ABSTRACT

Human auditory perception is perfectly capable to deal with time-invariant linear filter effects, such as those introduced by telephone handsets and telephone channels. We compared two different schemes for modeling human auditory time-frequency masking: RASTA filtering and the dynamic cepstrum representation (DCR). We used a small set of context-independent phone hidden Markov models for a recognition task of connected digit strings over the telephone. We found that RASTA filtering outperformed the Gaussian DCR approach, despite the fact that RASTA represents a more crude approximation of human forward masking. Our results may be influenced by the choice of the mel-frequency cepstral representation that we used. The superior performance of the RASTA technique may also be explained by the fact that the frequency response of the RASTA filter is better matched to the region of modulation frequencies where human auditory perception is most sensitive.

1. INTRODUCTION

For automatic speech recognition over the phone, the filtering effects of the handset microphone and the telephone channel may cause serious performance degradation. However, recognition by humans is very robust against stationary distortions of the speech signal. It has been suggested that the well documented time-frequency masking properties of the human auditory system contribute to this robustness. In this paper we discuss how models of human auditory time-frequency masking can be used in the acoustic front-end of an automatic speech recognizer to enhance channel robustness. We compare two different schemes for modeling time-frequency masking in the context of a connected digit recognition task: RASTA filtering [1] and the dynamic cepstrum representation [2].

The linear filtering characteristics introduced by the communication channel can be modelled in the cepstral domain as a constant additive bias vector. As has been suggested in [1], RASTA filtering a sequence of cepstral vectors effectively removes the DC-component, that can be attributed to the linear filtering characteristics introduced by the communication channel and the overall shape of the vocal tract of the speaker. It has been repeatedly shown that RASTA filtering substantially improves recognition performance when word models or context dependent models are used [1, 3, 4]. However, RASTA may be less suited for context independent (CI) modeling due to the strong left context dependency introduced by the long filter memory. The most important effect of the RASTA filter is its high-pass characteristic at low modulation frequencies. As argumented recently in [5], RASTA filtering may be interpreted as a crude means to approximate the temporal (forward) masking effect in human auditory perception.

A somewhat more sophisticated approach to model temporal masking in human auditory perception has been proposed by [2]. These authors introduced the dynamic cepstrum representation (DCR) as a model of forward auditory time-frequency masking and showed that consistently better recognition performance was achieved when comparing the dynamic cepstrum to the conventional cepstral representation. The ability of the dynamic cepstrum to enhance the spectral dynamics and to reduce the effects of the overall average spectrum make this approach also an attractive candidate for channel normalisation.

In this paper we describe a number of experiments where we trained context-independent (CI) phone model HMMs. We compared the recognition performance in case RASTA filtered and Gaussian DCR acoustic vectors were used. This paper is organised as follows. In sections 2 and 3 the signal processing technique and the speech material are described. In section 4, we discuss the type of HMMs, the cross-validation scheme to train these and the recognition syntax for our experiments. In section 5, we discuss the results for the different channel normalisation techniques that we studied. Finally, the conclusions are presented in section 6.

2. SIGNAL PROCESSING

Speech signals were digitized at 8 kHz and stored in A-law format. After conversion to a linear scale, preemphasis with factor 0.98 was applied. A 25 ms Hamming analysis window that was shifted with 10 ms steps was used to calculate 24 filterband energy values for each frame. The 24 triangular shaped filters were uniformly distributed on a mel-frequency scale. Finally, 12 mel-frequency cepstral coefficients (MFCC's) were derived. We did not apply liftering, because we were using continuous Gaussian mixture density HMMs with diagonal covariance matrices. For this class of models, multiplying a coordinate of the feature vector with some constant (say \( c \)) is equivalent to multiplying the probability used during the dynamic programming (DP) with \( c^{-1} \). In other words: the factor \( \log(c) \) is added to each negative log probability in the DP. Since this factor occurs in each negative log prob, the result of the DP is not affected. In addition to the twelve MFCC's we also used their first time-derivatives (delta-MFCC's), log-energy (logE) and its first time-derivative (delta-logE). In this manner we obtained 26-dimensional feature vectors. Feature extraction was done using HTK v1.4 [6].

For channel normalisation we applied the filtering (either the RASTA filter with integration factor 0.98 [1] or the Gaussian
The speech material for this experiment was taken from the Dutch POLYPHONE corpus [7]. Speakers were recorded over the public switched telephone network in the Netherlands. Hand-set and channel characteristics are not known; especially handset characteristics are known to vary widely. The speakers were selected in such a way that all major dialect backgrounds in the Netherlands are represented. None of the utterances used for training or test had a high background noise level.

