Chapter 10
A Likelihood-Maximizing Framework for Enhanced In-Car Speech Recognition Based on Speech Dialog System Interaction

Tristan Kleinschmidt, Sridha Sridharan, and Michael Mason

Abstract Speech recognition in car environments has been identified as a valuable means for reducing driver distraction when operating noncritical in-car systems. Under such conditions, however, speech recognition accuracy degrades significantly, and techniques such as speech enhancement are required to improve these accuracies. Likelihood-maximizing (LIMA) frameworks optimize speech enhancement algorithms based on recognized state sequences rather than traditional signal-level criteria such as maximizing signal-to-noise ratio. LIMA frameworks typically require calibration utterances to generate optimized enhancement parameters that are used for all subsequent utterances. Under such a scheme, suboptimal recognition performance occurs in noise conditions that are significantly different from that present during the calibration session – a serious problem in rapidly changing noise environments out on the open road. In this chapter, we propose a dialog-based design that allows regular optimization iterations in order to track the ever-changing noise conditions. Experiments using Mel-filterbank noise subtraction (MFNS) are performed to determine the optimization requirements for vehicular environments and show that minimal optimization is required to improve speech recognition, avoid over-optimization, and ultimately assist with semi-real-time operation. It is also shown that the proposed design is able to provide improved recognition performance over frameworks incorporating a calibration session only.

Keywords Automatic speech recognition (ASR) • In-car speech recognition • LIMA frameworks • Mel-filterbank noise subtraction (MFNS)

T. Kleinschmidt (✉) • S. Sridharan • M. Mason
Speech & Audio Research Laboratory, Queensland University of Technology, Brisbane, QLD, Australia
e-mail: t.kleinschmidt@qut.edu.au; s.sridharan@qut.edu.au; m.mason@qut.edu.au

J.H.L. Hansen et al. (eds.), *Digital Signal Processing for In-Vehicle Systems and Safety*, DOI 10.1007/978-1-4419-9607-7_10,

10.1 Introduction

With the increased desire from consumers to integrate electronic devices such as MP3 players, navigation systems, and mobile phones for use in their vehicles comes the need to provide more intuitive human–machine interfaces (HMI) than currently seen in low- to midrange vehicles. Automatic speech recognition (ASR) can provide a safe and easy-to-use HMI, and technological advancements have enabled low-cost hardware implementations of ASR systems – a key requirement to widespread adoption in the automotive industry.

Most ASR systems are trained for use in controlled scenarios (e.g., office environments or telephone-based systems) and fail to produce satisfactory performance under the continually changing noise conditions found in automotive environments [1]. This is a key challenge to deployment of in-car ASR – drivers demand high-accuracy recognition, but high levels of noise restrict recognition performance of conventional ASR systems.

Speech enhancement is a common method for making ASR systems more robust against noise. Enhancement techniques aim to reduce the noise levels present in speech signals, allowing clean speech models (which are easily trained due to the availability of large amounts of data) to be utilized by the recognizer. This is a popular approach as enhancement algorithms are typically easily integrated with existing ASR front-end processing, as well as requiring little-to-no prior knowledge of the operating environment in order to achieve improvements in recognition accuracy. Both of these aspects are particularly attractive for in-car applications where hardware and software overheads must be minimized and where the system is continually subjected to changes in acoustic conditions.

Popular speech enhancement algorithms such as filter-and-sum beamforming (using multiple microphone speech acquisition) and spectral subtraction were originally designed to improve intelligibility and/or quality of speech signals without considering the effects on other speech processing systems such as recognition [2]. Optimization of parameters in these algorithms focuses on signal-based measures (e.g., maximizing signal-to-noise ratio or minimization of the mean-squared signal error). Enhancement techniques operating in this manner may still produce word accuracy improvements, but these improvements are by-products of the optimization process rather than its objective [2].

Promising results have been shown in studies that use speech recognition likelihoods as the optimization criteria as opposed to quality or intelligibility measures [2–4]. Enhancement techniques are placed within likelihood-maximizing (LIMA) frameworks, which attempt to *jointly* optimize both the recognized acoustic state sequence as well as enhancement parameters. There are three main types of LIMA framework – calibrated, unsupervised, and supervised.

Calibrated LIMA frameworks require a known adaptation utterance in order to optimize the enhancement parameters. Adaptation is typically performed using a dedicated calibration session for each speaker, with the optimized enhancement parameters kept constant for all other utterances for that speaker [2, 3]. This approach assumes constant noise conditions and therefore has limited potential for achieving optimal performance in rapidly changing vehicular environments.

