1. Performance Evaluation of a Speech Interface for Motorcycle Environment

Iosif Mporas, Todor Ganchev, Otilia Kocsis, Nikos Fakotakis

Artificial Intelligence Group, Wire Communications Laboratory,

Dept. of Electrical and Computer Engineering, University of Patras, Rion 26500, Greece

{imporas, , tganchev, okocsis, fakotakis}@upatras.gr

Abstract: 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 motorcycle on the move. The performance of these algorithms is ranked in terms of the improvement they contribute to the speech recognition rate, when compared to the baseline result, i.e. without speech enhancement. The experimentations on the MoveOn motorcycle speech and noise database suggested 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.

1.1 Introduction

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 concentrate on the 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 [1], 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. 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 [2].

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.

In this study, the front-end speech enhancement system for a motorcycle on the move environment is investigated. The speech-based interface, as presented in this study, is part of a multi-modal and multi-sensor interface developed in the context of the MoveOn project. The performance of various speech enhancement algorithms in the non-stationary conditions of motorcycle on the move is assessed. Performance of assessed algorithms is ranked in terms of the improvement they contribute to the speech recognition rate, when compared to the baseline results (i.e. without speech enhancement). Following, a short overview of the MoveOn system, the enhancement methods evaluated, and the experimental setup and results are presented.

1.2 System Description

The MoveOn project aims at the creation of a multi-modal and multi-sensor, zero-distraction interface for motorcyclists. This interface provides the means for hands-free operation of a command and control system that enables for information support of police officers on the move. The MoveOn information support system is a wearable solution, which constitutes of a purposely designed helmet, waist and gloves. The helmet incorporates microphones, headphones, visual feedback, a miniature camera and some supporting local-processing electronics. It has a USB connection to the waist that provides the power supply and the data and control interfaces. The waist incorporates the main processing power, storage repository, TETRA communication equipment and power capacity of the wearable system, but also a number of sensors, an LCD display, and some vibration feedback actuators. Among the sensors deployed on the waist are acceleration and inclination sensors, and a 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 pre-processing, speech enhancement, speech recognition, and text-to-speech synthesis components, which are integrated into the multimodal dialogue interaction framework based on Olympus/RavenClaw [3, 4], but extended for the needs of multimodal interaction. Each component in the system is a server on itself, i.e. ASR, TTS, speech preprocessing, 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 interaction.

1.3 Speech Enhancement Methods

We consider eight speech enhancement techniques, which will be examined in the non-stationary motorcycle environment conditions:

- The spectral subtraction (SPECSUB) algorithm [5], which is a well-known technique will serve as a reference point. It 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.
- Spectral subtraction with noise estimation (SPECSUB-NE) [6]. This method 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 outperforms the SPECSUB in non-stationary environments.
- Multi-band spectral subtraction method (M-BAND) [7]. It 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.
- Speech enhancement using a minimum mean square error log-spectral amplitude estimator [8], which we refer to as (Log-MMSE). This method relies on a short-time spectral amplitude estimator for speech signals, which minimizes the mean-square error of the log-spectra.

- Speech enhancement based on perceptually motivated Bayesian estimators (STSA-WCOSH) of the speech magnitude spectrum [9]. This algorithm 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.
- Subspace algorithm with embedded pre-whitening (KLT) [10]. It 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.
- Perceptually-motivated subspace algorithm (PKLT) [11]. It incorporates a human hearing model in the suppression filter in order to reduce the residual noise. From a perceptual perspective, the perceptually based eigenfilter introduced here yields a better shaping of the residual noise. This method was reported to outperform the KLT method.
- Wiener algorithm based on wavelet thresholding (WIENER-WT) multi-taper spectra [12]. It uses a low-variance spectral estimators based on wavelet thresholding the multitaper spectra. Listening tests reportedly had shown that this method suppresses the musical noise and yielded better speech quality than the KLT, PKLT and Log-MMSE algorithms.

1.4 Experiments and Results

The speech front-end described in Section 1.2 was tested with each of the speech enhancement techniques outlined in Section 1.3. Different environmental conditions and configuration settings of the speech recognition engine were evaluated. In the following, we describe the speech data, the speech recognition engine and the experimental protocol utilized in the present evaluation. Finally, we provide the experimental results.

