which is reduced as the number of misclassified validation input pattern increases of a fixed β amount. Finally, δ_j, and δ_k represent the *sensitivity* [28] for a hidden, a output unit , respectively. The sensitivity with respect to the generic c_{jr} coefficient is indicated with δ_{jr}. The training phase ends when one of the two following conditions is verified: (1) the change in the number of misclassified training input vectors is higher than a preset value θ, or (2) the number of epochs is greater than a preset value N.

3) Adaptation Phase: During this phase the shape of the Hermitian-based hidden non-linearity is modified to better suit the speaker-specific features. The coefficients in Eq. 6 are the only parameters adapted, and all other parameters of the MLPs are held constant. The stochastic back-propagation algorithm with cross-entropy error function (J) is used to adapt the coefficients of Eq. 6. With respect to the generic c_{jr} coefficient, the change of the overall error is computed as follows

$$\frac{\partial J}{\partial c_{jr}} = \frac{\partial J}{\partial f_j} \frac{\partial f_j}{\partial c_{jr}} = \left[\sum_{k=1}^{N} w_{kj}^{(2)} \delta_k \right] h_r(z). \qquad (13)$$

where δ_k is the *sensitivity* of the kth output unit, as aforementioned. In Figure 4, the solid curve represents the activation function of a generic neuron after SI training whereas the dashed and dot curves represents the shape of the activation function for the same hidden neuron after adaptation for two different speakers, namely speaker 444 and speaker 447 from the Nov92 data available with the WSJ corpus [13]. It is interesting to note that the activation function assumes a different shape for different speakers. The structure of the proposed connectionist system is not modified in contrast to the above mentioned linear transformation approaches. Hence the number of adaptable parameters is not equal to either the input or output dimension of the neural architecture as is done for the LIN and LHN methods. Thus, the proposed system may be modified based on the amount of available adaptation data. The latter experiment is not carried out in the present work.

IV. Experimental Setup and Results

The ultimate goal of ASR is continuous speech recognition with a large vocabulary; it is therefore fundamental to compare and contrast the proposed adaptation approach with the best current adaptation techniques for hybrid HMM/ANN systems on a LVCSR task.

A. Experimental Setup

1) Corpus: The proposed approach is evaluated on a LVCSR task using the 5,000-word WSJ0 (5k-WSJ0) corpus. The SI84 data (7077 utterances, or 15.3 hours of speech from 84 speakers) are used during the training phase. The training material is separated into a 6877-sentence training set and a 200-sentence validation set. The testing phase uses the Nov92 evaluation data, which contains 330 utterances from 8 speakers. The *si_et_ad* set is used during the adaptation phase, and this set consists of 8 speakers with 40 utterances per speaker. The number of context-independent phonemes is 40.

2) Speech Parametrization: Split temporal context (STC) features [36], which makes use of long temporal context, are used. Mel filter bank energies are first computed in the conventional way. Then temporal evolutions of critical band spectral densities are taken around each frame using a context of 31 frames (310 ms) around the current frame is chosen. This context is split into 2 halves: left and right Contexts. The discrete cosine transform was applied two these two context to de-correlate and reduce dimensionality (only 11 coefficients are retained). Two single-hidden-layer MLPs are trained to produce phoneme-state posterior probabilities for both context parts. These two MLPs are referred to as lower networks. A third single-hidden-layer MLP merges the output of the two lower networks and produces the final set of phoneme-state posterior probabilities, which are used in the hybrid HMM/ANN system displayed in Figure 3. The three MLPs estimate phoneme-state posterior probabilities [36], [37]. Each phoneme was split into three states. In summary, a hierarchical structure of three MLPs computes $p(q_t|x_t)$, which are used in Eq. (3).

3) Neural Architecture: Two hybrid HMM/ANN systems using the STC features are compared, and they differ only in the type of hidden activation function. The Hermitian-based activation function is used in the MLP (HMLP) of the first system. The sigmoidal activation function is employed in the MLP (SMLP) of the second system. These LVCSR systems are referred to as HMM/HMLP and HMM/SMLP, respectively. All MLPs are trained using Algorithm 1, and word recognition employes a trigram language model. The number of hidden nodes is chosen on the performance on the validation sentences. The input dimension of the lower networks is equal to 253 after dimensionality reduction, and the dimension of the output layer is 120. The input dimension of the merger is 240, and the number of output classes is 120.

