Hierarchical Acoustic Modeling Based on Random-Effects Regression for Automatic Speech Recognition

Yan Han and Lou Boves
Department of Language and Speech, Radboud University of Nijmegen, the Netherlands
y.han@let.ru.nl, l.boves@let.ru.nl

Abstract
Recent research on human intelligence [1] suggests that the auditory system has a hierarchical structure, in which the lower levels store individual properties, and the upper levels store the group properties of utterances. However, most of the conventional automatic recognizers adopt a single-level model structure. In structure-based models, such as HMM and parametric trajectory models, only the group properties of utterances are modeled. In template-based models, only the individual properties of utterances are exploited. In this paper, we propose a novel hierarchical acoustic model to simulate the human auditory hierarchy, in which both the group and the individual properties of utterances can be explicitly addressed. Furthermore, we developed two evaluation methods, namely bottom-up and top-down test, to simulate the prediction-verification loops in human hearing. The model is evaluated on a TIMIT vowel classification task. The proposed hierarchical model significantly outperforms parametric trajectory models.

Index Terms: hierarchical acoustic model, random effects model, human speech recognition, automatic speech recognition

1. Introduction
In spite of substantial progress that has been made in the area of ASR, the performance in the transcription of spontaneous speech still leaves much to desire, even for good quality speech. Human listeners, however, can recognize highly variable speech with high accuracy, even under acoustically adverse conditions. Recent research on human intelligence [1] suggests that the big gap between human listener and automatic recognizers is due to their fundamentally different way in storing and modeling the group and individual properties of speech.

Given the fact that the same word can be spoken by different speakers, with different accents, at different speeds and different emotions, each utterance is unique. However, the unique utterances realizing the same linguistic unit also share some common properties as a group. Thus, both human listener and automatic recognizer has to account for two type of effects: the cross-utterance invariant group characteristics, and the cross-utterance variable individual characteristics. The neocortex of the human brain has a conceptually hierarchical structure. In the hierarchy, the lower levels store the cross-utterance variant patterns relating to individual properties of utterances, and the upper levels store the cross-utterance invariant patterns relating to the group properties. During speech perception the lower and upper levels of the neocortex interact bi-directionally to integrate the group and individual patterns. Contrary to human listener, most automatic recognizers only adopt a single-level model structure.

In single-level structure-based acoustic models, such as hidden Markov models (HMM) [2] and parametric trajectory models (PTM) [3], only the group properties invariant across utterance are accounted for. The invariant patterns include the time-varying means (e.g. the Gaussian means in HMM and the polynomial function in PTM), and the variance between individual utterances and the means. However, these invariant group properties can only roughly depict the cross-utterance variations, and this may hurt the discriminative power of these models. One such an example is illustrated in Fig.1. Given the fact that the same word can be spoken by different speakers, with different accents, at different speeds and different emotions, each utterance is unique. However, the unique utterances realizing the same linguistic unit also share some common properties as a group. Thus, both human listener and automatic recognizer has to account for two type of effects: the cross-utterance invariant group characteristics, and the cross-utterance variable individual characteristics. The neocortex of the human brain has a conceptually hierarchical structure. In the hierarchy, the lower levels store the cross-utterance variant patterns relating to individual properties of utterances, and the upper levels store the cross-utterance invariant patterns relating to the group properties. During speech perception the lower and upper levels of the neocortex interact bi-directionally to integrate the group and individual patterns. Contrary to human listener, most automatic recognizers only adopt a single-level model structure.

In single-level structure-based acoustic models, such as hidden Markov models (HMM) [2] and parametric trajectory models (PTM) [3], only the group properties invariant across utterance are accounted for. The invariant patterns include the time-varying means (e.g. the Gaussian means in HMM and the polynomial function in PTM), and the variance between individual utterances and the means. However, these invariant group properties can only roughly depict the cross-utterance variations, and this may hurt the discriminative power of these models. One such an example is illustrated in Fig.1. Consider a PTM for a certain phoneme with quadratic polynomial mean ($P$ in the figure) and constant variance over time (dotted line around $P$). Suppose the model is trained with utterances $T$. When evaluating the likelihood score of utterance $T'$, the model will generate a probability for $T'$ as high as that obtained from the training data $T$. However, $T'$ never appeared in training data and it may belong to another phoneme which the model should reject. This problem arises because PTMs cannot adequately model the within-utterance variation related to the shapes of the speech feature trajectories. In recognition of highly dynamic speech, the within-utterance variations are even more roughly addressed, and it may seriously degrade recognition performance.

In template-based recognizers, on the other hands, the individual properties of utterances can be explicitly accounted for. In the most basic form, the template-based models directly exploit utterances in the training data, without any modeling assumptions to account for the similarity with other utterances in the same group. This type of models actually have a single level structure, in which very detailed individual properties of utterances are stored. Previous research [4] already showed that...
using a distance measure weighted by some properties shared by a group of utterances can significantly increase the performance of template-based models. This suggests that modeling the cross-utterance invariant group properties is very important.

