SPEECH ENHANCEMENT FOR ROBUST SPEECH RECOGNITION IN MOTORCYCLE ENVIRONMENT

IOSIF MPORAS, TODOR GANCHEV†, OTILIA KOCSIS and NIKOS FAKOTAKIS

Artificial Intelligence Group, Wire Communications Laboratory,
Dept. of Electrical and Computer Engineering, University of Patras, Rion-Patras, 26500, Greece
{*imporas, okocsis, fakotaki*}*@upatras.gr,* †*tganchev@ieee.org*

Received (15 September, 2009)
Accepted (17 January 2010)

In the present work, we investigate the performance of a number of traditional and recent speech enhancement algorithms in the adverse non-stationary conditions, which are distinctive for motorcycles on the move. The performance of these algorithms is ranked in terms of the improvement they contribute to the speech recognition accuracy, when compared to the baseline performance, i.e. without speech enhancement. The experiments on the MoveOn motorcycle speech and noise database indicated that there is no equivalence between the ranking of algorithms based on the human perception of speech quality and the speech recognition performance. The Multi-band spectral subtraction method was observed to lead to the highest speech recognition performance.

Keywords: Fast-varying noise; speech enhancement; robust speech recognition.

1. Introduction

Technological advances in the Internet protocol (IP)-telephony domain have leaded to an increased interest to provide accessibility to the large domain of web application over the phone, with personal assistant-based dialogue systems offering higher comfort to the end-users[1,2]. Higher demands with respect to efficiency, comfort and safety are imposed to those services that have transited from the controlled office/home environment to the mobile outdoors environment. On the route, driver distraction can become a significant problem, thus highly efficient human-machine interfaces and interaction are required to enable car or motorcycle drivers to interact with mobile systems and services in an easy, risk-free way.

Spoken language dialogue systems considerably improve driver's safety and user-friendliness of human-machine interfaces, due to their similarity to the conversational activity with another human, a parallel activity to which the driver is used to and it allows him to concentrate on his main activity, the driving itself. Driving quality, stress and strain situations and user acceptance when using speech and manual commands to acquire certain information on the route has previously been studied[3], and the results have shown that, with speech input, the feeling of being distracted from driving is smaller, and road safety is improved, especially in the case of complex tasks. Although natural human-to-human interaction can be considered multimodal, as speech is usually

† Corresponding Author: Todor Ganchev (tganchev@ieee.org)

combined with free hand gestures, interpretation of gestures in a multimodal setting is challenging even in a controlled environment[4]. Moreover, assessment of user requirements from multimodal interfaces in a car environment has shown that when the car is moving the system should switch to the "speech-only" interaction mode, as any other safety risks (i.e. driver distraction from the driving task by gesture input or graphical output) must be avoided[5].

The performance of speech-based interfaces, although reliable enough in controlled environments to support speaker and device independence, degrades substantially in a mobile environment, when used on the road. There are various types and sources of noise interfering with the speech signal, starting with the acoustic environment (vibrations, road/fan/wind noise, engine noise, traffic, etc.) to changes in the speaker's voice due to task stress, distributed attention, etc. In the integration of speech-based interfaces within vehicle environments, the research is conducted in two directions: (i) addition of front-end speech enhancement systems to improve the quality of the recorded signal, and (ii) training the speech models of the recognizer engine on noisy, real-life, speech databases.

Preliminary speech/noise detection using front-end speech enhancement methods for noise suppression has shown promising results and currently benefits from the suppression of interfering signals by using a microphone array, which enables both spatial and temporal measurements[6]. The advantages of multi-channel speech enhancement can be successfully applied to the car environment, while in the motorcycle environment, due to processing power limitations, research is focused to one-channel speech enhancement. After more than three decades of advances on the one-channel speech enhancement problem, four distinct families of algorithms seem to have predominated in the literature: (i) the spectral subtractive algorithms[7], (ii) the statistical model-based approaches[8,9,10], (iii) the signal subspace approaches[11,12], and (iv) the enhancement approaches based on a special type of filtering[13].

As previously mentioned, the accuracy of the recognition task is also highly improved by using suitably trained speech models for the recognizer engine, including sufficient noise scenarios from the application domain. Dedicated speech corpora have previously been designed, recorded and annotated, starting with the car environment, and emerging with the motorcycle one. A European initiative, co-funding the design and implementation of speech databases in support of the development of multilingual speech recognition applications in the car environment, started in 1998 with the SPEECHDAT-CAR project[14]. The databases developed within this project were designed to include phonetically balanced speech for the needs of training generic speech recognition systems, but also domain-specific data, needed for adapting the acoustic models of speaker-independent automatic speech recognition systems to the automotive environment. Databases for ten European languages were collected within the SPEECHDAT-CAR project, with recordings from at least 300 speakers for each language, and seven characteristic environments (low speed, high speed, with audio equipment on, etc).

