DISCRIMINATIVE FEATURE EXTRACTION FOR SPEECH RECOGNITION IN NOISE

Ángel de la Torre, Antonio M. Peinado, Antonio J. Rubio, Pedro García

Dpto. de Electrónica y Tecnología de Computadores
Universidad de Granada, 18071 GRANADA (Spain)
e-mail: atv@hal.ugr.es telf: 34-58-243271 fax 34-58-243230

ABSTRACT

Signal representation is crucial for designing a speech recognizer. The feature extractor selects the information to be used by the classifier to perform the recognition.

In noisy environments, the data vectors representing the speech signal are changed and the recognizer performance is degraded by two main facts: (1) the mismatch between the training and the recognition conditions and (2) the degradation of the signal to be recognized. In such a situation, the representation of the speech signal plays an important role.

In this paper, we analyze the importance of the representation for speech recognition in noise. We apply the *Discriminative Feature Extraction* (DFE) method to optimize the representation. The experiments presented in this work show that the DFE method, which has been successfully applied in clean environments, leads also to improvements of the speech recognizers in noise.

1. INTRODUCTION

Noise that contaminates the speech signal changes the data vectors representing the speech. In noisy environments, the performance of speech recognizers is degraded by two main facts:

- The distribution of vectors representing the speech signal is affected by the noise type and level. So, if the signal in the operating environment is contaminated with a certain noise, it is optimal to train the system with a reference signal contaminated with the same type of noise and SNR. Otherwise, the mismatch between the reference and the operating environments degrades the recognizer performance.
- The contamination of the signal leads to a random modification of the feature vectors. Therefore, even in the case of training the system in the same conditions as the operating environment, a reduction of the recognition accuracy is observed.

The representation of the speech signal plays an important role in the noisy speech recognition. In order to improve the performance of speech recognizers in noise, recent studies have been carried out on the representation [1]. Some methods are oriented to the search of noise resistant features and robust distance measures. In these schemes, a robust feature vector is utilized for the representation of the speech signal, so that the effect of the noise is minimum.

Some works present comparative studies about the robustness of different parameters or distance measures against different types of noise. For example, for a cepstral representation, the *projection measure* is more robust than Euclidean distance because the angle between vectors is less affected by noise than their norms [2].

A different way for increasing the robustness of a certain representation consists in the application of *transformations* which enhances the most robust components of the feature vector. For example, for a cepstral representation, an adequate liftering window can increase the robustness of the recognizers. The *bandpass liftering window* proposed by Juang et al. [3] is shown to provide a good performance under different noise conditions.

The discriminative training have been applied for computing robust transformations of a set of parameters. The *Linear Discriminant Analysis* (LDA) [4] improves clean recognition and some comparative tests show that LDA leads also to more noise robust representations [5].

More recently, the Discriminative Feature Extraction (DFE) method has been proposed for improving the representation of the speech signal [6] [7]. The feature extractor is a transformation applied to the original feature vectors and the parameters of the transformation are discriminatively trained in order to minimize the error-rate. In the new representation space, the most discriminative components are enhanced and this increases the discrimination capability of the recognizer.

In this work, we compare the behavior of speech recognizers in noise for several representations and we apply the DFE method for improving the representation. The importance of the representation for speech recognizers in noise is discussed. The experiments show that the DFE method, which has been successfully applied in clean environments, also improves the representation for speech recognition in noise.

2. DISCRIMINATIVE FEATURE EXTRACTION

The DFE representation is obtained by applying a linear transformation to the original feature space, $\tilde{\mathbf{x}} = V\mathbf{x}$, (where $\mathbf{x}$ are the original feature vectors and $\tilde{\mathbf{x}}$ the transformed ones). The elements of the transformation $v_{n,p}$ are iteratively computed with the *Minimum Classification Error* criterion [8] by a gradient descent algorithm in order to minimize a cost function L which represents the classification error. At iteration k, $v_{n,p}$ is computed by gradient descent of the cost function,

$$v_{n,p}^{k} = v_{n,p}^{k-1} - \eta \frac{\partial L}{\partial v_{n,p}} \tag{1}$$

where η is the convergence coefficient. Let $\{X_1, \ldots, X_M\}$ be the set of training sequences and $\{\lambda_1, \ldots, \lambda_I\}$ the set of classes; the cost function can be defined as,

