Recognition of Spelled City Names in Automotive Environments

Andreas Korthauer

Robert Bosch GmbH, Stuttgart, Germany
Andreas.Korthauer@de.bosch.com

Abstract

This contribution presents the development and evaluation of a spelled letter recognizer for automotive environments. Speech data which have been collected in real-world driving situations are used for the training of the hidden Markov models. The feature extraction is optimized with regard to the recognition accuracy. The best results for both arbitrary spelling sequences and constrained city name recognition are achieved by a system with two-channel LDA and integrated noise reduction.

1. Introduction

Following the trend that more and more services are available in the car via mobile communication and on-board electronic systems, information and automotive technologies are moving closer together today. Therefore the development of ergonomic non-distracting user interfaces for complex cockpit functions is necessary. Automatic speech recognition (ASR) technology supplies (among others) a suitable tool for this challenging task.

Specifically, the spoken language dialog for the navigation system requires reliable recognition of thousands of city names. In this context the recognition of spelling sequences is needed as fall-back strategy and for the disambiguation of similar sounding names. Research in spelled letter recognition has mainly been done for name retrieval in telephone-based services so far [1, 2, 3]. In this paper we discuss the recognition of spelled city names in the context of automotive applications.

For that purpose we have developed a speaker-independent spelled letter recognizer on the basis of hidden Markov models (HMMs) using the HTK toolkit [4]. Several feature extraction schemes were investigated and compared with regard to the recognition performance of the system.

2. Car speech database

The recognizer has been trained on the Car Speech Data Collection (CSDC) of the MoTIV project [5]. This database collection contains speech signals of 630 speakers. It has been collected in 7 different cars with up to 4 parallel microphones (Table 1). The corpus comprises 3350 sequences with 39372 continuously spelled letters. The CSDC database was divided into three mutual exclusive sets for training (80 %), cross-validation (10 %), and evaluation (10 %). Only the evaluation results are presented here.

Additionally, 1428 spelling sequences with a total of 11412 spelled letters from the German SpeechDat-Car database [6, 7] were used for evaluation purposes only.

As the speech data was collected in real driving situations most of the occurring noises were recorded, too. The SNR has been measured with the algorithm given in [8]. It varies between -7 dB and +30 dB. Aiming at an optimal spelled letter recognizer for car environments the HMMs were trained using the noisy speech signals.

3. Optimized feature extraction

An optimal feature extraction has been developed for the recognition of spelled letter sequences in the car environment. There are two major problems to be solved:

1. Spoken letters are highly confusable within certain sets. The most relevant is the E-set with the German letters \{E, B, C, D, G, P, T, W\}. Linear discriminant analysis is a means to find ASR features with optimal class separability. The configuration of the feature transformation is discussed in section 3.1.

2. The speech signal \(d(k)\) is disturbed by additive noise \(n(k)\) from environmental sources. More formally, the microphone signal \(x(k)\) reads:

\[
x(k) = d(k) + n(k)
\]

(1)

Algorithms for noise reduction can be integrated in the acoustic front end of the recognizer in order to mitigate this problem. Given only the noisy signal \(x(k)\) it is better to integrate noise reduction into the feature extraction stage rather than using it as a mere preprocessing step for a recognizer which has been trained on clean signals. The noise reduction scheme is presented in section 3.2.

3.1. Linear discriminant analysis

The optimized feature extraction is derived from the general concept of feature transformation as shown in Fig. 1. Some primary feature vectors \(y\) (possibly more than one in parallel) are transformed in one joined feature vector \(\tilde{x}\) with reduced dimensionality. As in conventional feature extraction first order derivatives \(\Delta\tilde{x}\) can be appended to the transformed feature vector \(\tilde{x}\).

