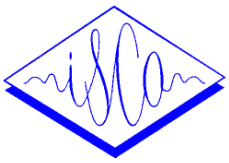

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

Word Recognition in the Car : Adapting Recognizers to New Environments

C. Mokbel*, L. Barbier*, Y. Kerlou* and G. Chollet* **

* Télécom-Paris, Dépt. Signal, CNRS URA-820,
46, rue Barrault, 75634 Paris Cedex 13, FRANCE
** IDIAP, Case Postale 609, 1920 Martigny, CH

Abstract

Several techniques for the adaptation of DTW and HMM speech recognizers to car environment are presented and tested. These techniques are : nonlinear spectral subtraction, spectral subtraction with neural networks, two strategies of feature parameters transformation using linear regression or neural networks, and finaly two approaches to adapt directly the HMM states distributions. Results given prove the superiority of parameters transformations on spectral subtraction, neural networks transformation on linear regression transformations, and adding noise to references on speech enhancement strategy. They prove also the importance of the choice of the spectral representation and the corresponding distance measure. A number of perspectives exists : defining adaptative transformations, transformations by class of noise, adapting the HMM parameters using NN.

10.21437/ICSLP.1992-239

I. Introduction

A number of real life applications are offered to speech recognition. Word recognizers loose their performances when changing conditions between learning and testing phases. Car environment applications to control hand free phones and on board computers are our major interest. In this case, learning is performed in clean conditions and recognition is done in noise varying conditions.

Noise effect can be divided in two parts: additive noise and Lombard effect. Firstly, the noise acts as an additive signal wich disturbs the clean speech signal. Thereby, we observe a transformation in the acoustical space described by the recognizer reference models. This transformation varies with the noise characteristics. Secondly, the speaker changes his or her manner of speaking in presence of noise : this is the Lombard effect.

Recently, a lot of work has been done to overcome the problems of recognition in noise [6]. Several techniques are proposed here to increase the robustness of recognizers in this case. They can be divided in three classes : choice of a robust speech representation and distance measure, adaptation of the recognizer by subtracting the noise from the test signals or adding it to the reference models.

Normaly, recognition is performed in a spectral representation space associated with a given distance measure. A first approach to improve recognition performance in presence of noise consists of defining an appropriate speech representation space and its relative distance measure. We have done comparisons of several representations and distance measures for the car environment. Results could be found in [9].

In the two other approaches, we try to adapt the recognizers to new environment conditions by speech enhancement or by transforming the references models in order to resemble the test conditions [1][8]. In fact, if we suppose that a given speaker define an acoustical subspace when pronouncing the recognition vocabulary, then the recognition problem consists to identify the trajectories in this subspace. Changes in conditions transform that subspace and deteriorate recognition performance. Looking to the problem from this side, three adaptation strategies appear. Firstly, we can superpose test acoustical subspace to reference acoustical subspace by speech enhancement if reference models were trained in clean conditions. Secondly, we can superpose reference acoustical subspace to test acoustical subspace by adding noise to references. A third method consists on projecting both of the reference and test acoustical subspaces on an adequate predefined subspace. This method is not considered in our work. Several techniques are used in our experiments : nonlinear spectral subtraction, linear regression, multilayer neural networks, transforming the HMM states distributions. These techniques and the way we use them to adapt the recognizers are described in section III. This section is devided in two subsections dealing with nonlinear spectral subtraction and feature parameter transformation. The second section is reserved to the description of the experimental conditions : the database, the recognizers, the feature set. In section IV, we give two sets of recognition results obtained with DTW and HMM recognizers. A comparaison is also done proving that adding noise to the references outperforms speech enhancement. Finally, some conclusions and perspectives are made.

II. Database and recognition systems

Our recognition experiments are speaker dependent and in the isolated word case. We are interested only by the problem of the noise. Our database has been recorded in a car (ESPRIT-ARS project). It is composed of four speakers: two males (GI, LA) and two females (CH, CO) and three noise conditions corresponding to different car speed: 0Km/h (considered clean), 90 Km/h and 130Km/h. For each condition and each speaker, 4 series of 43 words were recorded.

Two recognition systems are used in our tests : a DTW and a CDHMM. The speech signals are analysed frame by frame to give MFCC feature vectors. The length of the analysis window is 20 ms. These windows are shifted by 12 ms. The filter bank is formed of 24 sinusoidal filters uniformely distributed on a MEL scale. 8 MFCC are used with the DTW recognizer and 12 MFCC + 12 ∂MFCC with the CDHMM.