An unsupervised LIMA framework was also proposed in [2] whereby online optimization takes place on an utterance-by-utterance basis using the hypothesized transcription as opposed to the true transcription. Whilst this method removes the restriction of a calibration session and showed considerable reductions in word error rates [2], it is highly reliant on the initial accuracy of the speech recognizer. Whilst the word error rate of the recognizer used in these experiments was high (approximately 60%), the test recordings were obtained at relatively high signal-to-noise ratios in a constant noise environment. Systems operating in the nonstationary vehicular environment exhibit even higher word error rates, resulting in reductions in accuracy of the hypothesized transcriptions. Optimization on unreliable transcriptions should be avoided as it could lead to suboptimal parameter estimation and therefore further reductions in recognition performance.

In this chapter, we consider the third alternative (i.e., a supervised LIMA framework) and propose a dialog-based design that allows regular optimization iterations in order to track the ever-changing noise conditions. The chapter reviews LIMA frameworks employing Mel-filterbank noise subtraction (MFNS) specifically for in-car speech recognition. The analysis involves testing a number of calibrated adaptation scenarios, as well as development of a novel online optimization framework, based on speech dialog systems which exploit user confirmation of correctly recognized voice commands to provide adaptation data for the LIMA framework.

10.2 LIMA Mel-Filterbank Noise Subtraction for In-Car Environments

10.2.1 Likelihood Maximization

Speech enhancement algorithms aim to produce improvements in human intelligibility of speech signals. Automatic speech recognition systems hypothesize the most likely sequence of statistical models produced by the observed feature vectors. As a result, traditional optimization of spectral subtraction algorithms based on waveform criteria such as signal-to-noise ratio maximization [5, 6] does not necessarily translate into improvements in ASR word accuracy [2]. With the primary aim of using speech enhancement to improve speech recognition accuracy, Seltzer et al. [2] proposed a likelihood-maximization framework for enhancement parameter optimization. This framework was originally proposed for filter-and-sum beamforming but has since been applied to subtraction factors in multiband spectral subtraction [3].

In a recognition system incorporating speech enhancement, feature vectors are a function of the speech enhancement process. The recognition hypothesis provided by an optimal Bayes classifier regularly used in ASR systems is given by

$$\hat{w} = \arg\max_{w \in W} P(Z(\xi)|w) \cdot P(w), \tag{10.1}$$

where dependence of the feature vectors Z on the enhancement parameters ξ is clearly shown. The acoustic score $P(Z(\xi))$ is the measure of importance in LIMA systems as the transcription on which the optimization takes place is assumed to be known, and therefore, the language model score $P(w)$ will not change. The aim of likelihood maximization for MFNS is therefore to optimize the parameters to maximize the acoustic score of the recognized word sequence $\hat{w}$.

An initial decode pass is performed using default enhancement parameters to generate a state sequence s on which to optimize ξ. In order to find the optimal values of ξ, gradient-based optimization is used on the total log-likelihood of the observed features, which is defined by

$$L(\xi) = \sum_i \log(P(z_i(\xi)|s_i)). \tag{10.2}$$

For a Hidden Markov Model (HMM) speech recognizer using Gaussian mixture state models (as used in this chapter), the gradient of the total log-likelihood is given by [2]

$$\nabla_\xi L(\xi) = -\sum_i \sum_{m=1}^{M} \gamma_{im}(\xi) \frac{\partial z_i(\xi)}{\partial \xi} \sum_{im}^{-1} (z_i(\xi) - \mu_{im}), \tag{10.3}$$

where $\gamma_{im}(\xi)$ is the a posteriori probability of the mth mixture component in state s_i given the observed feature vector $z_i(\xi)$. The mean vector μ and covariance matrix $\sum$ from the acoustic model are required for each state i and mixture component m in order to calculate the gradient. The remaining term in Eq. 10.3 is the Jacobian matrix, $\partial z_i(\xi)/\partial \xi$, which consists of the partial derivatives of each element of the feature vector with respect to each of the enhancement parameters. Each Jacobian element is derived directly from the enhancement procedure (refer to Sect. 10.2.3). Once the gradient-based optimization converges, the new enhancement parameters are used to generate another set of feature vectors, and a subsequent decode pass is performed. A new state sequence is generated, and the enhancement parameters are further optimized for this new state sequence. The process continues until the recognition likelihood (and state sequence) converges, ensuring joint optimization of the recognized state sequence and the speech enhancement parameters.

10.2.2 Optimization Methods for In-Car ASR

10.2.2.1 Calibrated LIMA Framework

The simplest and most common approach for optimizing the enhancement parameters is to use a calibration session with a known transcription w_C. Previous studies used a single known utterance for each speaker in order to determine

optimal enhancement parameters for that particular speaker [2, 3]. Whilst this procedure ensures that optimization takes place on a state sequence which is correct, calibrated LIMA frameworks inherently assume that background noise conditions do not change between the calibration and testing sessions. This is a major challenge for in-car speech recognition since vehicular environments are subjected to continually changing noise levels and conditions which mean calibration utterances would be required every time noise conditions changed significantly from the previous optimization. To overcome this, optimized enhancement parameters could be stored for each common noise condition; however, this still requires a calibration utterance to be used at some point in the system. Since there is a wide range of noise conditions, the user would be continually asked to repeat the adaptation utterance in order to obtain the optimal set of parameters. This operation is an unnecessary annoyance for the driver and is likely to lead drivers to become frustrated with the speech dialog system; such emotions could lead to further repercussions on ASR and driving performance [7].

