Robust Speech Interaction in Motorcycle Environment

Iosif Mporas[1], Otilia Kocsis, Todor Ganchev and Nikos Fakotakis

Artificial Intelligence Group, Wire Communications Laboratory,

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

Abstract

Aiming at robust spoken dialogue interaction in motorcycle environment, we investigate various configurations for a speech front-end, which consists of speech pre-processing, speech enhancement and speech recognition components. These components are implemented as agents in the Olympus/RavenClaw framework, which is the core of a multimodal dialogue interaction interface of a wearable solution for information support of the motorcycle police force on the move. In the present effort, aiming at optimizing the speech recognition performance, different experimental setups are considered for the speech front-end. The practical value of various speech enhancement techniques is assessed and, after analysis of their performances, a collaborative scheme is proposed. In this collaborative scheme independent speech enhancement channels operate in parallel on a common input and their outputs are fed to the multithread speech recognition component. The outcome of the speech recognition process is post-processed by an appropriate fusion technique, which contributes for a more accurate interpretation of the input. Investigating various fusion algorithms, we identified the Adaboost.M1 algorithm as the one performing best. Utilizing the fusion collaborative scheme based on the Adaboost.M1 algorithm, significant improvement of the overall speech recognition performance was achieved. This is expressed in terms of word recognition rate and correctly recognized words, as accuracy gain of 8.0 % and 5.48 %, respectively, when compared to the performance of the best speech enhancement channel, alone. The advance offered in the present work reaches beyond the specifics of the present application, and can be beneficial to spoken interfaces operating in non-stationary noise environments.

Keywords: Speech recognition, speech enhancement, motorcycle environment, agent-based system, data fusion.

[1] Corresponding Author: Iosif Mporas, Wire Communications Laboratory, Dept. of Electrical and Computer Engineering, University of Patras, Greece, Tel. +30 2610 996496, Fax. +30 2610 997336, imporas@upatras.gr.

1. INTRODUCTION

1.1 Background

Human-computer interaction has a long history, during which various interfaces were developed, currently aiming to provide a more natural interaction, mimicking human to human interaction, such as 3D gesture or speech. Achievement of naturalness involves progress from command or menu-based (system driven) to user-driven dialog management and system intelligence, to allow adaptation to environment changes and user preferences. Technological advances in the Internet protocol (IP)-telephony domain, have leaded to an increased interest to provide accessibility to the large domain of web applications over the phone (Tsai, 2006). Personal assistant-based dialogue systems which offer higher comfort for the end-users were developed for the needs of various applications (Möller et al., 2004 and Paraiso and Barthes, 2006). Activities which traditionally were performed in an office or at home, i.e. in a well controlled environment, are now supported by mobile and embedded technologies and have migrated outdoors. There is an increased demand on services providing efficiency, empowered by high comfort and safety, in the new environment, taking into account that most of the times parallel activities, such as driving a car or a motorcycle, are performed. On the route, driver distraction can become a significant problem, thus highly efficient human-machine interfaces are required. In order to meet both, comfort and safety requirements, new technologies need to be introduced into the car environment enabling drivers to interact with mobile systems and services in an easy, risk-free way.

Spoken language dialogue systems considerably improve 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 been previously studied (Gartner, Konig, Wittig, 2001), 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 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 (Berton, Buhler, Minker, 2006).

Commercial use of speech recognition in the context of human-computer interaction has been intensified in the last decade, with major applications to “informative” systems: inquiries of time schedules for trains or movies, information regarding services and products, bank account or card balance inquiry, etc. Spoken communication, even between humans, is highly affected by noise pollution (Zahaeeruddin and Jain, 2008). Even more, the performance of speech recognition

systems, although reliable enough to support speaker and device independence in controlled environments, degrades substantially in a mobile environment 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 speaker's voice due to task stress, distributed attention, increased cognitive load, 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.

Speech recognition in car environment in the early 90's started with combinations of basic hidden Markov models (HMM) recognizers with front-end noise suppression, environmental noise adaptation and multi-channel concepts (Hansen, Clements, 1991, and Lockwood, Boundy, 1992). Preliminary speech/noise detection with front-end speech enhancement methods as noise suppression front-ends for robust speech recognition 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 (Visser, Otsuka, Lee, 2003). The advantages of multi-channel speech enhancement can be successfully applied to the car environment, while in the motorcycle environment 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 (Kamath and Loizou, 2002), (ii) the statistical model-based approaches (Ephraim and Malah, 1985, Loizou, 2005 and Hu and Loizou, 2004), (iii) the signal subspace approaches (Hu and Loizou, 2003, and Jabloun and Champagne, 2003), and (iv) the enhancement approaches based on a special type of filtering (Gannot, Burshtein and Weinstein, 1998).

The accuracy of the speech recognition task is highly improved by using suitably trained speech models for the recognizer engine. Sufficient noise scenarios, from the application domain, should be included in the training phase for the improvement of the performance. Dedicated speech corpora have been designed, recorded and annotated, starting with the car environment, and emerging with the motorcycle one. European initiative, supporting the development of corpora to support training and testing of multilingual speech recognition applications in the car environment started in 1998 with the SPEECHDAT-CAR project (Moreno et al., 2000). The databases developed are designed to include a phonetically balanced corpus to train generic speech recognition systems and an application corpus, providing enough data to adapt speaker independent recognition systems to the automotive environment. A total of 10 languages are supported, 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 a total of 500 speakers from United States of America and a natural conversational interaction (Hansen et al., 2003). The research of human-computer interaction in car environment has evolved to the multimodal mode (audio and visual), and adequate audio-visual corpus has been developed in the AVICAR database (Lee, Hasegawa-Johnson, Goudeseune, 2004) using a multi-sensory array of eight microphones and four video cameras. For the motorcycle environment, the SmartWeb motorbike corpus has been designed for a dialogue system dealing with open domains (Kaiser, Mogele, Shiel, 2006). Recently, a domain specific (police domain) database, dealing with the extreme conditions of the motorcycle environment, has been developed in the MoveOn project (Winkler et al., 2008). In the latest the focus is the specificity of the domain, where the cognitive load is quite high and the accuracy in recognition of commands in the context of a template driven dialog, in the motorcycle environment, is of high priority.