The evaluation of the front-end was carried out on the speech and noise database, created during the MoveOn project [13]. The database consists of approximately 40 hours of annotated recordings, most of which were recorded in three audio channels fed by different sensors, plus one channel for the audio prompts. Thirty professional motorcyclists, members of the operational police force of UK, were recorded when riding their motorcycles. 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. Motorcycles and helmets from various vendors were used, and the trace of road differed among sessions. The database includes outdoor recordings (city driving, highway, tunnels, suburbs, etc) as well as indoor (studio) recordings with the same hardware.

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

The Julius [14] speech recognition engine was employed for the present evaluation. The decoder of the recognition engine utilizes a general purpose acoustic model and an application-dependent language model. The acoustic model was built from telephone speech recordings of the British SpeechDat(II) database [15], by means of the HTK toolkit [16]. It 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 twelve Mel frequency cepstral coefficients [17], computed through a filter-bank of 26 channels, and the energy of each frame. The speech feature vector was of dimensionality equal to 39, since the first and second derivatives were appended to the static parameters. All HMMs were trained through the Baum-Welch algorithm [18], with convergence ratio equal to 0.001.

The language models were built by utilizing the CMU Cambridge Statistical Language Modeling (SLM) Toolkit [19]. Specifically, we used the transcriptions of the responses of the MoveOn end-user to the system [20] 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.

The performance of different enhancement methods, implemented as in [22], was assessed by evaluating their effect on the speech recognition results. Two different experimental setups were considered: (i) indoors and (ii) outdoors conditions. The performance of each enhancement method in the indoors condition was used as a reference, while the outdoors condition is the environment of interest. In contrast to previous work [21], 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 system by measuring the speech recognition performance. Specifically, the percentage of correctly recognized words (*CRW*) and the word recognition rates (*WRRs*) obtained in the speech recognition process after applying each enhancement method were measured. The *CRW* indicates the ability of the front-end to recognize the uttered message from the end-user, while the *WRR* points out the insertion of non uttered words, together with the word deletions and substitutions that the *CRW* measures. 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 [21].

We evaluated the speech recognition performance for each speech enhancement method in the indoors and outdoors conditions, with bi-gram and tri-gram language models. Table 1 presents the performance for the indoor experiments, in

Table 1. Performance (*WRR* and *CRW* in percentages) for various speech enhancement techniques for the indoors recordings.

Enhancement Techniques	2-gram LM		3-gram LM	
	WRR	**CRW**	**WRR**	**CRW**
Log-MMSE	76.75	81.41	70.29	81.36
No Enhancement	76.71	81.41	70.25	81.30
M-BAND	75.61	79.87	71.27	80.19
SPECSUB-NE	74.25	81.35	68.53	70.80
PKLT	74.10	80.07	67.85	79.88
WIENER-WT	73.48	80.31	67.15	80.24
KLT	69.69	78.32	63.95	78.09
STSA-WCOSH	66.16	77.30	59.10	77.11
SPECSUB	50.89	77.04	40.35	77.04

terms of *WRR* and *CRW* in percentages.

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. All remaining methods decreased the speech recognition performance. This is owed to the distortion that these speech enhancement methods introduce into the clean speech signal. Obviously, indoors, i.e. on noise-free speech, the general purpose acoustic model performs better without 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 is owed to the limited amount of data that were available for training the language models. Obviously the data were sufficient for training the bi-gram model but not enough for the tri-gram model.

In Table 2 we present the speech recognition performance in percentages for the outdoors scenario, in terms of *WRR* and *CRW*, for both the bi-gram and tri-gram language models.

In contrast to the indoors scenario, the speech enhancement in the noisy out-

Table 2. Performance (*WRR* and *CRW* in percentages) for various speech enhancement techniques for the outdoors recordings.

Enhancement Techniques	2-gram LM		3-gram LM	
	WRR	**CRW**	**WRR**	**CRW**
M-BAND	55.16	69.13	49.65	69.63
STSA-WCOSH	49.56	66.00	41.73	65.82
SPECSUB-NE	46.34	67.22	30.87	68.09
PKLT	39.76	58.11	29.40	58.48
Log-MMSE	39.22	64.17	27.83	64.90
KLT	39.20	64.16	27.84	64.92
WIENER-WT	35.64	54.07	29.06	54.59
SPECSUB	26.95	57.49	14.84	57.23
No Enhancement	23.77	54.95	14.29	55.17

doors scenario (motorcycles on the move) improved the speech recognition performance. Specifically, all speech enhancement methods demonstrated superior performance, when compared to the baseline result, i.e. without speech enhancement. As Table 2 presents, the multi-band speech enhancement technique, M-BAND, outperformed all other methods evaluated here. Similarly to the indoors case, the bi-gram language model provided more accurate recognition results. These results reveal, that the ranking of speech enhancement algorithms based on the human perception of speech quality (please refer to [21]) 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 *CWR* and *WRR*. Moreover, although the spectral subtraction with noise estimation algorithm, SPECSUB-NE, didn't perform well in the perceptual speech quality evaluation, here it has the second best performance in terms of *CRW*.

