ELSEVIER

Speech Communication 34 (2001) 127–139

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

HMM adaptation for applications in telecommunication

Hans-Günter Hirsch [a,b,*,1]

[a] *Centre of Speech Technology, Institute of Speech, Music and Hearing, Stockholm, Sweden*
[b] *Ericsson Eurolab Deutschland GmbH, Nordostpark 12, 90411 Nürnberg, Germany*

Abstract

The mismatch between the acoustic conditions during training and recognition often causes a performance deterioration in practical applications of speech recognition systems. Two important effects are the presence of a stationary background noise and the frequency response of the transmission channel from the speaker to the audio input of the recognizer. The original contribution of this work are two signal processing schemes for the estimation of the actual noise spectrum and the difference of the frequency responses between training and recognition. The estimated noise components are taken to adapt the cepstral parameters of the recognizer's references which are described by hidden Markov models (HMMs). The adaptation process is based on the parallel model combination (PMC) approach (M.J.F. Gales, Model based techniques for noise robust speech recognition, Dissertation at the University of Cambridge, 1995). For speaker independent connected or isolated word recognition considerable improvements can be achieved in the presence of just one type of noise as well as in the presence of both types together. Furthermore this adaptation scheme is integrated as part of a complete dialogue and recognition system which is accessible via the public telephone network. The usability and the gain in recognition performance is shown for this application in a real telecommunication scenario under consideration of all real-time aspects.

Zusammenfassung

Der Grund für die Verschlechterung der Leistungsfähigkeit von Spracherkennungssystemen in praktischen Anwendungen findet sich in unterschiedlichen akustischen Bedingungen während des Trainings und während der Erkennung. Zwei einflussreiche Effekte sind das Vorhandensein einer stationären Hintergrundstörung und der Frequenzgang des Übertragungskanals zwischen dem Sprecher und dem Audioeingang des Erkenners. Der Originalbeitrag der hier präsentierten Arbeiten beinhaltet zwei Signalverarbeitungsverfahren zur Schätzung des aktuellen Störspektrums und der Differenz der Frequenzgänge während des Trainings und der Erkennung. Die geschätzten Störkomponenten dienen der Adaption der cepstralen Parameter der Referenzen des Erkenners, wobei die Referenzen durch Hidden Markov Modelle beschrieben werden. Der Adaptionsvorgang basiert auf dem Ansatz der parallelen Modellkombination (PMC). Deutliche Verbesserungen können für die Erkennung von einzelnen Wörtern und Wortketten erzielt werden in der Gegenwart eines oder beider Störeffekte zur gleichen Zeit. Desweiteren ist dieses Adaptionsverfahren als Teil eines kompletten Dialog-und Erkennungssystems integriert worden, das über das öffentliche

* Tel.: +49-911-5217329; fax: +49-911-5217961.
E-mail address: hans-guenter.hirsch@eed.ericsson.se (H.-G. Hirsch).

[1] The author is with Ericsson Eurolab in Germany. This work was partly done during a research stay at the centre for speech technology (CTT) which is part of the royal institute of technology (KTH) in Stockholm.

0167-6393/01/$ - see front matter
PII: S0167-6393(00)00050-9

Telefonnetz zugänglich ist. Die Verwendbarkeit als auch der Gewinn in Bezug auf die Leistungsfähigkeit des Erkenners werden für diese Anwendung in einer realen Telekommunikationsumgebung unter Berücksichtigung aller Aspekte einer Implementierung in Echtzeit gezeigt.

Résumé

La diminution de l'efficacité des systèmes de reconnaissance vocale lors d'applications pratiques est due aux différentes exigences acoustiques pendant l'apprentissage et pendant la reconnaissance. Deux facteurs importants sont la présence de perturbations sonores en arrière plan et la transmission de fréquences entre le micro et l'entrée audio du reconnaisseur. L'étude présentée ici comporte deux modes de traitement de signaux pour estimer les perturbations sonores actuelles et la différence lors de la transmission de fréquences pendant l'apprentissage et pendant la reconnaissance. L'évaluation des perturbations sonores est utilisée pour une adaptation des paramètres cepstral des références du reconnaisseur, encore appelées HMM (Hidden Markov Models). Le mode d'aptation est basé sur la combinaison parallèle de modèles (PMC). Des améliorations importantes peuvent être apportées pour la reconnaissance d'un mot ou d'une chaîne de mots malgré la présence d'un ou des deux types de perturbations sonores simultanément. Cette procédure d'adaptation a de plus été intégrée à un système de dialogue et de reconnaissance qui est disponible sur le réseau téléphonique public. L'application et le gain d'efficacité du reconnaisseur sont démontrés pour cette application dans un environnement de télécommunication existant et en temps réel.

Keywords: Robust speech recognition; HMM adaptation

1. Introduction

Robustness is the most important factor limiting the application of speech recognition in a lot of real-life situations. Two important types of noises are illustrated in Fig. 1 which have a major influence on the recognition performance and which are present in almost all applications.

