Interpolation of Acoustic Models for Speech Recognition

Thiago Fraga-Silva\textsuperscript{1,2}, Jean-Luc Gauvain\textsuperscript{1}, Lori Lamel\textsuperscript{1}

\textsuperscript{1} LIMSI - CNRS, B.P. 133, 91403 Orsay, France
\textsuperscript{2} Université Paris-Sud, 91403 Orsay, France

\{thfraga, gauvain, lamel\} @limsi.fr

Abstract

Acoustic models for speech recognition are often trained on data coming from a variety of sources. The usual approach is to pool together all of the available training data, considering them all to be part of a unique training set. In this work, assuming that each source may have a different degree of relevance for a given target task, two techniques are proposed to weigh subsets of the training data. The first one is based on the interpolation of the model probability densities, while the other on data weighting. An method to automatically select the mixture coefficients is also proposed. The best technique presented here outperformed unsupervised MAP adaptation and led to improvements in word accuracy (up to 6\% relative) over the pooled model.

Index Terms: Acoustic modeling, model interpolation, adaptation

1. Introduction

In Large Vocabulary Continuous Speech Recognition (LVCSR) systems, the acoustic phenomena is usually modeled by hidden Markov models (HMM). The parameters of such models are often estimated by maximizing the likelihood function on the training data [1]. The accuracy of the models estimated via the maximum likelihood (ML) criterion are strongly dependent on at least two factors. First, the ML estimation assumes that the training and the test sets are drawn from the same distribution function. In other words, the acoustic model will be applied to recognize some data that are similar to the training set. Second, the method relies on the use of a large amount of data to generate robust estimates.

The most common approach used for acoustic modeling is to pool together all of the available training data, considering them all to be part of a unique training set. This method leads to good accuracy if the two aforementioned assumptions hold. However, gathering a fair amount of acoustic training data that matches well the target is a hard task for many different domains, such as conversational telephone speech recognition [2] or non-native speech recognition [3, 4]. This problem is often treated using an adaptation method [4, 5, 6], such as Maximum Likelihood Linear Regression [7] or Maximum a Posteriori (MAP) [8, 9] adaptation. In such cases, an initial (and more general) well-trained model is adapted to some specific domain using a training subset that matches better the target task.

For some tasks, such as broadcast recognition, a large amount of acoustic data can be easily collected from a variety of sources. In such a case, pooling the training data usually leads to overall good performance levels on test sets that also come from a wide variety of sources [10]. However, it is reasonable to consider that each of the training sources can match differently a given test set or another. In general, some improvement of performance can be obtained by adapting the pooled model to some data coming from the same source of the target set [11]. However, the adaptation techniques allow similarity levels to be adjusted to only one of the sources. This work hypothesizes that better parameter estimates can be obtained by considering different degrees of relevance for each of the training sets.

In this direction, this paper proposes a method to measure these degrees of relevance, assigning a mixture coefficient to each training source. These coefficients are estimated using an Expectation-Maximization (EM) algorithm and a held-out data set with its associated manual or automatic transcriptions. In the same manner, two methods to take into account these coefficients are presented. They are both derived from the linear interpolation of the component models associated to each of the training sources. In the first method, the combination of the component models is performed only at the decoding phase. In the second method, the combination of the models is performed at the training phase and requires the re-estimation of the model parameters. It follows that the latter method is approximately equivalent to perform a data weighting during the model parameter estimation.

The principle applied here is strongly inspired by a similar approach widely used on language model training. Such models are commonly obtained by the interpolation of component language models, each one estimated from a different text source. The mixture coefficients are usually obtained by maximizing the likelihood function (or equivalently, minimizing the perplexity) on some held-out data.

Interpolation of acoustic models has been used in other tasks related to speech processing. For instance, for speech synthesis, it has been used to combine models of different speaking styles [12] or models trained for different speakers [13]. For speech recognition, model interpolation has been used to combine native and non-native acoustic models for a non-native speech recognition task as an alternative to acoustic model adaptation [4, 14]. However, in the previously report work, the mixture coefficients were manually selected and the model interpolation technique was assessed only at the decoding phase. This work shows that better performance levels can be achieved by considering the interpolation coefficients during model parameter estimation. Furthermore, it is shown that optimal coefficients (in the likelihood sense) can be automatically estimated.

