Context-adaptive pre-processing scheme for robust speech recognition in fast-varying noise environment

Iosif Mporas, Todor Ganchev*, Otilia Kocsis and Nikos Fakotakis

Wire Communications Laboratory, Dept. of Electrical and Computer Engineering,

University of Patras, 26500 Rion-Patras, Greece

imporas@upatras.gr, * tganchev@ieee.org, okocsis@upatras.gr, fakotaki@upatras.gr

*Corresponding author at:

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

Tel.: +30 2610 969808
FAX: +30 2610 997336
E-mail address: tganchev@ieee.org
tganchev@wcl.ee.upatras.gr

* Corresponding author: Todor Ganchev (tganchev@ieee.org), Senior Member of the IEEE, Member of the EURASIP

Keywords: fast-varying noise; speech enhancement; speech pre-processing; speech recognition; motorcycle environment

List of unusual symbols used

<empty>

Number of pages: 26

Number of tables: 4

Number of figures: 1

List of Figures

List of Tables

Abstract

Based on the observation that dissimilar speech enhancement algorithms perform differently for different types of interference and noise conditions, we propose a context-adaptive speech pre-processing scheme, which performs adaptive selection of the most advantageous speech enhancement algorithm for each condition. The selection process is based on an unsupervised clustering of the acoustic feature space and a subsequent mapping function that identifies the most appropriate speech enhancement channel for each audio input, corresponding to unknown environmental conditions. Experiments performed on the MoveOn motorcycle speech and noise database validate the practical value of the proposed scheme for speech enhancement and demonstrate a significant improvement in terms of speech recognition accuracy, when compared to the one of the best performing individual speech enhancement algorithm. This is expressed as accuracy gain of 3.3% in terms of word recognition rate. The advance offered in the present work reaches beyond the specifics of the present application, and can be beneficial to spoken interfaces operating in fast-varying noise environments.

1. Introduction

Mobile systems, providing a large variety of services and interactions in a continuously changing environment, are nowadays a reality, and some of the most advanced applications have been ported to the mobile world [1]. Activities, which traditionally were performed in an office or at home, in a well-controlled environment, have now migrated outdoors, being supported by mobile and embedded technologies. The last results in an increased demand on services providing efficiency empowered by high comfort and safety in the new environment, taking into account that most of the time parallel activities, such as driving a car or a motorcycle, are performed. On the route, driver distraction can lead to significant risks, thus highly efficient human-computer interfaces are required.

1.1. *Problem formulation and main challenges in the motorcycle-on-the-move environment*

In order to meet both, comfort and safety requirements, new technologies need to be introduced into the mobile environment, enabling drivers to interact with mobile systems and services in an easy risk-free way. Driving quality, stress and strain situations and user acceptance when using speech and manual commands to acquire certain information on the route have previously been studied [2], 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 [3]. Even more, the use of graphical or gesture based interfaces although possible to some degree when driving a car, is highly limited when driving a motorcycle.

The performance of speech-based interfaces, although reliable enough in controlled environments to support speaker and device independence, degrades substantially in a non-stationary environment [4], reaching its worst in the motorcycle on the move environment. There are various factors, which contribute for severe degradation of the speech signal in a moving-motorcycle environment, among which are:

(i) the presence of additive interferences from the acoustic environment, such as rumble noises from road vibrations or from the friction between the tires and the road surface, other mechanical noise from fans, gears, horns, wind noise, engine noise, surroundings traffic noise, etc,

(ii) speech signal alteration related to changes in the speaker's voice and speaking style due to task stress, distributed attention, physical efforts, body vibrations, Lombard effect, etc.

In the present work, we focus our attention on compensating the speech signal degradation caused by the presence of additive interferences from the acoustic environment, which is somehow decoupled from the signal alteration due to changes in the cognitive load, the physical stress and the body vibration of the motorcyclist. Thus, any effort to deal with the alteration of speech due to the condition of the motorcyclist and the physical stress over his body remains beyond the scope of this work.

1.2. *Description of related work*

The accuracy of the automatic speech recognition is significantly improved by using suitably trained acoustic models for the speech decoder. Sufficient samples of the various noise scenarios and samples from the application domain should be included in the data set used for training of the acoustic models to achieve the improvement of the overall speech recognition accuracy. For that purpose, various dedicated speech databases, which are representative for a set of mobile voice-interaction applications, have been designed, recorded and annotated, starting with the car environment, and emerging with the motorcycle one. European initiative, aiming at the development of databases in support of multilingual speech recognition applications in the car environment started in 1998 with the SPEECHDAT-CAR project [5]. 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 ten European 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 the United States of America and a natural conversational interaction [6]. The research on 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 [7] 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 [8]. Recently, a domain-specific database (operations of the motorcycle police force), dealing with the extreme conditions of the motorcycle environment, has been developed in the MoveOn project [9]. In the latest, the focus is on the domain specificity of the moving motorcycle on-the-road environment, where the cognitive load is quite high and the accuracy in recognition of commands in the context of a template driven dialog is of high priority.

In addition to the use of dedicated speech databases for adapting the acoustic models of the speech decoders, it has been proved that addition of noise suppression front-ends contributes for the improvement of the speech recognition accuracy. In the early 90’, the first trials to perform speech recognition in car environment were done, and started with combinations of basic hidden Markov model (HMM) recognizers with front-end noise suppression, environmental noise adaptation and multi-channel concepts [10, 11]. 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 [12]. 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, since using microphone arrays is impractical.

After more than three decades of advances on the one-channel speech enhancement problem [13, 14], four distinct families of algorithms seem to have predominated in the literature: (i) the spectral subtractive algorithms [15, 16, 17], (ii) the statistical model-based approaches [18, 19, 20], (iii) the signal subspace approaches [21, 22], and (iv) the enhancement approaches based on a special type of filtering [23]. The references illustrating each of the aforementioned groups are indicative and here we do not claim exhaustiveness of the list for each family of algorithms. However, it is important to emphasize that although in each of the aforementioned families there are few algorithms which demonstrate high performance for a specific set of noise conditions, and although one can identify “the best performing algorithm among all families”, the speech accuracy gain obtained by this single method might still remain insufficient (and in a way suboptimal), when operation in highly non-stationary and fast-varying noise environments, such as the one associated with the motorcycle-on-the-move environment, is considered. This has been confirmed by a recent research [24], where evidence that a collaborative speech enhancement scheme outperforms the best individual algorithm was provided.