An example for this type of application is the installation of a recognition system at a switch in a telephone network where mainly two types of noise exist. 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. This type of noise is also referred to as "additive" noise. The second one is the frequency characteristic of the whole transmission channel, e.g. including the microphone and the telephone line. The term "convolutional" noise has been introduced for this type of distortion. Several approaches have been investigated to compensate these effects individually or both together (Gales and Young, 1995; Minami and Furui, 1996; Sankar and Lee, 1996; Stern et al., 1997).

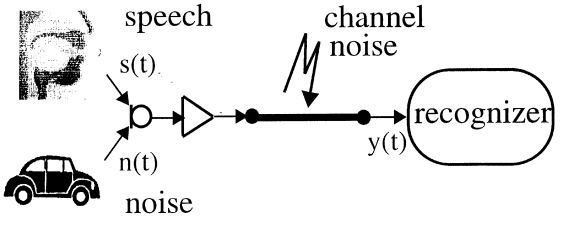

Fig. 1. Noise sources in the setup of a recognizer.

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

$$Y(f) = |H(f)|^2 S(f) + N(f), \tag{1}$$

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 estimates of $N(f)$ and $H(f)$ it is possible to adapt the spectral parameters of HMMs. The investigations of this study are based on cepstral features which are used in most of today's recognition systems. To apply the spectral adaptation, the cepstral parameters have to be transformed back to the linear spectral domain. The needed transformations back and forth are

well described in the parallel model combination (PMC) scheme.

In this paper two processing schemes are described to estimate the noise spectrum and the mismatch between frequency responses as well as their usage to adapt the cepstral parameters of a HMM based recognizer. Recognition experiments are done on artificially distorted data. The influence of additive stationary noise only is investigated first while extending the experiments to a combination of additive and convolutional noise later on. Furthermore, we consider the high mismatch situation of using different data bases for training and testing. Finally this adaptation approach is tested as part of a dialogue system in the telephone network with all constraints of a real-time implementation. The design of a recognition system based on HMM adaptation, which can be implemented in practical applications, is the final goal of this work. This practical aspect further differentiates this work from the more theoretical investigations on PMC (Gales, 1995, 1997).

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 modeling 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,
- the corresponding 12 delta cepstral coefficients.

The complete analysis scheme is shown in Fig. 2.

Feature vectors are calculated every 10 ms analyzing a 25 ms window. A preemphasis as well as a weighting with a Hamming window is applied to the samples inside each 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 (Young et al., 1996) on 5 consecutive frames of MFCC parameters.

Whole words are modeled by HMMs with the following features:

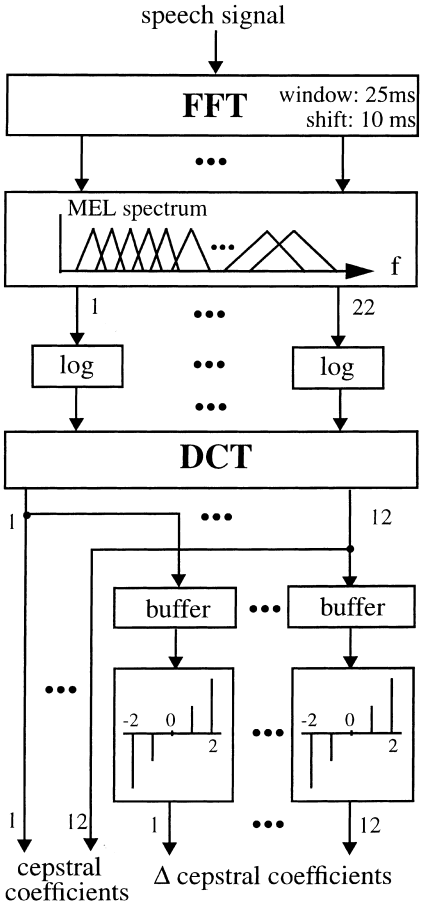

Fig. 2. Processing scheme of the feature extraction.

- 18 states per word,
- mixture of 4 (or 2) Gaussian distributions per state,
- simple left-to-right model,
- covariance matrix with only elements on the diagonal.

The training is done with the tools of the HTK software package (Young et al., 1996).

3. Adaptation of HMMs

The effects of additive and convolutional noise can be formally described in the linear spectral domain as shown in Eq. (1). For adaptation the cepstral parameters have to be transformed back to the spectral domain. Estimations of the noise

spectrum as well as of the frequency response are needed. In this approach the estimation processes are based on signal processing techniques. Both processing schemes and the adaptation scheme will be presented in this section.

3.1. Estimation of the noise spectrum

A measure is derived which is related to the signal-to-noise ratio (SNR) in subbands. This measure is taken to detect segments which contain only stationary background noise. The input consists of the short-term subband energies which are calculated by the feature extraction of the recognizer in the MEL frequency range. The processing scheme is illustrated in Fig. 3.

