Journal of Computer Science 3 (9): 685-692, 2007
ISSN 1549-3636
© 2007 Science Publications

# Arabic Speech Pathology Therapy Computer Aided System

[1]Z.A. Benselama, [2]M. Guerti and [3]M.A. Bencherif
[1]Electronic Dept., Saâd Dahleb University, Blida, Algeria BP-270 Blida ALGERIA, &
NationalPolytechnic College Electronic Dept., Algiers 10, Avenue Hassen Badi B.P. 182 16200 ALGERIA
[2]National Polytechnic College Electronic Dept., Algiers 10, Avenue Hassen Badi B.P. 182 16200 ALGERIA
[3]ALJOUF University P.O. Box: 2014, Aljouf, Skaka, SAUDI ARABIA

**Abstract:** This article concerns a computer aided pathological speech therapy program, based on speech models such as the hidden Markov model and artificial intelligence networks, in order to help persons, suffering from language pathologies, follow a correction learning process, with different interactive feedbacks, aiming to evaluate the degree of evolution of the illness or the therapy. We dealt with the Arabic occlusive sigmatism as a prime approach, which is the inability to pronounce the[s] or [ʃ]. Results obtained are satisfying and the therapy program is prepared, for autonomous use by patients, for deep analysis and verifications.

**Keywords:** Word segmentation, phoneme recognition, HMM/GMM, ANN

## INTRODUCTION

Computer aided therapy software is a new concept. in the field of medicine, specially in the speech process, It is an orientation to the introduction of different algorithms in the illness correction and default detection. This approach is essentially used in hospitals in pre-surgery [1,2] or in the default correction, by repeated learning sessions.

While treating a speech pathology, therapists try to detect the different defaults by some non-invasive methods, in order to preserve the real functioning conditions. The speech therapist intervenes in the treatment, using his level of knowledge and its acquired accuracy to detect the defaults, and correct them, before and/or after the use of the surgery.

Speech recognition techniques are, well, suited to the implementation, of a computer aided therapy system used, in the speech correction or language learning [3]. These techniques are based on the mixture of different approaches, such as the Gaussian mixture models, (GMM)[4], artificial neural networks [5,6], hidden Markov models [7], or some hybrid techniques like ANN/HMM,[8] HMM/GMM[9]. In this work, we used a complementary technique to get a more robust recognition system, as an extension to our previous work based only on HMM/GMM [10]

We will first focus on the speech pathology, related to our problem, then the following part concerns the material and methods section, the third section concerns the modeling parameters. The fourth part concerns the automatic word segmentation procedure followed by a discussion about the results and a conclusion

## SPEECH PATHOLOGY

The speech pathology deals essentially with the detection of the default pronouncing areas, in order to measure the degree of the illness, for a future adequate treatment.

The different organs attained by this pathology are divided into two main parts: the first part deals with peripheral parts, like the hare nozzle, palatal division, velar insufficiency, lingual defaults etc…, which are consistent constraining mechanical movements in the production of the phonemes, or problems related to the incapability to articulate phonemes in the systematic and permanent ways. [11]. the second part of pathologies concerns the source which is related to the speech vocal cords, this direction, is not considered in this article.
Generally, due to the defaults of the phonetic system, different pathologies occur, such as, the lisping, the hissing, the rhotacism, the stammering, the gammacism, etc…

## MATERIAL & METHODS

This work is similar, in the problem definition, to the learning methods of the English language by the non-native English Taiwanese or Japanese people [3,12]. These studies are oriented toward the detection of the bad pronounced phonemes, this default is intrinsically

**Corresponding Author:** Z.A. Benselama, Electronic Dept., Saâd Dahleb University, Blida, Algeria BP-270 Blida ALGERIA native within this region of the world, due to the absence of theses phoneme in their languages.

Different programs are available in the speech pathology domain, but none treats the Arabic language neither the occlusive stigmatism [13].

We tried to follow the same procedure; in order to end up with a similar approach that relates to the occlusive sigmatism, which concerns, the airflow from one side of the mouth. Due to a bad placement of the tongue, the similarity is regarded as learning a phoneme in a new language context, or learning how to pronounce a faulty phoneme in a pathology context.

