INTERSPEECH 2009 BRIGHTON

Improving the robustness with multiple sets of HMMs

Hans-Günter Hirsch, Andreas Kitzig

Department of Electrical Engineering and Computer Science,
Niederrhein University of Applied Sciences, Krefeld, Germany
hans-guenter.hirsch@hs-niederrhein.de, andreas.kitzig@hs-niederrhein.de

Abstract

The highest recognition performance is still achieved when training a recognition system with speech data that have been recorded in the acoustic scenario where the system will be applied. We investigated the approach of using several sets of HMMs. These sets have been trained on data that were recorded in different typical noise situations. One HMM set is individually selected at each speech input by comparing the pause segment at the beginning of the utterance with the pause models of all sets. We observed a considerable reduction of the error rates when applying this approach in comparison to two well known techniques for improving the robustness. Furthermore, we developed a technique to additionally adapt certain parameters of the selected HMMs to the specific noise condition. This leads to a further improvement of the recognition rates.

Index Terms: robust speech recognition, HMM adaptation

1. Introduction

A number of robust feature extraction schemes have been developed, e.g. [1],[2], that lead to an improved recognition in acoustic situations where background noise is present or the speech spectrum is modified by unknown frequency responses. Alternatively, several HMM adaptation techniques, e.g. [3],[4],[5], are available to adapt the parameters of HMMs to the acoustic scenario. Furthermore, other techniques like the reconstruction of missing features allow an improved recognition of noisy speech data.

Nevertheless, best recognition rates are still achieved by training the reference patterns on data that have been recorded in exactly the same acoustic environment like the one where the recognition system is applied, e.g. [6],[7]. But this approach is based on the knowledge of the acoustic scenario and on the assumption that the acoustic conditions will not change while using the recognizer. An example for a fairly stationary acoustic scenario is the application of a recognition system inside a car.

In case the recognition system will be applied in a number of different acoustic environments the system can be trained in a so called multi-condition mode, e.g. [8]. Speech data are used for training that have been recorded under all known conditions. Usually, a recognition performance can be achieved with this approach that is somewhere in between the performance of a system trained on clean data only and a system trained on the specific noise condition. As in case of training with data from a single noise condition, the noise scenarios for the multi-condition training have to be known in advance.

We investigated the approach to train several sets of HMMs that represent a broad range of noise conditions. Based on recordings of about 30 typical background noise situations we clustered these noise recordings in 5 categories. For each cluster we created noisy speech data at a SNR of 5 dB for training an individual set of HMMs. At each speech input this set of HMMs is selected that matches best the actual acoustic condition. This is done by comparing the pause segment before the speech starts with the pause HMMs of the 5 noise clusters. Additionally, the cepstral and energy mean parameters of the selected HMMs can be adapted to the specific noise condition at each speech input. The adaptation approach has been derived from [5].

In the next section we will present the motivation for our work by comparing the results of two robust recognition systems against the performance of training a system on specific noise conditions. Then, we introduce our approaches to use multiple sets of HMMs and as additional processing step to adapt the cepstral and energy parameters of the HMMs to a specific noise condition. The results of different recognition experiments demonstrate the applicability of the new approaches.

2. Motivation

To motivate the investigations presented in this paper we compared the performance of two robust recognition systems trained on clean data only with the recognition rates that can be achieved by training a system with speech data of a specific noise situation at a certain SNR (signal-to-noise ratio).

The speech data of two test conditions were taken from the experimental framework called “Aurora-5” [9]. The Aurora-5 set-up contains distorted versions of the TIDigits data. Besides the presence of background noise Aurora-5 includes also test sets that simulate the recording in hands-free mode and the speech transmission in cellular networks. For our investigations we looked at two conditions where “car” noise was added in combination with a G.712 filtering and where “interior” noise was added to the designated TIDigits test data. G.712 is an ITU recommendation to simulate the bandpass characteristics of telephone devices. Aurora-5 extends the earlier “Aurora-2” framework to a more complex but also more realistic set-up. The speech data for each noise condition have been created by randomly adding one out of a whole set of noise signals that represent the specific situation. For Aurora-2 only one noise recording was taken.