The noise spectrum is estimated as weighted sum of the actual and past short-term MEL spectra as long as no speech onset or the presence of a nonstationary segment is detected. The weighting function is an exponentially decaying curve, hence the actual spectrum gets a stronger weight than past spectra. It is realized as a simple recursive update.

$$\sqrt{\hat{N}(t_i,f)} = \alpha\sqrt{\hat{N}(t_{i-1},f)} + (1-\alpha)\sqrt{X(t_i,f)}, \tag{2}$$

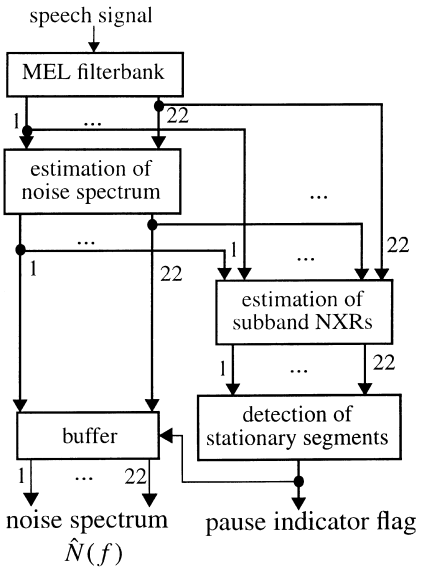

Fig. 3. Block diagram for the estimation of the noise spectrum.

where $\sqrt{\hat{N}(t_i,f)}$ is the estimated magnitude noise spectrum at time t_i and $\alpha = 0.9$ in the actual implementation.

The initialization of the estimation process is based on the assumption that the speech input to a recognizer is usually preceded by a pause segment where only the background noise is present. The update of the noise spectrum in an individual subband takes place as long as the input spectral component $\sqrt{X(t_i,f)}$ is below a threshold of $\beta\sqrt{\hat{N}(t_{i-1},f)}$ with $\beta = 1.75$ in the actual implementation. Exceeding the threshold corresponds to a certain rise of the subband energy which may exist due to a speech onset.

In parallel to the estimation of the noise spectrum a flag is determined indicating the presence of speech or of a nonstationary segment. This is based on the estimation of the measure

$$NX(f) = \sqrt{\hat{N}(t_i,f)}/\sqrt{X(t_i,f)}$$

that describes the noise-to-signal&noise ratio in a subband. A relative ratio is calculated as

$$NX_{\mathrm{rel}}(f) = \frac{NX(f) - NX_{\min}(f)}{NX_{\max}(f) - NX_{\min}(f)}, \tag{3}$$

with $NX_{\min}(f)$ and $NX_{\max}(f)$ as minimum respectively maximum of $NX(f)$ in all previous frames.

$NX_{\mathrm{rel}}(f)$ takes values between 0 and 1 and describes the relative noise-to-signal & noise ratio for the current range of $NX(f)$ values in the actual input utterance. $NX_{\mathrm{rel}}(f)$ takes small values in the presence of speech and values close to 1 in the presence of stationary segments. The presence of speech inside a subband is indicated when the relative ratio takes a value below an adaptive threshold of $[0.8 - NX_{\min}(f)]$. The speech flag is set when in at least three successive frames at least one subband indicates the presence of speech. In case the relative ratios in all subbands take a value above a threshold of 0.8 this frame is marked as pause. All mentioned constants and thresholds have been optimized in earlier investigations (Hirsch and Ehrlicher, 1995).

The estimated noise spectrum is copied into a buffer as long as the flag indicates the presence of a stationary segment. Furthermore the flag is used to

start the recognition and to trigger the HMM adaptation at speech onset. The whole estimation is individually applied to each input utterance of the recognizer.

3.2. Estimation of the frequency response

Several approaches exist for the estimation of the frequency response and its application to recognition in additive and convolutional noise (Gales and Young, 1995; Minami and Furui, 1996; Sankar and Lee, 1996; Stern et al., 1997). Some of these approaches cause a high computational load or need some special adaptation data. The method presented in this paper is computationally inexpensive, does not cause any delay and does not need any adaptation data.

Using Eq. (1) the actual frequency response can be estimated as

$$|\hat{H}_{\mathrm{act}}(f)|^2 = \frac{Y_{\mathrm{long}}(f) - \hat{N}(f)}{\hat{S}_{\mathrm{long}}(f)}. \tag{4}$$

Assuming a constant frequency response $H(f)$ and a constant noise spectrum $N(f)$ during a speech utterance the long-term spectrum $Y_{\mathrm{long}}(f)$ of this utterance can be introduced as description of the noisy input speech. In the same way, the short-term spectrum $S(f)$ can be substituted by the corresponding long-term spectrum $\hat{S}_{\mathrm{long}}(f)$ as estimate for the clean speech. This leads to a better estimation of the frequency response than using only the information of a single short-term spectrum. The long-term spectrum $Y_{\mathrm{long}}(f)$ 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 all segments which have been classified as speech by the recognizer. The estimated noise spectrum $\hat{N}(f)$ is determined as described above. The long-term spectrum $\hat{S}_{\mathrm{long}}(f)$ of the "clean" speech is estimated by using the spectral information which is contained in the HMMs. After the recognition of an utterance the Viterbi alignment is used to define the "best" sequence of HMM states which represents the input speech. For all states we consider the Gaussian mixture component with the smallest spectral distance from the input speech spectrum. The cepstral means of these Gaussians are transformed back to spectral parameters, and then averaged over all states to provide an estimate of $S_{\mathrm{long}}(f)$. We take the Euclidean distance between cepstral means as measure for the spectral distance. The whole process to determine the actual estimate $|\hat{H}_{\mathrm{act}}(f)|$ of the frequency response is shown in the block diagram of Fig. 4.