1.2 Our Goal and the Plan of the Present Contribution

In developing the speech interface in the MoveOn system, there are certain research challenges to overcome in order to achieve reliable and natural voice interaction in the motorcycle environment: zero-distraction interaction system for people who are moving on the road ("on the move", "eyes-busy" and "hands-busy") not being able to interact through a visual/tactile interface, such as a screen or a button pad, due to safety reasons. Target users that will benefit from the application of interest discussed here are police force motorcyclists and motorcycle drivers at large.

In the present work, we report on a challenging research and development effort for optimising the speech recognition accuracy of the MoveOn system's speech front-end. This development is based on a collaborative scheme, which relies on a number of speech enhancement channels and multithread automatic speech recognition component. The speech pre-processing, speech enhancement, speech recognition and data fusion components discussed in the following sections are implemented as interactive agents in the Olympus/RavenClaw framework, which is the core of the multimodal dialogue interaction system. The present work can be viewed as a natural continuation of an earlier study (Ntalampiras et al., 2008), where eight speech enhancement algorithms were evaluated in terms of objective and subjective quality of speech on the motorcycle speech database, referred to as MoveOn speech and noise database (Winkler et al., 2008). That earlier study provided useful indication about the potential usefulness of various speech enhancement algorithms and their performance in the target environmental conditions, and assisted us in selecting the potential candidates for best performing speech enhancement algorithms for the present work. However, since the performance of these speech enhancement algorithms in terms of usefulness for improvement of speech recognition performance is not known, and cannot be judged directly from the results of the aforementioned objective evaluation, we were motivated to consider the implementation of an optimized

system, taking advantage of multiple best performing algorithms, instead of simply relying on the top-performer in terms of perceptual quality.

Thus, in contrast to (Ntalampiras et al., 2008), in the present study the performance of the various speech enhancement schemes, on the MoveOn speech and noise database, is assessed by ranking their effect on speech recognition performance. Moreover, we demonstrate how the speech recognition performance can be boosted further by proper fusion of the outputs of several parallel speech enhancement channels, which are equipped with different speech enhancement techniques and accordingly adapted channel-specific acoustic models of the speech recognizer. The benefit of such collaborative scheme is experimentally demonstrated in terms of improved word recognition rate (WRR) and higher rates of correctly recognized words (CRW).

The remaining sections of this article are organized as follows: In Section 2 we introduce the MoveOn application, outline the architecture of the multimodal interaction dialogue system, and specify the requirements to the speech front-end. In Section 3 we discuss the implementation of the collaborative scheme for speech front-end, which is composed of multiple parallel speech enhancement channels, whose outputs are fused, in an attempt to improve the overall speech recognition performance. In Section 4 we detail the experimental setup, and Section 5 presents the experimental results. Finally, Section 6 offers discussion and concluding remarks.

2. SYSTEM ARCHITECTURE

In this section we briefly introduce the MoveOn application, the main design solution and the generic functionality requirements to the speech front-end.

2.1 Description of the MoveOn Application

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. This information support is obtained either remotely from the control centre in the police station, through a secure terrestrial trunked radio (TETRA) link, or locally through the functionality provided by the wearable computing environment developed within the project. This environment offers functionality 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.

2.2 Architectural Model of the MoveOn System

The architectural model of the multimodal dialogue interaction system, illustrated in Figure 1, and in the following referred to as the MoveOn system, is based on the functionality offered by the Olympus/RavenClaw dialog management framework (Bohus, Rudnicky, 2003, Bohus et al., 2007). This interaction framework, which builds on the CMU Communicator platform, relies on a centralized architecture, where a central unit, referred to as *hub*, provides synchronization for the rest of the components. Each component in the system functions as an agent, i.e. automatic speech recognizer (ASR), natural language understanding (NLU), natural language generation (NLG), text-to-speech convector (TTS), speech pre-processing, speech enhancement, etc., are agents, which communicate either directly with each other or through the central hub. The agents, the connections between them, as well as the triggering events handled by the hub are easily configurable and don't require recompilation of the system, thus allowing plug-in of new functionality.

Figure 1

The Olympus/RavenClaw framework was initially developed to provide management of spoken-based dialogs, but its modular architecture and modality independent task-based dialog management allow for easy integration of additional modalities. For the purpose of the MoveOn system, a series of new agents, handling different types of input/output modalities or additional necessary functionality, have been implemented and integrated to the architecture. For example, the speech enhancement component has been implemented as a new interactive agent, integrated within the Olympus/RavenClaw framework, and its operation is triggered by events sent by the pre-processing agent.

2.3 The Speech Front-end in the Context of the MoveOn System

The MoveOn system is implemented as a wearable solution, which constitutes of a purposely designed helmet, 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 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 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, 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 pre-processing, speech enhancement, speech recognition, and text-to-speech synthesis components. 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.

Since the noisy motorcycle environment constitutes a great challenge to the spoken dialogue interaction, in order to achieve robust communication we addressed the problem in a constructive manner. This involves optimization of all components involved in the human-computer interaction as well as the overall interaction strategy. Specifically, the interaction strategy was oriented towards minimizing the distraction of the motorcyclist and maximizing the robustness and efficiency of command and control functionality. Furthermore, the overall template-driven dialogue interaction was designed with special care for achieving reliable communication in adverse noise environment. This was done mainly by proper selection of the vocabulary of commands, command utterances and dialogue structure. Among all factors analyzed, the speech recognition performance proved to be the most crucial factor on which depends the overall success of interaction. Thus, in the following we will focus mainly on the optimization of the speech pre-processing, speech enhancement and speech recognition agents of the system, which collectively we refer to as the speech front-end of the dialogue interaction system.