III. Techniques used to adapt recognizers to new environments

Three classes of techniques have been experimented : nonlinear spectral subtraction, feature transformations, and HMM states distributions adaptation.

III-1. Nonlinear Spectral Subtraction (NSS)

Spectral subtraction was first used for speech enhancement[3]. It estimates the amplitude spectrum of the filtered signal by substracting the mean of the ambient noise amplitude spectrum from the amplitude spectrum of the noisy signal. The phase of the noisy signal is kept for the filtered signal. If we dispose of a unique microphone to capture the signal (as in our experiments), a speech activity detector must be used to determine the non-speech regions where the noise characteristics are updated. Thereby, noise is supposed stationnary enough to maintain its characteristics in the speech parts following the non-speech regions. This technique has given good results for applications with low level stationnary noise. Some problems occur when the SNR decreases and the noise looses its stationnarity. Typically it is the case of the car environment. In these conditions, a sort of musical noise appears in the filtered signals. This musical noise is mainly due to the variations of the noise instantaneous spectrum around its mean. To overcome this problem and improve the quality of the filtered signals, nonlinear spectral substraction can be used [3][4][7]. In that case, a nonlinear function surestimating the noise mean spectrum is defined. The surestimation is more important in frequency zones where SNR is weaker. This property of the surestimation function is derived from the auditory system characteristics. In fact, our auditory system is less sensitive to the valleys than to the peaks of the sound spectra.

Recently, NSS was used with great success to improve speech recognition in noise [7]. The results shown in [7] prove the importance of this technique and the nonlinear function used. In [9] we developped a new NSS procedure. In contrary to the classical functions our function surestimates more the noise in the zones where SNR is greater. It is derived by imposing the criterium of a fixed SNR at the output of the filter. We will not develop the calculation in this paper. The resulting function is only given here. For a frequency channel i, we add to the mean value of the noise spectrum the quantity $B_n(i)$ such as :

$$\begin{cases} B_n(i) = |E(i)| \; th \dfrac{B_n(i)\,|E(i)|}{s_b^2(i)} \\ -K\, s_b(i) \le B_n(i) \le K\, s_b(i) \end{cases}$$

where :

$E(i) = C\,[\,|X(i)| - \mu(i)\,]$,

$$C = \sqrt{\frac{2}{\pi}}\, s_b(i)\, E\left[\frac{1}{|S(i)|}\right],$$

$\mu(i)$ and $s_b(i)$ are the mean and standard deviation of the noise, and

X(i) and S(i) are the spectra of the noisy and clean signals.

Using this function, satifactory results are obtained when hearing the filtered signals. It is necessary to underline that the SNR in the car environment at 130Km/h is of the order of -5dB. The filtered signals are then passed to the CDHMM recognizer. The results are given in section IV.

Neural networks (NN) could be used to perform spectral subtraction. A new technique has been proposed in [2]. Noisy speech amplitude spectrum are presented at the entry layer of a multilayer perceptron. We ask our NN to produce the corresponding clean amplitude spectrum at its output layer. Our NN is composed of N partially interconnected layers. Each layer is formed of M cells, where M is the half of the spectrum length (158). Each cell i communicates with a window of 2L+1 cells (from i-L to i+L) of the previous layer. At the output of each layer, an estimation of the filtered amplitude spectrum is found. These estimations are successively refined from the entry layer to the output layer. Preliminary experiments permit to fix the parameters N and L to 3 and 4 respectively. This choice is constrained by a compromise between the error and speed of convergence in learning stage.

The NN is trained by an improved back-propagation procedure [2]. In order to obtain corresponding input output spectra, we have simulated noisy signals from clean ones. This was done by adding real car noise to obtain the same SNR and the same energy of the real noisy data at 90 and 130 Km/h. The pairs of simulated noisy and clean spectra are presented to the input and output of the NN respectively. The training algorithm adjust the NN parameters in order to minimise the mean of the differences between the log of the filtered and clean spectra at the NN output. The use of the log function associates the log-deviation distance to the spectra space. More details could be found in [2].

Using the obtained NN with real data, we reconstruct the filtered speech signal from the filtered amplitude spectrum and the noisy phase spectrum. The resulting signal is intelligible, the musical noise of the SS is largely reduced and the distorsions introduced are limited. This technique is tested with the DTW recognition system.

III-2. Parametric transformations : Linear Regression (LR) and Neural Network (NN).

