5

Evolutionary Speech Recognition

Anne Spalanzani
LIG Laboratory - INRIA Rhône Alpes
Grenoble, France

1. Introduction

Automatic speech recognition systems are becoming ever more common and are increasingly deployed in more variable acoustic conditions, by very different speakers. After a short adaptation, they are most of all robust to the change of speaker, to low background noise, etc. Nevertheless, using these systems in more difficult acoustic conditions (high background noise, fast changing acoustic conditions ...) still need an adapted microphone, a long time relearning, background noise computing, which make these systems fastidious to use.

The difference between learning and testing conditions is the major reason of the drop in the systems' performances. Therefore, a system that obtains good performances in a laboratory is not automatically performant in real conditions (in an office, a car, using a telephone, ...). The two main causes are the extreme acoustic variability of the conditions of use and the non exhaustive corpus faced with these multiple conditions of use. The speech signal production, the signal propagation in the acoustic environment and the way the listener perceives and interprets this signal constitute multiple sources of variability which limit the usability of these systems.

In order to ensure the adaptability of these systems to different sources of variability and in order to improve their robustness, a lot of research has been developped. Technics can operate at different levels. For example, an adaptation of the noisy signal can be done at the acoustic level by applying filters, semantic and context knowledge can be used also as well as prosodic or multimodal informations (gestures accompanying the words, lips movements, etc.)

This article presents an acoustical approache which concentrates on the adaptation of the system itself so that it recognizes noisy speech signals. It is organized as follows: section 2 presents the different sources of speech signal variability. Section 3 presents classical technics to overcome the problem of variability and a discussion on classical technics and the importance of exploring the evolutionary technics. In section 4, evolutionary algorithms and methods to combine them with neural networks are described. The application of these methods onto automatic speech recognition systems and results obtained are presented in section 5. Discussion on results and possible future directions are developped in section 6.

Source: Robust Speech Recognition and Understanding, Book edited by: Michael Grimm and Kristian Kroschel,
ISBN 987-3-90213-08-0, pp.460, I-Tech, Vienna, Austria, June 2007

2. Source of speech signal variability

Speech is a dynamical acoustic signal with many sources of variation. Sources of degradation of the speech signal can be classified in 3 main categories (cf. Figure 1.) which are speaker variability, channel distortions and to room acoustic (Junqua, 2000).

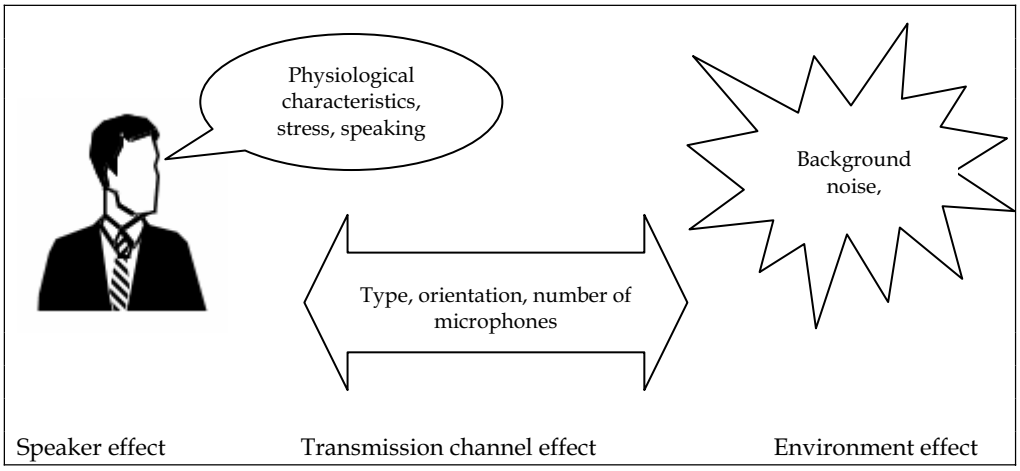


Figure 1. Schematic representation of the different sources of variability which can degrade the speech recognition system performances

The production of speech depends on many articulators that induce a lot of variability in the same linguistic message and for a unique speaker. These variations, known as speaking style, are due to factors like environment or social context. A speaker can change the quality of his voice, his speaking rate, his articulation according to some environmental factors. Stress conditions due to background noise increase the vocal effort.

The same message pronounced by two different speakers generate big variations. This is due most of all to physiological differences such as the vocal track length, male-female differences or child-adult differences.

The speech signal can be affected by the kind of microphone (or telephone) used, the number of microphones, the distance between the microphone and the speaker.

The characteristics of the environment such as the size of room, the ability of the materials used on the walls to absorb frequencies or modify the signal's properties, background noise, affect the signal quality. The characteristics of the spectrum resulting from the mix of noise and speech signal differ from the ones of a clean speech signal and the mix can be both additive (ambiant noise) or convolutive (room reverberation, microphone).

Robustness in speech recognition refers to the need to maintain good recognition accuracy even when the quality of the input speech is degraded, or when the acoustical, articulatory, or phonetic characteristics of speech in the training and testing environments differ.

The next section presents the main solutions proposed in order to deal with all these sources of degradation.

3. Classical proposed solutions in speech recognition robustness

Limiting the drop in performance due to acoustic environment changes remains a major challenge for speech recognition systems.
To overcome this problem of mismatch between a training environment and an unknown testing environment, many speech researchers have proposed different strategies which we classify in four categories:

- Speech signal processing approaches
- Model adaptation approaches
- Integration of multiple sources of information approaches
- Hybrid approaches

The first group focuses on a richer pre-processing of the input speech signal before presenting it to the model; e.g. by trying to suppress environment noises or by modifying acoustic parameters before the classification level (Bateman et al. 1992) (Mansour and Juang 1988).
In the second group, the adaptation to a new environment is achieved by changing the parameters of a speech recognition model (Das et al. 1994) (Gong 1995). For example, for a hidden markov model, this implies the modification of the number of states, of gaussians. Equivalently for a neural network, the number of units and the number of layers would be adjusted, as well as the weights' values.
The third group is increasingly addressed and corresponds to the combination of at least two sources of information to improve the robustness (Yuhas et al. 1989) (McGurk and MacDonald 1976). For example, the speaker's lips movements and their corresponding acoustic signals are processed and integrated to improve the robustness of the speech recognition systems.
The fourth group is based on the idea that hybridization should improve the robustness of systems by compensating the drawbacks of a given method with the advantages of the other (Junqua and Haton 1996). For example many contributions have focused on the combination of hidden markov models and neural networks to take advantage of the power of discrimination of neural networks while preserving the time alignment feature of hidden markov models.

