Frame-synchronous noise compensation for hands-free speech recognition in car environments

J.-T.Chien and M.-S.Lin

Abstract: It has become increasingly important to develop hands-free speech recognition techniques for the human–computer interface in car environments. However, severe car noise degrades the speech recognition performance substantially. To compensate the performance loss, it is necessary to adapt the original speech hidden Markov models (HMMs) to meet changing car environments. A novel frame-synchronous adaptation mechanism for in-car speech recognition is presented. This mechanism is intended to perform unsupervised model adaptation efficiently on a frame-by-frame basis instead of a conventional adaptation algorithm relying on batch adaptation data and supervision information. The proposed adaptation scheme is performed during frame likelihood calculation where an optimal equalisation factor is first computed to equalise the model mean vector and the input frame vector. This equalisation factor then serves as a reference index to retrieve an additional bias vector for model mean adaptation. As a result, a rapid and flexible algorithm is exploited to establish a new robust likelihood measure. In experiments on hands-free in-car speech recognition with the microphone far from the talker, this framework is found to be effective in terms of recognition rate and computational cost under various driving speeds.

1 Introduction

There is no doubt that the robustness issue is crucial in pattern recognition because a mismatch between training and testing data always exists and degrades the recognition performance considerably in real-world applications. For applications of speech recognition, the distortion sources come from inter- and intra-speaker variabilities, transducers/channels and surrounding noises. For instance, when the speech recogniser is designed for hands-free control of the car equipment including cellular telephones, air conditioning systems, Global Positioning Systems etc., the noise from the engine, music, babble, wind, echo etc. under different driving speeds will degrade the performance of the recogniser [10, 20]. Also, changes of speaker voice caused by abrupt alterations of car noise level (known as the Lombard effect) will damage the recogniser [21]. However, it is impractical to collect numerous training data from various noise conditions to generate speech models covering a wide range of environmental statistics. A feasible approach is to build an adaptive speech recogniser where the speech models can be adapted to new environments using environment-specific adaptation data.

The issue of adaptive speech recognition has been attracting many researchers [18, 22]. In the literature, the maximum *a posteriori* (MAP) adaptation of speech hidden Markov models (HMMs) [17] and maximum-likelihood linear regression (MLLR) [23] provided two main approaches to model adaptation. The MAP adaptation adjusted the HMM parameters directly using MAP estimation. This scheme adapted those HMM units with at least one sample appearing in adaptation data. Good asymptotic properties can be achieved for sufficient adaptation data. In addition, the MLLR indirectly adapted the HMM parameters through cluster-dependent transformation functions in which the parameters were obtained via maximum-likelihood (ML) estimation. This transformation-based adaptation is effective even using sparse adaptation data. In general, these two approaches are feasible for batch adaptation in a supervised manner. If the supervision of adaptation data is unknown, these methods have to estimate the supervision of adaptation data through one pass of the recognition process. Unsupervised adaptation is then performed according to the estimated supervision. The quality of estimated supervision can be assessed further to improve the unsupervised adaptation [1, 7, 19]. In [11], an unsupervised frame-synchronous variant of MLLR was presented. Moreover, the realistic environments are nonstationary due to the evolving nature of environmental statistics. It is difficult to capture the changing statistics using the batch adaptation data. The incremental adaptation technique is accordingly important for practical speech recognition systems [5].

To avoid waiting for long batch data and preparing data supervision, this paper presents a frame-synchronous unsupervised adaptation approach for robust car speech recognition. The approach is to perform the adaptation during the likelihood calculation. Specifically, when the observation likelihood is determined, an optimal equalisation factor is initially computed to equalise the HMM mean vector towards the speech frame. The bias of the equalisa-

IEE Proceedings online no. 20000693

DOI: 10.1049/ip-vis:20000693

Paper first received 21st February and in revised form 3rd July 2000

The authors are with the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan 70101

tion is subsequently compensated by a reference vector, which is extracted from the pre-stored reference function using the optimal equalisation factor as a relation index. Therefore, a novel adaptation scheme, called 'adaptation by reference', is developed. Herein, the reference function is trained through a limited set of speech from car noise environments without the need for data supervision. The relation indices between optimal equalisation factors and adaptation biases are correspondingly constructed using the pairs of training frames and HMM mean vectors. These indices are stored for table lookup in a testing session. In the experiments, the proposed noise compensation approach is found to be superior in terms of recognition rate and cost under different car driving conditions: standby, downtown and freeway.

2 Frame-synchronous noise compensation

2.1 Background of adaptive speech recognition

In this context of statistical recognition, the optimal word sequence $\hat{W}$ of an input utterance $\boldsymbol{X}=\{\boldsymbol{x}_t\}$ is determined according to the Bayes rule

$$\hat{W} = \arg\max_W p(W|\boldsymbol{X}) = \arg\max_W p(\boldsymbol{X}|W)p(W) \quad (1)$$

where $p(\boldsymbol{X}|W)$ is the accumulated likelihood of utterance $\boldsymbol{X}$ and $p(W)$ is the prior knowledge of word sequence, i.e. language model. Using the framework of continuous-density HMMs [26], the missing data of state sequence $S=\{s_t\}$ is incorporated into the calculation of accumulated likelihood written by

$$p(\boldsymbol{X}|W) = \sum_{\text{all } S} p(\boldsymbol{X}, S|W) = \sum_{\text{all } S} p(\boldsymbol{X}|S, W)p(S|W) \quad (2)$$

In general, the likelihood computation of eqn. 2 is very expensive and almost unattainable. One efficient approach is to apply the Viterbi algorithm [27] and decode the optimal state sequence $\hat{S}=\{\hat{s}_t\}$. The summation over all possible state sequences in eqn. 2 is accordingly approximated by the single most likely state sequence, i.e.

