Speech Recognition Under Noise Conditions: Compensation Methods

Angel de la Torre, Jose C. Segura, Carmen Benitez, Javier Ramirez,
Luz Garcia and Antonio J. Rubio
University of Granada
Spain

1. Introduction

In most of the practical applications of Automatic Speech Recognition (ASR), the input speech is contaminated by a background noise. This strongly degrades the performance of speech recognizers (Gong, 1995; Cole et al., 1995; Torre et al., 2000). The reduction of the accuracy could make unpractical the use of ASR technology in applications that must work in real conditions, where the input speech is usually affected by noise. For this reason, robust speech recognition has become an important focus area of speech research (Cole et al., 1995).

Noise has two main effects over the speech representation: it introduces a distortion in the representation space, and it also causes a loss of information, due to its random nature. The distortion of the representation space due to the noise causes a mismatch between the training (clean) and recognition (noisy) conditions. The acoustic models, trained with speech acquired under clean conditions do not model speech acquired under noisy conditions accurately and this degrades the performance of speech recognizers. Most of the methods for robust speech recognition are mainly concerned with the reduction of this mismatch. On the other hand, the information loss caused by noise introduces a degradation even in the case of an optimal mismatch compensation.

In this chapter we analyze the problem of speech recognition under noise conditions. Firstly, we study the effect of the noise over the speech representation and over the recognizer performance. Secondly, we consider two categories of methods for compensating the effect of noise over the speech representation. The first one performs a model-based compensation formulated in a statistical framework. The second one considers the main effect of the noise as a transformation of the representation space and compensates the effect of the noise by applying the inverse transformation.

2. Overview of methods for noise robust speech recognition

Usually the methods designed to adapt ASR systems to noise conditions are focused on the reduction of the mismatch between training and recognition conditions and can be situated in one of these three groups (Gong, 1995; Bellegarda, 1997):

Source: Robust Speech Recognition and Understanding, Book edited by: Michael Grimm and Kristian Kroschel, ISBN 987-3-90213-08-0, pp.460, I-Tech, Vienna, Austria, June 2007

- Robust representations of speech: if speech is represented with a parameterization that is minimally affected by noise, we can assume that the mismatch between training and recognition conditions can be ignored.
- Compensation of the noisy speech representation: if we know how the speech representation is affected by noise, the effect of the noise over the representation of the speech can be compensated, and a clean version of the speech representation can be processed by the recognizer.
- Adaptation of the clean speech models to the noise environment: taking into account the noise statistics, the speech models (trained in the reference clean environment) can be adapted to the recognition noisy conditions and the recognition can be performed using the noisy speech representation and the models adapted to noise conditions.

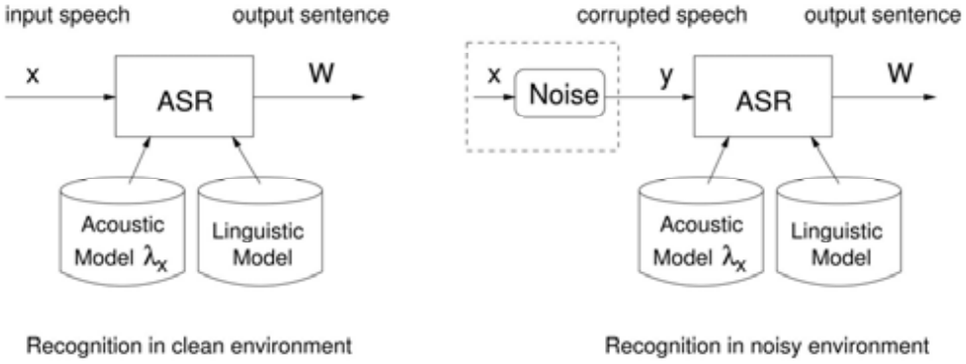


Figure 1. Block diagram of a speech recognition system processing clean or noisy speech with clean speech models

If we denote with x the clean speech representation, the effect of the noise produces a distortion and converts x into y. As shown in figure 1, the noise produces a mismatch between training and recognition conditions since corrupted speech y is recognized using clean models λ_x. According to the above classification, methods based on robust representation assume that $x \kappa\text{-} y$, which minimizes the impact of recognizing noisy speech y using clean models λ_x. Compensation and adaptation methods assume that noisy speech y is a distorted version of the clean speech $y = T(x)$, where T(-) models the distortion caused by noise. Compensation methods estimate for the inverse distortion function $T^{-1}(\bullet)$, and provide an estimation of the clean speech as:

$$\hat{x} = \hat{T}^{-1}(y) \tag{1}$$

and recognition is performed using the estimation of the clean speech x and the clean models λ_x. On the other hand, adaptation methods apply the estimated distortion function T(-) to the models:

$$\hat{\lambda_y} = \hat{T}(\lambda_x) \tag{2}$$

and recognition is performed using the noisy speech y and the estimation of the noisy models $\hat{\lambda_y}$.

2.1 Robust parameterizations

For those methods included in this category, speech parameterization is assumed to be independent of the noise affecting the speech. In order to improve robustness against noise a variety of methods have been proposed:

- Application of liftering windows: In LPC-cepstrum based representations, cepstral coefficients are not equally affected by noise. For this reason, the application of liftering windows to reduce the contribution of low-order coefficients (i.e. those more affected by noise) increases robustness against noise (Junqua & Wakita, 1989; Torre et al., 1997).
- Methods based on auditory models: Some authors have designed parameterizations based on human auditory models in order to increase robustness. In this group we can find, for example, PLP analysis (Perceptually-based Linear Prediction) (Hermanski et al., 1985; Junqua & Haton, 1996), the EIH model (Ensemble-Interval Histogram) (Ghitza, 1992; Ghitza 1994; Rabiner & Juang, 1993), or the synchronous auditory models like the Seneff auditory model 0ankowski et al., 1995) and the SLP (Synchronous Linear Prediction) proposed by Junqua 0unqua & Haton, 1996). Compared to LPC-cepstrum, parameterizations based on auditory models provides better recognition results under noise conditions, where auditory masking or lateral inhibition play an important role in speech perception.
- Mel-scaled cepstrum: The Mel-Frequency Cepstral Coefficients (MFCC) (Davis & Mermelstein, 1980) provide significant better results than LPC-cepstrum under noise conditions, and similar results to those provided by parameterizations based on auditory models, even though with significantly lower computational load (Jankowski et al., 1995). For this reason, high resolution auditory models are not considered for speech recognition applications that must work in real time, and MFCC parameterization is one of the most commonly used speech representations for robust speech recognition (Moreno, 1996).
- Discriminative parameterizations: Parameterizations based on discriminative criteria enhance those features containing more discriminative information, and this improves separability among classes. This improves recognition in both, clean and noisy conditions. In this group, Linear Discriminant Analysis (LDA) (Duda & Hart, 1973; Fukunaga, 1990) has been successfully applied for robust speech recognition (Hunt et al., 1991). Some comparative experiments shows that IMELDA representation (Integrated MEL-scaled LDA) are more robust to noise than LPC-cepstrum or MFCC. Discriminative Feature Extraction (Torre et al., 1996) has also been successfully applied to robust speech recognition (Torre et al., 1997).
- Slow variation removal: Most noise processes varies slowly in time (compared to the variations of the speech features). High-pass filtering of the speech features tends to remove those slow variations of the feature vectors representing speech, which increases accuracy of speech recognizers under noise conditions. RASTA processing (RelAtive SpecTrAl) performs this high-pass filtering of speech parameterization either in the logarithmically scaled power spectra (Hermanski et al., 1993) or in the cepstral domain (Mokbel et al., 1993). Some experiments show that RASTA processing reduces error rate when training and recognition conditions are very different, but increases it when conditions are similar (Moreno & Stern, 1994). An efficient and simple way to remove slow variations of speech parameters is the Cepstral Mean Normalization (CMN). CMN provides results close to those of RASTA processing without the

undesired degradation observed when training and recognition conditions are similar (Anastasakos et al., 1994). Currently, the use of CMN is generalized for robust speech parameterizations (Moreno, 1996; Young et al., 1997).

- Inclusion of time derivatives of parameters: Dynamic features or time derivatives (i.e. delta-cepstrum and delta-delta-cepstrum) (Furui, 1986) are usually included into the speech parameterization since they are not affected by slow variations associated to noise. The inclusion of the dynamic features improves the recognizer performance in both, clean and noisy conditions (Hernando & Nadeu, 1994).

