NON-LINEAR TRANSFORMATIONS OF THE FEATURE SPACE FOR ROBUST SPEECH RECOGNITION

Ángel de la Torre, José C. Segura, Carmen Benítez, Antonio M. Peinado, Antonio J. Rubio

Dpto. Electrónica y Tecn. Comp., Universidad de Granada, 18071 GRANADA (Spain)
Tel: +34.958.24.32.71 Fax: +34.958.24.32.30 e-mail: atv@ugr.es

ABSTRACT

The noise usually produces a non-linear distortion of the feature space considered for Automatic Speech Recognition. This distortion causes a mismatch between the training and recognition conditions which significantly degrades the performance of speech recognizers. In this contribution we analyze the effect of the additive noise over cepstral based representations and we compare several approaches to compensate this effect. We discuss the importance of the non-linearities introduced by the noise and we propose a method (based on the histogram equalization technique) specifically oriented to the compensation of the non-linear transformation caused by the additive noise. The proposed method has been evaluated using the AURORA-2 database and task. The recognition results show significant improvements with respect to other compensation methods reported in the bibliography and reveals the importance of the non-linear effects of the noise and the utility of the proposed method.

1. INTRODUCTION

The noise severely affects automatic speech recognition applications working in real conditions [1, 2]. The recognition systems, usually trained with clean speech do not model properly the speech acquired under noisy conditions. The noise significantly degrades the performance of speech recognizers mainly due to the mismatch between the training conditions and recognition conditions [3].

The methods proposed to make the speech recognizers more robust against the noise are mainly focussed on the minimization of the mismatch caused by the noise [3, 4, 5]. Some of them try to represent the speech signal using robust features in order to minimize the effect of the noise. Other methods try to compensate the effect of the noise over the representation and provide an estimation of the clean speech representation. There are also methods which adapt the recognizers to the noise conditions in order to evaluate the noisy speech representation with noisy speech models.

The noise introduces a distortion of the representation space which usually present a non-linear behavior. For example, cepstral based representations suffer non-linear distortions when the speech signal is affected by an additive noise [6, 7]. In this case, the frames with more energy are slightly affected but those frames with energy in the same range or smaller than the energy of the noise are severely affected. Even though linear methods (like the Cepstral Mean Normalization (CMN) [8] or Mean and Variance Normalization (MVN) [9]) provide significant improvements for cepstral based representations, these methods present important limitations due to the non-linear distortion. Methods oriented to the compensation of the noise effects over the speech representation should consider the non-linear effects and should be able to estimate the non-linear transformation providing the best estimation of the clean speech given the noisy speech.

In this work, we analyze the effect of the noise over the cepstral based representations. We show how linear methods (like CMN and MVN) are useful for the compensation of the convolutional noise and they also compensate some effects of the additive noise. The non-linear effects of the additive noise and the limitations of the linear compensation methods are also analyzed. In order to estimate non-linear transformations for proper noise compensation, we propose the application of the histogram equalization technique. We have adapted this technique (usually applied for image processing) for the compensation of the non-linear effects caused by the noise over the cepstral coefficients. We have carried out recognition experiments (using the AURORA-2 database and task [2]) to show the importance of the compensation of the non-linear effects and to evaluate the proposed compensation method.

2. NON-LINEAR EFFECTS OF THE NOISE

Currently, most of the automatic speech recognition systems make use of parameterizations based on Mel Frequency Cepstral Coefficients (MFCC) [10]. The MFCC coefficients are obtained from a bank of filters uniformly distributed in Mel frequency scale. For each frame, the MFCC coefficients are obtained as an orthonormal transformation (usually a DCT) of the output log-energies of the filterbank. If the speech and the additive noise are uncorrelated signals, for the frame t and filter b, the energy of the contaminated speech $Y_b(t)$ can be written as a function of the energies of the clean speech $X_b(t)$ and the noise $N_b(t)$ according to,

$$Y_b(t) = X_b(t) + N_b(t) \qquad (1)$$

If the signal is also affected by a convolutional noise (described by $H_b(t)$ for the frequency band b), the contaminated speech can be written as,

$$Y_b(t) = (X_b(t) + N_b(t)) \cdot H_b(t) \qquad (2)$$

and the relationship for the logarithmically scaled output of the filterbank ($x_b = \log(X_b)$) is given by,

$$y_b(t) = h_b(t) + \log[\exp(x_b(t)) + \exp(n_b(t))] \qquad (3)$$

In this domain, the convolutional noise introduces a global shift of the parameters representing the speech, while the additive noise introduces a non-linear transformation of the feature space. Since the MFCC coefficients are obtained by applying an

