COMPENSATION OF NOISE EFFECTS FOR
ROBUST SPEECH RECOGNITION IN CAR ENVIRONMENTS

Angel de la Torre\textsuperscript{1,2}, Dominique Fohr\textsuperscript{2}, Jean Paul Haton\textsuperscript{2}
\textsuperscript{1} Dpto. Electronica y Tecn. Comp., Universidad de Granada, 18071 Granada (Spain)
\textsuperscript{2} LORIA/INRIA, BP 239 - F54506 Vandoeuvre-les-Nancy (France)
e-mail atv@hal.ugr.es fohr@loria.fr jph@loria.fr

ABSTRACT

In this paper, we propose a novel method to compensate the effect of the noise for Automatic Speech Recognition in car environments. This method can be applied to recognizers using a standard MFCC front-end. We perform a channel-by-channel compensation of the noise effect in the filter-bank output domain. In a first stage, the parameters describing the noise are estimated and secondly, we estimate the expected value of the clean speech in a probabilistic framework. The compensated filter-bank outputs are then used to obtain a compensated version of the MFCC-based parameters representing the speech signal.

Recognition experiments using the French VODIS database (recorded in several cars running in real traffic situations) have been carried out to test the proposed compensation method. The results show the capability of the proposed method for the compensation of the noise effect in car environments.

1. INTRODUCTION

In most practical applications of Automatic Speech Recognition (ASR), the input speech is contaminated by a background noise, and this degrades significantly the performance of speech recognizers [1] [2]. The reduction of accuracy could make unpractical speech recognizers for real applications. For this reason, the design of an application including ASR must consider the environment in which the speech signal is acquired, and methods to adapt speech recognizers to noise conditions must be applied. Currently, an important research effort is dedicated to the compensation of the noise effects in ASR [1] [3] [4] [5] [6] [7].

The compensation of noise effects in car environments has attracted the interest, due to the amount of applications that could be developed using speech as input/output (like interaction with an on-board computer, dialing telephone numbers, or interacting with a remote information system) in a context in which hands and eyes are not free [8] [9] [10]. In car environments, the conditions in which the audio signal is acquired degrade significantly the quality of the input speech. The position of the microphone (that usually cannot be allocated close to the mouth of the speaker) and the contamination of the speech signal with sounds generated by other sources (like the engine, wheels, wind, other cars, radio, etc.) reduce the SNR, and, consequently, degrade the performance of the recognizers.

This work is dedicated to the compensation of the noise effect for ASR in the context of car environment. We analyze the effect of the noise in car environments over the representation of the speech signal and over the performance of automatic speech recognizers.

We also propose a novel compensation method for the effect of noise. The method can be applied to recognizers using a front-end based on Mel Frequency Cepstral Coefficients (MFCC) [11]. Based on the estimation of parameters describing the noise, a channel-by-channel compensation of the noise effect in the filter-bank outputs domain is performed. Recognition experiments to test the efficiency of the proposed method have been carried out. Experimental results are presented and discussed in this work.

2. THE NOISE IN CAR ENVIRONMENTS

Making use of the French VODIS database [12], we have analyzed how speech is affected by noise in car environments. This database has been recorded in car environments in different real traffic situations, and it includes a wide range of speakers, car models and driving conditions. Each sentence has been simultaneously recorded by two microphones, one situated close to the mouth of the speaker and the other at the windshield. Even though the quality of the signal recorded by the close-talk microphone is significantly better, for commercial applications using ASR in car environments the microphone must be usually allocated at the windshield.

The average SNR is 21 dB for the close-talk sentences and decreases to 11 dB when the signal is recorded with the far-talk microphone. The SNR varies in a wide range depending on the speaker, driving conditions, etc., taking sometimes values under 5 dB for the far-talk recordings. The analysis of the sentences in the database reveals that the speech signal is affected by several sources of distortion that could be classified as follows: (a) stationary or quasi-stationary additive noise (due to the engine, wheels, wind, air-conditioner, etc.); (b) non-stationary additive noise (due to the radio, people in the cabin, bumps, other cars, etc.); (c) reverberation (due to the resonances in the cabin); (d) convolutional noise (attenuation of different frequency bands due to the microphone response, its position, the orientation of the speaker, etc.); and (e) Lombard effect. The management of these sources of distortion with a fast variability is difficult because the estimation of the parameters describing the distortion source is not possible. For this reason, noise compensation methods in ASR are usually focused on the stationary (or quasi-stationary) part of the additive and the convolutional noises.