1.5 Conclusions

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. The experimental results showed 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. Finally, 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 supported by the MoveOn project (IST-2005-034753), which is partially funded by the European Commission.

References

[1] Gartner, U., Konig, W., Wittig, T. (2001). 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 Vechicle Design, pp. 7-13, CO.

[2] Berton, A., Buhler, D., Minker, W. (2006). SmartKom-Mobile Car: User Interaction with Mobile Services in a Car Environment. In SmartKom: Foundations of Multimodal Dialogue Systems, Wolfgang Wahlster (Ed.). pp. 523-537, Springer.

[3] Bohus, D., Rudnicky, A.I. (2003). RavenClaw: Dialog Management Using Hierarchical Task Decomposition and an Expectation Agenda. Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH):597-600.
[4] Bohus, D., Raux, A., Harris, T.K., Eskenazi, M., Rudnicky, A.I. (2007). 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.
[5] Berouti, M., Schwartz, R., Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise. In Proceedings of the IEEE ICASSP'79:208-211.
[6] Martin, R. (2001). Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Transactions on Speech and Audio Processing 9(5):504-512.
[7] Kamath, S., Loizou, P. (2002). A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. Proceedings of ICASSP'02.
[8] Ephraim, Y., Malah, D. (1985). Speech enhancement using a minimum mean square error log-spectral amplitude estimator. IEEE Transactions on Acoustics, Speech, Signal Processing 33:443-445.
[9] Loizou, P. (2005). Speech enhancement based on perceptually motivated Bayesian estimators of the speech magnitude spectrum. IEEE Transactions on Speech and Audio Processing 13(5):857-869.
[10] Hu ,Y., Loizou, P. (2003). A generalized subspace approach for enhancing speech corrupted by coloured noise. IEEE Trans. on Speech and Audio Processing 11:334-341.
[11] Jabloun, F., Champagne, B. (2003). Incorporating the human hearing properties in the signal subspace approach for speech enhancement. IEEE Transactions on Speech and Audio Processing 11(6):700-708.
[12] Hu, Y., Loizou, P. (2004). Speech enhancement based on wavelet thresholding the multitaper spectrum. IEEE Transactions on Speech and Audio Processing 12(1):59-67.
[13] Winkler, T., Kostoulas, T., Adderley, R., Bonkowski, C., Ganchev, T., Kohler, J., Fakotakis N. (2008). The MoveOn Motorcycle Speech Corpus. Proceedings of LREC'2008.
[14] Lee, A., Kawahara, T., Shikano, K. (2001). Julius -- an open source real-time large vocabulary recognition engine. Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH):1691-1694.
[15] Hoge, H., Draxler, C., Van den Heuvel, H., Johansen, F.T., Sanders, E., Tropf, H.S. (1999). SpeechDat Multilingual Speech Databases for Teleservices: Across the Finish Line. Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH):2699-2702.
[16] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P. (2005). The HTK Book (for HTK Version 3.3). Cambridge University.
[17] Davis, S.B., Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech and Signal Processing 28(4):357-366.
[18] Baum, L.E., Petrie, T., Soules, G., Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Annals of Mathematical Statistics 41(1):164–171.
[19] Clarkson, P.R., Rosenfeld, R. (1997). Statistical Language Modeling Using the CMU-Cambridge Toolkit. Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH): 2707-2710.
[20] Winkler, T., Ganchev, T., Kostoulas ,T., Mporas, I., Lazaridis, A., Ntalampiras, S., Badii, A., Adderley, R., Bonkowski, C. (2007). MoveOn Deliverable D.5: Report on Audio databases, Noise processing environment, ASR and TTS modules.
[21] Ntalampiras, S., Ganchev, T., Potamitis, I., Fakotakis, N. (2008). Objective comparison of speech enhancement algorithms under real world conditions. Proceedings of the PETRA 2008:34.
[22] Loizou P. (2007). Speech Enhancement: Theory and Practice, CRC Press, 2007.