3.1. Speech signal processing

3.1.1. Improving the signal-to-noise ratio

Speech enhancement techniques aim at recovering either the waveform or the parameter vectors of the clean speech embedded in noise (Gong 95). Among all the technics developped, methods working on noise filtering, spectal mapping and microphones arrays have been developped.
Different kinds of filters have been elaborated to reduce the noise. Classical filters such as Kalman or Wiener filters have been used for speech enhancement (Koo et al. 89). (Cardoso, 1989) (Jutten et al., 1991) have developped blind separation of noise and speech. These filters have shown a great capacity to reduce the noise in the spectrum but their drawback is that these methods alter also the speech signal. Bayesian methods have been developped to obtain an estimate of clean speech given noisy speech (Ephraim and Malah 1983) (Lim and Oppenheim, 1979) and the use of HMM to enhance noisy speech have been proposed in (Ephraim 92) (Seymour and Niranjan 94).

Spectral substraction is another technique which reduces the effect of added noise in speech. This method assumes that the noise and speech are uncorrelated and additive in the time domain (Gouvea and Stern, 1997) (Huerta and Stern, 1997).
The signal restoration via a mapping transformation exploits the direct correspondance between noisy and clean environments (Juang and Rabiner, 1987) (Barbier and Cholet 1991) (Gong and Haton 1994) (Guan et al. 1997) (Sagayama 1999). They propose to reduce the distance between a noisy signal and its corresponding clean signal. Unfortunalty, it requires knowing the clean speech which is not avalaible in most practical applications.
Further improvements in recognition accuracy can be obtained at lower signal-to-noise ratios by the use of multiple microphones (Silverman et al., 1997) (Seltzer, 2003).

3.1.2. Extracting robust features

Contrary to speech recognition systems, human beings are able to disregard noises to concentrate on one signal in particular. He is able to perform robust phonetical analysis which not depends on the speaker, transmissions channels or environment noises. Even if high levels are implied in the complex process of speech recognition, it seems that the auditory model has a major role.
Many ear models inspired by the human ear have been developped (Hermansky, 1990) (Allen, 1979) (Bourlard and Dupont, 1997).

3.2. Model adaptation

The noisy integration can be done by compensation that is by adapting the system to new noisy data (by restructurating the system structure and learning). Many techniques have been studied to proceed speaker adaptation (Schwartz and Kubala92) (Ström 94), channel adaptation such as telephone adaptation (Mokbel et al 93) or environment adaptation (Chiang 97). The major problem of these methods is the choice of the corpus. Very few methods propose to adapt the system on-line without knowledge on the new data to which the system should be adapted to (Kemp and Waibel 99). Practically, it is impossible to create an exhaustive corpus taking into account all the possible condition of use with all the noise characteristics. One other major problem is the systems' ability to learn correctly a huge amount of data.

3.3. Discussion

Among all these methods, very little are used in real automatic speech recognition systems. Actually, most of these methods are still at the stage of experimental research and do not provide enough convincing results to be integrated. The most common method is certainly the increase of the size of the database. As the computing power and the size of memory increase, it becomes possible to provide a good quantity of information for the training phase in order to have a powerful system in many conditions of use. However, in spite of the increase of their size, training databases are only small samples of the whole possible signal variabilities.
It is not possible to forecast all the testing conditions and it is necessary to explore new ways of search for a better comprehension of the problems involved in speech recognition (Bourlard, 1996).
We are conscious of the fragility of speech recognition. The communication between a speaker and his interlocutor is a complex mechanism. It implies acoustic signal exchanges in

a certain context (the acoustic environment) but it implies also exchanges of different kinds of signs with various levels of signification. 'Recognize' signifies for human beings 'identify' learned elements but also 'generalize' his knowledge of new elements. That is also 'recognize' missing or redondant features, 'adapt' to the environment, to the speaker or the situation. In hard conditions, the human listener takes into account the constraints offered by the linguistic code to reduce ambiguities, but he uses also information coming from the situation, redondancy, repetition. Generally, the speaker knows where he is and knows the environment's properties and limitations. He is able to take into account the semantic, pragmatic and lexical contexts (Caelen et al., 1996) and he is able to adapt his perception to the acoustic context.

Nature has created extremely complex systems with, sometimes, very simple principles. We can imagine that algorithms based on living being abilities (such as adaptation and evolution) could be able in the future to deal with all the parameters evolving during a dialogue, and reach human capacities.

The objective of the work presented in this article is to investigate innovative strategies using evolutionary algorithms inspired by Darwin. These algorithms evolve population of individuals in an environment supporting the survival and the reproduction of best suited individuals. The efficiency of this kind of algorithms is well known for the optimization of complex functions and for the resolution of the multi criteria problems. We aim at answering whether these algorithms can constitute a new approach for the adaptation of speech recognition systems to changing acoustic environments.

In order to study the possibility of incorporating evolutionary algorithms in the field of automatic speech recognition systems' robustness, we propose to remain at the acoustic level of the speech processing and more particularly at the level of the automatic speech recognition systems' adaptation with no assumption about the type of noise.

In this context, the robustness of the automatic speech recognition systems can be approached by two ways: dealing with structure or dealing with stimuli.

The first approach consists in adapting corrupted testing data so that they become close to training data (cf Figure 2.).

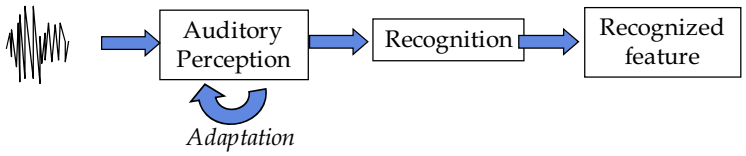


Figure 2. adaptation of the auditory perception

The speech recognition system is then able to provide good recognition rate. In this study, the system does not evolve anymore, only data do. The approach proposed first in (Spalanzani 99) processed a signal transformation via a mapping operator using a principal components analysis and evolutionary algorithms. This transformation attempted to achieve a self-adaptation of speech recognition systems under adverse conditions.

The second approach consists in adapting the speech recognition system itself (cf. Figure 3.).

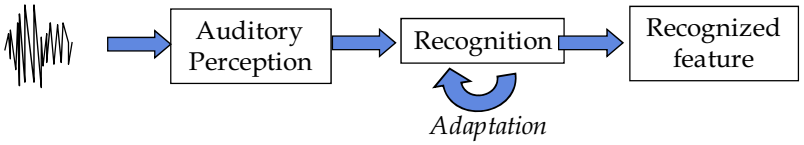


Figure 3. adaptation of the recognition system

Through the combination of evolutionary algorithms and backpropagation, a kind of relearning is operated to adapt the system to an environment different from the one in which it has been trained using the capacities of the system to adapt to changes of acoustic conditions using a local approach (by backpropagation of the gradient) and a global one (by evolution) in order to find an optimal system.
This article focuses on this second approach. Next section explains why and how combining neural networks and evolutionary algorithms.