2.2 Compensation of the noise

Compensation methods provide an estimation of the clean speech parameterization in order to reduce the mismatch between training (clean) and recognition (noisy) conditions. This way, the clean version of the speech is recognized using models trained under clean conditions. In this category, we can find the following methods:

- Parameter mapping: Based on stereo speech observations (the same speech signal acquired under both, clean and noisy conditions) this method estimates a mapping that transforms clean into noisy speech parameterizations. Linear mapping assumes that mapping is a linear function that is estimated based on minimum mean squared error criterion over the stereo speech observations. Usually stereo observations are obtained by adding noise to some clean speech observations (Mokbel & Chollet, 1991). Some authors have proposed non linear mappings (Seide & Mertins, 1994) or mapping based on neural networks (Ohkura & Sugiyama, 1991). The effectiveness of this method is limited because in practice there is no stereo speech material available for the estimation of the transformation, and clean speech must be contaminated with an estimation of the noise in the recognition environment.
- Spectral subtraction: Assumed that noise and speech are uncorrelated signals, and that noise spectral properties are more stationary than those of the speech, the noise can be compensated by applying spectral subtraction, either on the spectral domain, or in the filter-bank domain (Nolazco & Young, 1994). The effectiveness of spectral subtraction strongly depends on a reliable estimation of the noise statistics.
- Statistical enhancement: Clean speech can be considered a function of the noisy speech and the noise, where the noise parameters are unknown and randomly variable. Under this assumption, clean speech parameterization can be estimated in a statistical framework (Ephraim, 1992). Maximum A-Posteriori (MAP) methods computes the noise parameters maximizing the a-posteriori probability of the cleaned speech given the noisy speech and the statistics of the clean speech. Minimum mean squared error methods estimate noise parameters minimizing the distance between cleaned speech parameters and clean speech models, given the noisy speech observations. Usually, statistical enhancement based on an explicit model of the probability distributions of clean speech and noise involves numerical integration of the distributions, which implies practical problems for real time implementations.
- Compensation based on clean speech models: Under some approaches, compensation is formulated from a clean speech model based on a vector quantization codebook (Acero, 1993) or a Gaussian mixture model (Moreno, 1996; Stern et al., 1997). Under methods like CDCN (Codeword-Dependent Cepstral Normalization) (Acero, 1993) and RATZ (Moreno, 1996; Stern et al., 1997; Moreno et al., 1998) the transformation associated to

noise is computed for each Gaussian (or each region of the vector quantizer). Clean Gaussians and the corresponding noisy Gaussians provide an estimation of the clean speech from the noisy speech. The VTS (Vector Taylor Series) approach (Moreno, 1996; Moreno & Eberman, 1997) computes the correction for each Gaussian taking into account its parameters and the statistics of the noise, and performs the compensation taking into account the clean and the corresponding noisy Gaussian mixture models.

2.3 Adaptation of the models

The aim of adaptation methods is, as in the previous case, to minimize the mismatch between training (clean) and recognition (noisy) conditions. However, in this case, the mismatch is minimized by adapting the clean models to noise conditions.

- HMM decomposition: Under this approach, also called Parallel Model Combination (PMC) (Gales & Young, 1993; Gales, 1997), noisy speech is modeled with a hidden Markov model (HMM) with N X M states, where N states are used to model clean speech, and M are used to model the noise. This way, a standard Viterbi algorithm is applied to perform simultaneous recognition of speech and noise. In the case of non-stationary noises, several states M can be used to model the noise. In the case of stationary noises, one state would be enough to represent the noise. The probability distribution of the combined model at each state must take into account that one of the clean speech model and the one corresponding to the noise. One of the main drawback of this method is the computational load.
- State dependent Wiener filtering: Hidden Markov models allow segmentation of the speech signal into quasi-stationary segments corresponding to each state of the HMM. This adaptation method includes, for each state of the HMM, a Wiener filter to compensate for the noise effect in the recognition process, or alternatively, a correction of the probability distribution to implement the Wiener filtering (Vaseghi & Milner, 1997).
- Statistical adaptation of HMMs: This methods adapts the hidden Markov models to noisy conditions under a statistical formulation. Usually, mean and variances of the Gaussians are adapted taking into account stereo speech observations (if available) by iteratively maximizing the probability of the noisy speech being generated by the adapted models (Moreno, 1996).
- Contamination of the training database: Training with noise speech is obviously the most efficient way for adapting models to noise conditions. However, this cannot be done in practice because a-priori knowledge of the noise statistics is not available during recognition, and perform retraining with noisy speech would require estimation of the noise in the sentence to be recognized, contamination of the training database with such noise and training the recognizer with the noisy database. Training is a time consuming process and this procedure cannot be implemented in real time. However, recognition results under retrained conditions can be obtained in laboratory conditions. This kind of experiments provides an estimation of the upper limit in performance that can be obtained with the best method for robust speech recognition. Training with a specific type and level of noise significantly improves recognition performance when the speech to be recognized is affected for this kind and level of noise, but usually the performance degrades if the training and recognition noises do not match. Usually, if training is performed with a variety of noises, robustness improves and performance

under noise condition significantly improves. This is the philosophy of multi-condition training proposed in the standard Aurora II (Hirsch & Pierce, 2000).

3. Effect of the noise over the speech representation

3.1 MFCC speech representation

The effect of the noise depends on the speech representation and the type and level of noise. Currently, most of the representations for ASR are based on Mel Frequency Cepstral Coefficients (MFCC). Standard MFCC parameterization usually includes: (1) Pre-emphasis of the speech signal, in order to enhance high frequencies of the spectrum. (2) Segmentation of the signal into frames, typically with a duration from 20 to 40 msecs, using a Hamming window. (3) Using a Filter Bank, the output power in logarithm scale is obtained for each filter. These coefficients are known as Filter Bank Outputs (FBO). Usually, the Filter-Bank is composed of triangular filters distributed in the Mel frequency scale. (4) By applying a Discrete Cosine Transform (DCT), the FBO coefficients are transformed into the cepstral coefficients (the MFCC). In the MFCC domain, the correlations among the different coefficients is small. Also, high order MFCC parameters are removed in order to ignore the fine structure of the spectral shape. (5) Finally, coefficients describing the evolution in time of the MFCC parameters (Δ-cepstrum and ΔΔ-cepstrum) can be included in the parameterization. Additionally, the energy of the frame (and the Δ and ΔΔ associated parameters) are usually be included in the feature vectors representing the speech signal.

3.2 Additive noise in M FCC-based representations

(A) Distortion in the log-filter-bank domain: Let x_i and n_i be the samples of the speech signal and an additive noise, and $y_i = x_i + n_i$ the samples of the noisy speech. The energy of a frame can be written as:

$$E_y = \sum_{i=1}^{I} y_i^2 = \sum_{i=1}^{I} (x_i^2 + n_i^2 + 2x_i n_i) \tag{3}$$

and assuming statistical independence of the noise and speech signals:

$$\sum_{i=1}^{I} x_i n_i = 0 \quad \Rightarrow \quad E_y = E_x + E_n \tag{4}$$

This result can be applied to whatever parameter representing an energy of the noisy signal, and in particular, to the output energy of each filter of the filter-bank. Let $X_b(t)$, $Nb(t)$ and $Yb(t)$ be the output energy of the filter b at frame t corresponding to the clean speech, the noise and the noisy speech, respectively. The relationship among them is described by:

$$Y_b(t) = X_b(t) + N_b(t). \tag{5}$$

and for the logarithmically scaled output of the filter-bank ($x_b(t) = \log(X_b(t))$):

$$y_b(t) = \log[\exp(x_b(t)) + \exp(n_b(t))] \tag{6}$$

This equation describes how additive noise affects log-filter-bank outputs in MFCC based parameterizations. Figure 2 represents the effect of the additive noise in this domain. One can observe three effects associated to additive noise:

- Additive noise produces a non-linear distortion of the representation space.
- For those regions where noise level is greater than speech level, the log-energy of the noisy speech is similar to that of the noise. In that case, speech signal is masked by noise.

For those regions where speech level is greater than noise level, the noisy speech is only slightly affected by noise.

Since MFCC representation is obtained by a linear transformation (usually a discrete cosine transform) of the log-filter-bank energies, the above described effects are also present in MFCC domain.