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

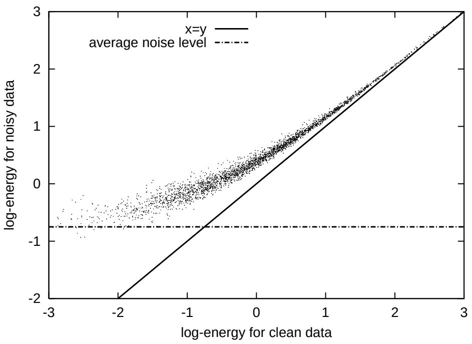

Fig. 1. *Effect of the additive noise over a log-energy coefficient.*

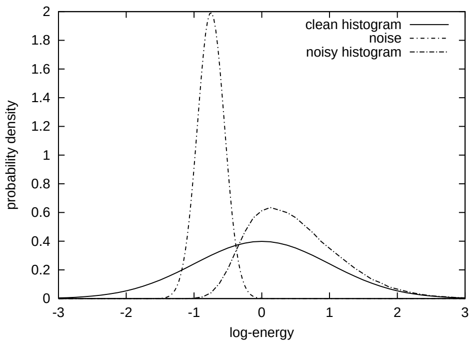

Fig. 2. *Histograms of the log-energy parameter for the clean values, the noise and the noisy values.*

orthonormal transformation to the log-filterbank outputs, the non-linear effects of the additive noise are also present in the MFCC domain.

We illustrate the non-linear effects and their consequences by means of a Monte Carlo simulation. We have randomly generated a set of clean log-energy values x_b according to a Gaussian probability distribution with mean $\mu_x = 0$ and standard deviation $\sigma_x = 1$. According to equation (3) these values have been contaminated in the log-energy domain with an additive noise randomly generated with a Gaussian distribution with $\mu_x = -0.75$ and standard deviation $\sigma_x = 0.2$. No convolutional noise was considered in this simulation ($h_b = 0$). Figure 1 shows the points $\{x_b, y_b\}$ obtained by the simulation. In this figure, it can be observed that for values x_b significantly greater than the noise, the clean values are not affected and y_b asymptotically tends to x_b. When the energy of the clean values is in the same range of the energy of the noise, the log-energy is severely affected, and when the energy is significantly smaller, y_b asymptotically tends to n_b, and then y_b shows the statistics of the noise independently of the x_b value. The additive noise causes a non-linear transformation of the feature space as can be clearly appreciated in this figure[1]. Therefore, an appropriate compensation method for the additive noise should provide a non-linear transformation to compensate the transformation caused by the noise.

The noise also affects the probability distribution of the parameters representing the speech. Figure 2 represents the normalized histograms of the clean values, the noise, and the contaminated values for the simulation. It can be observed that the transformation introduced by the noise modifies the histogram corresponding to the clean data. There is a compression of the low energy part of the clean histogram which causes a shift of the mean and a reduction of the variance of the noisy histogram. In addition, due to the non-linear effect of the noise, the shape of the histogram has been modified and it is not Gaussian.

3. COMPENSATION OF THE NOISE USING LINEAR AND NON-LINEAR TRANSFORMATIONS

Since one of the side effects of the additive noise is a shift of the mean of the probability distributions of the parameters representing the speech, the Cepstral Mean Normalization (CMN) partially compensates the mismatch caused by the noise. The combination of CMN with a normalization of the variance (Mean and Variance Normalization (MVN)) improves the compensation of the mismatch with respect to CMN, and an improvement of the recognition performance could be expected when this normalization is applied for robust speech recognition. However, these methods present the limitation that cannot compensate the non-linear effects caused by the noise.

