Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 50 (2008) 244–263
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

A new approach for the adaptation of HMMs to reverberation and background noise

Hans-Günter Hirsch *, Harald Finster

Niederrhein University of Applied Sciences, Department of Electrical Engineering and Computer Science, Reinarzstr. 49, 47805 Krefeld, Germany

Received 27 October 2006; received in revised form 6 September 2007; accepted 15 September 2007

Abstract

Looking at practical application scenarios of speech recognition systems several distortion effects exist that have a major influence on the speech signal and can considerably deteriorate the recognition performance. So far, mainly the influence of stationary background noise and of unknown frequency characteristics has been studied. A further distortion effect is the hands-free speech input in a reverberant room environment.

A new approach is presented to adapt the energy and spectral parameters of HMMs as well as their time derivatives to the modifications by the speech input in a reverberant environment. The only parameter, needed for the adaptation, is an estimate of the reverberation time. The usability of this adaptation technique is shown by presenting the improvements for a series of recognition experiments on reverberant speech data. The approach for adapting the time derivatives of the acoustic parameters can be applied in general for all different types of distortions and is not restricted to the case of a hands-free input.

The use of a hands-free speech input comes along with the recording of any background noise that is present in the room. Thus there exists the need of combining the adaptation to reverberant conditions with the adaptation to background noise and unknown frequency characteristics. A combined adaptation scheme for all mentioned effects is presented in this paper. The adaptation is based on an estimation of the noise characteristics before the beginning of speech is detected. The estimation of the distortion parameters is based on signal processing techniques. The applicability is demonstrated by showing the improvements on artificially distorted data as well as on real recordings in rooms.
© 2007 Elsevier B.V. All rights reserved.

Keywords: Robust speech recognition; HMM adaptation; Hands-free speech input; Reverberation

1. Introduction

The use of speech recognition as alternative input device is especially of interest where the user has not his hands available for controlling a keyboard or mouse. Quite often this input mode comes along with the need of a hands-free speech input. For reasons of practical usage and personal comfort it is especially of interest without the need of wearing a close-talking microphone.

The drawback of a hands-free speech input is a modification of the speech by the acoustic environment when the input takes place in a room. The influence of transmitting speech in a room can be modeled as superposition of the sound on the direct path from the talker's mouth to the recording microphone and multiple reflections of the sound at the walls and any equipment inside the room. For stationary conditions the transmission can be modeled as convolution of the speech with a room impulse response (RIR). But the RIR changes as soon as the talker moves in the room or room conditions change like in case of opening a door or a window or other people moving in the room. Thus the adaptive estimation of the RIR is a quite difficult and complex task.

* Corresponding author. Tel.: +49 2403 702596; fax: +49 2403 702597.
E-mail address: hans-guenter.hirsch@hs-niederrhein.de (H.-G. Hirsch).

0167-6393/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2007.09.004

Several approaches exist for enhancing speech that has been recorded in hands-free mode inside rooms. Some of these approaches have also been applied for the improvement of speech recognition. The methods can be separated in single and multi-channel processing techniques. Most of the multi-channel approaches (e.g. Omologo et al., 1998; Bitzer et al., 1999; Liu and Malvar, 2001; Seltzer et al., 2004), are based on a beamforming technique to reduce the influence of the reflected sound or on a correlation based multi-channel processing. A variety of different techniques are the basis for the single channel approaches (e.g. Avendano and Hermansky, 1996; Kingsbury, 1998; Yegnanarayana and Murthy, 2000; Gelbart and Morgan, 2002; Tashev and Allred, 2005; Wu and Wang, 2005; Kinshita et al., 2005). Some are based on modifying the envelope contours of subband energies.

Only a few approaches exist that try to improve speech recognition by modifying the pattern matching process (Couvreur et al., 2001; Palomäki et al., 2002). One method (Raut et al., 2005) is also based on an adaptation and modification of HMM parameters as in the approach presented here.

Looking at HMMs that are used for modeling speech in recognition systems, the detailed knowledge about the transmission as it is given by a RIR is not needed. Each state of a HMM represents a short speech segment with several tenth of milliseconds duration. The state contains information about the distribution of some spectral parameters within the segment. Usually a type of MEL filterbank is applied. In general HMMs describe speech with a quite low resolution with respect to time and frequency in comparison to the detailed description with a RIR. Thus the estimation of the RIR is not really needed to include the modifications of the HMM parameters that are caused by the hands-free speech input.

A new approach is presented in the next section for adapting the parameters of HMMs to the speech transmission inside a room. The method is based on the description of a room transmission by an impulse response with an exponentially decaying envelope as approximation for a real RIR. This approximation is applied to the fairly rough modeling of speech as a sequence of HMM states. As consequence of the exponentially decaying shape of the impulse response, the acoustic excitation at a certain point in time will also be seen at later time segments. The effect of reverberation is an temporal extension of an acoustic excitation. This extension is modeled by adding contributions of earlier states with respect to energy and spectral parameters. The only parameter, needed for the adaptation, is an estimate of the reverberation time T60 that defines the contour of the exponentially decaying RIR. Several recognition experiments have been performed to proof the usability of the new approach.

In most hands-free speech input situations background noise is present in the room. Thus, the HMM adaptation for the effects of a hands-free speech input is only useful when it can be combined with a technique for compensating the influence of stationary background noise and of an unknown frequency characteristic. During the recent years a lot of investigations have been carried out to reduce the deterioration of the recognition performance due to additive background noise and unknown frequency characteristics. The approaches are either based on the extraction of robust features in the front-end (e.g. Macho et al., 2002; Gadrudadri et al., 2002; ETSI, 2003), or on the adaptation of the HMM parameters to the noise conditions (e.g. Gauvain and Lee, 1994; Minami and Furui, 1996; Sankar and Lee, 1996; Gales and Young, 1996; Gales, 1997; Woodland, 2001). Most of the adaptation techniques try to estimate some kind of HMM parameter mapping with a maximization of the likelihood score as optimization criterion. This work uses signal processing approaches for estimating the distortion effects. The estimated distortion parameters are taken for the adaptation on the basis of a signal processing model that describes the spectral modifications due to the distortions. A frequency weighting can be caused, e.g. by the special characteristics of the recording microphone. In case of a hands-free speech input a frequency weighting might also occur due to the frequency-dependency of T60. This is not covered by the new approach which assumes one frequency independent value for T60 so far. But it can be compensated with an additional adaptation to unknown frequency characteristics.

This paper shows how the new adaptation technique can be combined with the adaptation of HMMs to additive noise and unknown frequency characteristics (Hirsch, 2001a). This approach is based on the well known PMC method (Gales, 1995). The results of several recognition experiments are presented in the last chapter that proof the usability of the combined adaptation to all distortion effects as they can occur in real application scenarios of speech recognition systems. The achieved results are compared to the application of the well-known MLLR (maximum likelihood linear regression) approach (Leggeter and Woodland, 1995) as an alternative adaptation technique.

2. Adaptation to hands-free speech input

The new approach for adapting the energy and spectral parameters of HMMs will be derived in this section. It is based on the approach of modeling the transmission in a reverberant room by an impulse response with an ideal, exponentially decaying shape. This will be presented in Section 2.1. Section 2.2 describes the adaptation of the static energy and spectral parameters of HMMs derived from the ideal modeling of the RIR. Finally Section 2.6 presents a new technique to adapt also the time derivatives of the energy and spectral HMM parameters.

2.1. Modeling the influence of a hands-free speech input

The multiple reflections of sound in a room can be ideally described by an exponential decay of the acoustic

energy which has been the result of early investigations in room acoustics (Kuttruff, 2000). This leads to a room impulse response $h(t)$ with an exponentially decaying shape

$$E(t) = E_0 \cdot \mathrm{e}^{-\frac{6\cdot\ln(10)}{T_{60}}\cdot t} \quad \text{with}$$

$$E_0 \equiv \text{energy of the acoustic excitation}$$

$$\Rightarrow\ h^2(t) \sim \mathrm{e}^{-\frac{6\cdot\ln(10)}{T_{60}}\cdot t}. \tag{1}$$

The only parameter for the description of the exponential shape is the reverberation time T60 that takes approximately values in the range of about 0.2–0.4 s for smaller rooms and of about 0.4–0.8 s for larger rooms. It can take values above 1 s for very large rooms like churches. The reverberation time depends on the interior in the room and the individual absorption characteristics of the walls.

The RIR can be transformed to the room transfer function by means of a Fourier transform. The room transfer function has a contour that changes very fast along frequency. Usually only the envelope of the room transfer function is of interest when looking at the filterbank approaches that are applied for extracting acoustic features in speech recognition. This effect can be covered by an adaptation to an unknown frequency characteristic.

More important for the frame based analysis in speech recognition is the influence on the contour of the short-term energy along time. The energy contours of a speech signal are shown in Fig. 1 before and after the transmission in a room. The energy is usually estimated as short-term energy in frames of about 20 ms duration. It can be seen that the reverberation leads to an artificial extension of each sound contribution. This extension occurs as so called reverberation tail with the exponentially decaying shape of the RIR. The same effect will also be seen when looking at the energy contours in single subbands of a MEL based filterbank that is usually applied in the front-end of a speech recognition system.

Transforming such energy contours to the so called modulation spectrum by means of a Fourier transform leads to the estimation of the modulation transfer function $m(F)$ (Houtgast et al., 1980) which can be mathematically described as

$$m(F) = \frac{1}{\sqrt{1 + \left(2\cdot\pi\cdot F\cdot\frac{T_{60}}{6\cdot\ln(10)}\right)^2}}. \tag{2}$$

The low pass characteristic of the modulation transfer function is shown in Fig. 2 for different values of T60. The cut-off frequency of the low pass characteristic is shifting to lower values of the modulation frequency for increasing values of T60. This corresponds to longer reverberation tails for higher values of T60. The artificial extension of sound contributions can lead to masking the acoustic parameters of low energy sounds by the parameters of a preceding sound with higher energy.

2.2. Adaptation of static parameters

Looking at a sequence of HMM states the acoustic excitation described by the parameters of a single state will also occur in succeeding states at a certain attenuation. This is based on the assumption that the HMMs have been trained on clean speech, recorded with a close talking microphone. Fig. 3 tries to visualize this effect.

Each state S_i of a HMM describes a speech segment with an average duration $\mathrm{dur}(S_i)$ that can be derived from the transition probability $p(S_i|S_i)$ to remain in this state.

$$\mathrm{dur}(S_i) = \frac{1}{1 - p(S_i|S_i)}\cdot t_{\mathrm{shift}}$$

$$\text{for all states } S_i \text{ with } 1 \leqslant i \leqslant \mathrm{NR_states}, \tag{3}$$

where t_{shift} is the time for shifting the analysis window in the feature extraction and NR_states is the number of

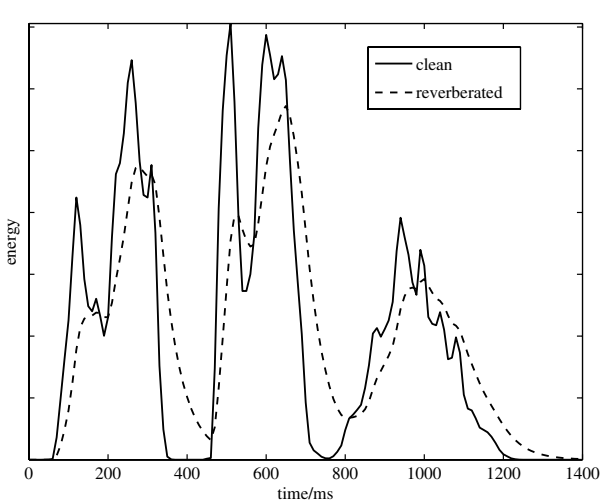

Fig. 1. Energy contours of a speech signal in clean and reverberant condition.