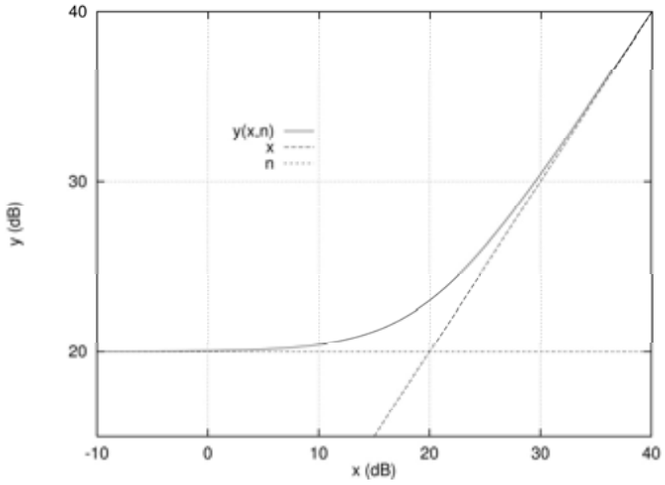


Figure 2. Distortion of the logarithmically scaled energy when a noise with constant level of 20 dB is added

(B) Distortion of the probability distributions: The previously described distortion of the representation space caused by additive noise also transforms the probability distributions. Let $p_x(x_b)$ be a probability density function (pdf) in the clean domain, and n_b the noise level affecting the clean speech. The pdf in the noise domain can be obtained as:

$$p_y(y_b) = p_x(x_b(y_b, n_b))\frac{\partial x_b}{\partial y_b} \tag{7}$$

where xb(yb, nb) and the derivative can be calculated from equation (6):

$$x_b(y_b, n_b) = y_b + \log(1 - \exp(n_b - x_b)) \tag{8}$$

$$\frac{\partial x_b}{\partial y_b} = \frac{1}{1 - \exp(n_b - y_b)} \tag{9}$$

Figure 3 represents a Gaussian probability distribution representing clean speech (mean 15dB; standard deviation 2dB) and the corresponding noisy pdf when speech is

contaminated with noise with different levels (0 dB, 5 dB, 10 dB, 15 dB). The following effects of the noise over the pdf can be observed:

- The additive noise causes a displacement of the mean.
- The standard deviation is reduced because the compression caused by noise is not uniform (is more important for the region of low energy).
- Due to the non-linear effect of the noise, the noisy pdf is distorted and it is not a Gaussian.

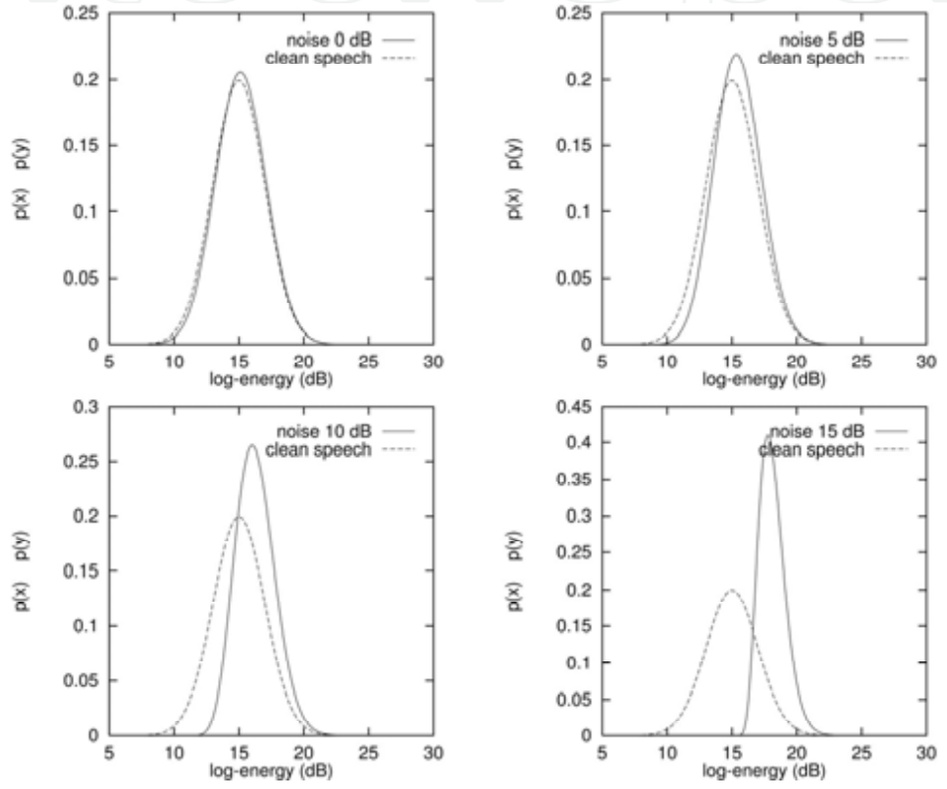


Figure 3. Distortion of the probability distributions caused by additive noise. Clean Gaussian with μ_x=15dB and σ_x=2dB. Noise with constant level OdB, 5dB, lOdB and 15dB

The impact of the mismatch caused by noise over classification is evident. Due to the distortion of the representation space caused by noise, the pdfs representing clean speech do not represent appropriately noisy speech. Let $p_x\ (x|\lambda_1)$ and $p_x\ (x|\lambda_2)$ be two pdfs representing class λ_1 and class λ_2 in the clean domain respectively. Due to the noise, both pdfs are distorted, and both clases would be represented by $p_y\ (y|\lambda_1)$ and $p_y\ (y|\lambda_2)$ in the noisy domain (that can be obtained by equation (7)). As illustrated in figure 4 the optimal boundaries are different in the clean and noisy representations. Mismatch is produced because noisy speech observations are classified using clean pdfs (and therefore boundary associated to clean pdfs), which increases the probability of classification error. In order to avoid this mismatch, the noise should be compensated on the speech representation (by applying the inverse transformation according to equation (6)) in order to obtain the clean

speech x_b, or alternatively, the models should be adapted (by applying equation (7)) in order to classify noisy speech with noisy models.

In figure 5 the area associated to the classification error in the previous example is shown. When the clean boundary is used to classify noisy speech, there is an increment in the error probability (this increment is associated to the mismatch). However when the noisy boundary is used, the probability error is exactly the same as in the case of clean speech.

(C) Randomness of noise: If noise was a constant level Uf,, the problems of noise compensation or noise adaptation would be easy to be solved. Using equation (6), an exact clean version of the speech would be obtained (or using equation (7) an exact noisy model could be used to classify noisy speech). In both cases, the probability of error would be independent of the noise level and similar to that obtained for clean speech (as can be observed in figure 4).

However, noise is a random process and therefore, the transformation is a function of random and unknown parameters. The energy of the noise cannot be described as a constant value n_b but as a pdf p_n (n_b), and therefore, for a given value of the clean speech x_b the noisy speech is not a value y_b, but a probability distribution given by:

$$p_y(y_b|x_b) = p_n(n_b(y_b, x_b))\frac{\partial n_b}{\partial y_b} \tag{10}$$

where n_b (y_b, x_b) and the partial derivative are given by equation (6):

$$n_b(y_b, x_b) = y_b + \log(1 - \exp(x_b - y_b)) \tag{11}$$

$$\frac{\partial n_b}{\partial y_b} = \frac{1}{1 - \exp(x_b - y_b)} \tag{12}$$

Figure 6 shows a Monte Carlo simulation representing how clean speech observations x_b are transformed into noisy speech observations y_b when noise is considered a random process described by a Gaussian distribution with different standard deviations. This figure illustrates the effects of the noise due to its randomness:

- For each value of clean speech x_b we do not obtain a value of noisy speech y_b but a probability distribution $p_y(y_b \mid x_b)$.
- For high energies of the clean speech (greater than the noise level) the noisy speech distribution is narrow.
- For low energies of the clean speech (compared to the noise level) the noisy speech distribution is wider.
- From a noisy speech observation y_b an estimation of the corresponding clean speech x_b is not possible. In the best case we could estimate the probability distribution $p_x(x_b \setminus y_b)$ and from it, the expected value $x_b = E[x_b \mid y_b]$ and the corresponding standard error. *In* other words, due to the randomness of the noise, there is an information loss that will increase the classification error.
- The information loss is more important as x_b is more affected by noise (for x_b with low energy compared to the noise).
- The information loss is more important as the noise level increases.