The number of hidden nodes and number of c_{rj} coefficients of the HMLP architecture could be selected by observing the effect of these values on the performance of the validation data. However, such an experiment would require several trials in which both these numbers have to be varied independently. Hence we decided to first tune the number of hidden nodes using the SMLP architecture on the validation data. Then, the number of c_{rj} coefficients for the HMLP architecture is found using the number of hidden nodes obtained for the SMLP architecture. Figure 5 shows that the SMLP architecture attains the best phoneme-state accuracy with a number of hidden nodes equal to 1500. A drop in the frame accuracy is observed when using more than 1500 hidden nodes. In order to have a reasonable trade-off between frame classification accuracy and speed of the training phase, the number of hidden neurons of all SMLPs in the hidden layer is set to 800. The same number of hidden nodes is used for the HMLP architecture. Figure 6 shows the phoneme-state classification accuracy curve in terms of the number of c_{jr} coefficients. This curve reaches a plateau at 10 coefficients and does not increase any further; therefore, the number of c_{jr} coefficients is set to 10.

Finally, the performance of the LOQUENDO connectionist HMM system trained on the SI-84 and evaluated on the same Nov92 data set reported in [11] is also included for comparison and to further assess the quality of the hybrid systems.

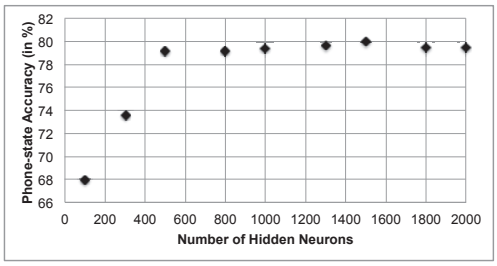

Fig. 5. *Phoneme-state classification accuracy in terms of number of hidden nodes.*

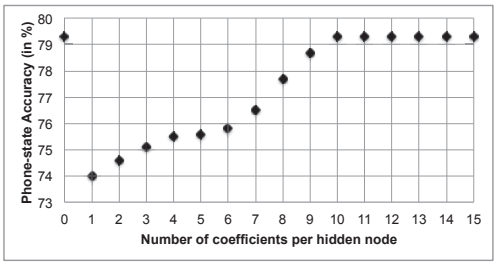

Fig. 6. *Phoneme-state classification accuracy on the validation data in terms of the number of c_{jr} coefficients. For these experiments, the number of hidden nodes was set to 800. The point for zero c_{jr} coefficients the performance of the SMLP network.*

B. Results

1) Phoneme Classification Results: In Table I, the frame accuracy rates (FARs) at the phoneme level, which are more meaningful than accuracies at a phoneme-state level, for the Hermitian-based (HMLP) and Sigmoidal-based (SMLP) STC architecture on the Nov92 evaluation set are given. Phoneme accuracies are obtained by folding back the three phoneme-state classes into a single phoneme class. The HMLP correctly classifies 85.0% of the test samples, as shown in the second row of Table I. The standard SMLP architecture correctly classifies 85.4% input patterns (see first row in Table I). The difference in the final frame accuracy between the SMLP and the HMLP may be due to the initialization of the c_{jr} coefficients, but we did not investigate on this point further since our final goal is the adaptation phase.

The performance of several connectionist ASR systems are presented and discussed in the following sections.

2) LVCSR Results: The LVCSR performance, in terms of word error rate (WER), for all hybrid HMM/ANN systems studied in this work is reported in Table II. The WER is computed as $(1-(H-I)/W)$, where W is the number of reference words and I, and H are the number of inserted and correctly recognized words, respectively. The second column of Table II shows the SI performance, the third column shows the performance of the adapted LVCSR systems, and the fourth column summarizes the relative improvement. LIN adaptation was applied to the HMM/SMLP system to obtain the SA result shown in the second row of Table II. The LIN adaptation layer was initialized with the identity matrix. The SA LOQUENDO perfomance is given as reported in [11] in the case of LIN adaptation. In [11], the authors also provide results for the

TABLE I
FRAME ACCURACY RATES OF THE HIERARCHICAL STRUCTURE OF MLPS ON THE NOV92 DATA.