2.1. Effect of additive noise

If the speech signal samples $x_t$ are contaminated with an additive noise $n_t$, the noisy speech samples are $y_t = x_t + n_t$. If we assume that speech and noise are uncorrelated signals ($E[x_t n_t] = 0$), a parameter
representing an energy verifies that,
\[
E_y = \sum_{i=1}^{I} y_i^2 = \sum_{i=1}^{I} (x_i^2 + n_i^2 + 2x_in_i) = E_x + E_n
\]  \hspace{1cm} (1)

This result can be applied to the energy of the frames, and it can also be extended to the output power of each filter of the filter bank in standard MFCC based parameterizations: \(Y_b = X_b + N_b\), where \(b\) represents the channel index in the Filter Bank. The Filter Bank Outputs (or FBO parameters), utilized to compute the MFCC parameters, are the output power of each filter logarithmically scaled\(^1\): \(x_b = \log(X_b)\). Therefore, the effect of an additive noise in the FBO domain is described by the equation,
\[
y_b = \log[\exp(x_b) + \exp(n_b)]
\]  \hspace{1cm} (2)

So, the effect of an additive noise in the FBO domain consists of transform the FBO parameters according to equation (2). In figure (1.A), \(y_b\) is represented as a function of \(x_b\) for a noise of 20 dB. This transformation of the representation of the speech in the FBO domain introduces a mismatch between the training (clean) and the recognition (noisy) conditions which degrades the performance of ASR. In this figure, a constant value was considered for the noise. In this case, a simple adaptation based on the estimation of the noise \(n_b\) and equation (2) could provide an exact reconstruction of the original clean speech \(x_b\). However, the noise is a random process and \(n_b\) does not take a constant value. So the relation between \(x_b\) and \(y_b\)

cannot be described as a transformation; it must be described in a probabilistic context. For this reason, for a value of the clean speech \(x_b\) the contaminated speech is described by a probability distribution \(p(y_b|x_b)\). Similarly, for a given value of the noisy speech \(y_b\), even in the case of a precise estimation of the noise statistics, \(p(n_b)\), we can only obtain a probability distribution for the clean speech \(p(x_b|y_b)\), or the expected value \(E[x_b|y_b]\), but not the exact value of the original clean speech. This situation is illustrated in figure (1.B). In addition to the mismatch caused by the additive noise, there is always a loss of information due to the noise being a random process, and therefore, even in the case of a precise compensation of the noise effect, the recognizers are affected by a loss of accuracy.

2.2. Effect of convolutional noise

A convolutional noise (due to the microphone response, its orientation, the resonances of the cabin, etc.) can be described as a multiplicative factor \(H_b\) modifying the output power of each filter, \(Z_b = Y_bH_b\), or an additive constant in the FBO domain, \(z_b = y_b + h_b\) (where \(h_b = \log(H_b)\)).

Similarly to the case of additive noise, the convolutional noise transform the feature space, causing a mismatch between the training and the recognition conditions. As the additive noise, the convolutional noise is a random process. However, the variation in time of the properties of the convolutional processes are in general very slow and assuming that \(h_b\) takes a constant value usually provides an acceptable compensation of convolutional noise.

3. COMPENSATION OF THE NOISE EFFECT IN CAR ENVIRONMENTS

The FBO domain is an adequate representation of speech to perform the compensation of the noise effect, since, \(a\) the additive and convolutional noise affecting each channel can be processed independently of the noise affecting the other channels, and, \(b\) in the logarithmically scaled power representation, the probability distribution for a wide range of additive noise processes can be approximated as a Gaussian pdf.

Therefore, we propose a channel-by-channel compensation algorithm for both, the convolutional and the additive noise. The compensation process includes two stages. In the first one, the parameters describing both noise processes are estimated, and secondly, noise effects are compensated in the FBO domain. The estimation is based on the decision of a Voice Activity Detector (VAD).

The average logarithmic power of speech, \(\mu(z_b)\), is estimated for each filter, from those frames labeled as speech by the VAD. From the comparison of the average logarithmic power of both, the speech to be compensated and the clean speech in a reference database, the parameters describing the convolutional noise are estimated: \(h_b = \mu(z_b) - \mu(x_b)\). Using those frames labeled as silence, the parameters describing the additive background noise (mean \(\mu(n_b)\) and standard deviation \(\sigma(n_b)\)) are estimated.

3.1. Compensation of the convolutional noise

The convolutional noise is compensated by subtracting \(h_b\) to the FBO parameters representing the speech signal, \(y_b = z_b - h_b\). Also, the convolutional effects are compensated in the estimation of the additive noise in order to perform a coherent compensation of it, 
\[
\mu(n_b) = \mu(n'_b) - h_b; \sigma(n_b) = \sigma(n'_b).
\]
Figure 2: (A) Estimation of the expected value of the clean speech $\hat{x}_b$ given the observed noisy speech $y_b$. (B) Transformation to obtain the estimation of the clean speech as a function of the noisy speech.