$$p(\boldsymbol{X}|W) \cong p(\boldsymbol{X}|\hat{S}, W)p(\hat{S}|W) = \pi_{\hat{s}_0}\prod_{t=1}^{T} a_{\hat{s}_{t-1}\hat{s}_t} b_{\hat{s}_t}(\boldsymbol{x}_t) \quad (3)$$

where $\pi_{\hat{s}_0}$ is the initial state probability, $a_{\hat{s}_{t-1}\hat{s}_t}$ is the state transition probability and $b_{\hat{s}_t}(\boldsymbol{x}_t)$ is the observation probability density function of $\boldsymbol{x}_t$ in state $\hat{s}_t$ which is modelled by a mixture of multivariate Gaussian densities

$$b_{\hat{s}_t}(\boldsymbol{x}_t) = p(\boldsymbol{x}_t|\hat{s}_t = i, W) = \sum_{k=1}^{K} \omega_{ik} f(\boldsymbol{x}_t|\theta_{ik}) = \sum_{k=1}^{K} \omega_{ik} N(\boldsymbol{x}_t|\mu_{ik}, \Sigma_{ik}) \quad (4)$$

Herein, ω_{ik} is the mixture weight, μ_{ik} and Σ_{ik} are, respectively, the $D\times 1$ mean vector and $D\times D$ covariance matrix of state $\hat{s}_t=i$ and mixture component k. The likelihood measure $f(\boldsymbol{x}_t|\theta_{ik})$ of frame $\boldsymbol{x}_t$ associated with HMM unit $\theta_{ik}=(\mu_{ik}, \Sigma_{ik})$ is expressed by

$$f(\boldsymbol{x}_t|\theta_{ik}) = (2\pi)^{-D/2}|\Sigma_{ik}|^{-1/2} \times \exp\left[-\frac{1}{2}(\boldsymbol{x}_t-\mu_{ik})'\Sigma_{ik}^{-1}(\boldsymbol{x}_t-\mu_{ik})\right] \quad (5)$$

If the acoustics of trained models θ_{ik} and test frame $\boldsymbol{x}_t$ come from the same acoustic environments, the likelihood measure defined in eqn. 5 is appropriate for estimating optimal word sequence $\hat{W}$ in eqn. 1. However, realistic car environments are adverse such that we cannot predict exactly the specific surrounding noise of car type and driving condition from the training material. Hence, it becomes necessary either to enhance the test frame $\boldsymbol{x}_t$ or to adapt the HMM parameters θ_{ik} to achieve robustness of speech recognition. In [3, 14], the techniques of spectral subtraction and Bayesian signal estimation were useful for speech enhancement. Using these techniques, the speech signal was enhanced to be acoustically near the trained speech models. On the other hand, the parallel model combination (PMC), which optimally combined the HMMs of speech and noise, was successfully employed for noisy speech recognition [16]. The adaptation of HMM parameters using an affine transformation is also popular for speaker adaptation [13, 23] and noise adaptation [28]. The resulting likelihood measure is given by

$$f(\boldsymbol{x}_t|\boldsymbol{A}_c, \boldsymbol{b}_c, \theta_{ik}) = (2\pi)^{-D/2}|\Sigma_{ik}|^{-1/2} \times \exp\left[-\frac{1}{2}(\boldsymbol{x}_t-\boldsymbol{A}_c\mu_{ik}-\boldsymbol{b}_c)'\Sigma_{ik}^{-1}(\boldsymbol{x}_t-\boldsymbol{A}_c\mu_{ik}-\boldsymbol{b}_c)\right] \quad (6)$$

where $\boldsymbol{A}_c$ and $\boldsymbol{b}_c$ are, respectively, a $D\times D$ scaling matrix and a $D\times 1$ bias vector of the cth HMM cluster. The cluster-dependent transformation parameters $(\boldsymbol{A}_c, \boldsymbol{b}_c)$ could be obtained either by ML estimation or by MAP estimation [8] using a period of adaptation data.

2.2 Optimal equalisation factor

Mansour and Juang [24] observed that the additive white noise would cause norm shrinkage of a speech cepstral vector. They consequently designed a distance measure where a scaling factor was introduced to compensate the cepstral shrinkage for cepstrum-based speech recognition. This approach was further extended to the adaptation of HMM parameters by detecting an equalisation scalar λ between the HMM unit θ_{ik} and noisy speech frame $\boldsymbol{x}_t$ [4]. The likelihood measure in eqn. 5 is modified to

$$f(\boldsymbol{x}_t|\lambda, \theta_{ik}) = (2\pi)^{-D/2}|\Sigma_{ik}|^{-1/2} \times \exp\left[-\frac{1}{2}(\boldsymbol{x}_t-\lambda\mu_{ik})'\Sigma_{ik}^{-1}(\boldsymbol{x}_t-\lambda\mu_{ik})\right] \quad (7)$$

One can determine the optimal equalisation factor λ_e by directly maximising the logarithm of the modified likelihood measure as follows:

$$\lambda_e = \lambda_e(\boldsymbol{x}_t, \theta_{ik}) = \arg\max_{\lambda} \log f(\boldsymbol{x}_t|\lambda, \theta_{ik}) = \frac{\boldsymbol{x}_t'\Sigma_{ik}^{-1}\mu_{ik}}{\mu_{ik}'\Sigma_{ik}^{-1}\mu_{ik}} \quad (8)$$

Geometrically, this factor is equivalent to the projection of $\boldsymbol{x}_t$ upon μ_{ik} weighted by Σ_{ik}^{-1}. The projection-based likelihood measure is subsequently obtained by substituting λ_e into eqn. 7, i.e. $f(\boldsymbol{x}_t|\lambda_e, \theta_{ik})$. The corresponding distance measure was referred to as the weighted projection measure (WPM) [4]. The projection-based likelihood measure was also expanded by additionally considering the adaptation of the covariance matrix and the variance adapted likelihood measure was generated [9]. However, the noise in car environments is non-white and is complex to characterise. It is difficult to adapt the HMM mean vector μ_{ik} properly by only applying the optimal equalisation scalar λ_e. Thus, we are stimulated to perform extra adaptation to compensate the adaptation bias of the HMM mean vector induced by λ_e. It is not possible to estimate the adaptation bias in a statistical sense using one speech frame. The following approach is accordingly exploited.

