AUTHOR QUERY FORM

ELSEVIER	**Book: Robust Automatic Speech Recognition Chapter: 00002**	**Please e-mail your responses and any corrections to: E-mail: s.li@elsevier.com**

Dear Author,

Any queries or remarks that have arisen during the processing of your manuscript are listed below and are highlighted by flags in the proof. (AU indicates author queries; ED indicates editor queries; and TS/ TY indicates typesetter queries.) Please check your proof carefully and answer all AU queries. Mark all corrections and query answers at the appropriate place in the proof (e.g., by using on-screen annotation in the PDF file http://www.elsevier.com/book-authors/science-and-technology-book-publishing/overview-of-the-publishing-process) or compile them in a separate list, and tick off below to indicate that you have answered the query. **Please return your input as instructed by the project manager.**

Location in Chapter	**Query/remark**	
AU:1, page 16	Please check and approve the edits made in the sentence: “These posteriors are	□
AU:2, page 24	Please provide better quality figure.	□
AU:3, page 32	Please provide volume and page range for Abdel-Hamid et al. (2014).	□
AU:4, page 38	Please provide volume and page numbers for Rumelhart et al. (1988).	□

CHAPTER 2

c0010 Fundamentals of speech recognition

CHAPTER OUTLINE

s0005 2.1 INTRODUCTION: COMPONENTS OF SPEECH RECOGNITION

p0005 Speech recognition has been an active research area for many years. It is not until recently, over the past 2 years or so, the technology has passed the usability bar for many real-world applications under most realistic acoustic environments (Yu and Deng, 2014). Speech recognition technology has started to change the way we live and work and has became one of the primary means for humans to interact with mobile devices (e.g., Siri, Google Now, and Cortana). The arrival of this new trend is attributed to the significant progress made in a number of areas. First, Moore's law continues to dramatically increase computing power, which, through multi-core processors, general purpose graphical processing units, and clusters, is nowadays several orders of magnitude higher than that available only a decade ago (Baker et al., 2009a,b; Yu and Deng, 2014). The high power of computation

Robust Automatic Speech Recognition. http://dx.doi.org/10.1016/B978-0-12-802398-3.00002-7

makes training of powerful deep learning models possible, dramatically reducing the error rates of speech recognition systems (Sak et al., 2014a). Second, much more data are available for training complex models than in the past, due to the continued advances in Internet and cloud computing. Big models trained with big and real-world data allow us to eliminate unrealistic model assumptions (Bridle et al., 1998; Deng, 2003; Juang, 1985), creating more robust ASR systems than in the past (Deng and O'Shaughnessy, 2003; Huang et al., 2001b; Rabiner, 1989). Finally, mobile devices, wearable devices, intelligent living room devices, and in-vehicle infotainment systems have become increasingly popular. On these devices, interaction modalities such as keyboard and mouse are less convenient than in personal computers. As the most natural way of human-human communication, speech is a skill that all people already are equipped with. Speech, thus, naturally becomes a highly desirable interaction modality on these devices.

p0010 From the technical point of view, the goal of speech recognition is to predict the optimal word sequence **W**, given the spoken speech signal **X**, where optimality refers to maximizing the *a posteriori* probability (maximum *a posteriori*, MAP) :

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmax}}\, P_{\Lambda,\Gamma}(\mathbf{W}|\mathbf{X}), \tag{2.1}$$

where Λ and Γ are the acoustic model and language model parameters. Using Bayes' rule

$$P_{\Lambda,\Gamma}(\mathbf{W}|\mathbf{X}) = \frac{p_{\Lambda}(\mathbf{X}|\mathbf{W})P_{\Gamma}(\mathbf{W})}{p(\mathbf{X})}, \tag{2.2}$$

Equation 2.1 can be re-written as:

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmax}}\, p_{\Lambda}(\mathbf{X}|\mathbf{W})P_{\Gamma}(\mathbf{W}), \tag{2.3}$$

where $p_{\Lambda}(\mathbf{X}|\mathbf{W})$ is the AM likelihood and $P_{\Gamma}(\mathbf{W})$ is the LM probability. When the time sequence is expanded and the observations $\mathbf{x}_t$ are assumed to be generated by hidden Markov models (HMMs) with hidden states θ_t, we have

$$\hat{\mathbf{W}} = \underset{\mathbf{W}}{\operatorname{argmax}}\, P_{\Gamma}(\mathbf{W}) \sum_{\theta} \prod_{t=1}^{T} p_{\Lambda}(\mathbf{x}_t|\theta_t)P_{\Lambda}(\theta_t|\theta_{t-1}), \tag{2.4}$$

where θ belongs to the set of all possible state sequences for the transcription W. The speech signal is first processed by the feature extraction module to obtain the acoustic feature. The feature extraction module is often referred as the front-end of speech recognition systems. The acoustic features will be passed to the acoustic model and the language model to compute the probability of the word sequence under consideration. The output is a word sequence with the largest probability from acoustic and language models. The combination of acoustic and language models are usually referred as the back-end of speech recognition systems. The focus of this book is on the noise-robustness of front-end and acoustical model, therefore, the robustness of language model is not considered in the book.

p0015 Acoustic models are used to determine the likelihood of acoustic feature sequences given hypothesized word sequences. The research in speech recognition has been under a long period of development since the HMM was introduced in 1980s as the acoustic model (Juang, 1985; Rabiner, 1989). The HMM is able to gracefully represent the temporal evolution of speech signals and characterize it as a parametric random process. Using the Gaussian mixture model (GMM) as its output distribution, the HMM is also able to represent the spectral variation of speech signals.

p0020 In this chapter, we will first review the GMM, and then review the HMM with the GMM as its output distribution. Finally, the recent development in speech recognition has demonstrated superior performance of the deep neural network (DNN) over the GMM in discriminating speech classes (Dahl et al., 2011; Yu and Deng, 2014). A review of the DNN and related deep models will thus be provided.

s0010 2.2 GAUSSIAN MIXTURE MODELS

p0025 As part of acoustic modeling in ASR and according to how the acoustic emission probabilities are modeled for the HMMs' state, we can have discrete HMMs (Liporace, 1982), semi-continuous HMMs (Huang and Jack, 1989), and continuous HMMs (Levinson et al., 1983). For the continuous output density, the most popular one is the Gaussian mixture model (GMM), in which the state output density is modeled as:

$$P_{\Lambda}(o) = \sum_{i} c(i)\mathcal{N}(o; \mu(i), \sigma^2(i)), \tag{2.5}$$

where $\mathcal{N}(o; \mu(i), \sigma^2(i))$ is a Gaussian with mean $\mu(i)$ and variance $\sigma^2(i)$, and $c(i)$ is the weight for the ith Gaussian component. Three fundamental problems of HMMs are probability evaluation, determination of the best state sequence, and parameter estimation (Rabiner, 1989). The probability evaluation can be realized easily with the forward algorithm (Rabiner, 1989).

p0030 The parameter estimation is solved with the maximum likelihood estimation (MLE) (Dempster et al., 1977) using a forward-backward procedure (Rabiner, 1989). The quality of the acoustic model is the most important issue for ASR. MLE is known to be optimal for density estimation, but it often does not lead to minimum recognition error, which is the goal of ASR. As a remedy, several discriminative training (DT) methods have been proposed in recent years to boost ASR system accuracy. Typical methods are maximum mutual information estimation (MMIE) (Bahl et al., 1997), minimum classification error (MCE) (Juang et al., 1997), minimum word/phone error (MWE/MPE) (Povey and Woodland, 2002), minimum Bayes risk (MBR) (Gibson and Hain, 2006), and boosted MMI (BMMI) (Povey et al., 2008). Other related methods can be found in He and Deng (2008), He et al. (2008), and Xiao et al. (2010).

p0035 Inspired by the high success of margin-based classifiers, there is a trend toward incorporating the margin concept into hidden Markov modeling for ASR. Several attempts based on margin maximization were proposed, with three major classes of methods: large margin estimation (Jiang et al., 2006; Li and Jiang, 2007), large margin HMMs (Sha, 2007; Sha and Saul, 2006), and soft margin estimation (SME) (Li et al., 2006, 2007b). The basic concept behind all these margin-based methods is that by securing a margin from the decision boundary to the nearest training sample, a correct decision can still be made if the mismatched test sample falls within a tolerance region around the original training samples defined by the margin.

p0040 The main motivations of using the GMM as a model for the distribution of speech features are discussed here. When speech waveforms are processed into compressed (e.g., by taking logarithm of) short-time Fourier transform magnitudes or related cepstra, the GMM has been shown to be quite appropriate to fit such speech features when the information about the temporal order is discarded. That is, one can use the GMM as a model to represent frame-based speech features.

p0045 Both inside and outside the ASR domain, the GMM is commonly used for modeling the data and for statistical classification. GMMs are well known for their ability to represent arbitrarily complex distributions with multiple modes. GMM-based classifiers are highly effective with widespread use in speech research, primarily for speaker recognition, denoising speech features, and speech recognition. For speaker recognition, the GMM is directly used as a universal background model (UBM) for the speech feature distribution pooled from all speakers. In speech feature denoising or noise tracking applications, the GMM is used in a similar way and as a prior distribution for speech (Deng et al., 2003, 2002a,b; Frey et al., 2001a; Huang et al., 2001a). In ASR applications, the GMM is integrated into the doubly stochastic model of HMM as its output distribution conditioned on a state, which will be discussed later in more detail.

p0050 GMMs have several distinct advantages that make them suitable for modeling the distributions over speech feature vectors associated with each state of an HMM. With enough components, they can model distributions to any required level of accuracy, and they are easy to fit to data using the EM algorithm. A huge amount of research has gone into finding ways of constraining GMMs to increase their evaluation speed and to optimize the tradeoff between their flexibility and the amount of training data required to avoid overfitting. This includes the development of parameter- or semi-tied GMMs and subspace GMMs.

p0055 Despite all their advantages, GMMs have a serious shortcoming. That is, GMMs are statistically inefficient for modeling data that lie on or near a nonlinear manifold in the data space. For example, modeling the set of points that lie very close to the surface of a sphere only requires a few parameters using an appropriate model class, but it requires a very large number of diagonal Gaussians or a fairly large number of full-covariance Gaussians. It is well known that speech is produced by modulating a relatively small number of parameters of a dynamical system (Deng, 1999, 2006; Lee et al., 2001). This suggests that the true underlying structure of speech is of a much lower dimension than is immediately apparent in a window that contains hundreds of coefficients. Therefore, other types of model that can capture better the properties of speech features are expected to work better than GMMs for acoustic modeling of speech. In particular, the new models should more effectively exploit information embedded in a large window of frames of speech features than GMMs. We will return to this important problem of characterizing speech features after discussing a model, the HMM, for characterizing temporal properties of speech next.

s0015 2.3 HIDDEN MARKOV MODELS AND THE VARIANTS

p0060 As a highly special or degenerative case of the HMM, we have the Markov chain as an information source capable of generating observational output sequences. Then we can call the Markov chain an observable (non-hidden) Markov model because its output has one-to-one correspondence to a state in the model. That is, each state corresponds to a deterministically observable variable or event. There is no randomness in the output in any given state. This lack of randomness makes the Markov chain too restrictive to describe many real-world informational sources, such as speech feature sequences, in an adequate manner.