4. Combining neural networks and evolutionary algorithms

In nature, living organisms are able to learn to adapt to the environment in order to survive and reproduce. Evolutionary algorithms are directly drawn from these principals of nature. Their principle consists of maintaining and manipulating a population of individuals (potential solutions) and implementing a "survival of the fittest" strategy (best solutions).
Neural networks are also a simplified way of simulating the ability of living organisms to adapt to their environment by learning.
It is appealing to consider hybrids of neural network learning algorithms with evolutionary search procedures, simply because Nature has so successfully done so.
The evolutionary algorithms proceed by globally sampling over the space of alternative solutions, while backpropagation proceeds by locally searching the immediate neighborhood of a current solution. This suggests that using the evolutionary algorithms to provide good initial weights sub-spaces from which backpropagation then continues to search will be effective (Belew, 1991).

4.1. Evolutionary algorithms

Evolutionary algorithms are stochastic search methods that mimic the metaphor of natural biological evolution. They operate on a population of potential solutions applying the principle of survival of the fittest to produce better and better approximations to a solution. At each generation, a new set of approximations is created by the process of selecting individuals according to their level of fitness in the problem domain and breeding them together using operators borrowed from natural genetics.
This process leads to the evolution of populations of individuals that are better suited to their environment than the individuals that they were created from, just as in natural adaptation.

4.1.1. Evolution

At the beginning of the computation a number of individuals (the population) are randomly initialized. The objective function (fitness function) is then evaluated for these individuals. The initial generation is produced.

If the optimization criteria are not met the creation of a new generation starts. Individuals are selected according to their fitness for the production of offspring. Parents are recombined to produce offspring. All offspring will be mutated with a certain probability. The fitness of the offspring is then computed. The offspring are inserted into the population replacing the parents, producing a new generation. This cycle is performed until the optimization criteria are reached.
The algorithm used to evolve to population is the following:

```
t = 0                              // time initialization
Init (P(t))                        // initial population creation
Eval (P(t))                        // population evaluation
While criteria non reached         // number of generations or good solution obtained
        P'(t) = Selection(P(t))    // mates selection
        P''(t) = Evolution (P'(t)) // mates' recombination and mutation
        Evaluation(P''(t))         // evaluation of the new population
        P(t+1) = P''(t)            // current population update
        t = t + 1                  // new generation
End While
```

Figure 4. evolutionary algorithm

4.1.2. Recombination

Recombination produces new individuals in combining the information contained in the parents. It consists in selecting a locus point and permutation the two right parts of the mates' genotypes.

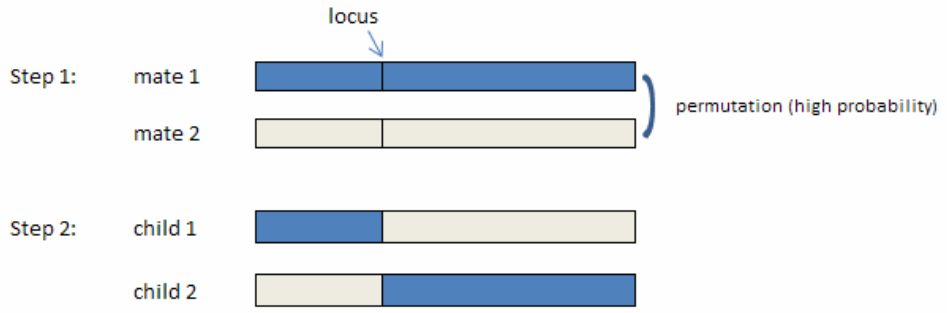


Figure 5. crossover principle

4.1.3. Mutation

After recombination, every offspring undergoes mutation. Offspring variables are mutated by small perturbations (size of the mutation step), with low probability. The representation of the genes determines the used algorithm. If the genes are binary, a bit-flip mutation is used. If the genes are real values, many possible mutations can be operated. Figure 6 shows an example of real value mutation.

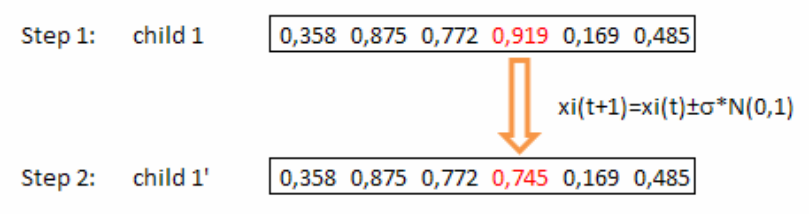


Figure 6: example of mutation

Next section will explain how to evaluate the individuals based on neural networks.

4.2. Neural networks

4.2.1. Backpropagation

Neural networks are collections of units (neurons) connected by weighted links (connection weights) used to transmit signals.

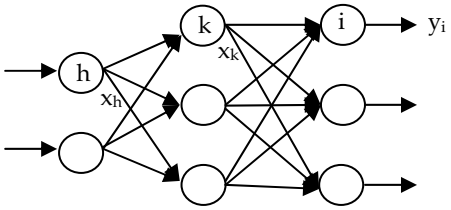


Figure 7: Extract of a multilayer neural network

The principle is based on the minimization of the quadratic error E which is function of the n desired outputs yd_i and the n outputs y_i given effectively by the network:

$$E = \sum_{i=1}^{n} (y_i - yd_i)^2 \tag{1}$$

Minimizing this error consists in modifiying the connection weights as follows:

$$\Delta W_{kh}^{(j)} = -a \cdot \delta_k^{(j)} \cdot y_h^{(j-1)} \tag{2}$$

With a the gain adaptation, $\delta_k^{(j)}$ the error of the neuron k of the j^{th} layer and $y_h^{(j-1)}$ the output of the neuron h of the $(j\text{-}1)^{th}$ layer.
For the neuron of the last layer:

$$\delta_k^{(j)} = y_h^{(j)} - yd_k \tag{3}$$

For the neuron of the hidden layers:

$$\delta_k^{(j)} = \left[\sum_{i \in layer(j+1)} \delta_i^{(j+1)} \cdot w_{ik}^{(j+1)} \right] \cdot \sigma'(p_k^{(j)}) \tag{4}$$

with : $p_k^{(j)} = \sum_i w_{ki} \cdot x_i$ where x_i corresponds to the output of neuron *i* and:

$$\sigma(p_k^{(j)}) = \left[\frac{1}{1 + \exp(-p_k^{(j)})} \right] \tag{5}$$

and σ' its' derivative.