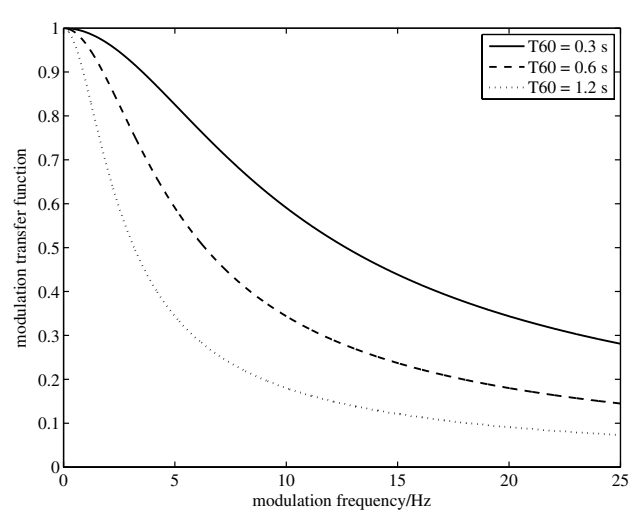

Fig. 2. Modulation transfer functions for different values of T60.

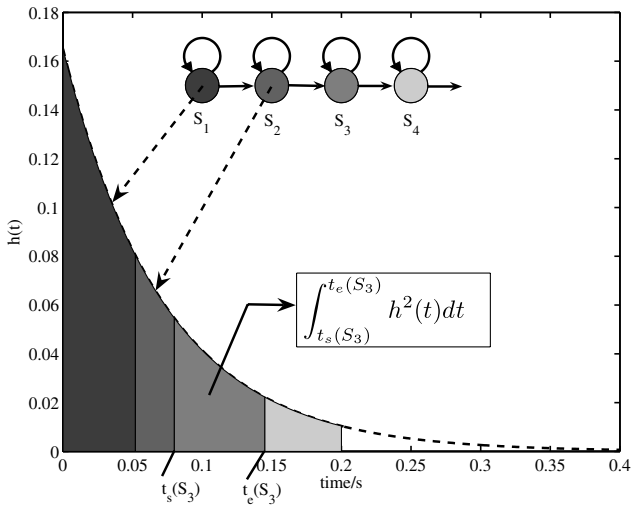

Fig. 3. The distribution of energy at state S_1 due to reverberation.

states of an individual HMM. t_{shift} takes a value of 10 ms for all analysis schemes applied in these investigations. This corresponds to a frame rate of 100 Hz.

The energy contribution of the acoustic excitation in the first state to a succeeding state S_i can be estimated by integrating the squared RIR as defined by Eq. (1) over the time segment of the succeeding state

$$\alpha_{i,1} = \int_{t_s(S_i)}^{t_e(S_i)} h^2(t)\partial t \quad \text{where } t_s(S_i) = \sum_{j=1}^{i-1} \text{dur}(S_j) \text{ and } t_e(S_i) = t_s(S_i) + \text{dur}(S_i) \text{ and } \int_0^{\infty} h^2(t)\partial t = 1. \tag{4}$$

Given an estimate of the reverberation time T60 the contribution factors can be individually calculated for all states of all HMMs. The approach for estimating T60 will be described later.

Usually each state of a HMM is defined by a set of spectral parameters like the MEL frequency cepstral coefficients (MFCCs) and an energy parameter. These parameters are the means of Gaussian distributions. The corresponding variances are needed as further parameters to completely define the shape of the Gaussians.

The mean of an energy parameter at an individual state S_i can be adapted by adding the energy contributions of the state itself and the preceding states

$$\tilde{E}(S_i) = \alpha_{i,i} \cdot E(S_i) + \alpha_{i,i-1} \cdot E(S_{i-1}) + \alpha_{i,i-2} \cdot E(S_{i-2}) + \cdots = \sum_{j=1}^{i} \alpha_{i,j} \cdot E(S_j). \tag{5}$$

In the same way the means of the power density spectra can be adapted. In case of using MFCCs, the cepstral coefficients have to be transformed back to the spectral domain first.

$$\{C_0, C_1, C_2, \ldots, C_{\text{NR_cep}}\} \xrightarrow{\text{IDCT}} \{\log(|X_1|), \log(|X_2|), \ldots, \log(|X_{\text{NR_mel}}|)\} \xrightarrow{\text{EXP}} \{|X_1|, |X_2|, \ldots, |X_{\text{NR_mel}}|\}, \tag{6}$$

where NR_cep is the highest index of the cepstral coefficients and NR_mel is the number of bands in the MEL frequency range. For NR_cep a value of 12 is chosen and NR_mel takes a value of 24 in our realization. $|X_k|$ represents the value of the magnitude spectrum in the MEL band with index k.

Then the power density spectra can be adapted in the same way as the energy parameter

$$|\tilde{X}_k(S_i)|^2 = \alpha_{i,i} \cdot |X_k(S_i)|^2 + \alpha_{i,i-1} \cdot |X_k(S_{i-1})|^2 + \alpha_{i,i-2} \cdot |X_k(S_{i-2})|^2 + \cdots = \sum_{j=1}^{i} \alpha_{i,j} \cdot |X_k(S_j)|^2 \quad \text{for } 1 \leqslant k \leqslant \text{NR_mel}. \tag{7}$$

The adapted spectra $\tilde{X}$ have to be transformed to the MFCCs again. In practice, mainly 2–3 preceding HMM states have an influence on the current state. This depends on the reverberation time and on the average durations of the HMM states.

The variances are not adapted. It turned out in earlier investigations (Gales, 1995; Hirsch, 2001a) that the modification of the variances has only a minor influence on the improvement of the recognition performance.

The effects of this adaptation approach are visualized in Fig. 4 in the spectral domain by comparing the spectrograms as representations of different HMMs. Spectrograms are shown in a three dimensional visualization mode.

The spectrogram of a HMM is estimated by transforming the MFCCs back to the linear spectral domain for all states. The transition probabilities $p(S_i|S_i)$ to remain in a state are taken to model the average duration of this state as defined by Eq. (3). In Fig. 4 three spectrograms of different HMM versions are shown for the word "six". Each HMM consists of 16 states where a single state is described by a set of cepstral coefficients including the zeroth cepstral coefficient C0. The spectrum of an individual HMM state is positioned with respect to its point in time t_{S_i} in the middle of the segment that is described by this state.

$$t_{S_i} = \sum_{j=1}^{i-1} \text{dur}(S_j) + \frac{\text{dur}(S_i)}{2}. \tag{8}$$

Furthermore a Spline interpolation is applied to the contour of the magnitude spectral values in each MEL band.

$$\{|X_k(t_{S_1})|, |X_k(t_{S_2})|, |X_k(t_{S_3})|, \ldots\} \xrightarrow{\text{Spline}} \{|X_k(0)|, |X_k(10\text{ ms})|, |X_k(20\text{ ms})|, \ldots\}. \tag{9}$$

Thus the spectrum can be recreated at a frame rate of 100 Hz as it is also defined by the window shift of 10 ms in the feature extraction.

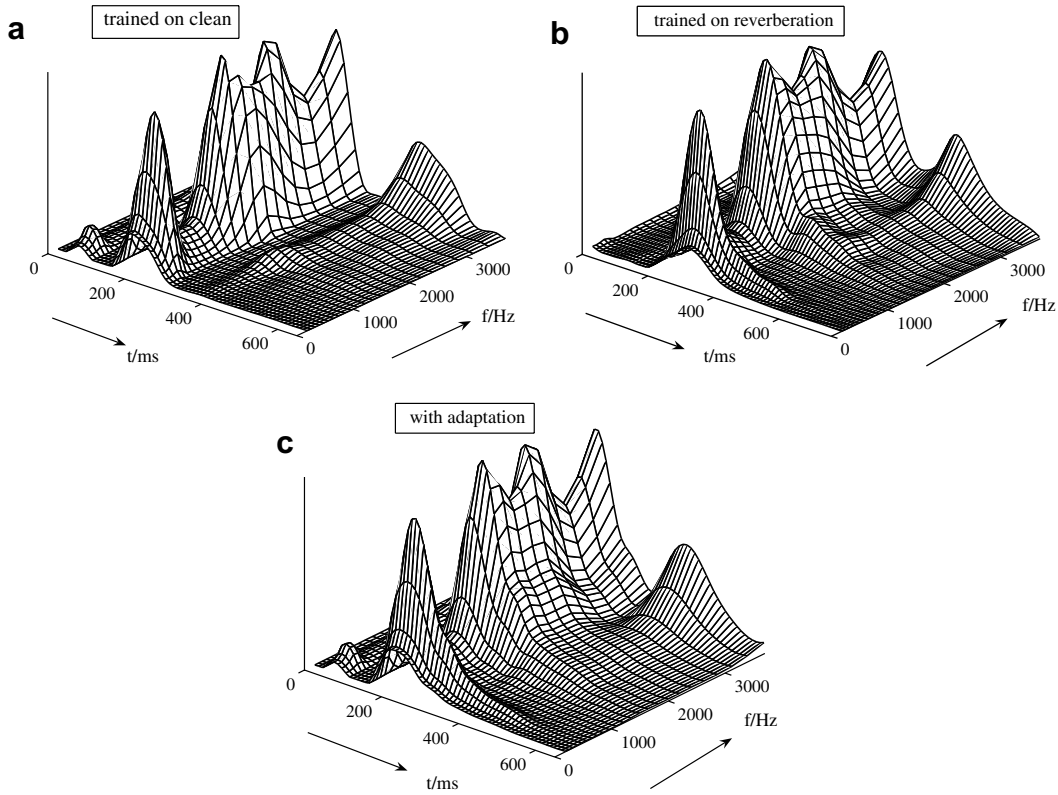

Fig. 4. Different spectrograms of HMMs for the word "six".

The spectrogram of the clean HMM is shown in graph (a) of Fig. 4 as it has been trained with the utterances of the TIDigits data base (Leonard, 1984). Only the utterances containing a single digit have been taken for the training. The contributions of the fricatives at the end can be clearly seen in the high frequency region where the formants of the vowel are visible in the middle of the word.

The spectrogram in graph (b) represents the HMM that has been trained on the TIDigits data after applying an artificial reverberation to the training utterances. The reverberation tails can be clearly seen when looking at the contours in individual MEL bands.

The spectrogram in graph (c) of Fig. 4 represents the HMM after adapting the clean HMM with the new approach. A fixed value is chosen for the reverberation time T60. The reverberation tails can also be seen in this figure. Comparing it with the spectrogram trained on reverberated data, a lot of similarities are visible. This shows that the new approach allows an adaptation of the static parameters that is comparable with training the HMM on data that have been recorded under reverberant conditions.

In practice each HMM state tries to model a certain speech segment with a mixture of Gaussian distributions for each feature component. In this case the energy parameter at state S_i and for the mixture component with the index mix_j is adapted by calculating

$$\begin{aligned}\tilde{E}(S_i, \text{mix}_j) &= \alpha_{i,i} \cdot E(S_i, \text{mix}_j) + \alpha_{i,i-1} \cdot \overline{E}(S_{i-1}) \\ &\quad + \alpha_{i,i-2} \cdot \overline{E}(S_{i-2}) + \cdots \\ &= \alpha_{i,i} \cdot E(S_i, \text{mix}_j) + \sum_{\ell=1}^{i-1} \alpha_{i,\ell} \cdot \overline{E}(S_\ell)\end{aligned}$$

$$\text{for all } S_i \text{ with } 1 \leqslant i \leqslant \text{NR_states} \quad \text{and all } \text{mix}_j \text{ with } 1 \leqslant j \leqslant \text{NR_mix}, \tag{10}$$

where $\overline{E}(S_\ell)$ is the average energy at state S_ℓ by weighting the energies of the different mixture components with the corresponding mixture weighting factors

$$\overline{E}(S_\ell) = \sum_{j=1}^{\text{NR_mix}} w(\text{mix}_j) \cdot E(S_\ell, \text{mix}_j) \quad \text{with} \quad \sum_{j=1}^{\text{NR_mix}} w(\text{mix}_j) = 1. \tag{11}$$

NR_mix is the number of Gaussian distributions for modeling this acoustic parameter in each individual HMM state.

Thus, the influence of an earlier state is considered by using the average energy at this earlier state.