p0065 The Markov property, which states that the probability of observing a certain value of the random process at time t only depends on the immediately preceding observation at $t-1$, is rather restrictive in modeling correlations in a random process. Therefore, the Markov chain is extended to give rise to a HMM, where the states, that is, the values of the Markov chain, are “hidden” or non-observable. This extension is accomplished by associating an observation probability distribution with each state in the Markov chain. The HMM thus defined is a doubly embedded random sequence whose underlying Markov chain is not directly observable. The underlying Markov chain in the HMM can be observed only through a separate random function characterized by the observation probability distributions. Note that the observable random process is no longer a Markov process and thus the probability of an observation not only depends on the immediately preceding observations.

s0020 2.3.1 HOW TO PARAMETERIZE AN HMM

p0070 We can give a formal parametric characterization of an HMM in terms of its model parameters:

o0005 **1.** State transition probabilities, $\mathbf{A} = [a_{ij}]$, $i,j = 1,2,\ldots,N$, of a homogeneous Markov chain with a total of N states

$$a_{ij} = P(\theta_t = j|\theta_{t-1} = i), \qquad i,j = 1,2,\ldots,N. \tag{2.6}$$

o0010 **2.** Initial Markov chain state-occupation probabilities: $\pi = [\pi_i]$, $i = 1,2,\ldots,N$, where $\pi_i = P(\theta_1 = i)$.

o0015 **3.** Observation probability distribution, $P(\mathbf{o}_t|\theta_t = i), i = 1,2,\ldots,N$. if $\mathbf{o}_t$ is discrete, the distribution associated with each state gives the probabilities of symbolic observations $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_K\}$:

$$b_i(k) = P[\mathbf{o}_t = \mathbf{v}_k | \theta_t = i], \qquad i = 1, 2, \ldots, N. \tag{2.7}$$

If the observation probability distribution is continuous, then the parameters, Λ_i, in the probability density function (PDF) characterize state i in the HMM.

p0075 The most common and successful distribution used in ASR for characterizing the continuous observation probability distribution in the HMM is the GMM discussed in the preceding section. The GMM distribution with vector-valued observations ($\mathbf{o}_t \in \mathcal{R}^{\mathcal{D}}$) has the mathematical form:

$$\begin{aligned} b_i(\mathbf{o}_t) &= P(\mathbf{o}_t | \theta_t = i) \\ &= \sum_{m=1}^{M} \frac{c(i,m)}{(2\pi)^{D/2} |\boldsymbol{\Sigma}(i,m)|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{o}_t - \boldsymbol{\mu}(i,m))^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(i,m)(\mathbf{o}_t - \boldsymbol{\mu}(i,m))\right] \end{aligned} \tag{2.8}$$

p0080 In this GMM-HMM, the parameter set Λ_i comprises scalar mixture weights, $c(i,m)$, Gaussian mean vectors, $\boldsymbol{\mu}(i,m) \in \mathcal{R}^{\mathcal{D}}$, and Gaussian covariance matrices, $\boldsymbol{\Sigma}(i,m) \in \mathcal{R}^{D \times D}$.

p0085 When the number of mixture components is reduced to one: $M = 1$, the state-dependent output PDF reverts to a (uni-modal) Gaussian:

$$b_i(\mathbf{o}_t) = \frac{1}{(2\pi)^{D/2} |\boldsymbol{\Sigma}_i|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{o}_t - \boldsymbol{\mu}(i))^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(i)(\mathbf{o}_t - \boldsymbol{\mu}(i))\right] \tag{2.9}$$

and the corresponding HMM is commonly called a (continuous-density) Gaussian HMM.

s0025 2.3.2 EFFICIENT LIKELIHOOD EVALUATION FOR THE HMM

p0090 Likelihood evaluation is a basic task needed for speech processing applications involving an HMM that uses a hidden Markov sequence to approximate vectorized speech features.

p0095 Let $\theta_1^T = (\theta_1, \ldots, \theta_T)$ be a finite-length sequence of states in a Gaussian-mixture HMM or GMM-HMM, and let $P(\mathbf{o}_1^T, \theta_1^T)$ be the joint likelihood of the observation sequence $\mathbf{o}_1^T = (\mathbf{o}_1, \ldots, \mathbf{o}_T)$ and the state sequence θ_1^T. Let $P(\mathbf{o}_1^T | \theta_1^T)$ denote the likelihood that the observation sequence $\mathbf{o}_1^T$ is generated by the model conditioned on the state sequence θ_1^T.

p0100 In the Gaussian-mixture HMM, the conditional likelihood $P(\mathbf{o}_1^T | \theta_1^T)$ is in the form of

$$\begin{aligned} P(\mathbf{o}_1^T | \boldsymbol{\theta}_1^T) = \prod_{t=1}^{T} b_i(\mathbf{o}_t) = \prod_{t=1}^{T} \sum_{m=1}^{M} \frac{c(i,m)}{(2\pi)^{D/2} |\boldsymbol{\Sigma}_{i,m}|^{1/2}} \\ \exp\left[-\frac{1}{2}(\mathbf{o}_t - \boldsymbol{\mu}(i,m))^{T} \boldsymbol{\Sigma}^{-1}(i,m)(\mathbf{o}_t - \boldsymbol{\mu}(i,m))\right] \end{aligned} \tag{2.10}$$

p0105 On the other hand, the probability of state sequence θ_1^T is just the product of transition probabilities, that is,

$$P(\boldsymbol{\theta}_1^T) = \pi_{\theta_1} \prod_{t=1}^{T-1} a_{\theta_t \theta_{t+1}}. \tag{2.11}$$

In the remainder of the chapter, for notational simplicity, we consider the case where the initial state distribution has probability of one in the starting state: $\pi_1 = P(\theta_1 = 1) = 1$.

p0110 Note that the joint likelihood $P(\mathbf{o}_1^T, \theta_1^T)$ can be obtained by the product of likelihoods in Equations 2.10 and 2.11:

$$P(\mathbf{o}_1^T, \boldsymbol{\theta}_1^T) = P(\mathbf{o}_1^T | \boldsymbol{\theta}_1^T) P(\boldsymbol{\theta}_1^T). \tag{2.12}$$

In principle, the total likelihood for the observation sequence can be computed by summing the joint likelihoods in Equation 2.12 over all possible state sequences θ_1^T:

$$P(\mathbf{o}_1^T) = \sum_{\boldsymbol{\theta}_1^T} P(\mathbf{o}_1^T, \boldsymbol{\theta}_1^T). \tag{2.13}$$

However, the computational effort is exponential in the length of the observation sequence, T, and hence the naive computation of $P(\mathbf{o}_1^T)$ is not tractable. The forward-backward algorithm (Baum and Petrie, 1966) computes $P(\mathbf{o}_1^T)$ for the HMM with complexity linear in T.

p0115 To describe this algorithm, we first define the forward probabilities by

$$\alpha_t(i) = P(\theta_t = i, \mathbf{o}_1^t), \quad t = 1, \ldots, T, \tag{2.14}$$

and the backward probabilities by

$$\beta_t(i) = P(\mathbf{o}_{t+1}^T | \theta_t = i), \quad t = 1, \ldots, T-1, \tag{2.15}$$

both for each state i in the Markov chain. The forward and backward probabilities can be calculated recursively from

$$\alpha_t(j) = \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} b_j(\mathbf{o}_t), \quad t = 2, 3, \ldots, T; \qquad j = 1, 2, \ldots, N \tag{2.16}$$

$$\beta_t(i) = \sum_{j=1}^{N} \beta_{t+1}(j) a_{ij} b_j(\mathbf{o}_{t+1}), \quad t = T-1, T-2, \ldots, 1; \qquad i = 1, 2, \ldots, N \tag{2.17}$$

Proofs of these recursions are given in the following section. The starting value for the α recursion is, according to the definition in Equation 2.14,

$$\alpha_1(i) = P(\theta_1 = i, \mathbf{o}_1) = P(\theta_1 = i) P(\mathbf{o}_1 | \theta_1) = \pi_i b_i(\mathbf{o}_1), \quad i = 1, 2, ...N \tag{2.18}$$

and that for the β recursion is chosen as

$$\beta_T(i) = 1, \qquad i = 1, 2, ...N, \tag{2.19}$$

so as to provide the correct values for β_{T-1}according to the definition in Equation 2.15.

p0120 To compute the total likelihood $P(\mathbf{o}_1^T)$ in Equation 2.13, we first compute

$$\begin{aligned} P(\theta_t = i, \mathbf{o}_1^T) &= P(\theta_t = i, \mathbf{o}_1^t, \mathbf{o}_{t+1}^T) \\ &= P(\theta_t = i, \mathbf{o}_1^t) P(\mathbf{o}_{t+1}^T | \mathbf{o}_1^t, \theta_t = i) \\ &= P(\theta_t = i, \mathbf{o}_1^t) P(\mathbf{o}_{t+1}^T | \theta_t = i) \\ &= \alpha_t(i) \beta_t(i), \end{aligned} \tag{2.20}$$

for each state i and $t = 1, 2, \ldots, T$ using definitions in Equations 2.14 and 2.15. Note that $P(\mathbf{o}_{t+1}^T | \mathbf{o}_1^t, \theta_t = i) = P(\mathbf{o}_{t+1}^T | \theta_t = i)$ because the observations are independent, given the state in the HMM. Given this, $P(\mathbf{o}_1^T)$ can be computed as

$$P(\mathbf{o}_1^T) = \sum_{i=1}^N P(\theta_t = i, \mathbf{o}_1^T) = \sum_{i=1}^N \alpha_t(i) \beta_t(i). \tag{2.21}$$

p0125 With Equation 2.20 we find for the posterior probability of being in state i at time t given the whole sequence of observed data

$$\gamma_t(i) = P(\theta_t = i | \mathbf{o}_1^T) = \frac{\alpha_t(i)\beta_t(i)}{P(\mathbf{o}_1^T)}. \tag{2.22}$$

Further, we can find for the posterior probability of the state-transition probabilities

$$\xi_t(i,j) = P(\theta_t = j, \theta_{t-1} = i | \mathbf{o}_1^T) = \frac{\alpha_{t-1}(i) a_{ij} P(\mathbf{o}_t | \theta_t = j) \beta_t(j)}{P(\mathbf{o}_1^T)} \tag{2.23}$$

These posteriors are needed to learn about the HMM parameters, as will be explained in the following section. AU1

p0130 Taking $t = T$ in Equation 2.21 and using Equation 2.19 lead to

$$P(\mathbf{o}_1^T) = \sum_{i=1}^N \alpha_T(i). \tag{2.24}$$

Thus, strictly speaking, the backward recursion, Equation 2.17 is not necessary for the forward scoring computation, and hence the algorithm is often called the forward algorithm. However, the β computation is a necessary step for solving the model parameter estimation problem, which will be briefly described in the following section.

s0030 2.3.3 EM ALGORITHM TO LEARN ABOUT THE HMM PARAMETERS

p0135 Despite many unrealistic aspects of the HMM as a model for speech feature sequences, one most important reason for its wide-spread use in ASR is the Baum-Welch algorithm developed in 1960s (Baum and Petrie, 1966), which is a prominent instance of the highly popular EM (expectation-maximization) algorithm (Dempster et al., 1977), for efficient training of the HMM parameters from data.

p0140 The EM algorithm is a general iterative technique for maximum likelihood estimation, with local optimality, when hidden variables exist. When such hidden variables take the form of a Markov chain, the EM algorithm becomes the Baum-Welch algorithm. Here we use a Gaussian HMM as the example to describe steps involved in deriving E- and M-step computations, where the complete data in the general case of EM above consists of the observation sequence and the hidden Markov-chain state sequence; that is, $[\mathbf{o}_1^T, \theta_1^T]$.