System	Accuracy (in %)
SMLP	85.4
HMLP	85.0

TABLE II
WER ON THE NOV92 TASK FOR SEVERAL CONNECTIONIST LVCSR SYSTEMS. A TRIGRAM LANGUAGE MODEL IS USED. LIN IS USED TO PERFORM ADAPTATION FOR LOQUENDO AND HMM/SMLP. IN THE PROPOSED APPROACH ADAPTATION IS ACCOMPLISHED USING EQ. (13), AND ONLY THE SHAPE OF THE HIDDEN ACTIVATION FUNCTIONS IS ADAPTED.

System	SI	SA	Rel. Imp.
HMM/HMLP	6.3 %	4.9%	22.2%
HMM/SMLP + LIN	6.3 %	5.2%	17.4%
LOQUENDO HMM/SMLP + LIN	6.5 %	5.6%	13.8%
HMM/SMLP (bias and slope)	6.3 %	6.3%	–

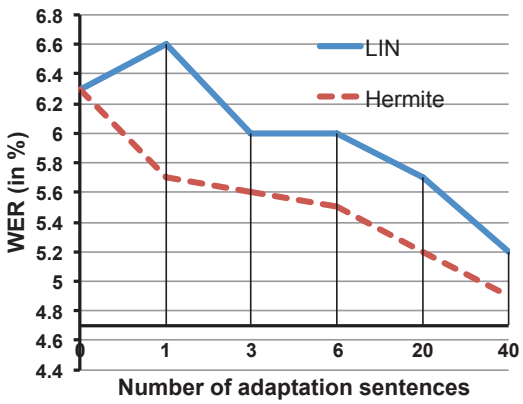

Fig. 7. *Influence of the amount of adaptation data on the WER of the SA ASR system. The dashed line refers to the Hermite-based ASR system, whereas the solid line is for the LIN system. The WER with zero adapted sentences represents the SI system performance.*

LHN technique and for the combination of LIN and LHN, but the conservative training (CT) technique is also applied. Conservative training seems to mitigate the issues that arise when a neural network is adapted with new data that do not adequately represent the knowledge included in the original training data. Conservative training [11] addresses this issue by not setting the value of the targets of the missing units in the adaptation data to zero. Another possibility to reduce over-fitting is to learn a shared affine transformation that is applied to each frame to the splicing that forms the MLP input, as reported in [10]. This will reduce the number of adaptation parameters to learn and can be also implemented for the Hermitian-based MLP. Nonetheless, applying this technique is out of the scope of this work, and the interested reader is referred to [10].

In our own experiments, we observed that LHN adaptation scheme without CT delivered a final performance lower than that obtainable with the LIN approach. Therefore, we decided to report LIN adaptation results only when CT is not adopted. For the sake of completeness, it should be pointed out that standard HMM/GMM approach can attain a WER as low as 4% when discriminative training techniques are adopted [38], [39]. Nevertheless, the proposed work is concerned with hybrid HMM/ANN where the ANN is a single hidden layer MLP.

The SI performance of the proposed HMM/HMLP system is given in the second row of Table II, and it is equal to a WER of 6.3%. This result is slightly better than that previously reported in [14], because a more conservative value is used for η in Algorithm 1. Furthermore, this recognition result equals that of the hybrid HMM/SMLP, as shown in the second row of Table II. The LOQUENDO connectionist system attains a WER of 6.5% before adaptation, and this results is reported in the third row of Table II. This first set of results demonstrates that a reasonable SI LVCSR system can be designed using HMLPs. Furthermore, the non-monotonic shapes of the Hermitian-based activation function does not harm recognition results.

LIN adaptation is performed on LOQUENDO yields a WER of 5.6%, which corresponds to a relative improvement of 13.8%. The WER is reduced to 5.2% from the initial 6.3% by using LIN over the hybrid HMM/SMLP system, and this reduction corresponds to a relative improvement of 17.5%. The hybrid SA HMM/HMLP LVCSR system attains a WER of 4.9% using the adaptation technique described in Section III-B1. This result corresponds to a relative improvement over the SI HMM/HMLP system of 22.2% and demonstrates the viability of the proposed technique. As shown in Table II, the proposed SA hybrid HMM/HMLP system is slightly superior to the SA LOQUENDO system and competitive with the SA HMM/SMLP system. In addition, the number of adaptation parameters involved in the proposed procedure is only 24,000 (i.e., 800 x 10 parameters for each of the three HLMPs), which is much lower than that involved in the LIN procedures, namely 186,364 for the hybrid HMM/SMLP system, and even more parameters for the LOQUENDO system. Furthermore, a minor gain in adopting our technique is that the SI structure of the HMLP need not to be modified to perform adaptation. In all adaptation experiments, the number of learning ephocs was set equal to 2 based on our experience with other task.