The research work presented in the following sections has been conducted in order to develop a robust and energy efficient speech interface in the MoveOn system [25] dedicated for command and control applications for police force motorcyclists. For the specific target group, a zero-distraction interaction system is aimed, since due to safety reasons the drivers are not able to interact through visual/tactile interfaces such as a screen or a button pad. Our research is supported by a dedicated speech database, which has previously been developed [26], and speech recognition accuracy has been improved by using suitably trained acoustic models. Recent research has shown that speech recognition accuracy can further be improved if a collaborative noise reduction scheme is employed, for the needs of the speech enhancement process [24, 27]. However, this advance relies on multiple speech enhancement channels that operate in parallel, and thus it was achieved at the cost of a significant increase of the computational and memory demands and a significantly increased complexity of the speech frond-end.

1.3. *Advance in the present work*

In the present work, we address the challenges imposed by the highly non-stationary and fast-varying noise environment in a constructional manner and develop a speech pre-processing scheme that adapts its configuration depending on the audio input. The present contribution builds on the fact that dissimilar speech enhancement algorithms perform differently for dissimilar types of interference and noise conditions [28], and on the idea that the most appropriate speech enhancement algorithm for each environmental condition can be selected dynamically during the run-time operation of the speech front-end, depending on the present audio input. The proposed adaptive scheme automatically selects only one speech enhancement channel, among all available, and thus alleviates scalability constraints inherent to earlier designs [24, 27], which assume a number of speech enhancement algorithms operating in parallel on a common input.

Specifically, the adaptive speech pre-processing scheme proposed here is organized as a two-stage process, where in the first stage the parameterized audio input is compared against a number of predefined clusters in the acoustic feature space to generate a new feature vector which consist of normalized log-likelihoods. The second stage employs a certain machine learning technique, which uses this new feature vector, in order to map the first-stage output to the most appropriate speech enhancement channel, among the available ones. In this manner each input speech utterance is redirected for processing to the most appropriate speech enhancement channel. Provided that a set of dissimilar speech enhancement methods are involved, and each of them offers an advantage in given noise conditions, we suppose that the overall accuracy improvement offered by the context-adaptive pre-processing scheme studied here will be higher than the accuracy of the best individual speech enhancement algorithm alone, when highly non-stationary and fast-varying noise environments are considered. To the best knowledge of the authors, the proposed selective scheme, based on GMM-clustering and subsequent mapping function for selection of the most appropriate speech enhancement channel has not previously been studied, and thus constitutes significant novelty in the manner that a speech front-end copes with highly non-stationary fast-varying noise conditions.

The practical usefulness of the proposed adaptive speech pre-processing scheme is investigated with the use of a number of traditional and recently developed speech enhancement algorithms. It is experimentally demonstrated that the proposed scheme contributes to a significant improvement of the speech recognition accuracy, when compared to the baseline result (the best individual speech enhancement method when used alone), while it needs only a fraction of the computational demands of the earlier design [24]. Thus, the proposed new design contributes to a significant reduction of the computational demand during operation and improves the energy efficiency of the speech front-end, which in the MoveOn application is part of a wearable solution that operates on battery power. The number and the actual choice of speech enhancement algorithms are application-specific issues, and their choice does not affect the logic of the operation of the proposed context-adaptive speech front-end.

The remaining of this article is organized as follows: In Section 2, we present the context-adaptive pre-processing scheme for speech enhancement. In Section 3, we briefly outline the MoveOn speech and noise database used in the experiments. In Section 4, we briefly outline the speech enhancement algorithms employed, detail on the experimental protocol that was followed, and explain the implementation and the settings of the various components. In Section 5, we present and discuss the experimental results and, finally, in Section 6, we conclude this work with a brief summary of work and results.

2. The context-adaptive speech pre-processing scheme

Fast-varying noise environments, which are typical for a motorcycle-on-the-move and are characterized by the superposition of interferences that vary vastly in both duration and spectral contents, are significant impediment for the use of speech recognition-based interaction services. This is because the speech enhancement process encounters significant difficulties due to the non-stationary interferences, originating from the acceleration and the deceleration of the engine, the vibrations from the road, the contrary wind, the change of location (city, tunnel, suburbs, etc), the different traffic conditions, etc, which are introduced to the speech signal. Specifically, as shown in [28], different speech enhancement algorithms demonstrate dissimilar performance for different types of noise. This fact is owed to the different manner in which the noisy speech signal is processed by each method, which in turn leads to advantageous performance of some speech enhancement algorithms for specific types of noise, but not for all types. Therefore, in non-stationary fast-varying noise environments, such as the one associated with motorcycles-on-the-move, the use of a single, even the best performing on average, speech enhancement method will be a suboptimal solution in the effort for achieving robust spoken interaction (this is validated experimentally in Section 5).

However, provided that a number of dissimilar speech enhancement methods are available, we suppose that selecting the most advantageous method for each environmental acoustic condition will contribute to the improvement of the overall accuracy, when compared to the accuracy of the best on average individual speech enhancement algorithm. Based on this concept, in the following we present a speech front-end, which relies on multiple speech enhancement channels and a channel selector, which selects the most appropriate speech enhancement algorithm, with respect to the predominant noise content in the input signal. Specifically, the channel selector is based on an unsupervised clustering of the feature vector space, and subsequent mapping between a cluster-based feature vector and that speech enhancement method which was found (on a training dataset) as the most appropriate for the specific type of input interference. The entire training process is data-dependent and since the mapping process is unsupervised, there is not any set of predefined named acoustic conditions assumed, but the corresponding acoustic condition of each audio input remains unnamed.

In the following subsections, we assume the voice inputs to the speech front-end to have the length of a short utterance, i.e., approximately one to two seconds, which allows us to assume that for the duration of that input, the environmental characteristics do not change significantly. This assumption is well supported by the operational conditions of the MoveOn command and control speech interaction system [9], where in all cases the overall duration of a speech utterance is less than three seconds. Another implicit assumption is that we possess a training dataset which is representative for the acoustic environment and the application of interest. In the present work, this assumption is correct since we possess the MoveOn noise and speech database which is briefly outlined in Section 3.

