Adaptation of HMMS in the presence of additive and convolutional noise

Hans-Günter Hirsch
Ericsson Eurolab Deutschland GmbH,
Nordostpark 12,
90411 Nuremberg, Germany
Email: hans-guenter.hirsch@eedn.ericsson.se

ABSTRACT - The performance of speech recognizers deteriorates in case of a mismatch between the conditions during training and recognition. One difference is the presence of a stationary background noise during recognition which is also referred to as "additive" noise. Furthermore the recognition is influenced by the frequency response of the whole transmission channel from the speaker to the audio input of the recognizer. The term "convolutional" noise has been introduced for this type of distortion. Several approaches are known to compensate these effects individually or both together [1]-[4]. This paper describes an approach which compensates both types of noise. The scheme is based on an estimation of the noise spectrum [5]. Furthermore the frequency response is iteratively estimated by using the alignment information of the best path in the Viterbi algorithm. The comparison between the spectra of the input signal and the spectra of the corresponding HMM (Hidden Markov Model) states is taken as basis for the filter estimation. The estimated additive and convolutional noise components are used as input to the well known Parallel Model Combination (PMC) approach [6] to adapt the whole word HMMs of a speaker independent connected word recognizer. Considerable improvements can be achieved in the presence of just one type of noise as well as in the presence of both types together.

1 Introduction

Robustness is the most important factor limiting the application of speech recognition in a lot of real-life situations. Considering the installation of a recognition system at a switch in a telephone network there exist mainly two types of noise. The first one is the stationary noise which is recorded as background noise at the caller's location and/or which is generated on the telephone line. The second one is the frequency characteristic of the whole transmission channel e.g. including the microphone and the telephone line.

The influence of additive and convolutional noise can be approximately described in the linear spectral domain by

\[ Y(f) = |H(f)|^2 \cdot S(f) + |N(f)| \]  \hspace{1cm} (1)

where \( S(f) \) is the power density spectrum of the clean speech and \( N(f) \) the spectrum of the noise. \( H(f) \) is the frequency response of the whole transmission system. \( Y(f) \) is considered as the input to the recognizer. It is assumed that \( N(f) \) and
H(f) are almost constant or only slowly changing over time. Given an estimate of 
N(f) and H(f) it is possible to adapt the parameters of HMMs. The investigations of 
this study are based on a cepstral analysis as it is used in most of today’s recognition 
systems. Cosine transformations are needed to apply the PMC adaptation scheme [6] 
on this type of parameters.

2 Features of the recognizer

The recognition system used throughout this study is based on a representation 
of speech by cepstral parameters and on the modelling of words by HMMs. A feature 
vector consists of
- 12 Mel frequency cepstral coefficients (MFCCs) including the zeroth cepstral 
  coefficient as representation of the short-term energy
- 12 Delta cepstral coefficients

Feature vectors are calculated every 10 ms analyzing a 25 ms window. The 
spectral analysis is based on a FFT. The power density spectrum is calculated for 22 
subbands in the MEL frequency range. Delta coefficients are calculated applying an 
often used regression formula [7] on 5 consecutive frames of MFCC parameters. 
Whole words are modelled by HMMs with the following features:
- 18 states per word
- mixture of 4 Gaussian distributions for each state
- simple left-to-right models
- covariance matrices with only elements on the diagonal

3 Estimation of noise spectrum and frequency response

The detection of noisy segments is realized by a processing scheme which 
estimates and evaluates the SNRs (signal-to-noise ratios) in subbands [5]. The input 
consists of the short-term energies which are calculated in the feature extraction of 
the recognizer in the MEL frequency range. The output of this detector is an 
indication of noisy segments respectively speech segments and an estimate of the 
noise spectrum. This estimation is individually done for each utterance.

Several approaches exist for the estimation of the frequency response H(f) and 
its application to recognition in additive and convolutional noise [1]-[4]. Some of 
these approaches cause a high computational load or need some special adaptation 
data. The method presented in this paper is computationally quite inexpensive, does 
not cause any delay and does not need any adaptation data.