In previous papers[1][8], we present approaches to adapt recognizers to new environments using linear regression and neural network techniques and first results. Suppose that the noise effects on the clean acoustical subspace are resumed by a transformation function F_n. If we denote X a clean feature vector and Y its corresponding vector in presence of a certain noise n, we write $Y = F_n(X)$. In order to adapt the recognizer to new environment, it is interesting to determine the function F_n or its inverse. This is purely an identification problem.

With LR we approximate this function by a linear transformation between the centered X and Y vectors. So we write :

$\overline{Y} = A\,\overline{X}$ where $\overline{Y} = Y - \mu_y$, $\overline{X} = X - \mu_x$ and
A is a transformation matrix.

In our case, {X} and {Y} are MFCC vectors. If A is a full matrix, this transformation is the well known Linear Multiple Regression (LMR). Firstly, noisy MFCC vectors are aligned with the clean MFCC vectors using a DTW module. From these couples of vectors, the means vectors μ_y and μ_x and the transformation matrix A [9] are computed. We have then two possibilities : passing from noisy to clean acoustical subspaces (A^c) which could be seen as a speech enhancement procedure (SE), or the inverse (A^n) which could be seen as adding noise to the references (ANR). Note that (A^n) is not exactly the inverse of (A^c) since the identification linear function is not exactly the function F_n which is not necessary inversible and the estimation of the linear transformation is an approximation that uses the MSE criterium on the learning data.

Our preliminary experiments have shown that a transformation corresponding to a mean level of noise is sufficient. In fact, very good results are obtained when learning the transformations with the 90Km/h conditions and testing with the conditions at 130Km/h. For the results presented here, we have used 20 words at 90Km/h

aligned with the corresponding words at 0Km/h to learn the transformation.

Once the transformations (A^c) and (A^n) learned, we transform the test utterances in (SE) scheme and we passed them to the recognition systems or we transform reference utterances in (ANR) scheme and we perform recognition on test utterances using the new reference models.

Another transformation function could be found using NN. Time alignment is done in the same way as for LMR but the transformations are performed by multilayer perceptrons. This network is composed of three layers and each layer is completly interconnected with the next one. The input and output layers are composed of M cells where M=8 is the number of speech MFCC parameters used for recognition with DTW. For the hidden layer, 16 cells was found a good compromise. Two networks NN^c and NN^n were trained for both of the schemes SE and ANR respectively. NN^c defines a nonlinear function and transforms noisy MFCC vectors to clean ones. In contrary, NN^n transforms clean MFCC vectors, generaly the reference vectors, in order to resemble to the noisy conditions. Using NN offers the advantage of defining a nonlinear transformation function extending thereby the identification capacities of the LMR. In fact, if the trained NN work in their linear parts, they perform exactely the same work as LMR matrices.

III-3.Adaptation of HMM states distributions

Previously [8] we have noticed that it is sufficient and more adequate to try to apply LR transformation in the filterbank space (FLT) and then to compute the transformed MFCC vectors from the obtained FLT ones. We have also noticed that these transformation matrices could be considered as speaker independent. In FLT space, the noise effect is supposed decorrelated between the 24 dimensions corresponding to adjacent frequency channels. This hypothesis permits to simplify the calculus and to diagonalise the (A^c) and (A^n) matrices. The computation costs are therefore reduced and the possibilities of adaptative transformations becomes realistic.

The good results obtained by these mean transformations and especially in the ANR scheme have encouraged us to develop a method in order to adapt directly the HMM parameters. The HMM we use in our tests are fixed variance. Neverthless, the method that we describe here permits to adapt classical CDHMM. We aim to adapt directely the states output distributions parameters (means and variances) knowing the diagonal transformation matrix A^{diag}. For this purpose, we define a dual FLT_HMM for each MFCC_HMM. These FLT models are trained in parallel with the MFCC models. In fact, while the MFCC models are trained, the means and variances of the FLT distributions are adjusted by substituting the MFCC vectors with the corresponding FLT vectors in the adaptation equations (Baum-Welch or Viterbi algorithms). Let T_{cos} be the inverse cosine matrix transforming the FLT vectors $\{X_{flt}\}$ in MFCC vectors $\{X_{mfc}\}$, and let $\{Y_{flt}\}$ and $\{Y_{mfc}\}$ be the adapted vectors in the ANR scheme. We can write :

$$X_{mfc} = T_{cos}\, X_{flt}$$
$$Y_{flt} = A^{diag}\, X_{flt} + B$$
$$Y_{mfc} = T_{cos}\, A^{diag}\, X_{flt} + T_{cos}\, B$$

where $B = \mu_y - A^{diag}\, \mu_x$

To determine the means and the covariance matrices of the adapted HMM, we write :

$$E(Y_{mfc}) = T_{cos}\, A^{diag}\, E(X_{flt}) + T_{cos}\, B \qquad (1)$$
$$Cov(Y_{mfc}) = T_{cos}\, A^{diag}\, Cov(X_{flt})\, A^{diag\,t}\, T_{cos}^{t}$$

