EURASIP Journal on Applied Signal Processing 2003:8, 814–823

On the Use of Evolutionary Algorithms to Improve the Robustness of Continuous Speech Recognition Systems in Adverse Conditions

Sid-Ahmed Selouani

Secteur Gestion de l'Information, Université de Moncton, Campus de Shippagan, 218 boulevard J.-D.-Gauthier, Shippagan, Nouveau-Brunswick, Canada E8S 1P6
Email: selouani@umcs.ca

Douglas O'Shaughnessy

INRS-Energie-Matériaux-Télécommunications, Université du Québec, 800 de la Gauchetière Ouest, place Bonaventure, Montréal, Canada H5A 1K6
Email: dougo@inrs-telecom.uquebec.ca

Received 14 June 2002 and in revised form 6 December 2002

Limiting the decrease in performance due to acoustic environment changes remains a major challenge for continuous speech recognition (CSR) systems. We propose a novel approach which combines the Karhunen-Loève transform (KLT) in the mel-frequency domain with a genetic algorithm (GA) to enhance the data representing corrupted speech. The idea consists of projecting noisy speech parameters onto the space generated by the genetically optimized principal axis issued from the KLT. The enhanced parameters increase the recognition rate for highly interfering noise environments. The proposed hybrid technique, when included in the front-end of an HTK-based CSR system, outperforms that of the conventional recognition process in severe interfering car noise environments for a wide range of signal-to-noise ratios (SNRs) varying from 16 dB to −4 dB. We also showed the effectiveness of the KLT-GA method in recognizing speech subject to telephone channel degradations.

Keywords and phrases: speech recognition, genetic algorithms, Karhunen-Loève transform, hidden Markov models, robustness.

1. INTRODUCTION

Continuous speech recognition (CSR) systems remain faced with the serious problem of acoustic condition changes. Their performance often degrades due to unknown adverse conditions (e.g., due to room acoustics, ambient noise, speaker variability, sensor characteristics, and other transmission channel artifacts). These speech variations create mismatches between the training data and the test data. Numerous techniques have been developed to counter this in three major areas [1].

The first area includes noise masking [1], spectral and cepstal substraction [2], and the use of robust features [3]. Robust feature analysis consists of using noise-resistant parameters such as auditory-based features, mel-frequency cepstral coefficients (MFCC) [4], or techniques such as relative spectral (RASTA) methodology [5]. The second type of method refers to the establishment of compensation models for noisy environments without modification to the speech signal. The third field of research is concerned with distance and similarity measurements. The major methods of this field are founded on the principle to find a robust distorsion measure that emphasizes the regions of the spectrum that are less influenced by noise [6].

Despite these efforts to address robustness, adapting to changing environments remains the major obstacle to speech recognition in practical applications. Investigating innovative strategies has become essential to overcome the drawbacks of classical approaches. In this context, evolutionary algorithms (EAs) are robust solutions, and they are useful to find good solutions to complex problems (artificial neural networks topology or weights for instance) and to avoid local minima [7]. Applying artificial neural networks, Spalanzani [8] showed that recognition of digits and vowels can be improved by using genetically optimized initialization of weights and biases. In this paper, we propose an approach which can be viewed as a signal transformation via a mapping operator using a mel-frequency space decomposition based on the Karhunen-Loève transform (KLT) and a genetic algorithm (GA) with a real-coded encoding (a part of EAs). This transformation attempts to adapt hidden Markov model-based CSR systems for adverse conditions. The principle consists of finding in the learning phase the principal axes generated by the KLT and then optimizing them for the projection of noisy data by genetic operators. The aim is to provide projected noisy data that are as close as possible to clean data.

This paper is organized as follows. Section 2 describes the basis of our proposed hybrid KLT-GA enhancement method. Section 3 describes the model linking the KLT to the evolution mechanism, which leads to a robust representation of noisy data. Then, Section 4 describes the database, the platform used in our experiments and the evaluation of the proposed KLT-GA-based recognizer in a noisy car environment and in a telephone channel environment. This section includes the comparison of KLT-GA processed recognizers to a baseline CSR system in order to evaluate performance. Finally, Section 5 concludes with a perspective of this work.

2. OVERALL STRUCTURE OF THE KLT-GA-BASED ROBUST SYSTEM

2.1. General framework

CSR systems based on statistical models such as hidden Markov models (HMM) automatically recognize speech sounds by comparing their acoustic features with those determined during training [9]. A bayesian statistical framework underlies the HMM-speech recognizer. The development of such a recognizer can be summarized as follows. Let w be a sequence of phones (or words), which produces a sequence of observable acoustic data o, sent through a noisy transmission channel. In our study, telephone speech is corrupted by additive noise. 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 $p(w \mid o)$ probability:

$$w' = \underset{w\in\Psi}{\operatorname{argmax}}\, p(w \mid o) = \underset{w\in\Psi}{\operatorname{argmax}}\, p(o \mid w)p(w), \qquad (1)$$

where Ψ is the set of all possible phone sequences, $p(w)$ is the prior probability, determined by the language model, that the speaker utters w, and $p(o \mid w)$ is the conditional probability that the acoustic chanel produces the sequence o. Let Λ be the set of models used by the recognizer to decode acoustic parameters through the use of the MAP. Then (1) becomes

$$w' = \underset{w\in\Psi}{\operatorname{argmax}}\, p(o \mid w, \Lambda)p(w). \qquad (2)$$

The mismatch between the training and the testing environments leads to a worse estimate for the likelihood of o given Λ and thus degrades CSR performance. Reducing 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. We are concerned with the feature space, and consider a transformation T that maps Λ into a transformed feature space. Our approach is to find T' and the phone sequence w' that maximize the joint likelihood of o and w given Λ:

$$[T', w'] = \underset{w\in\Psi}{\operatorname{argmax}}\, p(o \mid w, T, \Lambda)p(w). \qquad (3)$$

We propose a pseudojoint maximization over w and T, where the typical conventional HMM-based technique is used to estimate w, while an EA-based technique enhances noisy data iteratively by keeping the noisy features as close as possible to the clean data. This EA-based transformation aims to reduce the mismatch between training and operating conditions by giving the HMM the ability to "recall" the training conditions.