2.1. *Concept of the context-adaptive speech pre-processing scheme*

The proposed context-adaptive scheme is based on two stage processing of the input audio that dynamically selects the most appropriate speech enhancement method among all available. The selection process is based on GMM-clustering of the audio feature space in the first stage and, a mapping function in the second stage, which links the first stage output to one of the speech enhancement methods. In fact, the second stage operates on a newly generated feature vector consisting of the log-likelihoods estimated for a given audio input to belong to each cluster of the first stage. The mapping function in the second stage is implemented with a machine learning technique, which is capable to learn the relations between the output of the GMM-clustering and the speech enhancement method that was identified as the most advantageous for that type of audio input.

The clusters used at the first stage enclose a large number of audio inputs with similar acoustic characteristics. These clusters are learned in an unsupervised manner from a training dataset that is representative for the operational acoustic environment of interest. The mapping function, employed in the second stage, is also learned during training and depends on the particular design of the clustering stage and on the number and the performance of the speech enhancement techniques which are selected for the speech front-end. In fact the training of the second stage also involves efforts for establishing the performance of the different speech enhancement methods (in terms of speech recognition accuracy) for a number of audio inputs from the training dataset. Actually, during training, the mapping function learns the mapping between the output of the GMM-clustering stage, i.e. the feature vector consisting of the log-likelihoods estimated for each cluster, and the indexes of the available speech enhancement methods, based on the evidence about the appropriateness of each of these enhancement methods for each audio input, obtained during their evaluation.

The block diagram of the proposed speech front-end, with GMM-based clustering of the feature space and mapping to the most appropriate speech enhancement channel, is illustrated in Fig. 1. As the figure presents, the input audio is initially parameterized, and afterwards the sequence of speech feature vectors belonging to the input utterance is compared against the set of available cluster models. The outcome of this comparison is a vector of log-likelihood scores that is fed to the mapping function, which selects the most appropriate speech enhancement channel to be used. The audio input is then processed by the selected speech enhancement channel, and the enhanced speech signal is forwarded to the ASR, which utilizes the corresponding channel-dependent acoustic model. In such a manner, without significant increase of the computational demands compared to the case of using a single speech enhancement technique, the speech front-end is able to select the most advantageous among all available speech enhancement techniques, and thus to dynamically adapt to the changing acoustic environment. The functioning of the proposed selective scheme does not depend on the number and the type of speech enhancement algorithms involved in the individual channels, but works better for dissimilar speech enhancement channels that in some way are complimentary to each other.

Figure 1

2.2. *Formulation of the channel selector*

During training, the input audio signals are parameterized, and then clustered by a GMM in multiple clusters, each one corresponding to a portion of the audio feature space with similar acoustic characteristics. During operation, the channel selector estimates the log-likelihood of a given audio input to belong to each of the existing clusters. These log-likelihoods are concatenated to form a new feature vector, which is fed as input to the machine learning technique that selects the speech enhancement channel, which is the most appropriate to the specific audio input.

The entire process can be formalized as follows: Let us define a set of observation vector sequences $O=\left\{O^{j}\right\}$, $1\le j\le J$, where the observation vector sequence for the *j*-th audio input is $O^{j}=o_{1}^{j}o_{2}^{j}...o_{T_j}^{j}$, $1\le i\le T_{j}$ and the *i*-th observation vector (audio feature vector) of the *j*-th audio input is $o_{i}^{j}\in\mathbb{R}^{D}$. The observation sequences correspond to J audio inputs, which form the training dataset. This set of observation sequences, O, is utilized to train a set of M D-dimensional Gaussian distributions, and thus, the resulting Gaussian mixture model (GMM), Υ, clusters the feature space $\mathbb{R}^{D}$ to M clusters, each modelled by a GMM component $\mathrm{N}\left(.\right)$, i.e.

$$\Upsilon\left(O^{j}\right)=\sum_{m=1}^{M}c_{m}\mathrm{N}\left(O^{j},\mu_{m},\Sigma_{m}\right)=\prod_{i=1}^{T_j}\left(\sum_{m=1}^{M}c_{m}\mathrm{N}\left(o_{i}^{j},\mu_{m},\Sigma_{m}\right)\right), \qquad (1)$$

subject to $\sum_{m=1}^{M} c_m = 1$, with $0 \le c_m \le 1$ and $1 \le m \le M$. Here, the expression $\mathrm{N}\left(\cdot, \mu_m, \Sigma_m\right)$ stands for the m-th Gaussian density (m-th GMM component), which is one of the clusters in the feature space, with mean $\mu_m \in \mathbb{R}^{D}$ and covariance matrix $\Sigma_m \in \mathbb{R}^{D \times D}$. The training of the GMM, Υ, is performed using the Baum-Welch algorithm [29]. In brief, during training, the GMM is initially modelled by a single Gaussian component and, afterwards, each existing component is gradually being split to two new components that are retrained, resulting to $M = 2^{k-1}$ Gaussian distributions after k training steps, $1 \le k \le K$, including the initial step. In each step, the models are iteratively retrained through the Baum-Welch algorithm until the convergence ratio between two successive iterations reaches a predefined threshold.

After k training steps, k GMM models, $\Upsilon^{(k)}$, have been constructed, each one consisting of $M^{(k)} = 2^{k-1}$ components. Each set of components $\mathrm{N}^{(k)}\left(\cdot, \mu_{m_k}, \Sigma_{m_k}\right)$, $1 \le m_k \le M^{(k)}$, divides the feature space into $M^{(k)}$ clusters, independently from each other. For the case of the kth GMM model, $\Upsilon^{(k)}$, for every audio input we compute the normalized log-likelihood of belonging to each of the $M^{(k)}$ clusters, as

$$P^{(k)}\left(O^j\right) = \left\{ p_{m_k}^{(k)}\left(O^j\right) \right\}. \tag{2}$$

with $P^{(k)} \in \mathbb{R}^{M^{(k)}}$ and

$$p_{m_k}^{(k)}\left(O^j\right) = \frac{L\left(O^j \mid \mathrm{N}^{(k)}\left(O^j, \mu_{m_k}, \Sigma_{m_k}\right)\right)}{\sum_{m_k=1}^{M^{(k)}} L\left(O^j \mid \mathrm{N}^{(k)}\left(O^j, \mu_{m_k}, \Sigma_{m_k}\right)\right)}. \tag{3}$$

where $L\left(O^j \mid \mathrm{N}^{(k)}\left(O^j, \mu_{m_k}, \Sigma_{m_k}\right)\right)$, is the log-likelihood of the input speech utterance O^j, given the cluster m_k.