This paper is organized as follows. Section 2 briefly describes acoustic model training and its use in the recognition system. In Section 3, we propose two methods to interpolate component acoustic models. Section 4 describes the method used to estimate the mixture coefficients. Section 5 describes the recognition system. In the Section 6, the experiments and results are discussed. A brief conclusion is given in Section 7.
2. Acoustic models

In automatic speech recognition, the speech units are usually modeled by continuous density HMMs with Gaussian mixture state observation densities. Training these models requires an alignment between the audio stream and the associated phonemic representation. When manual reference transcriptions are available, the audio data is segmented in phones by a forced alignment procedure, using a pronunciation dictionary and an initial acoustic model. If no transcription is available, unsupervised training [15] can be applied. In this case, an initial system is used to decode the untranscribed data. The acoustic model was trained using the unsupervised method with the best hypothesis given by the decoder as ground truth.

The HMM parameters and the state Gaussian mixture model (GMM) parameters can be obtained by a ML estimation procedure, usually defined as:

$$\lambda^* = \arg\max_{\lambda} f(X|\lambda)$$

where $\lambda$ represents the model parameters, $X$ the observed feature vectors and $f(.)$ the likelihood function.

Equation 1 is an incomplete data problem [16] for both, the HMM and GMM parameters. This problem can be solved iteratively using an EM algorithm. The focus of this paper is on the estimation of the GMM parameters, but can be extended to the HMM parameter estimation.

3. Acoustic model interpolation

The acoustic model estimation is commonly performed considering the training data as a single homogeneous data set. The ML estimation is performed over all the available data. At the decoding phase, the likelihood function $f(x|\lambda)$ of the estimated model $\lambda$ is used together with a language and a pronunciation models in order to find the best possible sequence of words given an observed feature vector $x$.

The usual approach do not take into account the different degrees of relevance of each of the training sources. The following sections propose two methods to consider them. In this work, it was assumed that all the acoustic models have the same structure. Hence, the methods presented were used to interpolate between HMMs of the same model.

It was also assumed that the training data can be split into $K$ independent subsets, such as $X = \{X^1 \cdots X^K\}$. Each subset $X^k$ contains a training vector with dimension $T_k$, represented by $X^k = \{x^k_1 \cdots x^k_{T_k}\}$.

3.1. Gaussian mixture model interpolation

The first interpolation method proposed is straightforward. First, a model $\lambda_k$ is estimated on each of the $X^k$ subsets. The state GMMs are linearly interpolated at the decoding phase, assuming that each model contributes to the acoustic likelihood of an observed vector $x_t$ with a coefficient $\alpha_k$ as follows:

$$f_s,(x_t|\lambda) = \sum_{k=1}^{K} \alpha_k \cdot f_s,(x_t|\lambda_k)$$

where $f_s(.)$ is the GMM density function of a state $s_t$ at time $t$ and $\sum_{k=1}^{K} \alpha_k = 1$.

This interpolation approach can be used in two different ways. First, one can perform the interpolation of the component models $\lambda_k$ at runtime. It is also possible to build a model beforehand, by merging the parameters of the component models and adjusting them according to the interpolation coefficients. Despite some technical issues, these approaches are equivalent. In this work, the latter approach was used.

In comparison with the method presented in the next section, the GMM interpolation has the advantage that each of the component models $\lambda_k$ can be estimated only once independently of the target. By properly adjusting the coefficients, the interpolated model can be used for different tasks.

