ITG-Fachbericht 236: Sprachkommunikation · 26.-28.09.2012 in Braunschweig

Combining Different Recognition Schemes by Analyzing the Noise Condition

Hans-Günter Hirsch, Andre Ringl, Andreas Kitzig

Institute for Pattern Recognition, Niederrhein University of Applied Sciences, Krefeld, Germany
Email: {hans-guenter.hirsch,andreas.kitzig}@ hs-niederrhein.de
Web: dnt.kr.hs-niederrhein.de

Abstract

The degradation of the human performance is still considerably lower than the corresponding deterioration of automatic recognition systems when comparing the recognition of noisy versus clean speech. It can be observed that the degradation of the recognition rate is dependent on the applied recognition technique and the specific noise condition. We present an approach to select the appropriate recognition scheme by estimating the noise scenario at each speech input. Two different recognition schemes are applied. One is based on the extraction of robust features whereas the other approach contains an adaptation of HMMs (Hidden Markov Models). In case of extracting robust features we investigate the usage of multi-condition HMMs that have been trained on noisy speech signals. We verify that the process of selecting the appropriate scheme and the appropriate set of HMMs can be applied so that the lowest error rate is achieved for each acoustic condition.

1 Introduction

In comparison to humans, automatic recognition systems still show a considerable degradation of their performance in the presence of background noise and in case of a hands-free speech input in reverberant environments. A lot of different approaches have been developed to reduce the loss in recognition performance. Most schemes are either based on the extraction of robust features or the adaptation of the reference patterns. Knowing the noise scenario of an application in advance, the usage of reference patterns that have been trained with speech signals recorded in the application scenario leads to lowest error rates. But all robust schemes show deficiencies for certain acoustic conditions. For example, the training of HMMs on noisy speech signals comes along with a reduced recognition performance for clean signals.

Thus, we tried to combine the advantages of different recognition schemes in specific acoustic conditions. It has already been investigated to select an appropriate set of HMMs by estimating the noise environment [1], [2]. Our approach is based on the selection of a complete appropriate recognition scheme. As parameters for the selection the SNR (signal-to-noise ratio) and the spectral noise characteristics are estimated at each speech input.

In the following two sections we introduce the robust schemes and we present the recognition performance that can be achieved with these schemes. In Section 4, we describe our approach for selecting and combining the robust schemes. Its usage is verified by presenting the results for recognizing speech signals in different noise conditions.

2 Robust Recognition Schemes

We developed and applied two recognition schemes to improve the recognition performance in the presence of background noise and an unknown frequency weighting. One scheme consists of the extraction of robust features [3] whereas the second scheme is based on adapting the HMMs. The adaptation technique also covers the effects of a hands-free speech input in a reverberant environment [4].

2.1 Robust Feature Extraction

We set up a robust feature extraction based on an adaptive filtering in the spectral domain. To define the characteristics of the filter an estimation of the noise spectrum is needed. An own approach for estimating the noise spectrum is applied that is based on an evaluation of the energy contours in the DFT subbands [5]. The adaptive filtering contains a so called cepstro-temporal smoothing of the filter characteristics [6]. This type of smoothing is introduced for the purpose of speech enhancement to reduce the amount of musical tones and to improve the subjective quality of the noise reduced signal. A “blind equalization” as further processing block is applied to the cepstral coefficients. This processing block has been adopted from a robust front-end standardized by ETSI [7]. The equalization compensates the effect of an unknown frequency weighting, e.g., caused by the frequency response of the microphone or the transmission channel.

The recognition performance that can be achieved with the robust feature extraction scheme in the presence of additive background noise is almost the same as applying the ETSI scheme. But besides a more efficient implementation we observed a better performance for the condition of a hands-free speech input in a noisy environment.