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

$$|\hat{H}_{\mathrm{new}}(f)|^2 = \alpha|\hat{H}_{\mathrm{old}}(f)|^2 + (1-\alpha)|\hat{H}_{\mathrm{act}}(f)|^2, \tag{5}$$

where α is chosen to 0.9. The iterative updating generates a smoothed version of the frequency response and compensates estimation errors which might occur for an individual utterance.

The new estimate can be applied in the HMM adaptation scheme when recognizing 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 performed off-line, e.g. after recognizing an utterance so that it does not cause any delay.

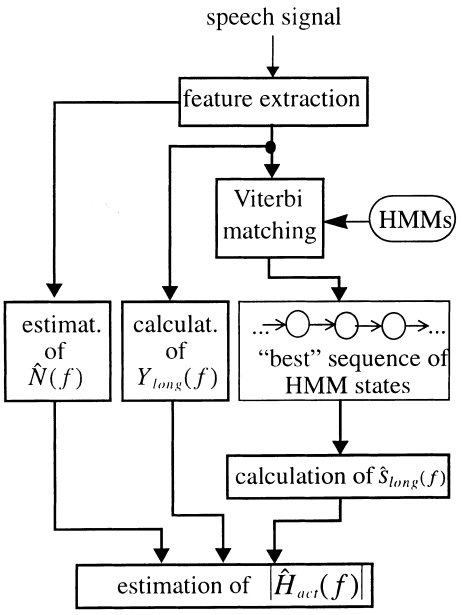

Fig. 4. Block diagram for the frequency response estimation.

In fact, this estimation procedure does not only estimate the frequency response of the transmission channel. Using the spectral information contained in the HMMs considers also the frequency response which was present during recording of the training data. Thus this processing estimates the whole mismatch between the frequency responses of training and recognition phase. Furthermore, the spectral characteristics of each individual speaker are compensated.

3.3. Adaptation of the cepstral parameters

Having the estimates of $N(f)$ and $H(f)$ the modification of the "clean" speech spectrum can be calculated according to Eq. (1). The application of this modification in the HMM modeling requires the transformation of the cepstral parameters back to the linear spectral domain. This processing is shown in Fig. 5 for the adaptation of the cepstral means.

The cepstral means of each mixture density component and each HMM state are transformed back to the linear spectral domain. The influence of convolutional and additive noise is compensated by

- multiplying the MEL power density spectrum with the estimated frequency response and
- adding the estimated noise spectrum.

The modified spectrum is transformed again to the cepstral domain.

For the realization in this study the actual estimate $\hat{N}(f)$ is applied for each individual speech input. Furthermore, the previous estimate of $|\hat{H}_{\text{new}}(f)|$ is taken. The adaptation is individually done once for each utterance at the onset of speech, which is detected as described in Section 3.1.

Besides adapting all HMMs an additional model is introduced which describes the background noise. This model consists of one state with a single density. The cepstral parameters are calculated from the preceding noisy segment.

Adapting only the cepstral means is called Log-add approximation in (Gales, 1995). It is based on the assumption that the stationary background noise can be described by a mean spectrum. Actually, the logarithm of the noise spectrum in each subband have approximately a Gaussian distribution. A more accurate adaptation can be achieved considering this distribution. The Gaussian distribution corresponds to a Log-normal distribution in the linear spectral domain. Looking at Eq. (1) adding noise has to be described more accurately as combining the distributions of speech and noise power spectra. Assuming that speech and noise are independent and additive the combination of distributions results in adding the spectral means and their corresponding covariance matrices (Gales, 1995).

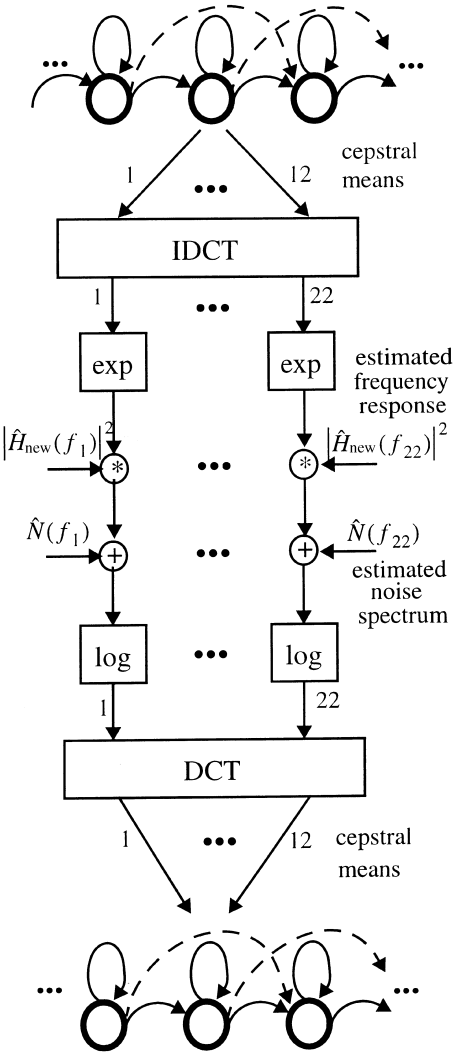

Fig. 5. The adaptation scheme.

