Robust HMM Training for Unified Dutch and German Speech Recognition

Rathi Chengalvarayan

Lucent Speech Solutions
Lucent Technologies Inc.
2000 Lucent Lane, Naperville
Illinois 60566, USA
rathi@lucent.com

Abstract

This paper describes our recent work in developing an unified Dutch and German speech recognition system in the SpeechDat domain. The acoustic component of the multilingual system is accomplished through sharing common phonemes without preserving any information about the languages. We propose a more robust MCE-based training algorithm, where only the language dependent phoneme models are allowed to be adjusted, according to the type of training data. Experimental results on Dutch and German subword recognition tasks clearly show an overall string error rate reduction of about 7% and 13% obtained by the newly trained unified recognizer in comparison with the conventional MCE-trained multilingual system.

1. Introduction

As the demand for speech recognition systems in multiple languages grows, the development of multilingual systems which combine the phonetic inventories by exploiting similarities among sounds of many languages into one single acoustic model set is of increasingly attractive [2, 6, 8, 16]. In real-time multilingual applications, a single decoder can be used. The benefits of such an approach are reduced complexity of systems by sharing models, improved language identification (LID) and bootstrapping acoustic models for unseen languages with sparse adaptation data [5, 17]. Alternative approaches usually do LID followed by a language specific recognizer or require multiple recognizers to run in parallel [9]. In this work, we focus on acoustic modeling and our aim is to develop an improved multilingual recognizer capable of decoding a word string in any of a given set of languages.

Combining acoustic models requires the definition of multilingual phonetic inventories [12]. Data driven clustering techniques have been proposed recently by exploiting similarities among sounds of different languages, without significant losses in word accuracy for multilingual tasks [1, 14]. For Dutch and German speech recognition we intend to share acoustic models among sounds of many languages into one single acoustic model set. The next section describes the design of a robust HMM training for a multilingual speech recognizer.

2. Robust HMM Training

We have used two methods for obtaining estimates of the HMM parameters namely the conventional maximum likelihood (MLE) algorithm, and a more effective minimum classification error (MCE) training procedure. For ML training, the segmental k-means training procedure was used [11]. The MCE training directly applies discriminative analysis techniques to string level acoustic model matching, thereby allowing minimum error rate training to be implemented at the string level [10]. A brief formulation of the MCE algorithm using generalized probabilistic descent (GPD) method is as follows:

- A discriminant function in MSE training is defined as

\[ g(O, S_k, \Lambda) = \log f(O, \Theta_{S_k}, S_k|\Lambda), \]

where \( S_k \) is the k-th best string, \( \Lambda \) is the HMM set used in the N-best decoding, \( \Theta_k \) is the optimal state
sequence of the \( k \)-th string given the model set \( \Lambda \), and 
\[
\log f(O, \Theta_{\Lambda_k}, S_k | \Lambda) = \sum \log \left( 1 + e^{-(S_{\Lambda_k} - S_{\Lambda_c}(O, \Lambda))} \right)
\]

is the related log-likelihood score on the optimal path of the \( k \)-th string.

- The misclassification measure is determined by

\[
d(O, \Lambda) = -g(O, S_c, \Lambda) + \log \left( \frac{1}{N - 1} \sum_{S_k \neq S_c} e^{g(O, S_k, \Lambda)} \right)
\]

which provides an acoustic confusability measure between the correct and competing string models.

- The loss function is defined as

\[
l(O, \Lambda) = \frac{1}{1 + e^{-\gamma d(O, \Lambda)}}
\]

where \( \gamma \) is a positive constant, which controls the slope of the sigmoid function.

- The model parameters are updated sequentially according to the GPD algorithm

\[
\Lambda_{n+1} = \Lambda_n - \epsilon \nabla l(O, \Lambda), \quad (1)
\]

\( \Lambda_n \) is the parameter set at the \( n \)-th iteration, \( \nabla l(O, \Lambda) \) is the gradient of the loss function for the training sample \( O \) which belongs to the correct class \( c \), and \( \epsilon \) is a small positive learning constant.

In this study, we trained four different subword acoustic models using the MCE-based discriminative training. Make a note that the MLE-trained models are used as the initial boot model for the subsequent MCE algorithm.

- **Dutch**: This monolingual model is trained by using all the Dutch training corpus and by using only the Dutch specific phonemes. Dutch has about 41 phonemes including the silence. During MCE training, the free running phoneme grammar is used to get the top N-best candidates.

- **German**: This monolingual model is trained by using all the German training corpus and by using only the German specific phonemes. German has about 45 phonemes including the silence. During MCE training, the free running phoneme grammar is used to get the top N-best candidates.