p0145 Each iteration in the EM algorithm consists of two steps for any incomplete data problem including the current HMM parameter estimation problem. In the E (expectation) step of the Baum-Welch algorithm, the following conditional expectation, or the auxiliary function $Q(\theta|\theta_0)$, need to be computed:

$$Q(\Lambda; \Lambda_0) = E[\log P(\mathbf{o}_1^T, \boldsymbol{\theta}_1^T|\Lambda)|\mathbf{o}_1^T, \Lambda_0], \tag{2.25}$$

where the expectation is taken over the "hidden" state sequence θ_1^T. For the EM algorithm to be of utility, $Q(\Lambda; \Lambda_0)$ has to be sufficiently simplified so that the M (maximization) step can be carried out easily. Estimates of the model parameters are obtained in the M-step via maximization of $Q(\Lambda; \Lambda_0)$, which is in general much simpler than direct procedures for maximizing $P(\mathbf{o}_1^T|\Lambda)$.

p0150 An iteration of the above two steps will lead to maximum likelihood estimates of model parameters with respect to the objective function $P(\mathbf{o}_1^T|\Lambda)$. This is a direct consequence of Baum's inequality (Baum and Petrie, 1966), which asserts that

$$\log\left(\frac{P(\mathbf{o}_1^T|\Lambda)}{P(\mathbf{o}_1^T|\Lambda_0)}\right) \geq Q(\Lambda; \Lambda_0) - Q(\Lambda_0; \Lambda_0).$$

p0155 After carrying out the E- and M-steps for the Gaussian HMM, details of which are omitted here but can be found in Rabiner (1989) and Huang et al. (2001b), we can establish the re-estimation formulas for the maximum-likelihood estimates of its parameters:

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}, \tag{2.26}$$

where $\xi_t(i,j)$ and $\gamma_t(i)$ are the posterior state-transition and state-occupancy probabilities computed from the E-step.

p0160 The re-estimation formula for the covariance matrix in state i of an HMM can be derived to be

$$\hat{\boldsymbol{\Sigma}}_i = \frac{\sum_{t=1}^{T} \gamma_t(i)(\mathbf{o}_t - \hat{\boldsymbol{\mu}}(i))(\mathbf{o}_t - \hat{\boldsymbol{\mu}}(i))^{\mathrm{T}}}{\sum_{t=1}^{T} \gamma_t(i)} \tag{2.27}$$

for each state: $i = 1, 2, \ldots, N$, where $\hat{\boldsymbol{\mu}}(i)$ is the re-estimate of the mean vector in the Gaussian HMM in state i, whose re-estimation formula is also straightforward to derive and has the following easily interpretable form:

$$\hat{\boldsymbol{\mu}}(i) = \frac{\sum_{t=1}^{T} \gamma_t(i)\mathbf{o}_t}{\sum_{t=1}^{T} \gamma_t(i)}. \tag{2.28}$$

p0165 The above derivation is for the single-Gaussian HMM. The EM algorithm for the GMM-HMM can be similarly determined by considering the Gaussian component of each frame at each state as another hidden variable. In a later section, we will describe the deep neural network (DNN)-HMM hybrid system in which the observation probability is estimated using a DNN.

s0035 2.3.4 HOW THE HMM REPRESENTS TEMPORAL DYNAMICS OF SPEECH

p0170 The popularity of the HMM in ASR stems from its ability to serve as a generative sequence model of acoustic features of speech; see excellent reviews of HMMs for selected speech modeling and recognition applications as well as the limitations of HMMs in Rabiner (1989), Jelinek (1976), Baker (1976), and Baker et al. (2009a,b). One most interesting and unique problem in speech modeling and in the related speech recognition application lies in the nature of variable length in acoustic-feature sequences. This unique characteristic of speech rests primarily in its temporal dimension. That is, the actual values of the speech feature are correlated lawfully with the elasticity in the temporal dimension. As a consequence, even if two word sequences are identical, the acoustic data of speech features typically have distinct lengths. For example, different acoustic samples from the same sentence usually contain different data dimensionality, depending on how the speech sounds are produced and in particular how fast the speaking rate is. Further, the discriminative cues among separate speech classes are often distributed over a reasonably long temporal span, which often crosses neighboring speech units. Other special aspects of speech include class-dependent acoustic cues. These cues are often expressed over diverse time spans that would benefit from different lengths of analysis windows in speech analysis and feature extraction.

p0175 Conventional wisdom posits that speech is a one-dimensional temporal signal in contrast to image and video as higher dimensional signals. This view is simplistic and does not capture the essence and difficulties of the speech recognition problem. Speech is best viewed as a two-dimensional signal, where the spatial (or frequency or tonotopic) and temporal dimensions have vastly different characteristics, in contrast to images where the two spatial dimensions tend to have similar properties.

The spatial dimension in speech is associated with the frequency distribution and related transformations, capturing a number of variability types including primarily those arising from environments, speakers, accent, and speaking style and rate. The latter induces correlations between spatial and temporal dimensions, and the environment factors include microphone characteristics, speech transmission channel, ambient noise, and room reverberation.

p0180 The temporal dimension in speech, and in particular its correlation with the spatial or frequency-domain properties of speech, constitutes one of the unique challenges for speech recognition. The HMM addresses this challenge to a limited extent. In the following two sections, a selected set of advanced generative models, as various extensions of the HMM, will be described that are aimed to address the same challenge, where Bayesian approaches are used to provide temporal constraints as prior knowledge about aspects of the physical process of human speech production.

s0040 2.3.5 GMM-HMMs FOR SPEECH MODELING AND RECOGNITION

p0185 In speech recognition, one most common generative learning approach is based on the Gaussian-mixture-model based hidden Markov models, or GMM-HMM (Bilmes, 2006; Deng and Erler, 1992; Deng et al., 1991a; Juang et al., 1986; Rabiner, 1989; Rabiner and Juang, 1993). As discussed earlier, a GMM-HMM is a statistical model that describes two dependent random processes, an observable process and a hidden Markov process. The observation sequence is assumed to be *generated* by each hidden state according to a Gaussian mixture distribution. A GMM-HMM is parameterized by a vector of state prior probabilities, the state transition probability matrix, and by a set of state-dependent parameters in Gaussian mixture models. In terms of modeling speech, a state in the GMM-HMM is typically associated with a sub-segment of a phone in speech. One important innovation in the use of HMMs for speech recognition is the introduction of context-dependent states (Deng et al., 1991b; Huang et al., 2001b), motivated by the desire to reduce output variability of speech feature vectors associated with each state, a common strategy for "detailed" generative modeling. A consequence of using context dependency is a vast expansion of the HMM state space, which, fortunately, can be controlled by regularization methods such as state tying. It turns out that such context dependency also plays a critical role in the recent advance of speech recognition in the area of discrimination-based deep learning (Dahl et al., 2011, 2012; Seide et al., 2011; Yu et al., 2010).

p0190 The introduction of the HMM and the related statistical methods to speech recognition in mid-1970s (Baker, 1976; Jelinek, 1976) can be regarded as the most significant paradigm shift in the field, as discussed and analyzed in Baker et al. (2009a,b). One major reason for this early success is the highly efficient EM algorithm (Baum and Petrie, 1966), which we described earlier in this chapter. This maximum likelihood method, often called Baum-Welch algorithm, had been a principal way of training the HMM-based speech recognition systems until 2002, and is still one major step (among many) in training these systems nowadays. It is interesting to note that the Baum-Welch algorithm serves as one major motivating example for the later development of the more general EM algorithm (Dempster et al., 1977). The goal of maximum likelihood or EM method in training GMM-HMM speech recognizers is to minimize the empirical risk with respect to the joint likelihood loss involving a sequence of linguistic labels and a sequence of acoustic data of speech, often extracted at the frame level. In large-vocabulary speech recognition systems, it is normally the case that word-level labels are provided, while state-level labels are latent. Moreover, in training GMM-HMM-based speech recognition systems, parameter tying is often used as a type of regularization. For example, similar acoustic states of the triphones can share the same Gaussian mixture model.

p0195 The use of the generative model of HMMs for representing the (piecewise stationary) dynamic speech pattern and the use of EM algorithm for training the tied HMM parameters constitute one of the most prominent and successful example of generative learning in speech recognition. This success has been firmly established by the speech community, and has been spread widely to machine learning and related communities. In fact, the HMM has become a standard tool not only in speech recognition but also in machine learning as well as their related fields such as bioinformatics and natural language processing. For many machine learning as well as speech recognition researchers, the success of HMMs in speech recognition is a bit surprising due to the well-known weaknesses of the HMM in modeling speech dynamics. The following section is aimed to address ways of using more advanced dynamic generative models and related techniques for speech modeling and recognition.

s0045 2.3.6 HIDDEN DYNAMIC MODELS FOR SPEECH MODELING AND RECOGNITION

p0200 Despite great successes of GMM-HMMs in speech modeling and recognition, their weaknesses, such as the conditional independence and piecewise stationary assumptions, have been well known for speech modeling and recognition applications since early days (Bridle et al., 1998; Deng, 1992, 1993; Deng et al., 1994a; Deng and Sameti, 1996; Deng et al., 2006a; Ostendorf et al., 1996, 1992). Conditional independence refers to the fact the observation probability at time t only depends on the state θ_t and is independent of the preceding states or observations, if θ_t is given.

p0205 Since early 1990s, speech recognition researchers have begun the development of statistical models that capture more realistically the dynamic properties of speech in the temporal dimension than HMMs do. This class of extended HMM models have been variably called stochastic segment model (Ostendorf et al., 1996, 1992), trended or nonstationary-state HMM (Chengalvarayan and Deng, 1998; Deng, 1992; Deng et al., 1994a), trajectory segmental model (Holmes and Russell, 1999; Ostendorf et al., 1996), trajectory HMM (Zen et al., 2004; Zhang and Renals, 2008), stochastic trajectory model (Gong et al., 1996), hidden dynamic model (Bridle et al., 1998; Deng, 1998, 2006; Deng et al., 1997; Ma and Deng, 2000, 2003, 2004; Picone et al., 1999; Russell and Jackson, 2005), buried Markov model (Bilmes, 2003, 2010; Bilmes and Bartels, 2005), structured speech model, and hidden trajectory model (Deng, 2006; Deng and Yu, 2007; Deng et al., 2006a,b; Yu and Deng, 2007; Yu et al., 2006; Zhou et al., 2003), depending on different "prior knowledge" applied to the temporal structure of speech and on various simplifying assumptions to facilitate the model implementation. Common to all these beyond-HMM model variants is some temporal dynamic structure built into the models. Based on the nature of such structure, we can classify these models into two main categories. In the first category are the models focusing on the temporal correlation structure at the "surface" acoustic level. The second category consists of deep hidden or latent dynamics, where the underlying speech production mechanisms are exploited as a prior to represent the temporal structure that accounts for the visible speech pattern. When the mapping from the hidden dynamic layer to the visible layer is limited to be linear and deterministic, then the generative hidden dynamic models in the second category reduce to the first category.

p0210 The temporal span in many of the generative dynamic/trajectory models above is often controlled by a sequence of linguistic labels, which segment the full sentence into multiple regions from left to right; hence the name segment models.