The word error rates are presented in tables 1 and 2 when recognizing the two versions of the TIDigits with car noise or with interior noise added at 4 SNRs. We applied the robust feature extraction scheme as it has been standardized by ETSI [1]. Furthermore, we looked at a second robust system based on an usual cepstral analysis without additional processing blocks to create robust features but including an adaptation of the HMM parameters to the specific background noise of each utterance [5]. Both recognition systems are trained on clean data only. Each test condition contains the recognition of the designated TIDigits test set containing 8700 utterances of adult speakers with a total of about 28000 digits.

SNR/dB	15	10	5	0
robust features	1.3	2.4	5.9	16.0
HMM adaptation	1.2	2.1	5.8	18.7

Table 1. *Word error rates (%) for "car" noise.*

SNR/dB	15	10	5	0
robust features	2.6	5.7	14.4	35.3
HMM adaptation	2.4	4.9	13.9	38.3

Table 2. *Word error rates (%) for "interior" noise.*

Both recognition systems show a comparable performance. The error rates are higher for the condition of interior noise due to the more "non stationary" noise characteristics in comparison to the fairly stationary car noise situation. Interior noise includes recordings, e.g. in a restaurant, in an office, e.t.c.

To determine the recognition performance with HMMs that have been trained on data from a specific noise condition we created noisy versions of the designated TIDigits training data for each noise condition at the desired SNRs [10]. We applied a usual cepstral analysis scheme where the 12 Mel frequency cepstral coefficients C1 to C12 and the logarithmic energy as well as the Delta and Delta-Delta features are extracted. Neither the feature extraction does contain specific blocks for improving the robustness nor the HMMs are adapted to the noise condition. The word error rates are presented in table 3 for all possible combinations of training and test conditions in case of car noise.

		SNR/dB of test data			
		15	10	5	0
SNR/dB of training data	15	**0.8**	2.0	8.5	32.5
	10	0.9	**1.4**	4.4	18.5
	5	2.3	1.8	**3.1**	10.5
	0	11.7	6.0	4.7	**8.9**

Table 3. *Word error rates (%) for "car" noise.*

The highest performance can be achieved when training the recognition system on the specific noise and SNR condition. Similar results occur when running these experiments on data with interior noise.

3. Recognition with multiple HMM sets

We took the results presented in the previous section as motivation to set up a recognition system with 5 sets of noisy HMMs as shown in figure 1.

As output of the analysis block the 12 Mel frequency cepstral coefficients C1 to C12 and the logarithmic energy are extracted from 25 ms segments of speech. The creation and the selection of one of the 5 HMM sets are described in the two following subsections. Each set contains 22 gender dependent HMMs for the 11 digits including the two versions "zero" and "oh" for the digit "0". Each HMM consists of 16 states and each state describes the occurrence of each acoustic parameter by a mixture of 2 Gaussian distributions. The pauses containing the background noise are modeled by a one state model with a mixture of 8 Gaussian distributions. The training is done with the corresponding tools of HTK [11].

The recognition of the digit sequences is based on the usual approach to calculate the probabilities that the observed sequence of feature vectors can be generated from a sequence of HMM states by means of the Viterbi algorithm. The complete recognition scheme as shown in figure 1 is implemented as modules in Matlab.

3.1. Creation of HMM sets

To create the 5 sets of noisy HMMs we took a collection of about 30 noise signals. These noise recordings reflect the typical scenarios where speech recognition systems might be applied. The noise signals were recorded

- inside different cars,
- inside buses, different types of trains,
- at public places like airports, train stations, exhibition halls, e.t.c.,
- in restaurants, offices, e.t.c.

To avoid the huge effort of creating an individual set of HMMs for each noise condition and based on the knowledge that some noise signals have similar spectral characteristics we clustered the noise signals in 5 categories. The clustering is based on estimating the spectral similarity between all noise signals. All signals are analyzed with the short-term cepstral analysis of the recognition system. Gaussian distribution density functions are estimated for each cepstral coefficient and each noise signal. A measure describing the similarity between 2 noise signals is derived from the comparison of the corresponding distribution functions. All similarity measures are taken as input for a k-means clustering. Looking at the results we find 5 categories as listed in table 4.