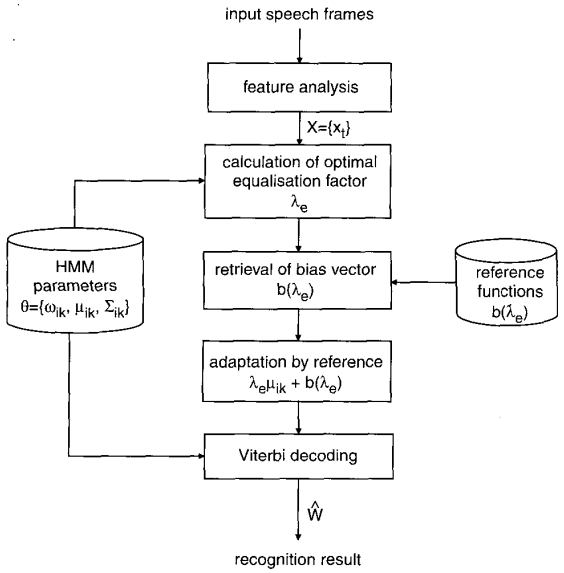

Fig. 1 *Flowchart of speech recognition system based on ABR method*

2.3 Adaptation by reference (ABR)

As indicated in eqn. 8, the optimal equalisation factor λ_e relates to the observation vector $\boldsymbol{x}_t$ and HMM unit $\theta_{ik} = (\mu_{ik}, \Sigma_{ik})$. This factor embeds the information of noise type, noise level and the relation to HMM parameters. In this study, this factor is regarded as a 'relational reference index' to perform the 'adaptation by reference' (ABR). A novel likelihood measure correspondingly results from merging an extra bias vector $\boldsymbol{b}(\lambda_e)$ into the likelihood measure. It turns out to be

$$f(\boldsymbol{x}_t|\lambda_e, \boldsymbol{b}(\lambda_e), \theta_{ik}) = (2\pi)^{-D/2}|\Sigma_{ik}|^{-1/2} \times \exp\left\{-\frac{1}{2}[\boldsymbol{x}_t - \lambda_e\mu_{ik} - \boldsymbol{b}(\lambda_e)]'\Sigma_{ik}^{-1}[\boldsymbol{x}_t - \lambda_e\mu_{ik} - \boldsymbol{b}(\lambda_e)]\right\} \quad (9)$$

Note that the bias vector $\boldsymbol{b}(\lambda_e)$ is shared by overall HMM units $\theta = \{\theta_{ik}\}$ and retrieved from the pre-trained reference function according to the index value λ_e. The reference function is estimated through a small set of in-car speech material. The estimation procedure is described in Section 3.2. Fig. 1 shows the flow chart of the recognition system based on the proposed ABR method. It can be seen that the optimal equalisation factor λ_e is calculated and used to extract the bias vector $\boldsymbol{b}(\lambda_e)$. Applying the new likelihood measure of eqn. 9 to the Viterbi decoding algorithm, the optimal word sequence $\hat{W}$ associated with the input speech data $\boldsymbol{X}$ can be found. In particular, the optimal equalisation factor λ_e is computed for each frame $\boldsymbol{x}_t$ and HMM unit θ_{ik} in the WPM and ABR methods.

3 Estimation of reference function

In this Section, the estimation procedures for the reference function are discussed. First, the in-car speech database used in the experiments is described.

3.1 Car speech database

The car speech database (CARNAV98) was collected in a joint project of the National Cheng Kung University and the Industrial Technology Research Institute, Taiwan, under contract no. G3-88037. This database contained 14 control commands and 100 Chinese names uttered by ten speakers (five males and five females) and recorded in a medium-sized car [Toyota Corolla 1.8 (car 1)] and a smaller car [Yulon Sentra 1.6 (car 2)]. Three sets of driving data for 'standby' condition, 'downtown' condition and 'freeway' condition with car speeds, respectively, of 0 km/h, 50 km/h and 90 km/h, were recorded via a high-quality MD Walkman of type MZ-R55 and using a hands-free Sony ECM-717 microphone far from the talker. During recording, the engine was kept on, air-conditioner was on, music was off and the windows were closed. Speech was digitised at 16-bit accuracy and 8 kHz sampling rate. For a speech frame, Hamming windowing was applied and a feature vector of 12-order LPC-derived cepstral coefficients (denoted by LPCC), 12-order delta cepstral coefficients, one delta log energy and one delta delta log energy was computed. In this database, car 1 had 122, 158 and 200 utterances of two males and two females and car 2 had 204, 263 and 324 utterances of three males and three females for driving conditions of standby, downtown and freeway, respectively. In total, 1271 utterances were gathered. To evaluate the outside performance of the proposed method, the speech data of car 1 were adopted for estimation of reference function and those of car 2 were adopted for recognition experiments. In this case, the environmental factors of speaker and car type are entirely different for reference function estimation and speech recognition. The segmental signal-to-noise ratio (SNR) values of various cars and driving conditions are calculated and listed in Table 1. It can be seen that car classes and driving speeds significantly change the noise levels. The lowest SNR of −10.14 occurs for the case of car 2 driven under freeway conditions.

3.2 Estimation procedures

Generally, the reference function is intended to reveal the adaptation bias behaviour of the projection-based likelihood measure induced by optimal equalisation factor λ_e. This function can be non-parametrically extracted from the training data of car 1. In a previous study [9], a stereo database was prepared by artificially adding the noise signal to clean speech to generate the pairs of clean frames and corresponding noisy frames. A reference function was trained according to the adaptation biases of these data pairs. However, it is hard to obtain the simulated stereo data in real-world environments. In particular, the stereo data in the presence of speaker variabilities and car noises are almost unrealisable. To relax the requirement for stereo data, the following estimation procedures for the reference function are presented.

First estimation procedure: Although the stereo data are not available, one can use the Viterbi decoding algorithm directly to search the state and mixture component tags associated with the car noisy frames $\boldsymbol{X} = \{\boldsymbol{x}_t\}$. To obtain better tags, the optimal equalisation factor was applied

Table 1. Comparison of segmental signal-to-noise ratios for different car types and driving conditions

	Toyota Corolla 1.8	Yulon Sentra 1.6
Standby, dB	10.31	5.63
Downtown, dB	0.34	−6.53
Freeway, dB	−3.77	−10.14