It is instructive to compare and contrast the well-know sigmoidal activation function against the proposed activation function in the speaker adaptation setting. To this end, the bias and slope hidden sigmoidal activation functions of the HMM/SMLP have been adapted. The last row of Table II shows that no improvement is gained by adapting slope and bias of the sigmoidal function, and that demonstrates that an effective adaptation scheme within the hybrid HMM/ANN framework can be established only by selecting a proper adaptation non-linearity, such as the Hermitian-based non-linearity proposed in this paper.

3) LVCSR Results with a Varying Amount of Adaptation Data: A crucial aspect of speaker adaptation is the amount of available adaptation data. In the series of experiments just discussed, the entire set of adapting sentences, i.e. 40 sentences, is used, yet it is often the case that the amount of adaptation data is much smaller. For example, only one or two sentences can be collected in the case of on-line adaptation. Thus, it is of fundamental importance to investigate the influence of the amount of adaptation data on the

TABLE III
ADAPTED SIGMOIDAL- AND HERMITIAN-BASED CONNECTIONIST LVCSR SYSTEMS WITH AND W/O CONSERVATIVE TRAINING USING SINGLE ADAPTATION UTTERANCE.

System	SA 1-sentence)	SA (1-sentence) + CT
HMM/HMLP	5.7 %	5.6%
HMM/SMLP	6.6 %	6.4%

SA system performance. Unfortunately, such an experimental setup is not available for the LOQUENDO system; therefore, the following results consider only the hybrid HMM/SMLP and HMM/HMLP systems.

Figure 7 compares the WERs on the Nov92 data for various amount of adaptation data for the Hermitian-based (dashed line), and LIN adaptation (dashed line) approaches. For large amounts of adaptation data, both the proposed adaptation technique and LIN adaptation of the MLP parameters perform quite well. For little amounts of adaptation data, the proposed technique leads to some improvement over LIN with a WER of 4.6% for 20 adaptation utterances against 5.4% for LIN, as an example. The robustness of the proposed approach becomes more evident when only a single adaptation sentence is available. Indeed, an increment in the WER over the SI hybrid HMM/SMLP system is observed when LIN is applied using only a single sentence; that is, LIN leads to a drop in performance moving from SI to SA conditions. In contrast, the proposed adaptation technique does not harm the SI performance when a single adaptation sentence is available. Finally, a visual inspection of the two adaptation curves shown in Figure 7 demonstrates that our approach not only involves less parameters, but it also outperforms the standard LIN approach as the amount of adaptation decreases.

4) ANN Adaptation with Conservative Training: As aforementioned, conservative training (CT) [11] mitigates the lack of adequately represent the knowledge included in the original training data. Table III compares the WERs of the LIN and Hermitian-based approaches on the Nov92 data when a single adaptation sentence is used along with the CT technique. A positive effect is observed on both hybrid ASR systems when adaptation is performed along with CT, and the WER drops from 5.7% down to 5.6% and from 6.6% to 6.4% for HMM/HMLP and HMM/SMLP systems respectively.

V. CONCLUSION & FUTURE WORK

The choice of hidden non-linearity in a feed-forward multi-layer perceptron (MLP) architecture is crucial to obtain good generalization capability and better performance [34]. Yet, little attention has been paid to this aspect in the ASR field. In [12], some initial studies adopting hidden activation functions based on orthonormal Hermite polynomials, which can change shape during training confirmed that using a non-monotonic activation function has beneficial effects on speech recognition. Indeed, a performance improvement is observed in continuous speech recognition with both matched and mismatched corpus conditions (see [12]). In this paper, that line of research have been extended, and the Hermitian MLPs has been evaluated in the context of speaker adaptation for LVCSR. It has been shown that the connectionist architecture of the hybrid HMM/ANN speech recognition system can be successfully adapted to a specific speaker while keeping the complexity and the structure of the SA and SI LVCSR systems equivalent with beneficial effects on the overall ASR performance. Furthermore, the proposed approach compares favorably with the standard LIN technique on the same task. Experimental evidence has also demonstrated that our approach is more robust to data scarcity than the conventional LIN approach.