The normalized log-likelihood vectors, $P^{(k)}$, $1 \le k \le K$, one for each of the K sets of components, are afterwards concatenated to a single vector, V, expressed as

$$V\left(O^j\right) = \left\{ P^{(k)}\left(O^j\right) \right\}, \text{ with } V\left(O^j\right) \in \mathbb{R}^{M_{TOT}}, \ M_{TOT} = \sum_{k=1}^{K} M^{(k)}, \tag{4}$$

which acts as input to a mapping function, f. Through this concatenated vector, the mapping function receives information about the environmental audio conditions at different levels of detail (depending on the number of clusters), and thus, any complementary information, which the different sets of clusters may contain, is used to the best advantage. The mapping function f selects the most appropriate speech enhancement method to be applied to the audio input O^j, i.e.

$$f\left(V\left(O^j\right)\right) = b, \tag{5}$$

where b, $1 \le b \le B$, is the index of the bth speech enhancement method, from a set of B speech enhancement channels that are available. The training of the mapping function f is performed on a bootstrap set of audio inputs, on which the speech recognition performance of the B existing enhancement methods has been evaluated in terms of WRR. The WRR is estimated individually for each speech enhancement algorithm, and, following, each audio sequence is linked to the speech enhancement channel that offered the highest WRR. When there are two or more speech enhancement channels which lead to the best WRR for that audio input, the enhancement channel with the highest average accuracy on the training dataset is selected. Once the training of the mapping function is completed the speech front-end is ready for operation.

3. The MoveOn speech and noise database

For the purpose of research and technology development in the MoveOn project, a dedicated speech database was recorded in environment typical for a motorcycle on-the-move [9]. Specifically, a group of thirty professional motorcyclists, members of the operational police force of UK, was recruited. While performing patrolling activates through the streets and suburbs of Birmingham, 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 earplugs.

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, emotion/affect 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. Experimental Setup

The context-adaptive speech front-end proposed in Section 2 was evaluated in different experimental setups: single channel and multiple channel speech enhancement, different number of clusters in the GMM, as well as for various channel mapping methods. In the following we describe in detail the settings of the experimental setup, the mapping algorithms and the experimental protocol of the present evaluation.

4.1. *Speech enhancement techniques*

In order to validate the practical usefulness of the proposed context-adaptive speech enhancement scheme (Section 2), we utilized a number of well-known and recent speech enhancement algorithms, which serve as independent channels in the speech front-end scheme discussed so far. A brief outline of these speech enhancement algorithms follows:

Spectral subtraction: The spectral subtraction (SPECSUB) algorithm [15] is a well-known technique, which 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. 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. Due to its well-known performance, the spectral subtraction algorithm serves here as an intuitive reference point.

Spectral subtraction with noise estimation: The spectral subtraction with noise estimation (SPECSUB-NE) method [16] 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.

Multi-band spectral subtraction: The multi-band spectral subtraction method (MBSS) introduced in [17] 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, 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 [17] as well as our previous experience with the MoveOn data [30] motivated us to consider this method.

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 [18], which we refer to as (MMSE-logSAE), 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 [30], it was observed that this method offers very good performance with respect to perceived enhancement on the MoveOn data, and therefore it is considered a strong candidate for achieving good speech recognition results.

Speech enhancement based on perceptually motivated Bayesian estimators of the speech magnitude spectrum: This method [19], which we refer to as SE-PMBE, utilizes Bayesian estimators of the short-time spectral magnitude of speech based on perceptually motivated cost functions. As it was demonstrated in [19], the estimators which implicitly take into account the auditory masking effect perform better in terms of having less residual noise and better speech quality, when compared to the MMSE-logSAE method. We consider this method due to its relatively good performance in [30], but also because we were interested to investigate if this advantage, when compared to MMSE-logSAE, will contribute for better speech recognition performance.

Subspace algorithm with embedded pre-whitening: The subspace algorithm with embedded pre-whitening (KLT) proposed in [21] 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. Although babble noise is not the primary interference in the present application domain, we consider this method since it offers complimentary features, which might be beneficial.

Perceptually-motivated subspace algorithm: The perceptually-motivated subspace algorithm (PKLT) introduced in [22] 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 yields a better shaping of the residual noise. This method was reported to outperform the KLT method.

Wiener algorithm based on wavelet thresholding multi-taper spectra: This algorithm, referred here as Wiener-WT, uses low-variance spectral estimators based on wavelet thresholding the multi-taper spectra [20]. Reported listening tests had shown that this method suppressed the musical noise and yielded better speech quality than the KLT, PKLT and MMSE-logSAE algorithms. Based on these reports, we expect that the Wiener-WT algorithm will be a strong candidate for the best performance.

All speech enhancement methods used here were implemented as in [31].

4.2. *The GMM-based clustering and the mapping function*

In the experimental evaluation, presented in Section 5, we study various design solutions for the two-stage channel selector discussed in Section 2. Specifically, we evaluated various configurations of the first-stage GMM-based clustering of the audio feature space by training a mixture model of 1 up to 64 continuous Gaussian distributions, using the HTK toolkit [32]. The audio feature space was parameterized by the same audio descriptors used in the speech recognition engine, i.e. the frame energy and 12 first Mel frequency cepstral coefficients with their first and second time derivatives appended (a detailed description of the feature extraction process is offered in Section 4.3). All GMMs were trained through the Baum-Welch algorithm [29], with training stop criterion (aka as convergence ratio) equal to 0.001. For the implementation of the second-stage mapping function we examined various classification algorithms:

(i) the support vector machines utilizing the sequential minimal optimization algorithm, here referred as to *SMO* [33],

(ii) the normalized Gaussian radial basis function network, here referred as to *RBFNetwork* [34],

(iii) the classification using regression methods, here referred as to *ClassificationViaRegression* [35],

(iv) the multi-layer perceptron neural network, here referred as to *MLP* [36],

(v) the bagging with fast decision tree learner, here referred as to *BaggingREPTree* [37],

(vi) the C4.5 decision tree, here referred as to *J48* [38],

(vii) the bagging with C4.5 decision tree, here referred as to *Bagging-J48* [37],

(viii) the Adaboost M1 method using the C4.5 decision tree, here referred as to *AdaboostM1-J48* [39], and

(ix) the *k*-nearest neighbours classifier, here referred as to *IBk* [40].