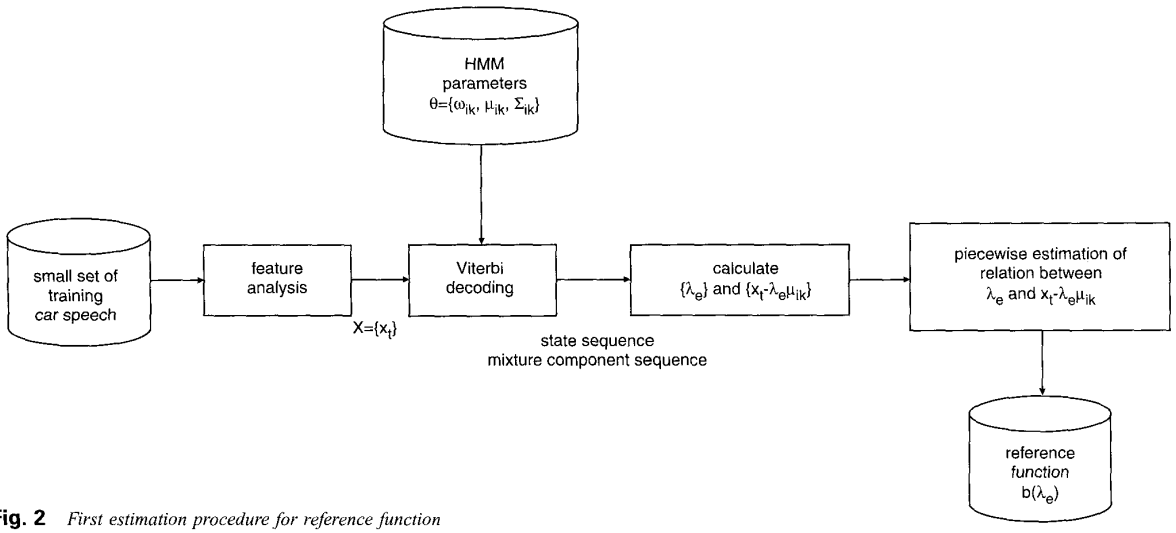

Fig. 2 *First estimation procedure for reference function*

Only pairs of observation frames $\{\boldsymbol{x}_t\}$ and corresponding HMM parameters $\{\theta_{ik}\} = \{\mu_{ik}, \Sigma_{ik}\}$ obtained by Viterbi decoding are included for estimation of reference function

during Viterbi decoding. The pairs of frames from car noise environments and associated HMM parameters $\{\theta_{ik}\} = \{\mu_{ik}, \Sigma_{ik}\}$ are accordingly produced. Given the pair data $\{\boldsymbol{x}_t, \theta_{ik}\}$, the optimal equalisation factors $\{\lambda_e\}$ of eqn. 8 and the corresponding adaptation biases $\{\boldsymbol{x}_t - \lambda_e\mu_{ik}\}$ are calculated. These pairs of $\{\lambda_e\}$ and $\{\boldsymbol{x}_t - \lambda_e\mu_{ik}\}$ are then plotted in a scatter diagram. Finally, the reference function $\boldsymbol{b}(\lambda_e)$ is piecewise estimated by averaging the scattered values $\{\boldsymbol{x}_t - \lambda_e\mu_{ik}\}$ where the step size of λ_e is specified by 0.01 [9]. Herein, the supervision of training data is known during Viterbi decoding. This estimation procedure is usually referred to as the 'piecewise constant approximation'. Fig. 2 displays the flow diagram of the first estimation procedure. It is shown that only pairs of observation frames $\{\boldsymbol{x}_t\}$ and corresponding HMM parameters $\{\theta_{ik}\} = \{\mu_{ik}, \Sigma_{ik}\}$ obtained by the Viterbi decoder are included for estimation of the reference function. Fig. 3 plots the estimated reference function of the first cepstral coefficient. This estimated function records important adaptation bias factors related to the optimal equalisation factor. Generally, there are two problems with this procedure. The first is the problem of data sparseness when only a small set of training frames is available. The other is the possibility of decoding inappropriateness of state and mixture component sequences even if supervision of training data is provided. To alleviate these problems, modifications were made to improve the quality of the estimated reference function.

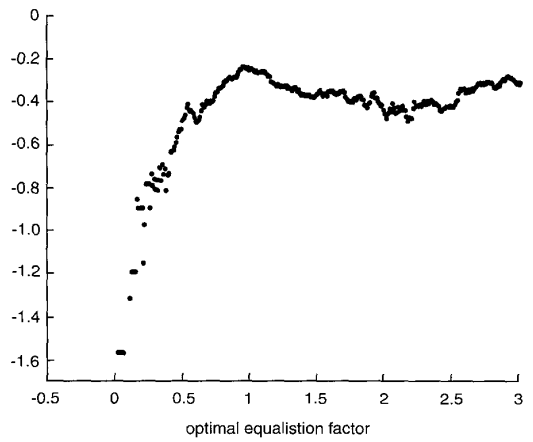

Fig. 3 *Reference function of first cepstral coefficient* $\{b_d(\lambda_e),\ d = 1\}$ *obtained by first estimation procedure*

Second estimation procedure: In the proposed ABR method, the optimal equalisation factor λ_e is calculated and $\boldsymbol{b}(\lambda_e)$ is retrieved for adaptation of the HMM mean vector. The parameters λ_e and $\boldsymbol{b}(\lambda_e)$ are varied for each likelihood measure $f(\boldsymbol{x}_t \mid \lambda_e, \boldsymbol{b}(\lambda_e), \theta_{ik})$. Basically, the optimal equalisation factor λ_e reflects the projection behaviour of $\boldsymbol{x}_t$ on $\theta_{ik} = (\mu_{ik}, \Sigma_{ik})$. It is insufficient just to couple the observation frame $\boldsymbol{x}_t$ with the corresponding HMM unit θ_{ik} obtained by the Viterbi decoder. Hence, it is better to couple each individual observation frame $\boldsymbol{x}_t$ with overall HMM units $\{\theta_{ik}\} = \{\mu_{ik}, \Sigma_{ik}\}$ for estimation of the reference function. In this case, the pairs of optimal equalisation factors $\{\lambda_e\}$ and adaptation biases $\{\boldsymbol{x}_t - \lambda_e\mu_{ik}\}$ are greatly increased so as to resolve the problem of data sparseness and to obtain a reliable reference function. Also, the execution of the Viterbi decoding algorithm and the need for training data supervision are ignored. Fig. 4 displays the flow diagram of the second estimation procedure of the reference function. The piecewise estimation technique mentioned above is still adopted herein. In Fig. 5, the histogram of optimal equalisation factor is plotted using the training data of car 1. Because car noise is coloured, it is observed that most optimal equalisation factors λ_e have values between 0 and 3, which are different from those in the presence of white noise [24]. In Figs. 6 and 7, the reference functions $\boldsymbol{b}(\lambda_e) = \{b_d(\lambda_e), d = 1, \ldots, D\}$ of the first and second cepstral coefficients, respectively, are illustrated. It can be seen that estimated curves are smooth and vary for different cepstral coefficients. The estimated reference function of the first cepstral coefficient using the first estimation procedure (Fig. 3) appears to be close to that using second estimation procedure (Fig. 6). Further, regarding the issue of memory cost, if the value λ_e of the 26-order reference function is limited between 0 and 3 with a step size of 0.01, the number of recorded floating points is 7800 (300×26). The occupied memory size is about 30 kbyte, which is small for a speech recogniser.