The variances of the cepstral parameters have to be transformed back to the logarithmic spectral domain to determine the covariance matrix in the linear spectral domain. Furthermore, it is shown in (Gales, 1995) how to approximately map the Gaussian distribution in the logarithmic domain to the Log-normal distribution in the linear domain. After combining the distributions in the linear domain the corresponding parameters in the cepstral domain have to be recalculated. This approach is called the Log-normal approximation in (Gales, 1995). It has to be considered that the transformations of the covariance matrices causes a much higher computational load than just transforming the cepstral means as in the Log-add approximation.

Furthermore the delta cepstral coefficients can be adapted. In this study a simple weighting according to (Gales, 1997) is applied to the corresponding spectral coefficients in the logarithmic domain as described by

$$\Delta\hat{S}_{lg}(f) \approx \frac{S(f)}{S(f)+\hat{N}(f)}\Delta S_{lg}(f), \tag{6}$$

$\Delta S_{lg}(f)$ represents the logarithmic spectral parameters when transforming back the delta cepstral coefficients, and $S(f)$ represents the power density spectrum of the corresponding cepstral means. The adaptation of the delta coefficients can be added to the Log-add approximation as well as to the Log-normal approximation. The Log-add as well the Log-normal approximation in combination without or with adapting the delta coefficients are investigated in the recognition experiments.

4. Recognition experiments

This study focuses on the speaker independent recognition of digit sequences and isolated digits. The whole word HMMs are determined from the training part of the TIDIGITS data base (Leonard, 1984) using the tools of the HTK package. 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. Each digit is modeled by a single HMM consisting of a mixture of four Gaussian components per state.

Recognition experiments are done on the designated part of the TIDIGITS and on the Bellcore digits. The TIDIGITS data are artificially distorted by adding noise and filtering the speech signals. The Bellcore data base contains isolated digits recorded via telephone lines. These data have a worse SNR than the clean TIDIGITS and contain all effects of recordings over telephone channels.

4.1. Recognition of the TIDIGITS

Two sets of recognition experiments are done on the designated TIDIGITS test data.

In the first set only the estimated noise spectrum is used for the HMM adaptation to investigate the recognition performance in the presence of stationary background noise. No adaptation of the frequency response is included.

The second set of recognition experiments takes the estimate of the noise spectrum and the estimate of the frequency response mismatch between training and test data as input for the HMM adaptation. The test data are filtered and stationary noise is added to show the performance in the presence of convolutional and additive noise.

Before describing the two sets of experiments the baseline performance of the recognizer is given as a starting point. A word error rate of **0.77**% can be achieved corresponding to a string error rate of **2.37**%. This test is done on the clean TIDIGITS without applying any type of adaptation.

4.1.1. Adaptation to additive noise

To investigate the influence of stationary background noise distorted versions of the TIDIGITS are created by artificially adding car noise at different SNRs. The car noise was recorded inside a car. An almost stationary segment is taken for the artificial distortion. The sub-segment which is actually added to an individual utterance is randomly extracted out of the noise recording. The results are plotted in Fig. 6 and listed in Table 1 when applying the Log-add approximation without and with adapting the delta coefficients.

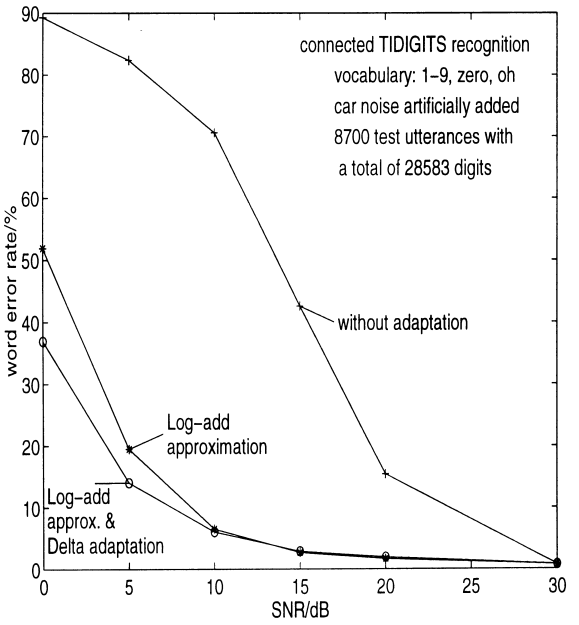

Fig. 6. Word error rates applying the Log-add approximation.

The results without adding noise are plotted at a SNR of 30 dB. The word error rate considerably increases when adding car noise and without applying any adaptation. A remarkable gain can be achieved over the whole range of SNRs when adapting the HMMs. The adaptation of the delta coefficients further decreases the error rates at low SNRs.

Similar results are achieved for the Log-normal approximation which are listed in Table 1.

The error rates improve at low SNRs but slightly increase at high SNRs in comparison to the Log-add approximation. In general, the adaptation of the delta coefficients as well as the adaptation of the variances seems to be worthwhile only at low SNRs for this special realization of a PMC approach. It has to be considered that the adaptation of the variances causes a high computational load.

