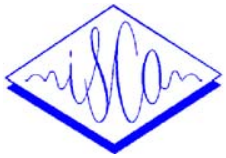

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

HMM ADAPTATION FOR TELEPHONE APPLICATIONS

Hans-Günter Hirsch
Ericsson Eurolab GmbH
Nordostpark 12, 90411 Nuremberg, Germany
hans-guenter.hirsch@eedn.ericsson.se
http://www.ericsson.de/ecc/fue/eurolab.html

ABSTRACT

The presence of background noise and the frequency response of a transmission line like in telephone applications have a major influence on the performance of speech recognition systems. An approach is presented in this paper to cope with both effects. It is based on an estimation of the stationary noise spectrum and an estimation of the mismatch between the frequency responses present during training and during recognition. These estimations are used in combination with the PMC scheme [1] to adapt the whole word HMMs for a speaker independent recognition of connected words. A considerable improvement can be achieved on recognizing distorted speech data. The technique is also used as part of a complete speech dialogue system over the telephone network where it could also proof its beneficial usability.

Keywords: robust speech recognition, HMM adaptation

1. INTRODUCTION

Robustness is the most important factor limiting the application of speech recognition in general and especially in telephone applications. Looking at the installation in a switch inside a telephone network the performance of a recognition system is mainly degraded by additive background noise and the transmission characteristics of the telephone network. This can be described in the linear spectral domain by

$$Y(f) = |H(f)|^2 \bullet S(f) + N(f) \qquad (1)$$

where S(f) is the power density spectrum of the clean speech and N(f) is the spectrum of all additive noise components. H(f) is the frequency response of the whole transmission system. Y(f) is considered as the input to the recognizer. N(f) and H(f) are assumed to be almost constant or only slowly changing over time. Two methods are presented in this paper to achieve good estimates for N(f) and H(f). Given those estimates it is possible to adapt the spectral parameters of HMMs based on the Parallel Model Combination (PMC) scheme [1]. Special care is taken of the ability to easily integrate these estimation techniques in a real-time recognition system.

2. THE RECOGNIZER

The recognition system used throughout this study is based on the representation of speech by cepstral parameters and the modeling of whole words by HMMs. The power density spectrum is calculated for 22 subbands in the Mel frequency range applying a FFT based analysis scheme. A feature vector consists of 24 components in total containing 12 Mel frequency cepstral coefficients (MFCCs) including the zeroth cepstral coefficient as representation of the short-term energy and the corresponding 12 Delta cepstral coefficients. Feature vectors are calculated every 10 ms analyzing a 25 ms window.
The recognition is based on the modeling of whole words by HMMs. The HMMs are simple left-to-right models and consist of 18 states per word where each state is described by a mixture of Gaussian distributions. Models are trained with the tools of the HTK software package [5].

3. THE ADAPTATION SCHEME

PMC is a well known technique for the adaptation of HMMs. Here it is combined with two processing schemes for the estimation of the stationary additive noise and the frequency response of the transmission line. The estimations are done for the 22 subbands in the Mel frequency range.
The cepstral means of each mixture density and each HMM state are transformed back to the linear spectral domain. The influence of the different noises is considered (see formula 1) by multiplying the Mel power density spectrum of the clean speech with the estimated frequency response and adding the estimated noise spectrum. Finally the modified spectrum is transformed again to the cepstral domain.

The estimation of the additive noise spectrum is based on the detection of the beginning of speech

during each input to the recognizer. The preceding stationary noisy segment is taken to calculate the average spectrum of this segment and taking it as estimate for the background noise. The speech detection is based on an analysis of the SNRs in the Mel frequency subbands [2]. Furthermore this detection is used to trigger the recognition process by starting the Viterbi match after the detection.

The processing to estimate the frequency response is based on the assumption that the frequency response is estimated after a speech input to the recognizer and after running the Viterbi match. The estimated response is used during the next speech input to the recognizer assuming a stationary behavior of the telephone line. Formula 1 can be rewritten it as

$$\left|\hat{H}(f)\right|^2 = \frac{Y_{long}(f) - \hat{N}(f)}{\hat{S}_{long}(f)} \qquad (2)$$

Assuming a constant frequency response H(f) and a constant noise spectrum N(f) during a speech utterance the short-term spectra Y(f) and S(f) can be substituted by their corresponding long-term spectra. The spectrum Y_{long} describes the noisy speech input and is calculated by transforming back the cepstral parameters to the spectral domain and summing up the short-term spectra over all segments which have been classified as speech by the matching of the recognizer. The spectrum S_{long} 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 represent the speech input. Only this density is considered in each state which has the smallest spectral distance to the input spectrum. Transforming back the corresponding cepstral means to the spectral domain the long-term spectrum can be calculated as sum over all corresponding HMM states. More details about the processing can be found in [3]. An advantage of this technique is that it can be done after a recognition and does therefore not cause any delay.

Analyzing the technique in further detail it can be noticed that it is not exactly the transmission characteristic which is estimated. Taking the spectral information contained in the HMMs means also a consideration of the frequency response which was present during recording the training data. Thus this processing estimates the whole mismatch between the frequency responses of training and recognition phase. Furthermore it includes also an adaptation to the speaker characteristics to some extent.