In all implementations of the mapping function, the input vector is as described in equation (4), i.e. the concatenated normalized log-likelihoods of each GMM cluster. Each of the nine evaluated classification algorithms, listed above, mapped every input audio instance, which is described in terms of the vector of equation (4), to one of the speech enhancement techniques described in Section 4.1. The selection of the most appropriate speech enhancement channel, in terms of word error rate deduction, will result to improvement of the speech recognition performance. In the evaluation of these implementations of the mapping function, we relied on their implementation in the WEKA machine learning toolkit [34]. All algorithms were evaluated for different sets of GMM clusters in the first-stage, i.e. for 4 (k=3), 8 (k=4), 16 (k=5), 32 (k=6) and 64 clusters (k=7).

4.3. *The speech recognition engine*

The speech recognition engine was configured with channel-specific acoustic models, each one of them adapted on speech data processed by the specific speech enhancement method of the corresponding channel. Furthermore a language model, which was common for all speech enhancement channels, was constructed.

In detail, the channel-specific acoustic models were obtained by means of maximum a posteriori [41] 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 [42]. This acoustic model consists of one context-dependent HMM model for each phone of the British SAMPA alphabet, [43], 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 first twelve Mel frequency cepstral coefficients (MFCC), i.e. the 0th one excluded. The MFCC implementation relied on a filter-bank of 26 filters [32]. The speech feature vector consisted of 39 parameters, which are the 13 static parameters outlined above together with their first- and second-order time-derivatives. All HMMs were trained through the Baum-Welch algorithm [29], with convergence ratio equal to 0.001.

The common language model was built using the annotations of the MoveOn speech and noise database. Specifically, the transcriptions of the responses of the MoveOn end-users, provided in [26], were used 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.

Finally, in order to obtain a reliable estimation of the accuracy of the various setups evaluated here, we followed a 10-fold cross validation protocol. In all experiments the speech decoder utilized the channel-specific acoustic models, adapted to the specific speech enhancement method, as described above. The speech recognition performance was tested both for bi-gram and tri-gram word-level language models. In all experiments the speech recognition performance was estimated in terms of word recognition rate (WRR).

5. Experimental Results

Following the experimental setup described in Section 4, we performed an experimental evaluation of the context-adaptive speech enhancement scheme proposed in Section 2. Firstly in Section 5.1 we present the evaluation of the performance of the individual speech enhancement methods, and afterwards in Section 5.2, we investigate the performance of the proposed context-adaptive scheme, for different settings of the GMM-based clustering and for different implementations of the mapping function. In all experiments, the performance is evaluated in terms of speech recognition accuracy, and is compared with the performance of the baseline. In section 5.1, as the baseline we consider the speech recognition accuracy obtained for an acoustic model adapted to the application domain and the operational acoustic environment but without speech enhancement pre-processing, and thus is referred to as “No enhancement”. In Section 5.2 the baseline is the speech recognition accuracy, obtained for the best individual speech enhancement method, discussed in Section 5.1.

5.1. *Performance of the individual speech enhancement methods*

As a first step, we examined the speech recognition accuracy obtained when the pre-processing of the input is performed by one of the speech enhancement methods described in Section 4.2 independently from the other methods. The speech recognition accuracy for both bi-gram and tri-gram language models, in terms of WRR are shown in Table 1. These results were obtained for the subset of MoveOn recordings corresponding to the outdoor scenario: motorcycle on-the-move. The notations of the speech enhancement methods in the first column of the table stand for: Speech enhancement based on perceptually motivated Bayesian estimators (SE-PMBE) [19], multi-band spectral subtraction (MBSS) [17], speech enhancement using a minimum mean square error log-spectral amplitude estimator (MMSE-logSAE) [18], spectral subtraction with noise estimation (SPECSUB-NE) [16], the original spectral subtraction (SPECSUB) [15], subspace enhancement with embedded pre-whitening (KLT) [21], Wiener algorithm based on wavelet thresholding multi-taper spectra (Wiener-WT) [20], perceptually-motivated subspace enhancement (PKLT) [22], and a speech recognizer without speech enhancement (“No enhancement”), where the last is considered as the baseline. In all cases adapted acoustic models were used.

As can be seen in Table 1, the highest speech recognition accuracy was achieved for the SE-PMBE enhancement method. 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, i.e. “No Enhancement”. The MBSS was ranked as the second-best speech enhancement method, since it achieved slightly lower speech recognition accuracy, but still improved the baseline performance by 2.35%. Next the speech recognition accuracy obtained for the Log-MMSE, SPECSUB-NE and SPECSUB methods was somehow lower but they still offer some advantage when compared to the baseline. Finally, speech recognition accuracy obtained for the PKLT, KLT and Wiener-WT methods was significantly inferior when compared to the use of adapted acoustic model without speech enhancement.

TABLE 1

As Table 1 presents, the speech recognition accuracy for the bi-gram language model was always better than the one for the tri-gram language model. This result can be explained by the limited amount of data 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. Thus, in the remaining of this work we report only speech recognition results obtained with the bi-gram language model.

In order to investigate the statistical significance among the different WRRs, reported in the second column of Table 1, we performed the Wilcoxon signed-rank test [44]. The highlighted cells correspond to recognition results, which are not statistically different, i.e. the speech enhancement with the Log-MMSE, SPECSUB-NE and SPECSUB methods do not differ significantly in terms of average speech recognition accuracy. However, the speech recognition accuracy for the best performing method, SE-PMBE, is statistically different from the one of the remaining methods.

