Vol. 5, Issue. 3, 2018, ISSN 2349-0780

Comparative and selective study of systems and libraries of speech recognition.

LAMKADAM Abdelmajid
LISTA, Faculty of Sciences, University of Sidi Mohamed Ben Abdellah,
P.O Box 1796 Atlas, Dhar El Mehraz,
Fez, Morocco.
E-Mail: Majidxy@yahoo.fr

KARIM Mohamed
LISTA, Faculty of Sciences, University of Sidi Mohamed Ben Abdellah,
P.O Box 1796 Atlas, Dhar El Mehraz,
Fez, Morocco.
E-Mail: Karim_lessi@yahoo.fr

Abstract: Voice recognition [1] or precisely that speech recognition is an option that is built into most multimedia devices today. Taking into account the progress noted in recent years in automatic speech processing, research studies done [6], the ASR affected a large number of applications in this field, those open or closed source, free or paid, some related to specific platforms and languages. These developments were favored by a sound performance evaluation system of speech recognition.

Taking into account our objective to choose one tool among others, what are the criteria and constraints of selection? So we need a complete description and a comprehensive review by putting more emphasis on the characteristics [1], [6], general unrelated to the specifications and to particular cases.

In this contribution, we quoted different software [6], systems [13] of speech treatment, and libraries [11] which are most used currently, in result were presented and described the various components, after we have the wisely organized according to different criteria [5], to reach a clear choice and convinced of a tool or library for our further studies in the field of speech recognition.

Keywords: Automatic Speech Recognition (A.S.R), Voice Recognition (V.R), ASR System, ASR Toolbox, Speaker Recognition (S.R), Average Error Rate (A.E.R), MFCC, HMM, GMM, JANUS, JULIUS, HTK, SPHINX, PYTHON, MATLAB.

I. INTRODUCTION

Today speech technologies [5], [6] sweep our daily lives, namely interactive voice services, embedded voice commands. Indeed, a large number of products are currently available on the public market, and field of general security such as mobile phones, tablets, avionics and robotic tools...

Most current systems [13] recognition are based on probabilistic and deterministic approaches, containing two major modules, namely the acoustic decoding module and language modeling module.

Now other programs go further and integrate the techniques of speech recognition to evaluate and test the pronunciation of the speaker and detect errors.

Some of the work is dedicated to the comparative and selective study of existing systems, to quote their contexts, report their consideration of the acoustic characteristics of the speech signal, their learning and extraction [4] methods, and their recognition algorithms [6], then attach to the correct and meaningful classification of these software of speech recognition.

II. MULTITUDE OF ASR APPLICATIONS

A. Features

A recognition application is divided into two main phases, a training phase with the aim of extracting acoustic vectors, and a classification phase acoustic fingerprints whose purpose is speech recognition. According to a detailed description in [1], [6], these applications offer usually three main features:

1) Voice dictation:

With this task, it is possible to write documents without having to type any keyboard key, the ability of particular interest to people for whom intensive practice a keyboard is still a tedious task.

2) Voice control:

Via this task, some software [13], [18] (as Dragon Naturally Speaking) provide the ability to fully direct its computer pronouncing a list of voice commands. This possibility will put computers within the reach of people with disabilities preventing them from using the keyboard or mouse.

3) Vocal synthesis:

Using this task which is the inverse operation to the previous two spots, the machine can give voice from text data. This allows the blind to use machines.

B. Review of ASR tools and software

In today's market (net), there are several software or tools [3], [13] of speech recognition can be used, but not all gratis or free, among the most common are mentioned:

- **PerlBox** : Based on Sphinx, can control the KDE environment through voice.
- **teliSpeech** : Professional software Telisma, multilingual, speaker-independent, pooled.
- **Crescendo** Property Owner Crescendo Systems, medical sector, multilingual, easy adaptation.
- **MacSpeech**: Owner Macintosh, paying, multilingual, depends on the speaker.
- **ViaVoice:** IBM Tool, multi speaker, large vocabulary, efficient, simple, effective.
- **Xvoice**: Paid software, uses a library ViaVoice of IBM.
- **CLIPS-Text-TK**: open source Tool box, learning and filtering data via web.
- **EMACOP**: Developped by GEOD CLIPS, mono locutor, speech corpus management.
- **Gnome-voice-control**: Common Tool integrated in GNOME.
- **Voice Reader Studio/Home**: paying, integrated tool for Windows, multilingual.
- **DigtaSoft Voice**: based on Dragon NaturallySpeaking, suited to professional users.
- **SPEESHLOGGER**: based on Google software, gratis, easy, editable, uses for translation.
- **Braina**: Software Brainasoft, secured (uses SSL), easy, efficient, compliant networks.
- **TAZTI**: edited by Voice Tech Group Inc, free, surf the internet, French language.
- **SONIC.**
- ...

C. Characteristics

These various applications [13] or tools [6] differ among other various characteristics [1] among:

- Platform used,
- Quality and cost of the equipment,
- Resistance to recording conditions,

- Learning speed models,
- Performance of the classification phase,
- Processing parameters used and their number,
- Type of extractors and classifiers,
- Modularity to add or remove a speaker,
- Robustness to changes in intra- and inter-speakers,
- Text dependence, language and vocabulary.

D. Problem

The mentioned tools or software above, are generally developed for a given language. Their use in another language requires another effort to the collection of the database [4] speech with sets of words and phrases spoken by a large sample of speakers.

Given the progress and the problems cited in [5], most of the work [6], [11] for the dominant or minority languages are still insufficient and are less interesting, there is still a big effort to provide.

More methods [4], [6] voice recognition is capable of recognizing spoken sentences differently, plus they are performing. We must therefore adapt the software to the learning phase of the corpus of speech.

So the questions are revealed face the plurality of the software, to fix the choice of one among the others are:

- What tool is reliable and effective?
- What tool is available and affordable?
- What tool is easy to use?
- What tool is suitable for a very specific case?

III. CONTEXT AND WORKING METHOFOLOGIE

A. Comparison criteria

Systems that in our study are distinguished from each other by several criteria, in this work we focus on:

- License and software source.
- The nature and amount of speech.
- Type of extractor and classifier.
- Average Error Rate (A.E.R).

B. Quality of recognition

The quality [4] of recognition depends directly on the quality of voice data, which is information relating to an own voice, as different phonemes, different words, and different ways of pronunciation...

If the information is important and known by the system, automatically his reaction and his choice to make are better.

C. Evaluation method: Error rate

According to studies carried out [6], [13], the evaluation method [2] and the performance level [4] of such a tool, is evaluated by the average error rate (A.E.R) which is the average of the error rate obtained by several testing of various studies [6], [11], [13], defined by:

AER = Σ (100-Recognition Rate) / Total Number of Tests (1)

D. Classification criteria

Criteria classification systems [3], [4] of voice recognition or just the systems of speech recognition are many and based on several parameters:

1) Classification by number of speakers:

- Mono speaker system.
- Multi speaker system.

2) Text dependence classification:

- Text dependent system.
- Text independent system.

3) Classification by quantity of data for learning:

- Continuous speech.
- Isolated words.

IV. COMPARISON OF ASR SYSTEMS

A. Review of the most used ASR systems

As shown in several studies [1], [6], [13], systems invaded the speech recognition market are the following:

1) DRAGON DICTATE 30K:

According to the study and [13], this system of speech recognition using phoneme modeled by HMM and a statistical language model: learning takes place on the segmentation of several thousand words chosen to represent all phonemes in most existing contexts. This system is suitable for most European languages.

According to [6], the speaker adaptation -having realized learning- is performed in successive recognitions of 1000 to 2000 words, which resulted in an error rate by 6%.

*2) TANGORA***:**

In view of section [6] has described this system as a system of the speaker dependent recognition, from a vocabulary of 5,000 to 20,000 words. It uses a model of tri-gram language and fénone as acoustic base unit. In the classification phase, it uses the basic discrete HMM models, based on a dictionary of 200 references. It is suitable for multiple languages.

For a test [6] with 5,000 words, the error rate was 2.9%, and with 20,000 words the rate was 5.4%.

3) PHILIPS RESEARCH SYSTEM:

According to a study [6] already achieved, this system is characterized by the use of a model of bi-gram language and basic acoustic unit is the phoneme. It is suitable for the long German for a large vocabulary (nearly 13,000 words). This system is dedicated to radiologists and lawyers.

The study specified in [6] for a learning 9 hours of speech per speaker, the error rate exceeded 10%.

4) DICTATION VOICE LIMSI of:

According to a study [6] already made in 1995, LIMSI Dictation system uses as HMM classifier, and MFCC and their first and second derivatives as an extractor of the acoustic parameters, the base unit is the

allophone, uses a model of language is bi-gram and tri-gram, modeled by multi-models (male and female models and different phonological variants (inter and intra words).

Research and testing [14] on a basic vocabulary of 20,000 words, resulted in the error rate equal 9%.

5) DRAGON NATURALLY SPEAKING:

This system [1] of Scansoft platform, very convenient for multi speaker recognition, with continuous speech, learning is based on the phoneme modeled by the HMM.

System presented on paid versions adapted to different languages and is compatible with other programs such as Microsoft Word; Notepad…This system is very useful to write a text or to order computers.

According to the study [1], and a project [17] in UHC (University Hospital Charles Nicolle) of Rouen, the error rate for speech recognition of doctors varies between 1% and 5%.

B. Summary of ASR systems

The Table (I) below summarizes the classification of voice recognition systems as relevant and evaluative features in the field of speech recognition.

Table (I) Classification of the most used of (A.S.R) systems.

Software	Locutor	Speech	Language	Basic processing unit	Extractor (Treatment)	Classifier (Recognition)	AER (%)
DRAGON-DICTATE 30K	Mono	Continue	-	Phoneme	-	HMM	6
TANGORA	Mono	Continue	Tri-gramme	Fenone	MFCC	HMM	4
PHILIPS RESEARCH SYSTEM	Multi	Continue	Bi-gramme	Phoneme	Divers	HMM	10
DICTÉE VOCALE DE LIMSI	Multi	Continue	Bi/Tri-gramme	Allophone	MFCC	HMM	9
DRAGON NATURALLY SPEAKING	Multi	Continue	Multi	Phoneme	-	HMM	3

The graphical representation of the table above is:

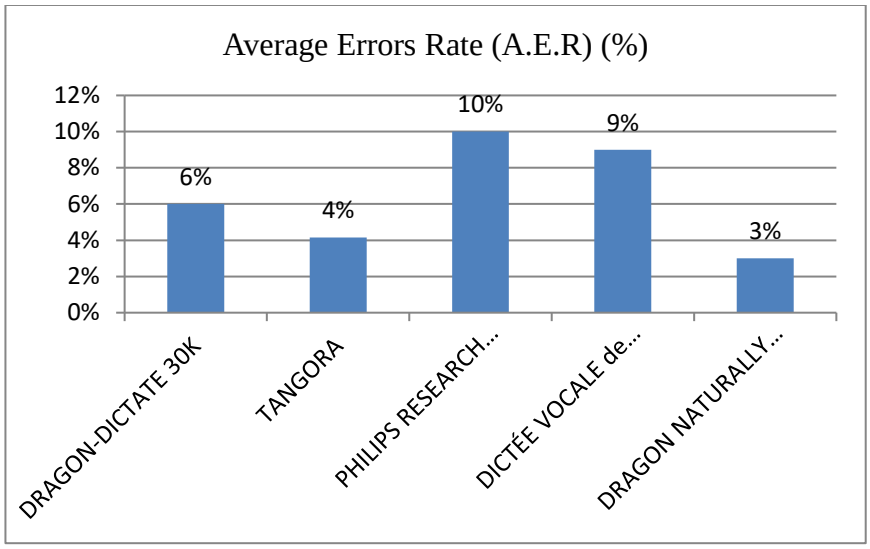

Fig. 1 Comparison of Speech Recognition systems.

C. *Discussions*

- Dragon NaturallySpeaking has a lower error rate compared to others.
- The different systems are based on continuous speech, and one classifier (HMM).
- Some systems extractors are confidential.
- Basic processing units are different from system to other.

V. COMPARISON OF THE ASR LIBRARIES

A. *Review of the most used ASR libraries*

In this part, we are interested to present some libraries [6] more known and used in the field of speech recognition; these toolkits are quickly contributed to the progress of systems of speech recognition:

1) Tool Box HTK:

HTK [7] (Hidden Markov Model Toolkit) is a set of portable modules written in C, enabling the creation and manipulation of HMM [6] (Hidden Markov Models). These various tools facilitate voice analysis, learning HMM, performing tests and analysis of results.

HTK is mainly used in research in the field of speech recognition, speech synthesis, handwriting recognition and the recognition of DNA sequences. This famous software is developed in CUED, and its derivatives are: GMTK, PVTK,

According to [7] published research, the mean error rate for the speaker recognition with six states in the HMM topology varies between 6% and 8%.

2) SPHINX: javaspeech

SPHINX [8], [9] is an open source tool, developed at CMU, written in Java and C variants, well known in the field of speech recognition, presented in several versions 2, 3, 4 and CMU.

Version 4 is highly configurable, recognition by Sphinx 4 supports including individual words and sentences continued. Its architecture is modular to allow for further research and test new algorithms. It is a recognition system, continuous speech, speaker independent. The system is based on the statistical description of each word using as a base unit, the allophone and a recursive finite state grammar.

The CMUSphinx tool box [8] is the leader in speech recognition with various tools used to create voice applications. It contains a number of packages for different tasks and applications.

According to research carried out [14], [8], an error rate of recognition varies from 2% to 4%, on the training set for the phonetic models HMM type that are learned about 100,000 sentences uttered by 500 male speakers.

3) PHYTON: Kaldi

Python as shown in [10] is a general programming language and scientific excellence by as MATLAB, very useful in case of security, multi-paradigm, interpreted at runtime. This is a solution with a free license without restrictions. Dynamic typing offers advantages in development time and flexibility of source code. The interpreter has been ported to all major platforms (Windows, OSX, UNIX, Android ...). Ease of integration, allows him to play the role of bridge between different programming languages.

According to a professional project [12] conducted at the University of Grenoble, the error rate is set at 10.03%.

4) JANUS: Tcl/Tk

Under Article [15] published in 1998, JANUS is a translation platform was developed in the laboratory of ISL universities: Carnegie Mellon and Karlsruhe, this system contains all the components needed to develop a system of recognition of phonemes large vocabulary, it is based the HMM and neural networks to classify the acoustic vectors.

The results of tests [15] on the corpus BRIEF-80 decreased the error rate to 9%.

5) JULIUS:

This system briefly described in [16] is an open source, with a free license; Designed for dictation, uses acoustic models and language models generated using the toolbox HTK, also uses GMM for extraction, distributed with Japanese or English language patterns.

After several attempts of adaptations, the error rate is decreased to 11% with a recognition system sound/word [16] continuously developed and evaluated within the framework of the European CompanionAble project.

6) ALIZE: SpkDet

ALIZE is free tools box for automatic recognition of speakers, the cross-platform compatibility under a multi-layer architecture facilitating function as GMM classifier of acoustic vectors.

This ALIZE project [18] developed with C++ in LIA (Computer Laboratory Avignon), provides a set of low-level and high-level frameworks for developing multitasking applications in the field of speaker recognition.

As described in the article [18], the performance of package: ALIZE/SpkDet is demonstrated as part of the survey NIST'06 SRE, an error rate around 5%.

7) MATLAB: voicebox

MATLAB [11] (Matrix LABoratory) a numerical calculation software, it was intended to facilitate access to matrix software developed in the LINPACK and EISPACK projects, thus optimized for the treatment of matrices. This software is widely used in laboratories and scientific research centers, it is available on multiple platforms, it uses a simple and effective language.

The error rate, whichever is announced by the study [12] on an Arabic corpus with Algerian students is 20%, while the work carried out [19] at the Opal University the errors rates vary from 1.5% to 12.4%.

B. Summary of libraries Toolboxes

The Table (II) below shows an incomplete classification of different libraries (toolbox), according to the most relevant criteria and attributes field of speech recognition.

Table (II) Classification of most used ASR Toolboxes.

Tools Box (Libraries)	Open Source	Continue Speech	Basic unit	Extractor	Classifier	Free License	AER) (%)
JULIUS	yes	yes	Phoneme	MFCC	HMM/GMM	yes	11
JANUS (Tcl/Tk)	yes	yes	Phoneme	HMC	HMM/ANN	No	9
HTK	No	Isolated word	Phoneme	MFCC	HMM	No	7
SPHINX (javaspeech)	yes	yes	Phone	LPC	HMM	yes	3
PYTHON (kaldi)	yes	yes	Phoneme	MFCC	HMM	yes	10
ALLIZE (SpkDet)	yes	yes	Phoneme	UBM/GMM	SVM	yes	5
MATLAB (voicebox)	yes	yes	Phoneme	MFCC, HTK	HMM/Viterbi	yes	10.75

The graphical representation of the table above is:

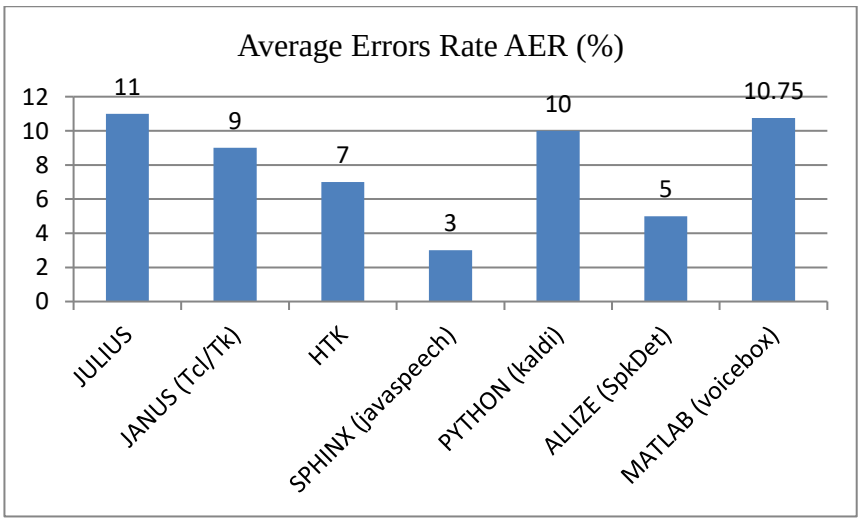

Fig. 2 Comparison of Speech Recognition libraries.

C. Discussions

- The corpus and the test conditions are totally different.
- These figures are means of various tests; they are just approximations for comparison.
- The performance illustrated in error rates vary approximately between 1.5% and 20%.
- For continuous speech and large vocabulary, HTK tool is unsuitable for the management of graphs and trees.
- The majority of toolkits use phonemes as the basic processing unit.
- Most of the toolboxes are mono extractor (MFCC), and mono classifier (HMM).

VI. WORK RESULTS AND FINAL CHOICE

A. Chronology of tests

At first we started our testing in HTK, after installing this tool on Windows and compile these functions directly under MSDOS, we was tested recognition and evaluation results, as these works are already done. Nevertheless, HTK was abandoned in favor of Sphinx-4 for several reasons:

- Many applications use Sphinx, especially JVoiceXML.
- Sphinx has a larger community; it is easier to get help.
- HTK rights are held by Microsoft that offers a restrictive license SPHINX.
- A study [9] sought to compare HTK and SPHINX. It turned out that these two systems are nearly equivalent in error rate, but SPHINX to the advantage of being slightly faster.
- Sphinx adopts LPC as one extractor and HMM as classifier, reflecting the lack of choice of extractors and recognizers,

SPHINX was abandoned to juggle freely with extractors and classifiers of acoustic vectors, our needs are provided by the MATLAB functions and scripts, and through other key features.

B. Why the choice of MATLAB?

Despite the MATLAB error rate towards the other, it was chosen as it allows interactive work is in command mode or programming mode; it is considered one of the best programming languages general, science, it also has the following features and characteristics, compared to other tools:

1) The matrix approach MATLAB can process data without any size limitation and perform numeric and symbolic computations quickly and reliably.

2) With graphics capabilities of MATLAB, it becomes very easy to interactively modify the parameters of the graphics to adapt to our needs.

3) MATLAB is a powerful environment, comprehensive and easy. It has hundreds of mathematical, scientific and technical. For the numerical calculation is much more concise than the archaic languages (C, Pascal, Fortran, Basic).

4) MATLAB also includes a set of integrated tools (Boxes tools) necessary for most users, and extend the MATLAB environment to solve specific types of problems.

5) The open approach enables MATLAB to build tools on the fly. Open for specific applications to specific domains (signal processing, statistical analysis, optimization ...), as can inspect the source code libraries and functions.

6) MATLAB has its own language, natural and intuitive allowing time savings, productivity and creativity, compared to other languages.

7) MATLAB can dynamically links to other programs, exchange data with other applications, or use as motor analysis and visualization.

8) Scientific and even artistic representations of the objects can be created on the screen using mathematical expressions.

9) Available and handled on multiple platforms: Sun, Bull, HP, IBM compatible PC (DOS, UNIX or Windows), Macintosh, iMac and several parallel machines.

10) MATLAB is available in Professional versions and student versions.

VII. CONCLUSION

Given the multitude of systems and toolkits, the study of some open source software has been performed.

First, it was interested in HTK and then move to the use of one of the most known in the world of voice recognition software, namely SPHINX (specifically version 4).

And given the failure of the latter, and the limitations they present about the limited choice of extractors, and comparison algorithms, we were forced to leave by adopting the famous software to perform and complete the tasks of speech recognition.

REFERENCES

[1] Hervé Haut, “La reconnaissance vocale, les systèmes de dicté continus,” Publication technique du SMALS-MVM, TECHN 29, Avril 2005, 1050 BRUXELLES, Belgique.

[2] Olivier Le Blouch, “Décodage acoustico-phonétique et applications à l'indexation audio automatique,” Présentée et soutenue le 12 Juin 2009, à l’Université Toulouse III - Paul Sabatier, France.

[3] V.Steinbiss, H.Hey, R.Heab-Umbach, B.-H.Tran, U.Essen, R.Kneser, M.Oerder, H.-G.Meier,X.Aubert, C.Dugast, D.Giller, “ The Philips Research System for Large-Vocabulary Continuous Speech Recongnition,” EUROSPEECH’93, Berlin, Germany, September 19-23, 1993.

[4] Luiza OROSANU, “ Reconnaissance de la parole pour l’aide à la communication pour les sourds et malentendants,” Soutenance le 11 Décembre 2015, Université de Lorraine, France.

[5] Joseph MARIANI, “Reconnaissance automatique de la parole : progrès et tendances,” LIMSI-CNRS, ORSAY CEDEX, France.

[6] Bruno JACOB, “Un outil informatique de gestion de Modèles de Markov Cachés : expérimentations en Reconnaissance Automatique de la Parole,” Soutenance le 27 septembre 1995. Université Paul Sabatier de Toulouse III, France.

[7] Preeti Saini, Parneet Kaur, Mohit Dua, “Hindi Automatic Speech Recognition Using HTK,” (IJETT)-Volume 4 Issue 6- June 2013.

[8] A.Sadiqui, N. Chenfour, “Système de Reconnaissance Automatique de l’arabe basé sur CMUSphinx,” Annals. Computer Science Series, 8th Tome 1st Fasc- January 2010.

[9] Josef R. Novak Paul R. Dixon Sadaoki Furui, "An Empirical Comparison of Sphinx and HTK models for Speech Recognition," Conference papers of Japanese Acoustic Society, March 2010.

[10] Frédéric Osterrath, "Utilisation du langage de programmation Python et de son écosystème dans le domaine de la science, une étude de cas pour la recherche dans le domaine de la reconnaissance de la parole," le 21 mai 2013, Centre de recherche informatique de Montréal, Québec.

[11]Hamza Fribia, Halima Babi, **"**Etude comparative entre les librairies de Reconnaissance vocale, **"**

[12] Elodie GAUTHIER, "Technologies mobiles pour la reconnaissance vocale des langues africaines," UFR SHS-IMSS, Projet professionnel, Année 2013-2014, Université de pierre Mendès, Grenoble, France.

[13] Baris Ulucinar, "Web WriteIt!," Master thesis report, Université de Fribourg Suisse, Avril 2007.

[14] G. Adda, M. Adda-Decker, J.L. Gauvain, L.F. Lamel, "Le système de dictée du LIMSI pour l'évaluation AUPELF'97," Journées Scientifiques et Techniques du réseau Francophone d'Ingénierie, April-1997, Avignon, France.

[15] Mohammad AKBAR, Jean CAELEN, "Parole et traduction automatique : le module de reconnaissance RAPHAEL," 17th International Conference on Computational Linguistics - Volume 1, COLING-ACL, Aout-1998.

[16] Pierrick Milhorat1, Dan Istrate3, Jérôme Boudy2, Gérard Chollet1, "Interactions sonores et vocales dans l'habitat," Atelier ILADI 2012: Interactions Langagières pour personnes Agées Dans les habitats Intelligents, pages 17–30, Grenoble, 4 au 8 juin 2012.

[17] Yves DROTHIER, "La reconnaissance vocale au service des médecins du CHU de Rouen," JDN Solutions, Copyright 2006 Benchmark Group - Boulogne Billancourt Cedex, France.

[18] J.-Bonastre, N. Scheffer, D. Matrouf, C. Fredouille, A. Larcher, A. Preti, G. Pouchoulin, N. Evans, B. Fauve, and J. Mason, "Alize/spkdet: A state-of-the art open source software for speaker recognition," IEEE Odyssey, The Speaker and Language Recognition Workshop, Jan 2008. Stellenbosch, South Africa.

[19] Fréjus A. A. Laleye, "Contributions à l'étude et à la reconnaissance automatique de la parole en Fongbe," Soutenue publiquement le 10 décembre 2016, Laboratoire d'Informatique, Signal et Image, Université du Littoral Côte d'Opale.