In the same way the power density spectra of individual mixture components are adapted by transforming back the set of average cepstral coefficients to the spectral domain first and adapting the subband energy in each Mel band in the same way as the energy parameter.

$$\overline{C}_m = \sum_{j=1}^{\text{NR_mix}} w(\text{mix}_j) \cdot C_m(\text{mix}_j) \quad \text{for } 1 \leqslant m \leqslant \text{NR_cep}, \tag{12}$$

$$\begin{aligned} &\{\overline{C}_0, \overline{C}_1, \overline{C}_2, \ldots, \overline{C}_{\text{NR_cep}}\} \\ &\quad \xrightarrow{\text{IDCT}} \{\log(|\overline{X}_1|), \log(|\overline{X}_2|), \ldots, \log(|\overline{X}_{\text{NR_mel}}|)\} \\ &\quad \xrightarrow{\text{EXP}} \{|\overline{X}_1|, |\overline{X}_2|, \ldots, |\overline{X}_{\text{NR_mel}}|\}, \end{aligned} \tag{13}$$

$$|\tilde{X}_k(S_i, \text{mix}_j)|^2 = \alpha_{i,i} \cdot |X_k(S_i, \text{mix}_j)|^2 + \sum_{\ell=1}^{i-1} \alpha_{i,\ell} \cdot |\overline{X}_k(S_\ell)|^2. \tag{14}$$

The adapted power density spectrum is transformed to the cepstral domain again.

2.3. Estimation of T60

The estimated reverberation time T60 is the only parameter that is needed for the adaptation as it has been described in the previous section. The recognition of an utterance is done with a set of adapted HMMs where the applied value of T60 has been estimated from the recognition of the previous utterance. T60 is estimated after the recognition of an utterance by a search for this set of adapted HMMs that leads to a maximum likelihood for another forced recognition of the already recognized sequence of HMMs. The restriction to the forced recognition of the already recognized HMM sequence is introduced to limit the computational costs. This iterative process is visualized in Fig. 5.

The sequence of buffered feature vectors is used to perform the match with the previously recognized HMM sequence. At the beginning this match is performed with adapted HMMs for the previous estimate of T60 and for values of T60 that differ by ±20 ms. The values for the probability, that the feature vectors match with the sequence of adapted HMMs, are taken as input for the search of this T60 value that leads to a maximum likelihood. Dependent on the achieved probabilities the estimated value of T60 is lowered or increased by another 20 ms or the search process is stopped in case the previous estimate of T60 leads to the maximum likelihood. In case the search process is continued, another set of adapted HMMs is derived from the clean HMMs where the adaptation can be restricted to the previously recognized HMMs. This newly adapted HMMs are used for another forced match and the search for the maximum likelihood. Because the hands-free conditions will usually alter only slowly in practical applications, the modification of T60 is restricted to the range of ±40 ms from the previous estimate. Thus only a few matches are needed, so that the computational costs are fairly low.

It turns out that the estimated value of T60 varies for different speakers even though the hands-free condition is the same. This seems to be dependent on the speaking rate. Thus, this adaptation technique includes also a kind of duration modeling to some extent.

2.4. Adaptation of delta parameters

Comparing the contours of the clean and the reverberant HMM at individual Mel bins it becomes obvious that also the Delta and Delta–Delta parameters as time derivatives of the static parameters are modified by the influence of the hands-free speech input. This can be seen for example in Fig. 4 where a "valley" is visible between the vowel and the succeeding phoneme for the clean HMM. This "valley" is filled by the reverberation tails for the reverberant HMM versions. This indicates that also the time derivatives will be different in this region.

The Delta parameters are calculated in the feature extraction for the frame at time t_i as sum of weighted differences between the static parameters of preceding and succeeding frames (Young et al., 2005). For example the calculation of the Delta logarithmic energy $\Delta\log E(t_i)$ is done as

$$\Delta \log E(t_i) = \frac{\sum_{j=1}^{3} j \cdot [\log E(t_{i+j}) - \log E(t_{i-j})]}{\text{norm}}$$
$$\text{with norm} = 2 \cdot \sum_{j=1}^{3} j^2, \tag{15}$$

where $\ldots, t_{i-1}, t_i, t_{i+1}, \ldots$ describe the window shift by 10 ms.

The Delta parameters of the reverberant speech are estimated as described below by looking at the adapted static parameters of all HMM states. The average logarithmic

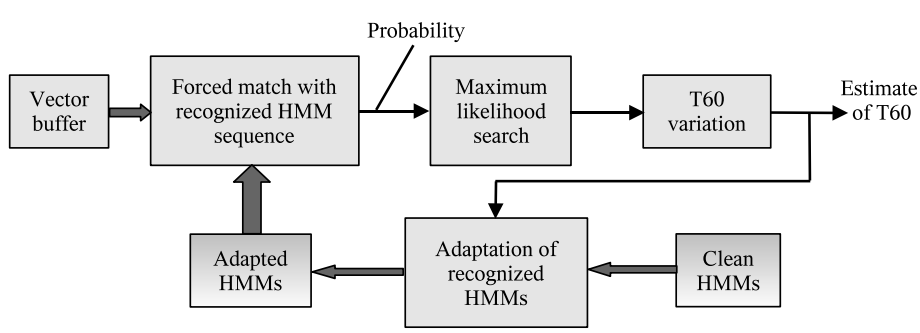

Fig. 5. Estimation of T60 by an iterative search of the maximum likelihood.

frame energies of all states are considered for a single HMM. A Spline interpolation is applied to recreate the average energy contour at a frame rate of 100 Hz.

$$\{\log \overline{E}(t_{S_1}), \log \overline{E}(t_{S_2}), \log \overline{E}(t_{S_3}), \ldots\} \xrightarrow{\text{Spline}} \{\log \overline{E}(0), \log \overline{E}(10\text{ ms}), \ldots, \log \overline{E}(t_i), \ldots\}. \quad (16)$$

In the same way an interpolated version of the average energy contour can be calculated for the average frame energies of the clean HMM.

The procedure for calculating the Deltas in the feature extraction, as described by Eq. (15), can be applied to the interpolated energy contours of the clean and the adapted HMM. Thus the average logarithmic Delta energies $\Delta \log \overline{E}_{\text{clean}}(t_i)$ and $\Delta \log \overline{E}_{\text{adapted}}(t_i)$ become available for the clean and for the adapted HMM version at the frame rate of 100 Hz ($t_i = 0, 10\text{ ms}, 20\text{ ms}, \ldots$). As the number of frames is equal for the clean and the adapted HMM of an individual word class the difference between the clean and the adapted Delta energies can be calculated

$$\Delta \log \overline{E}_{\text{diff}}(t_i) = \Delta \log \overline{E}_{\text{adapted}}(t_i) - \Delta \log \overline{E}_{\text{clean}}(t_i) \quad \text{for all } t_i = 0, 10\text{ ms}, 20\text{ ms}, \ldots \quad (17)$$

These values describe the average differences between the Delta logarithmic energies of the adapted and the clean HMM at each frame. By means of a Spline interpolation the average differences are calculated for all HMM states.

$$\{\Delta \log \overline{E}_{\text{diff}}(0), \Delta \log \overline{E}_{\text{diff}}(10\text{ ms}), \ldots, \Delta \log \overline{E}_{\text{diff}}(t_i), \ldots\} \xrightarrow{\text{Spline}} \{\Delta \log \overline{E}_{\text{diff}}(t_{S_1}), \Delta \log \overline{E}_{\text{diff}}(t_{S_2}), \Delta \log \overline{E}_{\text{diff}}(t_{S_3}), \ldots\}. \quad (18)$$

A weighted version of these average differences is added to the corresponding Delta parameters of the clean HMM to create a set of adapted Delta parameters.

$$\Delta \log \tilde{E}(t_{S_i}, \text{mix}_j) = \Delta \log E_{\text{clean}}(t_{S_i}, \text{mix}_j) + \beta \cdot \Delta \log \overline{E}_{\text{diff}}(t_{S_i}) \quad \text{for all } S_i \text{ with } 1 \leqslant i \leqslant \text{NR_states and all mix}_j \text{ with } 1 \leqslant j \leqslant \text{NR_mix}. \quad (19)$$

This is done individually for each state and for each mixture component. A factor β is introduced for the weighted summation of the differences. During recognition experiments we found a value of 0.7 for β to achieve highest performance.

The Delta cepstral parameters can be adapted in the same way. The average logarithmic Mel spectral values are taken as basis as they can be calculated by Eqs. (12) and (13) from the average cepstral coefficients for each HMM state. A Spline interpolation can be applied to recreate the contour of the logarithmic Mel magnitude in each Mel band at the frame rate of 100 Hz.

$$\{\log |\overline{X}_k(t_{S_1})|, \log |\overline{X}_k(t_{S_2})|, \log |\overline{X}_k(t_{S_3})|, \ldots\} \xrightarrow{\text{Spline}} \{\log |\overline{X}_k(0)|, \log |\overline{X}_k(10\text{ ms})|, \ldots, \log |\overline{X}_k(t_i)|, \ldots\} \quad \text{for } k = 1, 2, \ldots, \text{NR_mel}. \quad (20)$$

The logarithmic spectral domain seems to be the right domain for applying the Spline interpolation even though the interpolation could also be applied to the average cepstral parameters. The interpolated average logarithmic spectra are transformed to the cepstral domain

$$\{\log(|\overline{X}_1(t_i)|), \log(|\overline{X}_2(t_i)|), \ldots, \log(|\overline{X}_{\text{NR_mel}}(t_i)|)\} \xrightarrow{\text{DCT}} \{\overline{C}_0(t_i), \overline{C}_1(t_i), \ldots, \overline{C}_{\text{NR_cep}}(t_i)\} \quad \text{for } t_i = 0, 10\text{ ms}, \ldots \quad (21)$$

The Delta coefficients can be calculated for the contour of each individual average cepstral coefficient. This can be done again for the clean as well as for the adapted HMM so that the difference between these two versions can be estimated

$$\Delta \overline{C}_{m_{\text{diff}}}(t_i) = \Delta \overline{C}_{m_{\text{adapted}}}(t_i) - \Delta \overline{C}_{m_{\text{clean}}}(t_i) \quad \text{for all } t_i = 0, 10\text{ ms}, 20\text{ ms}, \ldots \text{ and for } 1 \leqslant m \leqslant \text{NR_cep}. \quad (22)$$

These values describe the average differences between the Delta cepstral coefficients of the adapted and the clean HMM at each frame. By means of a Spline interpolation the average differences are calculated individually for each cepstral coefficient for all HMM states

$$\{\Delta \overline{C}_{m_{\text{diff}}}(0), \Delta \overline{C}_{m_{\text{diff}}}(10\text{ ms}), \ldots, \Delta \overline{C}_{m_{\text{diff}}}(t_i), \ldots\} \xrightarrow{\text{Spline}} \{\Delta \overline{C}_{m_{\text{diff}}}(t_{S_1}), \Delta \overline{C}_{m_{\text{diff}}}(t_{S_2}), \Delta \overline{C}_{m_{\text{diff}}}(t_{S_3}), \ldots\} \quad \text{for } 1 \leqslant m \leqslant \text{NR_cep}. \quad (23)$$

The adapted cepstral coefficients can be calculated by adding a weighted version of the average differences to the Delta coefficients of the clean HMM

$$\Delta \tilde{C}_m(S_i, \text{mix}_j) = \Delta C_{m_{\text{clean}}}(S_i, \text{mix}_j) + \beta \cdot \Delta \overline{C}_{m_{\text{diff}}}(S_i) \quad \text{for all } S_i \text{ with } 1 \leqslant i \leqslant \text{NR_states and all mix}_j \text{ with } 1 \leqslant j \leqslant \text{NR_mix and for } 1 \leqslant m \leqslant \text{NR_cep}. \quad (24)$$

The adaptation of each cepstral coefficient is done individually for each state and for each mixture component. The value of β is the same as applied in Eq. (19) for the energy.

Fig. 6 summarizes and visualizes this new technique for calculating the differences between the Delta coefficients that can be derived from the static parameters of the clean and the adapted HMMs. It is not only applicable in case of adapting HMMs to the conditions of a hands-free speech input. This method can be applied in any case of adapting the static parameters of HMMs to the influence of distortion effects.