3.2. Data weighting

In the second method proposed, the mixture coefficients are considered during the parameter estimation by maximizing the likelihood function of the interpolated model given by Equation 2. It follows that this problem can be solved by maximizing the so-called auxiliary function $Q(\lambda, \lambda')$, which is in this case:

$$Q(\lambda, \lambda') = \sum_{k=1}^{K} \sum_{t=1}^{T_k} \sum_{i=1}^{K} \gamma_{it} \log(\omega_k \hat{\omega}_i f_s,(x^k_t|\mu_i, \Sigma_i))$$

where $\hat{\omega}_i$, $\hat{\mu}_i$ and $\Sigma_i$ are, respectively, the mixture coefficient, the mean vector and the covariance matrix of the $i$-th Gaussian component of the model $\lambda$. In this equation, $\gamma_{it} = P(i|x^k_t, \lambda)$ is the probability of being in the Gaussian $i$ at time $t$, given that the model $\lambda$ generates $x^k_t$.

The equations used to compute the new parameter estimates $\hat{\lambda} = \{\hat{\omega}_i, \hat{\mu}_i, \hat{\Sigma}_i\}$, can be obtained by taking the respective partial derivatives of $Q(\ldots)$. They can be expressed as:

$$\hat{\omega}_i = \frac{\sum_{k=1}^{K} \sum_{t=1}^{T_k} \gamma_{it} \alpha_k}{\sum_{k=1}^{K} N_k \alpha_k}$$

$$\hat{\mu}_i = \frac{\sum_{k=1}^{K} \sum_{t=1}^{T_k} \gamma_{it} \alpha_k x^k_t}{\sum_{k=1}^{K} \sum_{t=1}^{T_k} \gamma_{it} \alpha_k}$$

$$\hat{\Sigma}_i = \frac{\sum_{k=1}^{K} \sum_{t=1}^{T_k} \gamma_{it} \alpha_k (x^k_t - \hat{\mu}_i) (x^k_t - \hat{\mu}_i)^T}{\sum_{k=1}^{K} \sum_{t=1}^{T_k} \gamma_{it} \alpha_k}$$

The usual ML re-estimation equations can be obtained by setting $K = 1$ [17]. In practice, the presented method was approximated by performing the usual ML estimation with weighted training data sets.

In comparison with the precedent interpolation method, the data weighting method has the advantage to generate models with less parameters. On the other hand, it is necessary to re-estimate the model parameters for each different target set.

4. Choice of interpolation coefficients

The interpolation coefficients can be estimated by an approach that is quite similar to the one used on language model interpolation. At the first step, component acoustic models $\lambda_k$ are estimated for each training subset. Given some held-out data $x$, an EM algorithm is used to estimate the coefficients, based on the iterative formula [18]:

$$\alpha_k = \frac{\sum_{t=1}^{T} \alpha_k f_s,(x_t|\lambda_k)}{\sum_{k'=1}^{K} \sum_{t=1}^{T} \alpha_{k'} f_s,(x_t|\lambda_{k'})}$$

This task requires an alignment between the data stream and their associated transcriptions in order to calculate the like-
Table 1: Training corpora information. The size, given in hours, corresponds to the amount of raw data in each subset.

<table>
<thead>
<tr>
<th>Subset</th>
<th>Sources</th>
<th>Epoch</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>train1</td>
<td>RTP channel</td>
<td>2000</td>
<td>64</td>
</tr>
<tr>
<td>train2</td>
<td>RTP channel</td>
<td>2010</td>
<td>86</td>
</tr>
<tr>
<td>train3</td>
<td>Voice of America</td>
<td>2009</td>
<td>21</td>
</tr>
<tr>
<td>train4</td>
<td>Euronews</td>
<td>2009-2010</td>
<td>20</td>
</tr>
<tr>
<td>train5</td>
<td>Quaero data</td>
<td>2010</td>
<td>77</td>
</tr>
</tbody>
</table>

Table 2: Information of the data sets used for evaluation. Last column gives the size of the dev set associated to each of the eval set. Size is given in hours.

<table>
<thead>
<tr>
<th>Source</th>
<th>Epoch</th>
<th>Eval Name</th>
<th>Size</th>
<th>Dev Name</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>RTP channel</td>
<td>2000</td>
<td>eval1</td>
<td>1.3</td>
<td></td>
<td>3.9</td>
</tr>
<tr>
<td>Voice of America</td>
<td>2009</td>
<td>eval3</td>
<td>2.0</td>
<td></td>
<td>5.0</td>
</tr>
<tr>
<td>Euronews</td>
<td>2010</td>
<td>eval4</td>
<td>2.0</td>
<td></td>
<td>5.0</td>
</tr>
<tr>
<td>Quaero evaluation</td>
<td>2010</td>
<td>eval5</td>
<td>3.5</td>
<td></td>
<td>3.5</td>
</tr>
</tbody>
</table>

5. Task and System Overview

The experiments were carried out using the LIMSI speech recognition toolkit, with acoustic, lexical and language models developed for Portuguese and tested with broadcast data.