The CU-Move corpus consists of five domains, including digit strings, route navigation expressions, street and location sentences, phonetically balanced sentences and a route navigation dialog in a human Wizard-of-Oz like scenario, considering 500 speakers from United States of America and a natural conversational interaction[15].

For the motorcycle environment, the SmartWeb motorbike corpus has been designed for a dialogue system dealing with open domains[16]. Recently, a domain specific (police domain) database, dealing with the extreme conditions of the motorcycle environment, has been developed in the MoveOn project[17]. In this database, the focus is the specificity of the domain, where the cognitive load of motorcyclists is quite high and the accuracy of recognition of spoken commands in the context of a template driven dialog, in the motorcycle environment, is of high priority.

In the present study, we investigate the applicability of various speech enhancement algorithms for the specifics of the motorcycle-on-the-move environment. These algorithms are ranked in terms of the improvement they bring to the speech recognition performance, when compared to the case when no speech enhancement is employed. The speech enhancement component that is in the focus of the present study is a crucial part of front-end of a spoken dialogue interaction system, which provides hands-free information support to the motorcyclists. This speech-based interface is part of a multi-modal and multi-sensor interface developed in the context of the MoveOn project. Following, a brief overview of the MoveOn system, the speech enhancement methods evaluated, and the experimental setup and results are presented.

2. System Description

The MoveOn system is a multi-modal and multi-sensor, zero-distraction interface for motorcyclists that provides the means for hands-free operation of a command and control interface, which enables information support of police officers on the move. This information support is either obtained remotely from the control centre in the police station, through a secure terrestrial trunked radio (TETRA) link, or locally through the functionality provided by a wearable computing environment developed for that purpose. This environment offers functionalities such as navigation support, accessing local user-specific data repository, storing video and audio streams for reporting and evidence collection purposes, automated plate number recognition, automated logging and diary capabilities, information recall and storage on request, visualization and alert mechanisms, communication with colleagues on the road or in flying vehicles, etc. The remote information support guarantees command, control, and guidance support as well as access to forensic and other police databases located at the central police station.

The MoveOn system is implemented as a wearable solution, which constitutes of a set of purposely-designed accessories, such as a helmet, a waist and gloves. The helmet and the waist are connected through a flexible connector located just bellow the scruff of the neck. The gloves, which incorporate push-buttons and a scroller-based haptic interface, are connected to the waist through a flexible connection near the wrist. The helmet incorporates microphones, headphones, visual feedback, a miniature camera and some supporting local-processing electronics. It has a flexible connection, incorporating universal serial bus (USB) connector to the waist, which provides the power supply, and the data and control interfaces. The waist incorporates the main processing power, storage repository, the TETRA communication equipment and the power capacity of the wearable system, but also a number of sensors, a liquid crystal display (LCD), and some vibration feedback actuators. Among the sensors deployed on the waist are acceleration and inclination sensors, and a global positioning system (GPS) device, which provide the means for the context awareness of the system. Auxiliary microphone and headphone are integrated in the upper part of the waist, at the front side near the collar, for guaranteeing the spoken interaction and communication capabilities when the helmet is off.

The multimodal user interface developed for the MoveOn application consists of audio and haptic inputs, and audio, visual and vibration feedbacks to the user. Due to the specifics of the MoveOn application, involving hands-busy and eyes-busy motorcyclists, speech is the dominating interaction modality.

The spoken interface consists of multi-sensor speech acquisition equipment, speech