s0050 2.4 DEEP LEARNING AND DEEP NEURAL NETWORKS

s0055 2.4.1 INTRODUCTION

p0215 Deep learning is a set of algorithms in machine learning that attempt to model high-level abstractions in data by using model architectures composed of multiple non-linear transformations. It is part of a broader family of machine learning methods based on learning representations of data. The Deep Neural Network (DNN) is the most important and popular deep learning model, especially for the applications in speech recognition (Deng and Yu, 2014; Yu and Deng, 2014).

p0220 In the long history of speech recognition, both shallow forms and deep forms (e.g., recurrent nets) of artificial neural networks had been explored for many years during 1980s, 1990s and a few years into 2000 (Boulard and Morgan, 1993; Morgan and Bourlard, 1990; Neto et al., 1995; Renals et al., 1994; Waibel et al., 1989). But these methods never won over the GMM-HMM technology based on generative models of speech acoustics that are trained discriminatively (Baker et al., 2009a,b). A number of key difficulties had been methodologically analyzed in 1990s, including gradient diminishing and weak temporal correlation structure in the neural predictive models (Bengio, 1991; Deng et al., 1994b). All these difficulties were in addition to the lack of big training data and big computing power in these early days. Most speech recognition researchers who understood such barriers hence subsequently moved away from neural nets to pursue generative modeling approaches until the recent resurgence of deep learning starting around 2009-2010 that had overcome all these difficulties.

p0225 The use of deep learning for acoustic modeling was introduced during the later part of 2009 by the collaborative work between Microsoft and the University of Toronto, which was subsequently expanded to include IBM and Google

(Hinton et al., 2012; Yu and Deng, 2014). Microsoft and University of Toronto co-organized the 2009 NIPS Workshop on Deep Learning for Speech Recognition (Deng et al., 2009), motivated by the urgency that many versions of deep and dynamic generative models of speech could not deliver what speech industry wanted. It is also motivated by the arrival of a big-compute and big-data era, which would warrant a serious try of the DNN approach. It was then (incorrectly) believed that pre-training of DNNs using generative models of deep belief net (DBN) would be the cure for the main difficulties of neural nets encountered during 1990s. However, soon after the research along this direction started at Microsoft Research, it was discovered that when large amounts of training data are used and especially when DNNs are designed correspondingly with large, context-dependent output layers, dramatic error reduction occurred over the then state-of-the-art GMM-HMM and more advanced generative model-based speech recognition systems without the need for generative DBN pre-training. The finding was verified subsequently by several other major speech recognition research groups. Further, the nature of recognition errors produced by the two types of systems was found to be characteristically different, offering technical insights into how to artfully integrate deep learning into the existing highly efficient, run-time speech decoding system deployed by all major players in speech recognition industry.

p0230 One fundamental principle of deep learning is to do away with hand-crafted feature engineering and to use raw features. This principle was first explored successfully in the architecture of deep autoencoder on the “raw” spectrogram or linear filter-bank features (Deng et al., 2010), showing its superiority over the Mel-Cepstral features which contain a few stages of fixed transformation from spectrograms. The true “raw” features of speech, waveforms, have more recently been shown to produce excellent larger-scale speech recognition results (Tuske et al., 2014).

p0235 Large-scale automatic speech recognition is the first and the most convincing successful case of deep learning in the recent history, embraced by both industry and academic across the board. Between 2010 and 2014, the two major conferences on signal processing and speech recognition, IEEE-ICASSP and Interspeech, have seen near exponential growth in the numbers of accepted papers in their respective annual conferences on the topic of deep learning for speech recognition. More importantly, all major commercial speech recognition systems (e.g., Microsoft Cortana, Xbox, Skype Translator, Google Now, Apple Siri, Baidu and iFlyTek voice search, and a range of Nuance speech products, etc.) nowadays are based on deep learning methods.

p0240 Since the initial successful debut of DNNs for speech recognition around 2009-2011, there has been huge progress made. This progress (as well as future directions) has been summarized into the following eight major areas in Deng and Yu (2014) and Yu and Deng (2014): (1) scaling up/out and speedup DNN training and decoding; (2) sequence discriminative training of DNNs; (3) feature processing by deep models with solid understanding of the underlying mechanisms; (4) adaptation of DNNs and of related deep models; (5) multi-task and transfer learning by DNNs and related deep models; (6) convolution neural networks and how to design them to best exploit domain knowledge of speech; (7) recurrent neural network and its rich long short-term memory (LSTM) variants; (8) other types of deep models including tensor-based models and integrated deep generative/discriminative models.

s0060 2.4.2 A BRIEF HISTORICAL PERSPECTIVE

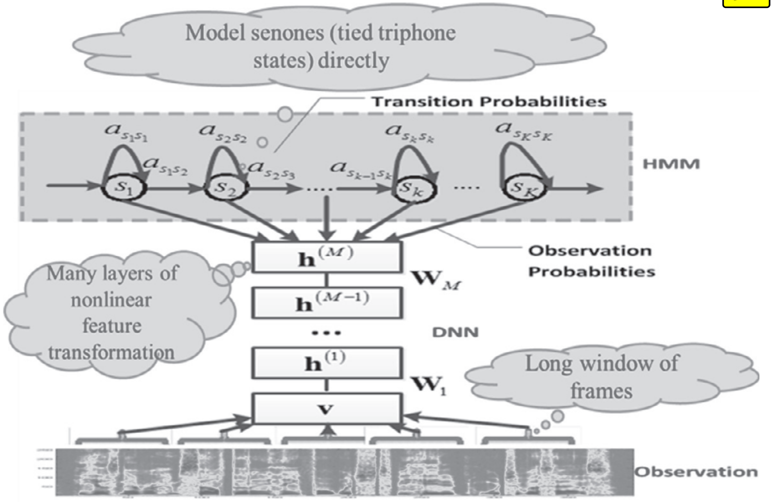
p0245 For many years and until the recent rise of deep learning technology as discussed earlier, speech recognition technology had been dominated by a "shallow" architecture—HMMs with each state characterized by a GMM. While significant technological successes had been achieved using complex and carefully engineered variants of GMM-HMMs and acoustic features suitable for them, researchers had for long anticipated that the next generation of speech recognition would require solutions to many new technical challenges under diversified deployment environments and that overcoming these challenges would likely require *deep* architectures that can at least functionally emulate the human speech recognition system known to have dynamic and hierarchical structure in both speech production and speech perception (Deng, 2006; Deng and O'Shaughnessy, 2003; Divenyi et al., 2006; Stevens, 2000). An attempt to incorporate a primitive level of understanding of this deep speech structure, initiated at the 2009 NIPS Workshop on Deep Learning for Speech Recognition (Deng et al., 2009), has helped create an impetus in the speech recognition community to pursue a deep representation learning approach based on the DNN architecture, which was pioneered by the machine learning community only a few years earlier (Hinton et al., 2006; Hinton and Salakhutdinov, 2006) but rapidly evolved into the new state of the art in speech recognition with industry-wide adoption (Deng et al., 2013b; Hannun et al., 2014; Kingsbury et al., 2012; Mohamed et al., 2009; Sainath et al., 2013a; Seide et al., 2014, 2011; Vanhoucke et al., 2013, 2011; Yu and Deng, 2011; Yu et al., 2010).

p0250 In the remainder of this section, we will describe the DNN and related methods with some technical detail.

s0065 2.4.3 THE BASICS OF DEEP NEURAL NETWORKS

p0255 The most successful version of the DNN in speech recognition is the context-dependent deep neural network hidden Markov model (CD-DNN-HMM) , where the HMM is interfaced with the DNN to handle the dynamic process of speech feature sequences and context-dependent phone units, also known as the senones, are used as the output layer of the DNN. It has been shown by many groups (Dahl et al., 2011, 2012; Deng et al., 2013b; Hinton et al., 2012; Mohamed et al., 2012; Sainath et al., 2013b, 2011; Tuske et al., 2014; Yu et al., 2010), to outperform the conventional GMM-HMMs in many ASR tasks.

p0260 The CD-DNN-HMM is a hybrid system. Three key components of this system are shown in Figure 2.1, which is based on Dahl et al. (2012). First, the CD-DNN-HMM models senones (tied states) directly, which can be as many as tens of thousands



f0005 **FIGURE 2.1**

Illustration of the CD-DNN-HMM and its three core components. AU2

of senones in English, making the output layer of the DNN unprecedentedly large. Second, a deep instead of a shallow multi-layer perceptrons are used. Third, the system takes a long and fixed contextual window of frames as the input. All these three elements of the CD-DNN-HMM have been shown to be critical for achieving the huge accuracy improvement in speech recognition (Dahl et al., 2012; Deng et al., 2013c; Sainath et al., 2011; Yu et al., 2010). Although some conventional shallow neural nets also took a long contextual window as the input, the key to the success of the CD-DNN-HMM is due to a combination of these components. In particular, the deep structure in the DNN allows the system to perform transfer or multi-task learning (Ghoshal et al., 2013; Heigold et al., 2013; Huang et al., 2013), outperforming the shallow models that are unable to carry out transfer learning (Lin et al., 2009; Plahl et al., 2011; Schultz and Waibel, 1998; Yu et al., 2009).

p0265 Further, it is shown in Seltzer et al. (2013) and many other research groups that with the excellent modeling power of the DNN, DNN-based acoustic models can easily match state-of-the-art performance on the Aurora 4 task (Parihar and Picone, 2002), which is a standard noise-robustness large-vocabulary speech recognition task, without any explicit noise compensation. The CD-DNN-HMM is expected to make further progress on noise-robust ASR due to the DNN's ability to handle heterogeneous data (Li et al., 2012; Seltzer et al., 2013). Although the CD-DNN-HMM is a modeling technology, its layer-by-layer setup provides a feature extraction strategy that automatically derives powerful noise-resistant features from primitive raw data for senone classification.

p0270 From the architecture point of view, a DNN can be considered as a conventional multi-layer perceptron (MLP) with many hidden layers (thus deep) as illustrated in Figure 2.1, in which the input and output of the DNN are denoted as **x** and **o**, respectively. Let us denote the input vector at layer l as $\mathbf{v}^l$ (with $\mathbf{v}^0 = \mathbf{x}$), the weight matrix as $\mathbf{A}^l$, and bias vector as $\mathbf{b}^l$. Then, for a DNN with L hidden layers, the output of the lth hidden layer can be written as

$$\mathbf{v}^{l+1} = \sigma(z(\mathbf{v}^l)), \quad 0 \le l < L, \tag{2.29}$$

where

$$\mathbf{u}^l = z(\mathbf{v}^l) = \mathbf{A}^l\mathbf{v}^l + \mathbf{b}^l \tag{2.30}$$

and

$$\sigma(\mathbf{x}) = 1/(1 + e^{\mathbf{x}}) \tag{2.31}$$

is the sigmoid function applied element-wise. The posterior probability is

$$P(o = s|\mathbf{x}) = \text{softmax}(z(\mathbf{v}^L)), \tag{2.32}$$

where s belongs to the set of senones (also known as the tied triphone states) . We compute the HMM's state emission probability density function $p(\mathbf{x}|o = s)$ by converting the state posterior probability $P(o = s|\mathbf{x})$ to

$$p(\mathbf{x}|o = s) = \frac{P(o = s|\mathbf{x})}{P(o = s)} p(\mathbf{x}), \tag{2.33}$$

where $P(o = s)$ is the prior probability of state s, and $p(\mathbf{x})$ is independent of state and can be dropped during evaluation.