The Delta–Delta parameters can be adapted in the same way as the Delta parameters. The Delta–Delta parameters

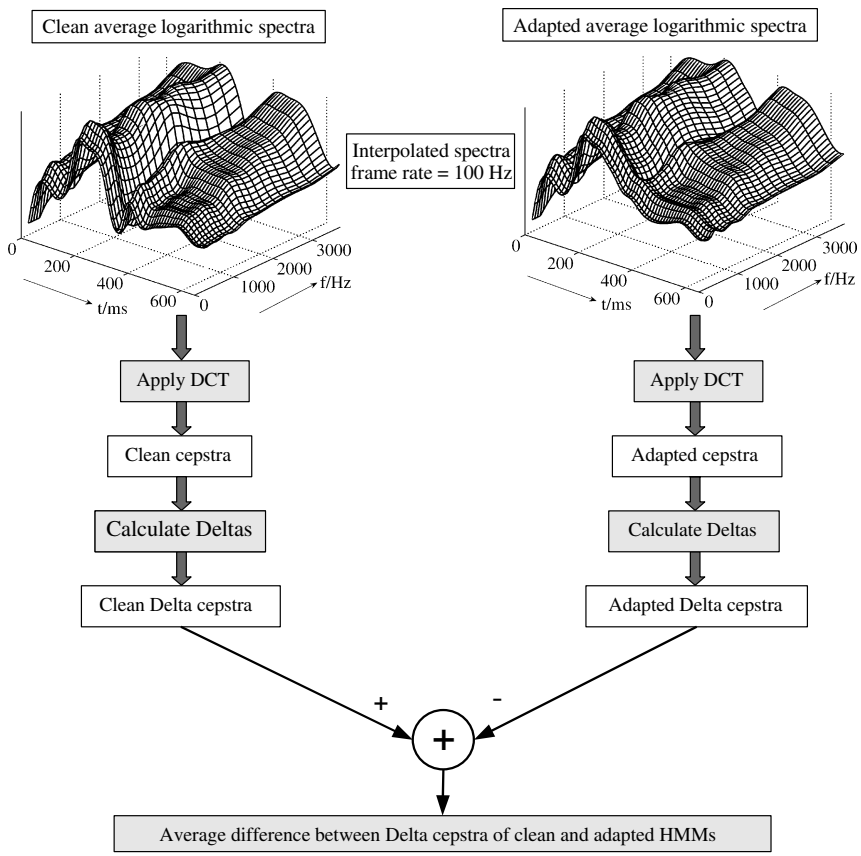

Fig. 6. Scheme for estimating the differences between the Delta parameters of clean and adapted HMMs.

are calculated from the Delta parameters in the same way the Delta parameters are determined from the static parameters. Looking at Eq. (15) the only difference is a value of 2 for the higher summation index. This results in a calculation of the Delta–Delta parameters over five sets of Delta parameters. Otherwise the adaptation of the Delta–Delta parameters is done as described by Eqs. (16)–(24).

2.5. Restrictions in case of a connected word recognition

The complete adaptation scheme as described above works well for whole word HMMs in case of an isolated word recognition. For the recognition of connected words the adaptation will not be perfect when uttering a sequence of words without even short pauses between the words. The beginning of a word is modified by the reverberation tail of the acoustic excitation at the ending of the preceding word. These effects can be taken into account only in a type of online adaptation. To get knowledge about the preceding word, the log likelihood is observed at the final states of all HMMs during the frame-wise recognition with the Viterbi algorithm. The adaptation of the first frames of all HMMs can be done when a high log likelihood is computed for the final states of a HMM so that is very likely that the corresponding word was spoken. Thus the energy and spectral parameters, that are contained in the final states of the HMM with the high likelihood, can be used to adapt the first states of all HMMs.

The implementation of this online technique is quite complex because it is based on a frame-wise decision process whether and which HMM creates a high likelihood at its final states. The authors implemented this approach but could find only small improvements for a connected digit recognition with respect to word accuracy. Also because of the high computational effort the approach was not investigated further.

2.6. Adaptation of triphone HMMs

Most often triphone HMMs are used in case of a phoneme based recognition which is used for the recognition of large vocabularies. A triphone HMM is applied to model the acoustic characteristics of a phoneme in the context of a specified preceding and a specified succeeding phoneme. Thinking about the adaptation of this triphone HMM, the knowledge about the preceding phoneme and the succeeding phoneme can be taken to apply the adaptation approach as it was presented for the application to whole word HMMs in the previous sections. Looking at a single triphone HMM as it is done for the triphone "s-I-k" in Fig. 7, the number of possible preceding and succeeding triphone HMMs is restricted.

The triphone HMM "s-I-k" for the vowel "i" in the context of a preceding "s" and a succeeding "k" will be

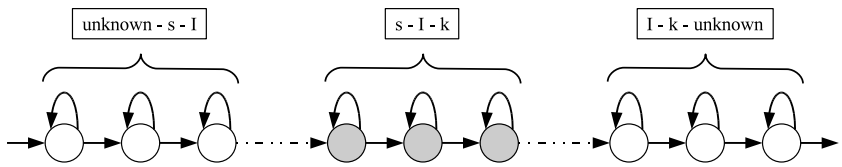

Fig. 7. Possible preceding and succeeding triphone HMMs for one selected triphone HMM.

preceded by a triphone HMM "unknown-s-I" for the fricative "s" with the vowel "i" following. Only the preceding phoneme of "s" is not defined. In the same way a triphone HMM "I-k-unknown" for the "k" will succeed the triphone HMM for the "i". This knowledge about the sequence of HMMs and hence also about the complete sequence of states enables the applicability of the new approach for the adaptation to reverberation. Looking at the sequence of three triphones as shown in Fig. 7, the complete sequence of nine HMM states for all three models is used for the adaptation. It is done in the same way as it has been described for the whole-word HMMs before.

Thus, the problem of being unable to adapt the first states of a whole word HMM, does not exist in case of triphone models where a certain knowledge about the preceding model is available.

This preceding model "unknown-s-I" is chosen from all available triphone HMMs by looking at the spectral similarity between the last state of each triphone HMM "unknown-s-I" and the first state of the HMM "s-I-k". The spectral similarity is estimated by calculating the Euclidean distance between the average cepstral coefficients of the first state of "s-I-k" and the corresponding coefficients of the last state of the preceding phoneme. The triphone HMM with the smallest spectral distance is selected

$$\underset{\text{unknown}\in\{\text{all phonemes}\}}{\text{MIN}} \sum_{m=1}^{\text{NR_cep}} \left[\overline{C}_m\{S_1(\text{“s-I-k”})\} - \overline{C}_m\{S_{\text{last}}(\text{“unknown-s-I”})\}\right]^2. \tag{25}$$

In case of modeling with a mixture of Gaussian distributions the average cepstral coefficients are calculated by taking into account the mixture weights

$$\overline{C}_m = \sum_{j=1}^{\text{NR_mix}} w(\text{mix}_j) \cdot C_m(\text{mix}_j) \quad \text{for } m = 1, 2, \ldots, \text{NR_cep}. \tag{26}$$

In the same way the preceding triphone is chosen by comparing the last state of "s-I-k" with the first state of all triphones "I-k-unknown"

$$\underset{\text{unknown}\in\{\text{all phonemes}\}}{\text{MIN}} \sum_{m=1}^{\text{NR_cep}} \left[\overline{C}_m\{S_{\text{last}}(\text{“s-I-k”})\} - \overline{C}_m\{S_1(\text{“I-k-unknown”})\}\right]^2. \tag{27}$$

The static and the Delta parameters are adapted for the states of the HMM "s-I-k" by looking at the whole state sequence of all three consecutive triphone HMMs and applying the adaptation technique as described for the whole-word HMMs. By transforming back the nine sets of cepstral coefficients to the spectral domain and applying a Spline interpolation the spectrogram for the complete segment of the three triphones is available. The knowledge of the preceding phoneme has the advantage that the energy and spectral information of these states can be used to adapt the succeeding states of the HMM "s-I-k" with respect to reverberation. Hence, the drawback of not knowing the preceding HMM for the adaptation of the first states does not exist as it occurs in case of whole-word HMMs. The knowledge of the succeeding triphone is only used for the better estimation of the adapted Delta parameters.

This adaptation technique can be applied to all triphone HMMs that are used for the recognition. In case of using state tying a further tying can be applied after the adaptation. This has not been investigated here. The authors only intended to demonstrate the principal applicability of the new adaptation method to the modeling with triphone HMMs.

3. Recognition experiments on hands-free speech input

Recognition experiments have been run to proof the applicability of the new adaptation approach and to quantify the improvements that can be achieved.

Some details about the applied feature extraction and the recognition are presented at the beginning of this chapter. After this the results are shown for a series of experiments with the intention to demonstrate the applicability to a word recognition based on whole word HMMs. Finally the improvements are presented separately for the recognition of a large vocabulary based on the use of triphone HMMs.

3.1. Feature extraction

The acoustic features are extracted from the speech signal by a cepstral analysis scheme that is similar to many realizations in this field. A pre-emphasis is applied to the speech signal by means of a first order FIR filter where the value of the preceding sample is weighted by a factor of 0.95 before subtracting it from the value of each sample. Short segments of speech are extracted with a 25 ms

Hamming window. The window is shifted by 10 ms which corresponds to a frame rate of 100 Hz. Each speech frame is transformed to the spectral domain by means of a 256 point DFT (Discrete Fourier Transform). The so called MEL spectrum is estimated by weighting the values of the DFT magnitude spectrum with triangular shaped functions and summing up the spectral magnitudes for each triangular. Thus, a MEL spectrum is computed for 24 nonlinearly distributed frequency bands in the range from 200 Hz up to 4000 Hz. The 24 logarithmic MEL spectral values are transformed to the cepstral domain by means of a DCT (Discrete Cosine Transform). Thirteen cepstral coefficients C0 to C12 are calculated. Thus, C0 is available as acoustic parameter in each state of all HMMs. C0 is only needed to transform back the cepstral coefficients to the spectral domain as part of the adaptation process. But C0 is not used for the recognition. Instead of C0 an energy parameter is estimated from the preemphasised and Hamming weighted speech samples. A preemphasis factor of -1 is applied here which leads to a slightly higher attenuation of the low frequency components. This is of advantage in the presence of background noise with its main energy at low frequencies. The short-term energy is calculated by summing up the squared values of all samples in each 25 ms frame.

The described analysis technique is applied to speech data sampled at 8 kHz. In case of data sampled at 16 kHz the same MEL filterbank is applied in the frequency range up to 4000 Hz. All filter characteristics like e.g. the preemphasis filtering or the MEL filters and all other individual settings like e.g. the frame length or the FFT length are chosen in an appropriate way so that the final cepstral coefficients are almost identical to the coefficients which result from an analysis of the same utterance sampled at 8 kHz. This approach has been investigated in earlier work (Hirsch et al., 2001b). It allows for example the recognition of speech data sampled at 8 kHz with HMMs that have been trained on data sampled at 16 kHz.

Besides the 13 cepstral coefficients and the energy parameter in the range up to 4 kHz two further parameters are calculated in case of data sampled at 16 kHz. These are two energy parameters that describe the energy in the frequency range from 4 to 5.5 kHz respectively in the range from 5.5 to 8 kHz. This is realized by summing up the corresponding components of the FFT power density spectrum. These additional coefficients can help to increase the recognition performance a bit in comparison to the case of recognizing data sampled at 8 kHz.

Twelve cepstral coefficients C1–C12 and the logarithm of the energy parameter are used as acoustic parameters for the recognition. Furthermore Delta and Delta–Delta coefficients are added as additional features where the Delta parameters are calculated as described by Eq. (15). The Delta calculation corresponds to the way of estimating Delta parameters in the HTK software package (Young et al., 2005).

Thus, finally a feature vector consists of 39 components in case of speech data sampled at 8 kHz and it consists of 45 components in case of data sampled at 16 kHz.

The parameters of the HMMs are determined by applying the available training tools of the HTK software package. The recognition is done either with an own C implementation of a Viterbi recognizer or with the corresponding tool of the HTK package. It has been verified that the own implementation leads to the same recognition results as the HTK recognizer. Furthermore the Viterbi recognizer has also been implemented as Matlab module. This was helpful during the development process of the adaptation algorithms due to the easier software development with Matlab and its graphical visualization properties.

The adaptation techniques have been implemented as Matlab and as C modules. The adaptation is individually applied to each speech utterance when detecting the beginning of speech. The applied voice activity detector takes the MEL magnitude spectrum as input. It will be described a bit more in detail later. The C modules for the analysis and the recognition are designed for an application in a real-time recognition and dialogue system (Hirsch, 1999). This means that the Viterbi match can be started and run in parallel to the feature extraction after the detection of speech. Techniques like a cepstral mean normalization on the whole utterance are not considered here because they would delay the beginning of the Viterbi match till the detection of the end of speech.