The illness or pathology, we are dealing with, concerns the lingual interposition in the production of the Arabic phonemes such as [s][س] and [ʃ] [ش]. This occlusive sigmatism concerns respectively the replacement of the sounds [ʃ][ش], [∅] [ج], [s] [س], [z] [ز] by [θ] [ث] and [d] [د] by [f] [ف] respectively.

As an example, taken from our own recorded database, due to the lack or absence of Arabic speech pathology databases, we have chosen, for this article, the sample word [ʃaχsija], where the occlusion pathology is intensively significant, this word has been pronounced by a patient representing the speech pathology; the transcribed heard sounds are shown in table 1.

Table 1: Phonetic transcription of some words containing the mispronounced phonemes [ʃ] and [s]

| Initial | Pronunciation | | Incorrect |
|---|---|---|---|
| word | Correct | Incorrect | word |
| شخصية<br>[C1] | [ʃ aχs i j a] | [θaχθija] | تخثــيـة |
| | | [θaχʃθija] | تخشثيــة |
| شمس<br>[C2] | [ʃ a m s ø] | [ʃ a m ʃ ø] | شمش |

The second word is given as an example, in order to illustrate, the confusion that may happen in other word pronunciations.

These phonemes present a huge misspelling, due to their close place of articulation.

Let us remark the changes at the phonemic level, which gives a nearby word; that is the correction target. In both situation, the heard sounds, give the impression of the right word, but with a phoneme 'shift', due to the tongue that is shifted from the post-alveolar to the alveolar position, ending sometimes in the dental position, pronouncing a faulty phoneme.

The therapy methodology, as shown in the flowchart of the figure 1, describes how are treated other similar therapies, as soon as the concerned correct and pathological databases are available.

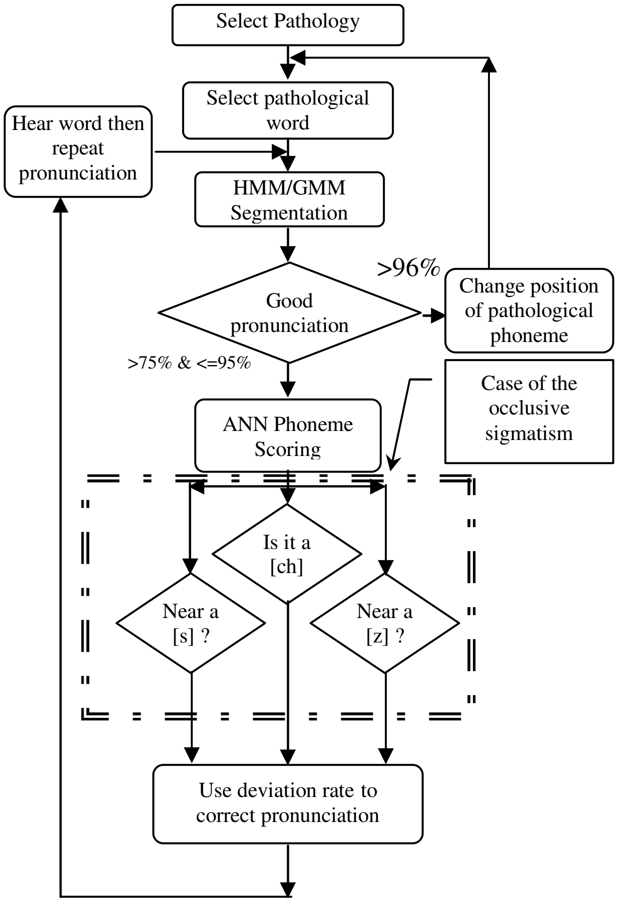

Fig.1: Flowchart of the therapy procedure

The choice of the parameters modeling the healthy or pathological speeches depends essentially on the desired recognition level.

In this work, we used two approaches, a global word segmentation approach based on a hidden Markov model (HMM), then a specific phonemic recognition level, in order to distinguish phonemes from their neighbors, based on a neural network rating process as shown in Figure.1.

## MODELLING PARAMETERS

**Choice of the cepstral coefficients:** In order to make a robust recognition, we worked within a medium noisy environment, since the dedicated tool, will be used in a hospital, or even at home, where the level of noise would be difficult to control or settle to its low level.

The Mel Frequency Cepstral coefficients called MFCC are a good compromise, they can sustain to noise and have a good intrinsic de-correlation factor, added to the fact that they model the hearing system in a filtering manner like does the cochlea.