An alternative solution is to calibrate once only for each driving session (e.g., a common startup utterance such as "Start dialog" could be used for adaptation), but this introduces the risk of inferior recognition in noise conditions significantly different to those present during calibration.

The calibration framework is also reliant on the words contained in the adaptation utterance; therefore, it is necessary for the adaptation utterance to be phonetically balanced and sufficiently long enough to provide as much acoustic model coverage as possible in order to generalize the optimized enhancement parameters. This is in direct conflict with the majority of dialog systems which promote simpler linguistic structures than human conversation and are therefore unlikely to be phonetically balanced. Thus, a separate utterance unrelated to the dialog transaction is required which is likely to be seen by the user as a further inconvenience and therefore impractical for this particular application.

10.2.2.2 Unsupervised LIMA Framework

The unsupervised LIMA framework proposed in [2] may be a more appropriate choice for in-car environments. Unsupervised adaptation removes the restriction of a calibration utterance (thereby making the adaptation process transparent to the user), and instead, optimization takes place on an utterance-by-utterance basis. The major issue with the unsupervised operation is that it uses a hypothesized transcription, w, rather than the true transcriptionw_C. The hypothesis transcription is highly reliant on the effectiveness of the underlying acoustic models and state sequence generated by Viterbi alignment; therefore, the hypothesis transcription is likely to be less than 100% correct.

Since the true transcription w_C is unknown, it is possible that states in the hypothesized transcription $\hat{w}$ are incorrect due to misrecognition and frame alignment errors (N.B. frame alignment errors will occur even when the transcription is known a priori, but should be limited). These inaccurate states will lead to the

resulting enhancement parameters being suboptimal since optimization is performed on the wrong state models. In turn, suboptimal enhancement parameters could lead to further decreases in accuracy in the subsequent decoding state. This effect is particularly likely when the number of incorrectly labeled frames is greater than the correctly labeled frames, as may be the case in high-noise conditions.

10.2.2.3 Proposed Dialog-Based LIMA Framework

Having identified problem with the existing LIMA frameworks, we propose to exploit a confirmation-based speech dialog system to drive optimization. Dialog systems requiring users to verify commands with simple "Yes/No" replies are a well-established mechanism in voice recognition applications. A block diagram of the proposed framework within the dialog exchange is shown in Fig. 10.1.

This system mimics the calibrated and unsupervised frameworks by performing an initial decode using default enhancement parameter values in the feature extraction stage. This framework differs from previous work following the initial ASR pass. Instead of immediately performing optimization, the hypothesized word sequence is first verified through the grounding process which is required in the dialog system in order to detect any misrecognition errors which need to be corrected prior to executing a desired action such as determining route navigation.

Since it is cumbersome for the dialog manager to request confirmation from the user after each response, grounding often occurs once the dialog systems have gathered a number of pieces of information, for example the suburb, street name, and number of a destination address. In the case where the user states the information is incorrect, the dialog manager will attempt to recover from these errors by either asking for corrections to specific information or restarting the dialog transaction. In this instance, the enhancement parameters remain unaltered.

It is also possible to incorporate knowledge of the state of the car environment to alter the enhancement parameters should the noise condition change drastically between optimizations. The purpose of this chapter is not to suggest how this should be done but to analyze the performance of existing and proposed LIMA frameworks and make recommendations on how these are best utilized in automotive environments.

When the user confirms the information to be correct, this affirmation is fed back to the dialog manager for further processing (e.g., a call to an external information source such as the navigation system) but also triggers the optimization of the enhancement parameters. In order to interface the optimization process with the grounding procedure, it is required to store the user responses as well as the hypothesized state sequences – this is reflected in Fig. 10.1. On confirmation, this stored information is used in the optimization process; if rejected, the stored state sequence is therefore unreliable, and so, the memory can be cleared in preparation for responses in the error-recovery stage.