3. SPEECH FRONT-END

The speech front-end mentioned in Section 2.3 considers single channel speech recognition. However, the optimized speech front-end system, implemented for the specific environment and present in here, is based on a collaborative scheme, which consists of four types of building blocks: speech pre-processing, speech enhancement, speech recognition

and fusion agents. As Figure 2 presents, in the proposed composite scheme, the speech front-end is designed to function as a parallel structure. In the operational mode, the speech pre-processing agent triggers the operation of the speech enhancement agent. The speech enhancement agent offers a multi-channel functionality, where each channel implements a different speech enhancement technique. It can be configured to operate in single channel mode, or multi-channel. In case of selection of single channel, the path followed further is classical, and doesn't involve fusion. In case of configuration of multi-channel mode, the output of the speech pre-processing agent represents a common input for the selected speech enhancement channels. The outputs of the speech enhancement channels are fed to a single multithread speech recognition agent, which utilizes channel-specific acoustic model and settings. Thus, the processing of speech in each channel follows a common processing flow, but differs in the employed algorithms for speech enhancement and in the channel-specific adaptation of the acoustic model utilized by the speech recognizer. The outputs of the multithread speech recognizer are fed to the data fusion agent.

Figure 2

In brief, the speech front-end operates as follows: In the pre-processing agent the input speech waveform is segmented to overlapping frame blocks and voice activity detection is performed on the frame sequence obtained so far. The voice activity detector labels the frames that correspond to speech activity and the remaining are considered as non-speech/ silence/ noise. Subsequently, the segmented speech signal is processed by the speech enhancement agents. The speech enhancement channels operate on the full length of the input signal, since some of them need periodical sampling of the noise statistics. However the speech recognition agent only considers the input signal segments in which speech activity was detected. This improves the speech recognition performance and contributes to the reduction of the overall computational load of the speech front-end.

Thus, only the enhanced speech frames that were labelled as speech are subject to speech parameterization, which converts each speech frame to a single feature vector. The sequence of feature vectors, computed so far, is further processed in the multithread speech recognition agent. The speech recognition agent incorporates a set of channel-specific acoustic models, each adapted for the corresponding speech enhancement method. These acoustic models are created in off-line mode by adapting a general purpose acoustic model, built from large amount of speech data. The adaptation consists in updating/retraining the adjustable parameters of the initial acoustic model, with certain amount of data that are processed through the corresponding speech enhancement technique. A common language model is utilized for all speech enhancement channels. The parallel outputs of speech recognition, agent, one for each speech enhancement channel, are

further combined in the fusion agent to obtain the final recognition result. The fusion scheme is learned in off-line mode and depends on the performance of the individual speech enhancement channels of the speech front-end.

*3.1 **Speech Enhancement Agent***

The speech enhancement algorithms considered here are plug-in components of the speech enhancement agent, each utilizing a standardised input/output interface. This makes possible easy addition of better performing algorithms in the future, by configuring them to function on an additional channel. In the present work, we consider the four top-performing speech enhancement techniques, among the eight investigated in (Ntalampiras et al., 2008), which were observed to perform well in the non-stationary motorbike environment conditions. Besides, we also consider the traditional spectral subtraction algorithm (Berouti, Schwartz, Makhoul, 1979), which is a well-studied technique, and which here serves as a reference point. Specifically, the speech enhancement techniques utilized in the present work are the:

1) Spectral subtraction (SS) method (Berouti, Schwartz, Makhoul, 1979). The SS method relies on the inherited additivity of power spectra of additive independent signals. This is approximately true for short-time estimates of the spectra as well. 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.

2) Spectral subtraction with noise estimation (SSNE) (Martin, 2001). 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. This algorithm is more appropriate for real world conditions, and outperforms the SS in non-stationary environments.

3) Multi-band spectral subtraction method (MBSS) introduced in (Kamath, Loizou, 2002). It is based on the well-known SS algorithm, but accounts for the fact that in real world conditions, interferences do not affect the speech signal uniformly over the entire spectrum, i.e. any real-world interference (which differs from the white noise) affects the speech spectrum differently at different frequencies. The MBSS method has been demonstrated to outperform the standard power spectral subtraction method resulting in superior speech quality and largely reduced musical noise. The results presented in (Kamath, Loizou, 2002) as well as our previous experience with the MoveOn data (Ntalampiras et al., 2008) motivated us to select this method.

4) Speech enhancement using a minimum mean square error log-spectral amplitude estimator (Ephraim, Malah, 1985), which we refer to as (MMSE-logSAE). This method 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 cate-

gory of statistical model-based algorithms. In previous work (Ntalampiras et al., 2008), it was observed to offer very good performance on the MoveOn data, therefore it is a strong candidate for achieving good speech recognition results.

5) Speech enhancement based on perceptually motivated Bayesian estimators (SE-PMBE) of the speech magnitude spectrum (Loizou, 2005). This algorithm utilized Bayesian estimators of the short-time spectral magnitude of speech based on perceptually motivated cost functions. In Loizou (2005), 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 MMSE-logSAE method. We selected this method due to its relatively good performance in (Ntalampiras et al., 2008), but also because we were interested to investigate if this advantage, when compared to MMSE-logSAE, will contribute for better speech recognition performance.