4.2.2. Learning

The learning algorithm used in this work is adapted from (Schalkwyk et Fanty, 1996) which permits a fast convergence toward a good solution. The learning phase stops when the error does not decrease anymore, that is when it is not able to learn anymore. At the end of this learning phase, the algorithm gives a recognition rate which corresponds to the number of recognized features divided by the total number of features to recognize and a number of learning iterations.

4.2.3. Recognition

The recognition phase is used to evaluate the neural network based individuals. Data used for recognition are different from the one used for learning. The recognition rate given by the individuals are used as fitness function.

4.3. Hybridization

Three major approaches combining neural networks and evolutionary algorithms are presented in the literature: finding the optimal weights of the neural network (using evolution only or combined with backpropagation), finding the optimal topology of the neural network and finding optimal learning parameters (Yao, 1995) (Whitley, 1995) (Hancock 1992).

4.3.1. The topology of the neural network

The design of the network's topology is not easy. Finding the optimal number of layers as well as finding the number of neurons on each layer is usually done by empiric methods (more than based on theoric fundations).

4.3.2. The optimal weights of the neural network

The initialization of the weights before learning is crucial. It influences the convergence speed and the quality of the obtained network.

- Learning time

The learning convergence time depends on the initial and final weight space. In fact, the more initial weights are close to their final value, the faster is the convergence (Thimm, 1994) (Chan and Nutter, 1991). For example (Lee et al. 93) have shown theorically that the probability of premature saturation at the beginning epoch of learning procedure in the backpropagation algorithm derived in terms of the maximum value of initial weights, the number of nodes in each layer and the maximum slope of the sigmoidal activation function.

- Quality of the resulting network

Learning by backpropagation can be seen as a function optimization where the weights are its parameters. Its convergence toward a local minimum can be global also. If it is not the case, we can consider that the learning has not been done correctly.

4.4. Neural network encoding

The representation of a neural network in a genotype has been studied deeply in the literature (Miller et al., 1989) (Gruau and Whitley, 1993) (Mandischer, 1993). Weights manipulated by the backpropagation algorithm are real values and can be encoded by different ways as described in (Belew, 1991). In order to be efficient with the crossover operator, weights having a high interaction should be close.

Let's consider a 3 layer-network with *n* inputs, *m+1-n* hidden and *p+1-m* outputs, *wij* the connection of the neuron *i* toward the neuron *j*. Figure 8 shows the representation of this network:

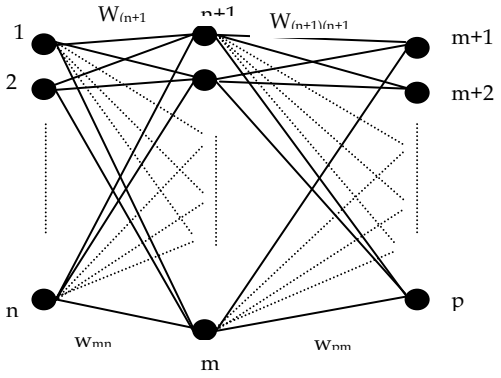


Figure 8: architecture of a neural network

The genotypic representation of this architecture is the following:

(w(n+1)1, w(n+1)2, …, w(n+1)n, w(n+2)1, w(n+2)2, …, w(n+2)n, …, w(m+1)m, …, wpm)

Figure 9. genotypic representation of the neural network architecture

4.5. Changing environment adaptation

The problem of adapting populations to changing environment is an interesting problem which gave place to a certain number of works. For example, (Nolfi and Parisi, 1997) investigated a robot adaptation in a changing environment controlled by a population of neural networks. (Cobb and Grefenstette, 1993) studied the evolution of a population tracking the optimum of complex functions (such as combination of sinusoids and gaussians) and the capacity of genetic algorithms to adapt, for example, to the translation of such functions.

Keeping diversity in the population seems to be the key for a good adaptation to environments changing quickly. (Cobb and Grefenstette, 1993) proposed a comparison of the performances of various strategies (high mutation rate (10%), triggered hypermutation and random immigrants). The high mutation rate generates a significant diversity, the triggered hypermutation varies the mutation rate according to the way the environment changes, its rate is weak when the changes are weak, high during abrupt changes, the random immigrants introduces randomness into a percentage of its population which generates diversity. It results from this work that each of these methods has advantages and disadvantages depending on the way the environment changes.
Methods based on the thermodynamic principles (Mori et al. 1996) can be found in the literature also. In addition, evolution strategy seems to be well fitted to adapt to the changes of environment (Bäck, 1996). Indeed, the integration of evolution's parameters in the genotype enables the adaptation of the mutation rates when it is necessary.

4.6. Lamarkism versus Darwinism

A certain number of works studied the influence of the heritage on the evolution of the populations (Turney et al., 1996). At the beginning of the century, two schools were confronted: The Darwinism is based on the idea that only the predisposition of the individuals to learn is transmitted to the children. Knowledge acquired by the parents is not transmitted then. Lamarckianism proposes that knowledge obtained by parents is directly transmitted to children. Hence, those parents' weights resulting from the training phase are transmitted to the following generation.
In the context of this debate, Baldwin proposed a kind of intermediary. He suggested the existence of a strong interaction between learning and evolution. The principal objective for the individuals is to integrate, by means of a neural network training, the information provided by the environment. The fittest individuals are selected for reproduction. So they transmit to their descendant their capacity to learn. If it is considered that an individual is all the more ready to integrate knowledge that the innate configuration of its neural network' weights is closed to that after training, we can consider then that knowledge is assimilated in its genes. Thus, the training time decreases. Once the individuals are able to acquire these concepts correctly, we can consider that those are comparable in the genotype.
(Nolfi et al., 1994) and (Nolfi and Spalanzani, 2000) showed that learning guides evolution (individuals having the best learning performances reproduce more often than the others), but also that evolution guides learning (the selected individuals have greater capacities to learn and these capacities are improved during generations). Thus, instinctive knowledge of the individuals is transmitted and improved during the evolution whereas, and it is the difference with the theory of Lamarck, the genotype is not affected directly by the training.
(Whitley et al., 1994) affirmed that under all the test conditions they explored, the Lamarckian evolution is much faster than that of Darwin and results are often better. Concerning the problem in which we are interested in, the adaptation to the changes of environment, (Sasaki and Tokoro, 1997) affirmed that the Darwinism is more adapted than Lamarckianism, whereas for a static environment, the opposite is noted. (Mayley 1996) proposed to penalise individuals having a long training phase. He affirmed also that knowledge assimilation can be done only if there is a neighbourhood correlation, i.e. a correlation between the distance from two genotypes and that of their associated phenotype.