The primary advantage of the proposed dialog-based LIMA framework is that optimization never takes place on inaccurate transcription hypotheses, which overcomes the limitation of the unsupervised framework. Another advantage is the

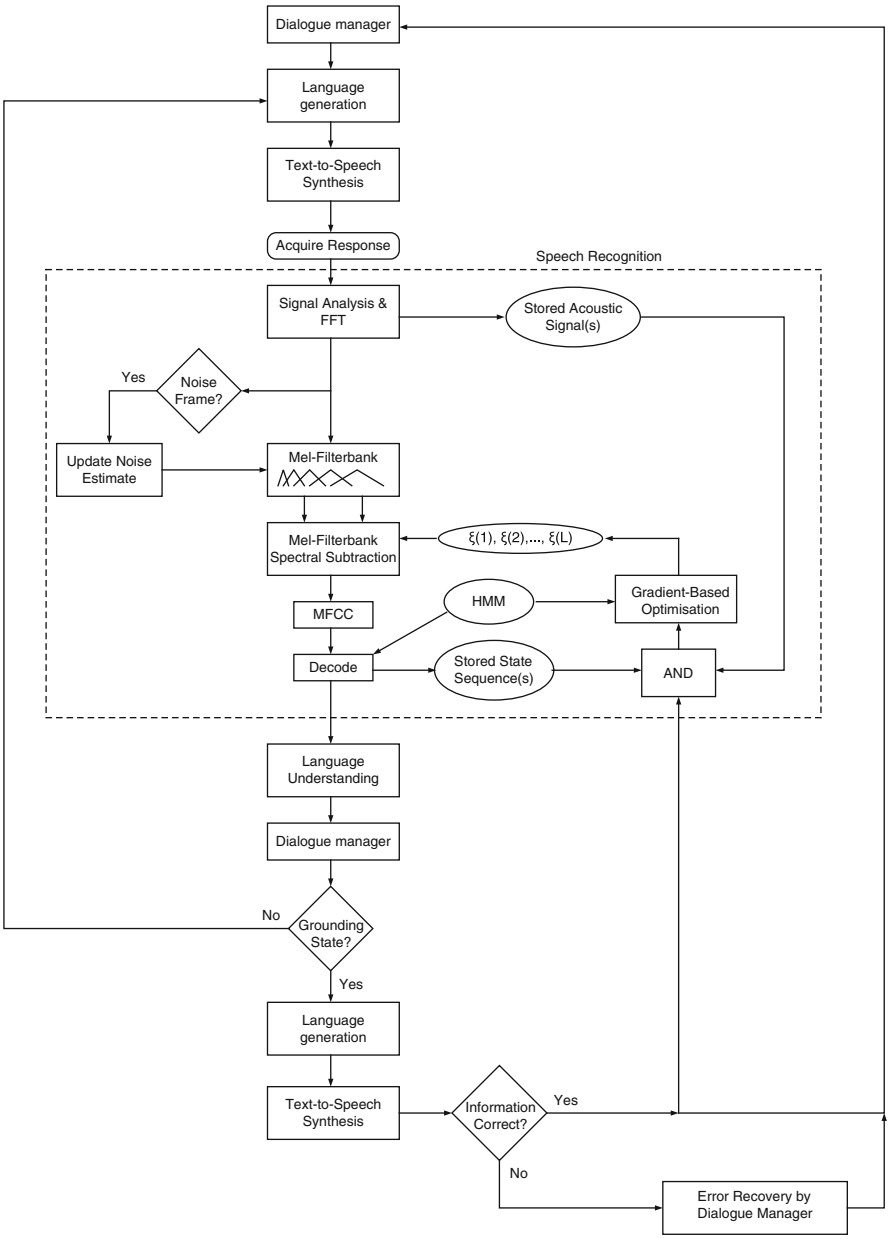

Fig. 10.1 Proposed confirmation-based speech dialog system for in-car speech recognition using LIMA speech enhancement

ability to continually update the enhancement parameters as the noise conditions inside the vehicle change. This is achieved by maintaining the previous enhancement parameters until the next successful dialog transaction, by which time the noise conditions may have changed. As a result, the dialog-based system is able to overcome the need for matched noise conditions required for calibrated operation to be fully effective.

10.2.3 Mel-Filterbank Noise Subtraction

In this chapter, we concentrate on spectral subtractive enhancement algorithms for this application. Spectral subtraction for speech enhancement was originally proposed by Boll in 1979 [8]. Enhancement is typically performed in the frequency domain; however, subband subtraction techniques such as the Mel-filterbank noise subtraction (MFNS) method proposed in [9] have become popular for use with recognition systems. BabaAli et al. [7] recently utilized the framework introduced in [2] to optimize the subtraction scaling factors in multiband spectral subtraction in the frequency domain.

In a noisy environment, speech $S(f)$ is assumed to be corrupted by uncorrelated additive background noise $D(f)$ to produce corrupted speech $Y(f)$:

$$Y^i(f) = S^i(f) + D^i(f), \tag{10.4}$$

where frequency spectra are obtained from the short-time Fourier transform of frame i.

Generally, an estimate of the background noise magnitude spectrum is subtracted from the magnitude spectrum of the noisy signal to give an estimate of the clean speech magnitude. Noise estimates are calculated during nonspeech periods and are typically kept constant throughout speech periods. In the following, the frame index i has been removed from the noise estimate to reflect this operation.