Cluster	Noise Signals
1	inside cars
2	at public places like airports, restaurants, …
3	inside cars
4	inside buses, trains
5	at public places like train stations, on the street, …

Table 4. *Noise categorization.*

We created noisy training data for each cluster by randomly selecting one of the corresponding noise signals and adding a randomly selected segment to each clean training utterance at

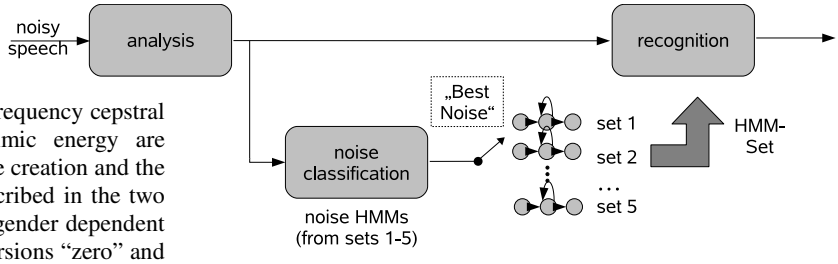

Figure 1: *Speech recognition with multiple HMM sets.*

a SNR of 5 dB. We focused on the SNR of 5 dB because we observed a fairly good performance over the whole SNR range in table 3 with HMMs trained at 5 dB. Furthermore, we will present a scheme for adapting the parameters of HMMs to an individual noise condition. Especially, the adaptation of the energy parameter should compensate a possible difference in SNR between training and test data.

3.2. Selection of HMM set

During recognition one of the 5 HMM sets is selected. The selection is based on calculating the probabilities that the noise segment at the beginning of each utterance can be generated by one of 5 Gaussian mixture models (GMMs). The GMMs are the single state HMMs that are determined in the training phase to model the pause containing the noise characteristics in each cluster. The selection is finished at the beginning of speech so that this approach can be applied in a real-time recognition system without causing any delay. The probabilities are calculated by looking at the 12 static cepstral coefficients C1 to C12 only. Thus, the selection is independent of the noise energy.

The results of the noise selection process are shown in figures 2 and 3 for the two Aurora-5 test sets with car respectively interior noise added at a SNR of 5 dB. The figures show the number of utterances from the total of 8700 that are mapped to each of the 5 clusters.

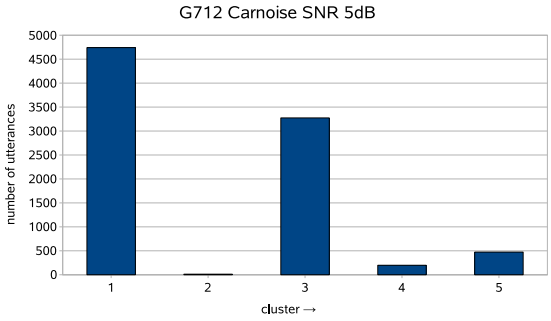

Figure 2: *Results of noise classification.*

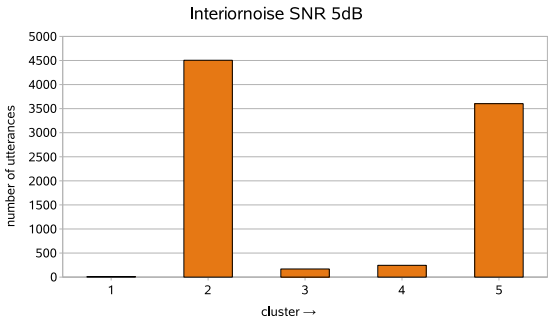

Figure 3: *Results of noise classification.*

The results proof that the mapping works quite well. HMM sets 1 and 3 are the ones that have been trained on car noise data. Sets 2 and 5 have been trained on speech data containing the different types of interior noise. Results are only shown for the SNR of 5 dB because the classification to the noise clusters is almost identical for the SNRs of 0, 10 and 15 dB. This is due to using the ceptral coefficients C1 to C12 only for estimating the similarity of the spectral shapes without taking into account the energy of the noise segment.

3.3. Recognition results without adaptation

The word error rates are shown in figures 4 and 5 when selecting a set of HMMs for the recognition of each individual utterance as described in the previous subsection. No further adaptation is applied on the selected HMMs.