$$L = \sum_{m=1}^{M} l_m(X_m) \tag{2a}$$

$$l_m(X_m) = \frac{1}{1 + e^{-\alpha d_m(X_m)}} \tag{2b}$$

$$d_m(X_m) = -g_{k(m)} + \frac{1}{\beta}\log\left[\frac{1}{I-1}\sum_{j\neq k(m)} e^{\beta g_j}\right] \qquad (2c)$$

where $g_i = g_i(X_m, \lambda_i)$ are the *discriminant functions* (the recognized class is the one whose discriminant function is the largest one) and $\lambda_{k(m)}$ is the correct class for the considered sequence X_m. This way, $l_m \to 0$ for a clearly correct classification and $l_m \to 1$ for an incorrect classification (l_m is a smooth and derivable classification error function for sequence X_m).

In order to compute $\partial L/\partial v_{n,p}$ it is necessary to know the discriminant functions, which are given by the definition of the classifier. According to the results discussed in [7], the transformation and the classifier are independently trained. Therefore, for the estimation of the transformation a very simple classifier is utilized, and all the training processes for the definitive classifier are performed in the new optimized representation space.

The classifier proposed for the estimation of the transformation, models the production of the feature vectors that belong to each class λ_i by one spherical Gaussian probability density function,

$$P(\tilde{\mathbf{x}}|\lambda_i) = \frac{1}{(2\pi\sigma_i^2)^{d/2}}\exp\left(-\frac{1}{2}\frac{\|\tilde{\mathbf{x}}-\tilde{\mathbf{y}}_i\|^2}{\sigma_i^2}\right) \qquad (3a)$$

$$\tilde{\mathbf{y}}_i = E_i[\tilde{\mathbf{x}}] \quad \sigma_i^2 = \frac{1}{d}E_i[\|\tilde{\mathbf{x}}-\tilde{\mathbf{y}}_i\|^2] \qquad (3b)$$

where d is the dimensionality of the representation space, and $E_i[\cdot]$ means average over all the vectors that belong to the class λ_i. According to this model, for a given sequence $X_m = \mathbf{x}_1, \ldots, \mathbf{x}_T$, the discriminant functions are constructed as,

$$g_i(X_m|\lambda_i) = \log P(X_m|\lambda_i) = \frac{1}{T}\sum_{t=1}^{T}\log P(\tilde{\mathbf{x}}_t|\lambda_i) \qquad (4)$$

From the definition of the cost function, and taking into account that $\tilde{\mathbf{x}} = V\mathbf{x}$, it is possible to compute $\partial L/\partial v_{n,p}$, and this allows the iterative estimation of the transformation by using equation (1). The DFE transformation enhances the most discriminative features, and this improves the representation of the speech signal.

3. EXPERIMENTS

3.1. Recognition task

A speaker independent Isolated Word Recognition task has been developed for testing the different representations under several noise conditions. The vocabulary is composed of 16 words (10 Spanish digits and 6 keywords) and the data base is composed of 3 utterances of every word recorded from 40 speakers (20 male and 20 female). In order to study the effect of noise, the original speech signal is contaminated with additive Gaussian white noise. Recognition experiments are performed for several SNRs ranging from 35dB to 0 dB.

3.2. Representation of the speech signal

The speech signal has been sampled (sample frequency f_s=8kHz) and segmented into frames of 32ms, overlapped 16ms. We have computed 20 cepstral coefficients (from 10 LPC coefficients), 20 delta cepstral ones, the energy and the delta energy for every frame of speech.

The feature vectors have been obtained from these coefficients by applying several liftering windows to the cepstral and delta cepstral vectors, or by applying a transformation computed with the DFE method. We have applied *rectangular, triangular* [2], *statistically weighted* [9] and *raised-sine* [3] liftering windows (these representations are labelled *Rect, Tri, SW* and *R-sin*, respectively). Two representations have been obtained by DFE, from two different initializations. The first one is initialized using a *Rect* liftering window and the second one, using a *SW* liftering window. These DFE representations are labeled *DFE-A* and *DFE-B* respectively.

3.3. Recognition systems

Two variants of HMM-based recognition systems have been used: Discrete Hidden Markov Models (DHMM) [10] and Multiple VQ Hidden Markov Models (MVQHMM) [11]. The DHMM system has been implemented using a 512 centroids codebook. For the MVQHMM system, a 32 centroids codebook is used for every class.

3.4. Recognition results

Some experiments have been developed for comparing the proposed representations. The differences among these experiments are related to the way the training data set was prepared.