p0275 Although recent studies (Senior et al., 2014; Zhang and Woodland, 2014) started the DNN training from scratch without using GMM-HMM systems, in most implementations the CD-DNN-HMM inherits the model structure, especially in the output layer including the phone set, the HMM topology, and senones, directly from the GMM-HMM system. The senone labels used to train the DNNs are extracted from the forced alignment generated by the GMM-HMM. The training criterion to be minimized is the cross entropy between the posterior distribution represented by the reference labels and the predicted distribution:

$$F_{\text{CE}} = -\sum_t \sum_{s=1}^{N} P_{\text{target}}(o = s|\mathbf{x}_t) logP(o = s|\mathbf{x}_t), \tag{2.34}$$

where N is the number of senones, $P_{\text{target}}(o = s|\mathbf{x}_t)$ is the target probability of senone s at time t, and $P(o = s|\mathbf{x}_t)$ is the DNN output probability calculated from Equation 2.32.

p0280 In the standard CE training of DNN, the target probabilities of all senones at time t are formed as a one-hot vector, with only the dimension corresponding to the reference senone assigned a value of 1 and the rest as 0. As a result, Equation 2.34 is reduced to minimize the negative log likelihood because every frame has only one target label s_t:

$$F_{\text{CE}} = -\sum_t \log P(o = s_t|\mathbf{x}_t). \tag{2.35}$$

p0285 This objective function is minimized by using error back propagation (Rumelhart et al., 1988) which is a gradient-descent based optimization method developed for neural networks. The weight matrix W and bias b of layer l are updated with:

$$\hat{\mathbf{A}}^l = \mathbf{A}^l + \alpha \mathbf{v}^l (\mathbf{e}^l)^T, \tag{2.36}$$

$$\hat{\mathbf{b}}^l = \mathbf{b}^l + \alpha \mathbf{e}^l, \tag{2.37}$$

where α is the learning rate. $\mathbf{v}^l$ and $\mathbf{e}^l$ are the input and error vector of layer l, respectively. $\mathbf{e}^l$ is calculated by back propagating the error signal from its upper layer with

$$e_i^l = \left[\sum_{k=1}^{N_{l+1}} A_{ik}^{l+1} e_k^{l+1}\right] \sigma'(u_i^l), \tag{2.38}$$

where A_{ik}^{l+1} is the element of weighting matrix $\mathbf{A}^{l+1}$ in the ith row and kth column for layer $l+1$, and e_k^{l+1} is the kth element of error vector $\mathbf{e}^{l+1}$ for layer $l+1$. N_{l+1} is the number of units in layer $l+1$. $\sigma'(u_i^l)$ is the derivative of sigmoid function. The error signal of the top layer (i.e., output layer) is defined as:

$$e_s^L = -\sum_t (\delta_{ss_t} - P(o = s|\mathbf{x}_t)), \tag{2.39}$$

where δ_{ss_t} is the Kronecker delta function. Then the parameters of the DNN can be efficiently updated with the back propagation algorithm.

p0290 Speech recognition is inherently a sequence classification problem. Therefore, the frame-based cross-entropy criterion is not optimal. The sequence training criterion has been explored to optimize DNN parameters for speech recognition. As the GMM parameter optimization with sequence training criterion, MMI, BMMI, and MBR criteria are typically used (Kingsbury, 2009; Mohamed et al., 2010; Su et al., 2013; Veselỳ et al., 2013). For example, the MMI objective function is

$$F_{\text{MMI}} = \sum_r \log P(\hat{\mathbf{S}}^r|\mathbf{X}^r), \tag{2.40}$$

where $\hat{\mathbf{S}}^r$ and $\mathbf{X}^r$ are the reference string and the observation sequence for rth utterances. Generally, $P(\mathbf{S}|\mathbf{X})$ is the posterior of path $\mathbf{S}$ given the current model:

$$P(\mathbf{S}|\mathbf{X}) = \frac{p^k(\mathbf{X}|\mathbf{S})P(\mathbf{S})}{\sum_{\mathbf{S}'} p^k(\mathbf{X}|\mathbf{S}')P(\mathbf{S}')} \tag{2.41}$$

$P(\mathbf{X}|\mathbf{S})$ is the acoustic score of the whole utterance, $P(\mathbf{S})$ is the language model score, and k is the acoustic weight. Then the error signal of MMI criterion for utterance r becomes

$$e_s^{(r,L)} = -k \sum_t \left(\delta_{s\hat{s}_t} - \sum_{\mathbf{S}^r} \delta_{ss_t} P(\mathbf{S}^r|\mathbf{X}^r) \right). \tag{2.42}$$

p0295 There are different strategies to update the DNN parameters. The batch gradient descent updates the parameters with the gradient only once after each sweep through the whole training set and in this way parallelization can be easily conducted. However, the convergence of batch update is very slow and stochastic gradient descent (SGD) (Zhang, 2004) usually works better in practice where the true gradient is approximated by the gradient at a single frame and the parameters are updated right after seeing each frame. The compromise between the two, the mini-batch SGD (Dekel et al., 2012), is more widely used, as the reasonable size of mini-batches makes all the matrices fit into GPU memory, which leads to a more computationally efficient learning process. Recent advances in Hessian-free optimization (Martens, 2010) have also partially overcome this difficulty using approximated second-order information or stochastic curvature estimates. This second-order batch optimization method has also been explored to optimize the weight parameters in DNNs (Kingsbury et al., 2012; Wiesler et al., 2013).

p0300 Decoding of the CD-DNN-HMM is carried out by plugging the DNN into a conventional large vocabulary HMM decoder with the senone likelihood evaluated with Equation 2.33. This strategy was initially explored and established in Yu et al. (2010) and Dahl et al. (2011), and has soon become the standard industry practice because it allows the speech recognition industry to re-use much of the decoder software infrastructure built for the GMM-HMM system over many years.

s0070 2.4.4 ALTERNATIVE DEEP LEARNING ARCHITECTURES

p0305 In addition to the standard architecture of the DNN, there are plenty of studies of applying alternative nonlinear units and structures to speech recognition. Although sigmoid and tanh functions are the most commonly used nonlinearity types in DNNs, their limitations are well known. For example, it is slow to learn the whole network due to weak gradients when the units are close to saturation in both directions. Therefore, rectified linear units (ReLU) (Dahl et al., 2013; Jaitly and Hinton, 2011; Zeiler et al., 2013) and maxout units (Cai et al., 2013; Miao et al., 2013; Swietojanski et al., 2014) are applied to speech recognition to overcome the weakness of the sigmoidal units. ReLU refers to the units in a neural network that use the activation function of $f(x) = \max(0, x)$. Maxout refers to the units that use the activation function of getting the maximum output value from a group of input values.

s0075

Deep convolutional neural networks

p0310 The CNN, originally developed for image processing, can also be robust to distortion due to its invariance property (Abdel-Hamid et al., 2013, 2012, 2014; Sainath et al., 2013a). Figure 2.2 (after (Abdel-Hamid et al., 2014)) shows the structure of one CNN with full weight sharing. The first layer is called a convolution layer which consists a number of feature maps. Each neuron in the convolution layer receives inputs from a local receptive field representing features of a limited frequency range. Neurons that belong to the same feature map share the same weights (also called filters or kernels) but receive different inputs shifted in frequency. As a result, the convolution layer carries out the convolution operation on the kernels with its lower layer activations. A pooling layer is added on top of the convolution layer to compute a lower resolution representation of the convolution layer activations through sub-sampling. The pooling function, which computes some statistics of the activations, is typically applied to the neurons along a window of frequency bands and generated from the same feature map in the convolution layer. The most popular pooling function is the maxout function. Then a fully connected DNN is built on top of the CNN to do the work of senone classification.

p0315 It is important to point out that the invariant property of the CNN to frequency shift applies when filter-bank features are used and it does not apply with the cepstral

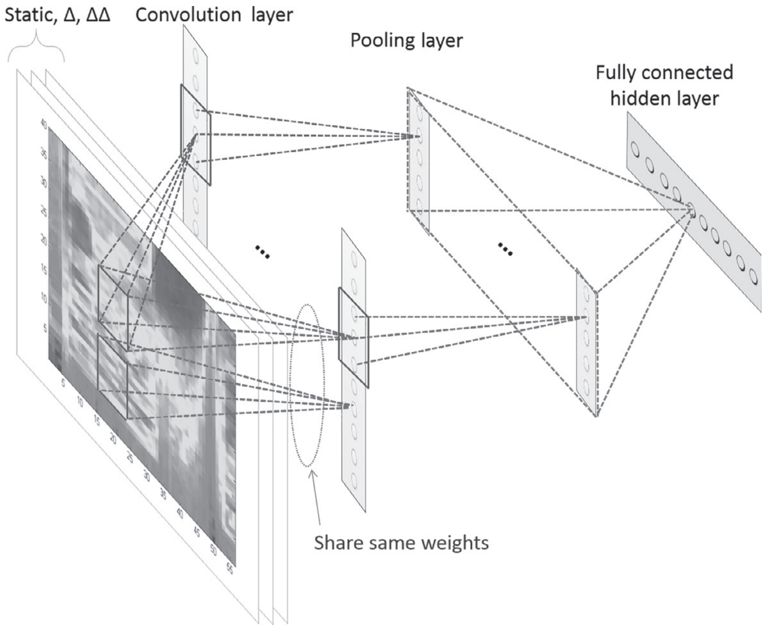


FIGURE 2.2

f0010 Illustration of the CNN in which the convolution is applied along frequency bands.

feature; see a detailed analysis in Deng et al. (2013a). Indeed using filter bank features as the input open a door for the CNN to exploit the structure in the features. It was shown that by using a CNN along the frequency axis they can normalize speaker differences and further reduce the phone error rate from 20.7% to 20.0% on the TIMIT phone recognition task (Abdel-Hamid et al., 2012). These results were later extended to large vocabulary speech recognition in 2013 with improved CNN architectures, pretraining techniques, and pooling strategies (Abdel-Hamid et al., 2013, 2014; Deng et al., 2013a; Sainath et al., 2013a,c). Further studies showed that the CNN helps mostly for the tasks in which the training set size or data variability is relatively small (Huang et al., 2015). For most other tasks the relative word error rate reduction is often in the range of 2-3%.

s0080 *Deep recurrent neural networks*

p0320 A more popular and effective deep learning architecture than the CNN in the recent speech recognition literature is a version of the recurrent neural network (RNN) which stacks one on another and which often contains a special cell structure called long short-term memory (LSTM). RNNs and LSTMs have been reported to work well specifically for robust speech recognition due to its powerful context modeling (Maas et al., 2012; Weng et al., 2014; Wöllmer et al., 2013a,b).

p0325 Here we briefly discuss the basics of the RNN as a class of neural network models where many connections among its neurons form a directed cycle. This gives rise to the structure of internal states or memory in the RNN, endowing it with the dynamic temporal behavior not exhibited by the basic DNN discussed earlier in this chapter.

p0330 An RNN is fundamentally different from the feed-forward DNN in that the RNN operates not only based on inputs, as for the DNN, but also on internal states. The internal states encode the past information in the temporal sequence that has already been processed by the RNN. In this sense, the RNN is a dynamic system, more general than the DNN that performs memoryless input-output transformations. The use of the state space in the RNN enables its representation to learn sequentially extended dependencies over a long time span, at least in principle.