In this chapter, however, we consider Mel-filterbank noise subtraction [9]. Using the Mel-frequency scale commonly used in speech recognition, the frequency spectrum is divided into a number of subbands with f_U^k and f_L^k being the upper and lower cutoff frequencies for the kth Mel-filterbank, respectively. Using this definition, Mel-filterbank noise subtraction is described by

$$\begin{aligned}
E_Y^i(k) &= \int_{f_L^k}^{f_U^k} |Y^i(f)| \mathrm{d}f \\
E_{\hat{D}}(k) &= \int_{f_L^k}^{f_U^k} |\hat{D}(f)| \mathrm{d}f \\
E_{\hat{S}}^i(k) &= \begin{cases} E_Y^i(k) - \alpha(k) E_{\hat{D}}(k) & E_Y^i(k) > \frac{\alpha(k)}{1-\beta} E_{\hat{D}}(k) \\ \beta E_Y^i(k) & \text{otherwise} \end{cases}
\end{aligned} \tag{10.5}$$

where $E_Y^i(k)$, $E_{\hat{D}}(k)$, and $E_{\hat{S}}^i(k)$ are the energies of the kth Mel-filterbank of the noisy speech, noise estimate, and the clean speech estimate, respectively. The scaling factor β enforces a maximum level of signal energy attenuation and ensures that output filterbank energies remain positive. Filterbank-dependent subtraction factors – $\alpha(k)$ – are included to compensate for estimation inaccuracies of the instantaneous noise energy. In the experiments that follow, only the subtraction factors are optimized, that is:

$$\xi = [\alpha_1, \alpha_2, ..., \alpha_K]. \tag{10.6}$$

The expression for the Jacobian elements $\partial z_i(\alpha(k))/\partial\alpha(k)$ for each enhancement parameter can be derived as per [10] to produce

$$\frac{\partial z_i(\alpha(k))}{\partial\alpha(k)} = -\frac{1}{2}\sum_{k=0}^{K-1}\frac{\Phi_{ck}E_{\hat{D}}^i(k)}{\hat{E}_S^i(k)} \times \left(1 + \frac{E_Y^i(k)(1-\beta) - \alpha(k)E_{\hat{D}}^i(k)}{|E_Y^i(k)(1-\beta) - \alpha(k)E_{\hat{D}}^i(k)|}\right), \tag{10.7}$$

where Φ_{ck} are elements of the DCT matrix for cepstral coefficient c.

10.3 Experimental Procedures

10.3.1 Experimental Data

Digit strings comprising the phone numbers task of the AVICAR database collected by the University of Illinois [11] were used as the test data. The AVICAR database contains real speech recorded in five different driving conditions: idle (IDL), 35 mph with windows up (35U) and down (35D), and 55 mph with windows up (55U) and down (55D). All experiments utilized an altered version of the first five experimental folds of the AVICAR evaluation protocol developed in [12]. The data for this evaluation consists of 38 speakers, all of which have at least one utterance available in all of the noise conditions.

10.3.2 Speech Recognizer

Utterance decoding was performed using the HMM Toolkit [13]. Speaker-independent, context-dependent 3-state triphone HMM acoustic models were trained using the Wall Street Journal 1 corpus. Each HMM state was represented using a 16-component Gaussian mixture model.

For each observation, 39-dimensional MFCC feature vectors were generated consisting of 13 MFCC (including C_0) plus 13 delta and 13 acceleration coefficients. Cepstral mean subtraction was applied to each feature. The elements of the Jacobian were derived from this feature representation as per Eq. 10.7.

The recognition task uses an open word loop grammar [12]; therefore, no restrictions are made to ensure that exactly ten digits are recognized.

All speech recognition results quoted in this chapter are word accuracies (in %) and are calculated as

$$\text{Accuracy} = \frac{N - D - S - I}{N} \times 100, \tag{10.8}$$

where N represents the total number of words, D the number of deletions, S the number of substitutions, and I the number of insertions [13].

10.3.3 Optimization Iterations

Since LIMA is an optimization problem, over-optimization of the enhancement parameters to a specific noise condition, speaker, or subset of acoustic state models is highly possible and should be avoided. This suggests that the number of optimization iterations should not be large in order to maintain generality across conditions, but too little iteration may result in the LIMA framework operating less effectively than a standard enhancement system. Considering real-time operation (another important consideration for in-car ASR) also points to limited iterations.

To address this issue, two experiments were designed to determine a suitable balance between ASR performance and pseudo real-time operation using the noise-only calibration framework described in Sect. 10.3.4. This framework was used since the belief was that noise conditions have a greater effect on the resulting enhancement parameters than individual speakers since speaker-independent acoustic models are being used.

In the first experiment, the number of gradient-descent iterations was varied whilst using a single joint optimization iteration (i.e., full recognition and parameter optimization cycles). The second experiment varied the number of joint optimization iterations whilst the gradient-descent iterations (determined from the former experiment) were kept constant. The combined outcomes of these experiments dictated the levels of optimization used for assessing the frameworks detailed in Sect. 10.3.4.