Among other things, the speakers were asked to read a connected digit string containing six digits. We divided this set of digit strings in two parts. For training we reserved a set of 960 strings, i.e. 80 speakers (40 females and 40 males) from each of the 12 provinces in the Netherlands (denoted trn960 in short). An independent set of 911 utterances (tst911; 461 females, 450 males) was set apart for testing. (In principle we again wanted to have 40 female and 40 male test speakers). For proper initialisation of the models, we manually corrected automatically generated begin- and endpoints of each utterance in the trn960 data set. We did not always use all training and testing material. Most of the time, we used only half the amount of training data (i.e. 480 utterances, trn480; 240 females, 240 males). For cross-validation during training we used a subset of 240 utterances taken from the test set (tst240; 120 females, 120 males). For evaluation of the models when training was completed we always used the full test set tst911. We list the number of available realisations of each digit for all of our data sets in columns 3 till 6 of Table 1.

### Table 1: Phonic transcriptions (column 2) and the number of realisations (columns 3 till 7) of each digit.

<table>
<thead>
<tr>
<th>Digit</th>
<th>Transcription</th>
<th>trn960</th>
<th>trn480</th>
<th>tst911</th>
<th>tst240</th>
</tr>
</thead>
<tbody>
<tr>
<td>nul</td>
<td>n Y I</td>
<td>590</td>
<td>294</td>
<td>548</td>
<td>136</td>
</tr>
<tr>
<td>een</td>
<td>e n</td>
<td>590</td>
<td>286</td>
<td>562</td>
<td>165</td>
</tr>
<tr>
<td>twe</td>
<td>t w e</td>
<td>591</td>
<td>296</td>
<td>597</td>
<td>181</td>
</tr>
<tr>
<td>drie</td>
<td>d r i</td>
<td>597</td>
<td>299</td>
<td>574</td>
<td>155</td>
</tr>
<tr>
<td>vier</td>
<td>v i r</td>
<td>569</td>
<td>284</td>
<td>523</td>
<td>135</td>
</tr>
<tr>
<td>vijf</td>
<td>v E i f</td>
<td>573</td>
<td>273</td>
<td>526</td>
<td>124</td>
</tr>
<tr>
<td>zes</td>
<td>z E S</td>
<td>578</td>
<td>301</td>
<td>536</td>
<td>136</td>
</tr>
<tr>
<td>zeven</td>
<td>z e v Q n</td>
<td>582</td>
<td>270</td>
<td>510</td>
<td>130</td>
</tr>
<tr>
<td>acht</td>
<td>a x t</td>
<td>554</td>
<td>297</td>
<td>525</td>
<td>151</td>
</tr>
<tr>
<td>neger</td>
<td>n e x Q n</td>
<td>534</td>
<td>281</td>
<td>556</td>
<td>121</td>
</tr>
</tbody>
</table>

3. DATABASE

The speech material for this experiment was taken from the Dutch POLYPHONE corpus [7]. Speakers were recorded over the public switched telephone network in the Netherlands. Hand-set and channel characteristics are not known; especially handset characteristics are known to vary widely. The speakers were selected in such a way that all major dialect backgrounds in the Netherlands are represented. None of the utterances used for training or test had a high background noise level.

Among other things, the speakers were asked to read a connected digit string containing six digits. We divided this set of digit strings in two parts. For training we reserved a set of 960 strings, i.e. 80 speakers (40 females and 40 males) from each of the 12 provinces in the Netherlands (denoted trn960 in short). An independent set of 911 utterances (tst911; 461 females, 450 males) was set apart for testing. (In principle we again wanted to have 40 female and 40 male test speakers). For proper initialisation of the models, we manually corrected automatically generated begin- and endpoints of each utterance in the trn960 data set. We did not always use all training and testing material. Most of the time, we used only half the amount of training data (i.e. 480 utterances, trn480; 240 females, 240 males). For cross-validation during training we used a subset of 240 utterances taken from the test set (tst240; 120 females, 120 males). For evaluation of the models when training was completed we always used the full test set tst911. We list the number of available realisations of each digit for all of our data sets in columns 3 till 6 of Table 1.

4. MODELS

4.1. Model topology

The digit set of the Dutch language was described using 18 context independent (CI) phone models (see second column of Ta-
From Figure 2, it appears that 528 Gaussians (i.e. 8 Gaussians per state) is the optimal acoustic resolution required for the given amount of training data that we used. Although the recognition performance increases monotonically as a function of the acoustic resolution, the improvements are not significant for 16 and 32 Gaussians per state. We found a word error rate (WER; defined as one minus the proportion of digits correct) of 3.7 ± 0.5% at 8 Gaussians per state.