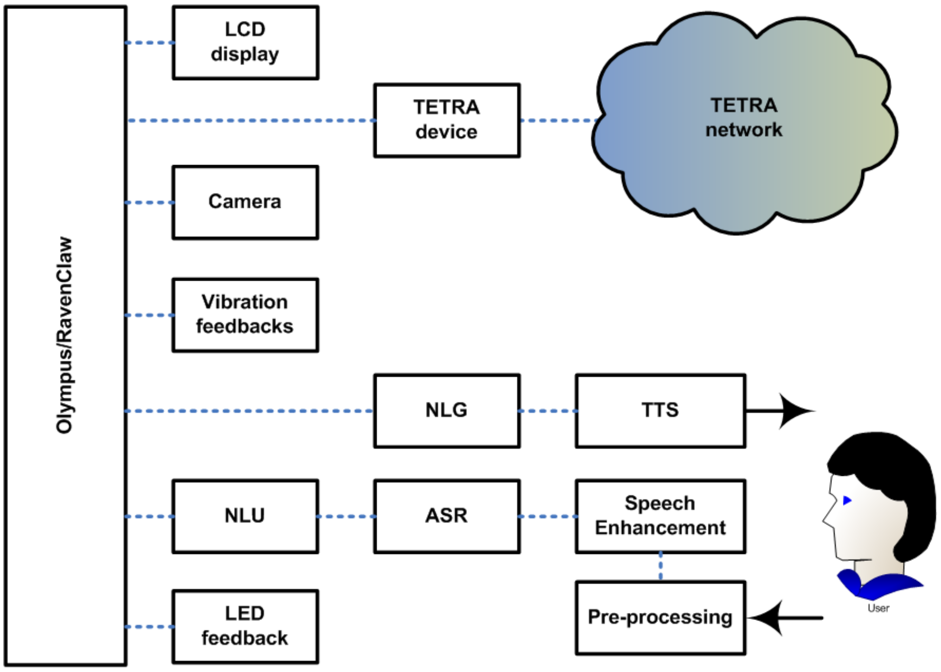

Fig. 1. Block diagram of the MoveOn multimodal interaction.

pre-processing, speech enhancement, speech recognition, and text-to-speech synthesis components etc. This interface has to provide the proper recognition and interpretation of the speech input and to deliver non-distractive, intelligible and naturally sounding feedback to the user. Achieving these objectives within the operational environment of the MoveOn application is not a trivial task, and it requires proper design and implementation of the speech front-end and the system's feedback to the user.

The spoken interface is build on a multimodal dialogue interaction framework, based on Olympus/RavenClaw[18,19], which was extended for the needs of multimodal interaction. Each component in the system is a server on itself (i.e. ASR, TTS, speech pre-processing, speech enhancement, etc are servers), communicating either directly with each other or through a central hub, which provides synchronization.

Since the noisy motorcycle environment constitutes a great challenge to the spoken dialogue interaction, a special effort is required to guarantee high speech recognition performance, as it proved to be the most crucial element for the overall success of the interaction. The robust operation of the speech recognition component depends on successful elimination of noise, while preserving the speech signal integrity. The fast-varying noise conditions and the huge number of interferences that may appear simultaneously, which is typical for motorcycle on the move, constitute the greatest challenge for the speech enhancement algorithms.

3. Speech Enhancement Methods

In the present study, we consider eight speech enhancement techniques, which are tested in the fast-varying conditions of the motorbike-on-the-move environment. In earlier work[20], these algorithms were evaluated in terms of objective assessment of the perceptual quality of de-noised speech. In the present work, the focus falls on the performance of these algorithms, evaluated in terms of the improvement of the speech recognition accuracy they add, when compared to the baseline speech recognition accuracy, i.e. obtained without speech enhancement.

3.1. *Spectral Subtraction*

The spectral subtraction (SPECSUB) algorithm[21], which is a well-known technique, is often used as a baseline against which other speech enhancement algorithms are compared. This algorithm relies on the fact that the power spectra of additive independent signals are also additive. Thus, in the case of stationary noise, in order to obtain a least squares estimate of the speech power spectrum, it suffices to subtract the mean noise power. Due to its low complexity and good efficiency, the spectral subtraction method is a standard choice for noise suppression at the pre-processing stage of speech recognition systems. Due to its well-known performance, the spectral subtraction algorithm serves here as an intuitive reference point.

3.2. *Spectral Subtraction with Noise Estimation*

The spectral subtraction with noise estimation (SPECSUB-NE)[22] tracks spectral minima in each frequency band without any distinction between speech activity and speech pause. Based on the optimally smoothed power spectral density estimate and the analysis of the statistics of spectral minima an unbiased noise estimator is implemented. Due to the last, this algorithm is more appropriate for real world conditions, and was reported to outperform the SPECSUB in non-stationary noise environments.

3.3. *Multi-Band Spectral Subtraction*

The multi-band spectral subtraction method (M-BAND)[7] is based on the SPECSUB algorithm but accounts for the fact that in real world conditions, interferences do not affect the speech signal uniformly over the entire spectrum. The M-BAND method was demonstrated to outperform the standard SPECSUB method resulting in superior speech quality and largely reduced musical noise. The results presented in [7] as well as our previous experience with the MoveOn data[20] suggested that this method may perform well in terms of improvement of the speech recognition accuracy.

3.4. *Speech Enhancement Using a Minimum Mean Square Error Log-Spectral Amplitude Estimator*

The speech enhancement using a minimum mean square error log-spectral amplitude estimator[8], which we refer to as (Log-MMSE), relies on a short-time spectral amplitude estimator for speech signals, which minimizes the mean-square error of the log-spectra. This speech enhancement method belongs to the category of statistical model-based algorithms. In previous work[20], it was observed to offer very good performance on the MoveOn data, and therefore it is a strong candidate for achieving excellent improvement of the speech recognition accuracy.