3.2 ***Speech Recognition Agent***

The parallel speech processing channels share an HMM-based speech recognition engine. In the MoveOn setup we identified the Julius decoder (Lee, Kawahara, Shikano, 2001) as the more appropriate among other speech recognizers (HTK, Sphinx-III), mainly due to the lower computational demands. The speech decoder utilizes an acoustic model together with an *n*-gram language model trained on data acquired in the environment of the target application domain. For the training of the acoustic models we exploited the HTK toolkit (Young et al., 2005), while for the language model we used the CMU Cambridge Statistical Language Modelling (SLM) Toolkit (Clarkson and Rosenfeld, 1997).

3.3 ***Fusion Agent***

The fusion agent of the speech front-end extracts the spoken utterance, based on processing of the channel-specific recognition results of the multithread speech recognition agent. In detail, the fusion agent operates as follows: for each input speech utterance the corresponding confidence scores obtained from the speech recognition agent, i.e. the acoustic score (AS), the language model score (LS) and the total score (TS), of each channel are used by a machine learning technique, which predicts the output with the minimum word error rate. The machine learning algorithm is trained in off-line mode from confidence scores vectors with class labels of the channel that minimizes the word error rate for the specific instance (input speech waveform). The recognition output of the channel that the fusion agent f has selected is considered as the recognized utterance and is forwarded for further process to the NLU agent, i.e.

$$f\left(\left\langle AS_1, LS_1, TS_1\right\rangle, ..., \left\langle AS_N, LS_N, TS_N\right\rangle\right) = c_j \text{, with } c_j \in C \quad (1)$$

where $\langle AS_i, LS_i, TS_i \rangle$ are the corresponding confidence scores of the *i*-th channel output, c_j is the channel that was predicted to have the lowest word error rate and $C = \{c_1, c_2, ..., c_N\}$ is the set of parallel channels in the speech front-end.

4. EXPERIMENTAL SETUP

The speech front-end proposed in Section 3 was evaluated in different configurations: single channel or multiple channel speech enhancement, different configuration settings of the speech recognition engine, and various fusion methods. Following, the speech data, the settings of the experimental setup, the fusion algorithms and the experimental protocol of the present evaluation are presented.

4.1 The MoveOn Speech and Noise Database

For the purpose of research and technology development in MoveOn project a dedicated speech database was recorded in the motorcycle environment (Winkler et al., 2008). Specifically, a group of thirty professional motorcyclists, members of the operational police force of UK, was recruited. 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 implemented as audio sequences that are played to the motorcyclists via earplug, while they were performing patrolling activates.

In total, the speech corpus consists of about forty hours of recordings, obtained in forty recording sessions. Different motorbikes and helmets were used, and the trace of the road differed among the sessions. Specifically, each session included in-city driving, highway, tunnels, suburbs, etc. In addition, there were ten recording sessions with the same hardware but in office environment.

Every session of the database consists of four audio channels recorded simultaneously: two from omni-directional microphones (AKG C 417”’) placed within the helmet – 10 cm one from another – at the two sides of the mouth; one channel from a throat microphone (Alan AE 38), and finally 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 for synchronization purposes during annotation. The language of all recordings is British English spoken by native speakers.

All 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 are labelled by their position and estimated magnitude. One additional

tier indicates when the helmet visor is open or closed, since this condition affects significantly the amount and the shaping of noise.

*4.2 **Speech Recognition Settings***

The configuration of the speech recognition engine mainly included the setup of the acoustic models, one for each speech enhancement channel, and the construction of a language model, common for all speech enhancement channels.

The channel-specific acoustic models were obtained by means of maximum a posteriori (Gauvain and Lee, 1994) adaptation of a general purpose British English acoustic model, with the enhanced speech recordings of the MoveOn database. The general purpose acoustic model was built from telephone speech recordings of the SpeechDat(II)-FDB4000-British database (Hoge et al., 1999). This acoustic model consists of one context-dependent HMM model for each phone of the British SAMPA alphabet, (Wells, 1997), provided with the lexicon of the British SpeechDat(II) database. All phone models are three state left-to-right HMMs without skipping transitions. Each state of the HMMs was modelled by a mixture of eight continuous Gaussian distributions. The state distributions were trained from speech feature vectors, estimated from speech waveforms after pre-processing and speech parameterization. The pre-processing of the speech signals, sampled at 8 kHz, consisted of frame blocking with length 25 milliseconds and step 10 milliseconds, and filtering with pre-emphasis coefficient equal to 0.97. The speech parameterization consisted in the computation of the energy of each frame, together with the twelve Mel frequency cepstral coefficients (MFCC), i.e. the 0th one excluded. The MFCC implementation relied on a filter-bank of 26 filters (Young et al., 2005). The speech feature vector consisted of the 13 static parameters computed so far together with their first- and second-order derivatives appended, which resulted to a total of 39 dimensions. All HMMs were trained through the Baum-Welch algorithm (Baum et al., 1970), with convergence ratio equal to 0.001.

For the training of the language model the MoveOn speech and noise database was used. The transcriptions of the responses of the MoveOn end-user, provided in (Winkler et al., 2007), were utilized to compute 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.

*4.3 **Evaluation Procedure***

Initially, the speech front-end was evaluated separately for each of the speech enhancement techniques, described in Section 3.2. Afterwards we performed comparison with the collaborative approach, proposed in Section 3, which relies on fusion of a number of parallel speech enhancement channels. For the purpose of data fusion we evaluated a number of well

known machine learning algorithms. For all the experimental tests, speech recognition performance was measured in terms of word recognition rate (WRR) and percentage of correctly recognized words (CRW).

In all experiments the speech decoder utilized the channel-specific acoustic models. The speech recognition performance was tested both for bi-gram and tri-gram word-level language models. The speech enhancement methods described in Section 3.2 were implemented as in (Loizou, 2007).