5.1. Corpora

The audio data used in this work contains about 268 hours of untranscribed data. The relevant information about the corpora is shown in Table 1. The training data were separated into five subsets, according to the source the shows come from and the epoch they were broadcast. For the first experiments, the first 4 sources, shown in the upper part of the table, were grouped into a unique set, henceforth *train1-4*. The *train5* set consists of data collected for the 2011 Quaero Programme Evaluation 1. Although it consists of data coming from different sources, they were considered to be homogeneous and treated as part of a unique subset. Manual transcriptions of a portion of the *train5* data were available, but they were not used here.

The systems were evaluated on four data sets, coming from the same sources. The relevant information about the evaluation sets is given in Table 2. For each of these sets, an associated development set was used to estimate the interpolation coefficients. The development set of the Quaero data is the only one for which manual transcriptions were available. All the system parameters were tuned on this set. Automatic transcriptions for all the dev sets were generated using the baseline system.

The language model training data include about 640 million words from nine different written sources, such as newspapers, newswires and blogs. These data cover the period from 1991 to 2010. About 30k words of manual transcriptions from RTP shows broadcast in 2000 were also used. The text data was normalized in order to convert numerical forms and abbreviations to spoken forms.

5.2. System description

The system used in this work is quite similar to other systems developed at LIMSI [19]. It makes use of a pronunciation dictionary, n-gram language models and acoustic models based on continuous density HMMs. Each phone is modeled by a tied-state left-to-right context-dependent triphone HMM, with Gaussian mixture observation densities. Only PLP [20] features were used. The PLP feature vector contains 39 components, including 12 cepstrum coefficients and log energy with their first and second derivatives. The phone set contains 35 phones, as well as special units for silence, breath and hesitation markers.

In this work, the vocabulary, language models and pronunciation dictionary described in [11] were used. The vocabulary was automatically selected based on interpolation of unigram language models. With the selected vocabulary, 2-, 3- and 4-gram component language models were estimated from each of the 10 different sources. The final language models were obtained by interpolation of the component models, with coefficients automatically chosen in order to minimize the perplexity on the Quaero development set. The pronunciation dictionary was obtained via a rule-based grapheme to phoneme converter. For all the experiments, the language models, pronunciation dictionary and decoding parameters were kept fixed. At each test, only the acoustic models were changed.

6. Experiments

The acoustic models were estimated using an unsupervised training approach with the best hypothesis taken as ground truth. The system used to transcribe the training data is described in [10]. In this work, only one iteration of unsupervised training was performed for each model created. All the models trained have the same structure, covering about 15.7k phone contexts and having about 11.5k tied-states. The speech units are modeled by GMMs with up to 32 components, while silence is modeled by mixtures of 1024 components.

6.1. Impact of interpolation coefficients

The first experiments were performed in order to validate the automatic choice of the interpolation coefficients. To better take into account the impact of the coefficients, the training data were separated only into two different sets, *train1-4* and *train5* (See Table 1). Different pairs of coefficients were manually set. Two pairs of coefficients were automatically obtained using the proposed method, with the estimation guided by either, the manual or the automatic transcriptions of the Quaero development set. For each pair of coefficients, a model was estimated using both interpolation methods proposed. The models were evaluated on the *eval5* set.