In a next step, 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 principle approaches. The first approach tries to make the feature extraction more robust against certain distortions. In the second approach the references are adapted with respect to the distortion without modifying the existing feature extraction. Some results for the noisy TIDIGITS are plotted in Fig. 7. Again car noise is considered as additive noise.

The spectral subtraction is applied as preprocessing of the noisy utterances before processing them in the recognizer. The method as described in Section 3.1 is used for the estimation of the noise spectrum. Spectral subtraction is done with an overestimation factor of 1 and without adding a noise floor (Hirsch and Ehrlicher, 1995).

Fig. 7 shows a considerable improvement when applying spectral subtraction in comparison to the case without adaptation. But the improvement is higher when applying the HMM adaptation as Log-add approximation without adapting the delta coefficients.

This result supports the following hypothesis. The modification of the feature extraction may reduce certain distortions but introduces artificial distortions of the speech. For example, some speech segments with low energy or a spectral

Table 1
Word error rates when applying the Log-normal approximation

SNR/dB	Without adaptation	Log-add approximation	Log-add and Delta adaptation	Log-normal approximation	Log-normal and Delta adaptation
Clean	0.77%	0.77%	0.77%	0.97%	0.96%
20	15.4%	1.69%	1.97%	2.75%	3.13%
15	42.6%	2.7%	2.92%	3.34%	3.75%
10	70.6%	6.52%	6.05%	5.96%	6.38%
5	82.4%	19.5%	14.01%	15.07%	14.11%
0	89.25%	51.87%	36.9%	38.45%	32.81%

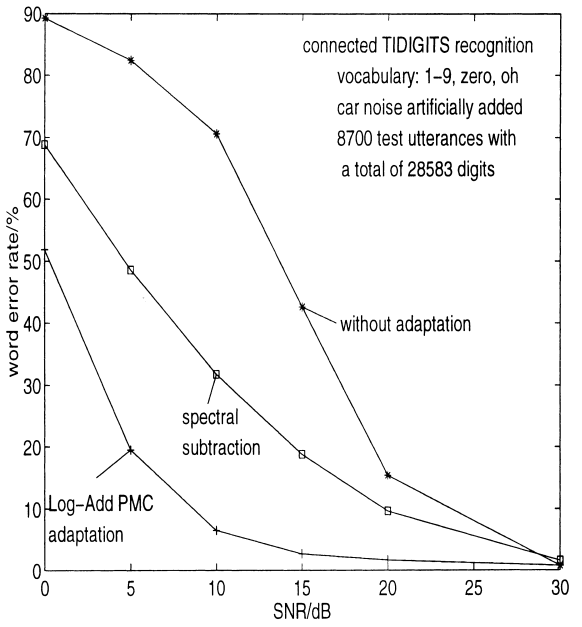

Fig. 7. Word error rates comparing the Log-add approximation against spectral subtraction.

characteristic similar to the additive noise will be attenuated or spectrally shaped in case of applying spectral subtraction. Those new artificial distortions have a negative effect on the recognition. On the other hand the adaptation of the references does not modify the input speech. The knowledge about the distortion is used to map the acoustic parameters contained in the references. Indeed, this may have a negative effect on the recognition too, if done incorrectly. Looking at the realization in this study the same noise estimation is used for spectral subtraction and for the HMM adaptation. Thus it gets obvious that one has to expect better results using the noise estimate to adapt the references instead of modifying the feature extraction.

Another noisy version of the TIDIGITS is artificially created showing the ability to adapt to changing noise situations and changing SNRs. Besides the car noise two further noises are considered. These are the helicopter noise and the stationary noise with a speech-like spectrum which were used in the NOISEX92 study (Varga and Steeneken, 1993).

Table 2
Word error rates recognizing the noisy TIDIGITS (car noise, helicopter noise, speech-like noise)

	Without adaptation	Log-add approximation
Word error rate	47.7%	5.07%

One of the three noises is randomly selected for the distortion of a digit sequence. Thus on average each noise is added to 1/3 of all TIDIGITS utterances. Furthermore the SNR is randomly chosen between 5 and 15 dB in steps of 1 dB. Thus the overall SNR is 10 dB on average. The recognition results are listed in Table 2 when applying the Log-add approach without adapting the delta coefficients.

The improvement is almost the same in comparison to the situation where only one type of noise is considered at a constant SNR. This shows that the method is applicable to situations where the noise as well as the SNR changes for consecutive utterances.

As conclusion of these investigations on additive noise it turns out that the adaptation of the static cepstral coefficients leads to a considerable improvement of the recognition performance. The adaptation of the variances seems to be not worthwhile when thinking about the limited further gain on one hand, but the high computational costs on the other hand.

4.1.2. Adaptation to additive and convolutional noise

In this section the estimated noise spectrum and the estimate of the frequency response mismatch are used to adapt the HMMs. The overall performance increases when applying the adaptation to the "clean" test data. The results are listed in Table 3. The adaptation of the delta coefficients is not included in 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. This is done according to a real application. A person calling a speech server based on recognition will

Table 3
Error rates when recognizing the clean data

	Without adaptation	Log-add approximation and filter estimation
String error rate	2.37%	1.98%
Word error rate	0.77%	0.65%

Table 4
Word error rates when recognizing the filtered data

	Without adaptation	Log-add approximation and filter estimation
Word error rate	4.23%	0.71%

use the system for a while before the next speaker is going to use it.