The fusion scheme, presented in Section 3.3, was evaluated in alternative implementations of the machine learning algorithms. In detail, we used a two-layered backpropagation multilayer perceptron neural network, MLP, (Mitchell, 1997), a support vector classifier with radial basis function kernel utilizing the sequential minimal optimization algorithm, SVM, (Keerthi et al., 2001), a *k*-nearest neighbour classifier, IBk, (Aha, Kibler, 1991) and a C4.5 decision tree learner, J48, (Quinlan, 1993). Except these, we employed two bagging algorithms, one using decision trees, Bagging (J48), and one using fast tree learner with reduced error pruning, Bagging (REPTree), (Breiman, 1996). Finally, a boosting algorithm combined with decision trees, Adaboost.M1 (J48), (Freund, Schapire, 1996) was used. For the evaluation of these fusion methods we relied on their implementation in the WEKA machine learning toolkit (Witten and Frank, 2005).

In order to ensure the reliability of our experimental results we performed ten-fold cross validation experimentations, using ninety percent of the data for training and ten percent for testing during each fold. Thus, there was no overlapping between the training and test data for any of the performed experiments.

5. EXPERIMENTAL RESULTS

In this section, the experimental results are presented, considering, in a first step, the performance of individual speech enhancement algorithms and, in a second step, the performance of a fusion collaborative scheme, applied to individual speech enhancement channels. The performance of the front-end speech enhancement is measured in terms of averaged word recognition rate (WRR) per speech utterance.

The performance results for each individual speech enhancement method in the motorcycle environment, in terms of WRR, are shown in Table 1, for both bi-gram and tri-gram language models. The notions of the speech enhancement methods in the first column of the table stand for: Speech enhancement based on perceptually motivated Bayesian estimators (SE-PMBE), Multi-band spectral subtraction (MBSS), Speech enhancement using a minimum mean square error log-spectral amplitude estimator (MMSE-logSAE), Spectral Subtraction with noise estimation (SSNE), the original Spectral subtraction (SS), and a speech recognizer without speech enhancement (no enhancement), which is considered as the baseline. Note that in all cases, including the case without speech enhancement, adapted acoustic models were used.

Table 1

As can be seen in Table 1, the SE-PMBE enhancement method led to the best speech recognition performance, when compared to the other enhancement algorithms. For the case of speech decoding with bi-gram language model, the SE-PMBE method improved the WRR by 3.32% when compared to the baseline performance. The SE-PMBE method is followed by the MBSS method, which expressed slightly lower speech recognition performance than the first one, but still improved the baseline performance by 2.35%. The remaining methods, i.e. the MMSE-logSAE, SSNE and SS, achieved lower speech recognition performance, but still offering some improvement when compared to the baseline results. It is noteworthy mentioning that the ranking of speech enhancement algorithms in terms of speech recognition performance, presented in Table 1, is partially aligned with the ranking in terms of speech quality evaluation reported in (Ntalampiras et al., 2008). 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 front-end, comparing to the subjective speech perception tests examined in (Ntalampiras et al., 2008).

As Table 1 presents, in all cases the performance of the speech recognition decoder when using bi-gram language model was higher, when compared to the one for tri-gram model. This is owed to the limited amount of domain-specific data that were available for training the language model. The limited amount of training data did not allow the accurate computation of tri-gram models – there were not enough occurrences for all word tri-grams to train robust models. Thus, in the following experiments we report results only for the bi-gram language model.

In order to investigate the statistical significance among the different WRRs, reported in the second column in Table 1, we performed the paired *t*-test. The results of the *t*-test are shown in Table 2. The highlighted cells correspond to statistically similar recognition results.

Table 2

As it can be seen in Table 2, the word recognition rates reported in column 2 in Table 1, are not statistically different among the cases when the speech enhancement is performed by the MMSE-logSAE, SSNE and SS methods. However, the speech recognition results for the MBSS and SE-PMBE are statistically different from the remaining methods.

As we already said, the speech recognition performance, presented in Table 1, accounted for the averaged WRRs. In order to investigate in-depth the practical usefulness of each speech enhancement method for improving the speech recognition performance, we computed the number of speech utterances where each speech enhancement algorithm outperformed the others. In Table 3 we present these results in terms of WRR, i.e. for how many speech utterances the specific speech enhancement algorithm showed the highest WRR. The total number of utterances used in this study was 10201.

Table 3

As Table 3 shows, the SE-PMBE method is not always the best performing method, although it leads to the maximum averaged WRR (refer to Table 1). In Table 3 we can see that for more than 20 % of all cases, i.e. for 2210 utterances out of the 10201, the SE-PMBE speech enhancement algorithm was outperformed by one or more of the other algorithms. This observation was the basis to consider a fusion collaborative scheme in order to improve the overall speech recognition performance. In this fusion scheme, the selection of the most accurate speech recognition output, in terms of word recognition rate, for each input instance would lead to increase of the overall speech recognition accuracy of the speech front-end.

Therefore, in the following experimentations we combined the speech recognition outcomes, obtained for the corresponding speech enhancement channels, by utilizing various machine learning algorithms: Adaboost.M1 (J48), IBk, Bagging (J48), J48, Bagging (REPTree), MLP, and SVM algorithms. In Table 4 we present the speech recognition performance in terms of WRR for these fusion methods and for a variable number of top-performing speech enhancement channels. Specifically, the 2-best case corresponds to the fusion of the two speech recognition outputs that correspond to the two channels implementing the best performing speech enhancement algorithms (SE-PMBE and MBSS – refer to Table 1), while the 6-best corresponds to the fusion of speech recognition results, obtained for all speech enhancement channels and the channel without speech enhancement.