2.2 HMM adaptation

As result of earlier work we have available a technique to adapt certain HMM parameters to an unknown acoustic environment. The adaptation of HMMs is an alternative approach for improving the robustness of a recognition system. The scheme shown in Figure 1 contains an adaptation to noise in the background, to an unknown frequency weighting and to a hands-free speech input in a reverberant environment.

We estimate the noise spectrum at each speech input as one parameter defining the acoustic environment. An unknown frequency weighting and the reverberation time are estimated as further parameters. The frequency weighting is estimated after each speech input by comparing the spectra as defined by the mean cepstral

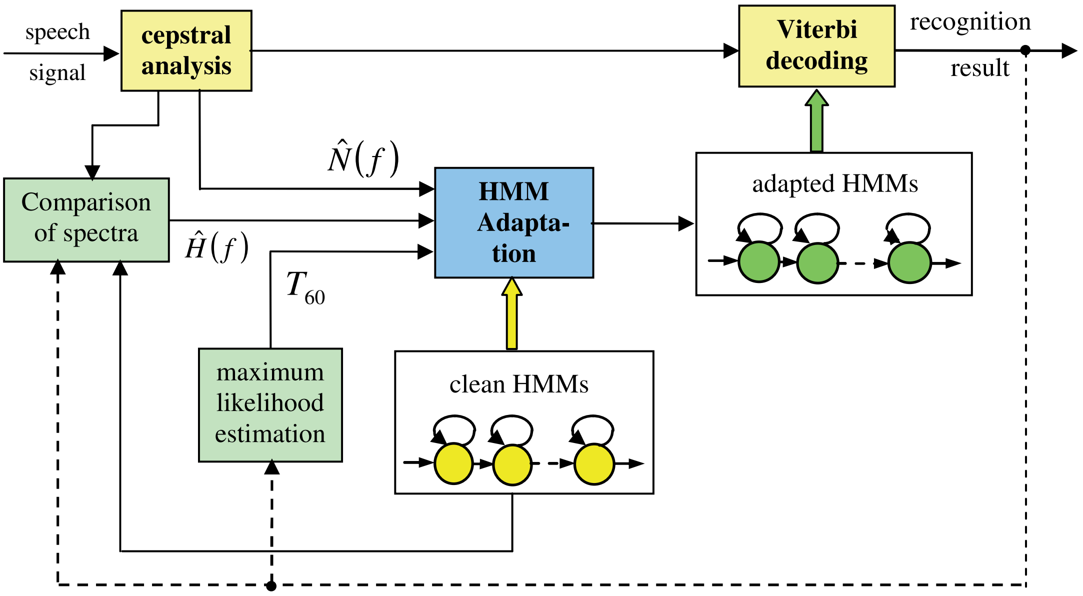

Figure 1: Robust recognition scheme based on HMM adaptation.

values within the states of the clean HMMs with the corresponding spectra of the input signal. The mapping of the feature vectors to the HMM states is taken from the Viterbi decoding. We estimate the reverberation time also after each new input with an iterative procedure by slightly modifying the earlier estimated reverberation time and creating a new set of adapted HMMs. We look for the maximum likelihood in the Viterbi decoding [4].

The adaptation is individually done at each speech input when the beginning of speech is detected. The logarithmic energy and the cepstral coefficients are calculated as acoustic parameters every 10 ms. We adapt the means of the energy and the cepstral coefficients as well as the means of the corresponding Delta and Delta-Delta coefficients as they occur as parameters in each HMM state. Without presenting all details of the processing [4], the adaptation is based on a transformation of the cepstral coefficients C0 to C12 back to the linear Mel spectrum as they occur as means of the Gaussian distributions in each HMM state S_i (i=state index).

$$\{\overline{C}_0, \overline{C}_1, \cdots, \overline{C}_{12}\} \overset{IDCT}{\Rightarrow} \{\log(S_{mel}(1)), \cdots, \log(S_{mel}(24))\}$$
$$\overset{EXP}{\Rightarrow} \{S_{mel}(1), S_{mel}(2), \cdots, S_{mel}(24)\}$$