These coefficients have been used intensively, in the recognition of the English digits in a noisy environment [14]. In the medical survey for the detection of different sounds [15]. In the comparison of the parameters in Arabic language recognition[7]. In the detection of the pathological speech[16]. In acoustic modeling techniques for embedded systems[17]. In the speaker recognition [18], and have shown good results in the identification of different complex phonetic features of the Arabic language[7], and essentially in the recognition and verification of the Japanese students while learning English [3].

The number of the MFCC coefficients is set to 13 then reduced to 12, considering that the first coefficient is the energy of the frame, which does not contribute to the discrimination process, in our situation. The other 12 coefficients represent the spectral envelope without high frequencies.

**Choice of the modeling parameters:** The Arabic language contains 34 phonemes; One Markov model may represent each phoneme [19].

The Markov representation known as the Hidden Markov Model, (HMM), allows synthesizing the information within a corpus via learning. The different occurrences represented as some probabilistic transitions within a graph model; each graph could be a phoneme, a word, depending on the required level of recognition.

The use of a huge set of data is very important, in order to have a complete selectivity in terms of multi-Gaussian distribution. The total set of data called '*O*' is approached by a Gaussian mixture model, for two main reasons, a compression side where the most important parameters would be the means and the variances of the Gaussians, added to that their degree of participation in the model. The second side concerns the natural distribution of data in the real world, which are best approximated by the Gaussian model.

The equation relating the data to the GMM is as follows:

$$p(O/\Theta)=\sum_{i=1}^{T}c_i.p_i(O/\theta_i) \quad (1)$$

And:

$$\Theta = \{c_1,c_2,...,c_T,\theta_1,\theta_2,...,\theta_T\} \quad (2)$$

with :

$$\theta_i = \{\mu_i,\sigma_i\} \quad (3)$$

Where :

$O$ : represents the acoustic observations

$\Theta$: represents the GMM, with $c1...c_T$ being respectively the degree of participation of the sub-models $\theta_1..\theta_T$, with means an variances respectively $\{\mu_i\ \sigma_i\}$, $i=1..T$.

$T$ : being the number of sub-models chosen.

We used initially the HMM/GMM approach for the automatic word segmentation, with different topologies represented by the number of states in the HMM model, and the number of Gaussians associated to each state, then an ANN network to distinguish between nearby phonemes.

The chosen topology of the HMM is a right to left word model, which stands for transition of the speech in one way, with 4,8 and 12 GMM associated to each state or phoneme.

From some previous studies [7,201], a model with 3 right to left states is also adequate for each phoneme, ending with a HMM model 3 times longer.

The different steps of construction of the GMM/HMM model are based upon the use of the Baum-Welch algorithm, variant of the Expectation Maximization, EM algorithm [20], which computes the expectation of an unknown random value against a known random variable. This method uses a recursive method, until a reasonable model adapts or fits well the data. This could be seen in the adjustment of the parameters of both the HMM and the GMM.

The global model is defined as follows:

$$\lambda = \{A,B,\pi\} \quad (4)$$

where:

$A$: represents the transition matrix containing the states $q_i$.

$B$: is the observation matrix containing the parameters of the GMM, and the observations are denoted by:

$$O = \{o_1,o_2,........,o_N\} \quad (5)$$

$\pi$: represents the initial probabilities of the HMM model.

In order to model the data, it is required to find the best path through the transition matrix maximizing:

$$P\ (O/\lambda) \quad (6)$$

The Viterbi approach is used, because instead of summing all the probability paths, only the maximal probability at each state is taken, this technique is based upon the fact that an optimal path is the sum of sub optimal paths.

At each step of the EM algorithm, we check if the new model brings amelioration in the adjustment of the data, that is verifying if the data fits the model at the $(n)^{th}$ step better than the $(n-1)^{th}$ step.

Initially we start by:

$$\tilde{\pi}_i = \gamma_i \qquad (7)$$

The equation (7) represents the relative frequency to be at the state $q_i$ initially.

The update of the parameters is done by completing the required data through the following process:

$$\tilde{a}_{ij} = \frac{\sum_{n=1}^{N-1} \xi_{ij}(n)}{\sum_{n=1}^{N} \gamma_i(n)} \qquad (8)$$

Where :

$\sum_{n=1}^{N-1} \xi_{ij}(n)$ : represents the number of transitions from the state $q_i$ to the state $q_j$ at iteration n.