When the noise is described as a pdf, the probability distribution of the noisy speech can be computed as:

$$p_y(y_b) = \int_{-\infty}^{\infty} p(y_b, x_b)dx_b = \int_{-\infty}^{\infty} p(y_b|x_b)p_x(x_b)dx_b \tag{13}$$

and taking into account equations (10), (11) and (12):

$$p_y(y_b) = \int_{-\infty}^{\infty} p_n(n_b(y_b, x_b))\frac{1}{1-\exp(x_b - y_b)}p_x(x_b)dx_b \tag{14}$$

Figure 7 shows the effect of considering the randomness of the noise over the distribution of the noisy speech, obtained by numerical integration of equation (14). It can be observed that distributions are wider as the distribution of the noise pdf is wider. This increment in the width of $p_y(y_b)$ increases the error probability and causes the information loss associated to the randomness of the noise.

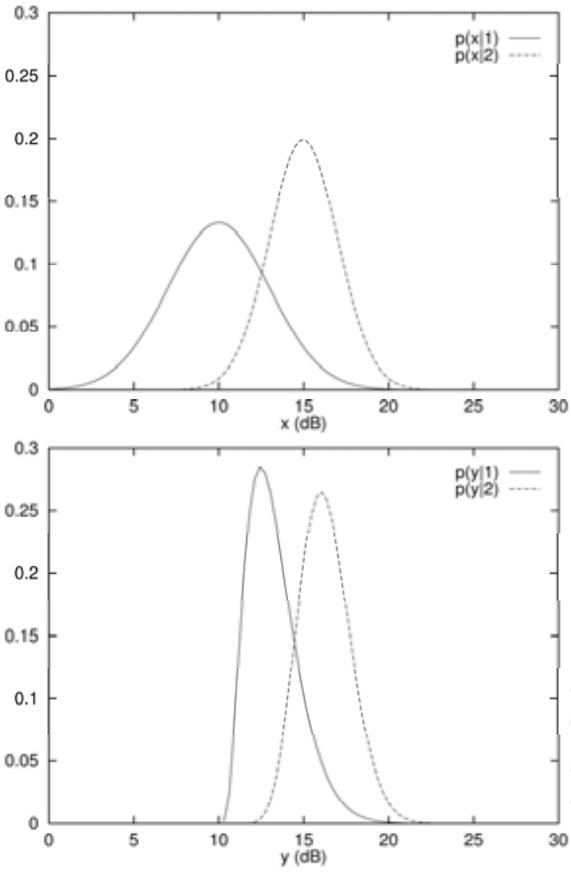


Figure 4. Displacement of the optimal decision boundary due to the noise. Clean distributions are Gaussians with μ_1=10dB, σ_1=3dB, μ_2=15dB, σ_1=2dB. Clean distributions are contaminated with a constant noise of lOdB. Optimal clean boundary at x_b =12.5317dB; optimal noisy boundary at y_b =14.4641dB

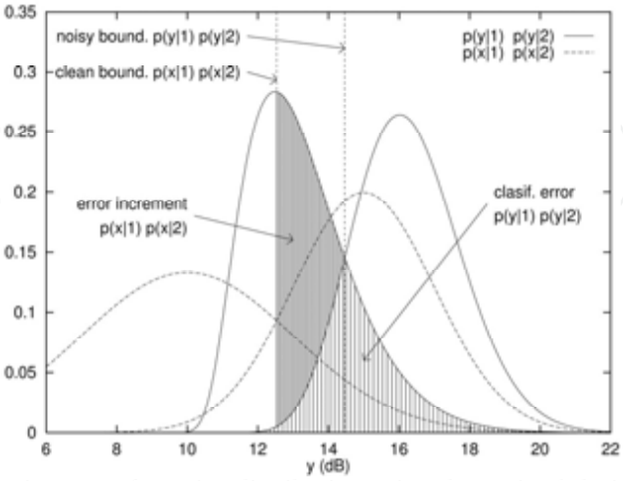


Figure 5. Classification error for noisy distribution using the optimal decision boundary y_b, and error increment when clean optimal decision boundary x_b is used

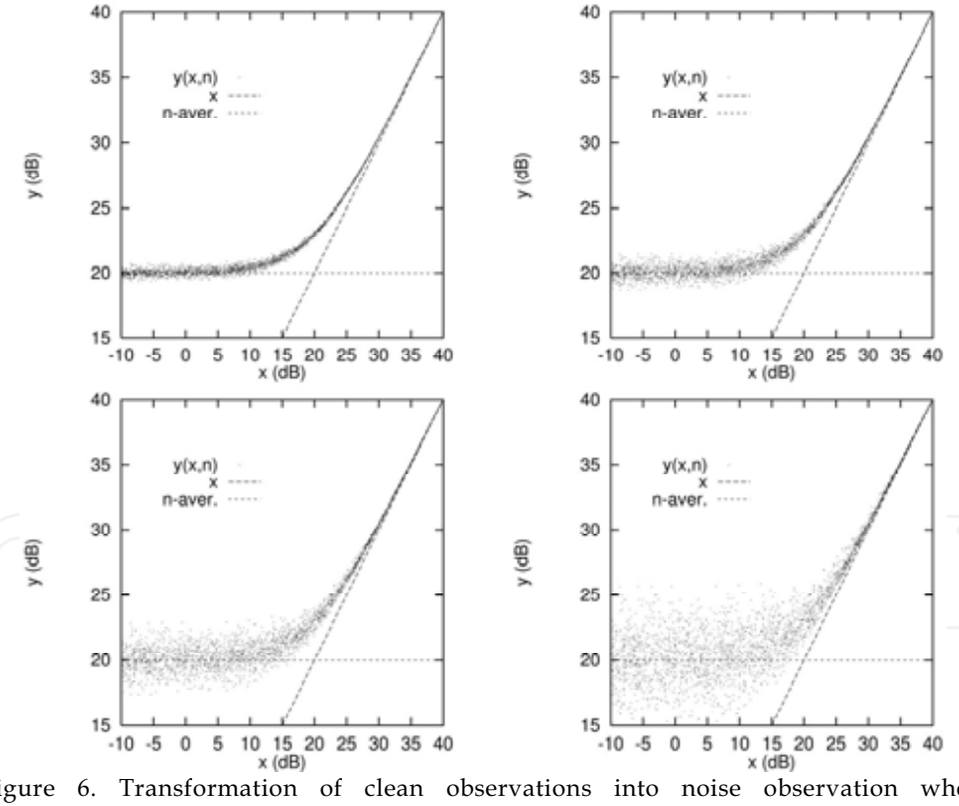


Figure 6. Transformation of clean observations into noise observation when contaminated with a noise with Gaussian distribution with $/z_n$=20dB and u_n equal to 0.25dB, O.SdB, IdB and 2dB

3.3 Effects of the noise over recognition performance

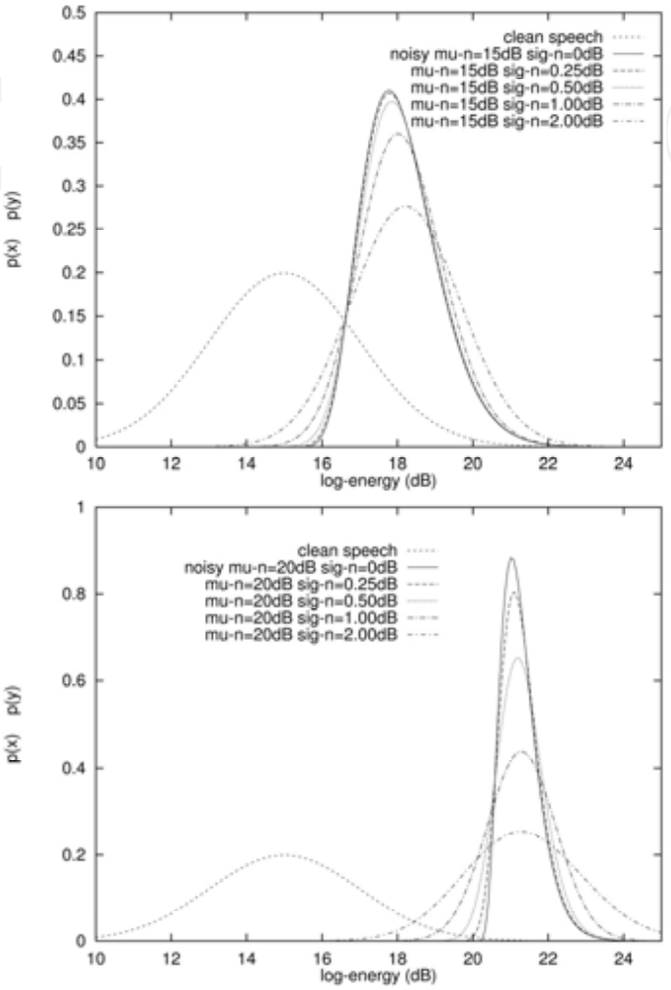


Figure 7. Effect of the randomness of noise over the probability distribution of the noisy speech: as the standard deviation of the noise increases, the noisy speech presents a wider distribution

According to the previous analysis, additive noise has two effects over speech recognition performance. On one hand, the distortion of the representation space produces a mismatch between training and recognition environments. On the other hand, the noise causes an information loss due to its implicit randomness. In order to study the role of each one over the error rate, recognition experiments can be performed using clean speech models and speech models retrained (using speech contaminated with the same noise affecting in the recognition environment). The increment of the error rate in the retrained conditions represents the degradation associated to the information loss caused by noise, while the increment of the error rate when using clean speech models represent the degradation due to both, the mismatch and the information loss.

