Applying the Aurora Feature Extraction Schemes to a Phoneme Based Recognition Task

Harald Finster, Hans-Günter Hirsch

Department of Electrical Engineering and Computer Science
Niederrhein University of Applied Sciences, Krefeld, Germany
{harald.finster, hans-guenter.hirsch}@hs-niederrhein.de

Abstract

The robustness of the ETSI (European Telecommunication Standards Institute) standardized feature extraction schemes is investigated for phoneme based recognition tasks of German speech data. The recognition tasks are an isolated command word recognition and the recognition of connected digits. The motivation of this work is the easy extensibility of a whole word recognition system by allowing also the recognition of phoneme based word HMMs (Hidden Markov Models). The recognition performance has been determined for different numbers of HMM states and different numbers of Gaussians per state. It turns out that fairly high recognition rates can be achieved also for noisy data when applying the second robust ETSI frontend.

1. Introduction

The authors have developed a speech dialogue system that is used for information services over the telephone [1]. The speech recognition of this system is based on whole word HMMs so that isolated or connected words can be recognized. One key feature of the system is a fairly high robustness to background noise and unknown frequency characteristics due to the use of a PMC (parallel model combination) adaptation scheme [2].

It has been investigated how the range of possible applications can be increased by introducing a phoneme based recognition. In many cases new applications with mostly natural dialogue require command words that do not exist in available databases. This causes the need of collecting new training data in case of using HMMs for whole words only. This time consuming task can be avoided by training a whole set of phoneme HMMs and concatenating word models from a sequence of phoneme models. The goal of this work is not the development of a phoneme based recognizer for continuous speech but the easy extensibility of a word based recognizer by e.g. command words that do not exist in available databases. It is investigated how the important HMM parameters, especially the number of states and the number of Gaussians per state have to be chosen for achieving a good recognition performance.

The acoustic features for the recognition are extracted with a cepstral analysis scheme. To make the results of this work more globally utilizable, the experiments have been run with the two feature extraction schemes that have been standardized by the ETSI Aurora working group [3], [4].

Both standardized frontends are shortly described in the next section. The modelling of phonemes as monophones or triphones with HMMs is presented in section 3. The recognition experiments and the achieved results are described in section 4.

2. ETSI frontends

Two feature extraction schemes have been standardized by the ETSI working group called “Aurora”.

The first standard [3] is an usual cepstral analysis scheme where 13 cepstral coefficients are determined based on a Mel filterbank with 23 channels. As an additional parameter the logarithmic frame energy is calculated in the time domain. The cepstral coefficients C_1 to C_{12} without the energy coefficient C_0 but including the frame energy are used as acoustic parameters for these investigations. Delta and Delta-Delta parameters are calculated as it has been realized in the HTK software [5]. Thus a feature vector consists of 39 components. We refer to this frontend by the abbreviation ETSI-1 in this paper.

The second standard [4] is also a cepstral analysis scheme that has been extended by additional signal processing blocks to reduce the influence of stationary background noise and of unknown frequency characteristics. This frontend contains also a slightly different method for calculating the Delta and Delta-Delta parameters. Again a feature vector consists of 39 parameters. The abbreviation ETSI-2 is used as reference to this analysis scheme throughout this paper.

The short-term analysis window is shifted by 10 ms in both standards so that 100 vectors per second are created as output.

3. Phoneme models

Phonemes are modelled by HMMs where the number of states and the number of Gaussians per state are the two main parameters that have been investigated within the scope of this work. The HMMs are determined from speech which has been recorded over the telephone, mainly in fixed networks. Telephone speech is taken for the training because of the target application in telephone based information services. A database with recordings of about 1000 speakers has been available for the training. In total this comes up to about 48 hours of speech for training. This telephone data represents the usual acoustic environment in fixed networks without containing a lot of noise.

We created a time labelling on the basis of phonemes for all used recordings. A Sampa like notation has been applied for the description of the whole set of phonemes. The labelling has been realized by initially training a subset of phonemes with a small amount of hand labelled spellings. By processing more and more data through several stages and increasing the number of phonemes at the same time, reliable labels could be created by applying a forced alignment technique.

We investigated the two modelling approaches of describing the phonemes as monophones or as triphones. In case of triphones we considered only five classes of phonemes for the description of the preceding and the succeeding phoneme. The members of the five phoneme classes are listed in table 1.