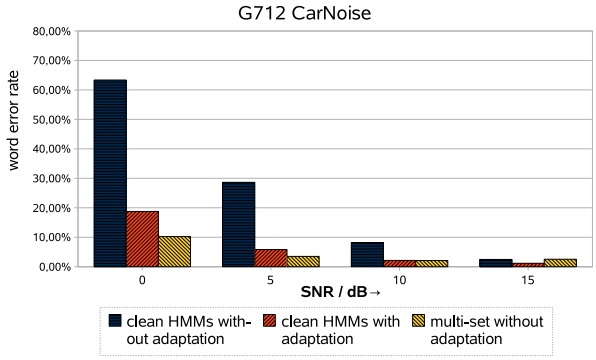

Figure 4: *Word error rates with multiple HMM sets.*

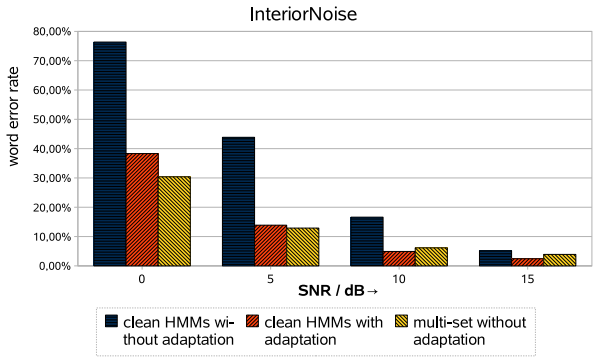

Figure 5: *Word error rates with multiple HMM sets.*

The results for the approach of using multiple sets of HMMs are compared against the recognition with a single set of HMMs that has been trained on clean data only. The error rates are presented for the single set of HMMs without and with an adaptation [5] to the specific noise condition. A considerable reduction of the error rates can be observed for the low SNRs of 0 and 5 dB due to training the sets of noisy HMMs on data with an SNR of 5 dB.

3.4. HMM adaptation

We derived a new approach from an existing adaptation scheme [5]. The earlier approach contains an adaptation of the acoustic parameters contained in HMMs that have been trained on clean data. We modified the existing scheme to adapt the means of the acoustic parameters as contained in the selected set of noisy HMMs to the specific noise characteristics of each utterance. One has to keep in mind that each set contains the characteristics of several slightly different noise signals at a SNR of 5 dB. The intention is an adaptation of the spectral and energy parameters to the specific noise condition.

We want to describe the basic idea without presenting all mathematical details. Estimating the noise spectrum and the noise energy from the pause segment at the beginning of each utterance we compare these estimates with the corresponding spectrum and energy of the selected pause HMM. The spectrum and the energy as contained in the pause HMM are derived from the average of the cepstral and energy mean

parameters. The differences in spectrum and energy are estimated by comparing the noise spectra and the noise energies of the speech input and the pause HMM. These differences are taken to adapt the means of the static cepstral and energy coefficients in each HMM state. The spectral adaptation is done in the linear Mel spectral domain by transforming the cepstral coefficients back and forth again.

Furthermore, an unknown frequency characteristic is estimated by comparing the short-term spectra of the input utterance to the corresponding spectra that can be derived after the recognition from the state sequence with highest probability. This is based on the legal assumption that the frequency characteristic usually changes only slowly from one speech input to the next one. The result of the adaptation process is visualized in figures 6 and 7. The figures show spectrograms as they can be derived from the average cepstral means that are contained in a HMM [5].

