DYNAMIC SELECTION OF A SPEECH ENHANCEMENT METHOD FOR ROBUST SPEECH RECOGNITION IN MOVING MOTORCYCLE ENVIRONMENT

Iosif Mporas, Todor Ganchev, Otilia Kocsis, Nikos Fakotakis

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

ABSTRACT
We present a speech pre-processing scheme (SPPS) for robust speech recognition in the moving motorcycle environment. The SPPS is dynamically adapted during the run-time operation of the speech front-end, depending on short-time characteristics of the acoustic environment. In detail, the fast varying acoustic environment is modeled by GMM clusters based on which a selection function determines the speech enhancement method to be applied. The correspondence between input audio and speech enhancement method is learned during the training of the selection function. The SPPS was found to outperform the best performing speech enhancement method by approximately 3.3% in terms of word recognition rate (WRR).

Index Terms— speech enhancement, acoustic noise, speech recognition, machine learning

1. INTRODUCTION
During the last years several technological applications have moved to the mobile world. A number of activities, which in the past were taking place indoors (office or home), have now also been extended outdoors. The new environment (outdoors) is increasing the demand for services which will provide efficiency, usability and safety, since most of the outdoors activities, like driving, are performed in parallel. Since the distraction of drivers is a matter of safety, the development of efficient human-machine interaction interface is crucial for the passengers’ safety.

When driving, the interaction of the driver with mobile/portable devices must be safe. Previous studies on acquiring certain information on the route with speech commands showed that the driver’s distraction is smaller and road safety is improved when voice is used. Moreover, when the car is moving, the driver should interact with devices only through voice since other communication modalities, as gestures or graphical interfaces are quite distracting [1].

In contrast to controlled environments, the performance of speech-based interfaces degrades substantially in the non-stationary environment. In the acoustic environment characteristics of a motorcycle on the move this is owed to the presence of additive interferences (e.g. rumble noises by road vibrations, friction from the tires-road contact, several mechanical noises, etc) as well as from variations in the driver’s voice and speaking style due to task-stress, distributed attention, physical effort, body vibrations, etc. In this work, we focus on the additive interferences coming from the acoustic environment.

In previous efforts for outdoors speech recognition, investment on both the collection of recordings from the same domain and noise-suppression methods has been done. Several speech databases, which are representative for a set of mobile voice-interaction applications, have been designed, recorded and annotated, for the car environment [2], as well as for the motorcycle on the move environment [3]. Except for the adaptation of the ASR acoustic models with domain-specific data, the ASR accuracy can be improved with speech enhancement pre-processing [4]. The speech enhancement methods can briefly be categorized in spectral subtractive algorithms [5], statistical model-based approaches [6, 7], signal subspace approaches [8], and enhancement methods based on a special type of filtering [9]. All these methods, although achieving high performance within specific noise conditions, reduce their performance in highly non-stationary and fast-varying acoustic environmental conditions, such as the characteristics for a moving motorcycle. In previous work [10], it was shown that a collaborative speech enhancement structure, which relies on parallel speech enhancement channels operating on a common input, offers advantage over the best individual speech enhancement method when used alone.

The SPPS is based on the conclusion made in [11] that dissimilar speech enhancement algorithms perform differently for dissimilar types of interference and noise conditions. Based on this we expect that the dynamic selection of the best-performing speech enhancement method for each input, will lead to increase of the overall speech recognition
accuracy. This approach employed in the present SPPS alleviates the scalability constraints of earlier designs [10, 12], where several speech enhancement algorithms operate in parallel for each audio input.

2. THE PRE-PROCESSING SCHEME

In the moving motorcycle environment, the captured input audio signal consists of the speech message together with additive interferences, which change rapidly both in time and spectral domain. In command and control applications, such as the MoveOn system [3], the user’s inputs are typically shorter than 3 seconds. For such durations we can assume that the environmental acoustic characteristics do not vary significantly, and therefore the selection of an appropriate speech enhancement algorithm for that environmental condition would contribute to the increase the overall speech recognition accuracy.

The block diagram of the dynamic SPPS studied here is illustrated in Fig. 1. As can be seen in the figure the incoming audio samples form frames, using a sliding Hamming window, and for each frame a vector consisting of a number of audio parametric features is computed. The sequence of feature vectors, corresponding to a certain audio input is next compared against a set of clusters which represent the acoustic space for the specific application and the log-likelihood of this sequence to belong to each cluster is computed. The computed log-likelihoods constitute a new feature vector, which is fed to a machine learning algorithm (denoted as selection function in Fig. 1), which maps the audio input to the most appropriate speech enhancement algorithm, among a number of available ones. The last requires that all candidate speech enhancement methods are evaluated on the training dataset in advance. Once the most appropriate speech enhancement algorithm for each audio input from the train dataset is learned, the speech front-end is ready for operation.