p0335 Let us now formulate the simple one-hidden-layer RNN in terms of the (noise-free) nonlinear state space model commonly used in signal processing. At each time point t, let $\mathbf{x}_t$ be the $K \times 1$ vector of inputs, $\mathbf{h}_t$ be the $N \times 1$ vector of hidden state values, and $\mathbf{y}_t$ be the $L \times 1$ vector of outputs, the simple one-hidden-layer RNN can be described as

$$\mathbf{h}_t = f(\mathbf{W}_{xh}\mathbf{x}_t + \mathbf{W}_{hh}\mathbf{h}_{t-1}), \tag{2.43}$$

$$\mathbf{y}_t = g(\mathbf{W}_{hy}\mathbf{h}_t), \tag{2.44}$$

where $\mathbf{W}_{hy}$ is the $L \times N$ matrix of weights connecting the N hidden units to the L outputs, $\mathbf{W}_{xh}$ is the $N \times K$ matrix of weights connecting the K inputs to the N hidden units, and $\mathbf{W}_{hh}$ is the $N \times N$ matrix of weights connecting the N hidden units from time $t-1$ to time t, $\mathbf{u}_t = \mathbf{W}_{xh}\mathbf{x}_t + \mathbf{W}_{hh}\mathbf{h}_{t-1}$ is the $N \times 1$ vector of hidden layer potentials, $\mathbf{v}_t = \mathbf{W}_{hy}\mathbf{h}_t$ is the $L \times 1$ vector of output layer potentials, $f(\mathbf{u}_t)$ is the hidden layer activation function, and $g(\mathbf{v}_t)$ is the output layer activation function. Typical hidden layer activation functions are Sigmoid, tanh, and rectified linear units while the typical output layer activation functions are linear and softmax functions. Equations 2.43 and 2.44 are often called the observation and state equations, respectively. Note that, outputs from previous time frames can also be used to update the state vector, in which case the state equation becomes

$$\mathbf{h}_t = f(\mathbf{W}_{xh}\mathbf{x}_t + \mathbf{W}_{hh}\mathbf{h}_{t-1} + \mathbf{W}_{\mathrm{yh}}\mathbf{y}_{\mathrm{t\text{-}1}}), \tag{2.45}$$

where $\mathbf{W}_{yh}$ denotes the weight matrix connecting from output layer to the hidden layer. For simplicity purposes, most RNNs in speech recognition do not include output feedback.

p0340 It is important to note that before the recent rise of deep learning for speech modeling and recognition, a number of earlier attempts had been made to develop computational architectures that are "deeper" than the conventional GMM-HMM architecture. One prominent class of such models are hidden dynamic models where the internal representation of dynamic speech features is generated probabilistically from the higher levels in the overall deep speech model hierarchy (Bridle et al., 1998; Deng et al., 1997, 2006b; Ma and Deng, 2000; Togneri and Deng, 2003). Despite separate developments of the RNNs and of the hidden dynamic or trajectory models, they share a very similar motivation—representing aspects of dynamic structure in human speech. Nevertheless, a number of different ways in which these two types of deep dynamic models are constructed endow them with distinct pros and cons. Careful analysis of the contrast between these two model types and of the similarity to each other will help provide insights into the strategies for developing new types of deep dynamic models with the hidden representations of speech features superior to both existing RNNs and hidden dynamic models. This type of multi-faceted analysis has been provided in the recent book (Yu and Deng, 2014), which we refer the readers to.

p0345 While the RNN as well as the related nonlinear neural predictive models saw its early success in small ASR tasks (Deng et al., 1994b; Robinson, 1994), it was not easy to duplicate due to the intricacy in training, let alone to scale them up for larger speech recognition tasks. Learning algorithms for the RNN have been dramatically improved since these early days, however, and much stronger and practical results have been obtained recently using the RNN, especially when the bidirectional LSTM architecture is exploited (Graves et al., 2013a,b) or when the high-level DNN features are used as inputs to the RNN (Chen and Deng, 2014; Deng and Chen, 2014; Deng and Platt, 2014; Hannun et al., 2014). The LSTM was reported to give the lowest PER on the benchmark TIMIT phone recognition task in 2013 by Grave et al. (Graves et al., 2013a,b). In 2014, researchers published the results using the LSTM on large-scale tasks with applications to Google Now, voice search, and mobile dictation with excellent accuracy results (Sak et al., 2014a,b). To reduce the model size, the otherwise very large output vectors of LSTM units are linearly projected to smaller-dimensional vectors. Asynchronous stochastic gradient descent (ASGD) algorithm with truncated backpropagation through time (BPTT) is performed across hundreds of machines in CPU clusters. The best accuracy is obtained by optimizing the frame-level cross-entropy objective function followed by sequence discriminative training. With one LSTM stacking on top of another, this deep and recurrent LSTM model produced 9.7% WER on a large voice search task trained with 3 million utterances. This result is better than 10.7% WER achieved with frame-level cross entropy training criterion alone. It is also significantly better than the 10.4% WER obtained with the best DNN-HMM system using rectified linear units. Furthermore, this better accuracy is achieved while the total number of parameters is drastically reduced from 85 millions in the DNN system to 13 millions in the LSTM system. Some recent publications also showed that deep LSTMs are effective in speech recognition in reverberant multisource acoustic environments, as indicated by the strong results achieved by LSTMs in a recent ChiME Challenge task involving speech recognition in such difficult environments (Weninger et al., 2014).

s0085 2.5 SUMMARY

p0350 In this chapter, two major classes of acoustic models used for speech recognition are reviewed. In the first, generative-model class, we have the HMM where GMMs are used as the statistical distribution of speech features that is associated with each HMM state. This class also includes the hidden dynamic model that generalizes the GMM-HMM by incorporating aspects of deep structure in the human speech production process used as the internal representation of speech features.

p0355 Much of the robust speech recognition studies in the past, to be surveyed and analyzed at length in the following chapters of this book, have been carried out based on the generative models of speech, especially the GMM-HMM model, as reviewed in this chapter. One important advantage of generative models of speech in robust speech recognition is the straightforward way of thinking about the noise-robustness problem: noisy speech as the observations for robust speech recognizers can be viewed as the outcome of a further generative process combining clean speech and noise signals (Deng et al., 2000; Deng and Li, 2013; Frey et al., 2001b; Gales, 1995; Li et al., 2007a) using well-established distortion models. Some commonly used distortion models will be covered in later part of this book including the major publications (Deng, 2011; Li et al., 2014, 2008). Practical embodiment of such straightforward thinking to achieve noise-robust speech recognition systems based on generative acoustic models of distorted speech will also be presented (Huang et al., 2001a).

p0360 In the second, discriminative-model class of acoustic modeling for speech, we have the more recent DNN model as well as its convolutional and recurrent variants. These deep discriminative models have been shown to significantly outperform all previous versions of generative models of speech, shallow or deep, in speech recognition accuracy. The main sub-classes of these deep discriminative models are reviewed in some detail in this chapter. How to handle noise robustness within the framework of discriminative deep learning models of speech acoustics, which is less straightforward and more recent than the generative models of speech, will form the bulk part of later chapters.

REFERENCES

Abdel-Hamid, O., Deng, L., Yu, D., 2013. Exploring convolutional neural network structures and optimization techniques for speech recognition. In: Proc. Interspeech, pp. 3366-3370.

Abdel-Hamid, O., Mohamed, A., Jiang, H., Penn, G., 2012. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Abdel-Hamid, O., Mohamed, A.r., Jiang, H., Deng, L., Penn, G., Yu, D., 2014. Convolutional neural networks for speech recognition. IEEE Trans. Audio Speech Lang. Process. AU3

Bahl, L.R., Brown, P.F., Souza, P.V.D., Mercer, R.L., 1997. Maximum mutual information estimation of hidden Markov model parameters for speech recognition. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 11, pp. 49-52.

Baker, J., 1976. Stochastic modeling for automatic speech recognition. In: Reddy, D. (Ed.), Speech Recognition. Academic, New York.

Baker, J., Deng, L., Glass, J., Khudanpur, S., Lee, C.H., Morgan, N., et al., 2009a. Research developments and directions in speech recognition and understanding, Part I. IEEE Signal Process. Mag. 26 (3), 75-80.

Baker, J., Deng, L., Glass, J., Khudanpur, S., Lee, C.H., Morgan, N., et al., 2009b. Updated MINDS report on speech recognition and understanding. IEEE Signal Process. Mag. 26 (4), 78-85.

Baum, L., Petrie, T., 1966. Statistical inference for probabilistic functions of finite state Markov chains. Ann. Math. Statist. 37 (6), 1554-1563.

Bengio, Y., 1991. Artificial Neural Networks and their Application to Sequence Recognition. McGill University, Montreal, Canada.

Bilmes, J., 2003. Buried markov models: A graphical modeling approach to automatic speech recognition. Comput. Speech Lang. 17, 213-231.

Bilmes, J., 2006. What HMMs can do. IEICE Trans. Informat. Syst. E89-D (3), 869-891.

Bilmes, J., 2010. Dynamic graphical models. IEEE Signal Process. Mag. 33, 29-42.

Bilmes, J., Bartels, C., 2005. Graphical model architectures for speech recognition. IEEE Signal Process. Mag. 22, 89-100.

Boulard, H., Morgan, N., 1993. Continuous speech recognition by connectionist statistical methods. IEEE Trans. Neural Networks 4 (6), 893-909.

Bridle, J., Deng, L., Picone, J., Richards, H., Ma, J., Kamm, T., et al., 1998. An investigation of segmental hidden dynamic models of speech coarticulation for automatic speech recognition. Final Report for 1998 Workshop on Language Engineering, CLSP, Johns Hopkins.

Cai, M., Shi, Y., Liu, J., 2013. Deep maxout neural networks for speech recognition. In: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 291-296.

Chen, J., Deng, L., 2014. A primal-dual method for training recurrent neural networks constrained by the echo-state property. In: Proc. ICLR.

Chengalvarayan, R., Deng, L., 1998. Speech trajectory discrimination using the minimum classification error learning. IEEE Trans. Speech Audio Process. (6), 505-515.

Dahl, G., Sainath, T., Hinton, G., 2013. Improving deep neural networks for LVCSR using rectified linear units and dropout. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8609-8613.

Dahl, G., Yu, D., Deng, L., Acero, A., 2011. Large vocabulary continuous speech recognition with context-dependent DBN-HMMs. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Dahl, G.E., Yu, D., Deng, L., Acero, A., 2012. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio Speech Lang. Process. 20 (1), 30-42.

Dekel, O., Gilad-Bachrach, R., Shamir, O., Xiao, L., 2012. Optimal distributed online prediction using mini-batches. J. Mach. Learn. Res. 13 (1), 165-202.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. 39 (1), 1-38.

Deng, L., 1992. A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal. Signal Process. 27 (1), 65-78.

Deng, L., 1993. A stochastic model of speech incorporating hierarchical nonstationarity. IEEE Trans. Acoust. Speech Signal Process. 1 (4), 471-475.

Deng, L., 1998. A dynamic, feature-based approach to the interface between phonology and phonetics for speech modeling and recognition. Speech Commun. 24 (4), 299-323.

Deng, L., 1999. Computational models for speech production. In: Computational Models of Speech Pattern Processing. Springer-Verlag, New York, pp. 199-213.

Deng, L., 2003. Switching dynamic system models for speech articulation and acoustics. In: Mathematical Foundations of Speech and Language Processing. Springer-Verlag, New York, pp. 115-134.

Deng, L., 2006. Dynamic Speech Models—Theory, Algorithm, and Applications. Morgan and Claypool, San Rafael, CA.