Recently, it has been demonstrated that adaptation of hybrid ASR system can be further boosted by using constrained MLLR (CMLLR) [18]. CMLLR transforms can be directly applied to the observed acoustic features, which can be easily incorporated into the adaptation HMM/ANN framework as the input transforms, e.g., [40]. Likewise, maximum likelihood vocal tract length normalization (VTLN) [41] implements a per speaker frequency scaling of the speech spectrum that can be directly applied to the speech features. We will explore CMLLR and/or VTLN in future work.

It should be remarked that ANNs are indeed enjoying a resurgence of interest among the speech researchers (e.g., [7], [42], [14], [43]) due to the great success of deep neural networks – MLPs with many hidden layers that adopts the pre-training approach proposed for generative deep belief nets [44] – in many speech applications (e.g., [8]). Although there is a concern of the deep neural network runtime efficiency compared to GMMs since a deep neural network has much more parameters to evaluate, researchers at Microsoft have recently conducted studies on how to reduce this runtime cost by restructuring the deep neural network with less parameters [45]. In the meanwhile, Google researchers have tried to predict multiple frame outputs with single frame input [46]. All these efforts allow efficient deployment of deep neural networks in commercial systems. Therefore, finding a good adaptation technique in this context can have a great impact on the hybrid HMM/ANN paradigm and thereby trigger a new trend in speech technology. A preliminary investigation on deep neural network adaptation is carried out in [47], where a feature-space maximum-likelihood linear regression technique is casted into the deep neural network framework. This method also adds one additional layer and freezes other parameters to perform adaptation, and it therefore has the same limitation of the LIN and LHN. The proposed approach, as already demonstrated for the single-hidden-layer MLP, overcomes LIN/LHN limitation while attaining superior performance especially with a single adaptation spoken utterance. Hermitian-based activation function can be easily integrated into deep neural networks. A recent study [48] shows the great power of deep neural networks when adapting all the network parameters with KL divergence regularizer. However, the adaption of all deep neural network parameters sometimes requires large amount of adaptation utterances and in practice it is hard to store an individual deep neural network for each speaker. In contrast, the study in this paper may bring a better but powerful way to adapt deep neural network since it only needs to adapt limited hidden activation function.

REFERENCES

[1] J. M. Baker, L. Deng, J. Glass, S. Khudanpur, C. Lee, N. Morgan, and D. OShaughnessy, "Research developments and directions in speech recognition and understanding, part 1," *IEEE Signal Process. Mag.*, vol. 26, no. 3, pp. 75–80, May. 2009.

[2] J. M. Baker, L. Deng, J. Glass, S. Khudanpur, C. Lee, N. Morgan, and D. OShaughnessy, "Research developments and directions in speech recognition and understanding, part 2," *IEEE Signal Process. Mag.*, vol. 26, no. 3, pp. 78–85, May 2009.

[3] C.-H. Lee, "On stochastic feature and model compensation approaches to robust speech recognition," *Speech Commun.*, vol. 25, no. 1-3, pp. 29–47, 1998.

[4] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Commun.*, vol. 16 (3), pp. 261–291, 1995.

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

[6] M. Afify, O. Siohan, and C.-H. Lee, "Upper and lower bounds on the mean of noisy speech: application to minimax classification," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 2, pp. 79–88, 2002.

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

[8] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, , and 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, pp. 82–97, Nov. 2012.

[9] J. Neto, L. Almeida, M. Hochberg, C. Martins, L. Nunes, S. Renals, and T. Robinson, "Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system," in *Proc. Eurospeech*, Madrid, Spain, Sept. 1995, pp. 2171–2174.

[10] V. Abrash, H. Franco, A. Sankar, and M. Cohen, "Connectionist speaker normalization and adaptation," in *Proc. Eurospeech*, Madrid, Spain, Sept. 1995, pp. 2183–2186.