Table 1: Phoneme classes for triphone modelling

Class	Description	Members
V	vowels	a: e: i: o: u: a E I O U y: oe Oe ae ar an E: @
F	Fricatives	h f v z x s S C
N	Nasals	m n l r j Y N Z On
S	Plosives	t g p d k b
D	Diphones	aI OY aU

The training of the HMMS has been performed with the HTK package [5] by applying the tool for creating initial models first and by refining the parameters with the two tools for a Baum-Welch reestimation on isolated units first and on embedded units later on.

Phoneme models with 3 states do not allow the skipping of a single state where models with more than 3 states allow the skipping over one state.

4. Recognition Experiments

The main target of these investigations is the extensibility of a robust word based recognizer by HMMs, that consist of concatenated phoneme models. Because of this, two recognition tasks have been considered. The first one is a simple isolated command word recognition. The vocabulary consists of 52 German command words. 7650 utterances are used in total which are also part of the training database. The abbreviation CMD will be used throughout this paper to refer to this task.

The second task is the recognition of connected German digits. Three sets of test data have been applied from three different databases.

The first set **S1** consists of about 2100 utterances containing about 10000 digits in total. This data has also been used for the training.

The second set **S2** contains about 1900 utterances with a total of about 7000 digits. This data has not been recorded over telephone but with a close talking microphone in a quiet environment and a direct connection to the recording system. The database contains recordings from 90 speakers that are different from the speakers in set S1. The speakers represent a wide range of German dialects. The results on this data can be used for showing the influence of a different acoustic environment on the recognition performance.

The third set **S3** of connected digits contains a few recordings of the German SpeechDatCar data collection. This subset has also been used for the experiments of the ETSI standardization activities. This data has been recorded in the noisy car environment. The set contains about 3000 utterances with a total of about 16500 digits. This data is intended to show the influence of a noisy background situation.

An overview about all speech data that has been applied for these investigations is given in figure 1.

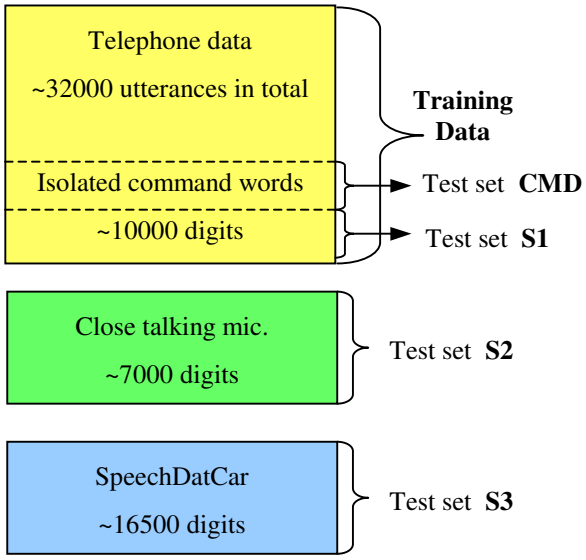

Figure 1: Overview about training and test data.

Most of the experiments are based on the use of monophones. Only a few results are presented to show and proof the gain of using triphone models instead.

4.1. Monophone models

Tables 2 and 3 contain the error rates for recognizing the isolated command words (CMD) when applying both ETSI frontends and the HMMs from monophone models. Tables 4 and 5 contain the error rates for the more complex task of recognizing digit set S1. Error rates include insertion and deletion errors.

As expected error rates are smaller for the easier task of a isolated word recognition. Comparing both frontends the recognition performance is slightly higher for the robust ETSI-2 frontend in general.

The results in all tables show an increasing performance for an increasing number of Gaussians per state where the gain is getting smaller for a higher number of Gaussians. Doubling the number of Gaussians comes along with also doubling the demand on memory and doubling the computational expense.