Deng, L., 2011. Front-end, back-end, and hybrid techniques for noise-robust speech recognition. In: Robust Speech Recognition of Uncertain or Missing Data: Theory and Application. Springer, New York, pp. 67-99.

Deng, L., Abdel-Hamid, O., Yu, D., 2013.a. A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, Canada.

Deng, L., Acero, A., Plumpe, M., Huang, X., 2000. Large vocabulary speech recognition under adverse acoustic environment. In: Proc. International Conference on Spoken Language Processing (ICSLP), vol. 3, pp. 806-809.

Deng, L., Aksmanovic, M., Sun, D., Wu, J., 1994a. Speech recognition using hidden Markov models with polynomial regression functions as non-stationary states. IEEE Trans. Acoust. Speech Signal Process. 2 (4), 101-119.

Deng, L., Chen, J., 2014. Sequence classification using high-level features extracted from deep neural networks. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Deng, L., Droppo, J., A.Acero., 2003. Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition. IEEE Trans. Speech Audio Process. 11, 568-580.

Deng, L., Droppo, J., Acero, A., 2002.a. A Bayesian approach to speech feature enhancement using the dynamic cepstral prior. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. I-829-I-832.

Deng, L., Erler, K., 1992. Structural design of hidden Markov model speech recognizer using multivalued phonetic features: comparison with segmental speech units. J. Acoust. Soc. Amer. 92 (6), 3058-3067.

Deng, L., Hassanein, K., Elmasry, M., 1994b. Analysis of correlation structure for a neural predictive model with applications to speech recognition. Neural Networks 7, 331-339.

Deng, L., Hinton, G., Kingsbury, B., 2013.b. New types of deep neural network learning for speech recognition and related applications: An overview. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, Canada.

Deng, L., Hinton, G., Yu, D., 2009. Deep learning for speech recognition and related applications. In: NIPS Workshop, Whistler, Canada.

Deng, L., Kenny, P., Lennig, M., Gupta, V., Seitz, F., Mermelsten, P., 1991a. Phonemic hidden Markov models with continuous mixture output densities for large vocabulary word recognition. IEEE Trans. Acoust. Speech Signal Process. 39 (7), 1677-1681.

Deng, L., Lennig, M., Seitz, F., Mermelstein, P., 1991b. Large vocabulary word recognition using context-dependent allophonic hidden Markov models. Comput. Speech Lang. 4, 345-357.

Deng, L., Li, J., Huang, J.T., Yao, K., Yu, D., Seide, F., et al., 2013c. Recent advances in deep learning for speech research at Microsoft. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, Canada.

Deng, L., Li, X., 2013. Machine learning paradigms in speech recognition: An overview. IEEE Trans. Audio Speech Lang. Process. 21 (5), 1060-1089.

Deng, L., O'Shaughnessy, D., 2003. Speech Processing—A Dynamic and Optimization-Oriented Approach. Marcel Dekker Inc., New York.

Deng, L., Platt, J., 2014. Ensemble deep learning for speech recognition. In: Proc. Annual Conference of International Speech Communication Association (INTERSPEECH).

Deng, L., Ramsay, G., Sun, D., 1997. Production models as a structural basis for automatic speech recognition. Speech Commun. 33 (2-3), 93-111.

Deng, L., Sameti, H., 1996. Transitional speech units and their representation by regressive Markov states: Applications to speech recognition. IEEE Trans. Speech Audio Process. 4 (4), 301-306.

Deng, L., Seltzer, M., Yu, D., Acero, A., Mohamed, A., Hinton, G., 2010. Binary coding of speech spectrograms using a deep auto-encoder. In: Proc. Annual Conference of International Speech Communication Association (INTERSPEECH).

Deng, L., Wang, K., A. Acero, H.H., Huang, X., 2002b. Distributed speech processing in MiPad's multimodal user interface. IEEE Trans. Audio Speech Lang. Process. 10 (8), 605-619.

Deng, L., Yu, D., 2007. Use of differential cepstra as acoustic features in hidden trajectory modelling for phonetic recognition. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 445-448.

Deng, L., Yu, D., 2014. Deep Learning: Methods and Applications. NOW Publishers, Hanover, MA.

Deng, L., Yu, D., Acero, A., 2006a. A bidirectional target filtering model of speech coarticulation: two-stage implementation for phonetic recognition. IEEE Trans. Speech Audio Process. 14, 256-265.

Deng, L., Yu, D., Acero, A., 2006b. Structured speech modeling. IEEE Trans. Speech Audio Process. 14, 1492-1504.

Divenyi, P., Greenberg, S., Meyer, G., 2006. Dynamics of Speech Production and Perception. IOS Press, Santa Venetia, CA.

Frey, B., Deng, L., Acero, A., Kristjansson, T., 2001a. ALGONQUIN: iterating Laplace's method to remove multiple types of acoustic distortion for robust speech recognition. In: Proc. Interspeech, pp. 901-904.

Frey, B., Kristjansson, T., Deng, L., Acero, A., 2001b. ALGONQUIN—learning dynamic noise models from noisy speech for robust speech recognition. In: NIPS, pp. 1165-1171.

Gales, M.J.F., 1995. Model-based techniques for noise robust speech recognition. Ph.D. thesis, University of Cambridge.

Ghoshal, A., Swietojanski, P., Renals, S., 2013. Multilingual training of deep-neural networks. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Gibson, M., Hain, T., 2006. Hypothesis spaces for minimum Bayes risk training in large vocabulary speech recognition. In: Proc. Interspeech.

Gong, Y., Illina, I., Haton, J.P., 1996. Modeling long term variability information in mixture stochastic trajectory framework. In: Proc. International Conference on Spoken Language Processing (ICSLP).

Graves, A., Jaitly, N., Mahamed, A., 2013a. Hybrid speech recognition with deep bidirectional LSTM. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, Canada.

Graves, A., Mahamed, A., Hinton, G., 2013b. Speech recognition with deep recurrent neural networks. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, Canada.

Hannun, A.Y., Case, C., Casper, J., Catanzaro, B.C., Diamos, G., Elsen, E., et al., 2014. Deep speech: Scaling up end-to-end speech recognition. CoRR abs/1412.5567. URL http://arxiv.org/abs/1412.5567.

He, X., Deng, L., 2008. DISCRIMINATIVE LEARNING FOR SPEECH RECOGNITION: Theory and Practice. Morgan and Claypool, San Rafael, CA.

He, X., Deng, L., Chou, W., 2008. Discriminative learning in sequential pattern recognition—A unifying review for optimization-oriented speech recognition. IEEE Signal Process. Mag. 25 (5), 14-36.

Heigold, G., Vanhoucke, V., Senior, A., Nguyen, P., Ranzato, M., Devin, M., et al., 2013. Multilingual acoustic models using distributed deep neural networks. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Hinton, G., Deng, L., Yu, D., et al., G.D., 2012. Deep neural networks for acoustic modeling in speech recognition. IEEE Sig. Proc. Mag. 29 (6), 82-97.

Hinton, G., Osindero, S., Teh, Y., 2006. A fast learning algorithm for deep belief nets. Neural Comput. 18, 1527-1554.

Hinton, G., Salakhutdinov, R., 2006. Reducing the dimensionality of data with neural networks. Science 313 (5786), 504-507.

Holmes, W., Russell, M., 1999. Probabilistic-trajectory segmental HMMs. Comput. Speech Lang. 13, 3-37.

Huang, J.T., Li, J., Gong, Y., 2015. An analysis of convolutional neural networks for speech recognition. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Huang, J.T., Li, J., Yu, D., Deng, L., Gong, Y., 2013. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Huang, X., Acero, A., Chelba, C., Deng, L., Droppo, J., Duchene, D., et al., 2001a. Mipad: a multimodal interaction prototype. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Huang, X., Acero, A., Hon, H.W., 2001b. Spoken Language Processing, vol. 18. Prentice Hall, Englewood Cliffs, NJ.

Huang, X., Jack, M., 1989. Semi-continuous hidden Markov models for speech signals. Comput. Speech Lang. 3 (3), 239-251.

Jaitly, N., Hinton, G., 2011. Learning a better representation of speech soundwaves using restricted Boltzmann machines. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5884-5887.

Jelinek, F., 1976. Continuous speech recognition by statistical methods. Proc. IEEE 64 (4), 532-557.

Jiang, H., Li, X., Liu, C., 2006. Large margin hidden Markov models for speech recognition. IEEE Trans. Audio Speech Lang. Process. 14 (5), 1584-1595.

Juang, B.H., 1985. Maximum-likelihood estimation for mixture multivariate stochastic observations of Markov chains. AT&T Tech. J. 64 (6), 1235-1249.

Juang, B.H., Hou, W., Lee, C.H., 1997. Minimum classification error rate methods for speech recognition. IEEE Trans. Speech Audio Process. 5 (3), 257-265.

Juang, B.H., Levinson, S.E., Sondhi, M.M., 1986. Maximum likelihood estimation for mixture multivariate stochastic observations of Markov chains. IEEE Trans. Informat. Theory 32 (2), 307-309.

Kingsbury, B., 2009. Lattice-based optimization of sequence classification criteria for neural-network acoustic modeling. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3761-3764.

Kingsbury, B., Sainath, T., Soltau, H., 2012. Scalable minimum Bayes risk training of deep neural network acoustic models using distributed Hessian-free optimization. In: Proc. Interspeech.

Lee, L.J., Fieguth, P., Deng, L., 2001. A functional articulatory dynamic model for speech production. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 2 , Salt Lake City, pp. 797-800.

Levinson, S., Rabiner, L., Sondhi, M., 1983. An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. Bell Syst. Tech. J. 62 (4), 1035-1074.

Li, J., Deng, L., Gong, Y., Haeb-Umbach, R., 2014. An overview of noise-robust automatic speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. 22 (4), 745-777.

Li, J., Deng, L., Yu, D., Gong, Y., Acero, A., 2007a. High-performance HMM adaptation with joint compensation of additive and convolutive distortions via vector Taylor series. In: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 65-70.

Li, J., Deng, L., Yu, D., Gong, Y., Acero, A., 2008. HMM adaptation using a phase-sensitive acoustic distortion model for environment-robust speech recognition. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4069-4072.

Li, J., Yu, D., Huang, J.T., Gong, Y., 2012. Improving wideband speech recognition using mixed-bandwidth training data in CD-DNN-HMM. In: Proc. IEEE Spoken Language Technology Workshop, pp. 131-136.

Li, J., Yuan, M., Lee, C.H., 2006. Soft margin estimation of hidden Markov model parameters. In: Proc. Interspeech, pp. 2422-2425.

Li, J., Yuan, M., Lee, C.H., 2007b. Approximate test risk bound minimization through soft margin estimation. IEEE Trans. Audio Speech Lang. Process. 15 (8), 2393-2404.

Li, X., Jiang, H., 2007. Solving large-margin hidden Markov model estimation via semidefinite programming. IEEE Trans. Audio Speech Lang. Process. 15 (8), 2383-2392.

Lin, H., Deng, L., Yu, D., Gong, Y., Acero, A., Lee, C.H., 2009. A study on multilingual acoustic modeling for large vocabulary ASR. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4333-4336.

Liporace, L., 1982. Maximum likelihood estimation for multivariate observations of Markov sources. IEEE Trans. Informat. Theory 28 (5), 729-734.

Ma, J., Deng, L., 2000. A path-stack algorithm for optimizing dynamic regimes in a statistical hidden dynamic model of speech. Comput. Speech Lang. 14, 101-104.