Table 4

As can be seen in Table 4, the most successful fusion scheme is based on the Adaboost.M1 (J48) method, since it offered the highest performance for any number of channels, *n*-best, with $n \in [2, 6]$. The IBk method led to slightly lower WRR comparing to Adaboost.M1, while for the 6-best case the two methods presented equal performance. Furthermore as the table presents, the Bagging (J48), J48, Bagging (REPTree) and MLP methods did not differ significantly in terms of

WRR. Finally, the SVM method proved to improve the WRR by 1.95% comparing to the best performing single channel, SE-PMBE. However, its performance was significantly lower than the other evaluated fusion methods.

As it was expected, the 6-best case offered the highest speech recognition performance across all fusion methods. The speech recognition performance for the fusion case 5-best was slightly lower, indicating that the exclusion from the fusion scheme of the baseline speech recognition result, i.e. speech recognition without prior speech enhancement, does not affect notably the overall performance of the speech front-end. Reduction of the number of parallel speech enhancement channels has the advantage of decreased computational demands. Furthermore, some reduction of the overall speech recognition performance was observed for the 4-best, 3-best and 2-best cases, when compared to the one for the 6-best. However, the competitive WRR results for the 4-best, 3-best and 2-best fusion schemes offer the opportunity WRR performance to be traded for computational complexity.

Since in the dialogue interaction process, the success of interaction depends on the correctly recognized utterances (which translates into correctly recognized uttered words), and doesn't explicitly depend on WRR, in Table 5 the result for the number of correctly recognized words (CRW) are presented.

Table 5

As Table 5 presents, the collaborative speech recognition scheme employed on the present task leads to significant improvement of the CRW, when compared to the CRW for the best performing speech enhancement channel, i.e. the one that incorporates the SE-PMBE speech enhancement algorithm. Here we present the CRW rates only for the best performing fusion algorithm – Adaboost.M1. Even the simplest fusion scheme, which exploits the speech recognition results obtained for the 2-best speech enhancement channels, increases the rate of CRW by 4.35%, and contributes for significant improvement of the overall success of interaction. By increasing the number of speech recognition outcomes employed in the fusion scheme to 6 (the maximum tested in this study), an improvement of 5.48% of the CRW is achieved, compared to the best performing individual speech enhancement method. These results provide the experimental validation of the practical significance of the proposed collaborative speech processing scheme, which is the core of the speech front-end studied in the present work.

6. DISCUSSION AND CONCLUSION

In the present contribution we studied a collaborative scheme for speech recognition, which is based on a multi-channel speech enhancement agent, implementing different speech enhancement techniques per channel. The outputs of the individual speech enhancement channels are fed to multithread speech recognition agent. The speech recognition outcomes, corresponding to the individual speech enhancement channels are post-processed by utilizing a fusion method that is trained to predict the best performing channel for each specific input. Alternative implementations of the fusion scheme were evaluated in order to identify the best performing one.

The presented experimental evaluation, using real-world database collected in the operations environment of the MoveOn application, validates the practical significance of the proposed collaborative scheme. Significant improvement of the overall speech recognition performance, both in terms of word recognition rate and correctly recognized words, was observed for fusion of the speech recognition results for the *n*-best speech enhancement channels, when compared to the one obtained by employing only the best performing speech enhancement method.

Due to practical limitations, among which is our striving towards portability we consider the 2-best fusion scheme as the most advantageous for the wearable MoveOn solution. This fusion scheme offers the best trade-off between computational demands and speech recognition performance, and complies well with the requirement for sufficient battery life.

As already discussed, the addition of more speech enhancement channels offers the chance for a further improvement of the speech recognition performance, at the price of additional computational demands. In future motorcycle-based applications, where the battery budged is permitting, we would suggest the use of the multiple best channels fusion scheme, which offers further advantages in terms of speech recognition performance improvement.

ACKNOWLEDGEMENTS

The research leading to these results was financially supported by the MoveOn project (IST-2005-034753), under the [European Community's] Sixth Framework Programme.

REFERENCES

Aha, D., & Kibler, D. (1991). Instance-based learning algorithms. Machine Learning, (6), 37-66.

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.

Berouti, M., Schwartz, R., & Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise. Proceedings of the IEEE ICASSP'79 (pp. 208-211). Washington, DC, USA.

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

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 (pp. 32-39). Roshester, NY, USA.

Bohus, D., & Rudnicky, A.I. (2003). RavenClaw: Dialog Management Using Hierarchical Task Decomposition and an Expectation Agenda. Proceedings of the Eurospeech 2003 (pp. 597-600). Geneva, Switzerland.

Breiman, L., (1996). Bagging predictors. Machine Learning, 24(2), 123-140.

Clarkson, P. R., & Rosenfeld, R. (1997). Statistical Language Modeling Using the CMU-Cambridge Toolkit. Proceedings of the Eurospeech 1997 (pp. 2707-2710). Rhodes, Greece.

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.

Freund, Y., & Schapire, R.E. (1996). Experiments with a new boosting algorithm. Proceedings of the 13th International Conference on Machine Learning (pp. 148-156). San Francisco, USA.

Gannot, S., Burshtein, D., & Weinstein, E. (1998). Iterative and sequential Kalman filter-based speech enhancement algorithms. IEEE Transactions on SAP, 6(4), 373-385.

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.

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

Hansen, J.H.L., & Clements, M.A. (1991). Constrained Iterative Speech Enhancement with Application to Speech Recognition. IEEE Transactions on ASSP, 39(4), 795-805.

Hansen, J.H.L., Zhang, X., Akbacak, M., Yapanel, U., Pellom, B., & Ward, W. (2003). CU-Move: Advances in in-vehicle speech systems for route navigation. Proceedings of the IEEE Workshop on DSP in Mobile and Vehicular Systems (19-45). Nagoya, Japan.

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 Eurospeech 1999 (pp. 2699-2702). Budapest, Hungary.