### 3.2. Compensation of the additive noise

According to above discussions, the compensation of additive noise is performed in a probabilistic framework. The estimation of the clean speech $\hat{x}_b$ is obtained by use of (a) the probability density function of the noise $p(n_0)$ (that can be obtained from $\mu(n_0)$ and $\sigma(n_0)$ by assuming a Gaussian distribution), (b) the distribution of the clean speech $p(x_b)$ (estimated from the training corpus) and (c) the observed contaminated speech $y_b$. The clean speech is estimated as the expected value $\hat{x}_b$:

$$\hat{x}_b = E[x_b|y_b, p(n_0), p(x_b)] \quad (3)$$

In our compensation procedure, $x_b$ is obtained by means of a Monte Carlo method by randomly generating values according to the pdfs $p(n_0)$ and $p(x_b)$. Each couple of random samples $\{x_b(i), n_0(i)\}$ provides a sample $y_b(i)$ according to equation (2), and the distribution of samples $\{x_b(i), y_b(i)\}$ can be utilized to obtain $\hat{x}_b$ as a function of $y_b$. The value $\hat{x}_b$ is obtained as the average of $x_b(i)$ for those couples $\{x_b(i), y_b(i)\}$ for which $y_b(i)$ in an interval centered in $y_b$ (see figure 2.A). For a fast implementation of the compensation procedure, the transformation can be obtained for a small set of points $\{y_b(k), \hat{x}_b(k)\}$, and the estimation of the clean speech $\hat{x}_b$ given the observed noisy speech $y_b$ is computed by interpolation using the points $\{y_b(k), \hat{x}_b(k)\}$ (see figure 2.B).

These transformations (like the one depicted in the figure 2.B) provide an estimation of the clean speech in the FBO domain that can be used to obtain a clean version of the MFCC-based coefficients representing the speech to be recognized.

### 4. EXPERIMENTAL RESULTS

The proposed compensation procedure has been tested in recognition experiments in which speech was recorded in car environments in real traffic situations. The experiments have been carried out using the French VODIS database. Two different partitions of the database have been prepared for training and recognition. The sentences used for training were recorded with the close-talk microphone, while far-talk versions of the sentences in the test partition were utilized for recognition. Three different recognition tasks where prepared for the experiments: (1) numbers (from, 0 to 1000), (2) connected digits and (3) dialing telephone numbers. Taking into account the small size of the vocabulary, each word is modeled as a continuous density hidden Markov model (CHMM) with 10 states and left-to-right topology. Each state is described as a mixture of 8 Gaussian pdfs. For each recognition task, an adequate grammar was prepared.

The representation of the speech is based on MFCC parameterization, and includes: sampling (8 kHz), pre-emphasis, segmentation into frames (Hamming windows of 32 ms and frame shift of 8 ms), estimation of the FBO’s (using 24 Mel-scaled triangular filters), and estimation of 12 MFCC parameters, 1st and 2nd order associated regression coefficients.

We have performed recognition experiments without applying any compensation method (Baseline), applying a standard cepstral mean normalization (CMN) [13] and a sentence-by-sentence compensation of the noise in the FBO domain according to the proposed method (FBO-Comp). Takin into account the limitations of the proposed method (because of the noise being a random process and due to the non-stationary sources of distortion), in order to reduce the mismatch we have combined the compensation strategies with training the recognizers using the far-microphone version of the training database (experiments Far-CMN and Far-FBO-Comp). In this last experiment, we have applied the proposed compensation method for both training and recognition. Figure 3 shows the results of the recognition experiments.

In the baseline experiments, the average sentence accuracy is very poor (33.8%). The application of CMN increases it to 50.6%. The proposed compensation procedure systematically improves the accuracy of the recognizers, increasing the average sentence accuracy up to 61.6%. The improvement of the proposed method with respect to CMN is also consistent when the compensation methods are combined with the above described training strategy. In this case, the average sentence accuracy is increased from 72.7% (Far-CMN) to 80.5% (Far-FBO-Comp).

Important aspects for a real time implementation of the proposed method have also been considered. For a proper estimation of the parameters describing the additive noise, 1/2 second of silence is sufficient, and 1 second of speech signal is enough for the estimation of the convolutional noise. Also, the computation load of the compensation method is small compared to the rest of the recognition process. This allows a real time sentence-by-sentence compensation of the noise effect, relevant in a context in which the acoustic and the noise conditions are modified rapidly, like car environments.

### 5. CONCLUSIONS

In this paper, we have analyzed the acoustic conditions in car environments, and the main sources of distortion degrading the quality of the speech signal and reducing the performance of ASR systems. We have proposed a novel method to compensate the effect of the convolutional and additive noise. This method, that can be applied to speech recognizers using MFCC-based parameterizations,
performs the compensation in the FBO domain by (1) estimating the parameters describing the noise and (2) compensating the FBO parameters according to those estimations. The proposed method has been tested in recognition experiments using sentences recorded in cars running in real traffic situations, obtaining significant improvements with respect to the standard CMN compensation method. In the performed experiments, the best results are achieved when the mismatch between training and recognition conditions is simultaneously reduced by both, training in conditions similar to the recognition ones and applying the proposed compensation method. Considerations about requirements for the estimation of the noise parameters and computational load show that a sentence-by-sentence compensation procedure can be considered for real-time applications.

6. ACKNOWLEDGMENTS

The experiments have been carried out with the ESPERE recognition toolbox developed at LORIA-INRIA [14].

This work has been partially supported by “Plan Propio de Investigacion 99-00” of the University of Granada.

7. REFERENCES