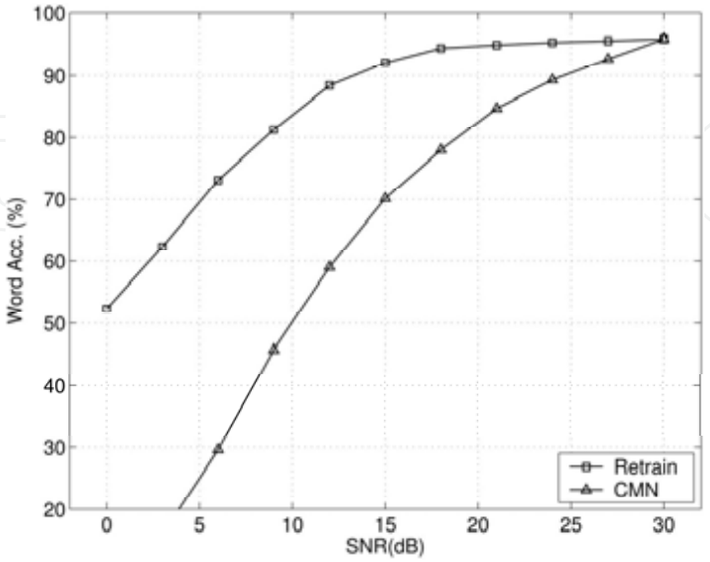


Figure 8. Reference recognition results (word accuracy versus SNR) for the baseline system (MFCC with CMN, clean training) and the retrained system

Figure 8 shows recognition performance for both, clean training and retraining conditions, for a connected digit recognition task (database of Spanish connected digits "DIGCON" (Torre et al., 2001)). The speech, represented with MFCC (including A and AA associated parameters and CMN) has been artificially contaminated with additive Gaussian white noise for SNRs ranging from 30 dB to -3 dB. Recognition experiments were carried out using a 256 Gaussians Semi-Continuous HMM speech recognizer. As observed in the figure, noise degrades performance of speech recognizer, due to both, the mismatch between training and recognition conditions and the information loss. The recognition results obtained under retraining conditions approaches the best results that could be achieved in the case of optimal compensation of speech representation or adaptation of speech models.

4. Model based compensation of the noise effects

4.1 Statistical formulation of noise compensation

Compensation of the noise effect can be formulated in a statistical framework, taking into account the probability distribution of the clean speech from a clean speech model. This way, the estimation of the clean speech could be obtained as the expected value of the clean speech, given the observed noisy speech, the model describing the clean speech and the model describing the noise statistics:

$$\hat{x}_b = E[x_b | y_b, \lambda_x, \lambda_n] \tag{15}$$

As clean speech model λx, a Gaussian mixture model (GMM) in the log-filter-bank domain can be trained using clean speech. The model describing the noise λn must be estimated from the noisy speech to be recognized. Usually the noise is represented as a Gaussian pdf in the log-filter-bank domain, and the parameters of the Gaussian are estimated from the first frames of the sentence to be recognized, or using the silence periods identified with a Voice Activity Detector (VAD). Different methods have been proposed to provide the

compensated clean speech under a statistical formulation. In the next section we describe the Vector Taylor Series (VTS) approach (Moreno, 1996; Moreno et al, 1997).

4.2 Vector Taylor Series approach

The effect of additive noise, described by equation (6), can be rewritten as:

$$y_b(t) = x_b(t) + g_b(t) \tag{16}$$

representing that noisy speech $y_b(t)$ is obtained by applying a correction $g_b(t)$ to the clean speech $x_b(t)$, where the correction is:

$$g_b(t) = \log(1 + \exp(n_b(t) + x_b(t))) \tag{17}$$

Let us ignore the frame index t for simplicity. We can define two auxiliary functions f_b and h_b as:

$$f_b \equiv \frac{1}{1 + \exp(x_b - n_b)} \tag{18}$$

$$h_b \equiv (1 - f_b) f_b \tag{19}$$

verifying that:

$$\frac{\partial g_b}{\partial x_{b'}} = -\frac{\partial g_b}{\partial n_{b'}} = -f_b \delta_{b,b'} \tag{20}$$

$$\frac{\partial^2 g_b}{\partial x_{b'} \partial x_{b''}} = \frac{\partial^2 g_b}{\partial n_{b'} \partial n_{b''}} = -\frac{\partial^2 g_b}{\partial x_{b'} \partial n_{b''}} = -h_b \delta_{b,b'} \delta_{b,b''} \tag{21}$$

$$\begin{aligned} y_b \approx x_b + g_0 + f_0[-(x_b - x_0) + (n_b - n_0)] + \\ \frac{1}{2} h_0[(x_b - x_0)^2 + (n_b - x_0)^2 - 2(x_b - x_0)(n_b - n_0)] \end{aligned} \tag{22}$$

where g_o, f_o and h_o are the functions g_b, f_b and h_b evaluated at x_O and n_O.

Using the Taylor series approach, we can describe how a Gaussian pdf in the log-filter-bank domain is affected by additive noise. Let us consider a Gaussian pdf representing clean speech, with mean $\mu_x(b)$ and covariance matrix $\Sigma_x(b, b')$, and let us assume a Gaussian noise process with mean $\mu_n(b)$ and covariance matrix $\Sigma_n(b, b')$. Taylor series can be expanded around $x_O = \mu_x(b)$ and $n_O = \mu_n(b)$. The mean and the covariance matrix of the pdf describing the noisy speech can be obtained as the expected values:

$$\mu_y(b) = E[y_b] \tag{23}$$

$$\Sigma_y(b, b') = E[(y_b - \mu_y(b))(y_{b'} - \mu_y(b'))] \tag{24}$$

and can be estimated as a function of $\mu_x(b)$, $\Sigma_x(b, b')$, $\mu_n(b)$ and $\Sigma_n(b, b')$ as:

$$\mu_y(b) \approx \mu_x(b) + g_0(b) + \frac{1}{2} h_0(b)[\Sigma_x(b, b) + \Sigma_n(b, b)] \tag{25}$$

$$\begin{aligned} \Sigma_y(b, b') \approx (1 - f_0(b))(1 - f_0(b'))\Sigma_x(b, b') + f_0(b) f_0(b') \Sigma_n(b, b') + \\ \frac{1}{2} h_0(b)(\Sigma_x(b, b') + \Sigma_n(b, b'))\delta_{b,b'} \end{aligned} \tag{26}$$

where $g_o(b)$, $f_o(b)$ and $h_o(b)$ are the functions g_b, f_b and h_b evaluated at $x_0 = \mu_x(b)$ and $n_o = \mu_n(b)$. Therefore, the Taylor series approach gives a Gaussian pdf describing the noisy speech from the Gaussian pdfs describing the clean speech and the noise.