In order to tackle the problem with the single-level structure-based and template-based models, we propose a novel acoustic model with hierarchical structure analogous to the scheme of human speech perception. The newly proposed model has two levels. On the bottom-level, individual-specific parameters are modeled by the distribution of the regression parameters generated from the bottom-level. The model parameters in the top-level and bottom-level can be learned simultaneously with an Expectation-Maximization (EM) procedure introduced in the theoretical framework of Random Effects Models [5]. In evaluating unknown utterances, we propose two methods, namely bottom-up and top-down processing, simulating the bi-directional interaction in the neocortex.

The paper is further organized as follows. Section 2 presents the theoretical framework of the proposed hierarchical model. In Sections 3, we show its application to vowel classification in TIMIT and show that it outperforms standard PTMs. Finally, in Section 5, we summarize our most important findings and draw our conclusions.

2. Methodology
2.1. Hierarchical Model Structure

Consider a data set comprising \( n \) tokens of a particular phoneme. Let \( Y_i \) be the \( i \)-th utterance consisting of \( m_i \) consecutive frames. For simplicity, the mathematics of the hierarchical model are presented for a one-dimensional acoustic feature. Assume we have a \( p \)-th order linear regression relationship between \( Y_i \) and \( X_i \) as

\[
Y_i = X_i \beta_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2 \mathbf{I})
\]

(1)

where \( Y_i \) is a \( m_i \times 1 \) vector of dependent observations, and \( X_i \) is a \( m_i \times (p + 1) \) design matrix defined as

\[
Y_i = \begin{bmatrix} Y_i(0) \\ Y_i(1) \\ \vdots \\ Y_i(m_i) \end{bmatrix} \quad \text{and} \quad X_i = \begin{bmatrix} 1 & \cdots & (m_i-1)^p \\ 1 & \cdots & (m_i)^p \\ \vdots & \ddots & \vdots \\ 1 & \cdots & (m_i)^p \end{bmatrix}
\]

\( \beta_i \) is a \( (p + 1) \times 1 \) vector of unknown regression coefficients, and \( \epsilon_i \) is a \( m_i \times 1 \) vector of residual errors that has a multivariate normal distribution with zero mean and covariance matrix \( \sigma^2 \mathbf{I} \), where \( \mathbf{I} \) is \( m_i \times m_i \) identity matrix. The regression model gives the conditional density

\[
p(Y_i|X_i, \beta_i, \sigma^2) = f(Y_i|X_i \beta_i, \sigma^2 \mathbf{I})
\]

(2)

in which \( f \) is the \( m_i \) dimensional, multivariate normal density with mean \( X_i \beta_i \) and covariance \( \sigma^2 \mathbf{I} \). Notice that each individual utterance has its own regression model through the parameters \( \beta_i = [\beta_i^1, \sigma_i^2]. \) The bottom-level of the hierarchy explicitly models individual utterances.

We assume that \( \beta_i \)’s in the individual regression models are random samples from a Gaussian distribution

\[
p(\beta_i | \phi) = g(\beta_i | \mu, \Lambda)
\]

(3)

where \( g \) is a \( p + 1 \) dimensional, multivariate normal distribution with mean \( \mu \) and covariance \( \Lambda \). At this level, the top-level of the hierarchy, the cross-utterance invariant patterns are modeled through the parameter \( \phi = [\mu, \Lambda] \).

Given the parameters \( \{ \beta_i, \sigma_i^2, \mu, \Lambda \} \), the conditional density of \( Y_i \) is

\[
p(Y_i|\theta_i, \phi) = p(Y_i|X_i \beta_i, \sigma_i^2)p(\beta_i | \phi) = f(Y_i|X_i \beta_i, \sigma_i^2 | g(\beta_i | \mu, \Lambda)
\]

(4)

2.2. EM Algorithm for Maximum Marginal Likelihood Estimates

The proposed hierarchical model consists of three type of objects: observed random variables \( Y_i \), unobserved random variables \( \beta_i \) and unknown fixed parameters \( \{ \sigma_i^2, \mu, \Lambda \} \). In Eq.(4), \( \beta_i \) figure as observed variables and it actually gives the likelihood \( L(Y_i, \beta_i | \sigma_i^2, \phi). \) This likelihood cannot be used to estimate the model parameters when \( \beta_i \) is unobserved. Thus, a maximum marginal likelihood method has been proposed [5], in which Eq.(4) is integrated over the space of \( \beta_i \)

\[
L(Y_i | \sigma_i^2, \mu, \Lambda) = \int f(Y_i | X_i \beta_i, \sigma_i^2 | g(\beta_i | \mu, \Lambda) d \beta_i = h(Y_i | X_i \mu, \sigma_i^2 \mathbf{I} + X_i \Lambda X^T_i)
\]