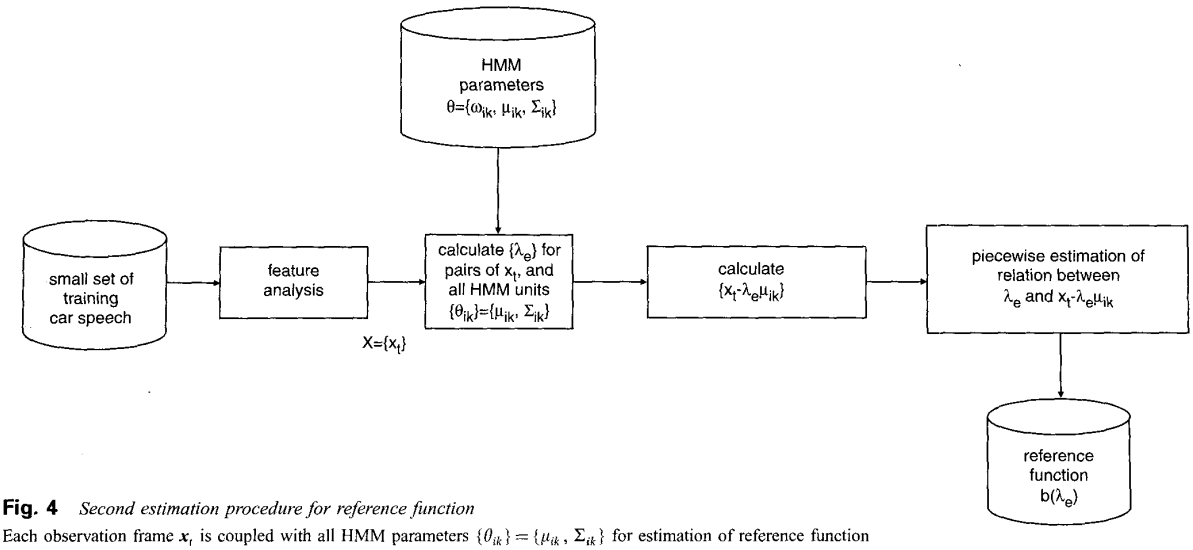

Fig. 4 *Second estimation procedure for reference function*
Each observation frame x_t is coupled with all HMM parameters $\{\theta_{ik}\} = \{\mu_{ik}, \Sigma_{ik}\}$ for estimation of reference function

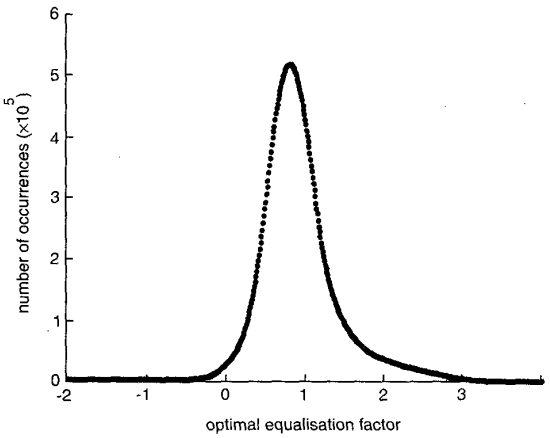

Fig. 5 *Histogram of optimal equalisation factor*

Note that the adaptation of the HMM mean vector in the ABR method has a form of affine transformation, which is similar to the MLLR adaptation in eqn. 6. The main differences between the proposed ABR method and the MLLR adaptation are threefold.

(i) The scaling factor of the model adaptation is a diagonal matrix with identical components $\lambda_e \boldsymbol{I}$ in the ABR method and a general matrix $\boldsymbol{A}_c$ in the MLLR adaptation.
(ii) In the MLLR adaptation, the transformation parameters $(\boldsymbol{A}_c, \boldsymbol{b}_c)$ are estimated using the ML principle through a set of speech frames. In contrast, the ABR method computes λ_e and looks up $\boldsymbol{b}(\lambda_e)$ for each frame likelihood calculation. The reference function $\boldsymbol{b}(\lambda_e)$ should be prepared beforehand.
(iii) If unsupervised adaptation using the MLLR is required, the cost of two-pass Viterbi decoding is needed. Conversely, ABR is a frame-synchronous unsupervised adaptation approach employed in each frame likelihood calculation.

There is also a parallel with the eigenvoice method, which can be related to MLLR [15, 25]. In the eigenvoice method, the bias was expressed as a linear combination of a series of basis vectors. The bias function $\boldsymbol{b}(\lambda_e)$, herein, can be viewed as a non-linear formulation of a mean shift vector, parameterised by λ_e for a single cluster of parameters. Moreover, in ABR, the adaptation parameters are obtained via memory association, wheareas in eigenvoice it is done using the ML principle on a per-session or utterance basis. In the present experiments, MLLR and ABR will be compared in terms of recognition time and recognition rate.