Each Mel spectral component Smel(k) (k=1,…,24) is adapted for each state Si by

- adding a certain amount of acoustic energy from previous states according to the smearing effect of reverberation,
- multiplying with a spectral modification characteristic H_{est} and
- adding the estimated noise spectrum N_{est}.

$$\hat{S}_{mel}(S_i,k) = \Big[\alpha(i,i) \cdot S_{mel}(S_i,k) +$$
$$\sum_{j=1}^{i-1} \alpha(j,i) \cdot S_{mel}(S_j,k)\Big] \cdot H_{est}(k) + N_{est}(k)$$

The coefficient α(j,i) defines this part of the acoustic energy that occurs in the segment of state S_i due to the reverberation and has been emitted in the corresponding speech segment of state with index j. We assume that the energy contour exponentially decays when turning off the acoustic excitation in a room. The logarithm of the adapted Mel spectrum is transformed again to the cepstral domain to determine the adapted cepstral means.

3 Recognition of noisy speech

Word error rates are presented in Table 1 for the recognition of different data sets containing distorted versions of the TIDigits as defined in the Aurora-5 setup [8]. We focused on distorted speech signals containing additive background noise only (referenced by the terms car and interior). But we included also two sets of speech signals containing only the effects of a hands-free recording in the reverberant environments of an office and a living room without noise in the background (HF-office and HF-living).

Results are shown for recognizing digits in the noisy car environment at different SNRs and in the presence of noises as they occur in typical application scenarios inside rooms (“interior” noise). Comparing the results of applying the adaptation or applying the robust feature extraction in combination with HMMs trained on clean data only, the HMM adaptation provides a fairly high performance for all noise conditions. The usage of the robust features offers a better performance especially at low SNR.

Furthermore, we present results for the usage of robust features in combination with HMMs trained on the features of noisy speech signals. It is well known that we can achieve a high performance when applying HMMs that have been trained on speech signals recorded in the acoustic environment where the recognition system will be operating. The prerequisite is the knowledge of the

acoustic condition	HMM adapt.	robust features		
		clean HMM	multi-cond. HMM (car)	multi-cond. HMM (interi.)
Clean	0,55	0,54	6,03	3,39
Car 15dB	1,15	1,23	1,16	1,19
Car 10dB	2,11	2,24	1,46	1,58
Car 5dB	5,83	5,87	2,73	3,01
Car 0dB	18,73	17,45	7,38	7,76
Interior15dB	2,14	2,92	2,89	2,43
Interior10dB	4,89	6,36	5,39	4,66
Interior 5dB	13,81	15,75	11,69	10,08
Interior 0dB	38,27	37,71	26,68	25,88
HF-office	2,63	4,36	9,78	6,92
HP-living	5,59	9,38	13,02	10,26

Table 1: Word error rates (%) for different noise conditions applying different recognition schemes.

expected noise scenario. But this is not known in advance for a lot of applications like the recognition with a mobile phone.

To avoid the dependency on an individual noise condition we took a set of 27 noise recordings from different environments like inside cars under various conditions, inside buses and trains and at public places like train stations, airports, shops and restaurants. We did not want to train an individual set of HMMs for each noise condition. Thus, we clustered the 27 noise signals into two categories based on a spectral distance measure. We found one class containing all car noises and some noises recorded in buses and trains. The second class contains all other noise signals mainly recorded at public places inside rooms. We refer to this class by the term “interior” noise.