In order to illustrate the importance of using the appropriate speech enhancement method for each type of noisy condition we computed the number of speech utterances for which each speech enhancement method outperformed the others. In Table 2 we present these results with respect to the 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 2

As Table 2 shows, the SE-PMBE method is not always the best-performing method, although it leads to the highest WRR on average (refer to Table 1). In Table 2 we can see that for more than 20 % of all cases, i.e. for 2273 utterances out of the 10201, the SE-PMBE speech enhancement algorithm was outperformed by one or more of the other algorithms. The last is in support of the assumption that the use of a number of dissimilar speech enhancement algorithms, which are selected dynamically depending on the input, could be beneficial in terms of overall improvement of speech recognition accuracy.

5.2. *Performance for the proposed context-adaptive speech pre-processing scheme*

Here we report results from the evaluation of the context-adaptive channel selector. Specifically, we investigated the accuracy of selecting the most appropriate speech enhancement method for various implementation of the mapping function, for different size of the input vector $\|V\|$=4, 12, 28, 60 and 124. The sizes of the input vector are obtained stacking up the outputs of GMMs with 4 (k=3), 8 (k=4), 16 (k=5), 32 (k=6) and 64 clusters (k=7). The results are shown in Table 3 in percentages, while the corresponding speech recognition accuracy in terms of WRR is presented in Table 4. In both tables, the best accuracy is indicated in bold. Our baseline is the speech recognition accuracy for the best in average speech enhancement method, SE-PMBE (Tables 1 and 2).

TABLE 3 & **TABLE 4**

As can be seen in both Tables 3 and 4, the highest accuracy was obtained when the mapping function was implemented via the *SMO* method [33]. It achieved the highest performance both in terms of detection accuracy and in terms of WRR. The *SMO* selector offered WRR equal to 90.0%, which is 1.1% higher than the second-best performing selector based on *RBFNetwork* [34] and 3.3% than the baseline, i.e. the best performing single enhancement channel (SE-PMBE). The *ClassificationViaRegression* [35], *MLP* [36] and *BaggingREPTree* [37] algorithms also improved the

WRR, when compared to baseline. On the other hand, the *J48* [38], *BaggingJ48* [39], *AdaboostM1-J48* [40] and *IBk* [40] methods did not offer any advantage when compared to the baseline.

Table 4 shows that most of the evaluated implementations of the mapping function achieved their highest classification accuracy as well as their highest word recognition rates for feature vector $\|V\|$=28, which consists of the log-likelihoods of GMMs with 4, 8 and 16 clusters, i.e. for k=5, Exceptions here are the *MLP* and *AdaboostM1-J48* mapping functions, which offered their highest accuracy for k=4 ($\|V\|$=12) and k=6 ($\|V\|$=60) respectively, however not significantly higher than these for the case k=5 ($\|V\|$=28). These results indicate that the clustering of the audio conditions characteristic for the motorcycle-on-the-move, as captured in the MoveOn database, can be modelled well with relatively small number of clusters. The use of fewer clusters, i.e. k<5, reduces the accuracy due to lack of capacity to distinguish among different types of interference. On the other hand, the use of more clusters, i.e. k>5, makes the mapping less consistent, and also significantly increases the dimensionality of the input vector which imposes difficulties related to the curse of dimensionality.

Summarizing the results discussed in Section 5.2, we conclude that the proposed context-adaptive pre-processing scheme for robust speech recognition is of significant practical value since it offers a significant advantage over the best performing single speech enhancement method.

6. Conclusion

Speech interaction between a motorcycle driver and a spoken dialogue system is often required for the needs of professional information support (as in police force operations) or for entertainment (web-access, control of music players or other personal devices, etc). The main difficulties for guaranteeing robust spoken dialogue interaction in these conditions are due to both the fast-varying noise environment and the changes in the properties of speech signal because of the body stress, the vibrations, and the increased cognitive load during driving. In the present work we focused on the negative effects caused by the interferences from the environment. Specifically, we studied a context-adaptive pre-processing scheme for speech enhancement which is part of a speech front-end operating in fast-varying open-air environment. The proposed scheme relies on a GMM-based clustering of the audio inputs and a mapping function that depending on the interferences in the input audio selects the most appropriate speech enhancement technique, among all available. It is worth mentioning that in the two-stage process proposed here, no explicit recognition of a certain named acoustic condition is performed for any given audio input. The clustering of the audio feature space is obtained in an unsupervised manner, and thus, the number of clusters does not correspond to distinct acoustic conditions. The mapping between the cluster-based feature vector and the most appropriate speech enhancement method is learned based on the observed speech recognition accuracy that each speech enhancement method demonstrated for each audio input. Thus, the estimation of the appropriateness of each of the candidate speech enhancement methods, for each particular setup, is obtained as a result of their performance on the acoustic condition characteristics of the application of interest. The entire training process is data-driven, and no prior knowledge is required for linking the available speech enhancement methods to specific audio conditions, and no prior labelling of the audio conditions is needed.

This proposed context-adaptive speech pre-processing scheme was found beneficial, since it offers an improved performance when compared to the best on average individual speech enhancement method. In the experimental validation of the proposed method, we observed that the highest performance of the proposed speech front-end is obtained when the first-stage of the channel selector is implemented by concatenation of the outputs of three sets of GMMs, with 4, 8 and 16 clusters, and when the second-stage mapping function is implemented via support vector machines.

Furthermore, the proposed context-adaptive speech pre-processing scheme is less demanding from computational perspective, when compared to methods proposed in earlier work [24, 27]. In addition, the proposed scheme has a generic nature, given by the independence with respect to the employed speech enhancement algorithms, and can be adapted straightforwardly to different applications. Thus, we deem that the usefulness of the proposed speech front-end goes beyond the limits of the MoveOn application, and that it can be successfully used in a number of applications which are characterized with fast-varying noise conditions, such as outdoor/street environment or open-air (open cabin or no cabin) vehicles, etc.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable comments, which significantly improved the quality of this article. The research that led to the results reported here was financially supported by the MoveOn project (IST-2005-034753), co-funded by the European Community under the Sixth Framework Programme (FP6).