3.4.1. Experiment 1

In the Experiment 1 the recognition systems are trained using clean speech. Figure 1 shows the recognition-rate versus the operating SNR for both DHMM and MVQHMM, for the representations *Rect, Tri, SW* and *R-sin*.

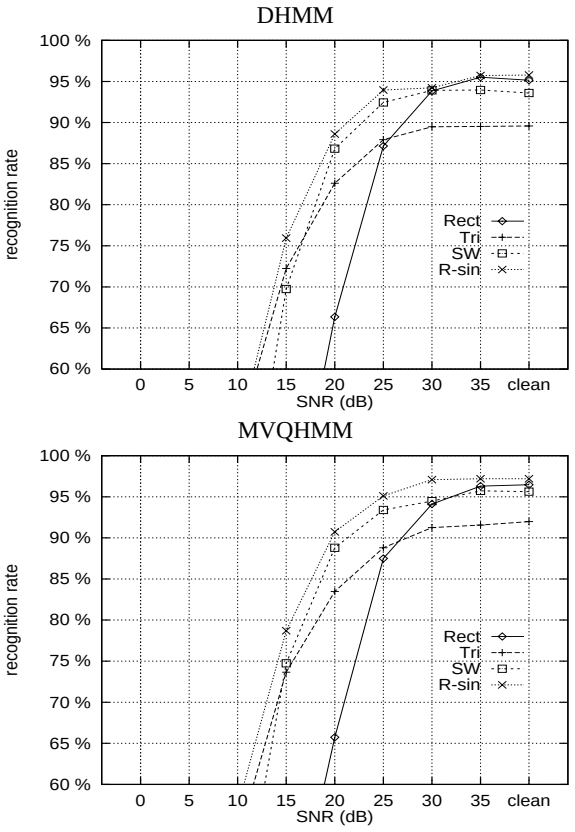

Figure 1: *Recognition results for the systems trained using clean speech*

For both recognition systems a similar behavior can be observed. The *Rect* representation provides a good result for

high SNR but the degradation is very fast. The *Tri* one is more robust but the performance for clean speech is poor. The *SW* and *R-sin* representations provide a good result for clean conditions, without the degradation observed for *Rect*.

In Table 1 the recognition results for the DFE representations are included in comparison to the other representations. In this experiment, since the DFE transformations are trained using clean speech, the improvements obtained with these transformations are less important as the SNR is smaller.

Table 1: *Recognition rate (%) for systems trained using clean speech*

SNR	Rect	Tri	SW	R-sin	DFE-A	DFE-B
DHMM						
Clean	95.15	89.58	93.59	95.78	97.34	96.61
35 dB	95.52	89.53	93.96	95.73	97.55	96.77
30 dB	93.85	89.48	93.91	94.22	96.77	96.56
25 dB	87.14	87.92	92.45	93.96	94.74	94.79
20 dB	66.35	82.61	86.82	88.60	90.31	90.16
15 dB	36.77	72.24	69.74	75.94	74.12	73.70
10 dB	17.40	51.93	33.59	51.88	41.93	42.14
5 dB	8.59	30.36	12.97	18.91	15.10	16.25
0 dB	6.77	12.40	7.29	7.24	9.74	10.52
MVQHMM						
Clean	96.46	91.98	95.62	97.19	98.75	98.65
35 dB	96.30	91.56	95.73	97.19	98.54	98.54
30 dB	94.11	91.25	94.48	97.08	98.23	98.13
25 dB	87.50	88.80	93.39	95.10	96.82	97.66
20 dB	65.73	83.49	88.80	90.73	92.81	94.06
15 dB	39.01	73.65	74.74	78.70	79.11	81.51
10 dB	18.80	53.44	41.51	57.29	49.37	48.60
5 dB	9.12	27.29	16.14	27.45	20.67	20.10
0 dB	6.61	9.95	9.17	8.85	10.21	10.94