3.5. *Speech Enhancement Based on Perceptually Motivated Bayesian Estimators of the Speech Magnitude Spectrum*

The speech enhancement based on perceptually motivated Bayesian estimators (STSA-WCOSH) of the speech magnitude spectrum[9] utilizes Bayesian estimators of the short-time spectral magnitude of speech based on perceptually motivated cost functions. It was demonstrated that the estimators, which implicitly take into account auditory masking effect, perform better in terms of having less residual noise and better speech quality, when compared to the Log-MMSE method. We selected this method due to its relatively good performance[20], but also because we were interested to investigate if this advantage, when compared to Log-MMSE, will contribute for better speech recognition performance.

3.6. *Subspace Algorithm with Embedded Pre-whitening*

The subspace algorithm with embedded pre-whitening (KLT)[10] is based on the simultaneous diagonalization of the clean speech and noise covariance matrices. Objective and subjective evaluations suggest that this algorithm offers advantage when the interference is speech-shaped or multi-talker babble noise.

3.7. *Perceptually-Motivated Subspace Algorithm*

The perceptually-motivated subspace algorithm (PKLT)[11] incorporates a human hearing model in the suppression filter in order to reduce the residual noise. From a perceptual perspective, the perceptually based eigen-filter introduced here yields a better shaping of the residual noise. This method was reported to outperform the KLT method.

3.8. *Wiener Algorithm Based on Wavelet Thresholding Multi-Taper Spectra*

The Wiener algorithm based on wavelet thresholding (WIENER-WT) multi-taper spectra[12] uses low-variance spectral estimators based on wavelet thresholding the multi-taper spectra. Reported listening tests had shown that this method suppressed the musical noise and yielded better speech quality than the KLT, PKLT and Log-MMSE algorithms. Based on these reports, we expect that the WIENER-WT algorithm will be a strong candidate for the best performance.

4. Experimental Setup

The speech interaction interface described in Section 2 was tested with each of the speech enhancement techniques outlined in Section 3. Different environmental conditions and configuration settings of the speech recognition engine were considered for the evaluation experiments. In the following, we describe the speech data, the speech recognition engine and the experimental protocol utilized in the present evaluation.

4.1. *The MoveOn Speech and Noise Database*

In the MoveOn project, a dedicated domain-specific speech database was recorded in the motorcycle environment for the purpose of research and technology development[17]. In detail, thirty professional motorcyclists, members of the operational police force of UK, participated in the recordings of the database. Each participant was asked to repeat a number of domain-specific commands and expressions, or to provide a spontaneous answer to questions related to time, current location, speed, etc. The prompt sheets, each one containing 302 prompts, were provided to the participants via earplug as sequences of pre-recorded audio prompts, while they were performing patrolling activates with their motorcycles.

The amount of collected audio data consists of approximately forty hours of recordings, distributed in forty recording sessions. During the recording of each session, different motorbikes and helmets were used, while the trace of the road differed among them. In detail, each session included in-city driving, highway, tunnels, suburbs, etc. Furthermore, ten sessions with the same hardware but in office environment (indoors) were recorded.

Four audio channels, recorded simultaneously, constitute each recording session: (i) two from omni-directional microphones (AKG C 417''') placed within the helmet -- 10 cm one from another -- at the two sides of the mouth; (ii) one channel from a throat microphone (Alan AE 38); and finally (iii) one channel that mixes the first of the in-helmet microphones with the audio prompts that were played to the speaker. This fourth audio channel served as reference for the synchronization of the channels during the annotation phase. The language of all recordings is British English spoken by native speakers.

The database was recorded at 44.1 kHz, with resolution 16 bits per sample. Later on, all recordings were down-sampled to 8 kHz for the needs of the present application.

The recordings were annotated in a multi-tier scheme. The annotations include different tiers for speech transcriptions, emotional tags, and various noise tags, such as: background noise, transient interferences (air-wind noise, engine noise, other noise, and sound events). The transient noises were labelled for their position and their estimated magnitude. One additional tier indicates when the helmet visor is open or closed, since this condition significantly affects both the amount and the shaping of noise.

4.2. *Speech Recognition*

In the present evaluation, we employed the Julius[23] speech recognition engine. The decoder of the recognition engine was utilized together with two different acoustic models, a general-purpose acoustic model and an adapted acoustic model. In addition, an application-dependent language model was used in all cases.