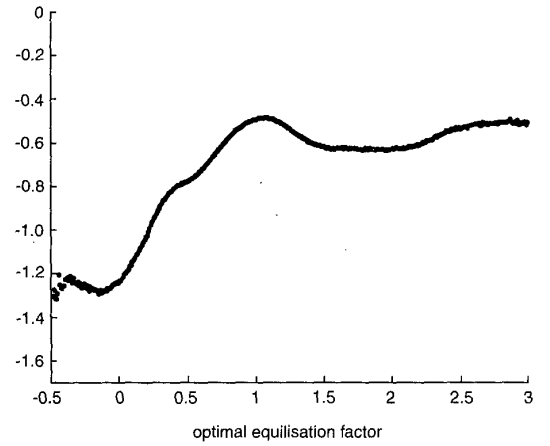

Fig. 6 *Reference function of first cepstral coefficient* $\{b_d(\lambda_e), d=1\}$ *obtained by second estimation procedure*

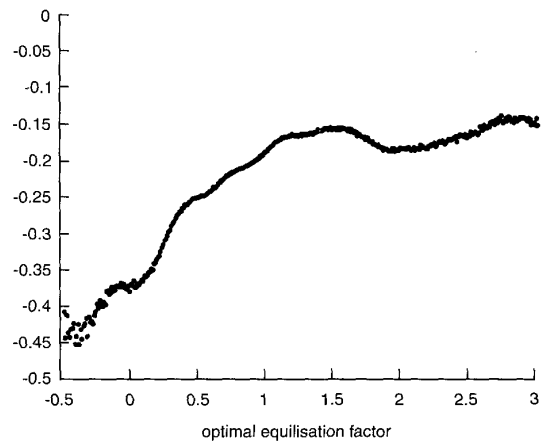

Fig. 7 *Reference function of second cepstral coefficient* $\{b_d(\lambda_e), d=2\}$ *obtained by second estimation procedure*

4 Experiments

4.1 Experimental setup and baseline system

The experiments conducted in this paper are aimed at recognition of Mandarin speech in car environments. Mandarin is a syllabic and tonal language. Without considering the tonal information, the overall number of Mandarin syllables is 408. In general, each Mandarin syllable can be divided into an initial (consonant) part and a final (vowel) part. When the syllable only has a final part, a null initial exists in practice. In this study, context-dependent subsyllable modelling was employed for constructing the HMM units of Mandarin speech [5]. Cumulatively, there were 93 context-dependent (CD) initials, 38 context-independent (CI) finals and 33 null initials generated in the experiments. CD initials, CI finals and null initials were arranged, respectively, by three, four and two left-to-right HMM states without state skipping. Hence, 498 HMM states (279 for CD initials, 152 for CI finals, 66 for null initials and 1 for background silence) were set up to cover all phonetics of 408 Mandarin syllables. Each HMM state contained four mixture components. During a training session, a speech database consisting of 5045 phonetically balanced Mandarin words uttered by 51 males and 50 females was collected. Each Mandarin word contained two to four syllables. This database was recorded in an office and via a high-quality microphone. It was applied to estimate the speaker-independent (SI) continuous-density HMM parameters. Basically, the speakers, microphone and ambient noise of the training database are completely different from those of the CARNAV98 database. The recognition system used in the experiment was intended to recognise the utterances of 14 control commands and 100 Chinese names under various driving conditions from a data set of car 2. A simulated human–car voice interface was demonstrated in [6]. Herein, speech recognition rates are averaged over three male and three female speakers. It is reported that the baseline system (i.e. using SI HMM parameters without any adaptation) attains the recognition rates of 70.1%, 36.6% and 18.2% for driving conditions of standby, downtown and freeway, respectively. In the following experiments, the effects of training data size and estimation procedure on reference function estimation are investigated. The recognition rates and recognition speeds of baseline, MLLR, WPM and ABR methods were compared for various driving conditions. Finally, the feature representation adopting a mel-scale frequency cepstral coefficient (MFCC) is evaluated. Cepstral mean subtraction (CMS) [2] is included for comparison.

4.2 Effects of training data size and estimation procedure

First, the recognition performance of the ABR method when the amount of training data is changed is examined for estimation of a reference function. To conduct such an examination, the utterances of one male and one female from car 1 data set were chosen at random to generate the 1M1F data size. All utterances of the car 1 data set containing utterances of two males and two females are used to form the 2M2F data size. Note that, those utterances including three driving conditions were collected using a car type and speakers which were different from those used for the recognition data. In this comparison, the first estimation procedure is applied. As shown in Fig. 8, the ABR method outperforms the baseline system substantially. Also, the ABR method with a training data size of

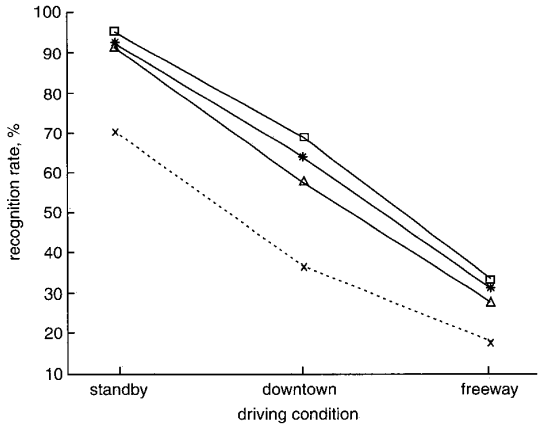

Fig. 8 *Comparison of recognition rates of baseline system and ABR under various driving conditions*

Different training data sizes and estimation procedures are examined for estimation of reference function
–□– ABR, 2M2F, second estimation
–*– ABR, 2M2F, first estimation
–△– ABR, 1M1F, first estimation
--×-- baseline

2M2F attains better recognition performance than that of 1M1F for various driving speeds. This is because the larger amount of training data provides richer statistics which enables better adaptation factors of the HMM parameters to be retrieved. However, the cost of data collection is also increased. For downtown driving, the recognition rates are improved from 57.7% for 1M1F to 63.9% for 2M2F, which are significantly better than the 36.6% of the baseline system.

On the other hand, the training data size was set to be 2M2F and the speech recognition rates were evaluated using different estimation procedures in the ABR method. From Fig. 8, it can be seen that the second estimation procedure achieves higher recognition rates than the first estimation procedure no matter what driving conditions are evaluated. For example, the recognition rates are increased from 92.5% using the first estimation procedure to 95% using the second estimation procedure in the standby condition. Such results reveal that integration of the pairs of each training frame $\boldsymbol{x}_t$ and all HMM units $\{\theta_{ik}\} = \{\mu_{ik}, \Sigma_{ik}\}$ into piecewise estimation of the reference function does improve the robustness of the estimated function $\boldsymbol{b}(\lambda_e)$. Hence, it is suggested that the second estimation procedure, without any need for data supervision, is a good choice for the ABR method.