(5)

where \( h \) is a \( m_i \) dimensional, multivariate normal distribution with mean \( X_i \mu \) and covariance \( \sigma_i^2 \mathbf{I} + X_i \Lambda X^T_i \). The log marginal likelihood for the complete data set is

\[
\mathcal{L}(Y | \sigma^2, \mu, \Lambda) = \log \prod_{i=1}^{n} h(Y_i | X_i \mu, \sigma_i^2 \mathbf{I} + X_i \Lambda X^T_i)
\]

(6)

In Eq.(6), \( \beta_i \) do not explicitly enter the likelihood function, but the estimates of \( \sigma_i^2 \) depend on the estimates of \( \beta_i \). Consequently, the likelihood function contains missing data, and cannot be directly derived in a one-step MLE algorithm. Therefore, we require an iterative EM algorithm. In the M-step, we have

\[
\hat{\beta}_i = (1/\sigma_i^2 X_i^T X_i + \Lambda^{-1})^{-1} (1/\sigma_i^2 X_i^T Y_i + \Lambda^{-1} \mu)
\]

(7)

and

\[
\hat{\mu} = \frac{\sum \hat{\beta}_i}{n}
\]

(8)

In the E-step,

\[
\hat{\sigma}_i^2 = \| Y_i - X_i \hat{\beta}_i \|^2 + \text{trace}(X_i^T \hat{\beta}_i X_i)
\]

(9)

\[
\hat{\Lambda} = \frac{\sum \hat{\beta}_i (\hat{\beta}_i - \mu) (\hat{\beta}_i - \mu)^T + V_i}{n}
\]

(10)

where \( V_i = (1/\sigma_i^2 X_i^T X_i + \Lambda^{-1})^{-1}. \)

The EM algorithm starts with randomly initialized parameters. Then the E-step and M-step are iteratively performed until convergence on Eq.(6). See [5] for the mathematical details of this EM algorithm.

2.3. Bottom-up and Top-down Testing

When recognizing an unknown utterance \( Y_u \), two alternative ways, namely bottom-up testing and top-down testing, can be applied, using the estimated model parameters for a particular phoneme \( s \), to compute the likelihood that \( Y_u \) belongs to \( s \).
In bottom-up testing, only the group patterns $\phi_s = \{\mu_s, A_s\}$ are assumed to be known. The individual patterns $\theta_s = \{\beta_s, \sigma_s^2\}$ given model $s$ are obtained directly from the unknown utterance $Y_u$. The estimation procedure is similar to the EM algorithm presented in Section 2.2, in which Eq.(7) and Eq.(9) are iteratively used to estimate $\beta_s$ and $\sigma_s^2$, until convergence of the log value of Eq.(5) is reached. The group parameters $\mu_s$ and $A_s$ are constants. Once the parameters $\theta_s = \{\beta_s, \sigma_s^2\}$ are obtained, the likelihood score of $Y_u$ given the model $s$ can be defined as the marginal likelihood

$$S_b(Y_u|s) = h(Y_u|X_u, \mu_s, \sigma_s^2) I + X_u A_s X_u'$$

In top-down testing, both the group parameters $\phi_s = \{\mu_s, A_s\}$ of phoneme $s$ and the individual patterns $\theta_i = \{\beta_s, \sigma_s^2\}$ for the previously observed tokens of phoneme $s$ are assumed to be known. The unknown utterance $Y_u$ is compared to each of the observed tokens, and the similarity between $Y_u$ and $Y_i$ given the model $s$ is measured in terms of the Mahalanobis distance, i.e.,

$$D^*_{s,i} = (\beta_u - \beta_i)' A_s^{-1} (\beta_u - \beta_i)^{-1}$$

The token most similar to $Y_u$ in $s$ is the one with the smallest Mahalanobis distance. The similarities can be used to adapt the results obtained by the bottom-up testing with respect to $s$

$$S_t(Y_u|s) = S_b(Y_u|s) \times \min \{\beta_u - \beta_i\} A_s^{-1} (\beta_u - \beta_i)$$

The bottom-up testing is similar to structure-based models, and the top-down testing is similar to template-based models. However, in the hierarchical model both the bottom-up and top-down testing explicitly utilize the group and the individual properties of the speech observations. The testing procedures will be discussed in more detail in Section 4 after the experimental results are presented.