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/octave is applied in the frequency range from 300 to 1000 Hz. The filter characteristic remains flat for frequencies between 1000 Hz and up to about 3000 Hz. The recognition results are listed in Table 4.

The influence of the filtering can be compensated almost completely by this type of iterative filter estimation. To get some idea about the filter estimation process the estimated frequency responses are plotted in Fig. 8 for the first 50 consecutive utterances of a recognition run.

All 50 utterances belong to the same speaker. The initial values of the estimate are 1. It can be seen that the estimated response adapts to a certain characteristic from its initial values and remains fairly stable for almost all frequency bands.

The attenuation of the low frequencies becomes obvious in Fig. 8, when comparing the frequency characteristic with the filter simulating the telephone channel. This is also true for high frequencies but it can not be seen in this view on the 3D plot.

Finally, the performance in the presence of additive and convolutional noise is investigated for SNRs in the range of 0–20 dB. Results are shown in Fig. 9.

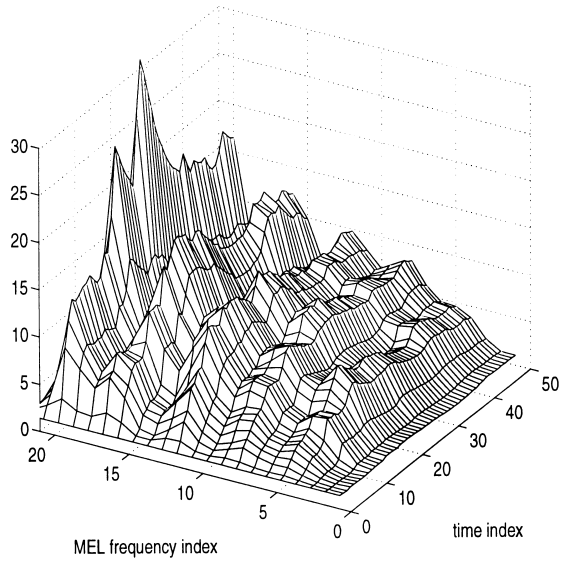

Fig. 8. Consecutive estimates of $|\hat{H}_{\text{new}}(f)|^2$ for a single speaker.

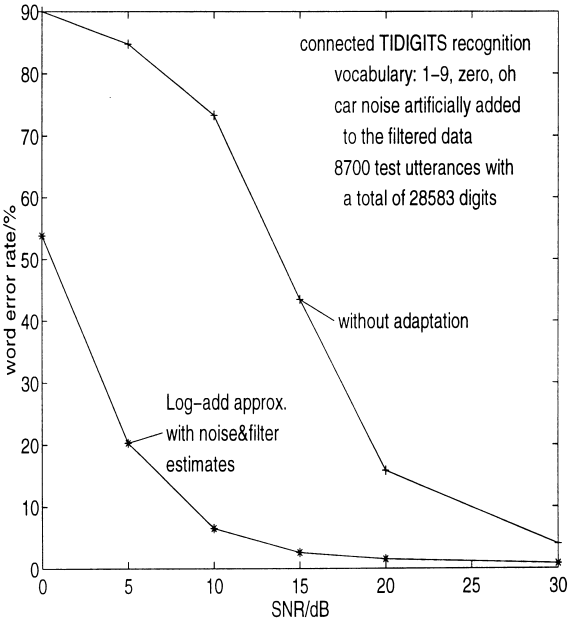

Fig. 9. Word error rates when recognizing the filtered data with car noise added.

The results for the filtered data without additive noise are plotted at a SNR of 30 dB. The results without and with adaptation look very similar to the ones presented in Fig. 6, where only noise is

added. Looking at a single utterance and a specific SNR it has to be mentioned that the absolute noise level is less in comparison to the condition where only noise is added without filtering. The energy of the speech is already reduced by the filtering which leads to a reduced noise level.

As result of previous experiments it turned out that the estimation of the frequency response is not influenced by degraded recognition results because of additive noise for SNRs down to 10 dB. Below 10 dB small degradations were found due to the higher number of mismatches introduced by the worse recognition. The estimation process is dependent on the correct number of matches because the matching information is the basis for the estimation. Therefore, the filter estimation is disabled below a certain SNR to avoid this degradation. The SNR is already determined when estimating the noise spectrum. In case of SNRs below a predefined threshold (<5 dB) the filter estimation is disabled for the next speech utterance. This technique is used to achieve the results presented in Fig. 9.

Summarizing, it turns out that the adaptation scheme is able to considerably improve the recognition performance in the presence of additive *and* convolutional noise.