5. Experimental results

5.1. Experimental plateform

Our simulations take place in EVERA (Environnement d'Etude de la Robustesse des Apprentis), our speech Artificial Life simulator which has been described in (Kabré and Spalanzani 1997). The main purpose is to provide a test-bed for the evaluation and improvement of the robustness of speech recognition systems. Two models are proposed in EVERA, a model of environment and a model of organisms (speech recognition systems) which evolve in it. In our study, the environment is a virtual acoustic environment. The virtual environment allows the simulation of the transfer function between acoustic sources (speaker and noise) and a microphone position. Thanks to Allen model of sound propagation in small rooms (Allen and Berkley 1979), by varying the reflection coefficient of walls, floor and ceiling, it is possible to control the reverberation time. This latter measures the time needed for the sound emitted in an acoustic environment to extinct. The difficulty of recognition increases with the reverberation time. The convolution between a speech signal taken from any database and the acoustic environment impulse response gives the speech signal for training the neural based speech recognition systems.

A population of speech recognition systems is built in an initial virtual acoustic environment. The creation of the initial population resorts to make a desired number of copies of a pre-trained neural network while making a random change of its weights. The environment is changed by either taking a new acoustic environment, a different speech database or a different noise (we considered different noises such as door closing, alarm, radio). The change is such that the systems are no longer well-suited to this environment. At this moment, after a period of training, an adaptation cycle starts thanks to genetic operators.

1. **Init** a population of automatic speech recognition systems.
2. **If** duration of simulation not elapsed **change** the acoustic environment **else goto** 6.
3. **Train** automatic speech recognition systems.
4. **Evaluate, Select** and **Reproduce** speech recognition systems.
5. **If** duration of adaptation elapsed **then goto** 2 **else goto** 3.
6. end.

Figure 10. Algorithm for evolving speech recognition systems so that they adapt to new virtual acoustic environments.

The acoustic environments are represented by a set of words to which noise and reverberation were added. Noises have been chosen thanks to their characteristics: PO for door closing (impulsive noise), AL for alarm and RE for alarm clock (narrow-band noises at different frequencies), FE for fire (white noise) and RA for change of radio frequency (non stationary noise).

An environment is defined by a triplet (type of noise, reverberation time, signal to noise ratio). The intelligibility of the signal is inversely proportional to the reverberation time and proportional to the signal to noise ratio. This is why a signal with a strong reverberation and a weak signal to noise ratio (for example (FE 0,7 -6)) is more difficult to recognize than a signal like (RA 0,4 20).

5.2. Comparison between Lamarckian and Darwinian evolution

First experiments determine the most effective method for our problem of adaptation to the environment. Since the opinions are divided concerning the method to use, we propose to test the performances of populations in term of quality of the individuals and in term of effectiveness. Within the framework of our experiments, this consists in studying the recognition rate of our population as well as the number of iterations necessary for each individual to optimise its training (i.e. the network converged). The objective of the individuals is to recognize a set of isolated words analysed by an acoustic analyser based on the model of the human ear (Hermansky, 1990). The vector resulting from the acoustic analysis represents the input of the networks having 7 input units, 6 hidden units and 10 output units. They are able to learn thanks to training algorithm based on backpropagation.

5.2.1 Quality of the results

On the general shape of the curves, we can also notice that results provided by the Lamarckian evolution are more stable and seem to drop less easily than those of the Darwinian population. The numerical results of the performances are presented in table 1. We can notice that the performances are quite equivalent in average. In average, the population evolving according to the method of Lamarck obtains 78% of recognition rate whereas that according to the method of Darwin obtains 76.6%. Concerning the best individual, less than 1% of improvement is noted since in average (for the 10 environments, that is to say 1000 generations), the best Darwinian individual obtains 80.1% whereas Lamarckian 80.9%.

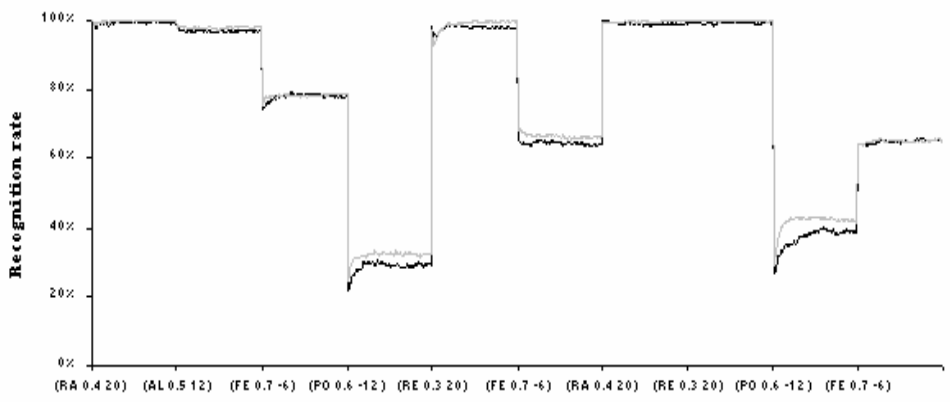


Figure 11. Recognition rates of the population of speech recognition systems evolving in a changing acoustic environment. 20 individuals evolve by genetic algorithms and neural training. Darwinian and Lamarckian heritage are compared on the average of the population.

Recognition rate	Worst	Best	Average
Lamarckian evolution	73.8 %	80.9 %	78 %
Darwinian evolution	70.5 %	80.1 %	76.6 %

Table 1. Comparison of the recognition rates between Lamarckian and Darwinian evolution.

5.2.2. Training efficiency

Concerning now the efficiency of the populations in the training phase, figures 12 shows the number of iterations necessary for a good convergence of the individuals' networks. Although, at each change of environment, the number of iterations increases in a more significant way during the evolution of Lamarck, the decrease of this number is more significant and, in average, the number of iterations is weaker. It is interesting to note that this number decreases throughout the Darwinian evolution, which can means that there is knowledge assimilation, and this without the use of penalty as proposed (Mayley 1996).
In both kinds of evolution, the number of iterations decreases during generations. Once more, we can notice that the evolution of Lamarck is more effective than that of Darwin. As indicates it table 2, Darwinian individual needs in average 91 iterations to learn correctly, whereas a Lamarckian individual need only 68.2 iterations. We can notice the differences between the best individuals (54.8 against 33.1) and worse (144 against 113.6).