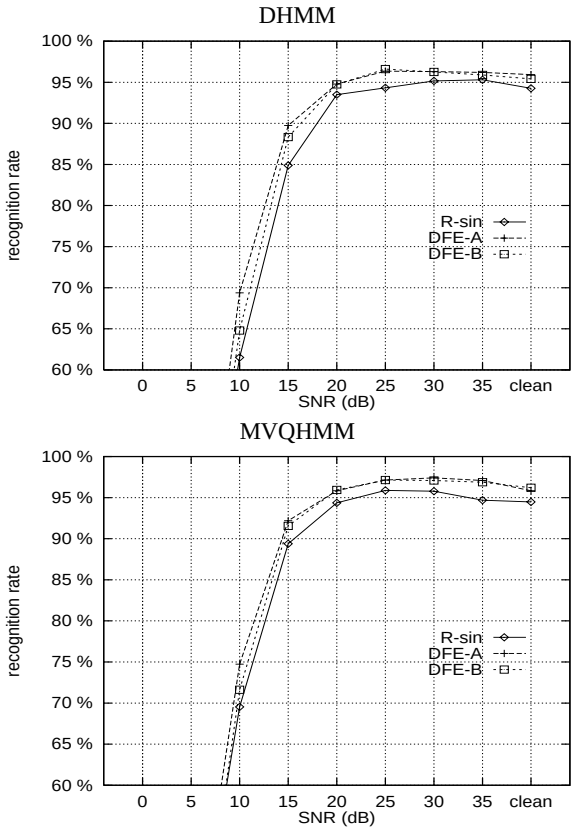

Figure 2: *Recognition results for the systems trained using corrupted speech (SNR=25dB)*

The improvement observed for the DFE representations for clean or not very degraded speech is due to the discriminative training of the feature extractor. But the discriminative training can degrade the performance when there is a mismatch between the training and test conditions. For this reason, the degradation is faster for the DFE representations when the speech signal is severely corrupted.

3.4.2. Experiment 2

In order to reduce the differences between the training and test conditions, in the *Experiment 2* we have trained the systems using a corrupted speech signal with SNR=25dB. This way, there is a slight degradation of the performance for clean speech, but under noisy environments, the degradation is smaller than the one observed when the systems are trained with clean speech. Figure 2 shows the recognition results versus the operating SNR for *R-sin* (the best one of the liftering windows) and for the DFE representations.

In this experiment, the DFE transformations were also estimated using the database corrupted at a SNR=25dB, and this increases the robustness of the representation in comparison to the *R-sin* representation. As can be observed, the accuracy of both recognizers is increased by the application of the DFE representations for SNRs greater than 10dB.

3.4.3. Experiment 3

In order to obtain more robustness against the different noise conditions, in the *Experiment 3* the systems are trained using the training database corrupted with different levels of noise. 8 databases were generated from the original one, by adding additive Gaussian white noise at different SNRs ranging from 35dB to 0dB. The system (and the DFE transformations) were trained in this experiment using a database composed of all these corrupted databases and the original one. This increases significantly the training computational load, but not the recognition process. The effect of such a training is the adaptation of the system (and the DFE representations) for a wide range of noise levels. Figure 3 shows the recognition-rate when the systems are trained using this strategy.

An important increment of the accuracy can be observed with respect to the experiments *1* and *2*, because this scheme of training reduces the mismatch between the reference and operating environments. The DFE method optimizes the representation for all the noise levels and the performance of the recognizers is improved with respect to the *R-sin* representation.

3.4.4. Discussion of the results

The results presented in the *Experiment 1* show the importance of the representation in noisy speech recognition. The representations utilized in this experiment only differs in the weights applied to the original coefficients. Therefore, a representation is more robust than other because it enhances the most robust coefficients and reduces the contribution to the distance measure of the least robust ones. For the rectangular liftering window, the contribution to the distance measure is mainly due to the low order cepstral and delta cepstral coefficients, which are severely affected by the noise. This causes the fast degradation observed