References

[1] S. Bohm, J. Koolwaaij, M. Luther, B. Souville, M. Wagner, M., Wibbels, Introducing IYOUIT, in: Proc. of the International Semantic Web Conference (ISWC’08), vol. 5318 of LNCS, Springer Verlag, 2008, pp. 804-817.

[2] U. Gartner, W. Konig, T. Wittig, Evaluation of manual vs. speech input when using a driver information system in real traffic, in: Driving Assessment 2001: The First International Driving Symposium on Human Factors in Driver Assessment, Training and Vechicle Design, CO. USA, 2001, pp. 7-13.

[3] A. Berton, D. Buhler, W. Minker, SmartKom-Mobile Car: User interaction with mobile services in a car environment, in: W. Wahlster (Ed.), SmartKom: Foundations of Multimodal Dialogue Systems, Springer, 2006, pp. 523-537.

[4] I. Cohen, B. Berdugo, Speech enhancement for non-stationary noise environments, Signal Processing 81(11) (2001) 2403-2418.

[5] A. Moreno, B. Linderberg, C. Draxler, G. Richard, K. Choukri, S. Euler, J. Allen, SPEECHDAT-CAR: A large speech database for automotive environments, in: Proc. of LREC 2000. Athens, Greece, 2000.

[6] J.H.L. Hansen, X. Zhang, M. Akbacak, U. Yapanel, B. Pellom, 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, Nagoya, Japan, 2003, pp. 19-45.

[7] B. Lee, M. Hasegawa-Johnson, C. Goudeseune, AVICAR: Audio-visual speech corpus in a car environment, in: Proc. of ICSLP 2004, Jeju Island, Korea, 2004, pp. 2489-2492.

[8] M. Kaiser, H. Mogele, F. Shiel, Bikers accessing the web: The SmartWeb motorbike corpus, in: Proc. of LREC 2006, Genoa, Italy, 2006.

[9] T. Winkler, T. Kostoulas, R. Adderley, C. Bonkowski, T. Ganchev, J. Kohler, N. Fakotakis, The MoveOn motorcycle speech corpus, in: Proc. of LREC 2008. Marrakech, Morocco, 2008.

[10] J.H.L. Hansen, M.A. Clements, Constrained iterative speech enhancement with application to speech recognition, IEEE Transactions on Audio, Speech and Signal Processing 39 (4) (1991) 795-805.

[11] P. Lockwood, J. Boundy, Experiments with a nonlinear spectral subtractor (NSS), HMMs and the projection, for robust speech recognition in cars, Speech Communication 11 (2-3) (1992) 215-228.

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

[13] N. Dal Degan, C. Prati, Acoustic noise analysis and speech enhancement techniques for mobile radio applications, Signal Processing 15 (1) (1988) 43-56.

[14] J. Li, S. Sakamoto, S. Hongo, M. Akagi, Y. Suzuki, Adaptive β-order generalized spectral subtraction for speech enhancement, Signal Processing 88 (11) (2008) 2764-2776.

[15] M. Berouti, R. Schwartz, J. Makhoul, Enhancement of speech corrupted by acoustic noise, in: Proc. of the IEEE ICASSP'79, Washington, DC, USA, 1979, pp. 208-211.

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

[17] S. Kamath, P. Loizou, P., A multi-band spectral subtraction method for enhancing speech corrupted by colored noise, in: Proc. of ICASSP-2002, Orlando, USA, vol. 4, 2002, pp. 4164-4167.

[18] Y. Ephraim, D. Malah, Speech enhancement using a minimum mean square error log-spectral amplitude estimator, IEEE Transactions on Acoustics, Speech, Signal Processing, 33 (1985) 443-445.

[19] P. Loizou, Speech enhancement based on perceptually motivated Bayesian estimators of the speech magnitude spectrum, IEEE Transactions on Speech and Audio Processing 13 (5) (2005) 857-869.

[20] Y. Hu, P. Loizou, Speech enhancement by wavelet thresholding the multitaper spectrum, IEEE Transactions on Speech and Audio Processing 12 (1) (2004) 59-67.

[21] Y. Hu, P. Loizou, A generalized subspace approach for enhancing speech corrupted by coloured noise, IEEE Trans. on Speech and Audio Processing 11 (2003) 334-341.

[22] F. Jabloun, B. Champagne, Incorporating the human hearing properties in the signal subspace approach for speech enhancement, IEEE Transactions on Speech and Audio Processing 11 (6) (2003) 700-708.

[23] S. Gannot, D. Burshtein, E. Weinstein, Iterative and sequential Kalman filter-based speech enhancement algorithms, IEEE Transactions on Speech and Audio Processing 6 (4) (1998) 373-385.

[24] I. Mporas, O. Kocsis, T. Ganchev, N. Fakotakis, Robust speech interaction in motorcycle environment, Expert Systems with Applications 37 (3) (2010) 1827-1835.

[25] B. Sarama, A. Khan, Refined Detailed Technical Specification of MoveOn System (D14), url: http://www.m0ve0n.net, 2008.

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

[27] V. Krishnan, P.S. Whitehead, D.V. Anderson, M.A. Clements, A framework for estimation of clean speech by fusion of outputs from multiple speech enhancement systems, in: Proc. of Interspeech-2005, 2005, pp. 2317-2320.

[28] Y. Hu, P. Loizou P., Subjective evaluation and comparison of speech enhancement algorithms, Speech Communication 49 (2007) 588-601.

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

[30] S. Ntalampiras, T. Ganchev, I. Potamitis, 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), F. Makedon, L. Baillie, G. Pantziou, and I. Maglogiannis (Eds.), PETRA-2008, vol. 282. ACM, New York, NY, 1-5, 2008.

[31] P. Loizou, Speech Enhancement: Theory and Practice. CRC Press, 2007.

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

[33] S.S. Keerthi, S.S., S.K. Shevade, C. Bhattacharyya, K.R.K. Murthy, Improvements to Platt's SMO algorithm for SVM classifier design, Neural Computation 13 (3) (2001) 637-649.

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

[35] E. Frank, Y. Wang, S. Inglis, G. Holmes, I.H. Witten, Using model trees for classification. Machine Learning. 32 (1) (1998) 63-76.

