VTS RESIDUAL NOISE COMPENSATION

J.C. Segura[1]*, M.C. Benítez*[1]*, A. de la Torre*[1]*,S. Dupont*[2]*, A.J. Rubio*[1]

[1]Dpto. de Electrónica y Tecn. de Comp. Universidad de Granada, Granada, SPAIN
[2]International Computer Science Institute, Berkeley, California, USA
{segura,carmen,atv,rubio}@ugr.es, dupont@icsi.berkeley.edu

ABSTRACT

The VTS approach for noise reduction is based on a statistical formulation. It provides the expected value of the clean speech given the noisy observations and statistical models for the clean speech and the additive noise. The compensated signal is only an approximation of the clean one and retains a residual mismatch. The main objective of this work is to characterize this residual noise and to propose techniques to reduce its unwanted effects. Two different approaches to this problem are presented in this paper. The first one is based on linear filtering the time sequences of compensated acoustic parameters; for this purpose we use LDA-based RASTA-like FIR filters. The second approach is based on canceling the distortion introduced into the probability distribution of acoustic parameters and uses the well-known technique of histogram equalization. Results reported on AURORA database show that the proposed methods increase the recognition performance.

1. INTRODUCTION

In the real world, there is a set of distortions that affect the speech signal since it is produced by the speaker until it is in digital form; these distortions are known as acoustical environment. A widely used model of the acoustical environment includes the two main sources of the distortion: additive noise and linear channel distortion. The distorted signal degrades significantly the performance of the speech recognition systems and methods to compensate the effect of the environment must be applied in order to perform an accurate enough recognition process. In this paper, we deal with the problem of reducing additive noise effects on ASR systems.

The effect of the additive noise consist of a non-linear transformation of the representation space in the log filter-bank-energy (log FBE) domain; VTS [1] is introduced to approximate this non-linear function by its Taylor series expansion. This approach for noise reduction is based on a statistical formulation. The compensation procedure is performed in the logarithmic filter-bank domain using a clean speech model and an estimation of the noise statistics. This method provides the expected value of the clean speech observations constrained to the observed noisy speech and the statistical models for both clean speech and additive noise.

Using VTS, a great reduction of the mismatch between noisy and clean speech is obtained. Nevertheless this compensation is not perfect and recognition accuracies are not as good as those obtained with clean speech. This is mainly due to a residual mismatch that remains after the compensation procedure is applied. In this work we try to characterize this mismatch as a residual noise (difference between compensated noisy speech and clean speech) in both time and modulation frequency domains. We also explore two possible approaches to deal with it.

As we will show later in this paper, the residual noise has two main components. One of them is a variable bias that appears at low modulation frequencies, and the other is the increment of noise components at high modulation frequencies. Recently data-driven techniques for designing filters for the time sequence of spectral parameters have been proposed; particularly the filters obtained using a LDA-based techniques [2] have the shape of a band-pass filter, suppressing both low and high modulation frequency components and enhancing the most discriminative range of modulation frequencies. By using such a filter after VTS, some part of the residual noise can be suppressed and therefore a performance increase could be expected.

The second proposed approach is focused on using a non-linear approach to deal with the non-linear nature of the residual noise. Histogram equalization is a well-known technique frequently used in image enhancement. Recently, some of the authors have applied this technique to noise compensation in ASR systems [3] with promising results. In this work we present results on applying it to equalize the probability distributions of cepstral coefficients obtained from VTS compensated speech.

2. VTS RESIDUAL NOISE

The most usual acoustic representation of speech for ASR is based on log FBE. In this domain, the effect of additive uncorrelated noise can be modeled as a non-linear transform of clean speech x and noise n to give the noisy observation y.

$$y = \log(e^x + e^n) \tag{1}$$

In VTS context, the clean speech is statistically modeled as a Gaussian mixture $p(x)$. Given an estimate of noise statistics $p(n)$, a linear approach of (1) is used to obtain a statistical model of corrupted speech $p(y)$. From it, an approximated maximum likelihood estimate of clean speech $\hat{x}$ is obtained (details of the implementation of VTS used in this paper can be found in [4])

$$\hat{x} = y - \sum_{k=1}^{K} P[k|y] \log(1 + \exp(\hat{\mu}_n - \mu_{xk})) \tag{2}$$

where $P[k|y]$ is the posterior probability of the k-th Gaussian given the noisy observation, $\hat{\mu}_n$ is the estimated noise mean and μ_{xk} is the mean of the k-th clean Gaussian. On a first approximation, $\hat{x}$ can be estimated using only the most likely Gaussian contribution

$$\hat{x} \approx y - \log(1 + \exp(\hat{\mu}_n - \hat{\mu}_x)) = y - \log(1 + e^{-\hat{s}}) \tag{3}$$

This work has been supported by the Spanish Government under the CICYT project TIC99-0583.