For all experiments, the enhancement parameters were initialized to $\alpha(k) = 1$ for all 26 Mel-filterbanks. These values were an appropriate initial guess since standard MFNS using these values provides improvements in speech recognition accuracy over a system without enhancement [10].

10.3.4 Likelihood-Maximization Frameworks

The AVICAR database enables analysis of LIMA frameworks based on speaker or noise calibration as well as a combination of both. The following LIMA frameworks have been tested:

- Calibrated LIMA framework using optimization on a noise-by-noise basis
- Calibrated LIMA framework using optimization on a speaker-by-speaker basis under a single, randomly chosen noise condition
- Calibrated LIMA framework using optimization for each speaker in each noise conditions (i.e., matched conditions)
- Proposed dialog-based LIMA framework without calibration
- Proposed dialog-based LIMA framework with a single calibration utterance in a random noise condition
- Proposed dialog-based LIMA framework with a single calibration utterance in the idle noise condition

The unsupervised LIMA frameworks were not assessed in this chapter as the overall performance of the speech recognizer is low (less than 50% average word accuracy), making the hypothesis transcriptions (and therefore the optimized parameters) unreliable.

Each calibrated LIMA framework used a single, randomly generated utterance treated as adaptation session. For the noise-only calibration framework, a random utterance from a random speaker was chosen for each experimental fold in the evaluation protocol. For speaker-based calibration (applied in both calibrated and dialog frameworks), a single utterance from a random noise condition was used for each speaker, with the remaining utterances ordered randomly to simulate realistic driving conditions.

The proposed dialog system was run using no prior calibration, and optimization occurred every time the decoder correctly recognized *all* ten digits in the phone number. Utterances which occur prior to the first optimization exhibit the same performance as the static MFNS system and are therefore ignored in the final evaluation (N.B. this is why baseline results differ across the experiments).

In order to simulate a priori knowledge relating to previously optimized enhancement parameters, the dialog-based framework was also tested using an initial adaptation utterance which was either randomly chosen or from the idle condition. The idle condition was chosen as this is a likely scenario for users to first communicate with the in-car speech dialog system – for instance, for entering a destination address before setting off on the journey. Again, all utterances which occurred prior to the first subsequent optimization (excluding calibration) were ignored in the evaluation.

10.4 Data Analysis and Recommendations

10.4.1 Gradient-Descent Iterations

The effect on ASR word accuracy as the number of gradient-descent iterations increases is shown in Table 10.1. Recognition results with no enhancement (baseline) and MFNS with static subtraction parameters ($\alpha(k) = 1$) are shown for comparison.

Analysis of these results shows that the optimal number of gradient-descent iterations is considerably different for each noise condition. For the more quiet conditions (idle and 35 mph with windows up), best performance is obtained with more than 20 iterations of gradient-descent optimization. For the noisier conditions, less than five optimization iterations provide the best performance (particularly for the 55-mph-with-windows-down noise condition). These three conditions also show trends of decreasing word accuracy as the number of iterations is increased above five. Since the noise conditions are approximately ordered by increasing levels of noise, it can be concluded that as the noise levels in the vehicle increase (i.e., higher speeds or open windows), the level of gradient-descent optimizations needs to be reduced in order to avoid over-optimization of the enhancement parameters.

The application of only one gradient-descent iteration provides a minimum of 0.3% improvement static MFNS, with both 35-mph scenarios improving by approximately 1%. A single iteration shows the effectiveness of a LIMA framework for improving ASR performance with minimal optimization.

The best overall performance across all five noise conditions is seen at three iterations. At this level of optimization, the 55-mph conditions both exhibit maximum performance, with two other noise conditions being only 0.1% below their best performance (IDL and 35D). The 35-mph-with-windows-up condition is the only

Table 10.1 ASR accuracies for increasing gradient-descent iterations used in parameter optimization

# Iterations	IDL	35U	35D	55U	55D
Baseline	70.4	48.8	36.2	41.8	23.5
$\alpha(k) = 1$	73.3	47.8	36.8	44.5	26.1
1	73.9	48.7	37.9	44.8	26.4
2	74.2	49.3	37.7	44.8	26.4
3	74.1	49.1	38.1	45.1	26.4
4	74.2	49.5	37.8	45.1	26.1
5	74.1	49.6	38.2	45.0	25.9
10	74.2	49.7	37.7	44.6	26.1
15	74.2	49.8	37.5	44.8	25.6
20	74.2	49.9	37.6	44.7	25.7
25	74.2	49.9	37.6	44.7	25.7

Table 10.2 ASR results for increasing number of joint optimization iterations

# Iterations	IDL	35U	35D	55U	55D
Baseline	70.4	48.8	36.2	41.8	23.5
$\alpha(k) = 1$	73.3	47.8	36.8	44.5	26.1
1	**74.1**	**49.1**	**38.1**	**45.1**	**26.4**
2	74.1	49.4	37.7	44.8	26.1
3	73.9	49.9	37.2	44.8	26.0
4	74.0	50.1	37.2	44.5	26.3
5	74.0	50.3	37.1	44.4	26.1
10	74.1	50.2	37.5	44.1	25.9

one which is well below its best performance (0.8%) but still provides improvement over the baseline and static MFNS systems. As a result, three gradient-descent iterations have been used for the remainder of the experiments in this chapter.