The general purpose acoustic model was trained from telephone speech recordings of the British SpeechDat(II) database[24], with the exploitation of the HTK toolkit[25]. The general purpose acoustic model consists of three-state left-to-right HMMs, without skipping transitions, one for each phone of the British SpeechDat(II) phone set. Each state is modelled by a mixture of eight continuous Gaussian distributions. The state distributions were trained from parametric speech vectors, taken out from speech waveforms after pre-processing and feature extraction.

The pre-processing of the speech signals, sampled at 8 kHz, consisted of frame blocking with length and step 25 and 10 milliseconds respectively, and pre-emphasis with coefficient equal to 0.97. The speech parameterization consisted in the computation of the first twelve Mel frequency cepstral coefficients[26], computed through a filter-bank of 26 channels[25], and the energy of each frame. The first and second derivatives of the 13 static speech parameters were appended to the initial vector, resulting to a parametric vector of dimensionality equal to 39. All HMMs were trained with the Baum-Welch algorithm[27], with convergence ratio equal to 0.001.

The adapted acoustic models, one for each speech enhancement method, were obtained by means of maximum a posteriori[28] (MAP) adaptation of the general-purpose British English acoustic model that was described above. The adaptation was performed with the exploitation of the corresponding enhanced speech recordings of the MoveOn database.

The language model was built with the utilization of the CMU Cambridge Statistical Language Modelling (SLM) Toolkit[29]. Specifically, we used the transcriptions of the responses of the MoveOn end-users to the system[30] to build bi-gram and tri-gram word models. Words included in the application dictionary but not in the list of n-grams were assigned as out-of-vocabulary words.

5. Experimental Results

The performance of different enhancement methods, implemented as in [31], was examined by evaluating their effect on the speech recognition accuracy. During the evaluation, we considered different environmental conditions as well as different experimental setups.

In detail, we examined the speech recognition performance under (i) indoors and (ii) outdoors recording conditions, after applying the speech enhancement methods, described in Section 3. The performance of each enhancement method in the indoors recording condition was used as a reference, while the outdoors condition is the environment of interest. In contrast to previous work[20], were the performance of enhancement algorithms was investigated on the basis of objective tests on the enhanced signals, here we examine directly the operational functionality of the ASR component by measuring the speech recognition performance. Specifically, the word recognition rates (WRRs) obtained by the speech recognition process after applying each speech enhancement method was measured. The WRR is an indicator of the amount of words that were deleted, substituted or inserted, comparing to the real word sequence that was uttered.

The speech recognition accuracy under the indoor recording conditions was examined with the exploitation of the general-purpose acoustic model, while for the case of the outdoor recordings the accuracy was examined considering both a general acoustic model and adapted acoustic. In terms of these performance measures, we assess the practical worth of each algorithm and its usefulness with respect to overall system performance. These results are compared against the quality measures obtained in earlier work[20].

5.1. *Speech Recognition using General Acoustic Model*

As a first step, we evaluated the speech recognition performance for each of the speech enhancement methods described in Section 3. The speech recognizer was tested with both bi-gram and tri-gram language models, using the general purpose acoustic model described in Section 4. The experimental results for the indoor recording conditions are shown in Table 1.

Table 1. Speech recognition performance (WRR in percentages) for various speech enhancement techniques for the indoors recordings, using general acoustic model.

Enhancement Techniques	2-gram LM	3-gram LM
Log-MMSE	76.75	70.29
No Enhancement	76.71	70.25
M-BAND	75.61	71.27
SPECSUB-NE	74.25	68.53
PKLT	74.10	67.85
WIENER-WT	73.48	67.15
KLT	69.69	63.95
STSA-WCOSH	66.16	59.10
SPECSUB	50.89	40.35

As can be seen in Table 1, the best performing method for the case of indoor recordings was the Log-MMSE together with the non-enhanced speech inputs (indicated as “No Enhancement”). In all remaining methods, the speech recognition performance was decreased. This reduction is owed to the distortion that these speech enhancement methods introduce into the natural/unprocessed speech signal. Obviously, under the indoor environmental conditions, where generally noise-free speech is captured by the microphones, the speech recognizer performs better without any speech enhancement pre-processing.

As Table 1 presents, the speech recognition performance for the bi-gram language model was better than the one for the tri-gram language model. This result can be explained by the limited amount of the application data that were available for the training of the language models. As the experimental results indicate, the data were sufficient for training the bi-gram model but not enough for the robust estimation of all tri-gram word probabilities.

In Table 2, we present the speech recognition performance for all the examined speech enhancement methods, for the case of outdoor recording conditions, i.e. the motorcycle on the move. The results are presented in terms of WRR, for both the bi-gram and tri-gram language models.