$\sum_{n=1}^{N} \gamma_i(n)$ : represents the number of transitions outgoing from state $q_i$ at iteration n.

For the GMM, the different parameters to be estimated are the variances and the means as well as the participation factor of each Gaussian in the global model at the state $q_i$ denoted as follows:

$$\tilde{c}_{il} = \frac{\sum_{n=1}^{N} \gamma_{il}(n)}{\sum_{n=1}^{N} \gamma_i(n)} \qquad (9)$$

Where «$l$» represents the $n^{th}$ Gaussian of the observation vectors at the state $q_i$.

$$\mu_{il} = \frac{\sum_{n=1}^{N} \gamma_{il}(n).o_t}{\sum_{n=1}^{N} \gamma_{il}(n)} \qquad (10)$$

While the equation (10) represents the mean of each Gaussian at the state $q_i$ and the term:

$$\sum_{il} = \frac{\sum_{n=1}^{N} \gamma_{il}(n).(o_t - \mu_{il})(o_t - \mu_{il})^T}{\sum_{n=1}^{N} \gamma_{il}(n)} \qquad (11)$$

That represents the variance of each Gaussian distribution within the state $q_i$.

**ANN Phoneme recognition/ distinction:** The multi layer perceptron (MLP) has been widely used in the area of pattern recognition, namely, document recognition, image processing and speech recognition [5,6]. It is the most successful neural network, known by its discrimination capability for pattern classification. It can approximate any function, such a model can easily associate the input shape to its class in a supervised manner.

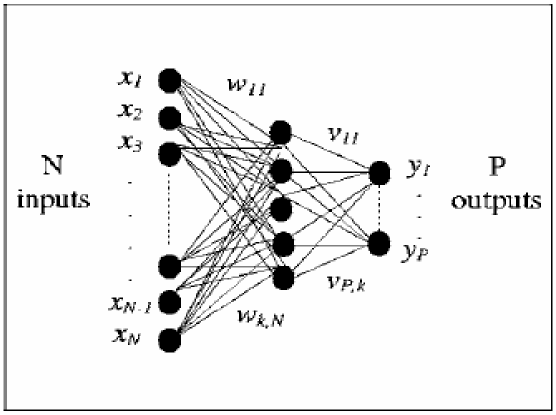

Fig.2: Multilayer network

The ANN is based upon two processes, training and test, which are modeled by the following equations:

The weight update is written as follows:

$$W_{ij}^{l}(k+1)= W_{ij}^{l}(k) + \eta X(k) \qquad 12$$

Where the vector $X(k)$ represents the direction of the minimum, within the gradient method, "$l$" the number of the layer and "$\eta$" represents the positive step constant value, of the learning process (training speed) required to minimise the quadratic output error designed by the difference between the desired value and the computed one.[21]

## AUTOMATIC WORD SEGMENTATION

**Best word pronunciation:** Our automatic speech therapy process is divided into two parts, a word best pronunciation score is given via the EM algorithm then a phoneme deviation rate evaluation is generated by an ANN network.

Initially we concentrate on the word completely, in order to make an adaptation of the tongue in a global context; then we start tuning the recognition at a phonemic level to distinguish between neighbor phonemes

We tested different HMM/GMM models to get a compromise between speed/precision, by using 12 MFCC coefficients, as well as their first Delta, and Delta-Delta dynamic components, with a mixture of 4, 8, 12 GM models, with a maximum of 35 iterations of the EM algorithm, as shown in figure 3.

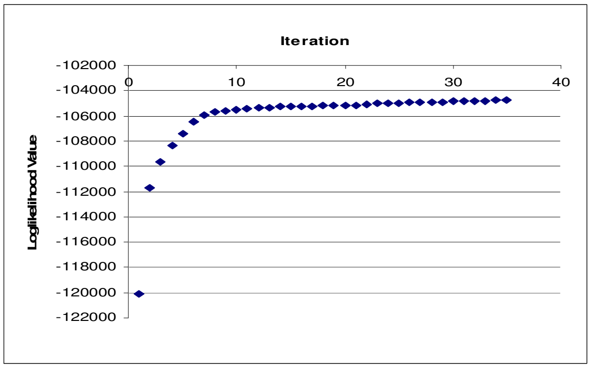

Fig.3: EM convergence

The Viterbi algorithm applied to one of the healthy pronunciations is shown in figure 4.