In order to compensate the non-linear effects, we propose the application of the histogram equalization (HEQ) technique, commonly applied for image processing [11], that we have adapted to the representation of the speech signal. The aim of this method is to provide a transformation $x(y)$ which converts the probability distribution of the noisy speech $p_y(y)$ into a reference probability distribution corresponding to the clean speech $p_x(x)$. It can be demonstrated that if $x(y)$ transforms $p_y(y)$ into $p_x(x)$, then the cumulative histograms verify that,

$$C_y(y) = C_x(x(y)) \qquad (4)$$

and therefore the transformation can be obtained from the cumulative histogram of the noisy speech and the reference cumulative histogram for the clean speech as,

$$x(y) = C_x^{-1}[C_y(y)] \qquad (5)$$

where C_x^{-1} represents the inverse function of C_x. In Figure 3, the transformations provided by the linear methods (CMN and MVN) and by the HEQ method for the described simulation are shown. The histograms resulting from the application of the transformations are also shown. In this figure, it can be observed that the CMN and the MVN are linear approaches that cannot compensate properly the noise effect. The HEQ method provides a transformation which compensates the non-linear effects of the noise and removes distortion from the probability distributions of the noisy data.

4. EXPERIMENTAL RESULTS

The three considered noise compensation methods (CMN, MVN and HEQ) have been compared in recognition experiments under noise conditions using the AURORA-2 database and task [2]. The task consists on the recognition of connected digits spoken in English. The speech is artificially contaminated at several SNRs with noise recorded for 10 different conditions. The recognition results at each SNR have been averaged over all the considered kinds of

[1] In addition to the transformation, due to the random behavior of the noise, the distribution $p(y_b|x_b)$ becomes wider as x_b is more affected by the noise, which causes an irreversible loss of information.

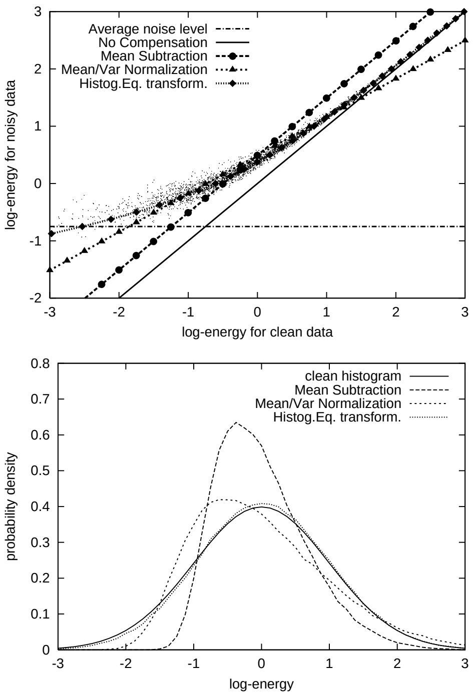

Fig. 3. *(A) Transformations to compensate the noise effect (Mean Normalization, Mean/Var Normalization and Histogram Equalization). (B) Histograms for the noisy data compensated with the different methods.*

noise. The speech recognizer is based on Hidden Markov Modeling. Each digit is modeled as a left-to-right Continuous Density HMM with 16 states and six Gaussians per state [12]. The speech recognizer has been trained under clean conditions and also using sentences contaminated with different kinds and levels of noise. Recognition experiments have been carried out using the clean training and the multicondition training recognizers according to the AURORA-2 task.

The speech representation is based on a MFCC parameterization. The speech signal, sampled at 8 kHz is segmented into frames and each frame is represented as a feature vector containing a log-energy coefficient, 12 cepstral coefficients and the 1st and 2nd associated regression coefficients, which amount to 39 components. In order to apply the considered compensation methods, in the three cases the transformations have been applied to each component of the cepstral vector. In each case, the estimation of the transformation for each component is based on the estimation of the mean (for CMN), the estimation of the mean and the variance (for MVN) and the estimation of the cumulative histogram (for HEQ). These estimations are obtained using all the frames in each sentence and the transformations are applied sentence by sentence. In the case of the HEQ, the considered reference probability density function is a Gaussian probability distribution with zero mean and unit variance. In the three cases, the compensation methods are applied for both, training and recognition.

Figure 4 shows the recognition results (Word Accuracy versus the SNR level) when each compensation method is applied. The results are averaged for the three sets (set A, set B and set C) considered in the AURORA-2 task. The figure includes clean training and multicondition training results. In the case of clean training results, the CMN compensation method slightly improves the baseline results. The normalization of the mean and variance provides a better compensation of the noise effect compared to the CMN method. In contrast to these linear compensation methods, the histogram equalization method is able to compensate the non-linear mismatch caused by the noise, which provides significant improvements with respect to the Baseline, the CMN and the MVN compensation methods. The clean training recognition accuracy (averaged over SNR levels between 20 dB and 0 dB) is 60.06%, 61.13%, 69.66% and 80.96% for the Baseline (no compensation method), and the CMN, MVN and HEQ compensation methods, respectively.