In contrast to the indoor environmental conditions, the application of speech enhancement techniques in the noisy outdoor conditions (motorcycles on the move) improved the speech recognition performance. All the evaluated speech enhancement methods demonstrated superior performance comparing to the baseline performance, i.e. without speech enhancement (indicated as “No Enhancement” in Table 2). This is owed to the fact that although the speech enhancement techniques introduce a distortion to the original speech signal (which is noisy in this case), their effect results to a processed signal that is acoustically more close to the phonetic patterns of the acoustic model than the non-enhanced one.

As shown in Table 2, the multi-band speech enhancement technique, M-BAND, outperformed all other methods evaluated here, achieving accuracy of approximately 55% for bi-gram language model. The perceptually motivated Bayesian estimator

Table 2. Performance (WRR in percentages) for various speech enhancement techniques for the outdoors recordings, using general acoustic model.

Enhancement Techniques	2-gram LM	3-gram LM
M-BAND	55.16	49.65
STSA-WCOSH	49.56	41.73
SPECSUB-NE	46.34	30.87
PKLT	39.76	29.40
Log-MMSE	39.22	27.83
KLT	39.20	27.84
WIENER-WT	35.64	29.06
SPECSUB	26.95	14.84
No Enhancement	23.77	14.29

enhancement technique, STSA-WCOSH, achieved the second-best performance, following the leading M-BAND technique by approximately 6% lower, in terms of WRR. Similarly to the indoors case, the bi-gram language model provided more accurate recognition results also in the outdoor environment.

These results reveal, that the ranking of speech enhancement algorithms based on the human perception of speech quality (please refer to [20]) differs from the ranking in terms of speech recognition performance. Specifically, the M-BAND algorithm, which was among the top-4 performers in terms of perceptual quality, is the best performing algorithm in terms of WRR.

5.2. *Speech Recognition using Adapted Acoustic Models*

As a second step, we evaluated the performance of the speech enhancement algorithms of interest, using adapted acoustic models. As described in Section 4, for each speech enhancement algorithm one adapted acoustic model was employed. The performance of the speech recognition engine using the adapted acoustic models was tested on the motorcycle environment, i.e. the outdoor recordings. The results, in terms of percentages of WRRs, are presented in Table 3, where the “No Enhancement” technique corresponds to MAP adaptation of the general acoustic model on the outdoor environment without the use of any speech enhancement algorithm.

As can be seen in Table 3, the STSA-WCOSH enhancement method achieved the highest speech recognition performance, when compared to the other evaluated enhancement algorithms. For the case of speech decoding with bi-gram language model, the STSA-WCOSH method improved the WRR by 3.32% when compared to the baseline performance (i.e. the “No Enhancement” case). The STSA-WCOSH method is followed by the M-BAND, which achieved slightly lower speech recognition performance than the first one, but still improved the baseline performance by 2.35%. The methods Log-MMSE, SPECSUB-NE and SPECSUB achieved lower speech recognition performance but still offer some improvement when compared to the baseline results. In contrast to these methods, the PKLT, KLT and WIENER-WT methods offered significantly inferior

Table 3. Performance (WRR in percentages) for various speech enhancement techniques for the outdoors recordings, using adapted acoustic models.

Enhancement Techniques	2-gram LM	3-gram LM
M-BAND	85.70	75.12
STSA-WCOSH	86.67	76.10
SPECSUB-NE	84.85	74.38
PKLT	77.93	65.53
Log-MMSE	84.86	74.23
KLT	82.41	70.34
WIENER-WT	81.29	70.29
SPECSUB	84.83	75.45
No Enhancement	83.35	71.62

speech recognition performance, when compared to the use of adapted acoustic model without speech enhancement.

As Table 3 presents, the performance of the speech recognition decoder was higher for the case of bi-gram language model, when compared to the one for tri-gram model. This is in agreement with the experimental results obtained when using general-purpose acoustic model, reported above.

In order to investigate the statistical significance among the different WRRs, reported in the second column of Table 3, we performed the Wilcoxon signed-rank test[32]. The highlighted cells correspond to statistically similar recognition results. As the Wilcoxon test indicated, the word recognition rates reported in the second column of Table 3 are not statistically different among the cases where the speech enhancement is performed by the Log-MMSE, SPECSUB-NE and SPECSUB methods. The speech recognition results for the best performing STSA-WCOSH method is statistically different from the remaining methods.

Similarly to the case with general-purpose acoustic model, the ranking of speech enhancement algorithms in terms of speech recognition performance, presented in Table 3, is partially aligned with the ranking in terms of speech quality evaluation reported in previous work[20]. The difference in these rankings is an indication of the dissimilarity between the human perception of speech quality and the automatic speech recognition process. Since here we are interested in the functionality of the dialogue system, the ranking of the enhancement methods in terms of speech recognition performance is a more appropriate criterion for configuring the speech front-end, when compared to the subjective speech perception tests, or the objective measures previously used[21].