Table 3 shows the main results obtained. Despite some small differences, the data weighting (DW) and the GMM interpolation (GMMI) methods led to equivalent WER performance levels. The first and last rows show the performances of the component models, trained only on the *train1-4* and *train5* sets. The model estimated on *train5* leads to a WER of 31.7%, which is 2.1% (absolute) smaller than the WER obtained with the model *train1-4*, even if this latter was trained on a subset 2.5 times bigger. Thus, as expected, the *train5* data match better the *eval5* data, since they come from the same source.
The WER obtained with the pooled (baseline) model is given in the second row. The pooled model is equivalent to the weighted model with the coefficients uniformly assigned to the training subsets. This model leads to a WER of 31.7%, which is the same performance obtained with the train5 model alone. Different WER performance levels were obtained by manually varying the interpolation coefficients, reaching a minimum (31.3%) with the coefficients 0.20/0.78. It turns out that these were the coefficients estimated via the manual transcriptions of the Quaero development data. Both interpolation methods led to the best WER performance in this case. The estimation guided by the automatic transcriptions led to a slight different pair of coefficients (0.22/0.78). However, these coefficients still led to the best performance with the data weighting method. A small absolute loss (0.1%) was obtained with the GMM interpolation method, although the performance obtained (31.4%) is still better than the baseline (31.7%).

These results suggest that the optimal interpolation coefficients can be automatically selected by maximizing the likelihood function on the development data. Furthermore, they can be chosen using an unsupervised approach, in which the estimation is guided by the automatic transcriptions of the dev set.

### Table 3: WER(%) on the eval5 set with acoustic models estimated using different pairs of mixture coefficients defined for the subsets train1-4 and train5. Data weighting (DW) and GMM interpolation (GMMI) methods were assessed.

<table>
<thead>
<tr>
<th>AM</th>
<th>Coefficients</th>
<th>WER(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>train1-4</td>
<td>train5</td>
</tr>
<tr>
<td>train5</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>baseline</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>auto, 1-best dev</td>
<td>0.22</td>
<td>0.78</td>
</tr>
<tr>
<td>auto, manual dev</td>
<td>0.20</td>
<td>0.80</td>
</tr>
<tr>
<td>train5</td>
<td>0.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

### Table 4: WER(%) on all the evaluation sets obtained with MAP adapted and interpolated models. The interpolated models were obtained using the data weighting (DW) and GMM interpolation (GMMI) methods. In both cases, equally set coefficients (equal) are compared with automatically selected (auto) ones.

<table>
<thead>
<tr>
<th>Test set</th>
<th>DW (pooled)</th>
<th>auto</th>
<th>GMMI</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>eval1</td>
<td>24.7</td>
<td>24.5</td>
<td>25.1</td>
<td>25.0</td>
</tr>
<tr>
<td>eval3</td>
<td>14.4</td>
<td>13.6</td>
<td>14.3</td>
<td>13.9</td>
</tr>
<tr>
<td>eval4</td>
<td>13.2</td>
<td>13.1</td>
<td>13.7</td>
<td>13.6</td>
</tr>
<tr>
<td>eval5</td>
<td>31.7</td>
<td>31.2</td>
<td>32.0</td>
<td>31.5</td>
</tr>
<tr>
<td>average</td>
<td>22.5</td>
<td>22.1</td>
<td>22.8</td>
<td>22.4</td>
</tr>
</tbody>
</table>

7. Conclusion

This paper has proposed an approach to take into account the relevance of different sources used on acoustic modeling. Two methods to combine subsets of the training data have been presented. In the first one, component models estimated on each source are interpolated at the decoding phase. The second method can be approximated by the parameter estimation with weighted training data. A method to estimate the optimal mixture coefficients using an EM approach has been proposed. It consists in maximizing the likelihood on some acoustic held-out data using either the manual or automatic transcriptions associated to them.

The data weighting method has led to the best WER performances on four different evaluation sets compared to the baseline pooled models, the maximum a posteriori adapted models and the GMM interpolated models. Absolute gains of performance up to 0.8% have been observed over the pooled models and up to 0.6% over the MAP adapted models.

This work can be extended in different ways. For instance, instead of estimating a global mixture coefficient for each component model, better performance levels might be achieved by estimating coefficients per phone or per phone model. Besides that, the GMM interpolation method could be improved to perform model adaptation at recognition time. It could be done by assigning mixture coefficients to each show (or speaker) after a first decoding pass, at the condition that each of the component models are well-trained.
8. References