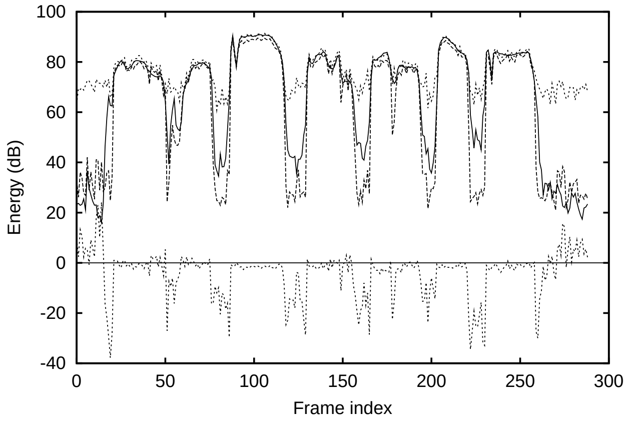

Fig. 1. *Typical temporal sequence of log FBE for clean speech (solid), noisy speech (dotted), VTS compensated speech (dashed) and residual noise (dotted bottom).*

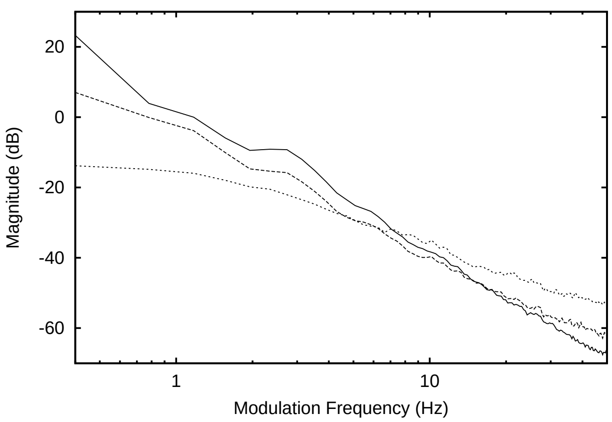

Fig. 2. *PSD of a typical log FBE for clean speech (solid), of the difference between clean and noisy speech (dashed) and of the residual noise (dotted).*

where $\hat{\mu}_x$ is the mean of the most probable Gaussian and $\hat{s} = \hat{\mu}_x - \hat{\mu}_n$ is an estimate of the instantaneous signal-to-noise ratio (SNR). Note that (3) is also the form of a Wiener filter in the log domain and therefore the following discussion is also valid for this compensation approach.

In the univariate case, considering a first order Taylor expansion of (1) around the true expected values of signal μ_x and noise μ_n we can write

$$y = \mu_x + \log(1+e^{-s}) + \frac{x-\mu_x}{1+e^{-s}} + \frac{n-\mu_n}{1+e^{s}} \qquad (4)$$

where $s = \mu_x - \mu_n$ is the actual SNR. Using (3) and (4) we have

$$\hat{x} = \mu_x + \log\left(\frac{1+e^{-s}}{1+e^{-\hat{s}}}\right) + \frac{x-\mu_x}{1+e^{-s}} + \frac{n-\mu_n}{1+e^{s}} \qquad (5)$$

from which we can obtain the expected value E and the variance V of the estimator $\hat{x}$.

$$E[\hat{x}] = \mu_x + \log\left(\frac{1+e^{-s}}{1+e^{-\hat{s}}}\right) \qquad (6)$$

$$V[\hat{x}] = \frac{\sigma_x^2}{(1+e^{-s})^2} + \frac{\sigma_n^2}{(1+e^{s})^2} \qquad (7)$$

These expressions show the two main components of the residual noise after compensation. First, equation (6) shows that the expected value of the estimator is the true value of clean speech plus a varying bias which is a non-linear function of both the actual and estimated SNR. Second, equation (7) shows that the variance of the estimator is a nonlinear interpolated value between the clean speech variance and the noise variance, and depends only on the actual SNR. Figure 1 shows this typical behavior for a sentence, where both effects are noticeable. residual noise has a great variance at initial and final silences while variance is smaller at non-masked speech parts. Where speech is heavily masked by noise the bias is large (due to an improper estimation of the SNR) while other parts exhibit a small bias.