The experimental results indicated the importance of the use of acoustic models, adapted to the environmental conditions of the motorcycle on the move, as well as of the speech enhancement technique that is utilized. Indeed, the use of adapted acoustic models improved the speech recognition performance by approximately 30%, when compared to the best performing enhancement method, using the general-purpose acoustic model. Furthermore, the use of an acoustic model, adapted to the environmental conditions, without the use of any enhancement algorithm offered approximately 28% improvement of the WRR, comparing to the non-adapted acoustic model.

6. Conclusion

Aiming at successful human-machine interaction in the motorcycle environment, we evaluated the recognition performance of a purposely-built speech front-end. Various speech enhancement techniques were assessed in an attempt to find the most appropriate pre-processing of the speech signal in the fast-varying noisy conditions. The experimental results demonstrated severe degradation of the speech recognition performance in the conditions of the motorcycle environment, compared to the clean-speech recordings conducted with the same hardware setup. The multi-band spectral subtraction method demonstrated the best performance among the eight evaluated techniques, when measured in terms of improvement of the speech recognition rate using a general-purpose acoustic model. In the case of using adapted acoustic models, the best performance was achieved by the speech enhancement method based on perceptually motivated Bayesian estimators (STSA-WCOSH) of the speech magnitude spectrum. Finally, the use of acoustic models that are adapted to the environmental conditions of the motorcycle on the move as well as the selection of an appropriate speech enhancement technique, proved to be essential for the successful interaction between the user and the dialogue system.

Acknowledgments

This work was partially supported by the MoveOn project (IST-2005-034753) (http://www.m0ve0n.net/), which is co-funded by the FP6 of the European Commission.

References

1. S. Möller, J. Krebber, A. Raake, P. Smeele, M. Rajman, M. Melichar, V. Pallotta, G. Tsakou, B. Kladis, A. Vavos, J. Hoonhout, D. Schuchardt, N. Fakotakis, T. Ganchev and I. Potamitis, INSPIRE: Evaluation of a Smart-Home System for Infotainment Management and Device Control", in *Proc. of the International Conference on Language Resources and Evaluation* (Lisbon, Portugal, 2004), vol.5, pp.1603-1606.
2. E.C. Paraiso and J.-P.A. Barthes, An intelligent speech interface for personal assistants in R&D projects, *J. Expert Systems with Applications*, **31**(4) (2006) 673-683.
3. U. Gartner, W. Konig and T. Wittig, Evaluation of Manual vs. Speech input when using a driver information system in real traffic, *Driving Assessment 2001: The First International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, CO, (2001) pp.7-13.
4. S. Kettebekov and R. Sharma, Understanding Gestures in Multimodal Human Computer Interaction, *J. International Journal on Artificial Intelligence Tools*, **9**(2) (2000) 205-223.
5. A. Berton, D. Buhler and W. Minker, SmartKom-Mobile Car: User Interaction with Mobile Services in a Car Environment, *SmartKom: Foundations of Multimodal Dialogue Systems*, ed. Wolfgang Wahlster (Springer, 2006) pp.523-537.