Ma, J., Deng, L., 2003. Efficient decoding strategies for conversational speech recognition using a constrained nonlinear state-space model. IEEE Trans. Audio Speech Process. 11 (6), 590-602.

Ma, J., Deng, L., 2004. Target-directed mixture dynamic models for spontaneous speech recognition. IEEE Trans. Audio Speech Process. 12 (1), 47-58.

Maas, A.L., Le, Q.V., O'Neil, T.M., Vinyals, O., Nguyen, P., Ng, A.Y., 2012. Recurrent neural networks for noise reduction in robust ASR. In: Proc. Interspeech, pp. 22-25.

Martens, J., 2010. Deep learning via Hessian-free optimization. In: Proceedings of the 27th International Conference on Machine Learning, pp. 735-742.

Miao, Y., Metze, F., Rawat, S., 2013. Deep maxout networks for low-resource speech recognition. In: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 398-403.

Mohamed, A., Dahl, G., Hinton, G., 2009. Deep belief networks for phone recognition. In: NIPS Workshop on Deep Learning for Speech Recognition and Related Applications.

Mohamed, A., Dahl, G.E., Hinton, G., 2012. Acoustic modeling using deep belief networks. IEEE Trans. Audio Speech Lang. Process. 20 (1), 14-22.

Mohamed, A., Yu, D., Deng, L., 2010. Investigation of full-sequence training of deep belief networks for speech recognition. In: Proc. Annual Conference of International Speech Communication Association (INTERSPEECH).

Morgan, N., Bourlard, H., 1990. Continuous speech recognition using multilayer perceptrons with hidden Markov models. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 413-416.

Neto, J., Almeida, L., Hochberg, M., Martins, C., Nunes, L., Renals, S., et al., 1995. Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system. In: Proc. European Conference on Speech Communication and Technology (EUROSPEECH), pp. 2171-2174.

Ostendorf, M., Digalakis, V., Kimball, O., 1996. From HMM's to segment models: A unified view of stochastic modeling for speech recognition. IEEE Trans. Speech Audio Process. 4 (5), 360-378.

Ostendorf, M., Kannan, A., Kimball, O., Rohlicek, J., 1992. Continuous word recognition based on the stochastic segment model. Proc. DARPA Workshop CSR.

Parihar, N., Picone, J., Institute for Signal and Infomation Processing, Mississippi State Univ., 2002. Aurora working group: DSR front end LVCSR evaluation AU/384/02.

Picone, J., Pike, S., Regan, R., Kamm, T., Bridle, J., Deng, L., et al., 1999. Initial evaluation of hidden dynamic models on conversational speech. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Plahl, C., Schluter, R., Ney, H., 2011. Cross-lingual portability of Chinese and English neural network features for French and German LVCSR. In: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 371-376.

Povey, D., Kanevsky, D., Kingsbury, B., Ramabhadran, B., Saon, G., Visweswariah, K., 2008. Boosted MMI for model and feature-space discriminative training. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4057-4060.

Povey, D., Woodland, P.C., 2002. Minimum phone error and I-smoothing for improved discriminative training. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 105-108.

Rabiner, L., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77 (2), 257-286.

Rabiner, L., Juang, B.H., 1993. Fundamentals of Speech Recognition. Prentice-Hall, Upper Saddle River, NJ.

Renals, S., Morgan, N., Boulard, H., Cohen, M., Franco, H., 1994. Connectionist probability estimators in HMM speech recognition. IEEE Trans. Speech Audio Process. 2 (1), 161-174.

Robinson, A., 1994. An application to recurrent nets to phone probability estimation. IEEE Trans. Neural Networks 5 (2), 298-305.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1988. Learning representations by back-propagating errors. Cognitive modeling. AU4

Russell, M., Jackson, P., 2005. A multiple-level linear/linear segmental HMM with a formant-based intermediate layer. Comput. Speech Lang. 19, 205-225.

Sainath, T., Kingsbury, B., Mohamed, A., Dahl, G., Saon, G., Soltau, H., et al., 2013a. Improvements to deep convolutional neural networks for LVCSR. In: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 315-320.

Sainath, T., Kingsbury, B., Soltau, H., Ramabhadran, B., 2013b. Optimization techniques to improve training speed of deep neural networks for large speech tasks. IEEE Trans. Audio Speech Lang. Process. 21 (11), 2267-2276.

Sainath, T., Mohamed, A., Kingsbury, B., Ramabhadran, B., 2013c. Deep convolutional neural networks for LVCSR. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8614-8618.

Sainath, T.N., Kingsbury, B., Ramabhadran, B., Fousek, P., Novák, P., Mohamed, A., 2011. Making deep belief networks effective for large vocabulary continuous speech recognition. In: Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp. 30-35.

Sak, H., Senior, A., Beaufays, F., 2014a. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In: Proc. Interspeech.

Sak, H., Vinyals, O., Heigold, G., Senior, A., McDermott, E., Monga, R., et al., 2014b. Sequence discriminative distributed training of long short-term memory recurrent neural networks. In: Proc. Annual Conference of International Speech Communication Association (INTERSPEECH).

Schultz, T., Waibel, A., 1998. Multilingual and crosslingual speech recognition. In: Proc. DARPA Workshop on Broadcast News Transcription and Understanding, pp. 259-262.

Seide, F., Fu, H., Droppo, J., Li, G., Yu, D., 2014. On parallelizability of stochastic gradient descent for speech DNNs. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Seide, F., Li, G., Yu, D., 2011. Conversational speech transcription using context-dependent deep neural networks. In: Proc. Annual Conference of International Speech Communication Association (INTERSPEECH), pp. 437-440.

Seltzer, M.L., Yu, D., Wang, Y., 2013. An investigation of deep neural networks for noise robust speech recognition. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7398-7402.

Senior, A., Heigold, G., Bacchiani, M., Liao, H., 2014. GMM-free DNN training. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Sha, F., 2007. Large margin training of acoustic models for speech recognition. Ph.D. thesis, University of Pennsylvania.

Sha, F., Saul, L., 2006. Large margin Gaussian mixture modeling for phonetic classification and recognition. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Stevens, K., 2000. Acoustic Phonetics. MIT Press, Cambridge, MA.

Su, H., Li, G., Yu, D., Seide, F., 2013. Error back propagation for sequence training of context-dependent deep networks for conversational speech transcription. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6664-6668.

Swietojanski, P., Li, J., Huang, J.T., 2014. Investigation of maxout networks for speech recognition. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Togneri, R., Deng, L., 2003. Joint state and parameter estimation for a target-directed nonlinear dynamic system model. IEEE Trans. Signal Process. 51 (12), 3061-3070.

Tuske, Z., Golik, P., Schluter, R., Ney, H., 2014. Acoustic modeling with deep neural networks using raw time signal for LVCSR. In: Proc. Annual Conference of International Speech Communication Association (INTERSPEECH).

Vanhoucke, V., Devin, M., Heigold, G., 2013. Multiframe deep neural networks for acoustic modeling. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Vanhoucke, V., Senior, A., Mao, M., 2011. Improving the speed of neural networks on CPUs. In: Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning.

Veselỳ., K., Ghoshal, A., Burget, L., Povey, D., 2013. Sequence-discriminative training of deep neural networks. In: Proc. Interspeech, pp. 2345-2349.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., Lang, K., 1989. Phoneme recognition using time-delay neural networks. IEEE Trans. Speech Audio Process. 37 (3), 328-339.

Weng, C., Yu, D., Watanabe, S., Juang, B., 2014. Recurrent deep neural networks for robust speech recognition. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Weninger, F., Geiger, J., Wöllmer, M., Schuller, B., Rigoll, G., 2014. Feature enhancement by deep LSTM networks for ASR in reverberant multisource environments. Comput. Speech Lang. 888-902.

Wiesler, S., Li, J., Xue, J., 2013. Investigations on hessian-free optimization for cross-entropy training of deep neural networks. In: Proc. Interspeech, pp. 3317-3321.

Wöllmer, M., Weninger, F., Geiger, J., Schuller, B., Rigoll, G., 2013a. Noise robust ASR in reverberated multisource environments applying convolutive NMF and long short-term memory. Comput. Speech Lang. 27 (3), 780-797.

Wöllmer, M., Zhang, Z., Weninger, F., Schuller, B., Rigoll, G., 2013b. Feature enhancement by bidirectional LSTM networks for conversational speech recognition in highly non-stationary noise. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6822-6826.

Xiao, X., Li, J., Cheng, E.S., Li, H., Lee, C.H., 2010. A study on the generalization capability of acoustic models for robust speech recognition. IEEE Trans. Audio Speech Lang. Process. 18 (6), 1158-1169.

Yu, D., Deng, L., 2007. Speaker-adaptive learning of resonance targets in a hidden trajectory model of speech coarticulation. Comput. Speech Lang. 27, 72-87.

Yu, D., Deng, L., 2011. Deep learning and its applications to signal and information processing. In: IEEE Signal Processing Magazine., vol. 28, pp. 145-154.

Yu, D., Deng, L., 2014. Automatic Speech Recognition—A Deep Learning Approach. Springer, New York.

Yu, D., Deng, L., Acero, A., 2006. A lattice search technique for a long-contextual-span hidden trajectory model of speech. Speech Commun. 48, 1214-1226.

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

Yu, D., Deng, L., Liu, P., Wu, J., Gong, Y., Acero, A., 2009. Cross-lingual speech recognition under runtime resource constraints. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4193-4196.

Zeiler, M., Ranzato, M., Monga, R., Mao, M., Yang, K., Le, Q., et al., 2013. On rectified linear units for speech processing. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3517-3521.

Zen, H., Tokuda, K., Kitamura, T., 2004. An introduction of trajectory model into HMM-based speech synthesis. In: Proc. of ISCA SSW5, pp. 191-196.

Zhang, C., Woodland, P., 2014. Standalone training of context-dependent deep neural network acoustic models. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Zhang, L., Renals, S., 2008. Acoustic-articulatory modelling with the trajectory HMM. IEEE Signal Process. Lett. 15, 245-248.

Zhang, T., 2004. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In: Proceedings of the twenty-first international conference on Machine learning.

Zhou, J.L., Seide, F., Deng, L., 2003. Coarticulation modeling by embedding a target-directed hidden trajectory model into HMM—model and training. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, Hongkong, pp. 744-747.

Non-Print Items

Abstract:

In this chapter, we introduced the fundamental concepts and component technologies for automatic speech recognition. The topics reviewed in this chapter include several important types of acoustic models—Gaussian mixture models (GMM), hidden Markov models (HMM), and deep neural networks (DNN), plus several of their major variants. The role of language modeling is also briefly discussed in the context of the fundamental formulation of the speech recognition problem.

The HMM with GMMs as its statistical distributions given a state is a shallow generative model for speech feature sequences. Hidden dynamic models generalize the HMM by incorporating some deep structure of speech generation as the internal representations of speech features. Much of the robust speech recognition studies in the past were carried out based on generative models of speech, since the noisy version of speech as the observation signal can be easily “generated” from clean speech using straightforward distortion models. Recently, the discriminative DNN, as well as its convolutional and recurrent variants, have been shown to significantly outperform all previous versions of generative models of speech in speech recognition. The main classes of these deep discriminative models are reviewed in some detail in this chapter. How to handle noise robustness within the framework of discriminative deep learning models of speech, which is less straightforward than the generative models of speech, will be covered in the later chapters of this book.

Keywords: Acoustic modeling, Language modeling, Gaussian mixture models, Hidden Markov models, Deep neural networks