As is shown in Figure 1, the idea is to manipulate the axes generating the feature representation space to achieve a better robustness on noisy data. MFCCs serve as acoustic features. A Karhunen-Loève decomposition in the MFCC domain allows obtaining the principal axes that constitute the basis of the space where noisy data is represented. Then, a population of these axes is created (corresponding to individuals in the initialization of the evolution process). The evolution of the individuals is performed by EAs. The individuals are evaluated via a fitness function by quantifying, through generations, their distance to individuals in a noise-free environment. The fittest individual (best principal axes) is used to project the noisy data in its corresponding dimension. Genetically modified MFCCs and their derivatives are finally used as enhanced features for the recognition process.

2.2. Cepstral acoustic features

The cepstrum is defined as the inverse Fourier transform of the logarithm of the short-term power spectrum of the signal. The use of a logarithmic function allows deconvolution of the vocal tract transfer function and the voice source. Consequently, the pulse sequence corresponding to the periodic voice source reappears in the cepstrum as a strong peak in the "frequency" domain. The derived cepstral coefficients are commonly used to describe the short-term spectral envelope of a speech signal. The computation of MFCCs requires the selection of M critical bandpass filters that roughly approximate the frequency response of the basilar membrane in the cochlea of the inner ear [4]. A discrete cosine transform, C_n, is applied to the output of M filters, X_k. These filters are triangular, cover the 156–6844 Hz frequency range, and are spaced on the mel-frequency scale. These filters are applied to the log of the magnitude spectrum of the signal, which is estimated on a short-time basis. Thus

$$C_n = \sum_{k=1}^{M} X_k \cos\left(\frac{\pi n}{M}(k - 0.5)\right), \quad n = 1, 2, \dots, N, \qquad (4)$$

where N is the number of the cepstral coefficients, M is the analysis order, and X_k, $k = 1, 2, \dots, M = 20$, represents the log-energy output of the kth filter.

2.3. KLT in the mel-frequency domain

In order to reduce the effects of noise on ASR, many methods propose to decompose the vector space of the noisy signal into a signal-plus-noise subspace and a noise subspace [10]. We remove the noise subspace and estimate the clean signal from the remaining signal space. Such a decomposition applies the KLT to the noisy zero-mean normalized data.

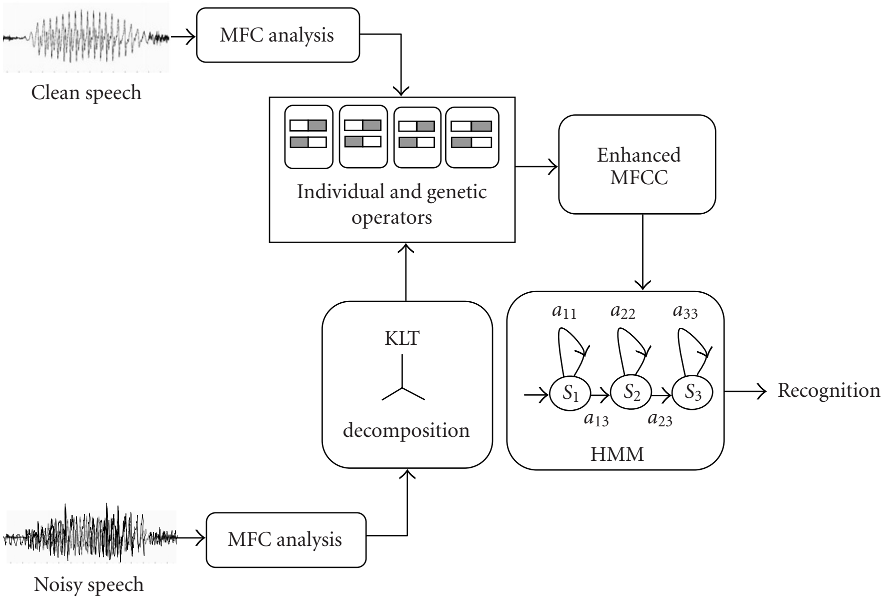


Figure 1: General overview of the KLT-EA-based CSR robust system.

If we apply such a decomposition over the noisy zero-mean normalized MFCC vector $\hat{\mathbf{C}} = [\hat{C}_1, \hat{C}_2, \ldots, \hat{C}_N]^T$ with the assumption that $\hat{\mathbf{C}}$ has a symmetric nonnegative autocorrelation matrix $R = \mathscr{E}[\hat{\mathbf{C}}^T\hat{\mathbf{C}}]$ with a rank $r \le N$, then $\hat{\mathbf{C}}$ can be represented as a linear combination of eigenvectors $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_r$, which correspond to eigenvalues $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_r \ge 0$, respectively. That is, $\hat{\mathbf{C}}$ can be calculated using the following orthogonal transformation:

$$\hat{\mathbf{C}} = \sum_{k=1}^{r} \alpha_k \boldsymbol{\beta}_k, \quad k = 1, \ldots, r, \tag{5}$$

where the coefficients α_k (principal components) are given by the projection of $\hat{\mathbf{C}}$ in the space generated by the r-eigenvector basis. Given that the magnitudes of low-order eigenvalues are higher than for the high-order ones, the effect of the noise on the low-order eigenvalues is proportionately less than that for high-order ones. Thus, a linear estimation of the clean vector $\mathbf{C}$ is performed by projecting the noisy vectors on the space generated by principal components weighted by a function W_k, which applies strong attenuation over higher-order eigenvectors depending on the noise variance [10]. The enhanced MFCCs are then given by

$$\tilde{\mathbf{C}} = \sum_{k=1}^{r} W_k \alpha_k \boldsymbol{\beta}_k, \quad k = 1, \ldots, r. \tag{6}$$