3.2. Recognition with whole-word HMMs

The TIDigits data base is taken as basis for the experiments on isolated and connected word recognition. A version of the TIDigits is used that has been downsampled at 8 kHz. All utterances from the adult speakers designated for training are taken to determine two gender dependent HMMs for each word. Each HMM consists of 16 states where each state is described by the mixture of two Gaussian distributions for each of the 39 acoustic features. A single state HMM with a mixture of eight Gaussians is used for modeling the pauses. The HMMs are defined as left-to-right models without skips over states. The recognizer is set up to recognize any sequence of digits with the restriction that a sequence contains only models from the same gender.

3.2.1. Recognition of single digits

A first series of experiments focused on these test utterances that contain only a single digit. These are about 2500 utterances in total. This is done to avoid the inter-word effects between fluently spoken words without pauses between the words. As already mentioned before, the reverberation will influence the beginning of a word by the ending of the preceding word.

The word error rates are shown in Fig. 8 where the recognition of connected words is still enabled so that also insertion errors can occur. Results are presented for three

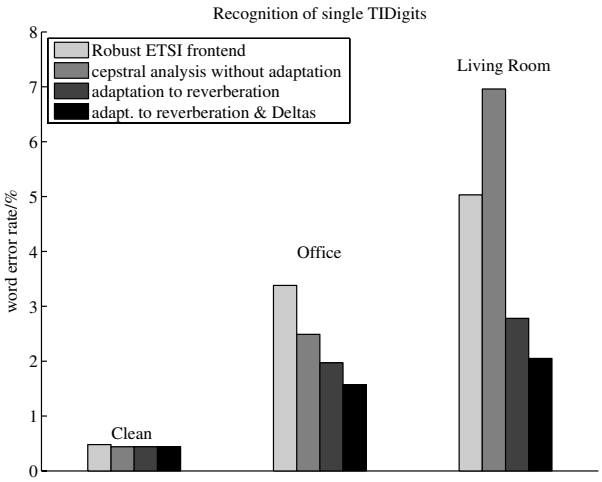

Fig. 8. Word error rates for these TIDigits utterances that contain only single digits.

different conditions. Besides the clean data the TIDigits have been processed with a tool for simulating the influence of a hands-free speech input (Hirsch and Finster, 2005). This tool creates fairly natural sounding reverberation. A Web interface exists to experience this tool (Finster, 2005). Test data have been created for the simulation of two different rooms, an office room with a reverberation time of about 0.4 s and a living room with a reverberation time of about 0.6 s. As expected the error rates are higher for the longer reverberation time.

For each condition four different processing methods are compared. The first one is based on the advanced front-end as it has been standardized by ETSI (2003). Robust acoustic features are extracted with this front-end. The term robustness refers to the presence of background noise and unknown frequency characteristics. This is realized by extending a cepstral analysis scheme by two further processing steps. The first one contains a Wiener filtering based on an estimation of the noise spectrum to reduce the influence of stationary background noise. A blind estimation and equalization of unknown frequency characteristics has been integrated as second processing block. Each feature vector consists of 39 parameters. These are 12 Mel frequency cepstral coefficients and an energy parameter as well as the corresponding Delta and Delta–Delta parameters. Feature vectors are computed at a rate of 100 Hz. This front-end is considered as a representative for a robust feature extraction and is taken as reference for comparing the results with the HMM adaptation.

The cepstral analysis scheme as it has been described in the previous section is investigated as second method. Word error rates are presented for the three cases where the recognition is done

- without any adaptation or
- with adaptation of the static parameters only or
- with adaptation of the static and the Delta and Delta–Delta parameters.

The error rates for the clean data are in the range of 0.4–0.5%. It can be seen that the influence of the reverberation leads to a considerable deterioration of the recognition rates for both feature extraction schemes. The high error rate of the cepstral analysis scheme in comparison to the ETSI front-end in case of the living room condition is due to a high number of insertion errors. The number of substitutions is even lower for the cepstral analysis scheme in comparison to the ETSI scheme.

Error rates can be reduced by applying the adaptation methods. Adapting also the Delta and Delta–Delta parameters leads to an additional gain in both reverberant situations.

The efficiency of the adaptation scheme is investigated by comparing the obtained results to the case of training the HMMs on reverberated data. Therefore a set of HMMs is trained with all TIDigits training utterances after applying the simulation of a reverberation in the living room.

Results are listed in the first two lines of Table 1 for the cases without adaptation and with adaptation of the static and Delta parameters and taking HMMs trained on clean data only. These are the error rates as already shown in Fig. 8. The third line of Table 1 contains the error rates for the case of applying the HMMs trained on reverberated data and without any adaptation.

The error rate decreases from about 7% to 1.7% for the living room condition when moving from the training on clean data and applying no adaptation to the training on reverberant data. Applying the adaptation scheme to the clean HMMs leads to an error rate quite close to the case of training on reverberated data. This can be taken as further proof for the usefulness of the applied adaptation method.

The drawback of training the HMMs on reverberant data is a considerable increase of the error rate to about 11% for the recognition of clean data. This indicates that the training has to be done on a mixture of conditions for a practical application. And this can only be done if the whole range of conditions is known in advance.

3.2.2. Connected word recognition

The word error rates are shown in Fig. 9 for the recognition of all TIDigits utterances that have been designated for recognition. These are 8700 utterances containing about 28 000 digits in total. The results are presented for

Table 1
Word error rates for the recognition of single TIDigits applying the cepstral analysis

	Condition		
	Clean (%)	Office room (%)	Living room (%)
Without adaptation	0.44	3.49	6.94
Adaptation to reverberation & Deltas	0.44	1.57	2.05
HMMs trained on living room	10.98	1.93	1.73

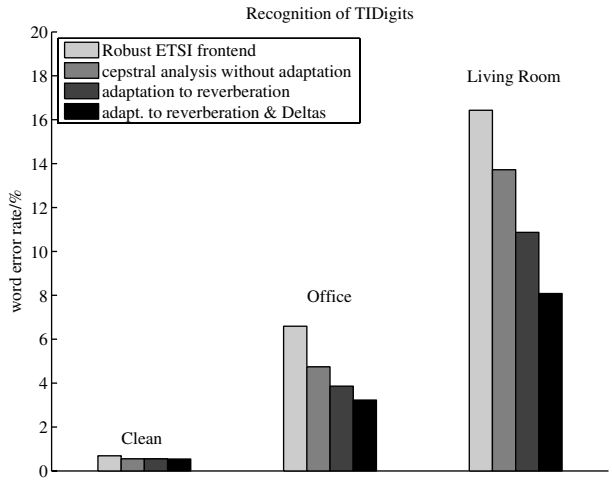

Fig. 9. Word error rates for the recognition of all TIDigits.

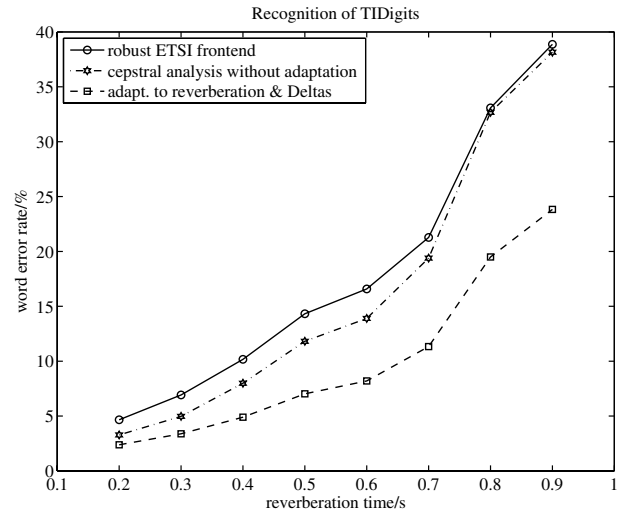

Fig. 10. Word error rates for a variation of the reverberation time in the living room.

the four conditions that have also been shown in the previous figure.

First of all the error rates for the robust ETSI front-end are higher in comparison to applying the cepstral analysis scheme without any adaptation. For the condition of a hands-free speech input in reverberant environments it looks like the ETSI front-end does not work as efficient as it does in the presence of background noise.

The application of the adaptation method leads also to a reduction of the error rates for the case of recognizing sequences of connected words. The relative improvement is not as high as in case of recognizing single digits. The authors mainly regard the superposition effect at the beginning of words as responsible for this. The acoustic information at the beginning of a word is modified by the acoustic information of the preceding word due to the reverberation. These "inter-word" modifications occur especially when sequences of words are spoken fluently with coarticulation effects. In general the speaking rate considerably varies between speakers when uttering a sequence of digits. This effect can only be approximately covered by modeling with HMMs with multiple mixture components. Analyzing the errors a bit more in detail, it turns out that about half of the errors are due to deletions in case of recognizing the living room data with adaptation. It seems to be difficult to recognize especially the "fast" speakers which create these co-articulation effects. The "inter-word" modifications are not compensated by this adaptation technique.

We observe again that the additional adaptation of the Delta and Delta–Delta parameters causes a further gain in recognition performance.

Further recognition results are presented in Fig. 10 for varying the reverberation time in the living room condition. The tool for simulating the hands-free speech input in noisy environments allows the variation of the reverberation time in a certain range. The RIR for the living room simulation is modified inside the tool so that it reflects the desired reverberation time.

As expected the error rate increases for higher reverberation time. The error rates for a reverberation time of 0.6 s are the ones shown in the previous figure. Again the recognition performance is slightly worse for applying the robust ETSI front-end in comparison to the cepstral analysis scheme. The improvement due to the adaptation is visible over the whole range of the reverberation time.

3.3. Recognition with triphone HMMs

Triphone HMMs are used for the recognition of a large vocabulary. The "Wall Street Journal" data base (WSJ0) (LDC, 1993) is taken as basis for these investigations as it has also been used for the evaluations inside the ETSI working group Aurora (Picone et al., 2004). Approximately 7200 utterances that have been recorded at a sampling rate of 16 kHz with a high quality microphone are taken for the training of triphone models. The triphones are modeled as HMMs with three states where each acoustic parameter of each state is described by a mixture of 4 Gaussian distributions. Training and recognition are done with HTK as it has been defined in (Au Yeung and Siu, 2004). The training procedure includes state tying to model the triphones with a total of about 3200 different states. The recognition is based on the usage of a dictionary containing the phoneme description of about 5000 words. The possible sequences of words are defined by a "bigram" model. The recognition process is speeded up by a state based pruning.

The cepstral analysis scheme is applied as it has been described before for data sampled at 16 kHz. We achieve a word error rate of 11.21% for the recognition of the 330 clean utterances that have been designated for testing. The word error rate can be reduced by applying HMMs that model acoustic parameters with a higher number of distributions. Because of the high computational costs,

Table 2
Word error rates for the recognition of the "Wall Street Journal" large vocabulary

Clean	Office without adaptation	Office with fixed adaptation
11.2%	48.8%	39.8%

only experiments have been run with HMMs modeling with a mixture of four distributions.

All 330 test utterances have been processed with the simulation tool to investigate the recording in an office room with a reverberation time of about 0.4 s. The word error rates for recognizing these data are listed in Table 2.

The word error rate increases considerably to a value of 48.8% in the hands-free mode.

The adaptation of the triphone HMMs is applied as described before. As the own implementation of the Viterbi recognizer does not support the use of complex language models, HTK is employed for the recognition. Thus, the estimation of the reverberation time T60 as well as the individual adaptation for each utterance is not applicable. The whole set of triphone HMMs is adapted once at the beginning with a fixed value for the estimated reverberation time T60 instead. The adaptation is implemented as Matlab functions. The intention of the authors is only the proof that the new adaptation approach can be applied to triphone HMMs in principal. The word error rate decreases by about 10% when applying the set of adapted triphone HMMs.

4. Combined adaptation to different distortion effects

The hands-free speech input in a room comes along with the recording of background noise as it is present in almost all applications of speech recognition systems. Furthermore the spectrum of the speech is modified by the frequency characteristics of the microphone and of an additional transmission channel, e.g. in case of transmitting the speech via telephone to a remote recognition system. This creates the need to compensate also these distortion effects.