We trained the parameters of HMMs by creating a set of noisy signals from the complete set of TIDigits designated for training. It consists of about 8600 utterances with a total of about 28000 digits. All noises in each class are equally applied. 50% of the noisy training data contain the noise at a SNR of 5 dB, 40% at a SNR of 10 dB and the remaining 10% at a SNR of 0 dB. Thus, we created two sets of so called multi-condition HMMs. The error rates in table 1 show a considerable improvement for the recognition of noisy signals especially at low SNRs when applying the multi-condition HMMs in comparison to using the HMMs trained on clean data only. The disadvantage is a relatively low recognition performance on clean data. We could compensate this deterioration to some extent by using also some clean signals as part of the training set. But we would not achieve the performance of applying HMMs trained on clean data only. Furthermore, our intention is a combination of different recognition schemes so that we are interested in highest performance at each acoustic condition.

Looking at the results for the two sets of reverberant speech signals highest performance is achieved applying the HMM adaptation. This could be expected because the robust feature extraction does not include any processing to reduce the effect of reverberation where the adaptation has especially been designed to compensate the influence of a hands-free speech input.

4 Combining Robust Recognition Schemes

To combine the advantages of the different recognition schemes we set up the system shown in figure 2. Based on an estimation of the SNR we decide in case of exceeding a certain SNR threshold to apply the recognition including the adaptation of HMMs. A threshold of 15 dB led to the highest average performance for all conditions. We estimate the SNR by taking the logarithmic frame energy that is calculated as acoustic parameter for the recognition anyway. Using the decision of a voice activity detector (VAD) we calculate the level N of the noise as average energy over all segments marked as non-speech. The level S of speech is estimated as average over all remaining segments containing speech. The value of 15 dB is not directly comparable with the SNR of the noisy signals from the Aurora-5 data base. The SNR of the Aurora-5 data takes into account a spectral weighting that restricts the calculation of S and N to the frequency range of “old” telephony from about 300 to 3400 Hz. The energy coefficient in the feature extraction is calculated after applying the high-pass filtering of a preemphasis.

Furthermore, two spectral distances are estimated in case the estimated SNR is below the threshold. The distances are calculated between the estimated noise spectrum of the speech input and the average spectra of all noises signals in two classes. The noises in each class are the ones that have been assigned by the clustering process described in the previous section. The average spectrum with the lower distance decides which set of multi-condition HMMs is applied. The error rates are presented in Table 2 and Table 3 for the approach of selecting a recognition scheme based on the estimation of the SNR and the noise category. To compare the results with the performance of the individual recognition schemes the lowest achievable error rates for each acoustic condition are listed.

In case of the hands-free conditions the error rates are exactly the ones that are presented in Table 1 for the recognition applying the HMM adaptation. All reverberant signals are selected as speech with high SNR.

		car noise			
	clean	15dB	10dB	5 dB	0 dB
Lowest error rate (table 1)	0,54	1,15	1,46	2,73	7,38
SNR & noise type selection	0,53	1,12	1,56	2,82	7,40

Table 2: Word error rates (%) for car noise condition.