4. OFF-LINE RECOGNITION

Some recognition experiments are performed off-line as simulations using already available speech data bases. The speaker independent recognition of digit sequences or isolated digits is considered as recognition task. HMMs are created for the digits "1" to "9", "zero" and "oh" from the training set of the TIDIGITS data base [4]. The training set consists of 8623 utterances from female and male adult speakers containing 28329 digits in total. These data were recorded at high SNR. The original data are downsampled to 8 kHz for these investigations. One HMM is created for each digit describing each state by 4 Gaussian distributions. A simple pause model is generated from the training data consisting of 1 state with a mixture of 4 Gaussians.

4.1 TIDIGITS

A first set of recognition experiments is carried out on the test part of the TIDIGITS which consists of 8700 utterances with 28583 digits in total. A word error rate of **0.77%** can be achieved as baseline performance of the recognition system without applying any type of adaptation. The error rate includes substitutions, deletions and insertions. This corresponds to a string error rate of **2.37%**.

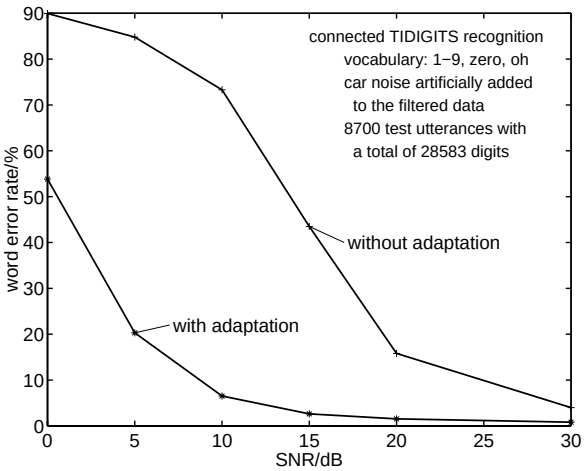

Figure 1. Word error rates for filtered noisy data

The TIDIGITS test data are artificially distorted to test the adaptation technique. At first the data are filtered with the typical frequency response of a telephone channel. Then car noise is added to the filtered data at different SNRs. The word error rates are shown in *Figure 1* without and with applying the adaptation technique. The results for the

filtered data without noise added are plotted at a SNR of 30 dB. A considerable improvement can be achieved by adapting the HMMs to the background noise and to the frequency characteristics of the telephone channel.

4.2 Bellcore digits

Besides the recognition of artificially distorted data the Bellcore digits are taken as test data which were recorded over telephone lines. The data partly contain background noise and the usual effects of different telephone lines and different handsets. The data base consists of 200 speakers uttering the 11 digits ("1" to "9", "zero" and "oh") as isolated words in real-life situations. Thus the task of recognizing isolated words is considered here. Still the same HMMs are used which are trained with the TIDIGITS as described above. Thus a situation is considered with a total mismatch between training and test data. The word error rates are listed in *Table 1* without and with applying the adaptation technique.

without adaptation	with adaptation
74.8 %	4.5%

Table 1. Word error rates for the Bellcore digits

Without adaptation the error rate is very high considering the 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 considerably decreases when applying the adaptation scheme. This indicates a good applicability of the described method in real-life applications.

5. APPLICATION IN THE TELEPHONE NETWORK

The recognition and adaptation scheme is integrated as part of a complete speech dialogue system which can be accessed over the public telephone network. The general structure of this dialogue system is shown in *Figure 2*.

A PC with Linux as operating system is taken as hardware basis. A passive ISDN card is used as connection to the telephone network. Looking at the software hierarchy you find a flexible dialogue controller on top. This controller is realized as a state machine and reads the definition of the dialogue from an ASCII file so that the dialogue can be easily designed and modified. Considering the application in the telephone network the two modules for speech recognition and speech output are the only modules for input and output of the dialogue system. The module for speech recognition is the same as used for the simulations with the only difference that the ALAW samples from the ISDN line are taken as input for the feature extraction. All software of this dialogue system is written in C. It is no problem to run the recognition scheme in real-time on the used PC with a clock frequency of 266 MHz.