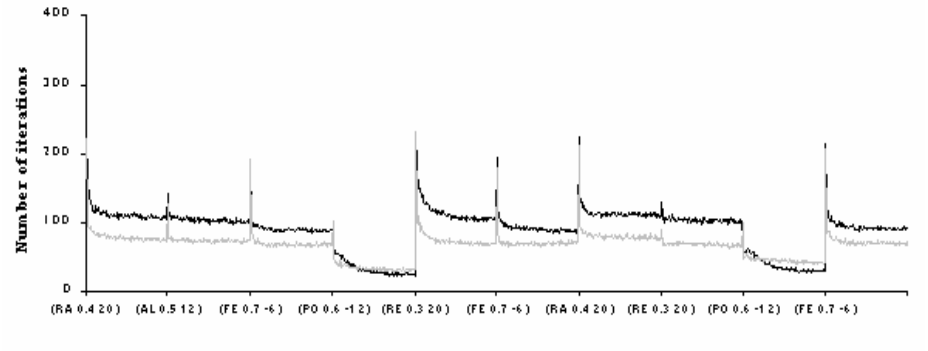


Figure 12. Learning times of ASRSs population evolving in a changing acoustic environment. 20 individuals evolve by genetic algorithms and neural training. Darwinian and Lamarckian heritage are compared on the average of the population.

Number of iterations	Worst	Best	Average
Lamarckian evolution	113.6	33.1	68.2
Darwinian evolution	144	54.8	91

Table 2. Comparison of the learning times.

5.3. Evolution during time

This experiment consists in presenting speech signals produced in several noisy environments. A sequence of several noisy environments is presented 5 times. This permits to test the systems' performances in identical acoustic conditions at different times.
Performances of 2 populations are tested. One population is made of neural networks evolving with the backpropagation algorithm only, the other is made of neural networks evolving with both backpropagation and lamarkian evolutionary algorithm. Results presented in figure 13 are the average of 9 simulations.

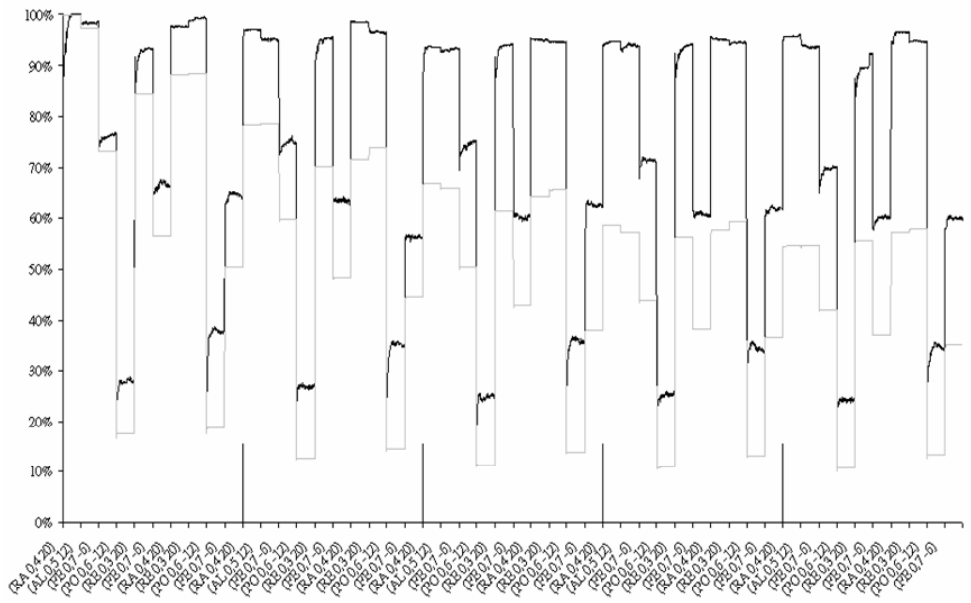


Figure 13. Evolution of the recognition rate during time while a sequence of 10 acoustic environments is presented 5 times. In gray, recognition rate of the "neural network population", in black, the recognition rate of the "evolutionary neural network population"

One can notice the regularity of the performances of the evoluted population and the constant decrease of the population using back-propagation only. Moreover, the recognition rates of the evoluted population are much higher than the ones given by the non evoluted population, whatever the acoustic environment.

6. Conclusion and perspectives

In this article, we have adapted neural networks based speech recognition systems using evolutionary algorithms in order to keep them adapted to new acoustic environments. We have compared two methods of heritage suggested by Darwin and Lamarck for the evolution of a population in changing acoustic environments. In term of speech recognition rate, both methods provide similar results. In term of learning speed, The Lamarckian

evolution seems to be more interesting in the context of changing acoustic environments. Results obtained in a long sequence of environments show that evolution provides stable results (i.e. for two identical acoustic conditions but at different times, results are similar) but systems without evolution do not.

The generic concepts presented in this article must not be limited to neural based methods. In this article, neural networks have been implemented because they are easy to controle and their convergence is quite fast. Moreover, the hybridization of neural networks and evolutionary methods have been studied deeply. But hidden markov models could be adapted also (Asselin de Beauville et al. 96) instead of neural network systems.

Moreover, (Lauri et al. 2003) have shown the efficency to combine evolutionary algorithms and Eigenvoices to adapt the system to new speakers. (Selouani and O'Shaughnessy 2003) combined hidden markov models and Karhonen-Loève transform to improve telephone speech recognition.

This work is at the frontiere between two very different fields which are automatic speech recognition and evolutionary algorithms. Work carried out comes from the idea that if the systems of recognition were able to self-modified in time, in order to adapt to the changes of acoustic environment, they could be much more robust. The idea was to take as a starting point the capacities of the alive beings to adapt to their environment to be the most powerful possible in order to survive.

Within the framework of speech recognition, we considered the automatic speech recognition systems, or filters, like individuals having to adapt to their acoustic environment changing. It appeared interesting to imagine a system able to adapt to the acoustic changes of conditions (characteristic of the speaker or the room for example) in order to remain performant whatever its conditions of use. The alive beings are able to adapt their manner of perceiving several signals while concentrating on the signal in which they are interested in. They are also able to update their knowledge of the environment when it is necessary. Considering that speech signal recognition for alive beings can be summarized in two main phases, namely perception of the signals and the attribution of entities to these signals, we have suggested to adapt one or the other independently by evolutionary technics.

We keep in mind that there is a strong interaction between these two processes but their adaptation by evolutionary algorithms in a parallel way seems, for the moment, impossible to control. Indeed, how to adapt our recognition system on data which adapt simultaneously to this one?

7. References

Allen J., B. & Berkley D. A. (1979). Image Method for efficiently simulating small-room acoustics. *JASA* 65(4):943-950.

Asselin de Beauville J-P, Slimane M., Venturini G., Laporte J-L and Narbey M. (1996). Two hybrid gradient and genetic search algorithms for learning hidden Markov models. *13th International Conference on Machine Learning*, Bari, Italy, pp.1-8.