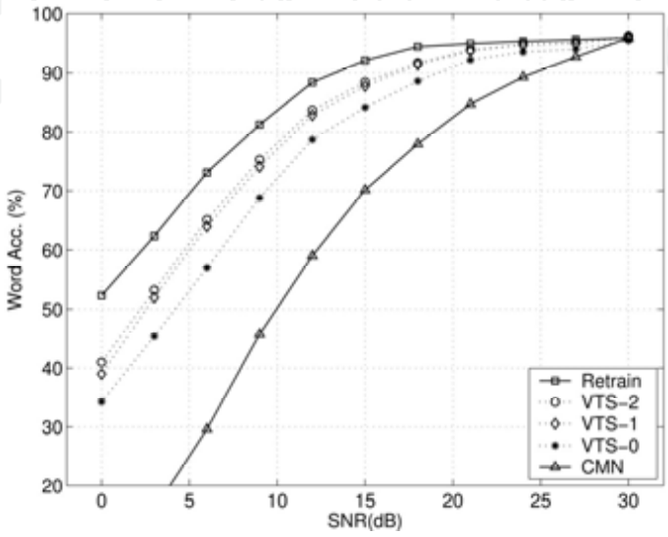


Figure 9. Recognition results obtained with VTS for different orders (0, 1 and 2) in Taylor series expansion

If the clean speech is modeled as a mixture of K Gaussian pdfs, the Vector Taylor Series approach provides an estimate of the clean speech vector $\hat{\mathbf{x}}$ given the observed noisy speech $\mathbf{y}$ and the statistics of the noise (μ_n and Σ_n) as:

$$\hat{\mathbf{x}} \approx \mathbf{y} - \sum_{k=1}^{K} P(k|\mathbf{y})\mathbf{g}(\mu_{x,k}, \mu_n) \tag{27}$$

where $n_{x,k}$ is the mean of the k^{th} clean Gaussian pdf and $P(k\,|\,y)$ is the probability of the noisy Gaussian k generating the noisy observation, given by:

$$P(k|\mathbf{y}) = \frac{P(k)\mathcal{N}(\mathbf{y}, \mu_{y,k}, \Sigma_{y,k})}{\sum_{k'=1}^{K} P(k')\mathcal{N}(\mathbf{y}, \mu_{y,k'}, \Sigma_{y,k'})} \tag{28}$$

where $P(k)$ is the a-priori probability of the k^{th} Gaussian and $\mathcal{N}(\mathbf{y}, \mu_{y,k}, \Sigma_{y,k})$ is the k^{th} noisy Gaussian pdf (with mean $\mu_{y,k}$ and covariance matrix $\Sigma_{y,k}$) evaluated at $\mathbf{y}$. The mean and covariance matrix of the k^{th} noisy Gaussian pdf can be estimated from the noise statistics and the k^{th} clean Gaussian using equations (25) and (26). This way, under VTS approach, the compensation process involves the following steps:

1. A Gaussian mixture model (GMM) with K Gaussians in the log-filter-bank domain is previously trained using clean speech
2. The noise statistics are estimated for the sentence to be compensated.
3. The clean GMM is transformed into the noisy GMM using equations (25) and (26).

4. The probability of each Gaussian generating each noisy observation $P(k|y)$ is computed using the noisy GMM (equation (28)).
5. The correction function *g(/J-x,ki M"n)* associated to each Gaussian of the GMM is computed.
6. The expected value of the clean speech is then computed for each frame using equation (27).

Figure 9 shows recognition results for the previously described connected digit task when VTS (K=V2&) is applied as noise compensation method. The reference plots correspond to the baseline clean trained and retrained systems. The other plots represent the error rate obtained when Oth, 1st and 2nd order VTS approach is applied for noise compensation. As can be observed, VTS significantly improves recognition performance, even though results are below that of the retrained system. Performance improves as the order in the expansion increases.

5. Non-linear methods for noise compensation

5.1 Limitations of model-based compensation

Model-based methods for noise compensation provide an appropriate estimation of the clean speech. They benefit from an explicit modelling of the clean speech and noise distributions as well as from an explicit modelling of the mechanism of distortion. Other effective way to face the noise compensation is to focus on the probability distribution of the noisy speech and apply the transformation that converts this distribution into the one corresponding to the clean speech. This approach does not take into account the mechanism of distortion, and only assumes that the compensated speech must have the same distribution as the clean speech. This can be considered as a disadvantage with respect to model based procedures (which would provide a more accurate compensation). However, if the mechanism of distortion if not completely known, a blind compensation procedure that is not restricted by a model of distortion can be useful. That is the case of Cepstral Mean Normalization (CMN) or Mean and Variance Normalization (MVN) (Viiki et al., 1998) that provides some compensation of the noise independently of the noise process and independently of the representation space where they are applied. This way, CMN compensates for the effect of channel noise, but is also able to partly reduce the effect of additive noise (since one side effect of additive noise is the displacement of the mean of the Gaussian pdfs). MVN allows compensation of mean and variance of the distributions, and therefore provides a more accurate compensation of additive noise than CMN.

One of the limitations of CMN and MVN is that they apply a linear transformation to the noisy noisy speech representation, and, as described in figure 2 the distortion caused by the noise present a non-linear behavior. In order to compensate for the non-linear effect of the noise, an extension of CMN and MVN methods has been formulated in the context of histogram equalization (HEQ) (Torre et al., 2005).

5.2 Description of histogram equalization

Histogram equalization was originally proposed in the context of digital image processing (Russ, 1995). It provides a transformation $x_1 = F(x_0)$ that converts the probability density function $p_0(x_0)$ of the original variable into a reference probability density function $p_1(x_1) = p_{ref}(x_1)$. This way, the transformation converts the histogram of the original variable into the reference histogram, i.e. it equalizes the histogram, as described below.

Let x_0 be an unidimensional variable following a distribution p_0 (x_0). A transformation x_1 (x_0) modifies the probability distribution according to the expression,

$$p_1(x_1) = p_0(x_0(x_1))\frac{\partial x_0}{\partial x_1} \tag{29}$$

where $x_0(x_1)$ is the inverse transformation of x_1 (x_0). The relationship between the cumulative probabilities associated to these probability distributions is given by,

$$\begin{aligned} C_0(x_0) &= \int_{-\infty}^{x_0} p_0(x_0')dx_0' = \int_{-\infty}^{x_0(x_1')} p_0(x_0(x_1))\frac{\partial x_0}{\partial x_1}dx_1' = \\ &= \int_{-\infty}^{x_0(x_1)} p_1(x_1')dx_1' = C_1(x_1(x_0)) \end{aligned} \tag{30}$$

and therefore, the transformation x_1 (x_0) which converts the distribution p_0 (x_0) into the reference distribution $p_1(x_1) = p_{ref}$ (x_1) (and hence converts the cumulative probability $C_0(x_0)$ into $C_1(x_1) = C_{ref}(x_1)$) is obtained from equation (30) as,

$$x_1(x_0) = C_1^{-1}[C_0(x_0)] = C_{ref}^{-1}[C_0(x_0)] \tag{31}$$

where $C_{ref}^{-1}[C]$ is the reciprocal function of the cumulative probability C_{ref} (x_1), providing the value x_1 corresponding to a certain cumulative probability C. For practical implementations, a finite number of observations is utilized and therefore cumulative histograms are utilized instead cumulative probabilities, and for this reason the procedure is named histogram equalization rather than probability distribution equalization.

The histogram equalization method is frequently utilized in Digital Image Processing in order to improve the brightness and contrast of the images and to optimize the dynamic range of the grey level scale. The histogram equalization is a simple and effective method for the automatic correction of too bright or too dark pictures or pictures with a poor contrast.

5.3 Noise compensation based on histogram equalization

The histogram equalization method allows an accurate compensation of the effect of whatever non-linear transformation of the feature space assumed that (1) the transformation is mono-tonic (and hence does not cause an information loss) and (2) there are enough observations of the signal to be compensated for an accurate estimation of the original probability distribution.

In the case of Digital Image Processing, the brightness and contrast alterations (mainly due to improper illuminations or non-linearities of the receptors) usually correspond to monotonic non-linear transformations of the grey level scale. On the other hand, all the pixels in the image (typically between several thousands and several millions) contribute to an accurate estimation of the original probability distributions. This makes the histogram equalization very effective for image processing.

In the case of automatic speech recognition, the speech signal is segmented into frames (with a frame period of about 10 ms) and each frame is represented by a feature vector. The number of observations for the estimation of the histograms is much smaller than in the case of image processing (typically several hundreds of frames per sentence) and also an independent

histogram equalization should be applied to each component of the feature vector. If the method is applied for noise compensation, one should take into account that the more speech is considered for the estimation of the histograms the more accurate transformation is obtained for the noise compensation. Additionally, the histogram equalization is intended to correct monotonic transformations but the random behavior of the noise makes the transformation not to be monotonic (which causes a loss of information in addition to the mismatch). The noise compensation based on histogram equalization (like the rest of the methods for noise compensation) can deal with the mismatch originated by the noise but not with the loss of information caused by the random behavior of the noise, and this limits the effectiveness of the noise compensation based on histogram equalization (like for other compensation methods).