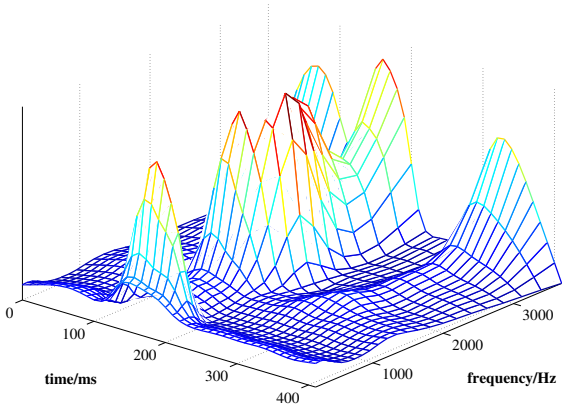

Figure 6: *Spectrogram of a noisy HMM for the digit "6".*

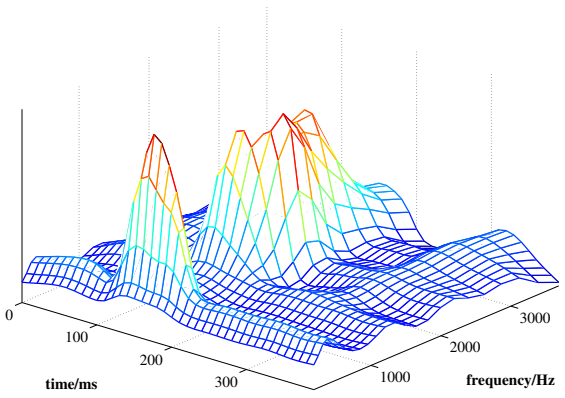

Figure 7: *Spectrogram of the adapted HMM.*

Figure 6 shows the spectrogram of a selected noisy HMM for the digit "6". The formant spectrum of the vowel "i" gets visible in the middle of the word where the fricatives at the beginning and at the end have their energy at higher frequencies. The average noise spectrum as contained in the training data can be seen when looking at the short-term spectrum at the end of the word. In case of a clean HMM the spectrum at the end would take values close to zero.
Comparing the noise spectra in figures 6 and 7 at the word endings we observe in figure 7 the adaptation to the specific spectral characteristics of the individual input utterance. The speech of this individual utterance has also been filtered with the G.712 bandpass characteristic. The adaptation to the bandpass characteristic is visible as attenuation of the fricatives' spectral features at higher frequencies.

3.5. Recognition results with adaptation

The word error rates are shown in figure 8 when applying the additional adaptation on the speech data distorted with car noise. We observe a further improvement of the recognition by adapting the HMMs of the selected set to the specific noise condition. Similar improvements occur in the case of interior noise.

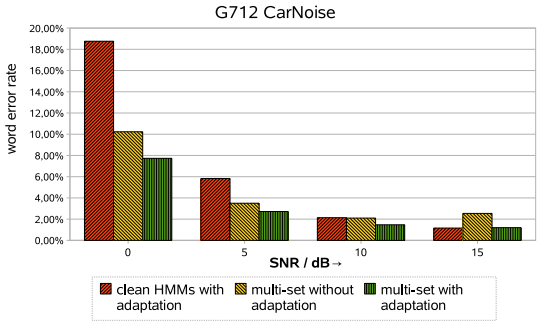

Figure 8: *Error rates with additional adaptation.*

4. Acknowledgements

The authors would like to thank the German ministry of education and research (BMBF) for supporting this work within the program FHProfUnd.

5. References

[1] 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, Nov. 2003.

[2] Gadrudadri, H., Hermansky, H., Morgan N. et. al., "Qualcomm-ICSI-OGI features for ASR," Proc. of ICSLP, pp. 21-24, 2002.

[3] Leggeter, C.J. , Woodland, P.C., "Maximum likelihood linear regression for speaker adaptation of continuous density Hidden Markov Models," Computer Speech and Language, Vol.9, pp. 171-185, 1995.

[4] Gales, M.J.F., Young, S.J., "Robust continuous speech recognition using parallel model combination," IEEE Trans. Speech and Audio Proc., Vol.4, pp.352-359, 1996.

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

[6] Xu, H., Tan, Z.H., Dalsgaard, P., Lindberg, B.: "Robust speech recognition based on noise and SNR classification – a multiple model framework", Proc. of Interspeech, 2005.

[7] Beritelli, F., Serrano, S., Russo, A., "A speech recognition system based on dynamic characterization of background noise", in IEEE Int. Symp. on Signal Proc. and Information Tech., 2006.

[8] Hirsch, H.G., Pearce, D., "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," Proc. of the ISCA workshop ASR2000, Paris, France, 2000.

[9] Hirsch, H.G., "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.

[10] Hirsch, H.G., "FaNT – Filtering and Noise adding Tool", Online: http://dnt.kr.hs-niederrhein.de/download.html, 2003

[11] Young, S. et.al., "The HTK book", version 3.3, Online: http://htk.eng.cam.ac.uk, 2005.