Other useful characterization of the residual noise can be done in the modulation frequency domain [5]. This characterization has been done by evaluating an averaged estimation of the power spectral density (PSD) of time sequences of residual noise. The PSD is obtained from the training set of AURORA for which stereo data is available; a subset of 402 sentences of this training set is used. The noisy conditions correspond to clean speech with CAR noise added at a mean SNR of 10dB. Figure 2 shows the average PSD of a typical log FBE for clean speech, of the difference between noisy and clean speech and of the residual noise. It can be seen that VTS reduces the mismatch between noisy and clean speech resulting in small values of the residual noise at low modulation frequencies. However, at the same time, high modulation frequency components of the residual noise have been increased.

3. RESIDUAL NOISE REDUCTION

In this section, two different approaches are presented to reduce the effect of residual noise. The first one is based on linear filtering the time sequences of log FBE's after VTS is applied. The second approach is based on the equalization of the probability distribution of cepstral parameters to a reference Probability Density Function (PDF).

3.1. Temporal filtering

In the previous section, we have shown that the residual noise has high modulation frequency components and also a non-zero mean value. Several approaches have been proposed for the design of temporal filters [2, 7] to enhance robustness of speech parameterization. These filters tend to suppress both high and low modulation frequency components and enhance those on the most discriminative range of modulation frequencies located in the 3-4 Hz region. Using such filters, some part of the residual noise can be suppressed.

In this work, we have used two temporal filters designed in an LDA-based data driven approach [8]. One is obtained from OGI stories database (referred in the following as LDAC) and the other one from the same database contaminated with Restaurant noise at a mean SNR of 10dB (referred as LDAN). The frequency responses of these filters are approximated with 27 tap FIR filters, and they are showed in figure 3. The difference between these two filters is that LDAN exhibits a greater attenuation at zero frequency and a narrower pass-band.

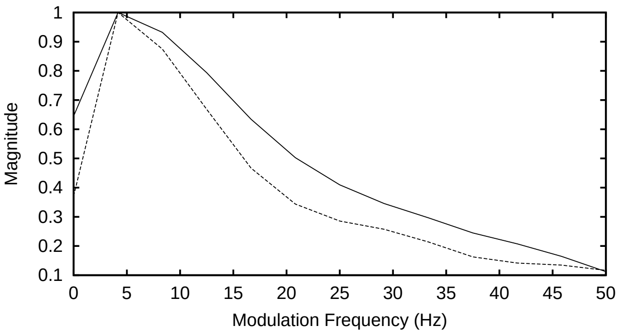

Fig. 3. *Frequency response of the LDA filters obtained from clean speech LDAC (solid) and from noisy speech LDAN (dashed).*

3.2. Histogram equalization

This technique was originally developed for digital image processing and its goal is to provide a transformation $x = F(y)$ which converts the probability density function (PDF) $p_y(y)$ of the original variable into a reference PDF $p_{ref}(x)$. Under certain conditions[1], the relation between the cumulative density functions (CDF) of x and y, and the transform function are [9]

$$C_Y(y) = C_{ref}(F(y)) \tag{8}$$

$$F(y) = C_{ref}^{-1}(C_Y(y)) \tag{9}$$

In practical situations, only a finite amount of data is available. Therefore, cumulative histograms are used instead of CDF's and for this reason this technique is named histogram equalization.

In this work, we apply histogram equalization (HEQ) in the cepstral domain. Once VTS is applied to reduce the mismatch in the log FBE domain, filter-bank energies are transformed into the cepstral domain and then each cepstral coefficients (and also the log energy term) are independently equalized to a common reference PDF. Cumulative histograms are estimated for each coefficient and each sentence to be equalized by considering 100 uniform intervals between $\mu_y - 4\sigma_y$ and $\mu_y + 4\sigma_y$ where μ_y and σ_y are the mean and standard deviation of the original values. The reference cumulative histogram is obtained from a normal distribution with zero mean and unity variance. The transformation (9) is tabulated for the points in the center of each interval and it is applied to the parameters to be compensated as a linear interpolation between the two closest tabulated points.

To illustrate the procedure of histogram equalization, figure 4 shows the histograms and transformation function used in the equalization of a non-linear transformed Gaussian distribution. Data has been generated using (1) with $\mu_x = 0, \sigma_x = 1, \mu_n = -0.75$ and $\sigma_n = 0.2$ and for the reference histogram we have used a Gaussian with zero mean and unity variance. Figures (4a) and (4b) show the histograms and cumulative histograms of y and the corresponding reference histograms and figure (4c) shows the transform function $F(y)$ that restores $p_y(y)$ to a normal distribution.