4.3 Comparison of recognition rates and recognition speeds for different methods

It is also interesting to compare the recognition rates of the ABR method with the WPM and MLLR methods. In MLLR, one regression function shared by all HMM units was estimated on a per-utterance basis. The regression matrix was simplified to be diagonal [23]. Only one expectation-maximisation (EM) iteration [12] was performed. The recognition comparison is demonstrated in Fig. 9. It can be seen that the WPM and MLLR methods make significant progress compared with the baseline system. Further, MLLR adaptation performs a little better than the WPM. However, when the ABR method is considered, it is obvious that the ABR method obtains the best recognition performance for different car noise environments. For freeway driving, the recognition rates of

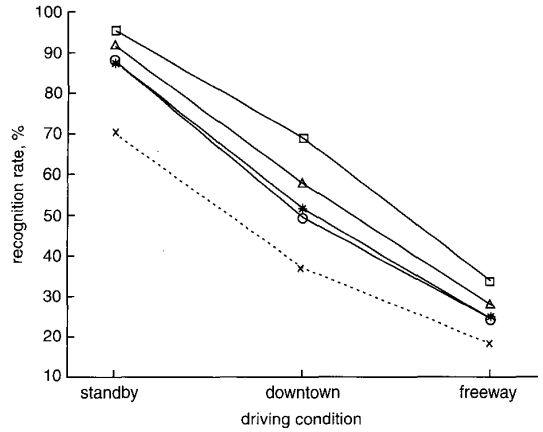

Fig. 9 *Comparison of recognition rates of baseline system, MLLR, WPM and ABR in cases of different training data sizes and estimation procedures for various driving conditions*

-□- ABR, 2M2F, second estimation
-△- ABR, 1M1F, first estimation
-○- WPM
-*- MLLR
--x-- baseline

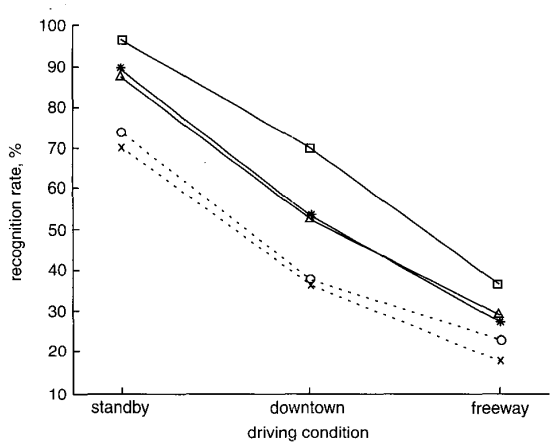

Fig. 10 *Comparison of recognition rates of baseline system, CMS, WPM and ABR in case of 2M2F data size and second estimation procedure for various driving conditions*

Feature vectors using LPCC and MFCC are evaluated
-□- ABR, MFCC
-*- WPM, MFCC
-△- CMS, MFCC
--○-- baseline, MFCC
--x-- baseline, LPCC

Table 2. Comparison of recognition times of baseline system, MLLR, WPM and ABR

Methods	Baseline	MLLR	WPM	ABR
Recognition time, s/utterance	0.41	0.91	0.67	0.74

WPM and MLLR are, respectively, 24.3% and 25%. The improvement using MLLR is not significant since the supervision is not reliably estimated in severe car-noisy environments. Conversely, using the ABR method, the recognition rates can be raised greatly to 33.6% when the 2M2F data size and the second estimation procedure are adopted. Nevertheless, the ABR method needs to collect and train a small data set to obtain a reference function. In addition recognition speeds of the baseline system, WPM, MLLR and ABR were determined. The recognition speeds are averaged over all test utterances and measured in seconds per utterance through simulating the algorithms on a Pentium II 350 personal computer. As reported in Table 2, the recognition cost of MLLR is the most expensive among the algorithms mentioned because two passes of Viterbi decoding are needed for supervision estimation and speech recognition. On the other hand, WPM and ABR are frame-synchronous methods in which model adaptation and speech recognition are carried out in the same session. The computational overheads of WPM and ABR are spent on the calculation of optimal equalisation factors and the adaptation of HMM parameters for each frame likelihood measure.

4.4 Effect of feature representation using MFCC

Furthermore, the ABR method was compared with another common feature representation based on MFCC. In the following experiments, the feature vector is switched to 12-order MFCC, 12-order delta MFCC, one delta log energy and one delta delta log energy. The baseline system, CMS, WPM and ABR, was carried out. The baseline results using LPCC are given for comparison. In ABR, the reference function is re-estimated using the 2M2F data size and following the second estimation procedure. The recognition rates are reported in Fig. 10. It can be seen that feature representation using MFCC performs better than that using LPCC. For freeway driving, the recognition rate is increased to 23.1%. This reveals that MFCC is a good feature representation for in-car speech recognition. Moreover, the performance of CMS and WPM is similar, but CMS is not frame-synchronous. The ABR is still significantly better than other methods in different driving conditions. In the standby condition, the recognition rates are 87.5% and 89.2% for CMS and WPM, respectively. However, ABR can achieve 96.5%. From all the experimental results, the superiority and feasibility of the proposed ABR method for hands-free speech recognition in car environments are confirmed.

5 Conclusions

This paper has presented a novel frame-synchronous model adaptation approach to hands-free car speech recognition. The proposed approach was designed to compute an optimal scaling factor to equalise the HMM mean vector towards the input observation vector for an individual frame likelihood measure. The adaptation bias caused by an optimal equalisation factor was subsequently compensated by a reference vector, which was retrieved according to the index of the optimal equalisation factor. During training, a small set of speech from a car noise environment was used to estimate the reference function containing important adaptation behaviours. After a series of experimental investigations, it was found that the increase of training data size did benefit the estimation of a statistically rich reference function. The estimation procedure could be improved by considering the data pairs of each observation frame associated with all HMM parameters. The needs for data supervision and Viterbi decoding were neglected. In addition, it was shown that the proposed method consumed moderate recognition time and achieved higher recognition rates compared with the baseline system or the CMS, WPM and MLLR methods. The feature representations using LPCC and MFCC were feasible for the proposed method. All investigations were

carried out across three kinds of driving conditions using a hands-free microphone far from the talker. In future work, we would like to enhance the individuality of reference function with respect to various HMM parameters to improve the performance further. Also, other types of relational index should be explored and more car classes should be tested for the proposed ABR method.