Various methods can find the adequate weighting function, particularly in the case of signal subspace decomposition [10]. The optimal order r fixing the beginning of the strong attenuation must be determined. In our new approach, GAs determine optimal principal components. No assumptions need to be made. Optimization is achieved when vectors $\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2, \ldots, \boldsymbol{\beta}'_N$, which do not correspond necessarily to the eigenvectors, minimize the Euclidean distance between $\hat{\mathbf{C}}$ and $\mathbf{C}$. The genetically enhanced MFCCs, $\tilde{\mathbf{C}}_{\text{Gen}}$, are

$$\tilde{\mathbf{C}}_{\text{Gen}} = \sum_{k=1}^{N} \alpha_k \boldsymbol{\beta}'_k, \quad k = 1, \ldots, N. \tag{7}$$

Determining an optimal r is not needed since the GA considers vectors $\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2, \ldots, \boldsymbol{\beta}'_N$ as the fittest individuals for the complete space dimension N. This process can be regarded as the mapping transform, T', of (3).

3. MODEL DESCRIPTION AND EVOLUTION

The use of GAs requires resolution of six fundamental issues: the chromosome (or solution) representation, the selection function, the genetic operators making up the reproduction function, the creation of the initial population, the termination criteria, and the evaluation function [11, 12]. The GA maintains and manipulates a family or population of solutions (the $\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2, \ldots, \boldsymbol{\beta}'_N$ vectors in our case) and implements a "survival of the fittest" strategy in its search for better solutions.

3.1. Solution representation

A chromosome representation describes each individual in the population. It is important since the representation scheme determines how the problem is structured in the GA and also determines the adequate genetic operators to use [13]. For our application, the useful representation of an individual or chromosome for function optimization involves genes or variables from an alphabet of floating-point numbers with values within the variables' upper and lower bounds (resp., +1 and −1). Michalewicz [14] has done extensive experimentation comparing real-valued and binary GAs, and has shown that real-valued representation offers higher precision with more consistent results across replications.

3.2. Selection function

Stochastic selection is used to keep search strategies simple while allowing adaptivity. The selection of individuals to produce successive generations plays an extremely important role in GAs. A common selection approach assigns a probability of selection, P_j, to each individual, j, based on its fitness value. Various methods exist to assign probabilities to individuals; we use the normalized geometric ranking [15]. This method defines P_j for each individual by

$$P_j = q'(1-q)^{s-1}, \tag{8}$$

where

$$q' = \frac{q}{1-(1-q)^P}, \tag{9}$$

where q is the probability of selecting the best individual, s the rank of the individual (1 being the best), and P the population size.

3.3. Genetic operators

The basic search mechanism of the GA is provided by two types of operators: crossover and mutation. Crossover transforms two individuals into two new individuals, while mutation alters one individual to produce a single solution. A float representation of the parents is denoted by $\overline{X}$ and $\overline{Y}$. At the end of the search, the fittest individual survives and is retained as an optimal KLT axis in its corresponding rank of $\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2, \ldots, \boldsymbol{\beta}'_N$ vectors.

3.3.1 Crossover

Crossover operators combine information from two parents and transmit it to each offspring. In order to avoid the extension of the exploration domain of the best solution, we preferred to use a crossover that utilizes fitness information, that is, a heuristic crossover [15]. Let a_i and b_i be the lower and upper bound, respectively, of each component x_i representing a member of the population ($\overline{X}$ or $\overline{Y}$). This operator produces a linear interpolation of $\overline{X}$ and $\overline{Y}$. New individuals $\overline{X'}$ and $\overline{Y'}$ (children) are created according Algorithm 1.

```
1. Fix $g = U(0,1)$, uniform random number
2. Compute fit[$\overline{X}$] and fit[$\overline{Y}$], fitness of $\overline{X}$ and $\overline{Y}$
3. If fit[$\overline{X}$] > fit[$\overline{Y}$]
      Then $\overline{X}' = \overline{X} + g(\overline{X} - \overline{Y})$ and $\overline{Y}' = \overline{X}$
      Estimate feasibility of $\overline{X}'$:
$\mathcal{F}(\overline{X}') = \begin{cases} 1 & \text{if } a_i \le x'_i \le b_i \quad \forall i \\ 0 & \text{otherwise} \end{cases}$
 $x'_i$ components of $\overline{X}'$,    $i = 1, \ldots, N$
4. If $\mathcal{F}(\overline{X}') = 0$
      Then generate  new $g$; goto 2
5. If all individuals reproduced then Stop
        else goto 1
```

Algorithm 1: The heuristic crossover used in the CSR robust system.

3.3.2 Mutation

Mutation operators tend to make small random changes in an attempt to explore all regions of the solution space [16]. The principle of a nonuniform mutation used in our application consists of randomly selecting one component, x_k, of an individual and setting it equal to a nonuniform random number,[1] x'_k:

$$x'_k = \begin{cases} x_k + (b_k - x_k)f(\text{Gen}) & \text{if } u_1 < 0.5, \\ x_k - (a_k + x_k)f(\text{Gen}) & \text{if } u_1 \ge 0.5, \end{cases} \tag{10}$$

where the function $f(\text{Gen})$ is given by

$$f(\text{Gen}) = \left(u_2\left(1 - \frac{\text{Gen}}{\text{Gen}_{\max}}\right)\right)^t, \tag{11}$$

where u_1, u_2 are uniform random numbers between $(0, 1)$, t a shape parameter, Gen the current generation, and $\text{Gen}_{\max}$ the maximum number of generations. The multi-nonuniform mutation generalizes the application of the nonuniform mutation operator to all the components of the parent $\overline{X}$. The main advantage of this operator is that the alteration is distributed on all individual components which lead to the extension of the search space and then permit to deal with any kind of noise.