<table>
<thead>
<tr>
<th>database</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>speakers</td>
<td>203</td>
<td>238</td>
<td>100</td>
<td>89</td>
<td>630</td>
</tr>
<tr>
<td>male</td>
<td>101</td>
<td>122</td>
<td>68</td>
<td>45</td>
<td>336</td>
</tr>
<tr>
<td>female</td>
<td>102</td>
<td>116</td>
<td>32</td>
<td>44</td>
<td>294</td>
</tr>
<tr>
<td>vehicles</td>
<td>BMW 7</td>
<td>Opel</td>
<td>VW</td>
<td>BMW 5</td>
<td>Audi</td>
</tr>
<tr>
<td>micros</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>2/4</td>
</tr>
</tbody>
</table>

Table 1: MoTIV Car Speech Data Collection CSDC
has been recorded simultaneously with at least two microphones. The speech signals of the CSDC and SpeechDat-Car databases can be combined with the signal segment index and the discrete cosine transform is substituted by the sine transform as feature transformation and the vector of Mel-frequency filter bank (MFFB) coefficients as input features. In a first step the discrete cosine transform is substituted by the linear discriminant analysis (LDA). LDA can be interpreted as a two-step operation [11]. First the within-class scatter matrix is transformed to the unity matrix performing a global de-correlation of the feature space and making it invariant against further linear transformations. Then the class separability is maximized by an eigen value decomposition of the between-class scatter matrix.

### 3.1.2. Dynamic features

As a convenient possibility to enlarge the temporal context which is included into the transformed vector the input vector of the LDA can be composed of time-delayed feature vectors, i.e. with the signal segment index \( m \) we have:

\[
y^{(u)}(m) = y^{(1)}(m-u+1), \quad 1 < u \leq U \tag{2}
\]

In contrast to static MFCC features which cover a context of only 20 ms the transformed vector \( \tilde{z} \) can include e.g. up to 100 ms. It has been proven advantageous to combine the transformation of a few sequential static feature vectors \( y \) with regression analysis of the transformed vector (i.e. \( \Delta \tilde{z} \)) as indicated in Fig. 2. The optimal context was found to be about 80 ms.

### 3.1.3. Multi-channel combination

The speech signals of the CSDC and SpeechDat-Car databases has been recorded simultaneously with at least two microphones (cf. Table 1). Far-talk microphones attached to the A-pillar and to the roof near the rear-view mirror has been used in all recordings. The spacing between these microphones is about 40 cm. The combination of these recording channels should provide additional information for the recognizer and improve recognition accuracy. The microphone signals \( x_1(k) \) and \( x_2(k) \) can be merged optimally in the feature space using the primary feature vectors \( y_1 \) and \( y_2 \) rather than in the spectral or time domain as e.g. in the classical beam forming approach [12]. According to Fig. 3 the two-channel LDA is combined with the extraction of dynamic features as discussed in the previous section.

### 3.2. Integrated noise reduction

The two-channel LDA feature extraction does not directly address the additive noise problem. Therefore a two-channel noise reduction has been integrated into the feature extraction. We started with a conventional coherence weighting algorithm [13]. In our implementation the short-time magnitude-squared coherence (MSC) [14]

\[
G(m, \nu) = \frac{\left| \Phi_{x_1x_2}(m, \nu) \right|^2}{\Phi_{x_1x_1}(m, \nu) \cdot \Phi_{x_2x_2}(m, \nu)} \tag{3}
\]

is applied to the mean estimate of the short-time power density spectra of both input signals

\[
\Phi_{dd}(m, \nu) = G(m, \nu) \cdot \left| X_1(m, \nu) \right|^2 + \left| X_2(m, \nu) \right|^2 / 2 \tag{4}
\]

where \( m \) and \( \nu \) denote the signal segment index and the discrete frequency variable respectively. The estimated power density spectrum \( \Phi_{dd}(m, \nu) \) of the speech signal \( d(k) \) is directly used for the calculation of MFCC features. In a second approach the spectral weighting \( G(m, \nu) \) is applied to each microphone channel separately. Then the estimated power density spectra \( \Phi_{dd_1}(m, \nu) \) and \( \Phi_{dd_2}(m, \nu) \) are transformed to MFFB coefficients and combined with the two-channel LDA. The overall block diagram of this two-channel feature extraction with integrated noise reduction is given in Fig. 4.

### 4. Experiments

The HMMs were trained on the CSDC training set for all experiments using Viterbi training for model initialization and embedded Baum-Welch re-estimation during the mixup procedure [4]. The recognizer works with plain continuous density HMMs and 20 mixtures per state. The basic system consists of whole-word models with 3–21 states. Phoneme models with 2–5 states were used with the LDA feature transformation. The
recognition of spelling sequences has been evaluated on the CSDC and the SpeechDat-Car corpus.