In earlier work (Hirsch, 2001a) the authors developed an adaptation scheme based on the well-known PMC (parallel model combination) approach. This scheme consists of an adaptation of the static Mel frequency cepstral coefficients. The cepstral coefficients are transformed back to the Mel spectral domain where the adaptation can be realized by a multiplication with a frequency weighting function as estimate for the frequency characteristics and by adding the estimated noise spectrum. The cepstral coefficients of all HMMs are individually adapted for each speech utterance when the beginning of speech is detected. Furthermore the energy parameter can be adapted with an estimate of the noise energy.

We present a short overview about the techniques for estimating the spectrum of the background noise and the frequency weighting function in the next section. Having these estimates as well as an estimation of T60, it will be shown that the earlier adaptation approach (Hirsch, 2001a) can be combined with the new method of adapting the spectra to a hands-free speech input.

4.1. Estimation of distortion parameters

The spectrum of the background noise is estimated by looking at a smoothed version of the Mel magnitude spectrum as it is calculated in the feature extraction. The contour of the spectral magnitude values is smoothed in each Mel subband by applying a first order recursive filtering

$$Xsmooth_k(t_i) = (1-\alpha)\cdot|Xi_k(t_i)| + \alpha\cdot Xsmooth_k(t_{i-1}) \quad \text{for } 1 \leqslant k \leqslant \text{NR_mel and } t_i = 0, 10\text{ ms}, 20\text{ ms}, \ldots \tag{28}$$

where $Xi_k(t_i)$ is the Mel spectrum of the analysis frame at time t_i as calculated in the feature extraction.

A VAD (voice activity detector) is applied that takes the Mel spectra as input. A speech onset is detected when the estimated signal-to-noise ratios exceed an adaptive threshold in several subbands for a certain number of frames. The VAD was developed for earlier investigations. More details can be found in (Hirsch and Ehrlicher, 1995; Hirsch, 2001a).

When the beginning of speech is detected the estimated noise spectrum is set to the smoothed spectrum of the last analysis frame that is marked as pause frame

$$N_k = Xsmooth_k(\text{last pause frame}) \quad \text{for } 1 \leqslant k \leqslant \text{NR_mel}. \tag{29}$$

Furthermore the energy of the noise is estimated as energy of the last pause frame

$$Enoise = Ei(\text{last pause frame}), \tag{30}$$

where $Ei(t_i)$ is the energy of the analysis frame at time t_i as calculated in the feature extraction.

The detection of speech begin is also taken as trigger point to perform the adaptation of all HMMs. For the simulation experiments we take the acoustic parameters of all frames from a recorded utterance as input for the Viterbi recognition. For the real-time version of the recognizer as it is applied in a speech dialogue system, we start the recognition process five frames earlier than the first frame detected as speech. Thus, the Viterbi calculation can be run almost in parallel with the feature extraction.

The frequency weighting function is estimated after the recognition of an utterance. It is applied for the recognition of the next utterance. This is based on the assumption that the frequency characteristics of the whole speech transmission will not change rapidly. Usually the microphone and the other transmission conditions do not change during a recognition session. The weighting function is estimated by comparing the long-term spectra of the noisy input speech and of the clean speech. The "best" sequence of

HMM states is considered as it is available after the Viterbi match by backtracking the path with the highest likelihood. The long-term spectrum of the noisy input speech is calculated for all analysis frames that are mapped on speech HMMs excluding the frames that are mapped on the pause model

$$Xlong_k = \frac{1}{\text{NR_speech}} \cdot \sum_{\text{speech frames}} |Xi_k(t_i)| \quad \text{for } t_i \in \{\text{feature vectors mapped on speech HMMs}\}, \tag{31}$$

where NR_speech is the total number of vectors mapped on speech HMMs.

In a similar way the long-term spectrum of the clean speech is estimated by looking at the spectral information contained in the HMM states at the path with highest likelihood. A set of adapted HMMs is used for the recognition. But for the estimation of the clean spectrum the spectral information is extracted from the corresponding clean HMMs. The cepstral coefficients of the corresponding clean HMM states are transformed back to the Mel spectral domain. In case of HMMs with multiple mixture components, the spectrum of this mixture component with the smallest spectral distance to the corresponding spectrum of the input signal is taken. Therefore, the estimated noise spectrum is subtracted from the spectrum of the input signal to compare it with the spectrum contained in a clean HMM. The spectral similarity is calculated as City block distance. So, the long-term spectrum of the clean speech can be estimated

$$Slong_k = \frac{1}{\text{NR_speech}} \cdot \sum_{\text{speech frames}} |X_k[\text{Mclean}(t_i), S(t_i), \text{mix}(t_i)]| - |\overline{N}sil_k| \quad \text{for } t_i \in \{\text{feature vectors mapped on speech HMMs}\} \tag{32}$$

with Mclean(t_i) and $S(t_i)$ as recognized model and state on the best path and mix(t_i) as mixture component with smallest spectral distance.

$\overline{N}sil$ is the Mel spectrum that can be derived from the single state pause model. Calculating the average cepstral values of the pause state according to Eq. (12), the spectrum can be determined by transforming the cepstral values to the spectral domain (Eq. (13)). The pause model contains the spectral information of the background noise that was present during the recording of the training data. In case of "clean" training data, $\overline{N}sil$ takes only small values. It is subtracted here to compensate its presence in the spectral parameters of all HMMs. In the rare case of getting a negative value after the subtraction the result is set to a fixed small positive value.

Subtracting the estimated noise spectrum from the long-term spectrum of the noisy input speech, the frequency weighting function can be estimated

$$W_k = \frac{Xlong_k - N_k}{Slong_k} \quad \text{for } 1 \leqslant k \leqslant \text{NR_mel}. \tag{33}$$

It turned out in earlier investigations that this type of estimating the spectral difference between the input signal and the clean HMMs works well. Because of comparing the spectral information from the input signal and the clean HMMs, the weighting function does not only contain the spectral characteristics of the recording equipment and the transmission line but also the frequency characteristics of the individual speaker to some extent.

In the same way the difference between the energy contours of the input speech and the best HMM sequence can be calculated. The energy values of the input signal are accumulated for those frames mapped on speech HMMs

$$Einput = \frac{1}{\text{NR_speech}} \cdot \sum_{\text{speech frames}} Ei(t_i) \quad \text{for } t_i \in \{\text{feature vectors mapped on speech HMMs}\}. \tag{34}$$

The energy parameters contained in the clean HMMs on the best path are accumulated as estimate for the clean energy. Models, states and mixture components are selected as described before (Eq. (32))

$$Eclean = \frac{1}{\text{NR_speech}} \cdot \sum_{\text{speech frames}} |E[\text{Mclean}(t_i), S(t_i), \text{mix}(t_i)]| - |\overline{E}sil| \quad \text{for } t_i \in \{\text{feature vectors mapped on speech HMMs}\}. \tag{35}$$

The average energy $\overline{E}sil$ of the single state pause model is subtracted to compensate the presence of background noise in the training data.

A weighting factor can be calculated that describes the average energy difference between the input signal and the energies contained in the sequence of HMM states on the best path

$$we = \frac{Einput - Enoise}{Eclean}. \tag{36}$$

This factor contains information about the loudness of the individual speaker.

4.2. Combined adaptation

Having estimates for the noise spectrum, the frequency weighting function and the reverberation time, the Mel spectra of the clean HMMs are adapted as shown in Fig. 11.

The cepstral coefficients of each state and mixture component are transformed back to the linear Mel spectrum for all clean HMMs. The Mel spectra are adapted to the estimated reverberation condition as described by Eq. (14). The estimated weighting function and the estimated noise spectrum are applied for the further adaptation

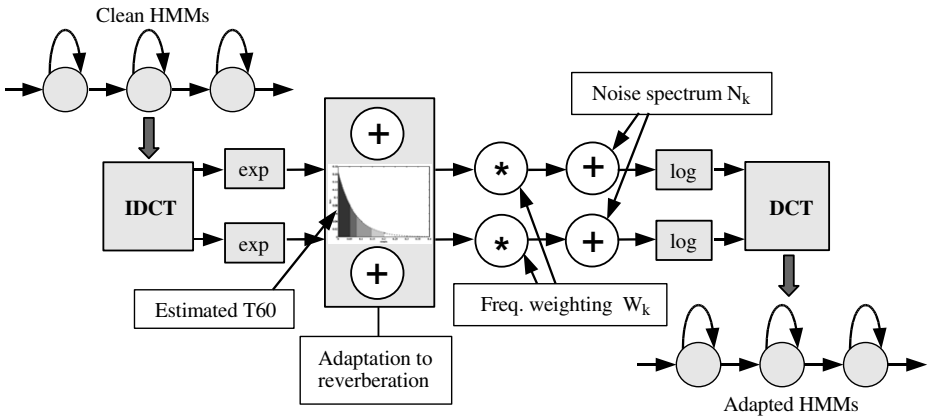

Fig. 11. Scheme for adapting HMMs to all distortion effects.

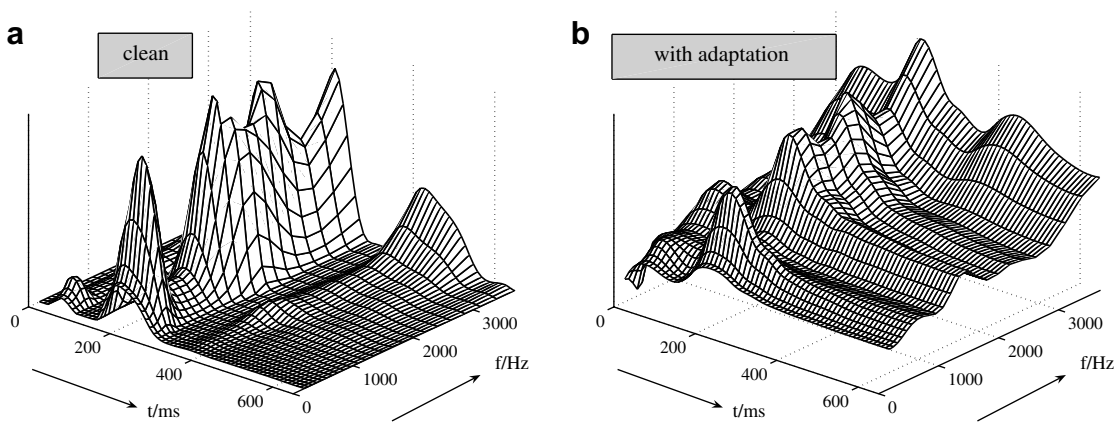

Fig. 12. Spectrograms of the clean and the adapted HMMs for the word "six".

$$\widehat{X}_k(S_i, \mathrm{mix}_j) = W_k \cdot \tilde{X}_k(S_i, \mathrm{mix}_j) + N_k \quad \text{for } 1 \leqslant k \leqslant \mathrm{NR_mel}. \tag{37}$$

The adapted Mel spectra are transformed to the cepstral domain again.

In the same way the energy parameter that has been adapted to the reverberation as stated in Eq. (10), is used as input for the further adaptation

$$\widehat{E}_k(S_i, \mathrm{mix}_j) = we \cdot \tilde{E}(S_i, \mathrm{mix}_j) + Enoise. \tag{38}$$

The adaptation to reverberation and noise is visualized by the spectrograms in Fig. 12. The spectrogram is shown in graph (a) as it can be calculated from the HMM of the word six trained on clean data. In graph (b) the adapted version of this HMM is visualized. The adapted HMM has been extracted during the recognition of artificially distorted TIDigits data. These data have been created from a simulation of the hands-free recording in a noisy living room environment. The noise spectrum as it is estimated for the individual input utterance, becomes visible as shift of the complete spectrogram. The reverberation tails can also be seen when looking at the contours along time in individual subbands.

5. Recognition experiments on hands-free speech input in noisy environments

Again the recognition of connected digits is employed to proof the applicability of the combined adaptation to several distortion effects. A data base called Aurora-2 (Hirsch and Pearce, 2000) exists that consists of noisy versions of the TIDigits. Noise signals have been artificially added at different SNRs. Furthermore a few test sets exist where the frequency characteristics have been modified to simulate the recording with audio devices in the telecommunication area. But Aurora-2 does not cover the effect of a hands-free speech input in noisy environments.

Thus new sets of distorted versions have been artificially created from the clean TIDigits by applying the already mentioned simulation tool (Finster, 2005). To make these data available for the research community, they have been put together as data base and have been combined with HTK based recognition experiments. They will become available under the title Aurora-5 (Aurora, 2006).