3.4. Evaluation function

The GA must search all the axes generated by the KLT of the mel-frequency space (that make the noisy MFCCs if they are projected into these axes) to find the closest to the clean MFCC. Thus, evolution is driven by a fitness function defined in terms of a distance measure between the noisy MFCC projected on a given individual (axis) and the clean MFCC. The fittest individual is the axis which corresponds to the minimum of that distance. The distance function applied to cepstral (or other voice representations) refers to *spectral distorsion measures* and represents the cost in a classification system of speech frames. For two vectors C and $\hat{C}$ representing two frames [6], each with N components, the geometric distance is defined as

$$d(C, \hat{C}) = \left(\sum_{k=1}^{N} (C_k - \hat{C}_k)^l\right)^{1/l}. \tag{12}$$

For simplicity, the Euclidean distance is considered ($l = 2$), which has been a valuable measure for both clean and noisy speech [6, 17]. Figure 2 gives for the first four best axes the evolution of their fitness (distorsion measure) through 300 generations. Note that $-d(C, \hat{C})$ is used as a distance measure because the evaluation function must be maximized.

[1] Otherwise, the original values of components are maintained.

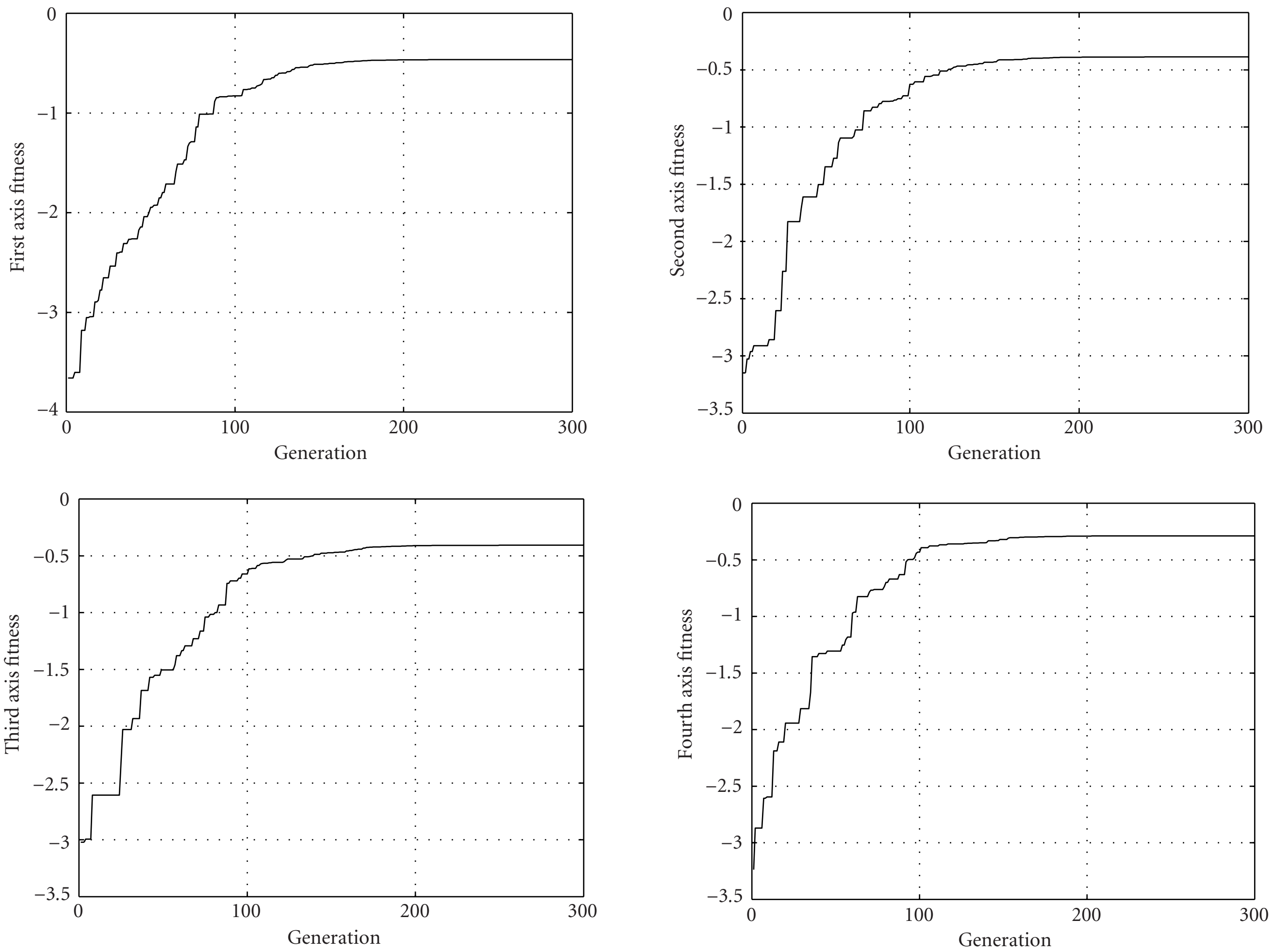


Figure 2: Evolution of the performances of the best individual during 300 generations. Only the four first axes are considered among the twelve.

3.5. Initialization and termination

The ideal, zero-knowledge assumption starts with a population of completely random axes. Another typical heuristic, used in our system, initializes the population with a uniform distribution in a default set of known starting points described by the boundaries (a_i, b_i) for each axis component. The GA-based search ends when the population gets homogeneity in performance (when children do not surpass their parents), converges according to the Euclidean distortion measure, or is terminated by the user if the number of maximum generations is reached. Finally, the evolution process can be summarized in Algorithm 2.