4.2. Recognition of the Bellcore digits

Additionally to the recognition of artificially distorted data a different set of speech data is recognized which was recorded over telephone lines. 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. It 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 introduced by the microphone and the usual effects of different telephone lines and different handsets. The recognizer is set up to recognize isolated words only. The word error rates are shown in Fig. 10 without and with adaptation. The Log-add approximation is applied as adaptation scheme. Results are shown

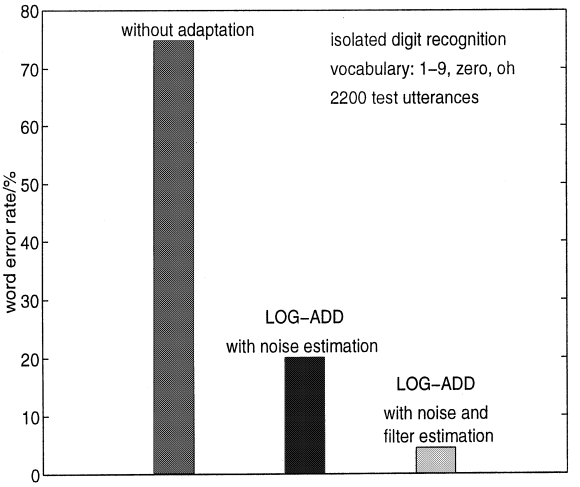

Fig. 10. Word error rates recognizing the Bellcore digits.

for compensating additive noise only, or both additive and convolutional noise.

The word error rate is about 75% without adaptation for this simple task of recognizing the 11 digits 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 to about 20% when applying the noise estimation for the Log-add approximation. A further reduction by a factor 4 resulting in a word error rate of about 4.5% 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. Application in the telephone network

The recognition and adaptation scheme is integrated as part of a complete speech dialogue system which is connected to the public telephone network. A PC with a clock frequency of 266 MHz and with Linux as operating system is taken as hardware basis. A passive ISDN card is used to connect the PC to the telephone network. The recognition module is the same as the one used for the simulation experiments with the only difference of taking the a-law samples (ITU, 1988) from the

ISDN line as input to the feature extraction. The static and the delta cepstral coefficients are adapted to the actual noise situation.

For a small field test 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 utter 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 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 Gaussian components. 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 are recorded up to now. The recognition system shows a 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 point in the dialogue the recognition of isolated command words only is considered as task. Furthermore, the system was called by some people which had fairly different accents in comparison to the speakers used to train 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 SNR of most speech signals is higher than 10 dB. The word error rates are shown in Table 5 without and with applying the adaptation scheme.

Table 5
Word error rates for telephone data

	Without adaptation	With adaptation
Word error rate	7.98%	3.47%

These off-line experiments are done by using the HMMs as described above which had been trained on nontelephone data. Again considerable improvements can be achieved by applying the adaptation technique. It has to be mentioned that garbage models are introduced to model nonstationary noises like e.g. breathing before and after the speech. Such garbage models help to improve the recognition in real-life applications.

6. 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. Thus an approach is shown how to cope with two degrading effects making it more difficult to achieve a robust recognition in a lot of real-world applications.

The processing is based on the PMC approach where the noise spectrum as well as the frequency response are estimated with signal processing techniques. Both estimation schemes are the original contribution of this paper. They work reliable and robust. It is shown that the recognition performance can be considerably improved for artificially distorted data as well as for real-life speech data. This improvement can already be achieved by adapting the cepstral means only. Adapting more accurately the distributions of the cepstral features leads to a further decrease of the error rate especially for low SNRs. But this causes a high computational load.

The adaptation scheme is integrated as part of a speech dialogue system in the public telephone network. It could prove its usability and its ability to improve the recognition performance also in this real-life scenario under all constraints of a real-time implementation.

References

Gales, M.J.F., 1995. Model based techniques for noise robust speech recognition. Dissertation at the University of Cambridge.

Gales, M.J.F., 1997. Nice model-based compensation schemes for robust speech recognition. In: ESCA Workshop on

Robust Speech Recognition for Unknown Communication Channels, Pont-a-Mousson, France, pp. 55–64.

Gales, M.J.F., Young, S.J., 1995. Robust speech recognition in additive and convolutional noise using parallel model combination. Computer Speech and Language 9, 289–307.

Hirsch, H.G., Ehrlicher, C., 1995. Noise estimation techniques for robust speech recognition. In: ICASSP95, Detroit, USA, pp. 153–156.

ITU recommendation G.711, 1988. Pulse code modulation of voice frequencies.

Leonard, R.G., 1984. A database for speaker-independent digit recognition. In: ICASSP, Vol. 84-3, p. 42.11.

Minami, Y., Furui, S., 1996. Adaptation method based on HMM composition and EM algorithm. In: ICASSP96, Atlanta, USA, pp. 327–330.

Sankar, A., Lee, C.H., 1996. A maximum-likelihood approach to stochastic matching for robust speech recognition. IEEE Trans. Speech and Audio Process. 4, 190–201.

Stern, R.M., Raj, B., Moreno, P.J., 1997, Compensation for environmental degradation in automatic speech recognition. In: ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels, Pont-a-Mousson, France, pp. 33–42.

Varga, A., Steeneken, H.J.M., 1993. Assessment for automatic speech recognition: II. Noisex92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Communication 12, 247–252.

Young, S. et al., 1996. The HTK Book. Manual for the HTK2.0 Software Package.