For the ∂MFCC vectors, we can write :

$$X_{\partial mfc} = T_{cos}\, X_{\partial flt}$$
$$Y_{\partial flt} = A^{diag}\, X_{\partial flt}$$
$$Y_{\partial mfc} = T_{cos}\, A^{diag}\, X_{\partial flt}$$

and the means and covariances become:

$$E(Y_{\partial mfc}) = T_{cos}\, A^{diag}\, E(X_{\partial flt}) \qquad (2)$$
$$Cov(Y_{\partial mfc}) = T_{cos}\, A^{diag}\, Cov(X_{\partial flt})\, A^{diag\,t}\, T_{cos}^{t}$$

Using the sets of equations (1) and (2), the noisy MFCC_HMM is easily deduced from the clean FLT_HMM. This increases the possibility of using an adaptative strategy to transform the HMM in order to superpose the ambient noise.

Another simple method is also used to adapt the HMM states means vectors. An estimation of the FLT noise vector could be done in the periods preceeding the word to recognize. This estimation N_{flt} is then used to adapt the FLT models states means additively using the equation :

$$E(Y_{mfc}) = T_{cos}\, \log\{\exp[E(X_{flt})] + \exp[N_{flt}]\}$$

IV. Experiments and results

We present in this paper two series of experiments according to recognizer system used : DTW or CDHMM based. With our DTW system, and to reduce the computation time, we used 8 MFCC weighted by the inverse of the global variances as feature vectors. The table 1 present the results in terms of recognition scores obtained by the DTW system when using the following techniques :

- MFCC(8) : initial results.
- SS_NN : spectral subtraction with NN.
- SE_NN : speech enhancement on MFCC tests.
- ANR_NN : adding noise to MFCC references.

Speaker	CH		CO		GI		LA	
Cond. (Km/h)	90	130	90	130	90	130	90	130
MFCC	69.8	30.6	29.4	33.3	92.2	81.4	73.6	70.5
SS_NN	93.6	88.9	64.7	59.4	98.8	88.9	93.6	91.9
SE_NN	96.1	93.0	96.9	96.1	93.3	89.1	94.5	96.1
ANR_NN	99.2	99.2	100	98.1	99.2	99.2	99.2	99.2

Table 1

Looking at this table we conclude that parametric transformation outperforms spectral subtraction techniques. Otherwise, ANR is a better strategy than SE. Table 2 compares the results on all the speakers comparing the NN to LR techniques :

- SE_NN : speech enhancement on MFCC tests (NN).
- ANR_NN : adding noise to MFCC references (NN).
- SE_LR : speech enhancement using LMR on MFCCs.
- ANR_LR : adding noise to references using LMR on MFCCs.

Results are given with confidence intervals in order to give an idea on their statistical validity.

Techn. / Cond.	90Km/h	130Km/h
MFCC	66.2 ± 3.5	55.0 ± 3.7
SE_NN	95.3 ± 1.5	93.9 ± 1.7
ANR_NN	98.8 ± 0.8	98.7 ± 0.8
SE_LR	95.0 ± 1.5	93.2 ± 1.0
ANR_LR	98.1 ± 1.0	98.5 ± 0.9

Table 2

For the HMM system, we present a first set of experiments showing the weak dependence on speakers

of the LR transformations in the FLT space with ANR scheme. For this purpose, we have computed a transformation by speaker and tested each of them with all the speakers. Table 3 gives the results.

Speaker	CH		CO		GI		LA	
Cond. (Km/h)	90	130	90	130	90	130	90	130
MFCC	85.5	62.2	51.7	49.4	97.1	87.2	81.4	90.7
Adap CH	98.3	97.1	95.3	92.4	99.4	99.4	98.8	98.8
Adap CO	91.9	86.0	100	97.7	100	100	99.4	100
Adap GI	88.4	82.6	93.6	83.1	100	100	100	100
Adap LA	82.6	72.1	89.5	54.7	99.4	98.8	100	100

Table 3

Table 4 sumarises the results for the following techniques :

- MFCC(12) + ∂MFCC(12) : Initial results.
- NSS : Nonlinear spectral subtraction (section 3).
- NSS_BP : NSS and bandpass lifter on MFCCs [5].
- ANR_FLT_LR : Adding noise to references using linear regression on FLT vectors.
- ANR_FLT : Adding noise to the models states means FLT vectors.