Figure 1: Proportion of words correct as a function of the number of Baum-Welch iterations. Dashed lines: ×, O, + = 1, 2, 4 Gaussians per state. Solid lines: ×, O, + = 8, 16, 32 Gaussians per state.

Figure 2: Recognition performance for test set tst911 as a function of the acoustic resolution of the models.

5.2. Channel normalisation

We replaced the twelve cepstral coefficients in our feature vectors by twelve RASTA filtered cepstral coefficients. In a separate experiment we calculated the Gaussian DCR for each of our twelve cepstral coefficients. In both cases, we retrained HMMs with the new feature vectors using exactly the same utterances for training and cross-validation. Again, the best models sets according to test set tst240 were evaluated with test set tst911. The recognition results are shown in Figure 3 together with the reference results. Figure 3 clearly indicates that both channel normalisation techniques improve the recognition performance: the proportion of digits correct is larger for each number of Gaussians that we tested. Notice that the improvements by using the RASTA filtering technique are significant at the 95% confidence level, but that those of the Gaussian DCR approach are not. Furthermore, the difference between RASTA and Gaussian DCR is significant. At 8 Gaussians per state we found WER = 3.3 ± 0.5% for Gaussian DCR features and WER = 2.6 ± 0.5% in case RASTA filtering was applied. In other words, due to the RASTA filtering the WER was reduced by 29% relative to the baseline performance obtained without channel normalisation. A last experiment was performed to verify that we used enough training data. To this aim models were trained with the trn960 data set using the RASTA filtered acoustic vectors. At 8 Gaussians per state, we found WER = 2.8 ± 0.5%. Because we did not observe a significant change in recognition performance, we concluded that data set trn480 was indeed large enough.

Figure 3: Recognition performance for three types of acoustic features: × = RASTA filtering, O = Gaussian DCR, + = no channel normalisation.

It has been reported that RASTA filtering did not appear to help increasing performance when using CI HMMs, due to the left-context dependency introduced by the long memory of the RASTA filter (for example see [8]). The fact that we did observe an improvement in our experiments may be explained by the fact that the number of different contexts for each phone model is not very large in our connected digit recognition task. In fact, as can be seen in the second column of Table 1, in 9 cases the phone models occur in one single context only, with an average number of different left contexts for each phone model as low as 1.6.

Our results can also be interpreted as follows. Recently, it has
In Figure 4 the frequency response of the RASTA filter and the Gaussian DCR are shown. First, it is important to notice that the attenuation of the DC-component by the RASTA filter is much better than that of the Gaussian DCR. Second, the passband (attenuation less than 3 dB) starts well below 2 Hz for the RASTA technique. However, in case of the Gaussian DCR, the passband starts at 5 to 6 Hz (depending on the cepstral coefficient). This means that especially at low modulation frequencies, the RASTA filter is much better tuned to the maximally sensitive region of human auditory perception. This may explain why we found that RASTA filtering outperformed the Gaussian DCR approach.

![Figure 4: Frequency responses for the RASTA filter (upper panel) and Gaussian DCR (lower panel). In case of the Gaussian DCR the curve starting at -12 dB corresponds to frequency response for the first cepstral coefficient. The frequency response starting at -5 dB corresponds to cepstral coefficient 12.](image)

### 6. CONCLUSIONS

In this paper we discussed how models of human auditory time-frequency masking can be used in the acoustic front-end of an automatic speech recognizer to enhance channel robustness. We compared two different schemes for modelling time-frequency masking: RASTA filtering and Gaussian DCR. We used a small set of CI phone HMMs for connected digit recognition over the phone. Without an explicit channel normalisation technique, we found WER = 3.7 ± 0.5% when using 8 Gaussians per state. Despite the fact that RASTA represents a more crude approximation of human forward masking, RASTA filtering outperformed the Gaussian DCR approach: WER = 2.6 ± 0.5% vs. WER = 3.3 ± 0.5% at 8 Gaussians per state. Our results may be influenced by the choice of the mel-frequency cepstral representation that we used: in its original form, Gaussian DCR modelling was proposed for cepstra based on a linear frequency scale. The superior performance of the RASTA technique may also be explained by the fact that the frequency response of the RASTA filter is better matched to the region of modulation frequencies where human auditory perception is most sensitive (2-16 Hz).

Acknowledgement

This work was funded by the Netherlands Organisation for Scientific Research (NWO) as part of the NWO Priority programme Language and Speech Technology.

---