Bäck T. (1996) *Evolutionary Algorithms in Theory and Practice*. Oxford University Press, New York.

Barbier L. et Chollet G. (1991). Robust Speech Parameters Extraction for Word Recognition in Noise Using Neural Networks. *ICASSP'91,* pp.145-148.

Bateman, D. C., Bye, D. K. & Hunt M. J. (1992). Spectral normalization and other spectral technics for speech recognition in noise. *Proceedings of the IEEE International conference. on Acoustic Speech Signal Processing*, (1)241-244. San Francisco.

Belew, R. K., McInerney, J. & Schraudolph, N. (1991). Evolving Networks : Using the Genetic Algorithm with Connectionist Learning. *In Proc. Second Artificial Life Conference*, pages 511-547, New York, Addison-Wesley.

Bourlard H. (1996). Reconnaissance Automatique de la Parole : Modélisation ou Description ? *XXIèmes Journées d'Etude sur la Parole*, Avignon, France, pp.263-272.

Bourlard H. and Dupont S. (1997). Subband-based Speech Recognition. *ICCASP'97*, pp. 1251-1254.

Caelen J., Kabré H. and Delemar O. (1996), Reconnaissance de la Parole :vers l'Utilisabilité. *XXIèmes Journées d'Etude sur la Parole*, Avignon, France, pp.325-329.

Cardoso J.F. (1989). Source separation using higher order moments. *ICASSP'89*, Glasgow, Scotland, vol. 4, pp. 2109-2112.

Chen C.L. and Nutter R.S. (1991). Improving the Training Speed of Three-Layer FeedForward Neural Nets by Optimal Estimation of the Initial Weights. *International Joint Conference on Neural Networks.*

Chiang T. (1997) Speech Recognition in Noise Using On-line HMM Adaptation. *Eurospeech'97, 5th European Conference on Speech Communication and Technology*, Rhodes, Grece.

Cobb H.G. and Grefenstette J.J. (1993). Genetic Algorithms for Tracking Changing Environments. *Fifth International Conference on Genetic Algorithms (ICGA 93),* 523-530, Morgan Kaufmann.

Das, S., Nadas, A., Nahamoo, D. & Picheny, M. (1994). Adaptation techniques for ambient noise and microphone compensation in the IBM Tangora speech recognition system. *In Proceedings of the IEEE International Conference On Acoustic Speech Signal Processing*. (1)21-23. Adelaide, Australia.

Ephraim Y. and Malah D. (1983). Speech enhancement using optimal non-linear spectral amplitude estimation. *ICASSP'83,* pp. 1118-1121.

Ephraim Y. (1992). Statistical-model-based speech enhancement systems. *Proceeding IEEE,* 80(10):1526-1555.

Gong Y. and Haton J-P. (1994). Stochastic Trajectory Modeling for Speech Recognition. *ICASSP'94*, pp. 57-60, Adelaide.

Gong, Y.: (1995) Speech recognition in noisy environments: A survey, *Journal of Speech Communication*, 16 : 261-291.

Gouvêa E.B. et Stren R.M. (1997) Speaker Normalization Through For-mant-Based Warping of the Frequency Scale. *Eurospeech'97, 5th European Conference on Speech Communication and Technology,* Rhodes, Greece.

Gruau F. et Whitley D. (1993). Adding Learning to the Cellular Development of Neural Networks : Evolution and the Baldwin Effect. *Evolutionary Computation* 1(3): 213-233.

Guan C-T., Leung S-H. et Lau W-H. (1997). A Space Transformation Approach for Robust Speech Recognition in Noisy Environments. *Eurospeech'97, 5th European Conference on Speech Communication and Technology*, Rhodes, Greece.

Hancock P.J.B. (1992). *Coding Strategies for Genetic Algorithms and NeuralNets*. Ph.D. Thesis, University of Stirling.

Hermansky, H. (1990). Perceptual Linear Predictive (PLP) Analysis of Speech, *Journal of Acoustic Society*, 87(4) 1738-1752.

Holland, H. (1975). *Adaptation in Natural and Artificial Systems*. The University of Michigan Press.

Juang B. et Rabiner L. (1987). Signal Restoration by Spectral Mapping. *ICASSP'87*, pp 2368-2371

Huerta J.M. et Stern R.M. (1997). Compensation for Environmental and speaker Variability by Normalization of Pole Locations. *Eurospeech'97, 5th European Conference on Speech Communication and Technology*, Rhodes, Greece.

Junqua J. C. (2000). *Robust Speech Recognition in Embedded Systems and PC Applications*, Kluwer Academic Publishers.

Junqua, J. C. & Haton, H. (1996). *Robustness in Automatic Speech Recognition*, Ed Kluwer Academic Publisher.

Jutten C., Hérault J., Comon P. et Sorouchyari E. (1991). Blind separation of sources. *Signal Processing*, vol. 24, pp. 1-29.

Kabré, H. & Spalanzani A. (1997). EVERA: A system for the Modeling and Simulation of Complex Systems. *In Proceedings of the First International Workshop on Frontiers in Evolutionary Algorithms, FEA'97*, 184-188. North Carolina.

Kabré, H.: (1996). On the Active Perception of Speech by Robots. *IEEE RJ/MFI (Multi-sensor Fusion and Integration for Intelligent Systems),* 775-785. Washington D.C.

Kemp T. et Waibel A. (1999) Unsupervised Training of a Speech Recognizer: Recent Experiments. *Proceedings of the Eurospeech'99*, pp. 2725-2728, Budapest.

Koo B., Gibson J. and Gray S. (1989) Filtering of colored noise for speech enhancement and coding. *ICASSP'89*, pp. 349-352. Glasgow.

Lauri F., Illina I., Fohr D. and Korkmazsky F. (2003) Using Genetic Algorithms for Rapid Speaker Adaptation, *Eurospeech'2003*, Genève, Switzerland.

Lee Y., Oh S.-H. et Kim M.W. (1993). An Analysis of Premature Saturation in BackPropagation Learning. *Neural Networks*, vol. 6, pp. 719-728.

Lim J.S. and Oppenheim A.V. (1979). All-pole modeling of degraded speech. *IEEE Transaction on Acoustics, Speech, and Signal Processing*, 26(3):197--210, 1979.

Mandischer M. (1993). Representation and Evolution in Neural Networks. *Artificial Neural Nets and Genetic Algorithms Proceedings of the International Conference at Innsbruck,* Austria, pages 643-649.

Mansour, D. & Juang, B. H. *(1988).* A family of distortion measures based upon projection operation for robust speech recognition. *IEEE International Acoustic Speech Signal Process*, 36-39. New York.