- **Multi-I**: In this case, the global phoneme set of 50 models are generated first by using the MLE training and then by using the MCE-based training. The set of 50 multilingual ASR phonemes are shown in Table 1. We have used both the Dutch and German databases for training this multilingual speech recognizer. During MCE training, there is no constraint is imposed on the phoneme sequence. The top N-best candidates for a given training sample is generated by using free-phoneme recognition. During German training data, the Dutch phonemes can become part of N-best candidate creation and vice-versa. Hence, all the language-dependent phonemes are free to be updated during any given training sample. We call this type of trained model as Multi-I.

- **Multi-II**: The training is done similar to Multi-I, but there is a new constraint incorporated during MCE training phase. Whenever a German training sample is presented, only the German phonemes are updated during MCE training by generating the top N-best candidate using only the German phoneme set (45 phonemes). The remaining five Dutch specific phonemes are untouched during this German data training. The five Dutch specific phonemes will not get into the top N-best candidate generation and hence will not be trained during MCE estimation. Similarly the nine German specific phonemes will not get updated during MCE training of Dutch data. The 35 common phonemes are always updated irrespective of the type of training data. The set of Dutch and German specific phonemes are shown in Table 2 and Table 3. Note that 70% of the phonemes are common to both the languages and the rest are specific to either Dutch or German language. We call this type of MCE-trained unified model as Multi-II.

In this paper, we report only the results obtained by sequential training. During the model training phase, we call one complete pass through the training data set as an epoch. For the case of string-by-string training, model parameters are updated several times over an epoch.

### Table 1: Unified mapping between SAMPA and ASR Phoneme set.

<table>
<thead>
<tr>
<th>SAMPA</th>
<th>ASR</th>
</tr>
</thead>
<tbody>
<tr>
<td>@</td>
<td>O</td>
</tr>
<tr>
<td>6</td>
<td>o</td>
</tr>
<tr>
<td>A</td>
<td>pf</td>
</tr>
<tr>
<td>a</td>
<td>p</td>
</tr>
<tr>
<td>b</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>9</td>
</tr>
<tr>
<td>Au</td>
<td>R</td>
</tr>
<tr>
<td>d</td>
<td>r</td>
</tr>
<tr>
<td>E</td>
<td>S</td>
</tr>
<tr>
<td>e</td>
<td>s</td>
</tr>
<tr>
<td>f</td>
<td>ts</td>
</tr>
<tr>
<td>g</td>
<td>U</td>
</tr>
<tr>
<td>9:</td>
<td>u</td>
</tr>
<tr>
<td>h</td>
<td>E</td>
</tr>
<tr>
<td>l</td>
<td>v</td>
</tr>
<tr>
<td>i</td>
<td>aU</td>
</tr>
<tr>
<td>j</td>
<td>OY</td>
</tr>
<tr>
<td>k</td>
<td>w</td>
</tr>
<tr>
<td>l</td>
<td>Y</td>
</tr>
<tr>
<td>m</td>
<td>Z</td>
</tr>
<tr>
<td>N</td>
<td>z</td>
</tr>
<tr>
<td>n</td>
<td>sil</td>
</tr>
<tr>
<td>n</td>
<td>#</td>
</tr>
</tbody>
</table>

This monolingual model is trained by using all the Dutch training corpus and by using only the Dutch specific phonemes. Dutch has about 41 phonemes including the silence. During MCE training, the free running phoneme grammar is used to get the top N-best candidates.
3. Speech Data

This section briefly describes the Dutch and German telephone speech databases. The SpeechDat (II) database is designed for development and assessment of Dutch and German speech recognizers [7]. This database was collected over the landline telephone network. Recordings were done using an ISDN telephone interface, yielding 8 KHz, 8-bit samples A-law coded signals and the environment is divided into home, office, street, factory and shop. In case of Dutch, the SpeechDat (II) contains 200 speakers for training and 50 for testing, while the PolyPhone consists of 4022 speakers for training and the remaining 1028 for evaluation [3]. Most of the SpeechDat (II) recordings of Dutch were done over the GSM digital mobile network, a few were done over the NMT analog mobile network. The German SpeechDat (I) contains 2000 speakers and all of the data is used for training, while the SpeechDat (II) contains 800 speakers for training and the rest of 200 speakers for testing.