3. Experiments

3.1. Experimental Set-up

To evaluate the modeling capability of the proposed hierarchical model, we performed experiments on a speaker independent vowel classification task for American English using the TIMIT corpus. The task includes 16 vowels: 13 monophthongs /i, iy, ih, ey, eh, ae, aa, ah, ao, ow, uh, us, er/ and 3 dipthongs /ay, oy, aw/. From the 630 available speakers, 462 are used for training and the remaining 168 are used for testing. We chose not to use the "sa" sentences in training or recognition as in [2], because they introduce a bias for certain vowels in certain contexts. In total, there are 31,865 training tokens and 11,606 test tokens. All tokens are hand-labeled. In acoustic feature extraction, 12 MFCC coefficients and the log-energy are computed using a 25ms Hamming window shifted with 5 ms steps. No delta coefficients are included in the acoustic features.

We compared the classification performance of the proposed hierarchical model with the PTMs proposed in [3]. After the tokens are extracted, both the PTMs and the hierarchical models were trained for each of the 16 vowels. We assumed that the feature dimensions of MFCCs are uncorrelated. Consequently, the PTMs employed a diagonal covariance matrix, and the hierarchical models were trained on basis of individual feature dimensions by using the EM algorithm presented in Section 2.2. The overall likelihood is the sum of the log-likelihoods in individual dimensions. The parameters in the hierarchical models were randomly initialized. The regression order in both the PTMs and the hierarchical models was set to

2. Mixture models for PTMs and for the hierarchical models were also trained. The EM [3] with successive splitting of components is used to train the mixture components of PTMs. Using the resulting mixture model of PTMs, we clustered the training tokens into subgroups based on the trajectory clustering algorithm [6]. After clustering, one hierarchical model was trained with the tokens in each subgroup. The weights of the components in the hierarchical model mixture were set to the corresponding weights of PTM components.

In our first experiment, we evaluated the hierarchical models with bottom-up testing. Since the segment boundaries are known, the unknown test utterance $Y_u$ with length $m_u$ can be classified as vowel $s$ by

$$\hat{s} = \arg\max_s \{\log S_b(Y_u|s) + \alpha \log p(m_u|s)\}$$

where $S_b(Y_u|s)$ are given by Eq.(11), and $p(m_u|s)$ are the duration probabilities, which were computed from a histogram of the training token durations. In order to match the dynamic ranges of $\log S_b(Y_u|s)$ and $p(m_u|s)$, a weighting factor $\alpha$ is applied to the duration term, which is experimentally tuned to optimize the classification performance. The same duration probabilities were also applied to PTMs in [3]. In this experiment the polynomial regressions are quadratic in both PTMs and the proposed hierarchical model.

We then applied the top-down testing as post-processing of the bottom-up results of the single mixture hierarchical models. Based on the likelihoods obtained in bottom-up testing, we generate an $N$-best list of vowel hypotheses for each of the test utterances. The likelihood score $S_t(Y_u|s)$ given in Eq.(13) was used to adapt the result obtained from bottom-up testing. In this experiment, we set $N$ equal to 3. Moreover, we trained cubic regressions in the hierarchical model. We did not apply duration probabilities, neither in top-down nor in bottom-up testing.

3.2. Experimental Results

Table 1. shows the classification results for the hierarchical models with bottom-up testing and for the PTMs. From this table it can be seen that the hierarchical models give significantly higher classification rates than PTMs. Using mixture models increases the classification rates for both the hierarchical models and PTMs; the hierarchical models outperform the PTMs for all numbers of mixtures.

Table 1: Classification rate (%) for bottom-up testing of HTMs with comparison to PTMs (95% Confidence Interval: ±0.9).

<table>
<thead>
<tr>
<th>Method</th>
<th>Mixtures</th>
<th>1</th>
<th>2</th>
<th>4</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTM</td>
<td></td>
<td>56.08</td>
<td>56.74</td>
<td>58.37</td>
<td>57.57</td>
</tr>
<tr>
<td>PTM</td>
<td></td>
<td>53.48</td>
<td>54.04</td>
<td>56.77</td>
<td>56.62</td>
</tr>
</tbody>
</table>

Table 2. illustrates the classification results for the bottom-up testing as post-processing after bottom-up testing. It can be seen that the top-down adaptation significantly improves the classification rates obtained from the bottom-up testing.

3.3. Analysis of Results

In our results the hierarchical models with bottom-up testing significantly outperformed conventional structure-based models. The improvement is due to the fact that the within-utterance variation in a group of utterances is explicitly addressed. Compare the testing likelihood used in bottom-up testing given in Eq.(11) with the likelihood used in PTMs:
of simulating the prediction-verification loop in human speech recognition. In this paper, we present a novel hierarchical acoustic model, which is analogous to the hierarchical structure of the auditory system in the neocortex of the human brain. We proposed two evaluation procedures, namely top-down and bottom-up testing to simulate the prediction-verification loops in the neocortex hierarchy. The hierarchical model was evaluated on a TIMIT vowel classification task. The top-down testing gives significant improvement over the classification rates obtained with PTMs.

6. References