Hu, Y., & Loizou, P. (2003). A generalized subspace approach for enhancing speech corrupted with colored noise. IEEE Transactions on SAP, 11(4), 334-341.

Hu, Y., & Loizou, P. (2004). Speech enhancement by wavelet thresholding the multitaper spectrum. IEEE Transactions on SAP, 12(1), 59-67.

Jabloun, F., & Champagne, B. (2003). Incorporating the human hearing properties in the signal subspace approach for speech enhancement. IEEE Transactions on SAP, 11(6), 700-708.

Kaiser, M., Mogele, H., & Shiel, F. (2006). Bikers Accessing the Web: The SmartWeb motorbike corpus. Proceedings of LREC 2006. Genoa, Italy.

Kamath, S., & Loizou, P. (2002). A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. Proceedings of ICASSP-2002 (vol. 4, pp. 4164-4167). Orlando, USA.

Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., & Murthy, K.R.K. (2001). Improvements to Platt's SMO Algorithm for SVM Classifier Design. Neural Computation, 13(3), 637-649.

Lee, A., Kawahara, T., & Shikano, K. (2001). Julius -- an open source real-time large vocabulary recognition engine. Proceedings of the Eurospeech 2001 (pp. 1691-1694). Aalborg, Denmark.

Lee, B., Hasegawa-Johnson, M., & Goudeseune, C. (2004). AVICAR: Audio-visual speech corpus in a car environment. Proceedings of ICSLP 2004 (pp. 2489-2492). Jeju Island, Korea.

Lockwood, P., & Boundy, J. (1992). Experiments with a Nonlinear Spectral Subtractor (NSS), HMMs and the projection, for robust speech recognition in cars. Speech Communication, 11(2-3), 215-228.

Loizou, P. (2005). Speech enhancement based on perceptually motivated Bayesian estimators of the speech magnitude spectrum. IEEE Transactions on SAP, 13(5), 857-869.

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

Martin, R. (2001). Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Transactions on SAP, 9(5), 504-512.

Mitchell, T.M. (1997). Machine Learning. McGraw-Hill International Editions.

Moreno, A., Linderberg, B., Draxler, C., Richard, G., Choukri, K., Euler, S., & Allen, J. (2000). SPEECHDAT-CAR. A Large Speech Database for Automotive Environments. Proceedings of LREC 2000. Athens, Greece.

Möller, S. Krebber, J., Raake, A., Smeele, P., Rajman, M., Melichar, M., Pallotta, V., Tsakou, G., Kladis, B., Vavos, A., Hoonhout, J. Schuchardt, D., Fakotakis, N., Ganchev, T., & Potamitis, I., (2004). "INSPIRE: Evaluation of a Smart-Home System for Infotainment Management and Device Control", Proceedings of LREC 2004 (vol.5, pp.1603-1606). Lisbon, Portugal.

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). Athens, Greece.

Paraiso, E.C., Barthes, J.-P.A. (2006). An intelligent speech interface for personal assistants in R&D projects, Expert Systems with Applications, 31(4), 673-683.

Quinlan, R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.

Tsai, M.J. (2006). VoiceXML dialog system of the multimodal IP-Telephony – The application for voice ordering service. Expert Systems with Applications, 31(4), 684-696.

Visser, E., Otsuka, M., & Lee, T.W. (2003). A spatio-temporal speech enhancement scheme for robust speech recognition in noisy environments. Speech Communication, 41(2-3), 393-407.

Zaheeruddin & Jain, V. K. (2008). An expert system for predicting the effects of speech interference due to noise pollution on humans using fuzzy approach. Expert systems with Applications, vol.35(4), 1978-1988.

Wells, J.C. (1997). SAMPA computer readable phonetic alphabet. In D. Gibbon, R. Moore, & R. Winski, Handbook of Standards and Resources for Spoken Language Systems (Part IV, section B). Berlin and New York: Mouton de Gruyter.

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.

Winkler, T., Kostoulas, T., Adderley, R., Bonkowski, C., Ganchev, T., Kohler, J., & Fakotakis, N. (2008). The MoveOn motorcycle speech corpus. Proceedings of LREC 2008. Marrakech, Morocco.

Witten, I.H., & Frank, E. (2005). Data mining: practical machine learning tools and techniques (2nd ed, Morgan-Kaufman Series of Data Management Systems). San Francisco: Elsevier.

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.