6 Acknowledgments

The authors acknowledge the valuable comments of anonymous reviewers, which considerably improved the paper presentation. This research has been partially supported by Computer & Communication Research Laboratories, Industrial Technology Research Institute, Taiwan under contract no. G3-88037.

7 References

1 ANASTASAKOS, T., and BALAKRISHNAN, S.V.: 'The use of confidence measure in unsupervised adaptation of speech recognizers'. Proceedings of international conference on *Spoken Language Processing (ICSLP)*, 1998, pp. 2303–2306
2 ATAL, B.: 'Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification', *J. Acoust. Soc. Am.*, 1974, **55**, pp. 1304–1312
3 BOLL, S.F.: 'Suppression of acoustic noise in speech using spectral subtraction', *IEEE Trans. Acoust. Speech Signal Process.*, 1979, **ASSP-27**, pp. 113–120
4 CARLSON, B.A., and CLEMENTS, M.A.: 'A projection-based likelihood measure for speech recognition in noise', *IEEE Trans. Speech Audio Process.*, 1994, **2**, pp. 97–102
5 CHIEN, J.-T.: 'Online hierarchical transformation of hidden Markov models for speech recognition', *IEEE Trans. Speech Audio Process.*, 1999, **7**, (6), pp. 656–667
6 CHIEN, J.-T., and LIN, M.-S.: 'Frame synchronous noise compensation for car speech recognition'. Proceeding of 12th conference on *Research on Computational Linguistics (ROCLING)*, Hsinchu-Taiwan, 1999, pp. 239–251 (in Chinese)
7 CHIEN, J.-T., and JUNQUA, J.-C.: 'Unsupervised hierarchial adaptation using reliable selection of cluster-dependent parameters', *Speech Commun.*, 2000, **30**, (4), pp. 235–253
8 CHIEN, J.-T., and WANG, H.-C.: 'Telephone speech recognition based on Bayesian adaptation of hidden Markov models', *Speech Commun.*, 1997, **22**, pp. 369–384
9 CHIEN, J.-T., LEE, L.-M., and WANG, H.-C.: 'Noisy speech recognition using variance adapted likelihood measure'. Proceedings of IEEE international conference on *Acoustics, Speech and Signal Processing (ICASSP)*, 1996, pp. 45–48
10 COMPERNOLLE, D.V.: 'Speech recognition in the car: from phone dialing to car navigation'. Proceedings of European Conference on *Speech Communication and Technology (Eurospeech)*, 1997, **5**, pp. 2431–2434
11 DELPHIN-POULAT, L., and MOKBEL, C.: 'Frame-synchronous adaptation of cepstrum by linear regression'. Proceedings of 1997 IEEE workshop on *Automatic Speech Recognition and Understanding*, 1997, pp. 420–427
12 DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B.: 'Maximum likelihood from incomplete data via the EM algorithm', *J. R. Statist. Soc. B*, 1977, **39**, pp. 1–38
13 DIGALAKIS, V., RTISCHEV, D., and NEUMEYER, L.G.: 'Speaker adaptation using constrained estimation of Gaussian mixtures', *IEEE Trans. Speech Audio Process.*, 1995, **3**, pp. 357–366
14 EPHRAIM, Y.: 'A Bayesian estimation approach for speech enhancement using hidden Markov models', *IEEE Trans. Signal Process.*, 1992, **40**, (4), pp. 725–735
15 GALES, M.J.F.: 'Cluster adaptive training for speech recognition'. Proceedings of international conference on *Spoken Language Processing (ICSLP)*, 1998, **5**, pp. 1783–1786
16 GALES, M.J.F., and YOUNG, S.J.: 'Robust continuous speech recognition using parallel model combination', *IEEE Trans. Speech Audio Process.*, 1996, **4**, pp. 352–359
17 GAUVAIN, J.L., and LEE, C.-H.: 'Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains', *IEEE Trans. Speech Audio Process.*, 1994, **2**, pp. 291–298
18 GONG, Y.: 'Speech recognition in noisy environments: a survey', *Speech Commun.*, 1995, **16**, pp. 261–291
19 HOMMA, S., TAKAHASHI, J., and SAGAYAMA, S.: 'Improved estimation of supervision in unsupervised speaker adaptation'. Proceedings of IEEE international conference on *Acoustics, Speech and Signal Processing (ICASSP)*, 1997, pp. 1023–1026
20 HUNT, M.J.: 'Some experience in in-car speech recognition'. Proceeding of workshop on *Robust Methods for Speech Recognition in Adverse Conditions*, 1999, Tampere, Finland, pp. 25–31
21 JUNQUA, J.-C.: 'The influence of acoustics on speech production: a noise-induced stress phenomenon known as the Lombard reflex', *Speech Commun.*, 1996, **20**, pp. 13–22
22 LEE, C.-H.: 'On stochastic feature and model compensation approaches to robust speech recognition', *Speech Commun.*, 1998, **25**, (1–3), pp. 29–47
23 LEGGETTER, C.J., and WOODLAND, P.C.: 'Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models', *Comput. Speech Language*, 1995, **9**, pp. 171–185
24 MANSOUR, D., and JUANG, B.-H.: 'A family of distortion measures based upon projection operation for robust speech recognition', *IEEE Trans. Acoust. Speech Signal Process.*, 1989, **37**, pp. 1659–1671
25 NGUYEN, P., WELLEKENS, C., and JUNQUA, J.-C.: 'Maximum likelihood eigenspace and MLLR for speech recognition in noisy environments'. Proceedings of European conference on *Speech Communication and Technology (Eurospeech)*, 1999, **6**, pp. 2519–2522
26 RABINER, L.R.: 'A tutorial on hidden Markov models and selected applications in speech recognition', *Proc. IEEE*, 1989, **77**, pp. 257–286
27 VITERBI, A.J.: 'Error bounds for convolutional codes and an asymptotically optimum decoding algorithm', *IEEE Trans. Inf. Theory*, 1967, **IT-13**, pp. 260–269
28 WOODLAND, P.C., GALES, M.J.F., and PYE, D.: 'Improving environmental robustness in large vocabulary speech recognition'. Proceedings of IEEE international conference on *Acoustic, Speech and Signal Processing (ICASSP)*, 1996, **1**, pp. 65–68