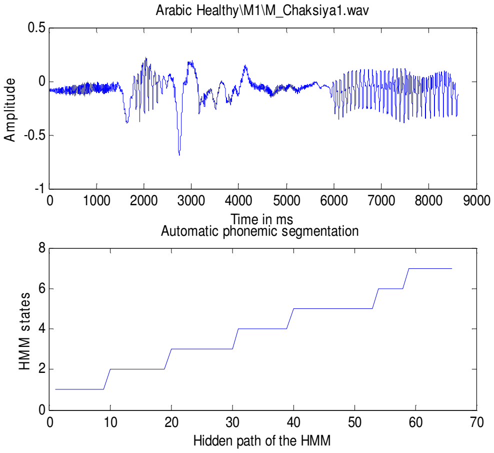

Fig.4: Segmentation of the healthy word [ʃaχsija]

The figure 4 shows that behind the HMM automatic segmentation there is mainly a phoneme correspondence, which is used to compute a second deviation score added to the word evaluation score, in our case the log-likelihood, as presented in figure 5.

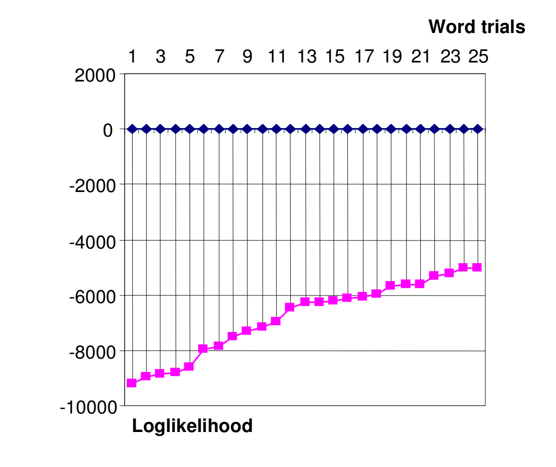

Fig.5: Global word evolution process of the trials

The best pronunciation as shown in figure 5 corresponds to the higher log-likelihood, in this case at trial 25.

In order to make the patient pronounce well, different trials are recorded, analyzed then feed backed at each time to the therapist, as well as to the patient, then a global view is presented visually, to help the correction.

This score might be given to the therapist in different forms, like a grading over 100.

Afters some hours of recording process, this might be days or even weeks, followed each time by the necessary visual and hearing feedbacks; either of the patient himself, in order to hear his pronunciation, or a good pronunciation with a tongue and lisp movement videos, and/or images; we remarked that the patient started to pronounce well the pathological word, as shown in figure 6. Both the worst and best pronunciations are illustrated, aiming to capture the visual change in the speech therapy process; this also could give an impression on the patient to manage the word or phoneme stress.

At this level, a deeper analysis is tackled, answering the question, which phoneme has been really pronounced and to which extent it is morphing toward the good one? And to which degree of likelihood, our system could be liable? to ensure correctness in limited trials.

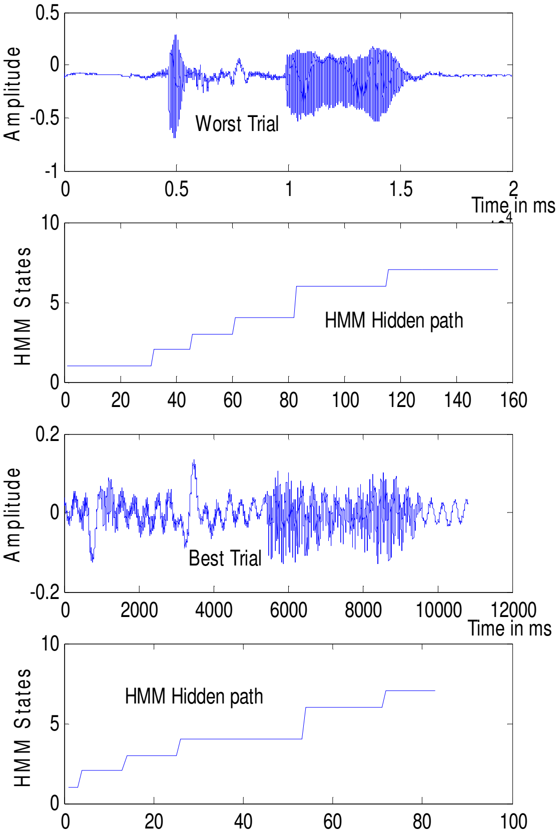

Fig.6: Different occurrences of the word [ʃaχsija] during the therapy trials, (worst and best cases respectively), before and after different visual and hearing feedbacks.