[11] R. Gemello, F. Mana, S. Scanzio, P. Laface, and R. De Mori, "Linear hidden transformations for adaptation of hybrid ANN/HMM models," *Speech Commun.*, vol. 49, no. 10-11, pp. 827–835, 2007.

[12] S. M. Siniscalchi, T. Svendsen, S. Sorbello, and C.-H. Lee, "Experimental studies on continuous speech recognition using neural architectures with "adaptive" hidden activation functions," in *Proc. ICASSP*, Dallas, TX, USA, Mar. 2010, pp. 4882–4885.

[13] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," in *Proc. Workshop on Speech and Natural Language*, Banff, Canada, Oct. 1992, pp. 899–902.

[14] S. M. Siniscalchi, J. Li, and C.-H. Lee, "Hermitian based hidden activation functions for adaptation of hybrid HMM/ANN models," in *Proc. Interspeech*, Portland, OR, USA, Sept. 2012.

[15] G. Zavaliagkos, R. Schwartz, and J. Makhou, "Batch, incremental and instantaneous adaptation techniques for speech recognition," in *Proc. ICASSP*, Detroit, MI, USA, May 1995, pp. 676–679.

[16] C. J. Legetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of the parameters of continuous density hidden Markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, 1995.

[17] M. J. F. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech & Language*, vol. 10, pp. 249–264, 1996.

[18] V. V. Digalakis, D. Rtischev, and L. G. Neumeye, "Speaker adaptation using constrained estimation of gaussian mixtures," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 4, pp. 357–366, Sept. 1995.

[19] M. J. F. Gales, "Cluster adaptive training of hidden Markov models," *Computer, Speech, and Language*, vol. 12, pp. 75–98, 1998.

[20] J. Gauvain and C. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, 1994.

[21] K. Shinoda and C.-H. Lee, "A structural Bayes approach to speaker adaptation," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 276–287, 2001.

[22] O. Siohan, C. Chesta, and C.-H. Lee, "Joint maximum a posteriori adaptation of transformation and HMM parameters," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 4, pp. 417–428, 2001.

[23] Q. Huo and C.-H. Lee, "On-line adaptive learning of the correlated continuous density hidden Markov model for speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 4, pp. 386–397, 1998.

[24] A. Gunawardana and W. Byrne, "Discriminative speaker adaptation with conditional maximum likelihood linear regression," in *Proc. Eurospeech*, Aalborg, Denmark, Sept. 2001.

[25] S. Tsakalidis, V. Doumpiotis, and W. Byrne, "Discriminative linear transforms for feature normalization and speaker adaptation in HMM estimation," in *Proc. ICSLP*, Denver, CO, USA, Sept. 2002.

[26] D. Povey, M. J. F. Gales, D. Y. Kim, and P. Woodland, "MMI-MAP and MPE-MAP for acoustic model adaptation," in *Proc. Interspeech*, Geneva, Switzerland, Sept. 2003.

[27] X. Li and J. Bilmes, "Regularized adaptation of discriminative classifiers," in *Proc. ICASSP*, Toulouse, France, May 2006, pp. 237–240.

[28] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. John Wiley & Sons, 1973.

[29] B. D. Ripley, *Pattern Recognition and Neural Networks*. Cambridge Univ. Press, 1996.

[30] Y. Kharin, *Robustness in Statistical Pattern Recognition*. Kluwer Academic Publishers, 1996.

[31] B.-H. Juang, S. E. Levinson, and M. M. Sondhi, "Maximum likelihood estimation for multivariate mixture observations of Markov chains," *IEEE Trans. Inf. Theory*, vol. 32, no. 2, pp. 307–309, 1986.

[32] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

[33] S. Gaglio, G. Pilato, F. Sorbello, and G. Vassallo, "Using the hermite regression formula to design a neural architecture with automatic learning of the "hidden" activation functions," in *AI*IA*, Bologna, Italy, Sept. 1999, pp. 226–237.

[34] L. Ma and K. Khorasani, "Constructive feedforward neural networks using hermite polynomial activation functions," *IEEE Trans. Neural networks*, vol. 16 (4), pp. 821–833, 2005.

[35] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Macmillan, 1994.