Speaker	CH		CO		GI		LA	
Cond. (Km/h)	90	130	90	130	90	130	90	130
MFCC	85.5	62.2	51.7	49.5	97.1	87.2	81.4	90.7
NSS	91.2	87.2	89.5	78.4	98.3	98.3	96.5	100
NSS_BP	99.4	94.2	97.7	89.5	96.5	96.5	93.0	100
ANR_FLT	98.8	92.4	96.5	89.0	100	100	98.8	100
ANR_FLT_LR	98.2	97.1	100	97.7	100	100	100	100

Table 4

V. Comparaisons and conclusions

This work compares several techniques proposed [2][9] to adapt recognizers to new environments. Linear regression is compared to neural networks transformations. Speech enhancement strategy is compared to adding noise to references. Spectral subtraction is compared to feature parameters transformations. Looking at the experiments results we can make the following conclusions :

- The parametric transformations are more adequate than the tested spectral subtraction techniques in terms of speech recognition.
- Adapting the recognizer by transforming the references in order to resemble to the test conditions (ANR strategy) outperforms the speech enhancement tested techniques (SE strategy).
- The results with NN and LR techniques are very close. Neverthless, the NN transformations show some superiority. Unfortunately, we have not tested these networks to adapt directely the HMM parameters.
- A simple method to adapt the HMM parameters using linear regression on the outputs of a filterbank is proposed and tested. Results are satisfactory. This transformation reveals weak dependence on the speaker.
- For a given noise class, considering a transformation for a mean level is sufficient.
- The differences between MFCC results in tables (1,2) and (3,4) and between NSS and NSS_BP in table 4 prove that an adequate choice of a spectral representation and the corresponding distance measure is necessary to perform robust speech recognition in presence of noise.
- In table 4, linear regression shows some superiority to adding directely the noise vector to the means of the states although the second approach is adaptative. In fact, two explainations could be given. Firstly, the linear regeression technique adapt both the means and covariances of the states distributions. Secondly, it is probable that linear regression simulates both the additive noise and the Lombard effect.

Although parametric transformations are more adequate to speech recognition, it is more important to continue the work on spectral subtraction for its double utility in speech enhancement and recognition.

A perspective to this work is to define several transformations in dependence of the different kind of noise possibly found in the car environment and to mix the transformed references in HMM.

Otherwise, it is interesting to make these transformations adaptative in time. Some work is to be done in this direction.

Looking to the interesting results obtained with NN, we will study the possibilities of using these networks for the direct adaptation of CDHMM parameters.

We have given some theoretical justifications for the superiority of the ANR strategy on SE strategy in [9]. These studies will be continued in order to find a final answer to the question : is it better to add noise to the references or to subtract noise from the test utterances?

Acknowlegement: This work was partly supported by the EEC within the ESPRIT-ARS project.

References

[1] Barbier L., Chollet G.,"Robust speech parameters extraction for word recognition in noise using neural networks", ICASSP, N° S2.27, pp. 145-148, 1991.

[2] Barbier L., "Recognizers adaptation to new environments by neuronal techniques", PhD thesis, Télécom-Paris (in french), Paris, 1992.

[3] Berouti M., Schwartz R. et Makhoul J., "Enhancement of speech corrupted by acoustic noise", ICASSP, pp. 208-211, 1979.

[4] Compernolle D.V., "Noise adaptation in a hidden Markov model speech recognition system", Computer Speech and Language, Vol. 3, pp. 151-167, 1989.

[5] Juang B.H., Rabiner L. R. and Wilpon J.G., "On the use of bandpass liftering in speech recognition", IEEE trans. on ASSP, Vol. 35, N° 7, pp. 947-954, 1987.

[6] Juang B.H., "Recent developments in speech recognition under adverse conditions", ICSLP, pp. 1113-1116, 1990.

[7] Lockwood P., Boudy J., "Experiments with non-linear spectral subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars", Eurospeech, Vol. 1, pp. 79-82, 1991.

[8] Mokbel C., Chollet G., "Speech recognition in adverse environments : Speech enhancement / Spectral transformations", ICASSP, pp. 925-928, 1991.

[9] Mokbel C., "Speech recognition in presence of noise : Speech enhancement / Adding noise to references", PhD thesis (in french), Télécom-Paris, Paris, 1992.