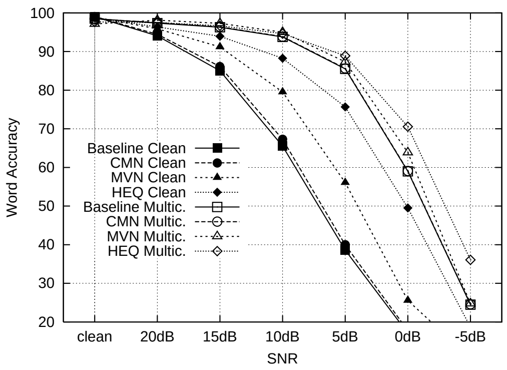

Fig. 4. *Recognition results obtained for the AURORA-2 database. Average over the different noises for clean training results and multicondition training results.*

In the case of multicondition training, the recognition results present the same tendency (average recognition accuracy of 86.39%, 86.50%, 88.33% and 89.66% for Baseline, CMN, MVN and HEQ, respectively), even though the differences in the performance for the different compensation methods are significantly smaller (because the multicondition training drastically reduces the mismatch between the training and recognition conditions).

The recognition results show the importance of the non-linear effects caused by the noise. The HEQ method provides important improvements in the recognition performance with respect to the baseline system and also with respect to the linear compensation methods because it is able to compensate the non-linear effects caused by the noise. These improvements are comparable to those provided by the best compensation methods proposed for the AURORA-2 task presented at the EUROSPEECH-2001 Conference (see Table 1) [13]. Additionally, the formulation of the HEQ method does not depend on the kind of noise or the parameterization utilized for the representation of the speech signal. Therefore, the HEQ method could provide improvements in speech recognition under noise conditions for a wide range of noise processes and for different parameterizations of the speech signal. The HEQ method could also be successfully combined with other noise compensation methods and additional improvements could be expected in this case.

5. CONCLUSIONS

In this work, we have analyzed the non-linear effects caused by the noise over the representation of the speech signal in the context of robust speech recognition under noise conditions. Linear

Aurora 2 Multicondition Training - Results

	A					B					C				Percentage
	Subway	Babble	Car	Exhibition	Average	estauran	Street	Airport	Station	Average	ubway M	Street M	Average	**Overall**	**Improvement**
Clean	97,97	97,67	97,85	97,53	97,76	97,07	97,67	97,85	97,53	97,53	98,13	97,70	97,92	**97,70**	**-57,15%**
20 dB	97,82	97,31	97,82	97,66	97,65	97,02	97,73	97,23	97,35	97,33	97,88	97,43	97,66	**97,53**	**5,04%**
15 dB	96,47	96,16	97,08	96,42	96,53	96,22	97,19	96,78	96,54	96,68	96,78	97,07	96,93	**96,67**	**6,98%**
10 dB	94,72	94,11	95,94	93,49	94,57	93,74	95,22	95,11	95,06	94,78	94,07	95,01	94,54	**94,65**	**11,42%**
5 dB	90,33	88,78	91,05	87,07	89,31	87,14	89,33	89,71	89,63	88,95	88,21	87,82	88,02	**88,91**	**22,24%**
0 dB	74,58	65,57	75,72	71,46	71,83	65,24	71,31	73,49	70,69	70,18	68,35	69,20	68,78	**70,56**	**27,05%**
-5dB	39,79	28,42	41,66	43,60	38,37	26,77	38,45	36,92	35,30	34,36	34,63	35,22	34,93	**36,08**	**15,14%**
Average	90,78	88,39	91,52	89,22	**89,98**	87,87	90,16	90,46	89,85	**89,59**	89,06	89,31	**89,18**	**89,66**	
	18,04%	**3,62%**	**37,09%**	**9,94%**	**17,75%**	**16,98%**	**24,05%**	**22,84%**	**32,31%**	**24,15%**	**34,70%**	**31,84%**	**33,32%**		**24,04%**