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

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

[38] X. He, L. Deng, and W. Chou, "A novel learning method for hidden markov models in speech and audio processing," in *Proc. MMSP*, Victoria, BC, Oct. 2006.

[39] S. Wiesler, G. Heigold, M. Nussbaum-Thom, R. Schlueter, and H. Ney, "A discriminative splitting criterion for phonetic decision trees," in *Proc. Interspeech*, Makuhari, Japan, Sept. 2010.

[40] B. Li and K. C. Sim, "Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems," in *Proc. Interspeech*, Makuhari, Chiba, Japan, Sept. 2010.

[41] L. Lee and R. C. Rose, "Speaker normalization using efcient frequency warping procedures," in *Proc. ICASSP*, Atlanta, GA, USA, May 1996.

[42] S. M. Siniscalchi, D.-C. Lyu, T. Svendsen, and C.-H. Lee, "Experiments on cross-language attribute detection and phone recognition with minimal target specific training data," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 3, pp. 875–887, 2012.

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

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

[45] J. Xue, J. Li, and Y. Gong, "Restructuring of deep neural network acoustic models with singular value decomposition," in *Proc. Interspeech*, Lyon, France, Aug. 2013.

[46] V. Vanhoucke, M. Devin, and G. Heigold, "Multiframe deep neural networks for acoustic modeling," in *Proc. ICASSP*, Vancouver, Canada, May 2013.

[47] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. ASRU*, Hawaii, USA, Dec. 2011, pp. 24–29.

[48] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Proc. ICASSP*, Vancouver, Canada, May 2013.

[49] C.-H. Lee, "From knowledge-ignorant to knowledge-rich modeling: A new speech research paradigm for next generation automatic speech recognition," in *Proc. Interspeech*, Jeju Island, Korea, Oct. 2004, pp. 109–112.

[50] I. Bromberg, J. Morris, and E. Fosler-Lussier, "Joint versus independent phonological feature models within crf phone recognition," in *Proc. NAACL-HLT (short paper session)*, NY, USA, 2007.

[51] S. M. Siniscalchi, P. Schwarz, and C.-H. Lee, "High-accuracy phone recognition by combining high-performance lattice generation and knowledge based rescoring," in *Proc. ICASSP*, Honolulu, HI, USA, Apr. 2007, pp. IV–869–V–872.

[52] S. M. Siniscalchi, J. Reed, T. Svendsen, and C.-H. Lee, "Exploring universal attribute characterization of spoken languages for spoken language recognition," in *Proc. Interspeech*, Brighton, UK, Sept. 2009, pp. 168–171.

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 2006, 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 at the Department of Electronics and Telecommunications under the guidance of Prof. T. Svendsen. In 2010, he was a Researcher Scientist at 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.

Jinyu Li (M) joined Microsoft Corporation in 2008. He received the B.Eng and M.Eng degrees in electrical engineering and information system from University of Science and Technology of China, Hefei, in 1997 and 2000, and the Ph.D. degree from the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, in 2008. From 2000 to 2003, he was a Researcher in the Intel China Research Center and iFlytek Speech, China. Currently, he is a senior lead scientist in Microsoft Corporation, Redmond, WA. His major research interests cover several topics in speech recognition, including discriminative training, noise robustness, feature extraction, and machine learning methods.

Chin-Hui Lee Chin-Hui Lee is a professor at School of Electrical and Computer Engineering, Georgia Institute of Technology. Dr. 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.

Dr. 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 of Engineering at Georgia Institute of Technology.

Prof. Lee has participated actively in professional societies. He is a member of the IEEE Signal Processing Society (SPS), and 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 promote the SPS Multimedia Signal Processing Technical Committee in which he is a founding member.

Dr. Lee is a Fellow of the IEEE, and has published close to 400 papers and 30 patents. 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. Dr. 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. He won the IEEE SPS's 2006 Technical Achievement Award for "Exceptional Contributions to the Field of Automatic Speech Recognition". He was one of the four plenary speakers at IEEE ICASSP, held in Kyoto, Japan in April 2012. More recently, he was awarded the 2012 ISCA Medal for "pioneering and seminal contributions to the principles and practices of automatic speech and speaker recognition, including fundamental innovations in adaptive learning, discriminative training and utterance verification."