4. EXPERIMENTS

4.1. Speech material

The following experiments used the TIMIT database [18], which contains broadband recordings of a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States, each reading 10 phonetically rich sentences. To simulate a noisy environment, car noise was added artificially to the clean speech. To study the effect of such noise on the recognition accuracy of the CSR system that we evaluated, the reference templates for all tests were taken from clean speech. The training set is composed of 1140 sentences (114 speakers) of *dr1* and *dr2* TIMIT subdirectories. On the other hand, the *dr1* subset of the TIMIT database, composed of 110 sentences, was chosen to evaluate the recognition system.

In a second set of experiments, and 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 [19]. It was created by transmitting speech from the TIMIT database over long-distance telephone lines. Previous work has demonstrated that telephone line use increases the rate of recognition errors; for example, Moreno and Stern [20] report a 68% error rate by using a version of SPHINX-II [21] as CSR system, TIMIT as training database, and NTIMIT database, for the test.

```
Fix the number of generations Gen$_{max}$ and boundaries of axes
    Generate for each principal KLT component a population of axes
        For Gen$_{max}$ generation Do
            For each set of components Do
                Project noisy data using KLT axes
                Evaluate global Euclidean distance for clean data
            End For
            Select and Reproduce
        End For
 Project noisy data onto space generated by the best individuals
```

Algorithm 2: The evolutionary search technique for best KLT axes.

4.2. CSR platform

In order to test the recognition of continuous speech data enhanced as described above, the HTK-based speech recognizer [22] was used. HTK is an HMM-based toolkit used for isolated or continuous whole-word-based recognition systems. The toolkit supports continuous-density HMMs with any number of state and mixture components. It also implements a general parameter-tying mechanism which allows the creation of complex model topologies. Twelve MFCCs were calculated using a 30-millisecond hamming window advanced by 10 milliseconds for each frame. To do this, an FFT calculates a magnitude spectrum for each frame, which is then averaged into 20 triangular bins arranged at equal mel-frequency intervals. Finally, a cosine transform is applied to such data to calculate the 12 MFCCs which form a 12-dimensional (static) vector. This static vector is then expanded after enhancement to produce a 36-dimensional (static + first and second derivatives: MFCC_D_A) vector upon which the HMMs, that model the speech subword units, were trained. Regarding the used frame length, the 1140 sentences of *dr1* and *dr2* TIMIT subsets provided 342993 frames that were used for the training. The baseline system used a triphone Gaussian mixture HMM system. Triphones were trained through a tree-based clustering method to deal with unseen context. A set of binary questions about phonetic contexts is built; the decision tree is constructed by selecting the best question from the rule set at each node [23].

4.3. Results and discussion

4.3.1 GA parameters

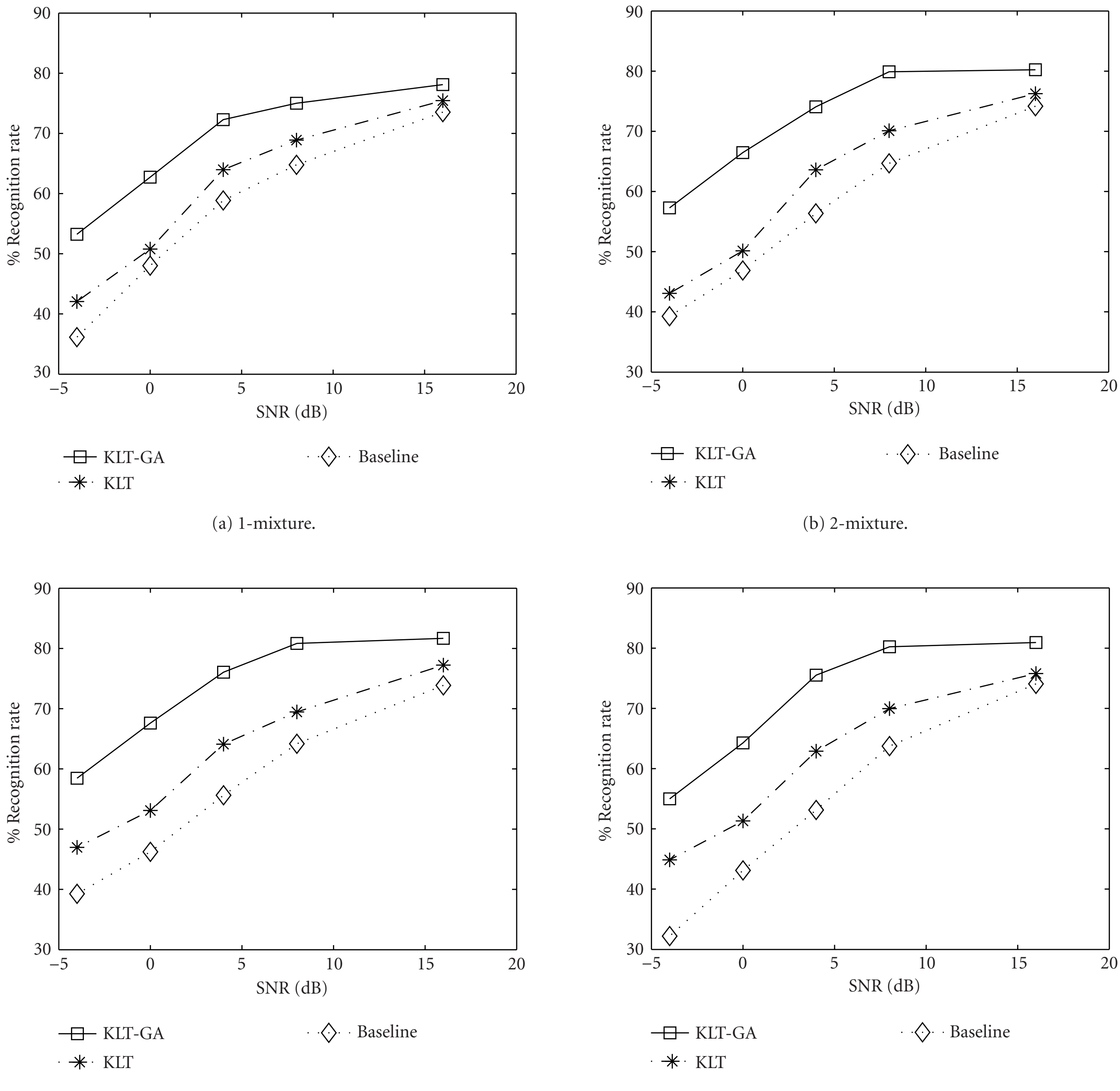
A population of 150 individuals is generated for each $\boldsymbol{\beta}'_k$ and evolves during 300 generations. The values for the GA parameters given in Table 1 were selected after extensive cross-validation experiments and were shown to perform well with all data. The maximum number of generations needed and the population size are well adapted to our problem since no improvement was observed when these parameters were increased. At each generation, the best individuals are retained to reproduce. In the end of the evolution process, the best individuals of the best population are considered as the optimized KLT axes. This method is used by Houk et al. in [15]. For this purpose, data sets are composed of 114331 frames extracted from the TIMIT training subset and corresponding noisy frames extracted from the noisy TIMIT and NTIMIT databases.