10.4.2 Joint Optimization Iterations

Having established the most effective number of gradient-descent iterations, the number of joint optimization iterations was analyzed. Table 10.2 shows these results with the best performance across all noise conditions highlighted for clarity.

Apart from the 35-mph-with-windows-up noise condition, the results indicate that only one joint optimization iteration is required for in-car speech recognition. This result indicates that only minor changes are made to the decoded state sequences and therefore appears to be no advantage in performing more than one joint optimization iteration. Relating this observation to the results of the gradient-descent iterations experiment, if the state sequence did not change at all, the parameter optimization would continue from exactly the same position that it finished previously, and therefore, over-optimization is likely to occur as the number of joint optimization iterations increases.

This result combined with that of Sect. 10.4.1 indicates that over-optimization is a serious issue for LIMA frameworks operating in vehicular environments. It is therefore suggested that optimization iterations be kept to a minimum in order to keep the enhancement parameters generalized. The practical advantage of these findings is the ability to achieve improved ASR using LIMA frameworks whilst creating minimal processing delays due to the need for only a few optimization iterations.

10.4.3 LIMA Frameworks

The LIMA frameworks listed in Sect. 10.3.4 were tested using the results obtained in the previous experiments. Table 10.3 presents the ASR results for all three

Table 10.3 ASR results for the calibrated LIMA frameworks

Adaptation condition	IDL	35U	35D	55U	55D
Baseline	70.4	48.8	36.2	41.8	23.5
$\alpha(k) = 1$	73.3	47.8	36.8	44.5	26.1
Noise	74.1	49.1	38.1	45.1	26.4
Speaker	73.6	49.5	38.2	44.9	26.5
IDL	**73.7**	49.3	37.8	44.6	26.8
35U	73.8	**49.9**	38.6	45.0	27.0
35D	73.0	49.4	**39.2**	45.1	26.7
55U	74.2	49.7	37.9	**45.5**	26.8
55D	73.1	49.1	38.2	44.7	**27.1**

calibrated frameworks. The matched calibrate-test conditions for speaker-based calibration are highlighted for clarity. Regardless of the calibration method used, the results show a global improvement over an enhancement system which does not utilize a LIMA framework.

Using matched conditions for speaker-based adaptation (i.e., employing calibration for each speaker in each noise condition) provides the best word accuracies in all cases except idle. Whilst the idle noise condition shows a 0.5% absolute decrease in word accuracy in its matched condition (as opposed to optimizing in 55U), the word accuracy performance is still an improvement over the static MFNS case (73.7% versus 73.3%). As a result, this is not seen to be a significant issue at this point in time.

In order to assess the effectiveness of the proposed dialog-based LIMA framework, all utterances occurring prior to the first optimization (or first optimization after calibration) for each speaker were ignored. This approach was required since the proposed technique requires 100% word accuracy in order to trigger optimization, a result which was achieved on only 3% of all utterances and mostly in the idle noise condition. This low number of optimization instances is due to the relatively low performance of the ASR system and nature of the recognition task which requires all ten digits to be recognized correctly.

These results of this final evaluation are summarized in Table 10.4. It should be noted that word accuracies in this table are better than in previous tables because this analysis removed a lot of utterances exhibiting poor ASR performance.

Almost all comparisons in Table 10.4 show that the proposed dialog-based LIMA framework for in-car ASR provides improved performance over the baseline enhancement system. Applying this framework can also recover losses in word accuracy incurred when using standard Mel-filterbank noise subtraction (e.g., in the two 35-mph noise conditions).

The results of this evaluation also prove the effectiveness of the proposed dialog-based framework when used with or without explicit calibration even though there are a very low number of optimization instances. For the case without calibration – which is the ideal operational behavior of such a framework since the user would be completely unaware of adaptation – global improvements over both baseline systems can be observed, with the best relative performance improvement over a

Table 10.4 ASR results for all LIMA frameworks

Framework	IDL	35U	35D	55U	55D
Baseline	79.1	55.8	42.1	49.8	27.6
$\alpha(k) = 1$	81.8	53.9	41.6	51.7	30.1
Proposed dialog system	82.6	55.9	42.3	53.1	31.1
Baseline	80.7	55.5	43.3	49.5	28.6
$\alpha(k) = 1$	81.4	53.3	45.3	50.0	33.6
Calibrated system (random)	82.5	55.7	46.4	52.5	33.3
Proposed dialog (random)	82.3	57.7	45.5	52.7	32.3
Baseline	80.4	57.7	44.7	53.3	28.4
$\alpha(k) = 1$	82.2	52.5	42.9	53.9	30.3
Calibrated system (IDL)	82.4	55.4	44.6	54.9	31.0
Proposed dialog (IDL)	82.9	55.9	46.0	55.5	30.9

system without enhancement being 16.7% in the idle condition. This particular result demonstrates the true potential of the framework to improve ASR accuracy, since utterances spoken during idle are most likely to trigger the optimization process. In comparison to the baseline enhancement system, the proposed framework shows relative improvements of between 1.2% and 4.4% in this mode of operation.