Table 2: Word error rates (%) for isolated command word recognition (CMD) with ETSI-1 frontend

Number of states	Number of Gaussians per state				
	1	2	4	8	16
3	10,4	5,4	4,6	3,5	2,9
5	8,1	5,5	4,2	3,0	2,7
7	5,5	3,5	2,4	1,9	1,7

Table 3: Word error rates (%) for isolated command word recognition (CMD) with ETSI-2 frontend

Number of states	Number of Gaussians per state				
	1	2	4	8	16
3	9,9	4,9	4,0	2,8	2,7
5	7,1	4,3	3,5	2,8	2,4
7	4,8	3,1	2,2	1,5	1,3

Table 4: Word error rates (%) for connected digit recognition (set S1) with ETSI-1 frontend

Number of states	Number of Gaussians per state				
	1	2	4	8	16
3	19,2	18,2	14,0	12,5	11,2
5	22,0	16,7	14,9	12,8	11,8
7	11,6	7,9	6,2	5,1	4,4

Table 5: Word error rates (%) for connected digit recognition (set S1) with ETSI-2 frontend

Number of states	Number of Gaussians per state				
	1	2	4	8	16
3	20,2	16,3	11,7	9,6	10,7
5	20,3	14,3	12,5	10,3	8,9
7	9,4	6,4	4,7	4,3	3,8

The error rates decrease for an increasing number of HMM states. Especially for the more complex task of a connected digit recognition there can be seen a considerable gain when applying models with 7 states in comparison to models with 5 states. We ran a few experiments with 9 states per HMM. It turned out that we have already reached the status of a certain saturation with 7 states because no further improvement could be obtained with 9 state HMMs. We achieved e.g. an error rate of 1,6 % on the command word task and an error rate of 3,9 % on the connected digit set S1 with HMMs with 9 states and 16 Gaussians per state when applying the ETSI-2 frontend.

It has to be considered that these results have been achieved for recognizing fairly clean data that has also been used for training. They are intended to give a first overview about the principal behaviour in dependency of the chosen HMM structure. More interesting is the recognition of speech data, that has been recorded in a different acoustic environment and that has not been used for training. For better readability further recognition results are presented as graphs instead of tables.

Word error rates are shown in figures 2 and 3 for the recognition of all data sets when applying the ETSI-1 respectively the ETSI-2 analysis scheme. The number of Gaussians is kept constant at a value of 16. The number of HMM states is varied between 3 and 9 states.

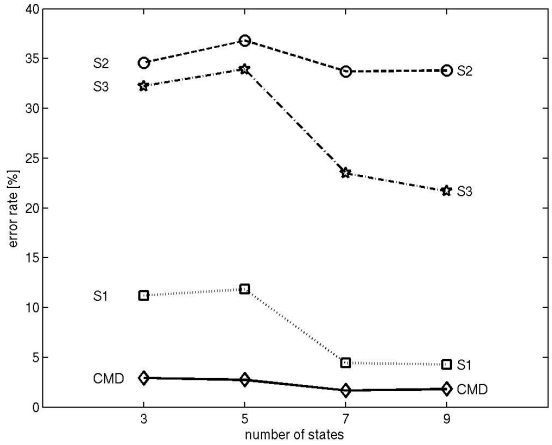

Figure 2: Word error rates for a varying number of HMM states (with 16 Gaussians) applying the ETSI-1 frontend.

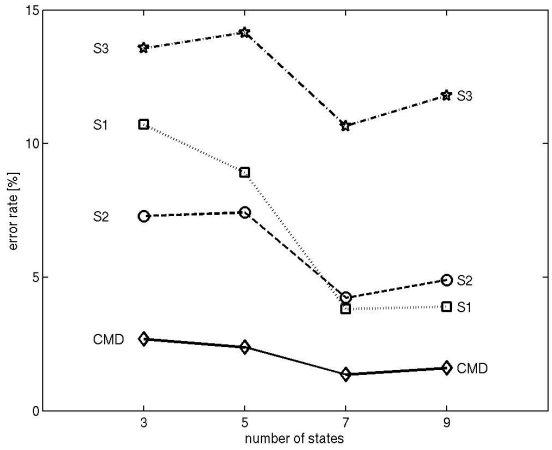

Figure 3: Word error rates for a varying number of HMM states (with 16 Gaussians) applying the ETSI-2 frontend.

As already mentioned, it can be seen, that no additional gain can be achieved by increasing the number of HMM states from 7 to 9. The results are even getting worse in most cases.

Comparing the results of figure 2 where the ETSI-1 frontend has been applied with the results of figure 3 where the robust ETSI-2 analysis scheme has been used, it turns out that there exist only small differences for the command word recognition and for test set S1, which have been part of the training data.

