ROBUSTNESS OF SPEECH RECOGNITION USING GENETIC ALGORITHMS AND A MEL-CEPSTRAL SUBSPACE APPROACH

S.A. Selouani[*], *D. O'Shaughnessy*[**]

[*]Université de Moncton, Campus de Shippagan, Canada,
[**]INRS-EMT, Montréal, Canada,
selouani@umcs.caT, dougo@inrs-emt.uquebec.ca

ABSTRACT

This paper presents a method to compensate cepstral coefficients (MFCCs) for a HMM-based speech recognition system evolving under telephone-channel degradations. The technique we propose is based on the combination of the Karhonen-Loève Transform (KLT) and Genetic Algorithms (GA). The idea consists of projecting the band-limited MFCCs onto a subspace generated by the genetically optimized KLT principal axes. Experiments show a clear improvement when the method was applied to the NTIMIT telephone speech database. Word recognition results obtained on the HTK toolkit platform using *N*-mixture tri-phone models and a bigram language model are presented and discussed.

1. INTRODUCTION

Limiting the decrease of Continuous Speech Recognition (CSR) system performances due to acoustic environment changes constitutes a very important issue. It has been observed that when modifying a CSR system whose models were trained in a clean conditions to handle real world environments, its accuracy dramatically degrades. Mismatches between training and test data are the roots of this drawback [1][5].

In order to face this difficulty many techniques have been developed. They are centered upon two major problems. One is how to establish a compensation method for clean models in order to adapt to new environments. Another assumes that noisy data is available and proposes to retrain a robust set of models. However, most of the current approaches assume that the speech and noise are additive in the linear power domain and the noise is stationary [1].

Investigating innovative strategies becomes essential in order to overcome the limits of noise-dependant methods. In this context, Genetic Algorithms (GAs) can constitute robust solutions since they demonstrate their power to investigate beyond the classical space of solutions by exploring a wide range of promising areas [3][4]. The approach we propose can be viewed as a signal transformation via a mapping operator using a Mel-Frequency subspace decomposition and Genetic Algorithms. This transformation attempts to achieve an adaptation of CSR systems under a telephone-channel degradation.

The idea consists of projecting noisy data, without any assumption about noise, onto an optimized subspace generated by principal axes acquired in a canonical environment (i.e., without noise) through the use of the KLT. The optimization of principal axes is performed using genetic operators such as mutations and crossovers in order to adapt the CSR to the new (telephone channel) environment.

This paper is organized as follows. In section 2 we describe the general framework of our approach. Section 3 reports the model linking the KLT to the evolution paradigm; then we proceed in Section 4 to describe the genetic operators and other evolution parameters that we used in our system. Section 5 presents and discusses the results obtained by using the proposed KLT-GA-based CSR system on telephone speech.

2. GENERAL FRAMEWORK

The recognition process aims to provide the most likely phone sequence w' given the acoustic data o. This estimation is performed by maximizing *a posteriori* (MAP) the $p(w/o)$ probability.

$$w' = \arg\max_{w \in \Psi} p(w/o), \quad (1)$$

where w is the reference sequence of phones (or words) that produces a sequence of observable acoustic data o, sent through a noisy transmission channel. Ψ is the set of all possible phone sequences. If we consider $p(w)$, the prior probability determined by the language model and $p(o/w)$ the conditional probability that the acoustic channel produces the sequence o, equation (1) can be written :

$$w' = argmax_{w \in \Psi} p(o/w)p(w). \quad (2)$$

Let Λ be the set of models used by the recognizer to decode acoustic parameters through the use of the MAP. Then equation (2) can be written as follows:

$$w' = argmax_{w \in \Psi} p(o/w,\Lambda)p(w). \quad (3)$$

The mismatch between the training and the testing environments yields a corresponding mismatch in the likelihood of o given Λ and consequently involves a breakdown of CSR systems. Decreasing this mismatch should increase the correct recognition rate.

The mismatch can be viewed by considering the signal space, the feature space, or the model space. In our method we are concerned with the feature space. We consider a transformation T that maps Λ into a transformed feature space. Our approach to decreasing the mismatch between o and Λ is to find T' and the phone sequence w' that maximizes the joint likelihood of o and w given Λ:

$$[T',w'] = argmax_{w \in \Psi} p(o/w,T,\Lambda)p(w). \quad (4)$$

In the approach we propose, a pseudo-joint maximization over w and T is performed where the typical conventional Hidden Markov Models-based technique is used to estimate w, and where a GA-based technique is used to enhance noisy data iteratively by keeping noisy features as close as possible to the clean data. This GA-based transformation aims at reducing the mismatch between training and operating conditions by giving the HMM the ability to recall the training conditions.

The individuals of the population used in the GA are composed of principal axes obtained after a KLT over noisy Mel-Frequency Cepstral Coefficients (MFCCs). These axes evolve through generations and become evolutionarily adapted to the noisy environment. The fittest individuals (best principal axes) are used to project the new incoming noisy data and then these enhanced data are fed into the HMM-based recognizer.