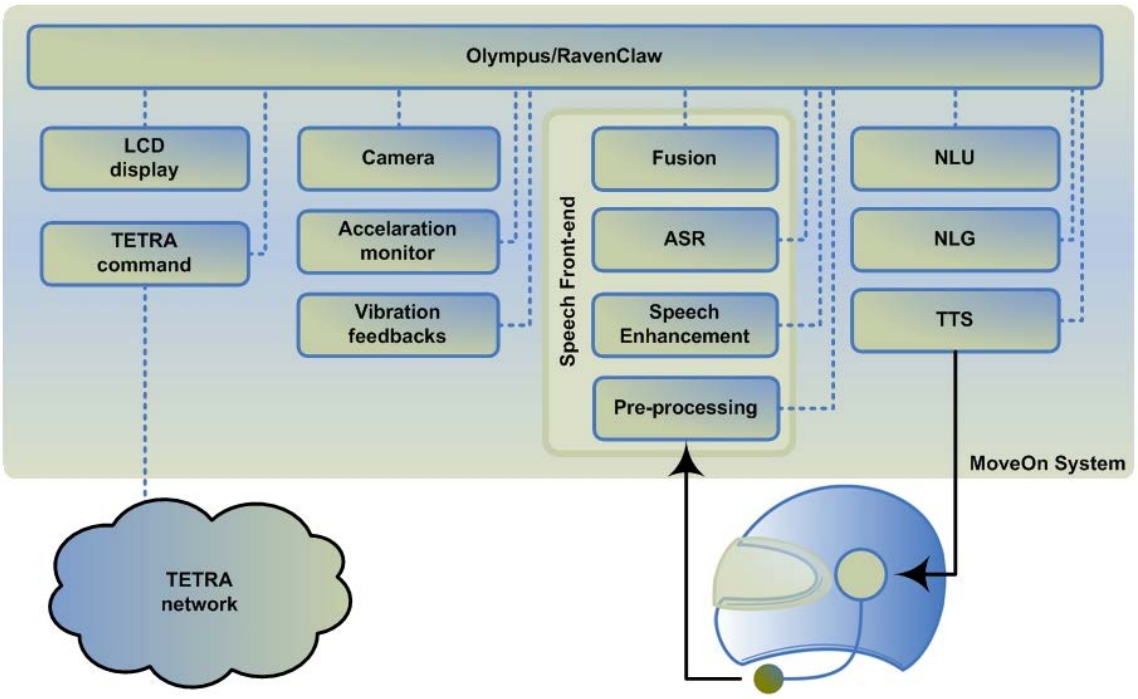

Figure 1. Architectural model of the MoveOn system.

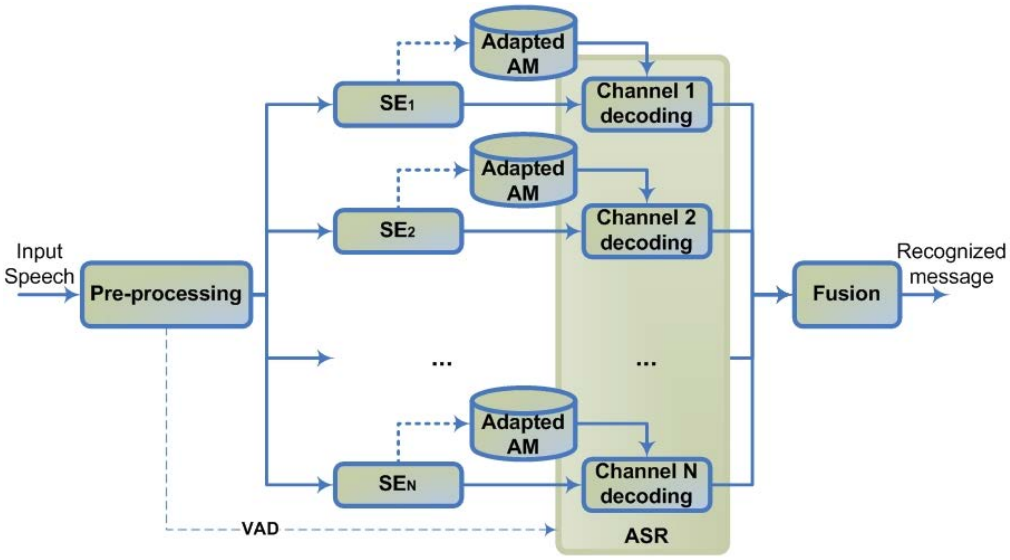

Figure 2. Functional diagram of the MoveOn speech front-end. Speech enhancement agents are denoted as SEs, acoustic models are denoted as AMs and speech recognition agent is denoted as ASR.

Table 1. Speech recognition performance (in terms of WRR) for different speech enhancement methods and bi-gram and tri-gram language models.

Enhancement method	2-gram LM [%]	3-gram LM [%]
SE-PMBE	86.67	76.10
MBSS	85.70	75.12
MMSE-logSAE	84.86	74.23
SSNE	84.85	74.38
SS	84.83	75.45
no enhancement	83.35	71.62

Table 2. The paired *t*-test for the speech recognition results obtained for bi-gram language model.

$\|t\|>1.64$	SE-PMBE	MBSS	SSNE	SS	no enhancement	MMSE-logSAE
SE-PMBE	---	7.20e+00	1.05e+01	1.02e+01	1.47e+01	1.05e+01
MBSS	---	---	3.44e+00	2.98e+00	6.68e+00	3.43e+00
SSNE	---	---	---	-4.04e-01	3.54e+00	4.01e-02
SS	---	---	---	---	-4.15e+00	4.19e-01
no enhancement	---	---	---	---	---	-3.52e+00

Table 3. Number of speech utterances for which the specific speech enhancement method led to the highest WRR.

Method	# utterances	% of all cases
SE-PMBE	7991	78.36
MBSS	1136	11.14
MMSE-logSAE	550	05.39
SSNE	290	02.84
SS	197	01.93
no enhancement	37	00.36

Table 4. Speech recognition performance (in terms of WRR) for various fusion methods and different number of speech enhancement/recognition channels combined.

Fusion method	2 – best [%]	3 – best [%]	4 – best [%]	5 – best [%]	6 – best [%]
AdaboostM1 (J48)	92.18	93.58	93.67	94.29	94.67
IBk	92.12	93.54	93.61	94.28	94.67
Bagging (J48)	91.59	93.01	93.07	93.73	94.08
J48	91.42	92.73	92.85	93.39	93.63
Bagging (REPTree)	91.54	92.65	92.67	93.11	93.37
MLP	91.20	92.19	92.08	92.72	92.62
SVM	87.92	88.02	88.26	88.53	88.62

Table 5. Correctly recognized words (CRW) in percentage from all utterances for the best performing speech enhancement channel and for different number of collaborating channels, when the fusion is based on the Adaboost.M1 method.

Fusion	**CRW [%]**
Best ASR channel (with SE-PMBE method)	91.26
2-best fusion	95.61
3-best fusion	96.30
4-best fusion	96.34
5-best fusion	96.59
6-best fusion	96.74

List of Figures

List of Tables