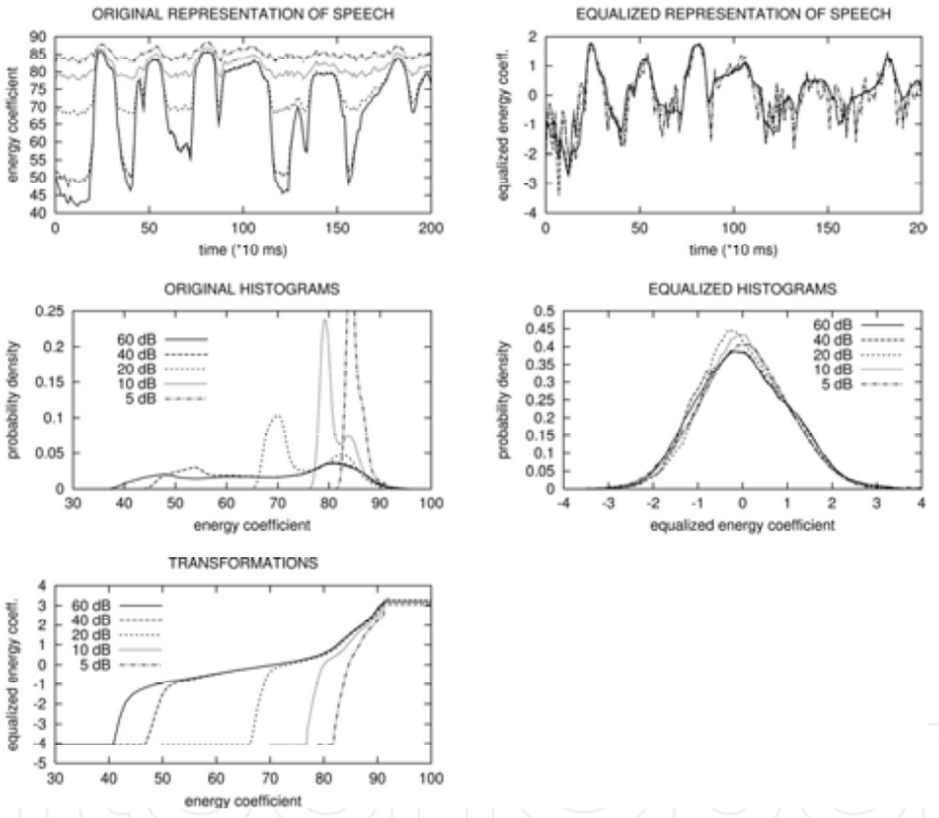


Figure 10. Effect of the histogram equalization over the representation of the speech for the energy coefficient. In the first row, the evolution over time of the energy (at different SNRs) is represented before (left) and after (right) histogram equalization, for a sentence. Histograms are represented in the second row. Transformations provided by histogram equalization procedure are represented in the last row

Compared to other compensation methods, the histogram equalization presents the advantage that it does not require any a-priori assumption about the process affecting the speech representation and therefore it can deal with a wide range of noise processes and can be

applied to a wide range of speech representations. We have applied the histogram equalization to each component of the feature vector representing each frame of the speech signal. As reference probability distribution, we have considered a normal probability distribution for each component. The histogram equalization is applied as a part of the parameterization process of the speech signal, during both, the training of the acoustic models and the recognition process. In Figure 10 we show how the histogram equalization method compensates the noise effect over the speech representation. We have contaminated the speech signal with additive Gaussian white noise at SNRs ranging from 60 dB to 5 dB. In the figure we have represented the effect of the noise and the histogram equalization for the energy coefficient. As can be observed, the noise severely affects the probability distributions of the speech causing an important mismatch when the training and recognition SNRs do not match. Histogram equalization significantly reduces the mismatch caused by the noise. However, it cannot remove completely the noise effect due to its random behavior.

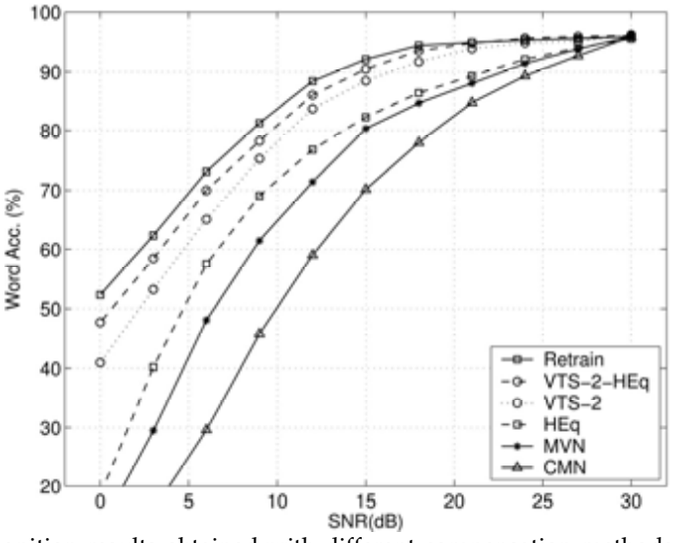


Figure 11. Recognition results obtained with different compensation methods, including 2nd order VTS (VTS-2), histogram equalization (HEq), and the combination of both (VTS-2-HEq). Compensation based on mean and variance normalization has also been included as reference

5.4 Application of histogram equalization to remove residual noise

One of the main features of histogram equalization is that no assumption is made with respect to the distortion mechanism. This reduces its effectiveness with respect to methods like VTS. However, this allow to use histogram equalization in combination with other methods. Usually, after applying a compensation method (for example, VTS, spectral subtraction, Wiener filtering, etc.) a residual noise is still present. This residual noise is difficult to be modeled because the mechanism of distortion becomes more complex than for additive or channel noise. In this case, histogram equalization can be applied since no assumption is required about the distortion process, and the compensation is performed taking only into account the probability distributions of the clean an noisy speech representations.

In figure 11 recognition results are presented for the previously described recognition task, including the reference results (clean training and retraining), histogram equalization, VTS

and the combination of both. Results applying Mean and Variance Normalization (MVN) have also been included as reference. As can be observed, histogram equalization provides better results than reference (CMN) and also better than MNV. This is consistent with the fact that histogram equalization can be considered an extension of CMN and MVN. Results provided by histogram equalization are worse than those of VTS, which shows that a model-based compensation method provides a more accurate compensation of the noise noise (particularly in these experiments, where additive Gaussian white noise was artificially added and therefore noise distortion match with the model proposed for VTS noise compensation). One can also observe that the combination of both, VTS and histogram equalization, provides an improvement with respect to VTS, showing that after VTS there is a residual noise that can be reduced by histogram equalization.

6. Conclusions

In this chapter, we have presented an overview of methods for noise robust speech recognition and a detailed description of the mechanism degrading the performance of speech recognizers working under noise conditions. Performance is degraded because of the mismatch between training and recognition and also because of the information loss associated to the randomness of the noise. In the group of compensation methods, we have described the VTS approach (as a representative model-based noise compensation method) and histogram equalization (a non-linear non-model-based method). We have described the differences and advantages of each one, finding that more accurate compensation can be achieved with model-based methods, while non-model-based ones can deal with noise without a description of the distortion mechanism. The best results are achieved when both methods are combined.

7. Acknowledgements

This work has received research funding from the EU 6th Framework Programme, under contract number IST-2002-507943 (HIWIRE, Human Input that Works in Real Environments) and SR3-VoIP projects (TEC2004-03829/TCM) from the Spanish government. The views expressed here are those of the authors only. The Community is not liable for any use that may be made of the information contained therein.

8. References

Acero, A. (1993). Acoustical and environmental robustness in automatic speech recognition, Kluwer Academic Publishers, 1993.

Anastasakos, A.; Kubala, E; Makhoul, J. & Schwartz, R. (1994). Adaptation to new microphones using tied-mixture normalization. Proceedings ofICASSP-94,1994.

Bellegarda, J.R. (1997). Statistical techniques for robust ASR: review and perspectives. Proceedings of EuroSpeech-97, Rhodes, 1997.