Since VTS compensation introduces a residual noise that presents important non-linearities, the use of histogram equalization to force a reference PDF for each cepstral coefficient could be expected to reduce the mismatch.

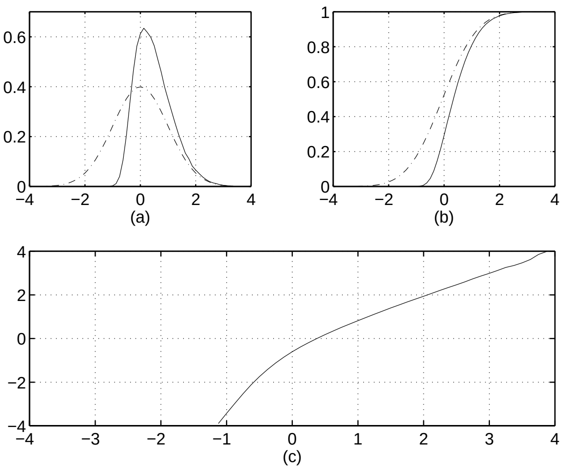

Fig. 4. *Histogram equalization. (a) Original (solid) and reference (dashed) histograms. (b) Original (solid) and reference (dashed) cumulative histograms. (c) Equalization transform.*

4. EXPERIMENTAL RESULTS

The proposed algorithms have been tested on the AURORA II database and task [6]. This database is based on TI-DIGITS. Two training sets are available; one of them (clean condition training) contains only clean data. The other one (multicondition training) contains the same data but corrupted with different noises (Subway, Babble, Car and Exhibition Hall) at mean SNR's from 20dB to 5dB. There are three test sets (A, B and C). Set A contains data corrupted with the same noises and SNR's used for the multicondition training data. Set B contains data corrupted at same SNR's that set A but with four different noises (Restaurant, Street, Airport and Train Station)). Finally, set C contains two of the noises of set A (Subway and Street) but has also a channel distortion. Two types of experiments are defined for this database. The first one uses the clean condition training set to train acoustic models and the second one uses the multicondition training set. Results are presented as an average value for five SNR conditions (from 20dB to 0dB) across the three test sets (A, B and C).

The baseline recognition system is based on HTK and use continuous density HMM models with six Gaussians per state. There are 11 digit models with 16 states, one silence model with 3 states and a inter-digit pause model with only one state. Basic parameter extraction is obtained with a set of 23 MEL-spaced triangular filters covering the frequency range from 64Hz to 4KHz at a frame rate of 100Hz. These parameters are transformed to cepstral domain using DCT, retaining only cepstral coefficients C1-C12, and log Energy is appended. Finally, this basic set of 13 parameters is augmented with its corresponding delta and acceleration coefficients obtained with regression lengths of 7 and 11 respectively.

The proposed techniques are implemented in the parameterization stage and therefore they are applied for both training and testing. VTS is applied in the log FBE domain as in [4]. This is a non-iterative implementation of VTS using a zero order approach of the mismatch function (1). The input feature vector is composed of the 23 log FBE's plus log energy. Clean speech is modeled as a mixture of 128 multivariate Gaussians estimated from the clean training partition of AURORA database and the noise is characterized by its mean value obtained from 20 frames of each sentence

[1] $F(y)$ must be single-valued and monotonically increasing

Clean Condition Training				
	A	B	C	Average
Baseline	61.34	55.75	66.14	60.06
VTS	80.66	80.74	77.99	80.16
VTS+LDAN	81.60	81.73	78.87	81.11
VTS+LDAC	81.88	81.55	79.69	81.31
VTS+HCEP	85.33	85.16	83.11	84.82
Multi Condition Training				
Baseline	87.82	86.27	83.78	86.39

Table 1. *Results for the proposed algorithms and baseline systems, averaged for SNR conditions from 20dB to 0dB.*

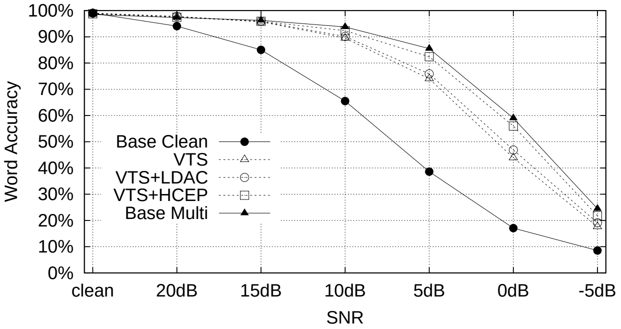

Fig. 5. *Mean results over sets A, B and C for baseline clean condition (Base Clean), baseline multicondition (Base Multi), and clean condition for VTS, VTS+LDAC and VTS+HCEP.*

that are considered as silence (10 at the beginning and 10 at the end). Features are compensated on a frame-by-frame basis and then transformed to the cepstral domain by means of a DCT transform.