Table 1: Values of the parameters used in the GA.

Parameter	Parameter value
Number of generations	300
Population size	150
Probability of selecting the best, q	0.08
Heuristic crossover rate	0.25
Multi-nonuniform mutation rate	0.06
Number of runs	50
Number of frames	114331
Boundaries $[a_i, b_i]$	$[-1.0, +1.0]$

4.3.2 CSR under additive car noise environment

Experiments were done using the noisy version of TIMIT at different values of SNR, from 16 dB to −4 dB. Figure 3 shows that using the KLT-GA-based optimization to enhance the MFCCs that were used for recognition with N-mixture Gaussian HMMs for $N = 1, 2, 4, 8$ with triphone models leads to a higher word recognition rate. The CSR system including the KLT-GA-processed MFCCs performs significantly better than its MFCC_D_A- and KLT-MFCC_D_A-based CSR systems, for low and high noise conditions. The system which contains enhanced MFCCs achieves 81.67% as the best word recognition rate ($\%C_W$) for 16-dB SNR and four Gaussian mixtures. In the same conditions, the baseline system dealing with noisy MFCCs and the system containing KLT-processed MFCCs achieve, respectively, 73.89% and 77.25%. The increased accuracy is more significant in low SNR conditions, which attests to the robustness of the approach when acoustic conditions become severely degraded. For instance, in the −4-dB SNR case, the KLT-GA-MFCC-based CSR system has accuracy higher than KLT-MFCC- and MFCC-based CSR systems, respectively, by 12% and 20%. The comparison between KLT- and KLT-GA-processed



(a) 1-mixture.

(b) 2-mixture.

(c) 4-mixture.

(d) 8-mixture.

Figure 3: Percent word recognition performance (%C_{Wrd}) of the KLT- and KLT-GA-based CSR systems compared to the baseline HTK method (noisy MFCC) using (a) 1-mixture, (b) 2-mixture, (c) 4-mixture, and (d) 8-mixture triphones for different values of SNR.

MFCCs shows that the proposed evolutionary approach is more powerful whatever is the level of noise degradation. Considering the KLT-based CSR, inclusion of the GA technique raised accuracy by about 11%. Figure 4 plots the variations of the first four MFCCs for a signal that has been chosen from the test set. It is clear from the comparison illustrated in this figure that the processed MFCCs, using the proposed KLT-GA-based approach, are less variant than the noisy MFCCs and closer to the original ones.

4.3.3 Speech under telephone channel degradation

Extensive experimental studies characterized the impairments induced by telephone networks [24]. When speech is recorded through telephone lines, a reduction in the analysis bandwidth yields higher recognition error, particularly when the system is trained with high-quality speech and tested using simulated telephone speech [20]. In our experiments, the training set (*dr1* and *dr2* subdirectories of TIMIT) (1140 sentences and 342993 frames) was used to train a set of clean