6. E. Visser, M. Otsuka and T.W. Lee, A spatio-temporal speech enhancement scheme for robust speech recognition in noisy environments, *J. Speech Communication*, **41**(2003) 393-407.
7. S. Kamath and P. Loizou, A multi-band spectral subtraction method for enhancing speech corrupted by colored noise, in *Proc. of ICASSP'02*, 2002.
8. Y. Ephraim and D. Malah, Speech enhancement using a minimum mean square error log-spectral amplitude estimator, *J. IEEE Transactions on Acoustics, Speech, Signal Processing*, **33**(1985) 443-445.
9. P. Loizou, Speech enhancement based on perceptually motivated Bayesian estimators of the speech magnitude spectrum, *J. IEEE Transactions on Speech and Audio Processing*, **13**(5) (2005) 857-869.
10. Y. Hu and P. Loizou, A generalized subspace approach for enhancing speech corrupted by coloured noise, *J. IEEE Trans. on Speech and Audio Processing*, **11**(2003) 334-341.
11. F. Jabloun and B. Champagne, Incorporating the human hearing properties in the signal subspace approach for speech enhancement, *J. IEEE Transactions on Speech and Audio Processing,* **11**(6) (2003) 700-708.
12. Y. Hu and P. Loizou, Speech enhancement based on wavelet thresholding the multitaper spectrum, *J. IEEE Transactions on Speech and Audio Processing*, **12**(1) (2004) 59-67.
13. S. Gannot, D. Burshtein and E. Weinstein, Iterative and sequential Kalman filter-based speech enhancement algorithms, *J. IEEE Transactions on Speech and Audio Processing*, **6**(4) (1998) 373-385.
14. A. Moreno, B. Linderberg, C. Draxler, G. Richard, K. Choukri, S. Euler and J. Allen, SPEECHDAT-CAR. A Large Speech Database for Automotive Environments, in *Proc. of 2nd International Conference on Language Resources and Evaluation* (*LREC 2000*), Athens, 2000.
15. J.H.L. Hansen, X. Zhang, M. Akbacak, U. Yapanel, B. Pellom and W. Ward, CU-Move: Advances in in-vehicle speech systems for route navigation, in *Proc. of the IEEE Workshop on DSP in Mobile and Vehicular Systems*, 2003, Nagoya, Japan, pp.19-45.
16. M. Kaiser, H. Mogele and F. Shiel, Bikers Accessing the Web: The SmartWeb motorbike corpus, in *Proc. of LREC'2006,* 2006.
17. T. Winkler, T. Kostoulas, R. Adderley, C. Bonkowski, T. Ganchev, J. Kohler and N. Fakotakis, The MoveOn Motorcycle Speech Corpus, in *Proc. of LREC'2008*, 2008.
18. D. Bohus and A.I. Rudnicky, RavenClaw: Dialog Management Using Hierarchical Task Decomposition and an Expectation Agenda, in *Proc. of the European Conference on Speech Communication and Technology* (*EUROSPEECH 2003*), pp.597-600.
19. D. Bohus, A. Raux, T.K. Harris, M. Eskenazi and A.I. Rudnicky, Olympus: an open-source framework for conversational spoken language interface research, *Bridging the Gap: Academic and Industrial Research in Dialog Technology workshop at HLT/NAACL 2007*, (2007).
20. S. Ntalampiras, T. Ganchev, I. Potamitis and N. Fakotakis, Objective comparison of speech enhancement algorithms under real world conditions, in *Proc. of the 1st International Conference on Pervasive Technologies Related to Assistive Environments* (*PETRA-2008*) (Athens, Greece, July 16-18, 2008). F. Makedon, L. Baillie, G. Pantziou, and I. Maglogiannis, Eds. PETRA-2008, vol. 282. ACM, New York, NY, 1-5. Available on-line at DOI= http://doi.acm.org/10.1145/1389586.1389627.
21. M. Berouti, R. Schwartz and J. Makhoul, Enhancement of speech corrupted by acoustic noise, in *Proc. of the IEEE ICASSP'79*, (1979) pp.208-211.

22. R. Martin, Noise power spectral density estimation based on optimal smoothing and minimum statistics, *J. IEEE Transactions on Speech and Audio Processing*, **9**(5) (2001) 504-512.

23. A. Lee, T. Kawahara and K. Shikano, Julius -- an open source real-time large vocabulary recognition engine, in *Proc. of the European Conference on Speech Communication and Technology* (*EUROSPEECH 2001*) pp.1691-1694.

24. H. Hoge, C. Draxler, H. Van den Heuvel, F.T. Johansen, E. Sanders and H.S. Tropf, SpeechDat Multilingual Speech Databases for Teleservices: Across the Finish Line, in *Proc. of the 6th European Conference on Speech Communication and Technology* (*EUROSPEECH 1999*), pp.2699-2702.

25. S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, *The HTK Book* (*for HTK Version 3.3*) (Cambridge University, 2005).

26. S.B. Davis and P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, *J. IEEE Transactions on Acoustics, Speech and Signal Processing*, **28**(4) (1980) 357-366.

27. L.E. Baum, T. Petrie, G. Soules and N. Weiss, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *J. Annals of Mathematical Statistics*, **41**(1) (1970) 164-171.

28. J.L. Gauvain and C. Lee, Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains, *J. IEEE Transactions on SAP*, **2**(1994) 291-299.

29. P.R. Clarkson and R. Rosenfeld, Statistical Language Modeling Using the CMU-Cambridge Toolkit, in *Proc. of the 5th European Conference on Speech Communication and Technology* (EUROSPEECH 1997), pp.2707-2710.

30. T. Winkler, T. Ganchev, T. Kostoulas, I. Mporas, A. Lazaridis, S. Ntalampiras, A. Badii, R. Adderley and C. Bonkowski, Report on Audio databases, Noise processing environment, ASR and TTS modules, *MoveOn Deliverable D.5* (2007).

31. P. Loizou, *Speech Enhancement: Theory and Practice* (CRC Press, 2007).

32. F. Wilcoxon, Individual comparisons by ranking methods, *J. Biometrics*, **1**(1945) 80-83.