[36] T.M. Mitchell, Machine Learning, McGraw-Hill International Editions, 1997.

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

[38] R. Quinlan, C4.5: Programs for machine learning, Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[39] Y. Freund, R.E. Schapire, Experiments with a new boosting algorithm, in: Proc. of the 13th International Conference on Machine Learning, San Francisco, USA, 1996, pp. 148-156.

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

[41] J.L. Gauvain, C. Lee, Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains, IEEE Transactions on Speech and Audio Processing 2 (1994) 291-299.

[42] H. Hoge, C. Draxler, H. Van den Heuvel, F.T. Johansen, E. Sanders, H.S. Tropf, SpeechDat multilingual speech databases for teleservices: Across the finish line, in: Proc. of Eurospeech 1999, Budapest, Hungary, 1999, pp. 2699-2702.

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

[44] F. Wilcoxon, Individual comparisons by ranking methods, J. Biometrics 1 (1945) 80-83.

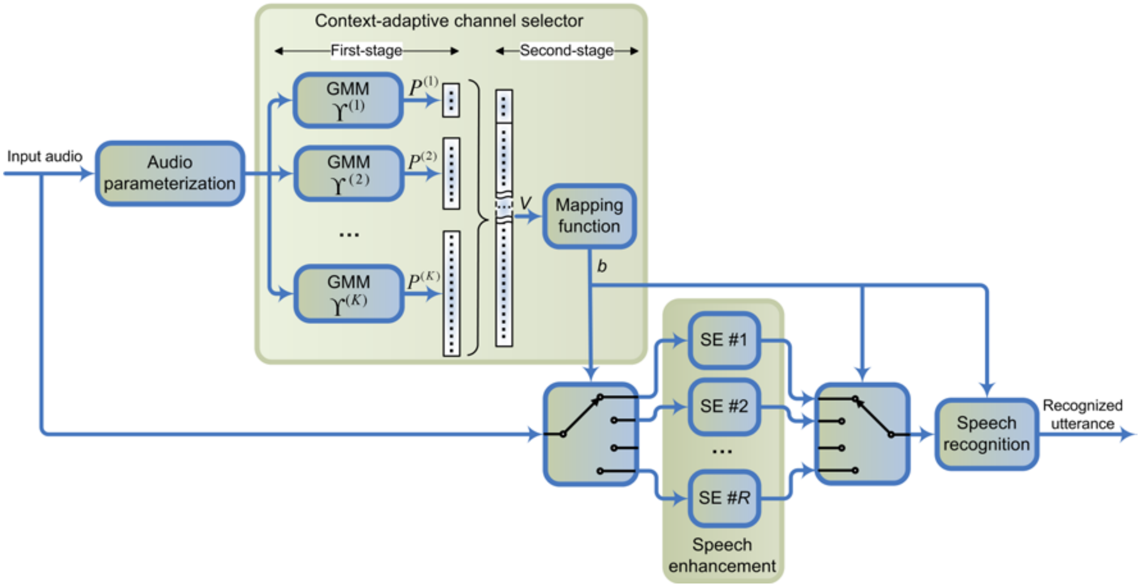

Fig. 1. Block diagram of the context-adaptive speech front-end with GMM-based clustering of the feature space in the first stage and mapping function for speech enhancement channel selection in the second stage.

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

Speech enhancement method	***bi-gram LM***	***tri-gram LM***
SE-PMBE	86.7	76.1
MBSS	85.7	75.1
MMSE-logSAE	84.9	74.2
SPECSUB-NE	84.9	74.4
SPECSUB	84.8	75.5
KLT	82.4	70.3
Wiener-WT	81.3	70.3
PKLT	77.9	65.5
No enhancement	83.4	71.6

Table 2. Number of speech utterances for which the specific speech enhancement method led to the highest word recognition rate (WRR).

Speech enhancement method	***# best-cases***	***best-case in %***
SE-PMBE	7928	77.72
MBSS	1116	10.94
MMSE-logSAE	534	5.23
SPECSUB-NE	232	2.27
SPECSUB	179	1.75
KLT	91	0.89
Wiener-WT	57	0.56
PKLT	42	0.41
No enhancement	22	0.22
Total	10201	100.00

Table 3. Accuracy of selecting the proper speech enhancement method (in percentages) for various implementations of the mapping function, for different size of the input vector $\|V\|$ = 4, 12, 28, 60, and 124.

Input vector length ***Mapping function***	$\|V\|$ ***=4***	$\|V\|$ ***=12***	$\|V\|$ ***=28***	$\|V\|$ ***=60***	$\|V\|$ ***=124***
SMO	77.3	80.0	**81.8**	80.7	79.7
RBFNetwork	76.3	79.7	**80.7**	80.1	78.7
ClassificationViaRegression	77.6	79.6	**80.6**	80.4	79.6
MLP	76.3	80.3	**80.4**	79.2	76.6
BaggingREPTree	78.4	79.2	**80.0**	79.6	78.3
J48	76.6	78.2	**79.0**	78.6	74.9
Bagging-J48	76.7	77.0	**78.8**	78.7	78.5
AdaboostM1-J48	74.1	74.1	75.2	75.2	**75.6**
IBk	65.9	66.1	**66.7**	66.1	66.2

Table 4. Word recognition rates (in percentages) for various implementations of the mapping function, for different size of the input vector $\|V\|$ = 4, 12, 28, 60, and 124.

Input vector length / ***Mapping function***	$\|V\|$ ***=4***	$\|V\|$ ***=12***	$\|V\|$ ***=28***	$\|V\|$ ***=60***	$\|V\|$ ***=124***
SMO	85.2	88.4	**90.0**	88.8	87.2
RBFNetwork	83.9	87.2	**88.9**	88.0	86.3
ClassificationViaRegression	85.4	87.1	**88.5**	88.0	87.9
MLP	84.0	**88.3**	88.2	87.3	84.1
BaggingREPTree	86.2	87.3	**88.1**	87.0	86.4
J48	84.3	86.1	**86.2**	86.0	82.5
BaggingJ48	84.2	84.6	**86.1**	86.0	85.1
AdaboostM1-J48	81.2	81.1	82.3	**82.5**	82.2
IBk	72.4	72.5	**73.5**	72.3	72.1