3. ROBUSTNESS USING THE KARHONEN-LOÈVE TRANSFORM

The principle of the KLT, applied in the context of noise reduction, is based on the decomposition of the space of the noisy signal into a signal-plus-noise subspace and a noise subspace. As presented in [1], enhancement is performed by removing the noise subspace and estimating the clean signal from the remaining signal space. This estimation is done by projecting the noisy vectors in the subspace generated by the low-order components of KLT, given the fact that the high-order eigenvalues are more sensitive to noise than the low-order ones. We have applied the KLT in the Mel-scale domain in the context of additive car noise [6]. The KLT using the zero-mean noisy MFCC vector, $\tilde{C}$, can be expressed as follows:

$$\tilde{C} = \sum_{k=1}^{N} \alpha_k \beta_k', \quad (5)$$

where the coefficients α_k are the principal components of the KLT. They are given by the projection of the vector in the space represented by the N-eigenvectors basis, β'. The dimension of the MFCCs is N. Our idea in this paper is to consider β' as an initial population of individuals for an evolution process. The components of these vectors are then viewed as the genes of these individuals and are submitted to genetic operators such as mutations and crossovers in order to find the best β' according an evaluation function (fitness).

4. GENETIC PARAMETERS & OPERATORS

GAs have become an increasingly appreciated and well-understood paradigm beyond the ALife community. Their principle consists of maintaining and manipulating a population of solutions and implementing a 'survival of the fittest' strategy in their search for better solutions. The fittest individuals of any population are encouraged to reproduce and survive to the next generation, thus improving successive generations. However, a proportion of inferior individuals can, by chance, survive and also reproduce. A more complete presentation of GAs can be found in the book of Michalewicz [4].

For any GA, a chromosome representation is needed to describe each individual (axis) in the population. The representation scheme determines how the problem is structured in the GA and also determines the genetic operators that are used. Our application involves genes (a component of an axis) from an alphabet of floating point numbers with values within the variables upper and lower bounds. The real-valued GAs are preferred to binary GAs since real-valued representation offers higher precision with more consistent results across replications [4].

4.1. Initial and final conditions

The ideal, zero-knowledge assumption is to start with an initial population composed of KLT principal axes issued from a set of noisy MFCCs. We choose to end the evolution process when the population gets homogeneity in performances. In other words when we observe that children do not surpass their parents the evolution process is ended. Our stop criteria can viewed as the convergence according to a stabilization of performances.

4.2. Evolving process

In order to keep evolving strategies simple while allowing adaptation behavior, stochastic selection of individuals is used. The selection of individuals to produce successive generations is based on the assignment of a probability of selection, P_j to each individual, j according to its fitness value. The roulette wheel selection method [3] [4] can be used. The probability P_j is calculated as follows:

$$P_j = \frac{F_j}{\sum_{k=1}^{PopSize} F_k}, \quad (6)$$

where F_k equals the fitness of individual k and *PopSize* is the population size. The fitness function is defined in terms of a distance measure between noisy MFCCs projected on a given candidate axis β' and the clean MFCCs. The Euclidian distance is used considering the fact that it is the most adapted measure in the cepstral domain. The general algorithm describing the evolution process is given in Figure 1.

4.3. Genetic operators

Genetic operators are used to create new solutions from the available solutions in the population. Crossovers and mutations constitute the basic types of operators. A crossover creates from two individuals (parents) two new individuals (children) while a mutation changes the genes of one individual to produce a new one (mutant).

A simple crossover method can be used. It generates a random number r from a uniform distribution and does an exchange of the genes of the parents (X and Y) on the children's genes (X' and Y'). It can be expressed by the following equations:

$$\begin{cases} X' = rX + (1-r)Y \\ Y' = (1-r)X + rY \end{cases} \quad (7)$$

The mutation operator consists of randomly selecting from a given percentage of individuals a number of their components (genes) and setting them equal to uniform random numbers.

We have shown in [6] that the use of the heuristic crossover and the non-uniform mutation [3] are appropriate in the case of additive noise. In the context of telephone speech the experiments using simple crossover and mutation seems sufficient to get the best β' given in equation (5). However, the number of generations is greater in the case of telephone speech. The values for the genetic parameters given in Table 1 were selected after extensive cross-validation experiments and were shown to perform well with all data.

Fix the number of generations Gen_{max} and boundaries of axes
Generate for each KLT component a population of axes
For Gen_{max} generations **Do**
    For each set of components **Do**
        Project noisy data using KLT axes
        Evaluate the global fitness function
    End for
    Select and **Reproduce**
End For
Project noisy data onto space generated by the best individuals

Fig. 1. Algorithm for an evolutionary search technique for the best KLT axes.

5. EXPERIMENTS & RESULTS

5.1 Speech material and CSR platform

In order to study the impact of telephone-channel degradation on recognition accuracy of both baseline and enhanced CSR systems, the NTIMIT database was used. The NTIMIT database, described in [2], was created by transmitting sentences in the TIMIT database over long distance telephone lines.