Most items are read, some are spontaneously spoken. The data distribution of the training and testing set is shown in Tables 4 and 5. Only the valid speech utterances were selected for training and testing. The Dutch has 19187 speech utterances for training and 2521 test utterances. The German training database consists of about 17089 strings and the testing database has a total of 1343 strings. The lexicon file, alphabetically ordered list of distinct lexical items which occur in the corpus with the corresponding pronunciation information, of about 23870 Dutch and 17145 German words are provided in the SpeechDat (II) database. The phonetic transcription is written with SAMPA phoneme symbol [13]. We mapped the SAMPA phoneme symbol to a single character ASR phoneme set as shown in Table 1. There are about 41 ASR phonemes in Dutch and 45 ASR phonemes in German after mapping the nasal vowels to oral vowels.

4. Experimental Results

The recognizer feature set consists of 39 features that includes the 12 liftered linear predictive cepstral coefficients, log-energies, their first and second order derivatives [4]. The initial boot model was trained using one iteration of conventional MLE training procedure [11]. Training included updating all the parameters of the model, namely, means, variances and mixture gains using five epochs of MCE training [10]. The MCE training as discussed in section 2 is applied to all training data with sequential modes resulting in four different models (Dutch, German, Multi-I and Multi-II). The Dutch and German models are monolingual and the Multi-I and Multi-II are multilingual models. Each training utterance is signal conditioned by applying cepstral mean subtraction prior to being used in MCE training. The number of competing string models was set to four and the step length was set to one. The multilingual unified subword model set used in the recognition consists of 50 context independent units. Each subword is modeled by a three state left-to-right continuous density HMM with only self and forward transitions. A mixture of Gaussians with diagonal covariance is employed to estimate the density function for each state. A maximum of 12 mixtures per state is allowed. The silence/background is modeled with a single state, 32 Gaussian mixture HMM. The lexical representations of the sentences are obtained by preprocessing the sentences orthographic transcriptions through a SpeechDat-II dictionary.

Tables 6 and 7 show the average string accuracy for three different acoustic models with various MCE algorithm. The monolingual system (German), trained by the MCE method, yields an overall string accuracy of about 93.74% and that of Dutch gives about 93.26%. The monolingual models yield the best result among the multilingual speech recognition system. The Multi-I model yields about 92.41% for German and 92.38% for Dutch testing data. The Multi-II model provides a string accuracy of about 93.37% for German and 92.90% for Dutch testing data,
Table 6: German subword speech recognition test results using the conventional and unified MCE training.

<table>
<thead>
<tr>
<th>Testing Databases</th>
<th>String Accuracies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>German</td>
</tr>
<tr>
<td>Company Names</td>
<td>90.00%</td>
</tr>
<tr>
<td>City Names</td>
<td>89.55%</td>
</tr>
<tr>
<td>Family Names</td>
<td>97.18%</td>
</tr>
<tr>
<td>Keywords</td>
<td>95.63%</td>
</tr>
<tr>
<td>Overall</td>
<td>93.74%</td>
</tr>
</tbody>
</table>

Table 7: Dutch subword speech recognition test results using the conventional and unified MCE training.

<table>
<thead>
<tr>
<th>Testing Databases</th>
<th>String Accuracies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dutch</td>
</tr>
<tr>
<td>Company Names</td>
<td>97.44%</td>
</tr>
<tr>
<td>City Names</td>
<td>88.10%</td>
</tr>
<tr>
<td>Family Names</td>
<td>95.36%</td>
</tr>
<tr>
<td>Keywords</td>
<td>93.25%</td>
</tr>
<tr>
<td>Overall</td>
<td>93.26%</td>
</tr>
</tbody>
</table>

yielding about 13% and 7% string error rate reduction when compared to Multi-I model set. Also the Multi-II recognition performance is comparable to the best monolingual systems and the model size of Multi-II is about half of that of combined German and Dutch monolingual systems. This shows that by incorporating the constrained phoneme grammar during MCE training provides an efficient way of modeling multilingual phoneme models. Further, the language dependent phoneme characteristics are well preserved during MCE training by using a single Viterbi decoder. An asymptotic trend by increasing the number of languages, both in terms of absolute recognition performance and in terms of the performance gap between monolingual and multilingual models, was observed.

5. Conclusions

We constructed a robust, compact, and efficient multilingual system, which uses a single decoder for Dutch and German sentences, and is capable of recognizing sentences with words from both languages. To create a robust multilingual acoustic models we incorporated a special grammar constraint during discriminative training. It provides flexibility and ease in recognizing bilingual sentences [5, 15]. The experimental results suggest that the proposed acoustic modeling procedure can be used for robust multilingual speech recognition and further the results are comparable with the matched monolingual system.

Acknowledgements

The author would like to thank Dr. Rafid Sukkar of Lucent Speech Solutions for his support in the early stages of this work.

6. References