A few details about the new data base will be listed in the next section before presenting the recognition results for these data. Finally results will be shown for the recog-

nition of data that have been recorded in a reverberant meeting room in hands-free mode.

5.1. Distorted data of the Aurora-5 experiment

The new data base focuses on two application scenarios for speech recognition. The first one is the application inside the noisy interior of a car, the second one the hands-free speech input in an office or a living room, to control e.g. electronic devices like a telephone or audio/video equipment. In comparison to Aurora-2 where only 1000 utterances were selected, each test set contains all 8700 utterances here.

The usage of telephone like devices is assumed in general for the car scenario by filtering all data with a G.712 frequency characteristic first (Campos-Neto, 1999). G.712 is a characteristic that attenuates all frequency components outside the range from about 300 to 3400 Hz. Car noise is added to the filtered data at different SNRs in the range from 0 to 15 dB. The noise segment for distorting a single utterance is randomly selected out of eight recordings that were made in different cars and under different conditions like e.g. windows open or closed. Three different versions exist for the car noise condition. The first version contains additive noise only according to the recording with a close talking microphone. The second one considers the recording with a hands-free microphone. The third version is like the second one but containing a further transmission over a GSM cellular telephone network. This reflects the usage of an information retrieval system located at a remote position in the telephone network. The GSM transmission is simulated by randomly selecting an AMR (adaptive multi-rate) speech coding mode and the channel conditions of the cellular channel. These options exist as part of the simulation tool. In total 15 test sets exist for the five different SNR conditions including the clean case and the three versions.

For the second scenario randomly selected noise segments are added from five recordings inside different rooms like e.g. an office room or a restaurant. The same range of SNRs is considered. Three different versions exist. The first one looks at additive noise only simulating the recording with a close talking microphone. The second one considers the recording in an office room where the reverberation time is randomly varied in the range from 0.3 to 0.4 s. In the third version the recording in a living room is simulated where the reverberation time randomly varies in the range from 0.5 to 0.6 s. This comes up again to 15 test sets in total.

In general the Aurora-5 data contain a bigger variance of the distortion conditions inside each test set in comparison to Aurora-2. For example only a single noise recording has been taken for Aurora-2 to create one test set.

5.2. Recognition of artificially distorted digits

Cepstral parameters are extracted again as acoustic features, as described and applied before for the experiments with reverberation as the single distortion effect. Also the same gender dependent HMMs are used that have been created with a training on the clean TIDigits. The word error rates are presented in Fig. 13 for the three different versions containing car noise.

Looking at the condition with additive noise only, shown in graph (a), the expected improvement can be seen when comparing the results for the robust ETSI front-end against the results for a cepstral analysis. Further small improvements are achieved when adapting the HMMs to all distortion effects as described in the previous chapter. Furthermore the error rates are shown for the unsupervised HMM adaptation with MLLR as it is available as part of the HTK Viterbi recognizer. An incremental MLLR is performed after each utterance. We observed a worse recognition performance when applying MLLR every two or more utterances. The adaptation is performed on the HMMs containing the features of the cepstral analysis so that the results can be immediately compared to the new adaptation approach. The error rates for MLLR are only a little bit worse in this case.

The improvement, comparing the new adaptation approach against the ETSI front-end, is higher when looking at the condition of a hands-free speech input in the noisy car environment. This is shown in the graph (b). The reverberation time is fairly small in a car in comparison to rooms. The major impact of the hands-free recording inside a car is a modification of the frequency characteristics. The adaptation seems to compensate these effects to a higher extent than the robust feature extraction except for the low SNR of 0 dB. The error rates for MLLR are again slightly worse.

The adaptation scheme shows its usability also for the case of an additional transmission over the GSM cellular network as shown by the results in graph (c). In this case the speech is further modified by the encoding and decoding and the transmission errors on the cellular channel. The adaptation technique seems to cover this type of distortion considerably better than the robust ETSI front-end. The performance of MLLR is extremely low for the SNR of 0 dB. This has been observed in several experiments where the performance without adaptation was already quite low. MLLR seems to be unable to find the right feature mapping in such cases and seems to adapt the features in the wrong direction.

The cases with car noise do not include the major effects of a hands-free speech input in a reverberant room environment. The word error rates presented in Fig. 14 do include such effects. These experiments investigate recordings of speech inside a noisy room environment.

In the case of additive noise only, shown in graph (a), the new adaptation scheme leads to similar error rates like the robust front-end. In general the recognition performance is lower in comparison to the case with car noise because the interior noise signals contain more non stationary segments.

A considerable improvement is observed when comparing the new adaptation technique against the robust

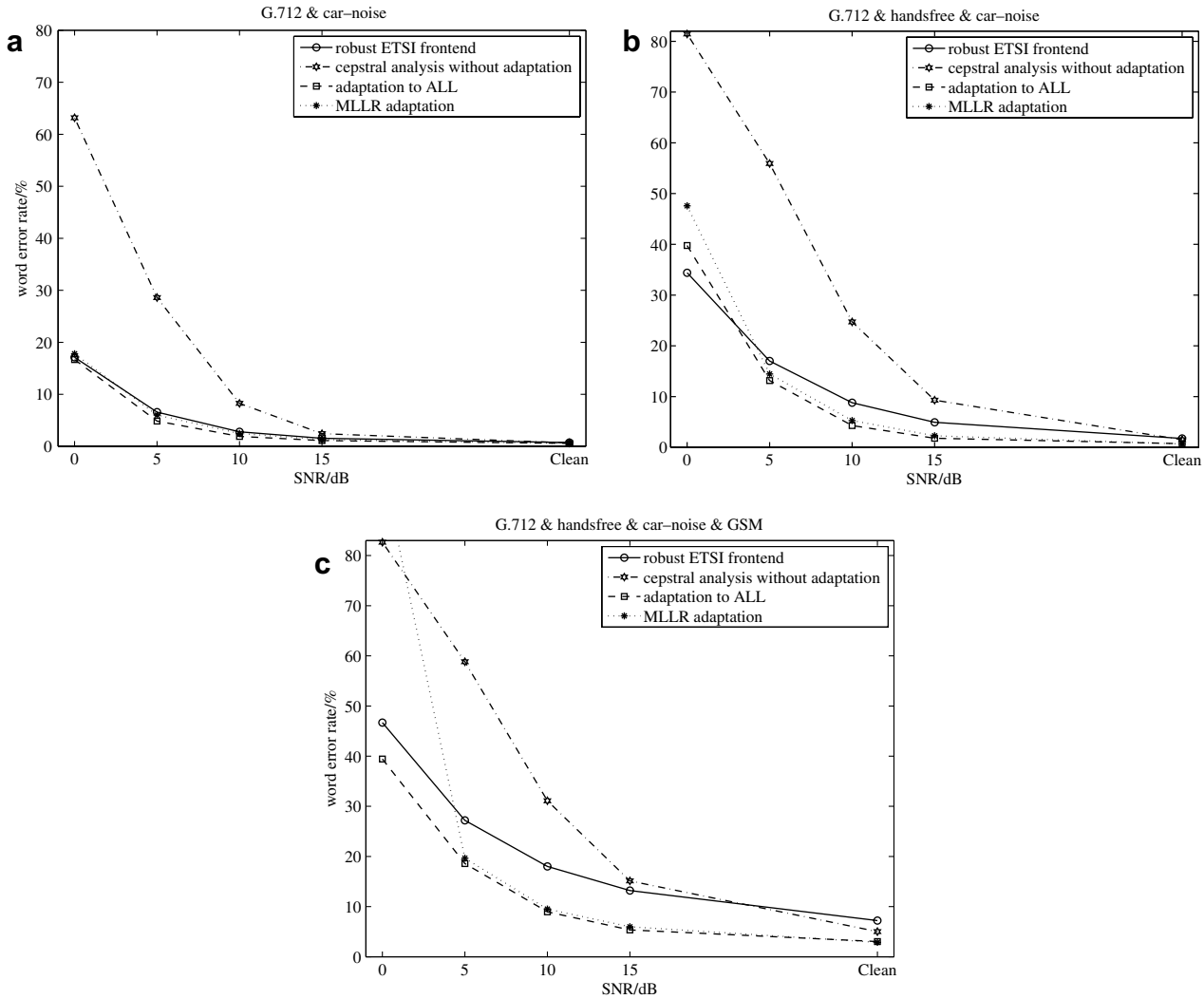

Fig. 13. Word error rates for different recording conditions inside a car.

front-end for the cases of a hands-free speech input in an office or a living room as shown in graphs (b) and (c). The additional adaptation to reverberation causes this improve- ment.

MLLR adaptation leads to worse results especially for SNRs below 15 dB. It looks like the mapping on the basis of a linear regression is not able to completely compensate the sum of spectral modifications caused by background noise and reverberation. While additive noise and a frequency weighting can be modeled as a stationary modification of each frame, reverberation includes modifications along the time axis. In sum this can not be completely compensated with a linear mapping. As already observed for the car noise conditions, MLLR seems to adapt into the wrong direction in case we obtain a low performance without adaptation.

5.3. Recognition of digits recorded in application scenarios

All results presented so far have been derived from a recognition of artificially distorted speech data. The authors believe that their simulation of recording conditions represents the situation of applying a recognition system in a real scenario quite well.

Thus the improvements on artificially distorted data should also be visible in real application scenarios. This is proofed by recognizing speech data that have been recorded in different situations.

The first experiment is run on the so called Bellcore digits. These are speech data that have been recorded over telephone lines. There exist the recordings of 220 American speakers that have spoken the 10 English digits as isolated words. The recordings contain the usual distortions that occur in case of transmitting speech over the telephone. This includes the usage of telephone devices with different frequency characteristics and the presence of some background noise. Word error rates are shown in Fig. 15 for an isolated word recognition using the set of HMMs that have been trained on the clean TIDigits.

For the cepstral analysis and the HMMs trained on clean TIDigits the recognition performance is low with an error rate of about 30%. The error rate can considerably

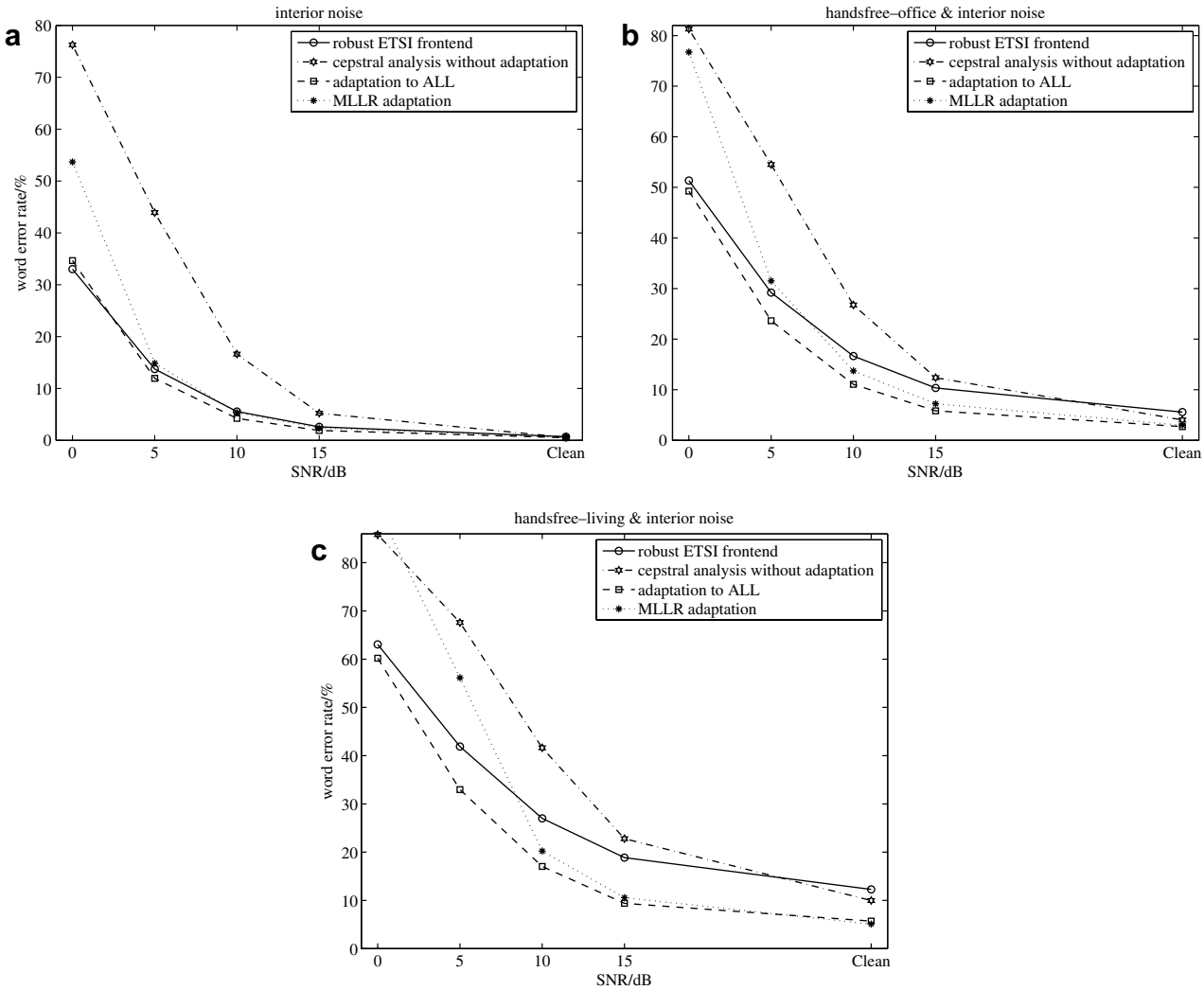

Fig. 14. Word error rates for different recording conditions inside rooms.