In our experiments, the training set composed of *dr1* and *dr2* subdirectories of the TIMIT database was used to train a set of clean speech models. The speech recognition system used the *dr1* subdirectory of NTIMIT as a test set. 12 MFCCs were calculated on a 30-msec Hamming window advanced by 10 msec. This static vector is expanded after the KLT-GA processing to produce a 36-dimensional (static+first+second derivatives) noted MFCC_D_A, the vector upon which the HMMs were tested. We have evaluated KLT- and KLT-GA-based CSR systems and a baseline HMM system by using the HTK platform [7] and tri-phone Gaussian *N*-mixture models.

5.2. Speech under telephone-channel degradation

Previous work has demonstrated that the use of speech over telephone lines increases the rate of recognition errors. For instance, Moreno and Stern [5] report an approximately 30% recognition rate by using TIMIT as a training database and the NTIMIT database for the test. When speech is recorded through telephone lines, a reduction in the analysis bandwidth yields higher recognition error, particularly when the system is trained

Genetic Parameters	Value
Number of generations	500
Population size	250
Crossover rate	0.25
Mutation rate	0.06
Number of runs	55
Number of frames	114331
Boundaries	[-1.0 , +1.0]

Table 1. Values of the parameters used in the genetic algorithm.

with high-quality speech and tested using simulated telephone speech.

In our experiments a population of 250 individuals is generated for each β'_k and evolves during 500 generations. The maximum number of generations needed and the population size are well adapted to our problem since no improvement was observed when these parameters were increased. The values of the GA parameters used in our experiments are given in Table1. Through these experiments, we found that using the KLT-GA approach to enhance the MFCCs that were used for recognition with *N*-mixture Gaussian HMMs for *N*=1, 2, 4 and 8, using tri-phone models, leads to an important improvement in the accuracy of the word recognition rate. As shown in Figure 2, this improvement can reach 27% when MFCC_D_A- and KLT-GA-MFCC_D_A-based CSR systems are considered. Experiments show that substitution and insertion errors are considerably reduced when the KLT-GA-based approach is included, which gives more effectiveness to the CSR system. A correct rate of 45% is reached by the KLT-GA-MFCC_D_A-based CSR system when the baseline and the KLT-baseline systems achieve 18% and 17% respectively.

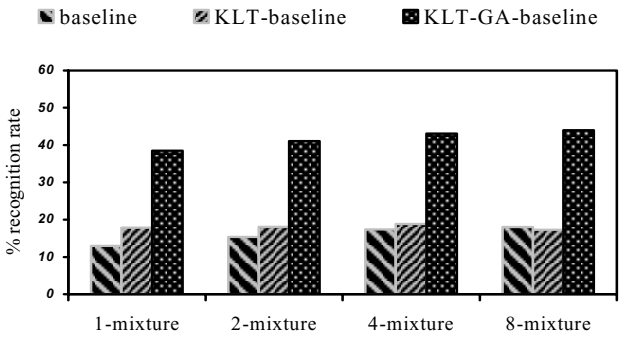

Fig. 2. Percentages of word recognition rate of the MFCC_D_A_-, KLT-MFCC_D_A-, KLT-GA-MFCC_D_A-based HTK CSR systems using 1-mixture, 2-mixture, 4-mixture and 8-mixture tri-phone models. The MFCC_D_A is the baseline system. The training is carried out on the TIMIT database and tested on the NTIMIT database.

6. CONCLUSION

We have illustrated the suitability of GAs for an important real-world application. The approach we have proposed overcomes many of the limitations found in existing CSR systems when they are submitted to a telephone-channel degradation. An important improvement (about 27 %) is reached when we compared a baseline HMM-based system, a classic KLT-based method and our KLT-GA-based technique. The main advantage of our method is that it does not require any *a priori* knowledge about the noise. In the near future, experiments will be carried out in order to apply our approach to speaker adaptation. In addition, and in order to gain more insight into the important question concerning the fitness function, we will test the feasibility of an online evaluation function linked to the phone identification accuracy.

7. REFERENCES

[1] Y. Ephraim, and H.L. Van Trees, "A signal subspace approach for speech enhancement", IEEE trans. On Speech and Audio Process., Vol.3, pp. 251-266, 1995.

[2] W.M. Fisher, et al, " The DARPA speech recognition database: specification and status", DARPA workshop on speech recognition, pp. 93-99, 1986.

[3] C.R Houk, J. A. Joines, and M. G. Kay, "A genetic algorithm for function optimization: a MATLAB implementation", North Carolina University-NCSU-IE, technical report 95-09, 1995.

[4] Z. Michalewicz, *Genetic Algorithms + Data Structure = Evolution Programs Adaptive,* AI series, Springer-Verlag, New York, 1996.

[5] P.J. Moreno and R Stern "Sources of degradation of speech recognition in the telephone network", Proceedings of ICASSP'94, Vol. 1 pp. 109-112, 1994.

[6] S.A Selouani and D. O'Shaughnessy, "Noise-robust speech recognition in car environments using genetic algorithms and a Mel-Cepstral subspace approach", Proceedings of ICSLP'02, pp. 2173-2176, 2002.

[7] Cambridge university Speech Group, "The HTK book (version 3.0)", Cambridge University Group, 2000.