4.1. Arbitrary spelling sequences

In a first set of experiments no constraints are put on the Viterbi algorithm besides conventional beam search pruning so that arbitrary spelling sequences are recognizable. Each letter sequence is considered to be equally probable a priori. The word correctness for the evaluation sets of the CSDC and SpeechDat-Car databases are shown in Tables 2 and 3.

Table 2: Word correctness (WC) on the CSDC database

<table>
<thead>
<tr>
<th>feature extraction</th>
<th>context</th>
<th>WC / %</th>
</tr>
</thead>
<tbody>
<tr>
<td>basic system MFCC</td>
<td>60 ms</td>
<td>77.2</td>
</tr>
<tr>
<td>coherence weighting + MFCC</td>
<td>60 ms</td>
<td>80.4</td>
</tr>
<tr>
<td>mono-channel LDA</td>
<td>80 ms</td>
<td>82.1</td>
</tr>
<tr>
<td>two-channel LDA</td>
<td>80 ms</td>
<td>83.9</td>
</tr>
<tr>
<td>coherence weighting + two-channel LDA</td>
<td>80 ms</td>
<td>84.6</td>
</tr>
</tbody>
</table>

Table 3: Word correctness (WC) on the SpeechDat-Car database

<table>
<thead>
<tr>
<th>feature extraction</th>
<th>context</th>
<th>WC / %</th>
</tr>
</thead>
<tbody>
<tr>
<td>basic system MFCC</td>
<td>60 ms</td>
<td>68.2</td>
</tr>
<tr>
<td>coherence weighting + MFCC</td>
<td>60 ms</td>
<td>69.8</td>
</tr>
<tr>
<td>mono-channel LDA</td>
<td>80 ms</td>
<td>75.2</td>
</tr>
<tr>
<td>two-channel LDA</td>
<td>80 ms</td>
<td>75.0</td>
</tr>
<tr>
<td>coherence weighting + two-channel LDA</td>
<td>80 ms</td>
<td>76.5</td>
</tr>
</tbody>
</table>

Figure 4: Two-channel feature extraction with integrated noise reduction

Clearly, within the same temporal context (60 ms resp. 80 ms) the two-channel feature extraction with coherence weighting has an advantage over the systems with MFCC or mono-channel LDA. Moreover the systems with 80 ms context included into the feature vector generally perform better than those with 60 ms.

4.2. City name recognition with integrated tree search

Now the recognition is fully constrained to a given list of names so that only “valid” entries can be recognized. The name list is represented as a tree network [15] which is integrated into the Viterbi search.

For the evaluation we used a list with 41764 German city names which was merged with the spelling sequences spoken in the databases. We also investigated the dependence of the recognition performance on the size of the name list. Fig. 5 shows the sentence correctness as a function of the number of list entries for the basic system with MFCC feature extraction on the A-pillar recordings of the SpeechDat-Car database. From the results can be concluded that the recognition accuracy decreases almost proportional to the logarithm of the list size.

Figure 5: Sentence correctness as a function of list size

The results for fully constrained recognition on the CSDC and SpeechDat-Car databases are presented in Tables 4 and 5. For small lists the improvement according to the optimized fea-
ture extraction is hidden by the effect of the highly constrained Viterbi search. But for the complete list of city names also the two-channel LDA with integrated noise reduction yields the best results. As in the case of arbitrary spelling sequences the recognition performs worse on the SpeechDat-Car data because the HMMs have been trained on the CSDC database.

5. Conclusions
We presented results from the evaluation of a spelled letter recognizer which has been developed for the noisy car environment. The optimized two-channel feature extraction including coherence-based noise reduction and LDA feature transformation consistently yields the best performance for the recognition of arbitrary spelling sequences as well as for the constrained recognition of 42000 German city names.

6. Acknowledgments
The Car Speech Data Collection (CSDC) has been funded by the German Ministry of Education, Science, Research and Technology (BMBF) in the scope of the MoTiV MMI Project, contract 19K6122B. Part of the SpeechDat-Car project has been funded by the Commission of the European Communities, Telematics Application Programme, Language Engineering, contract LE4-8334.

7. References