Mayley G. (1996). *Landscapes*, Learning Costs and Genetic Assimilation. Special *Issue of Evolutionary Computation on the Baldwin Effect*, vol. 4, n. 3.

McGurk, H., MacDonald, J. (1976). Hearing Voices and Seeing Eyes, *Nature*, 264:746-748.

Miller G.F., Todd M. et Hegde S.U. (1989). Designing Neural Networks using Genetic Algorithms. *Proceedings of the Third Conference on Genetic Algorithms*, San Mateo.

Mokbel, C., Monné, J. and Jouvet, D. (1993). On-line adaptation of a speech recognizer to variations in telephone line conditions. *EUROSPEECH'93*, 1247-1250.

Mori N., Kita H. et Nishikawa Y. (1996). Adaptation to a Changing Environment by Means of the Thermodynamical Genetic Algorithm. *4th Conference on Parallel Problem Solving from Nature*, Berlin, Allemagne.

Mühlenbein, H. & Schlierkamp-Voosen, D. (1995). Evolution as a Computational Process. *Lecture Notes in Computer Science*, 188-214, Springer, Berlin.

Nolfi S., Elman J.L. & Parisi D. (1994). Learning and evolution in neural networks. *Adaptive Behavior*, (3) 1:5-28.

Nolfi S. & Parisi D. (1997). Learning to adapt to changing environments in evolving neural networks. *Adaptive Behavior*, (5) 1:75-98,

Nolfi S. & Spalanzani A. (2000). Learning and evolution: On the effects of directional learning. Artificial Life. *Technical Report*, Institute of Psychology, Rome, Italy.

Sagayama S. (1999). Differential Approach to Acoustic Model Adaptation. *Workshop on robust Methods for Speech Recognition in Adverse Conditions*, Tampere, Finlande.

Sasaki T. and Tokoro M. (1997). Adaptation toward Changing Environments : Why Darwinian in Nature ? *Fourth European Conference on Artificial Life*.

Schalkwyk J. and Fanty M. (1996). The CSLU-C Toolkit for automatic speech recognitions. Technical Report n. CSLU-012-96.

Schwartz R. and Kubala F. (1992). Hidden Markov Models and Speaker adaptation. Speecj Recognition and Understanding. *Recent Advances*, *Trends and Applications.* Springer-Verlag, 1992.

Selouani S. and O'Shaughnessy D. (2003). On the Use of Evolutionary Algorithms to Improve the Robustness of Continuous Speech Recognition Systems in Adverse Conditions. *EURASIP Journal on Applied Signal Processing*, 2003:8, 814–823.

Seltzer M.L (2003). *Microphone Array Processing for Robust Speech Recognition*, Ph.D. Thesis, ECE Department, CMU.

Seymour C.W. and Niranjan M. (1994). An Hmm-Based Cepstral-Domain Speech Enhancement System, *In Proceedings ICSLP*, pages 1595—1598.

Silverman H.F., Patterson W.R., Flanagan J.L. et Rabinkin D. (1997). A Digital Processing System for Source Location and Sound Capture by Large Micro-phone Arrays. *ICASSP 97*, Volume 1, Page 251.

Spalanzani A. & Kabré H. (1998). Evolution, Learning and Speech Recognition in Changing Acoustic Environments. *5th Conference on Parallel Problem Solving from Nature*, Springer Verlag, pp. 663-671, Amsterdam, Netherland.

Spalanzani, A. (1999). *Algorithmes évolutionnaires pour l'étude de la robustesse des systèmes de reconnaissance automatique de la parole*, Ph.D. Thesis, Joseph Fourier University, Grenoble.

Spears, W.M., De Jong, K.A., Bäck, T., Fogel, D. and De Garis, H. (1993). An Overview of Evolutionary Computation. *In Proceedings of the European Conference on Machine Learning*, (667) 442-459.

Ström, N. (1994) Experiments with new algorithm for fast speaker adaptation. In *ICSLP*, pp. 459-462.

Thimm G. et Fiesler E. (1994). High Order and Multilayer Perceptron Initialization. *IDIAP technical report* 94-07.

Turney P., Whitley D. et Anderson R. (1996). Evolution, Learning, and Instinct : 100 Years of the Baldwin Effect. *Special Issue of Evolutionary Computation on the Baldwin Effect*, vol. 4, n. 3.

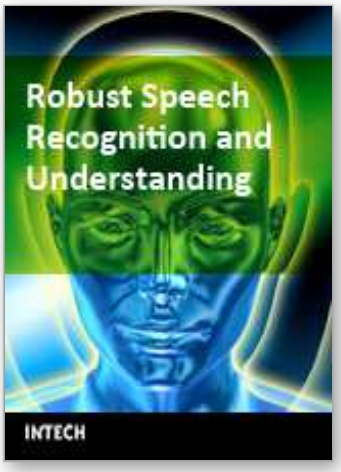
Wessels, L and Barnard, E. (1992). Avoiding False Local Minima by Proper Initialization of Connections. *IEEE Transactions on Neural Networks,* vol. 3, No 6.

Whitley D. (1995). Genetic Algorithms and Neural Networks. *Genetic Algorithms in Engineering and Computer Science.* Ed. J. Periaux et G. Winter.

Whitley D., Gordon S. et Mathias K. (1994). Lamarckian Evolution, the Baldwin Effect and Function Optimization. *Parallel Problem Solving from Nature* III. pp. 6-15. Springer-Verlag.

Yao X. (1995). Evolutionary Artificial Neural Networks. *Encyclopedia of Computer Science and Technology*. Ed.A. Kent et al., vol. 33, pp 137-170, Marcel Dekker Inc.

Yuhas, B.P., Goldstein, M.H. & Sejnowski, T.J. (1989). Interpretation of Acoustic and Visual Speech Signal using Neural Networks. *IEEE Common Magazine.*



Robust Speech Recognition and Understanding

Edited by Michael Grimm and Kristian Kroschel

ISBN 978-3-902613-08-0
Hard cover, 460 pages
Publisher I-Tech Education and Publishing
Published online 01, June, 2007
Published in print edition June, 2007

This book on Robust Speech Recognition and Understanding brings together many different aspects of the current research on automatic speech recognition and language understanding. The first four chapters address the task of voice activity detection which is considered an important issue for all speech recognition systems. The next chapters give several extensions to state-of-the-art HMM methods. Furthermore, a number of chapters particularly address the task of robust ASR under noisy conditions. Two chapters on the automatic recognition of a speaker's emotional state highlight the importance of natural speech understanding and interpretation in voice-driven systems. The last chapters of the book address the application of conversational systems on robots, as well as the autonomous acquisition of vocalization skills.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821