**Phoneme level recognition:** We noticed, that in the whole process of the segmentation, we dealt with [ʃa] instead of [ʃ], this is mainly due to the co-articulation that tends to decrease the effectiveness of the automatic segmentation, compared to the manual process. This later is based upon visual, hearing as well as human experience, while the automatic process is mainly based over distributions and probabilities; our aim is to make a small bridge between the two approaches.

**Phonemic deviation rate:** Data issued from the HMM word model, are reused for comparison in another database consisting of elementary phonemes [ʃ], [s], [θ].

In order to get a good deviation rate score, we tested different configurations, the first configuration concerns 12 inputs, one hidden layer and 4 classes, while the second configuration is composed of 12 inputs, two hidden layers, and 4 output classes. Both networks used the same TIMIT database with 70% as training data and 30% as testing data, as shown in the table 2 and table 3.

For one hidden layer network, we obtained the recognition rates shown in table 2:

Table 2: Choice of the ANN optimal parameters

| Hidden layer | 8 N | 16N | 32N | 64N | 128N |
|---|---|---|---|---|---|
| Phoneme | | Recognition rate % | | | |
| [س][s] | 99 | 10 | 78 | 94 | 89 |
| [ش][ʃ] | 5 | 89 | 62 | 45 | 85 |
| [ث][θ] | 14 | 56 | 60 | 38 | 52 |

For a two hidden layer network, we obtained the recognition rates shown in table 3:

Table 3: Choice of the ANN optimal parameters

| Hidden layer | 16:16:3 | 16:32:3 | 32:16:3 | 32:32:4 | 16:16:3 |
|---|---|---|---|---|---|
| Phoneme | | Recognition rate % | | | |
| [س][s] | 83.7 | 87 | 78 | 80 | 83.7 |
| [ش][ʃ] | 84 | 88 | 88 | 85 | 84 |
| [ث][θ] | 54 | 48 | 58 | 62 | 54 |

From the above results, the chosen configuration is the network made of 2 hidden layers of 32 neurons each, and 4 outputs. Nevertheless, this is not a final rate or score, because we still do not know the real phoneme deviation, that is why a second process is implemented, based on 2 parallel networks that best detect the neighborhood limit of the phonemes, giving a kind of phonemic distance or deviation rate

## RESULTS AND DISUSSIONS

We were concerned, initially, by the first and fourth phonemes of the word C1, the second phoneme [a] (vowel) is being absorbed by the [ʃ], while the third phoneme is of no concern in this pathology, we did a comparison with several occurrences or trials, in order to know the degree of correction and similitude that is being performed by the patient.

During the therapy process, different other words have been used in order to reinforce the phoneme stress; Correction occurred, mainly, when the patients hear themselves, even if the pronunciation is faulty, this helped a lot in the word segmentation. As a second point, the more the database contains different occurrences of the good pronunciation; the better is the segmentation, and the phoneme recognition. That is why; the phoneme database was partly taken from the TIMIT source.

At the end of the speech therapy process, we noticed that patients corrected themselves during the training sessions. This might be a procedure for other pathologies, such as the Facio-Scapulo-Humeral illness, which tends some patients to replace the phoneme [b] by [d] or [f] by [θ].

## CONCLUSION

In this work, we worked on a new approach that deals with Arabic speech pathology computer aided therapy system, we designed an automatic application in this processing context; that helps in the word segmentation and phoneme recognition, of some well selected words, identifying the targeted illness, using a simple computer and a basic recording tool.

This therapy approach is based upon some visual and hearing feedbacks; this may help the patient as well the therapist, to find out the speech illness regions, and follow the evolution of the illness, by a history or log file concept, as well as with some comparative methods; the whole system is designed to be autonomous with less and less need to the continuous presence of the therapist.