But the robustness of the ETSI-2 frontend gets obvious when comparing the results for test sets S2 and S3 that have been recorded in different acoustic environments than the training data. The data of set S2 differs mainly with respect to the frequency characteristic and it contains less noise than the data used for training. Set S3 has been recorded in the car environment so that this data is distorted by additive background noise. The error rates for the ETSI-1 frontend are in the range up to almost 40 % where the rates for the ETSI-2 frontend are below 15 %.

The recognition performance is almost the same for test sets S1 and S2 with 7 state HMMs in case of the ETSI-2 analysis scheme eventhough this data has been recorded in different acoustic scenarios. This impressive robustness can also be observed when looking at the curves in figure 4. Figure 4 contains the results for 7 state HMMs where the number of Gaussians is increased from 1 to 16. The curves for tests S1 and S2 are close together. It can be seen that no major additional gain can be achieved when increasing the number of Gaussians above a value of 8.

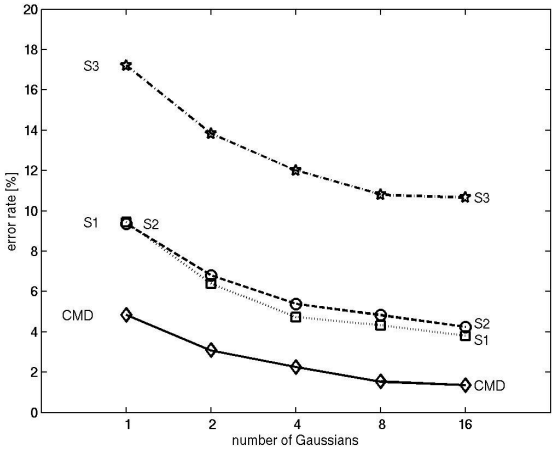

Figure 4: Word error rates for a varying number of Gaussians (with 7 state HMM) applying the ETSI-2 frontend.

4.2. Triphone models

The error rates are presented in figure 5 when applying triphone in comparison to monophone models as they have been decribed before. Again only the ETSI-2 frontend is applied and the HMMs consist of 7 states so that the results for the monophone models are identical to the curves shown in figure 4. Results for set S2 are not presented because they are almost the same as for set S1.

An expected gain can be achieved in all experiments because of the better phoneme modelling as triphones. The error rates are already quite low for a small number of Gaussians so that the relative improvement for an increasing number of Gaussians is not as high as for monophones.

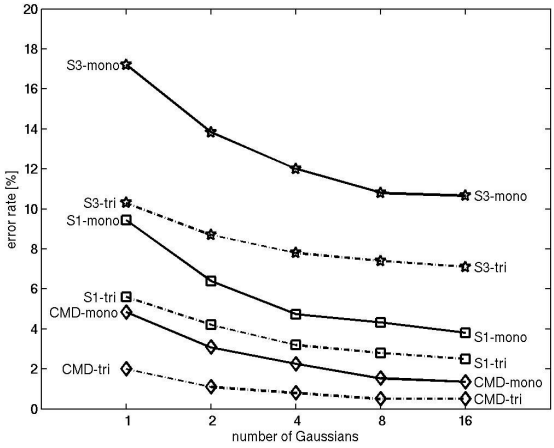

Figure 5: Word error rates for monophone ("mono") and triphone ("tri") models (with 7 state HMM) applying the ETSI-2 frontend.

5. Conclusions

Recognition results are presented for the phoneme based recognition of isolated command words and connected digits when applying the feature extractions schemes as they have been standardized by ETSI inside the Aurora working group. It turns out that a fairly high robustness can be achieved with the second robust ETSI frontend for recognizing speech data that has been recorded in different acoustic environments than the training data. The use of phoneme HMMs with 7 states and about 8 Gaussians per state offers already a good recognition performance without the need of high computational expense. An additional gain can be obtained by applying triphone models instead of monophone models.

6. References

[1] Hirsch, H.G., "HMM adaptation for applications in telecommunication", Speech Communication 34, pp. 127-139, 2001

[2] Gales, M.J.F., Young, S., "Robust speech recognition in additive and convolutional noise using parallel model combination", Computer, Speech and Language 9, pp. 289-307, 1995

[3] ETSI standard document, "Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithm", ETSI ES 201 108 v1.1.2 (2000-04), Apr. 2000.

[4] ETSI draft standard document, "Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Advanced Front-end feature extraction algorithm; Compression algorithm", ETSI ES 202 050 v0.1.0 (2002-04), Apr. 2002.

[5] Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P., "The HTK book – version2.2", Entropic, 1999.