Looking at formula 1 the frequency response can be estimated as

$$|\hat{H}_{\text{est}}(f)|^2 = \frac{Y_{\text{long}}(f) - \hat{N}(f)}{S_{\text{long}}(f)}$$

Assuming a constant frequency response H(f) and a constant noise spectrum N(f)
during a speech utterance the long-term spectrum $Y_{long}(f)$ of this utterance can be introduced as description of the noisy input speech. This long-term spectrum is calculated by transforming back the cepstral parameters of the noisy input speech to the spectral domain and summing up the short-term spectra over the whole speech utterance. The estimated noise spectrum $\hat{N}(f)$ is determined as described above. The long-term spectrum $\hat{S}_{long}(f)$ of the “clean” speech is estimated by using the spectral information which is contained in the HMMs. After having recognized an utterance the matching information of the Viterbi alignment is used to define the “best” sequence of HMM states which represents the input speech. Transforming back the static cepstral parameters of those HMM states to the spectral domain the long-term spectrum can be calculated as sum over all corresponding HMM states. The whole processing to determine $|\hat{H}_{act}(f)|$ is visualized in the block diagram of figure 1.

![Block diagram for the estimation of the frequency response](image)

Figure 1: Block diagram for the estimation of the frequency response

The estimate $|\hat{H}_{act}(f)|$ of the individual utterance is used to iteratively update the former estimate $|\hat{H}_{old}(f)|$. The new estimate is defined as

$$|\hat{H}_{new}(f)|^2 = \alpha \cdot |\hat{H}_{old}(f)|^2 + (1 - \alpha) \cdot |\hat{H}_{act}(f)|^2$$

where $\alpha$ is chosen less but close to 1. The iterative updating generates a
smoothed version of the frequency response and compensates errors due to estimation errors for an individual utterance.

This new estimate can be used for the recognition of the next utterance. The estimation of the long-term spectra requires inverse transformations of the corresponding cepstral coefficients into the linear spectral domain. The estimation process can be completely done after recognizing an utterance so that it does not cause a delay. Actually this estimation procedure does not precisely estimate the frequency response of the transmission channel. The whole mismatch between training and test data is considered.

Having the estimates of \( N(f) \) and \( H(f) \) a set of adapted HMMs can be calculated with the PMC method. The adaptation is individually done for each utterance when the beginning of speech is detected as described above. The actual estimate of \( N(f) \) and the previous estimate of \( H(f) \) are applied.

4 Recognition experiments

The whole word HMMs are determined for the recognition of digit sequences by using the training part of the TIDIGITS data base. The data base consists of the digits "1" to "9", "zero" and "oh". All data were recorded at a high SNR. The original data are downsampled to 8 kHz for these investigations. Training is done with the HTK package [7].

4.1 Recognition of the TIDIGITS

Applying the clean test data of the TIDIGITS a baseline performance of \( \text{2.37\%} \) string error rate can be achieved for the recognizer as described before without any adaptation. The corresponding word error rate is \( \text{0.77\%} \). The word error rate includes substitution, deletion and insertion errors.

Applying only the noise estimate

A first set of experiments is done where only the estimated noise spectrum is used for the PMC. Noisy versions of the TIDIGITS are created by artificially adding car noise at different SNRs. The car noise was recorded inside a car. The PMC approach is considered which is called the Log-add approximation in [6]. The Log-add approximation is based on an inverse cosine transformation of the static cepstral coefficients back to the spectral domain. The estimated noise spectrum is added in the linear domain to perform the adaptation. Furthermore an adaptation of the delta cepstral coefficients is investigated by applying a simple weighting to the corresponding spectral coefficients in the logarithmic domain according to [8]:

\[
\Delta \hat{S}_{Lg}(f) = \frac{S(f)}{S(f) + \hat{N}(f)} \cdot \Delta S_{Lg}(f)
\] (4)

Some results are plotted in figure 2 when applying the Log-add approximation. A considerable gain can be achieved over the whole range of SNRs. The result for
the clean data is plotted at a SNR of 30 dB. The adaptation of the Delta coefficients further decreases the error rates at low SNRs.

Figure 2: Word error rates applying the Log-add approximation

The adaptation of HMMs based on the PMC method is compared against the well known technique of spectral subtraction. Spectral subtraction is a noise reduction scheme which can be integrated in the feature extraction of the recognizer. Thus this is also a comparison of two principal approaches. The first approach tries to make the feature extraction more robust against certain distortions. In the second approach the references are adapted in respect of the distortion without modifying the existing feature extraction. Some results for the noisy TIDIGITS are plotted in figure 3.