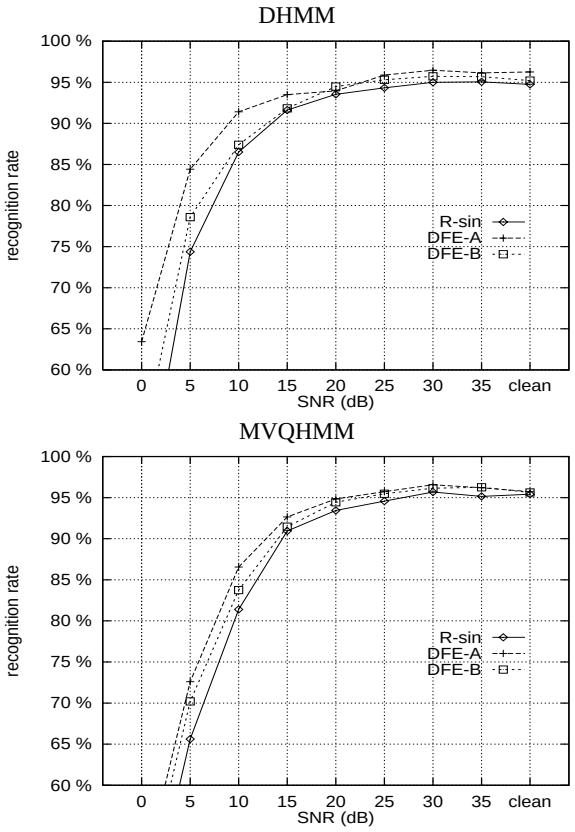

Figure 3: *Recognition results for the systems trained using the database corrupted with several noise levels*

for this representation. The triangular liftering window reduces the contribution of the low order cepstral and delta cepstral coefficients. But these coefficients are also useful for the recognition. For this reason, even though this representation is more robust than others, the performance is poor in relatively clean operating environments. The Statistically Weighted representation is usually a good starting point, because it equalizes the contribution of the different components in the feature vector, and for the cepstrum-LPC, the raised-sine liftering window provides good results as shown in several studies and observed in this experiment.

The main effect of the DFE transformation over the low order cepstral and delta cepstral coefficients is a reduction of their contribution to the distance measure (like the *Tri*, *SW* and *R-sin* liftering windows). However, there is also a small adjustment of the weights in order to optimize the representation for the training conditions. This makes the DFE representations improve the other ones in an operating environments close to the training conditions, but be non optimal when the operating and training environments are very different.

The DFE representations improve the recognition accuracy for all the proposed experiments with respect to the other liftering windows when the mismatch between training and recognition is not excessive.

4. CONCLUSIONS

The experimental results show the importance of the speech representation for speech recognition in noise. The DFE method, which improves the representation in clean conditions, improves also the representation in noisy environments. The improvements with respect to the best non-discriminative representation (*R-sin*) could be considered not very important in comparison with the noise degradation. However, the DFE method provides an improvement without an increment of the computational complexity of the recognizer, and has improved the best non-discriminative representation without a-priori information and from different initializations.

5. ACKNOWLEDGEMENTS

This work is funded by CICyT (Spanish Governmental Research Agency) under project TIC96-0956-C04-04.

6. REFERENCES

[1] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, pp. 261–291, Apr. 1995.

[2] J. Junqua and H. Wakita, "A Comparative Study of Cepstral Lifters And Distance Measures for All Pole Models of Speech in Noise," in *Proc. of ICASSP-89*, pp. 476–479, 1989.

[3] B. H. Juang, L. R. Rabiner, and J. G. Wilpon, "On the use of bandpass liftering in speech recognition," *IEEE Trans. on ASSP*, vol. 35, pp. 947–954, July 1987.

[4] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

[5] M. J. Hunt and C. Lefebvre, "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech," in *Proceedings of ICASSP'89*, vol. 1, pp. 262–265, 1989.

[6] A. Biem and S. Katagiri, "Feature extraction based on Minimum Classification Error/Generalized Probabilistic Descent method," in *Proc. of ICASSP '93*, vol. 2, pp. 275–278, 1993.

[7] A. de la Torre, A. M. Peinado, A. J. Rubio, V. E. Sánchez, and J. E. Díaz, "An application of Minimum Classification Error to feature space transformations for Speech Recognition," *Speech Communication*, vol. 20, pp. 273–290, Dec. 1996.

[8] B. Juang and S. Katagiri, "Discriminative Learning for Minimum Error Classification," *IEEE Trans. on Signal Processing*, vol. 40, pp. 3043–3054, Dec. 1992.

[9] Y. Tohkura, "A weighted cepstral distance measure for speech recognition," *IEEE Trans. on ASSP*, vol. 35, no. 10, pp. 1414–1422, Oct. 1987.

[10] L. R. Rabiner, J. G. Wilpon, and F. K. Soong, "High Performance Connected Digit Recognition Using Hidden Markov Models," *IEEE Trans. on ASSP*, vol. 37, pp. 1214–1225, Aug. 1989.

[11] J. Segura, A. Rubio, A. Peinado, P. García, and R. Román, "Multiple VQ Hidden Markov Modelling for Speech Recognition," *Speech Communication*, vol. 14, pp. 163–170, April 1994.