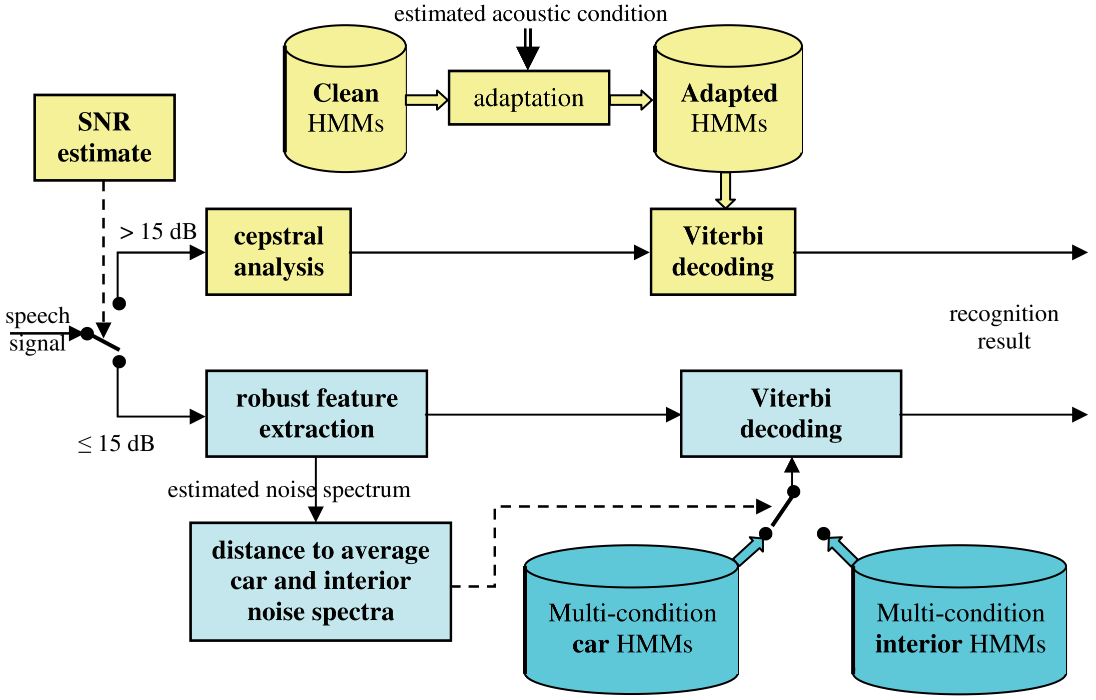

Figure 2: Combining different robust recognition schemes.

	interior noise			
	15 dB	10 dB	5 dB	0 dB
Lowest error rate (table 1)	2,14	4,66	10,08	25,88
SNR & noise type select.	2,22	4,68	10,21	23,97

Table 3: Word error rates (%) for interior noise condition.

In general, it can be seen that the selection process works almost optimal because we achieve a performance that comes close to or is even sometimes higher than the performance of the best recognition scheme in the specific acoustic condition. We are able to combine the advantages of the recognition scheme including the HMM adaptation and the scheme based on the usage of robust features and multi-condition HMMs. We achieve low error rates for clean as well as for noisy speech signals. Furthermore, the advantage of the adaptation scheme can be included to compensate the effect of reverberation.

5 Conclusions

A speech recognition setup is presented that is based on the combination of different recognition schemes. The estimation of the noise characteristics is taken to select the appropriate scheme for a specific acoustic condition. This setup causes almost no computational overhead due to the selection at the beginning of speech.

References

[1] Xu, H., Tan, Z.H., Dalsgaard, P., Lindberg, B., “Robust speech recognition based on noise and SNR classification – a multiple model framework“, *Interspeech*, 2005.

[2] H.G. Hirsch, A. Kitzig, “Improving the robustness with multiple sets of HMMs”, *Interspeech*, 2009.

[3] H.G. Hirsch, A. Kitzig, “Robust Speech Recognition by Combining a Robust Feature Extraction with an Adaptation of HMMs”, *ITG Fachtagung Sprachkommunikation*, Bochum, 2010.

[4] H.G. Hirsch, F. Finster, “A new approach for the adaptation of HMMs to reverberation and background noise”, *Speech Communication*, Vol.50, pp. 244-263, 2008.

[5] H.-G. Hirsch, C. Ehrlicher. Noise estimation techniques for robust speech recognition, *ICASSP*, 1995.

[6] C. Breithaupt, T. Gerkmann, R. Martin, “Cepstral smoothing of spectral filter gains for speech enhancement without musical noise”, *IEEE Signal Processing Letters*, 2007.

[7] ETSI standard document. Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Advanced Front-end feature extraction algorithm; Compression algorithm. *ETSI document ES 202 050 v1.1.3* (2003-11), Nov. 2003.

[8] H.G. Hirsch, “Aurora-5 experimental framework for the performance evaluation of speech recognition in case of a hands-free speech input in noisy environments”, Online: http://aurora.hs-niederrhein.de, data available from ELDA: http://www.elda.org, 2007.