Cole, R.; Hirschman, L.; Atlas, L.; Beckman, M.; Biermann, A.; Bush, M.; Clements, M.; Cohen, J.; Garcia, O.; Hanson, B.; Hermansky, H.; Levinson, S.; McKeown, K.; Morgan, N.; Novick, D.G.; Ostendorf, M.; Oviatt, S.; Price, P.; Silverman, H.; Splitz, J.; Waibel, A.; Weinstein, C.; Zahorian, S.; & Zue, V (1995). The challenge of spoken language systems: research directions for the nineties. IEEE Trans, on Speech and Audio Processing, 3,1,1995,1-21.

Davis, S.B. & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuosly spoken sentences. *IEEE Trans, on Acoustic,Speech and Signal Processing,* 28,4,1980,357-366.

Duda, R.O.; Hart, RE. (1973). *Pattern Classification and Scene Analysis,* J. Wiley and Sons, 1973.

Ephraim, Y. (1992). A bayesian estimation approach for speech enhancement using hidden Markov models. *IEEE Trans, on Acoustic, Speech and Signal Processing,* 40, 4, 1992, 725-735.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition,* Academic Press, 1990.

Furui, S. (1986). Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Trans, on Acoustic, Speech and Signal Processing,* 34,1986,52-59.

Gales, M.F.J. & Young, S.J. (1993). HMM recognition in noise using parallel model combination. *Proceedings ofEuroSpeech-93,*1993.

Gales, M.F.J. (1997). 'Nice' model-based compensation schemes for robust speech recognition. *Proceedings ofESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Chanels,* 1997.

Ghitza, O. (1992). Auditory nerve representation as a basis for speech processing. In: *Advances in Speech Signal Processing,* 453-486, Dekker, 1992.

Ghitza, O. (1994). Auditory models and human performance in tasks related to speech coding and speech recognition. *IEEE Trans, on Speech and Audio Processing,* 2,1,1994,115-132.

Gong, Y. (1995). Speech recognition in noisy environments: A survey. *Speech Communication,* 16,3,1995,261-291.

Hermanski, H.; Hanson, B.A. & Wakita, H. (1985). Low dimensional representation of vowels based on all-pole modeling in the psychophysical domain. *Speech Communication,* 4, (1-3), 1955,181-188.

Hermanski, H.; Morgan, N. & Hirsch, H.G. (1993). Recognition of speech in additive and convolutional noise based on RASTA spectral processing. *Proceedings ofICASSP-94,* 1994.

Hernando, J. & Nadeu, C. (1994). Speech recognition in noisy car environment based on OSALPC representation and robust similarity measuring techniques. *Proceedings of ICASSP-94,*1994.

Hirsch, H.G. & Pierce, D. (2000). The AURORA experimental framework for the performance evaluation of speech recognition systems under noise conditions. *ISCAITRW ASR2000, Automatic Speech Recognition: Challenges for the Next Millenium,* 2000.

Hunt, M.J.; Richardson, S.M.; Bateman, D.C. & Piau, A. (1991). An investigation of PLP and IMELDA acoustic representations and their potential for combination. *Proceedings of ICASSP-91,*1991.

Jankowski, C.R.; Hoang-Doan, J. & Lippmann, R.P. (1995). A comparison of signal processing front ends for automatic word recognition. *IEEE Trans, on Speech and Audio Processing,* 3, 4,1995, 286-293.

Junqua, J.C. & Wakita, H. (1989). A comparative study of cepstral lifters and distance measures for all pole models of speech in noise. *Proceedings ofICASSP-89,*1989.

Junqua, J.C. & Haton, J.P. (1996). *Robustness in automatic speech recognition,* Kluwer Academic Publishers, 1996.

Mokbel, C. & Chollet, G. (1991). Speech recognition in adverse environments: speech enhancement and spectral transformations. *Proceedings of ICASSP-91,*1991.

Mokbel, C.; Monn, J. & Jouvet, D. (1993). On-line adaptation of a speech recognizer to variations in telephone line conditions. *Proceedings ofEuroSpeech-93,*1993.

Moreno, P.J.; Stern, R. (1994). Source of degradation of speech recognition in telephone network. *Proceedings ofICASSP-94,*1994.

Moreno, P.J. (1996). *Speech Recognition in Noisy Environments,* PhD thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania, 1996.

Moreno, P.J.; Eberman, B. (1997). A new algorithm for robust speech recognition: the delta vector taylor series approach. *Proceedings of EuroSpeech-97,*1997.

Moreno, P.J.; Raj, B. & Stern, R. (1998). Data-driven environmental compensation for speech recognition: a unified approach. *Speech Communication,* 24, 4,1998,267-288.

Nolazco-Flores, J.A. & Young, S. (1993). *CSS-PMC: a combined enhancement/compensation scheme for continuous speech recognition in noise,* Cambridge University Engineering Department. Technical Report CUED/F-INFENG/TR.128,1993.

Ohkura, K. & Sugiyama, M. (1991). Speech recognition in a noisy environment using a noise reduction neural network and a codebook mapping technique. *Proceedings oflCASSP-91,*1991.

Rabiner, L.R. & Juang, B.H. (1993). *Fundamentals of Speech Recognition,* Prentice-Hall, 1993.

Russ, J.C. (1995). *The Image Processing Handbook,* CRC Press, 1995.

Seide, F. & Mertins, A. (1994). Non-linear regression based feature extraction for connected-word recognition in noise. *Proceedings ofICASSP-94,*1994.

Stern, R.; Raj, B. & Moreno, P.J. (1997). Compensation for environmental degradation in automatic speech recognition. *Proceedings of ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Chanels,* 1997.

Torre, A.; Peinado, A.M.; Rubio, A.J.; Sanchez, V.E. & Diaz, J.E. (1996). An application of Minimum Classification Error to feature space transformations for speech recognition. *Speech Communication,* 20,3-4,1996,273-290.

Torre, A.; Peinado, A.M.; Rubio, A.J. & Garcia P. (1997). Discriminative feature extraction for speech recognition in noise. *Proceedings ofEuroSpeech-97,* Rhodes, 1997.

Torre, A.; Fohr, D. & Haton, J.P. (2000). Compensation of noise effects for robust speech recognition in car environments. *Proceedings oflCSLP 2000,* Beijing, 2000.

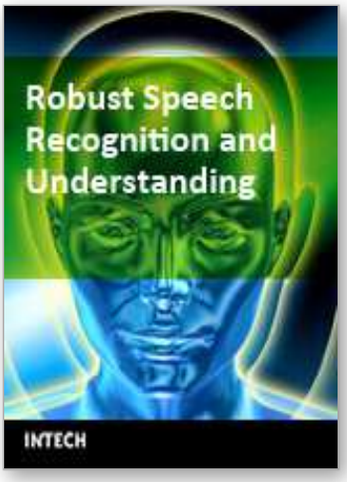
Torre, A.; Peinado, A.M.; Rubio, A.J. (2001). *Reconocimiento automdtico de voz en condiciones de ruido,* University of Granada, 2001.

Torre, A.; Peinado, A.M.; Segura, J.C.; Perez-Cordoba, J.L.; Benitez, C. & Rubio, A.J. (2005). Histogram equalization of the speech representation for robust speech recognition. *IEEE Trans, on Speech and Audio Processing,* 13,3,2005,355-366.

Vaseghi, S.V. & Milner, B.P. (1997). Noise compensation methods for hidden Markov model speech recognition in adverse environments. *IEEE Trans, on Speech and Audio Processing, 5,*1,1997,11-21.

Viiki, O.; Bye, B. & Laurila, K. (1998). A recursive feature vector normalization approach for robust speech recognition in noise. *Proceedings ofICASSP-98,*1998.

Young, S.; Odell, J.; Ollason, D.; Valtchev, V. & Woodland, P. (1997). *The HTK Book,* Cambridge University, 1997.



Robust Speech Recognition and Understanding

Edited by Michael Grimm and Kristian Kroschel

ISBN 978-3-902613-08-0
Hard cover, 460 pages
Publisher I-Tech Education and Publishing
Published online 01, June, 2007
Published in print edition June, 2007

This book on Robust Speech Recognition and Understanding brings together many different aspects of the current research on automatic speech recognition and language understanding. The first four chapters address the task of voice activity detection which is considered an important issue for all speech recognition systems. The next chapters give several extensions to state-of-the-art HMM methods. Furthermore, a number of chapters particularly address the task of robust ASR under noisy conditions. Two chapters on the automatic recognition of a speaker's emotional state highlight the importance of natural speech understanding and interpretation in voice-driven systems. The last chapters of the book address the application of conversational systems on robots, as well as the autonomous acquisition of vocalization skills.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen