# Interface of an Automatic Recognition System for Dysarthric Speech

Brahim-Fares Zaidi
Laboratory of Speech Communication and Signal Processing (LSCSP)
University U.S.T.H.B
Algiers, Algeria
fares.zaidi.1991@gmail.com

Sid-Ahmed Selouani
Laboratory of Research in Human-System Interaction (LARHSI)
University of Moncton, Shippagan Campus
Moncton, Canada
sid-ahmed.selouani@umoncton.ca

Malika Boudraa
Laboratory of Speech Communication and Signal Processing (LSCSP)
University U.S.T.H.B
Algiers, Algeria
mk.boudraa@gmail.com

Djamel Addou
Laboratory of Speech Communication and Signal Processing (LSCSP)
University U.S.T.H.B
Algiers, Algeria
daddou@usthb.dz

***Abstract*—This paper addresses the realization of a Human/Machine (H/M) interface including a system for automatic recognition of the Continuous Pathological Speech (ARSCPS) and several communication tools in order to help frail people with speech problems (Dysarthric speech) to access services providing by new technologies of information and communication (TIC) while making it easier for the doctors to achieve a first diagnosis on the patient's disease. In addition, an ARSCPS has been improved and developed for normal and pathology voice while establishing a link with our graphic interface which is based on the box tools Hidden Markov Model Toolkit (HTK), in addition to the Hidden Models of Markov (HMM). In our work we used different techniques of feature extraction for the speech recognition system in order to improve the dysarthric speech intelligibility while developing an ARSCPS which can perform well for pathological and normal speakers. These techniques are based on the coefficients of ETSI standard Mel Frequency Cepstral Coefficient Front End (ETSI MFCC FE V2.0); Perceptual Linear Prediction coefficients (PLP); Mel Frequency Cepstral Coefficients (MFCC) and the recently proposed Power Normalized Cepstral Coefficients (PNCC) have been used as a basis for comparison. In this context we used the Nemours database which contains 11 speakers that represents dysarthric speech and 11 speakers that represents normal speech.**

***Keywords*—Automatic Recognition System of Continuous Pathological Speech (ARSCPS); ETSI standard Mel frequency Cepstral Coefficient Front End (ETSI MFCC FE V2.0); Hidden Markov Model Toolkit (HTK); Hidden Models of Markov (HMM); Human/Machine (H/M); Technologies of Information and Communication (TIC); Mel Frequency Cepstral Coefficients (MFCC); Perceptual Linear Prediction (PLP); Power Normalized Cepstral Coefficients (PNCC)***

## I. INTRODUCTION

An interface for the control of Automatic Recognition System of Continuous Pathological Speech (ARSCPS) [1] can be very useful for speakers with dysarthria which is a neurological disorder of speech that affects millions of people. A dysarthria man has significant difficulty in communication, according to Aronson [2]; Dysarthria [3] covers various speech disorders resulting from neurological disorders. These disorders are related to the disturbance of the brain and stimulus nerves of the muscles involved in the production of speech. The characteristics of this pathology are slow, weak, imprecise or uncoordinated speech musculature movements [4]; the result is unintelligible speech.

The main objective of this work is to help the dysarthria people with the proposed interface that allows doctors to make their first diagnosis according to the type of dysarthria. Thus, patient intelligibility rate can be classified and other clinical assessments can be given to patients.

Our platform includes an ARSCPS based on the Hidden Models of Markov (HMM) [5] built by us. Furthermore, a pathological database NEMOURS [6] is integrated for improving the speech intelligibility of dysarthria patients.

This paper is outlined as follows: The following section presents the NEMOURS database. Section III explains the realization steps of an ARSCPS with Hidden Markov Model Toolkit (HTK) [5], [7]. Section IV, V and VI shows the results of the recognition with different techniques of features speech extraction. Section VII presents the different parts of our own ARSCPS which is the principal blocks of our H/M interface. We conclude this work in Section VIII.

## II. NEMOURS DATABASE

The Nemours database is constituted by 74 sentences spoken with varying degrees of dysarthria for each one of the 11 male speakers.

The composition of the NEMOURS database is 11 male speakers with different degrees of dysarthria and 11 normal

male speakers; 74 sentences spoken for each one of the 11 male speakers; as a result, we have 814 sentences for pathological voice and 814 sentences for normal voice; look at Table 1.

TABLE I. NEMOURS DATABASE

| NEMOURS DATABASE | Normal | Pathology |
|---|---|---|
| Number of Speakers | 11 | 11 |
| Number of Sentences For Each Speaker | 74 | 74 |
| Number of Total Sentences | 814 | 814 |

## III. STEPS OF REALIZATION OF AN ARSCPS WITH HTK

Our adapted ARSCPS for the NEMOURS database is based on the monophone models. These models require several successive treatment steps for their realization (see Fig. 1). However, during the development of our system, an additional step, which we call Step 0 has been necessary for its realization.

- **Step 0:** Transcription of texts;
- **Step 1:** Grammar of ARSCPS;
- **Step 2:** Dictionary;
- **Step 3:** Sound data;
- **Step 4:** Creating transcript files;
- **Step 5:** Acoustic parameterization of the system and data encoding (MFCC [8], ETSI FE V2.0 [9], [10], PLP [8], and PNCC [11], [12]);
- **Step 6:** Creation of HMM models;
- **Step 7:** The pause model;
- **Step 8:** Realignment of the training data;
- **Step 9:** Recognition of test corpus.

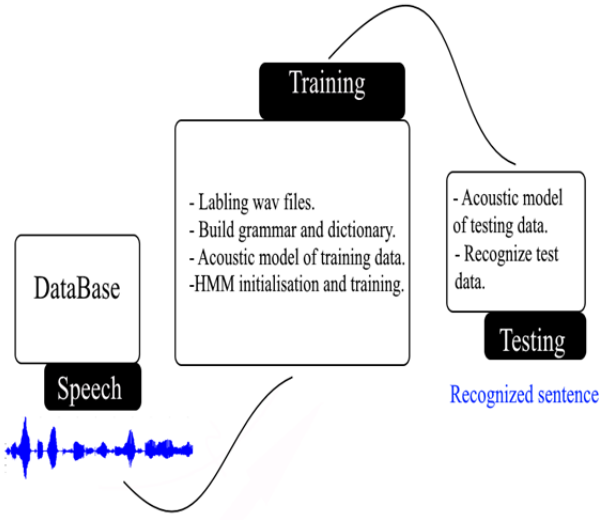


Fig. 1. Steps of an automatic speech recognition system [13].

For training steps (Fig. 2) we used 814 sentences, (for example from 1 to 74 sentences for each speaker), for testing steps (Fig. 3), we tested our ARSCPS with 352 sentences, (for example from 1 to 32 sentences for each speaker); look at Table 2.

TABLE II. TOTAL NUMBER OF SENTENCES FOR TRAINING/TESTING

| NEMOURS DATABASE | Normal | Pathology |
|---|---|---|
| Total Number of Sentences for Training (70%) | 814 | 814 |
| Total Number of Sentences for Testing (30%) | 352 | 352 |

We note that the sentences taken for the test are those that are already exist in the training, because our goal is the realization of an interface H/M that allows doctors to make their first diagnosis; and to make this diagnosis on the patient's disease, the patient have to pronounce sentences imposed by the doctor.

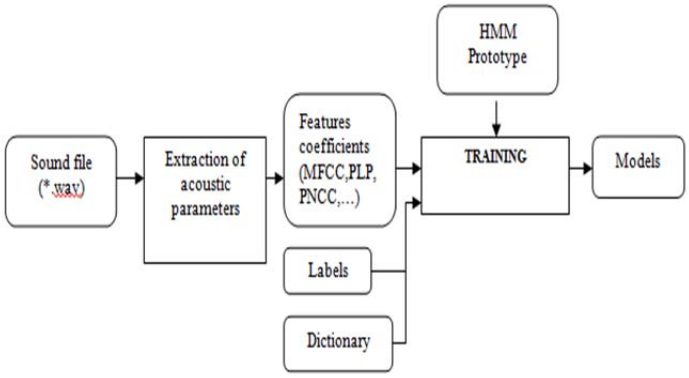


Fig. 2. Training phase.

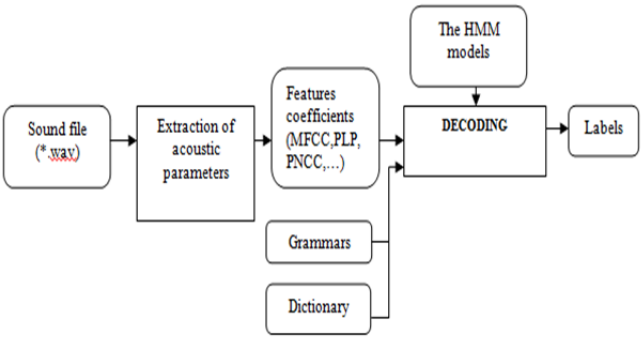


Fig. 3. Recognition phase.

## IV. EVALUATION OF RECOGNITION RESULTS

The generation of the test data acoustic model has been done, the deletion errors (D), the substitution errors number (S) as well as the number of insertion errors (I) are calculated (see Fig. 4).

Three types of recognition rates that correspond to sentence correction (1); word correction (2)' and word accuracy (3) are calculated.

The equations used are respectively:

- For SENT:

$$\text{Sentence Correction \%} = \frac{N-S}{N} \times 100 \quad (1)$$

N: Is the overall number of sentences.

- For WORD:

$$\text{Word Correction \%} = \frac{N-D-S}{N} \times 100 \quad (2)$$

$$\text{Word Accuracy \%} = \frac{N-D-S-I}{N} \times 100 \quad (3)$$

N: Is the overall number of words.

```
==================== HTK Results Analysis =======================
  Date: Mon Mar 20 12:55:54 2017
  Ref : C:\Users\ZAIDI\Desktop\ASRN\RECOGNIZER\labels\N1.MLF
  Rec : C:\Users\ZAIDI\Desktop\ASRN\test_new/result_clean.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=73.86 [H=260, S=92, N=352]
WORD: %Corr=96.54, Acc=93.09 [H=2039, D=45, S=28, I=73, N=2112]
================================================================
```

Fig. 4. HTK Results of ARSCPS.

To achieve this step of performance, a lot of scripts have been used for building our system.

## V. Acoustic Model of the Training/Testing Data

The objective of the extraction of the useful information done by the speech of NEMOURS database is to solve the problems of the speech recognition. ETSI MFCC FE, PLP, PNCC, and MFCC features are extracted from the speech.

## VI. Experimental Results

Table 3 shows the sentence correction, word correction, and the accuracy of the word from the four techniques of feature extraction for the system of speech recognition with voice normal samples. The best sentence correction, word correction, and the accuracy of the word with the proposed features can be found using ETSI FE V2.0 (MFCC_0_D_A) which is 86.65 %, 99.15 % and 97.40 %, respectively.

For the samples of voice pathological, Table 4 shows the sentence correction, word correction, and the accuracy of the word by using four techniques of feature extraction. The best sentence and word correction as well as the word accuracy are 73.86%, 96.54% and 93.09%, respectively which is for applying ETSI FE V2.0 (MFCC_0_D_A) feature with 39 coefficients. The sentence and word correction as well as the word accuracy of MFCC_0_D_A and PLP_0_D_A with 39 coefficients are almost the same, the word correction of MFCC_0_D_A and PNCC_0_D_A with 39 coefficients are the same.

In addition, for PNCC_0_D_A with 39 coefficients, the accuracy is 84.33%, and represents the worst result. For testing the system performance, we have applied the four feature extractions for speech recognition system with normal and pathological voice as shown in Tables 3 and 4. The ETSI FE V2.0 (MFCC_0_D_A) outperforms the previously techniques in terms of sentence correction, word correction and word accuracy.

TABLE III. Experiments with Voice Normal Samples

| **Normal voice (NEMOURS Database)** | | | | |
|---|---|---|---|---|
| **Features extraction** | **MFCC** | **PLP** | **ETSI FE V2.0** | **PNCC** |
| **Sentence correction (%)** | 85.23 | 81.25 | 86.65 | 83.52 |
| **Word correction (%)** | 99.10 | 98.11 | 99.15 | 98.72 |
| **Word accuracy (%)** | 96.73 | 96.16 | 97.40 | 96.54 |

TABLE IV. Experiments with Voice Pathological Samples

| **Pathological voice (dysarthria) (NEMOURS Database)** | | | | |
|---|---|---|---|---|
| **Features extraction** | **MFCC** | **PLP** | **ETSI FE V2.0** | **PNCC** |
| **Sentence correction (%)** | 67.05 | 66.19 | 73.86 | 62,50 |
| **Word correction (%)** | 95.60 | 95.98 | 96.54 | 95,60 |
| **Word accuracy (%)** | 86.84 | 86.41 | 93.09 | 84.33 |

## VII. Automatic Recognition System Interface of Continuous Pathological Speech (ARSCPS)

The interface includes an ARSCPS based on HMM models and the HTK toolbox [14].

We describe the different parts of an ARSCPS control interface using the DELPHI software and the PASCAL programming language [15].

### *A. Part 1: Management of the NEMOURS Database Textual and Sound Files*

For realizing this part (Fig. 5), we have written programs under DELPHI for:

- Loading sound files which allowed us to create the text files containing the links for access to the NEMOURS database.
- Listening to the sound files with extension ".wav".
- The listened Files transcription.
- Display the automatic recognition result of the listened continuous speech files to text format (Speech to Text).
- Synthesis of the recognized text files (Text to Speech).

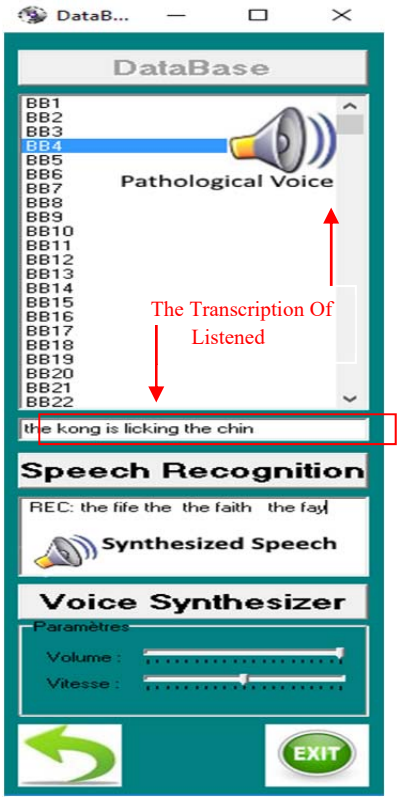


Fig. 5. DataBase (NEMOURS), speech recognition and voice synthesizer.

## B. *Part 2: Result of the Voice Recognition*

In this part (Fig. 6) of the interface, our system:

- Make the link with Part 1.
- Allowing the passage between ARSCPS and our interface.
- Compare the recognition of the NEMOURS database sentence.
- Show of the obtained results in percentage format.

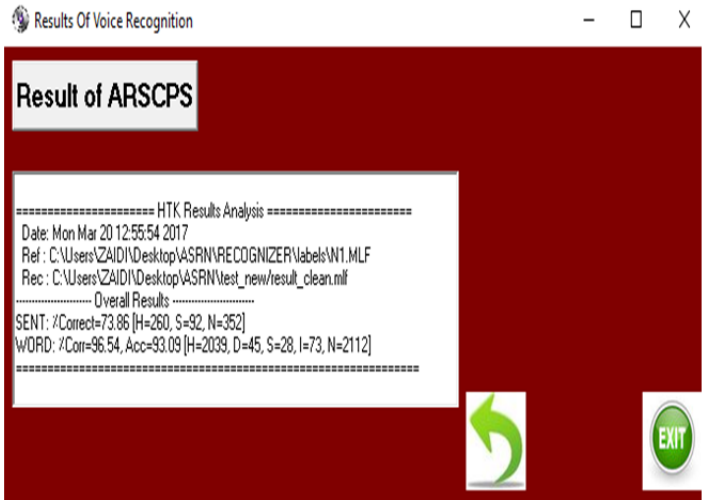


Fig. 6. Automatic recognition system result of continuous pathological speech (ARSCPS) with ETSI FE V2.0 (MFCC_0_D_A).

## C. *Part 3: Automatic Speech Recognition System in Real Time*

The objective of this interface part (Fig. 7) is to realize an automatic dictation system and attempt to make the system independent of the NEMOURS database speakers and always working in real time.

### *1) Part 3.1: Interface of Sound File Recording*

To do this, a recording module has been programmed with different acquisition parameters:

- The type of channels (mono or stereo).
- Bit (16 Bit).
- Sampling frequency (8000 Hz ou 16000 Hz).

### *2) Part 3.2: Control Interface of ARSCPS in Real Time*

In this part, our ARSCPS in real time:

- Makes the link with interface of Sound file recording.
- Allows passage between the new ARSCPS in real time and our interface.
- Compares recorded sentence with sentences of NEMOURS database.
- Displays the recognized sentence.


Fig. 7. Speech recognition in real time.

## VIII. CONCLUSION

In this paper, we proposed an ARSCPS using HTK. For the feature extraction we use ETSI FE, PLP, PNCC, and MFCC. For 22 speakers (for both normal and pathology) 114 words and 46 phonemes are applied in the experiments. The highest accuracy 97.40% is obtained ARSCPS by HTK using ETSI FE feature in the normal voice case. For the pathological voice, we achieved the highest accuracy 93.09% using ETSI FE features.

The ARSCPS that we wanted to implement must satisfy certain conditions which are:

- Continuous speech.
- Independent to speakers (Multi-speakers).
- Large vocabulary.
- Especially dedicated to the recognition of pathological speech (Dysarthria).

The proposed interface allowed us:

- The text and sound files management of the database.
- Listening to the sound files of extension ".wav".
- The transcription of listened Files.
- Display the result of ARSCPS of listened files to text format (Speech to Text).
- Synthesis of recognized text files (Text to Speech).
- Display of results of ARSCPS in percentage format.
- Make the system independent from speakers and make it working in real time.

Each of these objectives is a challenge to overcome. Therefore, realizing a complex system that treats a large number of complications requires: The writing of several scripts (XML, HTML, PASCAL, and HTK) [16]-[18], manipulation and management of a large number of scripts lines, and the management of sound files, text and binary data.

Today our challenge is to provide an interactive platform including the new TIC, making them accessible to people with communication disorders and proposing to the doctors a diagnosis system of the disease severity in order to take a quick decision.

## REFERENCES

[1] M. J. Alam, P. Kenny, P. Dumouchel, and D. O'Shaughnessy, "Robust feature extractors for continuous speech recognition," IEEE Xplore. European Signal Processing Conference (EUSIPCO), Lisbon, Portugal, November 2014.

[2] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential diagnostic patterns of dysarthria," Journal of Speech, Language, and Hearing Research, vol 12, pp. 246-269, June 1969.

[3] K. T. Mengistu, and F. Rudzicz, "Comparing humans and automatic speech recognition systems in recognizing dysarthric Speech," Springer, C. Butz, and P. Lingras, Berlin, Heidelberg, vol 6657, pp. 291–300, 2011 [Canadian Conference on Artificial Intelligence, p. 291, 2011].

[4] K.M Yorkston, D.R.Beukelman, K.R Bell, "Clinical management of dysarthric speakers," Journal of Neurology, Neurosurgery, and Psychiatry (JNNP), Boston, vol 51, issue 11, p.1467,1988.

[5] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK book," version 3.1, pp. 1-277, 2006.

[6] X. Menéndez-Pidal, J B. Polikoff, S M. Peters, J E. Leonzio,and H T .Bunnell, "The Nemours database of dysarthric speech," IEEE Xplore. Fourth International Conference on Spoken Language Proceedings (ICSLP), Philadelphia, PA, USA, August 2002.

[7] Y. J. Kim, "A simple example of generating a recognition system using HTK, for a word level recognition", Hanbat National University, pp. 1-41.

[8] A. H. Gharbi, "Sélection de paramètres acoustiques pertinents pour la reconnaissance de la parole- Selection of acoustic parameters relevant to speech recognition," PhD thesis, 09 december 2012.

[9] L. Guo, X. He, Y. Lu, and Y. Zhang, "A low cost robust front end for embedded ASR system," ISCA Archive, Kent Ridge, Singapore, 2006 [International Symposium on Chinese Spoken Language Processing, December 2006].

[10] Nitisha, and A. Bansal "Speaker recognition using MFCC front end analysis and VQ modeling technique for Hindi words using MATLAB," International Journal of Computer Applications, Sonipat, Haryana, India, Vol 45, May 2012.

[11] C. Kim, and R. M. Stern, "Power Normalized Cepstral Coefficients (PNCC) for robust speech recognition" IEEE Transactions. Audio, Speech, and Language Processing, Vol 24, pp. 1315-1329, July 2016.

[12] M. J. Alam, P. Kenny, and D. O'Shaughnessy, "Robust feature extraction for speech recognition by enhancing auditory spectrum", INRS-EMT, INTERSPEECH, University of Quebec, Montreal, Quebec, Canada, 2012.

[13] A. Mohammed, A. Mansour, M. Ghulam, Z. Mohammed, T. A. Mesallam, K. H. Malki, F. Mohamed, M. A. Mekhtiche, and B. Mohamed, "Automatic speech recognition of pathological voice," Indian Journal of Science and Technology, Vol 8, November 2015.

[14] M. Dua, R. K. Aggarwal, V. Kadyan, and S. Dua, "Punjabi automatic speech recognition using HTK," International Journal of Computer Science Issues (IJCSI), vol 9,issue 4, July 2012

[15] Club des développeurs et IT pro- Club developers and IT pro. Les meilleurs cours et tutoriels sur la programmation et l'informatique professionnelle- The best courses and tutorials on programming and business computing (Cours et tutoriels sur la programmation Delphi- Courses and tutorials on Delphi programming). http://delphi.developpez.com/cours/?page=langage[January, 20th, 2016].

[16] B. F. Zaidi, S. Selouani, M. Boudraa, and G. Hamdani, "Human/machine interface dialog integrating new information and communication technology for pathological voice," IEEE Xplore. Future Technologies Conference (FTC), San Francisco, CA, USA, January 2017.

[17] M. Nebra, "Apprenez à créer votre site web avec HTML5 et CSS3- Learn how to create your website with HTML5 and CSS3'', pp. 1-248, June 2013.

[18] Voxeo An Aspect Company, "XML DevelopmentLanguages Documentation", W3C, 685 Clyde Avenue Mountain View, CA 94043, version 2.1, pp. 1-254, January, 20th, 2015.