There are also noticeable improvements of the calibration-only LIMA framework, particularly one performing calibration during idle. In this case, the relative improvements range from 1.2% to 2.8% (excluding the marginal decrease in performance in the 55D noise condition). Given that most users will first speak to in-car dialog systems when entering their vehicle, this result verifies the potential of the proposed framework to be incorporated with a calibration session to produce further improvements in system performance.

Considering the operation of the proposed dialog-based system, there is potential for a loss of generality if a particular noise condition is consecutively optimized (as per the results in Table 10.2). The consistent improvements in Table 10.4, however, indicate that this is not an issue as regular changes in noise conditions seem to allow the optimization process to effectively track the internal noise conditions and set the enhancement parameters appropriately.

10.5 Conclusions

This chapter has reviewed likelihood-maximizing frameworks using Mel-filterbank noise subtraction for in-car speech recognition. A new LIMA framework based on a user-confirmation speech dialog system has been proposed. This framework has been evaluated against calibrated LIMA frameworks utilizing different adaptation scenarios.

Experiments have shown that with the proposed LIMA framework, minimal optimization is required for the best average recognition performance in car environments. This permits pseudo real-time operation of LIMA frameworks whilst

still providing improvements over standard speech enhancement techniques. The proposed dialog-based framework provides improved recognition performance over calibration-only systems; this effect is attributed to the ability to continually update enhancement parameters according to changes in noise conditions.

Acknowledgments Parts of the work presented here were funded through the Australian Cooperative Research Centre for Advanced Automotive Technology (AutoCRC).

References

1. Kleinschmidt T, Mason M, Wong E, Sridharan S (2009) The Australian english speech corpus for in-car speech processing. In: Proceedings of ICASSP. IEEE Computer Society, Washington, DC, pp 4177–4180
2. Seltzer ML, Raj B, Stern RM (2004) Likelihood-maximizing beamforming for robust hands-free speech recognition. IEEE Trans Speech Audio Process 12(5):489–498
3. BabaAli B, Sameti H, Safayani M (2009) Likelihood-maximizing-based multiband spectral subtraction for robust speech recognition. EURASIP J Adv Signal Process 878105:1–15
4. Shi G, Aarabi P, Jiang H (2007) Phase-based dual-microphone speech enhancement using a prior speech model. IEEE Trans Audio Speech Lang Process 15(1):109–118
5. Lockwood P, Boudy J, Blanchet M (1992) Non-linear spectral subtraction (NSS) and hidden Markov models for robust speech recognition in car noise environments. In: Proceedings of the ICASSP, San Francisco, pp 265–268
6. Wu K-G, Chen P-C (2001) Efficient speech enhancement using spectral subtraction for car hands-free applications. In: Proceedings of IEEE international conference on consumer electronics, Los Angeles, pp 220–221
7. Kleinschmidt T, Boyraz P, Bořil H, Sridharan S, Hansen JHL (2009) Assessment of speech dialog systems using multi-model cognitive load analysis and driving performance metrics. In: Proceedings of IEEE international conference on vehicular electronics & safety, Pune, pp 167–172
8. Boll S (1979) Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans Acoust Speech Signal Process 27(2):113–120
9. Nasersharif B, Akbari A (2006) A framework for MFCC feature extraction using SNR-dependent compression of enhanced Mel-filter bank energies. INTERSPEECH, 1632-Mon1A20.3
10. Kleinschmidt T (2010) Robust speech recognition using speech enhancement. PhD thesis, Queensland University of Technology
11. Lee B, Hasegawa-Johnson M, Goudeseune C, Kamdar S, Borys S, Liu M, Huang T (2004) AVICAR: audio-visual speech corpus in a car environment. In: Proceedings of INTERSPEECH, Jeju Island, pp 2489–2492
12. Kleinschmidt T, Dean D, Sridharan S, Mason M (2007) A continuous speech recognition protocol for the AVICAR database. In: Proceedings of ICSPCS, Gold Coast, pp 339–344
13. Young S, Evermann G, Gales M, Hain T, Kershaw D, Liu X, Moore G, Odell J, Ollason D, Povey D, Valtchev V, Woodland P (2006) The HTK Book. Cambridge University: Engineering Department, Cambridge