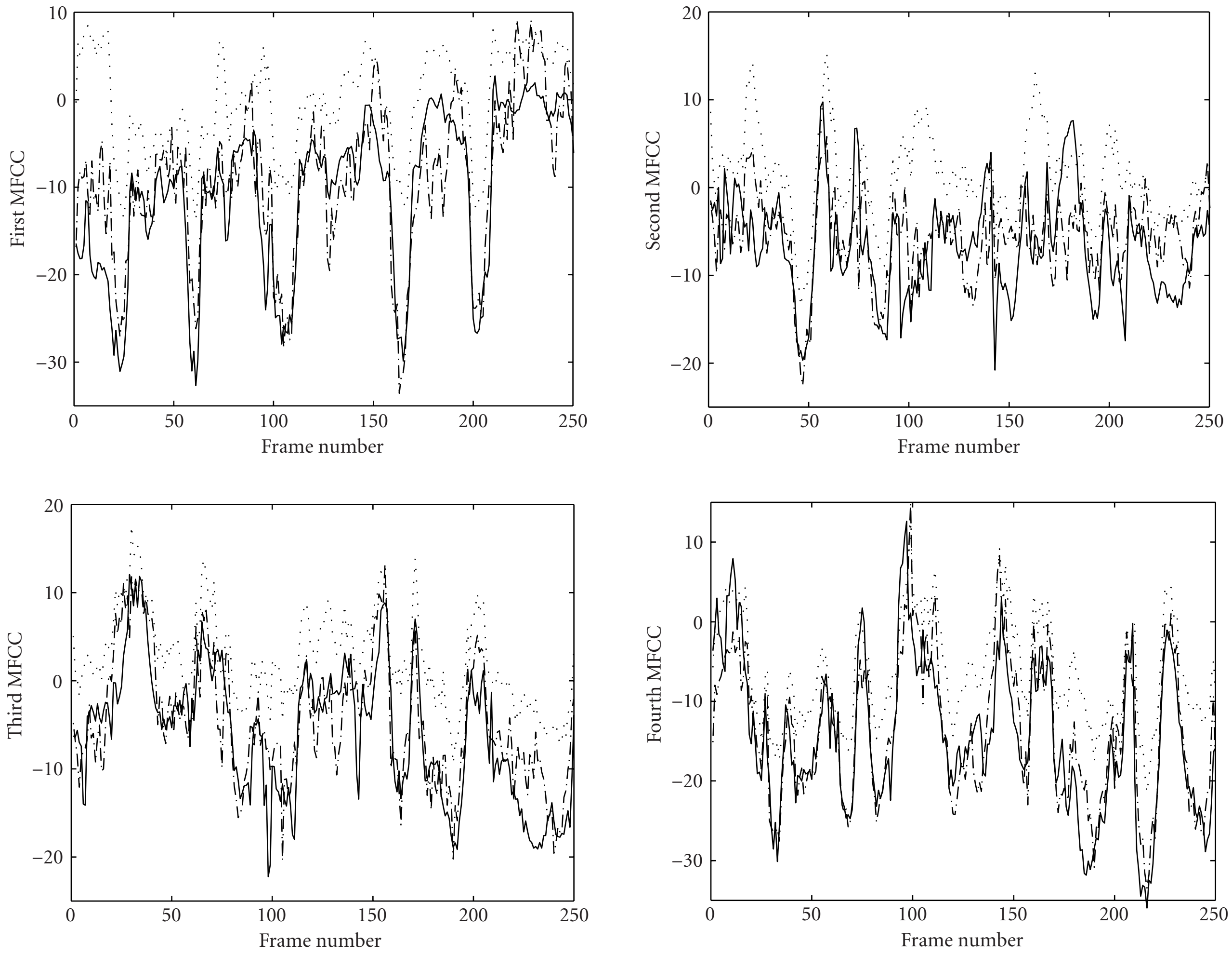


Figure 4: Comparison between clean, noisy, and enhanced MFCCs represented by solid, dotted, dashed-dotted lines, respectively.

speech models. The *dr1* subdirectory of NTIMIT was used as a test set. This subdirectory is composed of 110 sentences and 34964 frames. Speakers and sentences used in the test were different than those used in the training phase. For the KLT- and KLT-GA-based CSR systems, we found that using the KLT-GA as a preprocessing approach to enhance the MFCCs that were used for recognition with N-mixture Gaussian HMMs for $N = 1, 2, 4$, and 8, using triphone models, led to an important improvement in the accuracy of the word recognition rate. Table 2 showed that this difference can reach 27% for MFCC_D_A- and KLT-GA-MFCC_D_A-based CSR systems. Table 2 shows that substitution and insertion errors are considerably reduced when the evolutionary approach is included, which gives more effectiveness to the CSR system.

5. CONCLUSION

We have illustrated the suitability of EAs, particularly the GAs, for an important real-world application by presenting a new robust CSR system. This system is based on the use of a KLT-GA hybrid enhancement noise reduction approach in the cepstral domain in order to get less-variant parameters. Experiments show that the use of the enhanced parameters using such a hybrid approach increases the recognition rate of the CSR process in highly interfering car noise environments for a wide range of SNRs varying from 16 dB to −4 dB and when speech is submitted to the telephone channel degradation. The approach can be applied whatever the distorsion of vectors under the condition to identify the fitness function. The front-end of the proposed KLT-GA-based CSR system does not require any a priori knowledge about the nature of the corrupting noisy signal, which allows dealing with any kind of noise. Moreover, using this enhancement technique avoids the noise estimation process that requires a speech/nonspeech preclassification, which could not be accurate for low SNRs. It is also interesting to note that such a technique is less complex than many other enhancement techniques, which need to either model or compensate for the noise. However, this enhancement technique requires

Table 2: Percentages of word recognition rate (%C_{Wrd}), insertion rate (%ϵ_{Ins}), deletion rate (%ϵ_{Del}), and substitution rate (%ϵ_{Sub}) of the MFCC_D_A-, KLT-MFCC_D_A-, KLT-GA-MFCC_D_A-based HTK CSR systems using (a) 1-mixture, (b) 2-mixture, (c) 4-mixture, and (d) 8-mixture triphone models.

(a) %C_{Wrd} using 1-mixture triphone models.

	%ϵ_{Sub}	%ϵ_{Del}	%ϵ_{Ins}	%C_{Wrd}
MFCC_D_A	82.71	4.27	33.44	13.02
KLT-MFCC_D_A	77.05	5.11	30.04	17.84
KLT-GA-MFCC_D_A	54.48	5.42	25.42	**40.10**

(b) %C_{Wrd} using 2-mixture triphone models.

	%ϵ_{Sub}	%ϵ_{Del}	%ϵ_{Ins}	%C_{Wrd}
MFCC_D_A	81.25	3.44	38.44	15.31
KLT-MFCC_D_A	78.11	3.81	48.89	18.08
KLT-GA-MFCC_D_A	52.40	4.27	52.40	**43.33**

(c) %C_{Wrd} using 4-mixture triphone models.

	%ϵ_{Sub}	%ϵ_{Del}	%ϵ_{Ins}	%C_{Wrd}
MFCC_D_A	78.85	3.75	38.23	17.40
KLT-MFCC_D_A	76.27	4.88	39.54	18.85
KLT-GA-MFCC_D_A	49.69	5.62	25.31	**44.69**

(d) %C_{Wrd} using 8-mixture triphone models.

	%ϵ_{Sub}	%ϵ_{Del}	%ϵ_{Ins}	%C_{Wrd}
MFCC_D_A	78.02	3.96	40.83	18.02
KLT-MFCC_D_A	77.36	5.37	34.62	17.32
KLT-GA-MFCC_D_A	48.41	6.56	26.46	**45.00**

a large amount of data in order to find the "best" individual. Many other directions remain open for further work. Present goals include analyzing evolved genetic parameters, evaluating how performance scales with other types of noise (nonstationary, limited band, etc.).