We focused initially, as a new trend, on the replacement of the phoneme [ʃ], badly pronounced [θ] or [s].

The designed system allows an automatic recognition of different pathological phonemes, with a deviation rate score, that identifies the "how" a neighbor phoneme is pronounced instead of the original phoneme.

Let us remark also, that from our experiments, we deduced that beyond the computer process, the visual feedbacks have an intensive impact on the speech process as well as the psychological side of the experiment. In fact, patients were at each moment saying: 'why did not we try this before?', and this helped us a lot in the speech therapy.

## REFERENCES

1. Jeffery A. Jones, Munhall, K. G., 2003. Learning to produce speech with an altered vocal tract, Acoustical Society of America, 113(1): 532-543.
2. K. Shahin, 2002. Remarks on the speech of Arabic-Speaking Children with cleft palate, California Linguistic notes, 27(1):1-10.
3. Y. Tsubota, T. Kawahara and M. Dantsuji, 2002. Recognition and verification of English by Japanese students for computer-assisted language learning, School of informatics, Center for information and multimedia studies, Kyoto University.
4. D. J. Burr, 1992. Comparison of Gaussian and neural network classifiers on vowel recognition using the discrete cosine transform. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2:365-368.
5. A. Aktas, O. Schmidbauer, K.-H. Maier et W. H. Feix, 1990. Classification of coarse phonetic categories in continuous speech: statistical classifiers vs. temporal flow connectionist network. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, pp: 89-92.
6. T.ltosaar et M. Karjalainen, 1991. Event-based recognition and analysis of speech by neural networks. Proceedings of the European Conference on Speech Communication and Technology, pp: 1031-1034.
7. A.R. Elobeid Ahmed, 1998. Performance Tests on Several Parametric representations for an Arabic phoneme recognition system using HMM's, Transactions on Information and communications Technologies, 20.
8. D.J. Kershaw, 1997. Phonetic Context-Dependency in a Hybrid ANN/HMM Speech Recognition System, Doctoral Thesis, St John's College, University Cambridge.
9. J.L. Gauvain and C.H. Lee, 1994. Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains. In IEEE Transactions on Speech and Audio Processing, 2(2): 291–298.
10. Z.A. Benselama, M.Guerti, M.A.Bencherif, 2006. Système d'aide a la correction du sigmatisme occlusif , The First International Meeting on Electronics & Electrical Science and engineering, University of Djelfa, ALGERIA.
11. S. Poitoux, 2002. Etude des mesures de confiance dans le traitement de la parole avec application en logopédie, Polytechnic faculty (power) of Lausanne, Swiss.
12. O. Deroo, C. Ris, S. Gielen and J. Vanparys, 2000. Automatic detection of mispronounced phonemes for language learning tools, Proceedings of. ICSLP, 1: 681–684
13. http://www.bungalowsoftware.com
14. B. LE Viet, 2002. Reconnaissance automatique de digits en anglais en conditions bruitées, Thesis of DEA of data processing, Systems and communications, INP of Grenoble, France.

15. D.M. Istrate, 2003. Détection et reconnaissance des sons pour la surveillance médicale, Doctoral Thesis, CLIPS- IMAG, France, pp: 95-129.
16. A.A. Dibazar, S. Narayanan & T.W. Berger, 2002. Feature analysis for automatic detection of pathological speech, Proceedings Engineering Medicine and Biology symposium 02, 2:182-183.
17. C. Levy, G. Linares, P. Nocera, 2003. Comparison of several acoustic modelling techniques and decoding algorithms for embedded speech recognition systems, LIA/CERI, STEPMIND, France.
18. L. Lefort, T. Merlin, J. Bonastre et P. Nocera, 2002. Le projet MTM–Reconnaissance de la parole et du locuteur sur une plateforme embarquée, Computer Laboratory of Avignon.
19. L.R. Rabiner, Fellow IEEE, 1989. A tutorial on Hidden Markov Models and selected Applications in Speech recognition, Proceedings of the IEEE, 77(2):257-286.
20. J.A. Bilmes, 1998. A Gentle Tutorial of the EM Algorithm and its applications to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, Technical Report, ICSI-TR-97-021, International Computer Science Institute, Berkeley.
21. Joe Tebelskis. 1995. Speech Recognition using Neural Networks. PhD thesis, School of Computer Science, Pittsburgh.