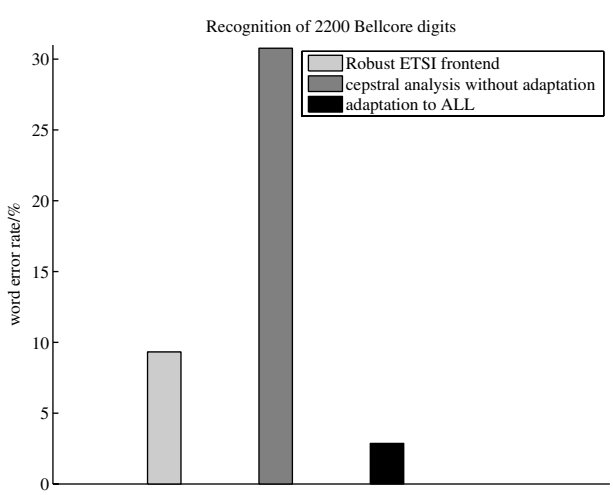

Fig. 15. Word error rates for the Bellcore single digits with HMMs trained on the TIDigits.

be reduced with the robust ETSI front-end to a rate less than 10%. Applying the adaptation technique leads to a further relative error rate reduction by about 70% in comparison to the error rate for the robust front-end.

MLLR adaptation is applied on the HMMs for the case of cepstral analysis, where these results are not shown in the figure. We obtain the highest performance with an error rate of 12% when performing the adaptation every five utterances. An interesting result is achieved when applying the MLLR adaptation on the HMMs trained on the features of the ETSI front-end. We obtain an error rate of 2.3% when performing the adaptation for each utterance. This is even slightly better than the new adaptation technique. It might indicate that it is possible to combine a robust feature extraction with an additional adaptation.

Another experiment is run on some recordings of the meeting recorder project (Janin et al., 2003). Speech data have been recorded during meetings in a meeting room where the microphones were placed in the middle of a table. Thus these data contain reverberation besides a low amount of background noise. The speakers uttered also sequences of English digits which are used for this

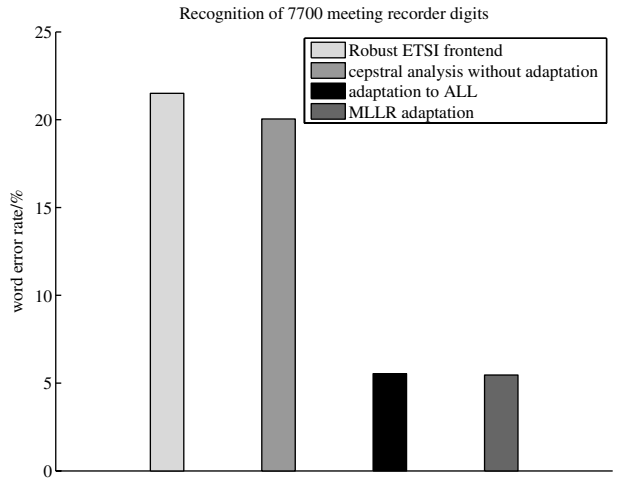

Fig. 16. Word error rates for the meeting recorder digits with HMMs trained on TIDigits.

experiment. Only the recordings of native English speakers are used resulting in about 2350 utterances with about 7700 digits in total. Word error rates are shown in Fig. 16.

In this case the performance of the robust front-end is slightly worse in comparison to a standard cepstral analysis. This effect has already been observed in most of the experiments on reverberant data presented in the preceding sections. Applying the adaptation scheme, the error rate can be reduced by about 70% in relation to the cepstral analysis without adaptation. We achieve almost the same performance when performing MLLR adaptation for each utterance.

6. Conclusion

We present a new technique for adapting the acoustic parameters of HMMs to the condition of hands-free speech input in reverberant rooms. This approach can be combined with existing techniques for the adaptation on noise and unknown frequency characteristics. Furthermore we introduce a new method for adapting the Delta parameters based on a preceding adaptation of the static parameters.

Applying the new adaptation approach on artificially distorted data or on real recordings in noisy conditions, we achieve a considerably higher recognition performance in comparison to the case without adaptation. The error rates are also lower than the ones that are achieved with the robust ETSI front-end that can be considered as representative for a robust feature extraction.

Especially in conditions where additive noise and reverberation distort the speech signal, we obtain a higher recognition performance with the new adaptation technique in comparison to MLLR adaptation. The linear regression seems to compensate the nonlinear distortion effects worse than the new approach.

In the future we will investigate whether and how robust feature extraction schemes can be combined with HMM adaptation techniques.

Acknowledgments

The investigations have been partly carried out during a research stay at the International Computer Science Institute in Berkeley, CA 94704, USA. H.G. Hirsch would like to thank Nelson Morgan and the whole speech group for fruitful discussions and their support.

This work is supported by the German research funding organization DFG (Deutsche Forschungsgemeinschaft).

References

Aurora project, 2006. <http://aurora.hsnr.de>.

Au Yeung, S.-K., Siu, M.-H., 2004. Improved performance of Aurora-4 using HTK and unsupervised MLLR adaptation. In: Proc. ICSLP.

Avendano, C., Hermansky, H., 1996. Study on the dereverberation of speech based on temporal filtering. In: Proc. ICSLP, pp. 889–892.

Bitzer, J., Simmer, K.U., Kammeyer, K.D., 1999. Multi microphone noise reduction techniques for hands-free speech recognition – a comparative study. In: Proc. Internat. Workshop on Robust Methods for Speech Recognition in Adverse Conditions, Tampere, Finland, pp. 171–175.

Campos-Neto, S., 1999. The ITU-T software library. Internat. J. Speech Technol., 259–272.

Couvreur, L., Dupont, S., Ris, C., Boite, J.M., Couvreur, C., 2001. Fast adaptation for robust speech recognition in reverberant environments. In: Proc. Internat. Workshop on Adaptation Methods for Speech Recognition, Sophia Antipolis, France.

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

Finster, H., 2005. Web interface to experience the simulation of acoustic scenarios. <http://dnt.kr.hsnr.de/sireac.html>.

Gadrudadri, H., Hermansky, H., Morgan, N., et al., 2002. Qualcomm-ICSI-OGI features for ASR. In: Proc. ICSLP, pp. 21–24.

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

Gales, M.J.F., 1997. Nice model-based compensation schemes for robust speech recognition. In: Proc. ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels, Pont-a-Mousson, France, pp. 55–64.

Gales, M.J.F., Young, S.J., 1996. Robust continuous speech recognition using parallel model combination. IEEE Trans. Speech Audio Proc. 4, 352–359.

Gauvain, J.L., Lee, C.H., 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Trans. Speech Audio Proc. 2, 291–298.

Gelbart, D., Morgan, N., 2002. Double the trouble: handling noise and reverberation in far-field automatic speech recognition. In: Proc. ICSLP, pp. 2185–2188.

Hirsch, H.G., 1999. HMM adaptation for telephone applications. In: Proc. European Conf. on Speech Communication and Technology, Vol. 1, pp. 9–12.

Hirsch, H.G., 2001a. HMM adaptation for applications in telecommunication. Speech Comm. 34, 127–139.

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

Hirsch, H.G., Finster, H., 2005. The simulation of realistic acoustic input scenarios for speech recognition systems. In: Proc. Interspeech Conf., pp. 2697–2700.

Hirsch, H.G., Pearce, D., 2000. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: Proc. ISCA Workshop ASR2000, Paris, France.

Hirsch, H.G., Hellwig, K., Dobler, S., 2001b. Speech recognition at multiple sampling rates. In: Proc. European Conf. on Speech Communication and Technology, pp. 1837–1840.

Houtgast, T., Steeneken, H.J.M., Plomp, R., 1980. Predicting speech intelligibility in rooms from the modulation transfer function, I. General room acoustics. Acustica 46, 60–72.

Janin, A. et al., 2003. The ICSI meeting corpus. In: Proc. ICASSP.

Kingsbury, B., 1998. Perceptually inspired signal processing strategies for robust speech recognition in reverberant environments. Dissertation at UC Berkeley, USA.

Kinshita, K., Nakatani, T., Miyoshi, M., 2005. Efficient blind dereverberation framework for automatic speech recognition. In: Proc. Interspeech Conf., Lisbon, Portugal, pp. 3145–3148.

Kuttruff, H., 2000. Room Acoustics. Spon Press.

LDC, 1993. Speech Data Base CSR-I (WSJ0). Wall Street Journal, http://www.ldc.upenn.edu.

Leggeter, C.J., Woodland, P.C., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density Hidden Markov Models. Comput. Speech Lang. 9, 171–185.

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

Liu, J., Malvar, H.S., 2001. Blind deconvolution of reverberated signals. In: Proc. ICASSP, Vol. 5, pp. 3037–3040.

Macho, D., Mauuary, L., Pearce, D., et al., 2002. Evaluation of a noise robust DSR front-end on Aurora databases. In: Proc. ICSLP, pp. 17–20.

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

Omologo, M., Svaizer, P., Matassoni, M., 1998. Environmental conditions and acoustic transduction in hands-free speech recognition. Speech Comm. 25, 75–95.

Palomäki, K.J., Brown, G.J., Barker, J., 2002. Missing data speech recognition in reverberant conditions. In: Proc. ICASSP, pp. 65–68.

Picone, J., Parihar, N., Hirsch, H.G., Pearce, D., 2004. Performance analysis of the Aurora large vocabulary experiment. In: Proc. European Signal Processing Conference, Vienna, Austria.

Raut, C.K., Nishimoto, T., Sagayama, S., 2005. Model adaptation by state splitting of HMM for long reverberation. In: Proc. Interspeech Conf., Lisbon, Portugal, pp. 277–280.

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

Seltzer, M.L., Raj, B., Stern, R.M., 2004. Likelihood-maximizing beamforming for robust hands-free speech recognition. IEEE Trans. Speech Audio Proc. 12 (5), 489–498.

Tashev, I., Allred, D., 2005. Reverberation reduction for improved speech recognition. In: Proc. Workshop on Hands-free Speech Communication, Rutgers, USA.

Woodland, P.C., 2001. Speaker adaptation for continuous density HMMs: a review. In: Proc. Internat. Workshop on Adaptation Methods for Speech Recognition, Sophia Antipolis, France.

Wu, M., Wang, D., 2005. A two-stage algorithm for enhancement of reverberant speech. In: Proc. ICASSP, Vol. I, pp. 1085–1088.

Yegnanarayana, B., Murthy, P.S., 2000. Enhancement of reverberant speech using LP residual signals. IEEE Trans. Speech Audio Proc. 8, 267–281.

Young, S. et al., 2005. The HTK Book (version 3.3). Cambridge University Engineering Department, http://htk.eng.cam.ac.uk.