Figure 3 indicates a considerable improvement when applying spectral subtraction in comparison to the case without an adaptation or without a modification of the feature extraction. But the improvement is higher when applying the PMC method with e.g. the Log-add approximation where only the static cepstral means are adapted.

Applying the estimates of the noise and the frequency response

A second series of experiments is run by additionally applying the estimated
frequency response $H(f)$. First of all the performance increases when applying it to the clean data. The word error rate decreases from $0.77\%$ without adaptation to $0.65\%$ when applying the Log-add approximation including the filter estimate. The adaptation of the Delta coefficients is not included in this and all further experiments. The main reason for the improvement can be seen in the adaptation to the speaker's volume and the speaker's long-term spectral characteristics. It has to be mentioned that the test utterances are consecutively processed for each speaker.

Now all test data are filtered with a frequency characteristic simulating a telephone channel. Frequencies below 300 Hz and above 3400 Hz are attenuated by 40 dB. An amplification of about 3 dB/Oct. is applied in the frequency range from 300 to 1000 Hz. The filter characteristic remains flat for frequencies between 1000 up to about 3000 Hz. Some recognition results are listed in table 1. The influence of the filtering can be compensated almost completely by this type of iterative filter estimation.

<table>
<thead>
<tr>
<th></th>
<th>without adaptation</th>
<th>Log-add approximation &amp; filter estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>word error rate</td>
<td>4.23 %</td>
<td>0.71 %</td>
</tr>
</tbody>
</table>

Table 1: Word error rates when recognizing the filtered data
To investigate the performance in the presence of additive and convolutional noise finally car noise is added to the filtered test data at a SNR of 10 dB. Results are listed in table 2.

<table>
<thead>
<tr>
<th></th>
<th>without adaptation</th>
<th>Log-add approximation &amp; filter estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>word error rate</td>
<td>58.1 %</td>
<td>4.2 %</td>
</tr>
</tbody>
</table>

Table 2: Error rates when recognizing the filtered data with car noise added at a SNR of 10 dB

Again a remarkable improvement is achieved. The improvement is still a little bit better in comparison to the condition where only additive noise is considered and the PMC adaptation is applied with a noise estimate only.

The adaptation scheme is able of considerably reducing the deterioration in the presence of additive and convolutional noise.

### 4.2 Recognition of the Bellcore digits

Besides the recognition of artificially distorted data furthermore a different set of data is recognized which was recorded over telephone lines. Still the same HMMs are used which are trained on the "clean" TIDIGITS. Thus a situation is considered with a total mismatch between training and test data. A part of the Bellcore digits data base is used here. This consists of 200 speakers uttering the 11 digits ("1" to "9", "zero", "oh") as isolated words in real-life situations. The data partly contain background noise recorded by the microphone and the usual effects of different telephone lines and different handsets. This time the recognizer is set up to recognize isolated words only. The word error rates are listed in table 3 when applying the Log-add PMC method with estimating the noise only and in case of estimating the noise and the frequency response.

<table>
<thead>
<tr>
<th></th>
<th>Log-add approximation &amp; noise estimation</th>
<th>Log-add approximation &amp; noise and filter estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>word error rate</td>
<td>74.8 %</td>
<td>4.5 %</td>
</tr>
</tbody>
</table>

Table 3: Word error rates when recognizing the 2200 digits of the Bellcore digits

The word error rate is about 75 % without adaptation for this simple task of recognizing 11 words as isolated words in a speaker independent mode. This shows impressively the problem in case of a total mismatch between training and test data.
The error rate decreases considerably when applying the noise estimation together with the Log-add approximation. A further reduction of about a factor of 4 in error rate is achieved when applying the noise estimate and the estimate of the mismatch filter response. This result shows the applicability of the described method on real-life applications.

5 Conclusions

A method is presented which adapts the HMMs to stationary background noise as well as to the "frequency response" mismatch between training and test data. The processing is based on the PMC approach where the noise spectrum as well as the frequency response are estimated. Both estimation schemes work reliable and robust. All investigations are done with respect to an easy implementation in a real-time recognizer.

Acknowledgements

Most of this work was done during a research stay at the centre for speech technology (CTT) which is part of the royal technical university (KTH) in Stockholm. The author would like to thank Rolf Carlson and Björn Granström for their support and hospitality.

References