Temporal filtering is applied in the log FBE domain after VTS; and histogram equalization is applied in the cepstral domain after VTS and DCT, but before the calculation of delta and acceleration parameters. Table 1 shows the results obtained for the baseline system, the VTS compensation algorithm alone and for the proposed residual noise compensation approaches. Tests labeled VTS+LDAC and VTS+LDAN correspond to the application of LDA filters trained on the clean and noisy database respectively and VTS+HCEP are results for VTS in combination with cepstral coefficient equalization. Figure 5 shows these results as a function of the SNR (Baseline results for multicondition training are also sown for reference).

From these results, it can be concluded that compensation of the residual noise improves the performance of VTS. The improvement is more noticeable at SNR's below 10 dB where there is a great amount of masked speech and VTS is less effective, leading to a large residual noise.

Recognition accuracies for VTS+LDAN and VTS+LDAC are very similar. However, when applying LDAN and LDAC alone we have noticed that much better results where obtained for the first (averaged word accuracy of 70.75% for LDAN and only 61.52% for LDAC). This can be explained by the fact that VTS compensated parameters are much more similar to clean ones than to noisy ones.

VTS+HCEP performs significantly better than VTS+LDAC (the best choice of LDA filter). This result seems to indicate that histogram equalization is a better choice for residual noise compensation than linear filtering. Nevertheless, two important aspects must be taken into account when comparing these results. First, temporal filter uses only 27 frames (270 ms of speech) while histogram equalization uses the whole sentence in the compensation process. Second, LDA filters have not been optimized for VTS compensated speech.

5. SUMMARY AND CONCLUSIONS

We have introduced the concept of residual noise as the residual mismatch after speech parameter compensation in the log FBE domain, and we have studied its approximate analytical form and presented an experimental study of it in the modulation frequency domain.

To deal with this residual noise, two different approaches have been proposed. One based on linear processing of time sequences of log FBE and the other is based on histogram equalization in the cepstral domain. Both techniques have been tested on the AURORA noisy TI-DIGITS task giving improved recognition accuracy. The higher performance of histogram equalization seems to indicate that non-linear techniques could be a better choice than linear ones to deal with residual noise.

6. REFERENCES

[1] R.M. Stern, B. Raj, and P.J. Moreno. ”Compensation for environmental degradation in automatic speech recognition”. *ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Chanels*, pages 33–42, April 1997.

[2] S. Van Vuuren and H. Hermansky. ”Data-driven design of RASTA-like filters”. *Proc. EUROSPEECH 1997, Rhodes, Greece*, pages 409–412, 1997.

[3] Angel de la Torre. ”Técnicas de mejora de la representación en los sistemas de Reconocimiento Automático de Voz”. *PhD thesis*, Universidad de Granada, España, April 1999.

[4] J.C. Segura, A. de la Torre, M.C. Benitez, A.M. Peinado. ”Model-based compensation of the additive noise for continuous speech recognition. Experiments using the Aurora-II database and tasks”. *Proc. of EuroSpeech-2001*, pages 221–224, Sep 2001.

[5] B.A. Hanson, T.H. Appelbaum, J,C, Junqua. ”Spectral dynamics for speech recognition in adversre conditions”. *In: Advanced Topics in Automatic Speech and Speaker Recognition* C.H. Lee, F.K. Soong (Eds.). Kluwer Academic Publishers, Dordrecht, 1999.

[6] H.G. Hirsch and D. Pearce. ”The AURORA experimental framework for the performance evaluation of speech recognition systems under noise conditions”. *ISCA ITRW ASR2000 ”Automatic Speech Recognition: Challenges for the Next Millenium”*, Paris, France, September 2000.

[7] C. Nadeu, D. Macho, J. Hernando. ”Time and frequency filtering of filter-bank energies for robust HMM speech recognition”. *Speech Communication* vol 34, pages 93-114, 2001.

[8] C. Benítez, L. Burget, B. Chen, S. Dupont, H. Garudadri, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, S. Sivadas ”Robust ASR front-end using spectral-based and discriminant features: experiments on the Aurora tasks” *Proc. of EuroSpeech-2001*, pages 429-432, Sep 2001.

[9] J.C. Russ. *The image processing handbook*. CRC Press, 1995.