2.1. The GMM-based environment models

In order to model the acoustic environment space we performed unsupervised clustering of the feature space with Gaussian mixture models (GMMs). In detail, utilizing a sufficient number of recordings, we trained several GMMs with \( C = 2^n \) number of components. As a result each component corresponds to a cluster of the acoustic environment space, with similar acoustic characteristics, however no distinct environmental condition is named. The number of components affects the degree of detail in clustering. Indeed, increasing the number of components results to acoustic clusters with more similar acoustic characteristics, thus to a more detailed description of each acoustic sub-space. On the other hand, due to the restricted amount of training data and due to the nature of the environment (fast varying superposed interferences), the use of few components results to more robust cluster models, however with greater intra-cluster variance. In order to overcome this, we compute the log-likelihoods of GMMs with different levels of detail, i.e. \( \{C = 2^2, C = 2^3, \ldots, C = 2^{N/2}, C = 2^N\} \) and append the normalized log-likelihoods of each level into a common vector.

2.2. The selection function

The SPPS does not attempt to identify certain named acoustic conditions. Instead, the set of GMM clusters generates a log-likelihood vector, which corresponds to the acoustic characteristics of the audio input. This vector is forwarded to a machine learning algorithm, which selects to the most appropriate speech enhancement algorithm, from a repository of available speech enhancement methods. The training of the selection function is performed on a bootstrap set of audio inputs, where the ground-truth label for each audio input is considered as the enhancement method with the highest word recognition rate (WRR) for that input.

Provided that a set of dissimilar speech enhancement methods are involved and that each of them is advantageous in specific noise conditions, we suppose that the overall WRR obtained with the SPPS will be superior, when compared to the one of best-individual speech enhancement method alone, especially when highly non-stationary and fast varying noise environments, like the moving motorcycle domain, are considered.

3. EXPERIMENTAL SETUP

The proposed scheme was evaluated using multiple speech enhancement algorithms, various GMM cluster numbers
and several implementation of the selection function. Comparison with the best-performing individual speech enhancement method was performed as well. The description of the settings of the experimental setup is described below.

### 3.1. The MoveOn application and database

Within the MoveOn project, a speech database recorded in the moving motorcycle environment was recorded for research and development purposes [3]. In brief, thirty professional motorcyclists, members of the operational police force of UK, were recruited. Each of the motorcyclists 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. Several types of motorbikes and helmets were employed, while the trace of the road differed among the sessions (in-city driving, highway, tunnels, suburbs, etc). The speech corpus consists of approximately 40 hours of recordings. The language of all recordings is British English spoken by native speakers.

### 3.2. Speech enhancement algorithms

The speech enhancement algorithms employed here are [13]: (i) the spectral subtraction (SPECSUB) algorithm, (ii) the spectral subtraction with noise estimation (SPECSUB-NE) method, (iii) the multi-band spectral subtraction method (MBSS), (iv) the minimum mean square error log-spectral amplitude estimator (MMSE-logSAE), (v) the enhancement based on perceptually motivated Bayesian estimators of the speech magnitude spectrum (SE-PMBE), (vi) the subspace algorithm with embedded pre-whitening (KLT), (vii) the perceptually-motivated subspace algorithm (PKLT), and (viii) the Wiener algorithm based on wavelet thresholding multi-taper spectra (Wiener-WT).

### 3.3. Implementations of the selection function

For the implementation of the selection function we examined several machine learning algorithms: (i) the support vector machines utilizing the sequential minimal optimization algorithm (SMO), (ii) the multi-layer perceptron neural network (MLP), (iii) the C4.5 decision tree (J48), (iv) the k-nearest neighbors classifier (IBk). The implementation of the selection functions was based on the WEKA machine learning toolkit [14].

### 3.4. The speech recognition engine

The speech recognition engine was implemented with the HTK toolkit [15]. A general-purpose British English acoustic model (AM), trained from telephone speech recordings of the SpeechDat(II)-FDB4000-British database [16], was used as basis to build one AM for each speech enhancement method with MAP adaptation. Three-state left-to-right HMMs without skipping transitions, modeled by a mixture of eight continuous Gaussian distributions per state, were used. The pre-processing of the speech signals, sampled at 8 kHz, consisted of frame blocking with length 25 milliseconds and step 10 milliseconds, and pre-emphasis filtering with coefficient equal to 0.97. The speech parameterization consisted of energy and MFCC [15] computation. The speech feature vector consisted of 39 parameters (13 static, 20 dynamic and 6 time-derivatives).

A common language model (LM) was built using the annotations of the MoveOn speech and noise database. Due to the limited size of the database we considered both bigram and tri-gram word models.

In all experiments we followed the 10-fold cross validation procedure.

### 4. EXPERIMENTAL RESULTS

Based on the experimental setup described in the previous section we examined the WRR performance of each enhancement method, as well as the WRR accuracy of the SPPS. In Table, we present the WRR in percentages for each of the speech enhancement methods considered here, for bi-gram and tri-gram LMs.

As can be seen in Table 1, the highest WRR score was achieved for the SE-PMBE enhancement method. For the case of speech recognition with bi-gram language model, the SE-PMBE method improved the WRR by approximately 3.3% when compared to the “No Enhancement”, which is considered as the baseline WRR accuracy. The higher performance of the bi-gram LMs is probably owed to the amount of data, which was not enough to estimate robustly tri-gram word probabilities. Although $SE$-PMBE was the best in average method, it achieved the highest WRR in only 77% of the 10201 utterances evaluated here. In the remaining 23% of the audio inputs, other speech enhancement methods achieved the highest WRR. This indicates the potential for improvement of the overall WRR, which a flexible speech enhancement scheme, such as the SPPS, has.

In Table 2 we present the WRR with bigram LMs for the SPPS, for several implementations of the selection function. As can be seen, the best performance was achieved when the selection function was implemented with the SMO algorithm and with size of the log-likelihood vector equal to 28, i.e. when the log-likelihoods for each input utterance were computed against 3 GMM models with 4, 8, and 16 clusters ($4+8+16=28$). Following the SMO, the MLP implementation with vector size 12 (i.e. $4+8=12$ clusters) achieved 88.3% in terms of WRR. Both the SMO and MLP implementations achieved higher speech recognition performance than the best-performing SE-PMBE enhancement method, when used alone. Next, the selection function implemented with the J48 algorithm with vector size 28, achieved accuracy of 86.2% in terms of WRR, which is
Table 1. Word recognition rates for different speech enhancement methods, for bi-gram and tri-gram language models.

<table>
<thead>
<tr>
<th>Speech enhancement method</th>
<th>bi-gram LM</th>
<th>tri-gram LM</th>
</tr>
</thead>
<tbody>
<tr>
<td>SE-PMBE</td>
<td>86.7</td>
<td>76.1</td>
</tr>
<tr>
<td>MBSS</td>
<td>85.7</td>
<td>75.1</td>
</tr>
<tr>
<td>MMSE-logSAE</td>
<td>84.9</td>
<td>74.2</td>
</tr>
<tr>
<td>SPECSUB-NE</td>
<td>84.9</td>
<td>74.4</td>
</tr>
<tr>
<td>SPECSUB</td>
<td>84.8</td>
<td>75.5</td>
</tr>
<tr>
<td>KLT</td>
<td>82.4</td>
<td>70.3</td>
</tr>
<tr>
<td>Wiener-WT</td>
<td>81.3</td>
<td>70.3</td>
</tr>
<tr>
<td>PKLT</td>
<td>77.9</td>
<td>65.5</td>
</tr>
<tr>
<td>No enhancement</td>
<td>83.4</td>
<td>71.6</td>
</tr>
</tbody>
</table>

Table 2. Word recognition rates (in percentages) for various implementations of the selection function, for different size of the log-likelihood vector.

<table>
<thead>
<tr>
<th>Selection function</th>
<th>4</th>
<th>12</th>
<th>28</th>
<th>60</th>
<th>124</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMO</td>
<td>85.2</td>
<td>88.4</td>
<td>90.0</td>
<td>88.8</td>
<td>87.2</td>
</tr>
<tr>
<td>MLP</td>
<td>84.0</td>
<td>88.3</td>
<td>88.2</td>
<td>87.3</td>
<td>84.1</td>
</tr>
<tr>
<td>J48</td>
<td>84.3</td>
<td>86.1</td>
<td>86.2</td>
<td>86.0</td>
<td>82.5</td>
</tr>
<tr>
<td>IBk</td>
<td>72.4</td>
<td>72.5</td>
<td>73.5</td>
<td>72.3</td>
<td>72.1</td>
</tr>
</tbody>
</table>

slightly worse than the accuracy of the SE-PMBE enhancement method. Finally, the selection function implemented with IBk algorithm did not manage to learn a beneficial mapping of the acoustic space to the best-performing speech enhancement method, which resulted in a significant reduction of the speech recognition accuracy.

The experimental results indicate that the use of a number of speech enhancement methods with a dynamic selection of the most appropriate one for each input improves the overall speech recognition performance, when compared to the individual best-performing speech enhancement method. It was shown that the SMO and MLP implementations of the selection function were able to learn a beneficial mapping between the unsupervised GMM clustering of the acoustic space and the best-performing speech enhancement method. Moreover, the GMM clustering with relatively low number of clusters (28) was able to provide a good description of the short-time acoustic condition characteristics for the fast varying environment, which is typical for a motorcycle on the move. Obviously, in different applications the optimal number of GMM clusters will depend on the variability and diversity of the existing interferences.

5. CONCLUSIONS

In acoustic environments characterized by fast varying interferences, such as the motorcycle on the move environment, the selection of a single speech enhancement method for the de-noising of the input signal is a suboptimal solution. In this paper, we presented a speech pre-processing scheme which dynamically selects the most appropriate among several available speech enhancement methods for each input utterance. This selection depends on the short-time environmental characteristics of the audio input. The experimental results on the MoveOn database showed an improvement of the speech recognition performance by 3.3% in terms of WRR. This improvement demonstrates the practical worth of the proposed dynamic speech pre-processing scheme in fast varying acoustic environments.

REFERENCES