Aurora 2 Clean Training - Results

	A					B					C				Percentage
	Subway	Babble	Car	Exhibition	Average	estauran	Street	Airport	Station	Average	ubway M	Street M	Average	**Overall**	**Improvement**
Clean	98,83	98,61	98,99	98,89	98,83	98,83	98,61	98,99	98,89	98,83	98,80	98,67	98,74	**98,81**	**-23,71%**
20 dB	96,38	95,95	97,38	95,59	96,33	95,73	96,70	96,57	96,64	96,41	95,95	96,86	96,41	**96,38**	**26,60%**
15 dB	92,94	93,44	95,50	92,47	93,59	93,64	95,07	94,57	94,26	94,39	93,06	94,65	93,86	**93,96**	**49,79%**
10 dB	87,60	87,48	89,65	83,86	87,15	87,75	89,72	89,95	89,66	89,27	87,63	89,30	88,47	**88,26**	**62,39%**
5 dB	77,46	73,37	77,24	70,56	74,66	74,70	77,60	78,11	77,20	76,90	73,93	76,42	75,18	**75,66**	**59,26%**
0 dB	53,98	44,56	51,21	48,94	49,67	47,10	53,60	51,77	48,35	50,21	46,09	49,70	47,90	**49,53**	**38,92%**
-5dB	20,97	15,27	17,98	23,45	19,42	16,46	21,19	18,79	16,82	18,32	16,09	19,56	17,83	**18,66**	**11,01%**
Average	81,67	78,96	82,20	78,28	**80,28**	79,78	82,54	82,19	81,22	**81,43**	79,33	81,39	**80,36**	**80,76**	
	39,94%	**58,02%**	**54,81%**	**37,25%**	**48,98%**	**57,36%**	**54,62%**	**61,91%**	**57,68%**	**58,05%**	**38,92%**	**45,06%**	**41,99%**		**51,81%**

Table 1. *Recognition results obtained for the AURORA-2 database by applying the histogram equalization compensation method.*

compensation methods like CMN or MVN partially compensate the transformation caused by the noise, but they are not able to provide non-linear transformations. We have proposed the application of the histogram equalization (HEQ) technique as a method to estimate the non-linear transformations which optimally compensate the noise effects.

The linear and non-linear compensation methods have been compared and evaluated with the AURORA-2 speech recognition database and task. The experimental results show the importance of the non-linear effects when the speech signal is affected by noise, and the necessity of compensation methods which are able to compensate the non-linearities introduced by the noise. The HEQ compensation method provides a recognition performance (averaged for SNR levels between 20 dB and 0 dB) of 80.76% for clean training and 89.66% for multicondition training. The proposed method significantly improves the performance of speech recognition systems under noise conditions and is shown as a competitive method compared with the ones proposed for the AURORA-2 task [13].

Since the formulation of HEQ does not make any assumption about the contamination process, it could compensate different noise processes. Additionally, the HEQ method does not depend on the parameterization utilized for the speech representation and therefore, it could be combined with other noise compensation methods in order to obtain additional improvements.

6. REFERENCES

[1] R. Cole et. al. The challenge of spoken language systems: research directions for the nineties. *IEEE Trans. on Speech and Audio Processing*, 3(1):1–21, January 1995.

[2] 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 Millennium"*, Paris, France, September 2000.

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

[4] J.R. Bellegarda. Statistical techniques for robust ASR: review and perspectives. *Proc. of EuroSpeech-97*, pages KN 33–36, 1997.

[5] J.C. Junqua and J.P. Haton. *Robustness in automatic speech recognition*. Kluwer Academic Publishers, 1996.

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

[7] A. de la Torre, D. Fohr and J.P. Haton. Compensation of noise effects for robust speech recognition in car environments. *Proc. of ICSLP 2000*, Oct 2000.

[8] C.R. Jankowski, Jr. Hoang-Doan, and R.P. Lippmann. A comparison of signal processing front ends for automatic word recognition. *IEEE Trans. on Speech and Audio Processing*, 3(4):286–293, July 1995.

[9] P. Jain and H. Hermansky. Improved mean and variance normalization for robust speech recognition. *Proc. of ICASSP 2001*, Salt Lake City, 2001.

[10] S.B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. on Acoustic, Speech and Signal Processing*,

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

[12] S. Young, J. Odell, D. Ollason, V. Valtchev and P. Woodland. *The HTK Book*. Cambridge University, 1997.

[13] EUROSPEECH 2001, Sessions A41 and B11. Noise Robust Recognition: Front-end and Compensation Algorithms *Proc. of EUROSPEECH 2001*, pp 184-236 and 421-440, Sep 2001.