REFERENCES

[1] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, no. 3, pp. 261–291, 1995.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[3] D. Mansour and B. H. Juang, "A family of distortion measures based upon projection operation for robust speech recognition," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1659–1671, 1989.

[4] S. B. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[5] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "RASTA-PLP speech analysis technique," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, pp. 121–124, San Francisco, Calif, USA, March 1992.

[6] J. Hernando and C. Nadeu, "A comparative study of parameters and distances for noisy speech recognition," in *Proc. Eurospeech '91*, pp. 91–94, Genova, Italy, September 1991.

[7] C. R. Reeves and S. J. Taylor, "Selection of training data for neural networks by a Genetic Algorithm," in *Parallel Problem Solving from Nature*, pp. 633–642, Springer-Verlag, Amsterdam, The Netherlands, September 1998.

[8] A. Spalanzani, S.-A. Selouani, and H. Kabré, "Evolutionary algorithms for optimizing speech data projection," in *Genetic and Evolutionary Computation Conference*, p. 1799, Orlando, Fla, USA, July 1999.

[9] D. O'Shaughnessy, *Speech Communications: Human and Machine*, IEEE Press, Piscataway, NJ, USA, 2nd edition, 2000.

[10] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech, and Audio Processing*, vol. 3, no. 4, pp. 251–266, 1995.

[11] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, Reading, Mass, USA, 1989.

[12] J. Holland, *Adaptation in Natural and Artificial Systems*, The University of Michigan Press, Ann Arbor, Mich, USA, 1975.

[13] L. B. Booker, D. E. Goldberg, and J. H. Holland, "Classifier systems and genetic algorithms," *Artificial Intelligence*, vol. 40, no. 1-3, pp. 235–282, 1989.

[14] Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs*, AI series. Springer-Verlag, New York, NY, USA, 1992.

[15] C. R. Houk, J. A. Joines, and M. G. Kay, "A genetic algorithm for function optimization: a Matlab implementation," Tech. Rep. 95-09, North Carolina State University, Raleigh, NC, USA, 1995.

[16] L. Davis, Ed., *The Genetic Algorithm Handbook*, chapter 17, Van Nostrand Reinhold, New York, NY, USA, 1991.

[17] B. H. Juang, L. R. Rabiner, and J. G. Wilpon, "On the use of bandpass liftering in speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 765–768, Tokyo, Japan, April 1986.

[18] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: specifications and status," in *Proc. DARPA Speech Recognition Workshop*, pp. 93–99, Palo Alto, Calif, USA, February 1986.

[19] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: A phonetically balanced, continuous speech telephone bandwidth speech database," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, pp. 109–112, Albuquerque, NM, USA, April 1990.

[20] P. J. Moreno and R. M. Stern, "Sources of degradation of speech recognition in the telephone network," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, pp. 109–112, Adelaide, Australia, April 1994.

[21] X. D. Huang, F. Alleva, H. W. Hon, M. Y. Hwang, K. F. Lee, and R. Rosenfeld, "The SPHINX-II speech recognition system: An overview," *Computer, Speech and Language*, vol. 7, no. 2, pp. 137–148, 1993.

[22] Cambridge University Speech Group, *The HTK Book (Version 2.1.1)*, Cambridge University Group, March 1997.

[23] L. R. Bahl, P. V. de Souza, P. S. Gopalakrishnan, D. Nahamoo, and M. A. Picheny, "Decision trees for phonological rules in continuous speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pp. 185–188, Toronto, Canada, May 1991.

[24] W. D. Gaylor, *Telephone Voice Transmission. Standards and Measurements*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1989.

Sid-Ahmed Selouani received his B.E. degree in 1987 and his M.S. degree in 1991, both in electronic engineering from the University of Science and Technology of Algeria (U.S.T.H.B). He joined the Communication Langagière et Interaction Personne-Système (CLIPS) Laboratory of Université Joseph Fourier of Grenoble, taking part in the Algerian-French double degree program and then he got a Docteur d'État degree in the field of speech recognition in 2000 from the University of Science and Technology of Algeria. From 2000 to 2002, he held a post-doctoral fellowship in the Multimedia Group at the Institut National de Recherche Scientifique (INRS-Télécommunications) in Montréal. He had teaching experience from 1991 to 2000 in the University of Science and Technology of Algeria before starting to work as an Assistant Professor at the Université de Moncton, Campus de Shippagan. He is also an Invited Professor at INRS-Télécommunications. His main areas of research involve speech recognition robustness and speaker adaptation by evolutionary techniques, auditory front-ends for speech recognition, integration of acoustic-phonetic indicative features knowledge in speech recognition, hybrid connectionist/stochastic approaches in speech recognition, language identification, and speech enhancement.

Douglas O'Shaughnessy has been a Professor at INRS-Télécommunications (University of Quebec) in Montreal, Canada, since 1977. For this same period, he has been an Adjunct Professor in the Department of Electrical Engineering, McGill University. Dr. O'Shaughnessy has worked as a Teacher and Researcher in the speech communication field for 30 years. His interests include automatic speech synthesis, analysis, coding and recognition. His research team is currently working to improve various aspects of automatic voice dialogues in English and French. He received his education from the Massachusetts Institute of Technology, Cambridge, MA (B.S. and M.S. degrees in 1972; Ph.D. degree in 1976). He is a Fellow of the Acoustical Society of America (1992) and an IEEE Senior Member (1989). From 1995 to 1999, he served as an Associate Editor for the IEEE Transactions on Speech and Audio Processing, and has been an Associate Editor for the Journal of the Acoustical Society of America since 1998. Dr. O'Shaughnessy has been selected as the General Chair of the 2004 International Conference on Acoustics, Speech and Signal Processing (ICASSP) in Montreal, Canada. He is the author of the textbook *Speech Communications: Human and Machine* (IEEE press, 2000).