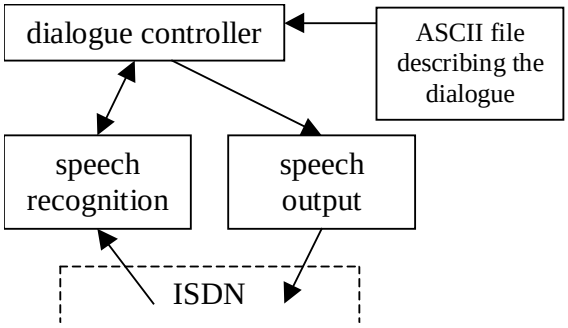

Figure 2.Speech dialogue system

A little demo is set up mainly aiming at the recognition of English digits and some command words like e.g. yes, no and help. Callers are asked to say e.g. the result of simple calculations or their phone number. Two gender dependent HMMs are created for each word from a speech data base containing the utterances of mainly German and Swedish people speaking English. These training data were recorded by using a close-by-talking microphone. They do not include the effects and limitations of telephone speech. Thus a mismatch is given between training and incoming speech data. HMMs consist of 18 states where each state is described by 2 Gaussians. The speech input to the recognizer is also stored on disk so that it can be used as training and test data later on. About 170 callers were recorded up to now. All over the recognition system shows a quite good performance. It is difficult to give any numbers for the performance. First of all there are different recognition tasks while using the system. The major task is the recognition of digit sequences. But at some points in the dialogue the recognition of isolated command words is only considered as task. Furthermore the system was called by some people which had fairly different accents in comparison to the speakers used for training the system. Another problem are non-cooperative speakers which just tried to fool the system.

To get some objective measures about the effect of the adaptation scheme all recorded utterances

containing only sequences of digits are taken as test data for an off-line recognition. These are about 1700 utterances from about 170 speakers containing in total 6976 digits. The utterances contain between 1 and 20 digits. The word error rates are shown in *Table 2* without and with applying the adaptation scheme.

without adaptation	with adaptation
7. 98%	3.47%

Table 2. Word error rates for own telephone data

These off-line experiments are done by using the HMMs as described above which had been trained on non-telephone data. Again a considerable improvement can be achieved by applying the adaptation technique. It has to be mentioned that also some garbage models are introduced to model instationary noises like e.g. breathing before and after the speech. Such garbage models help a lot to improve the recognition in real-life applications. Running the experiment without the adaptation and without the garbage models the word error rate increases to 13.73%.

To get insight into the estimation of the frequency response the estimated response is stored after each recognition while a caller is using the system. Such a sequence of consecutive estimates is presented in *Figure 3*. The estimated frequency response $|H(f)|^2$ as defined in formula 2 is shown versus time and Mel frequency. The time index describes consecutive recognitions which can be the recognition of a single command word or the recognition of a digit sequence.

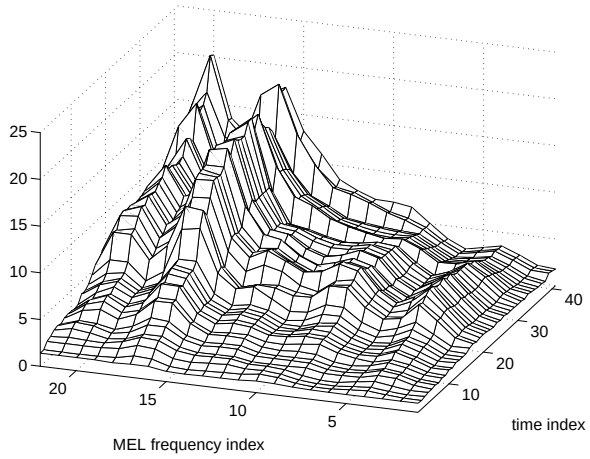

Figure 3. Consecutive estimates of $|H(f)|^2$

The figure shows that the system adapts fairly fast and remains then more or less stable. Looking at different speakers the estimated responses look similar in prinicipal but differ in detail caused by the specific characteristics of the telephone line and the characteristics of the speaker's voice.

The performance can be of course increased in principal by training or retraining the recognizer with telephone data. The adaptation technique helps a lot but it can not compensate all effects caused by the transmission over the telephone network. A retraining was actually done after recording approximately 80 people. At the beginning sometimes problems occurred recognizing some of the command words like "yes" and "no". This improved considerably after retraining the HMMs.

6. CONCLUSION

Two techniques are presented in this paper to estimate the stationary background noise and the mismatch between the frequency responses of training and recognition phase. These can be used in combination with the PMC scheme to adapt the HMMs of a speech recognition system. The usability of this approach is proofed by running recognition experiments on artificially distorted data as well as by integrating it as component of a speech dialogue system which can be accessed via the public telephone network.
Besides the considerable improvement of the recognition the main advantage of the presented estimation and adaptation scheme is the ability of an easy implementation in a real-time recognition system without the need of e.g. a high computational power and a separate preceding adaptation phase with special training data.

7. REFERENCES

[1] Gales, M.J.F (1997), Nice Model-Based Compensation Schemes for Robust Speech Recognition, *Proceedings of ESCA workshop Robust Speech Recognition for Unknown Communication Channels*, Pont-a-Mousson, France, pp. 55-64

[2] Hirsch, H.G., Ehrlicher, C. (1995), Noise Estimation Techniques for Robust Speech Recognition, *Proceedings of ICASSP*, Detroit, pp. 153-156

[3] Hirsch, H.G. (1997), Adaptation of HMMs in the Presence of Additive and Convolutional Noise, *Proceedings of IEEE workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, USA, pp. 412-419

[4] Leonard, R.G. (1984), A Database for Speaker-Independent Digit Recognition, *Proceedings of ICASSP*, San Diego, Vol. 3, p. 42.11

[5] Young, S. et al. (1999), The HTK book, *manual for the HTK2.2 software package*