Vocal Pathologies Detection and Mispronounced Phonemes Identification: Case of Arabic Continuous Speech

Naim Terbeh, Mounir Zrigui
LaTICE Laboratory
Monastir-Tunisia
naim.terbeh@gmail.com, mounir.zrigui@fsm.rnu.tn

Abstract

We propose in this work a novel acoustic phonetic study for Arabic people suffering from language disabilities and non-native learners of Arabic language to classify Arabic continuous speech to pathological or healthy and to identify phonemes that pose pronunciation problems (case of pathological speeches). The main idea can be summarized in comparing between the phonetic model reference to Arabic spoken language and that proper to concerned speaker. For this task, we use techniques of automatic speech processing like forced alignment and artificial neural network (ANN) (Basheer, 2000). Based on a test corpus containing 100 speech sequences, recorded by different speakers (healthy/pathological speeches and native/foreign speakers), we attain 97% as classification rate. Algorithms used in identifying phonemes that pose pronunciation problems show high efficiency: we attain an identification rate of 100%.

Keywords: Arabic healthy/pathological speech, language disabilities, phonetic model, forced alignment, artificial neural network, pronunciation problems

1. Introduction

Analysis of characteristics generated from speech signal produced by a speaker may be used to classify this speech to healthy or pathological. Our project focuses in introducing a new probabilistic approach that aims to detect vocal pathologies in the Arabic speech and identify phonemes that pose pronunciation problems. Nowadays, speech therapists use different medical techniques in vocal pathologies detection. Laryngoscopy, Electromyography and Videokimography are the most used (Majidnesha, 2013). But these methods possess a number of disadvantages; in application of these diagnostic methods, the patients feel much discomfort which can distort the produced signal so that it may lead to incorrect diagnosis (Alenso, 2001; Jollife, 2008). Acoustic and phonetic modeling seems the most appropriate and efficient in this area. We use the phonetic transcription to generate a phonetic model of Arabic speech: the percentage of occurrence of each bi-phoneme in Arabic spoken language. The comparison between the reference phonetic model (numerical model (Zouaghi, 2008)) of Arabic speech and that specific to the concerned speaker lead to classify the speech produced by this latter to healthy or pathological.

2. Stat of the Art

In the literature, there are several studies that treat human speeches to detect pronunciation disorders. Also, there are several approaches which are based on features contained in the speech signal:

- Vahid and al. in (Majidnesha, 2013), propose an ANN based approach to classify speeches to healthy or pathological. The proposed method accounts three stages which are extraction of MFCC (Ihichaichareon, 2012) coefficients vector, using the PCA method (Ihichaichareon, 2012; Jollife, 2008) to reduce feature vector (MFCC vector) and use an ANN to classify speeches in input (healthy or pathological).
- Little and al. in (little, 2006) combine between linear classification and biophysics of speech production to online vocal pathologies detection.
- In the (Majidnesha, 2012) work, Vahid proposes a HMM-based approach to classify speech to healthy or pathological. This method accounts three steps which are extraction of MFCC vector, use the LBG algorithm (patane, 2001) to extract the quantization vector and based on HMM model (Bréhilin) the speech in input has been classified to healthy or pathological.
- Kukharchik and al. use in the (Kukharchik, 2007) work, the change of the wavelet characteristics (Kukharchik, 2007) and the Support Vector Machines (Archaux, 2004) to classify a speech sequence in input to healthy or pathological.

Our proposed approach consists in introducing a new probabilistic approach based on phonetic distance (angle which separates two different phonetic models) and artificial neural network to classify Arabic speech to healthy or pathological and identify problematic phonemes.

3. Methodology

The principal objective in this work is to create a platform to assist people with language disabilities and non-native learners of Arabic spoken language to improve their mispronunciations. In this context, we want to follow an acoustic phonetic method to classify Arabic speeches to healthy or pathological. In the case of pathological classified speeches, two proposed algorithms will be used to identify phonemes that pose pronunciation problems. The proposed method to classify Arabic speeches consists to compare between the phonetic model proper to speaker and that of the Arabic spoken language (the reference phonetic model).

The proposed method can be summarized as following:

1. In the first stage, based on n corpus of Arabic healthy speech, an Arabic acoustic model and the Sphinx-align tool (kurki, 2008), we generate n phonetic models (one phonetic model for each corpus).

2. Calculation of the maximum distance between these phonetic models (maximum angle which separates between these models) presents the main objective of the second stage.

3. The third stage is dedicated to generate the reference phonetic model (the average of the n models previously generated).

4. In the fourth stage, for each new speech sequence to be classified, we generate the phonetic model proper to concerned speaker (speaker can be native, foreign, healthy, with disability…).

5. In the fifth stage, based on the comparison between reference phonetic model and model proper to speaker, an ANN classifies the speech in input into two classes: healthy or pathological.

6. Finally, if the Arabic speech is classified as pathological, we use our proposed algorithms to identify phonemes that pose mispronunciations.

The block diagram of our approach is illustrated in the following figure:

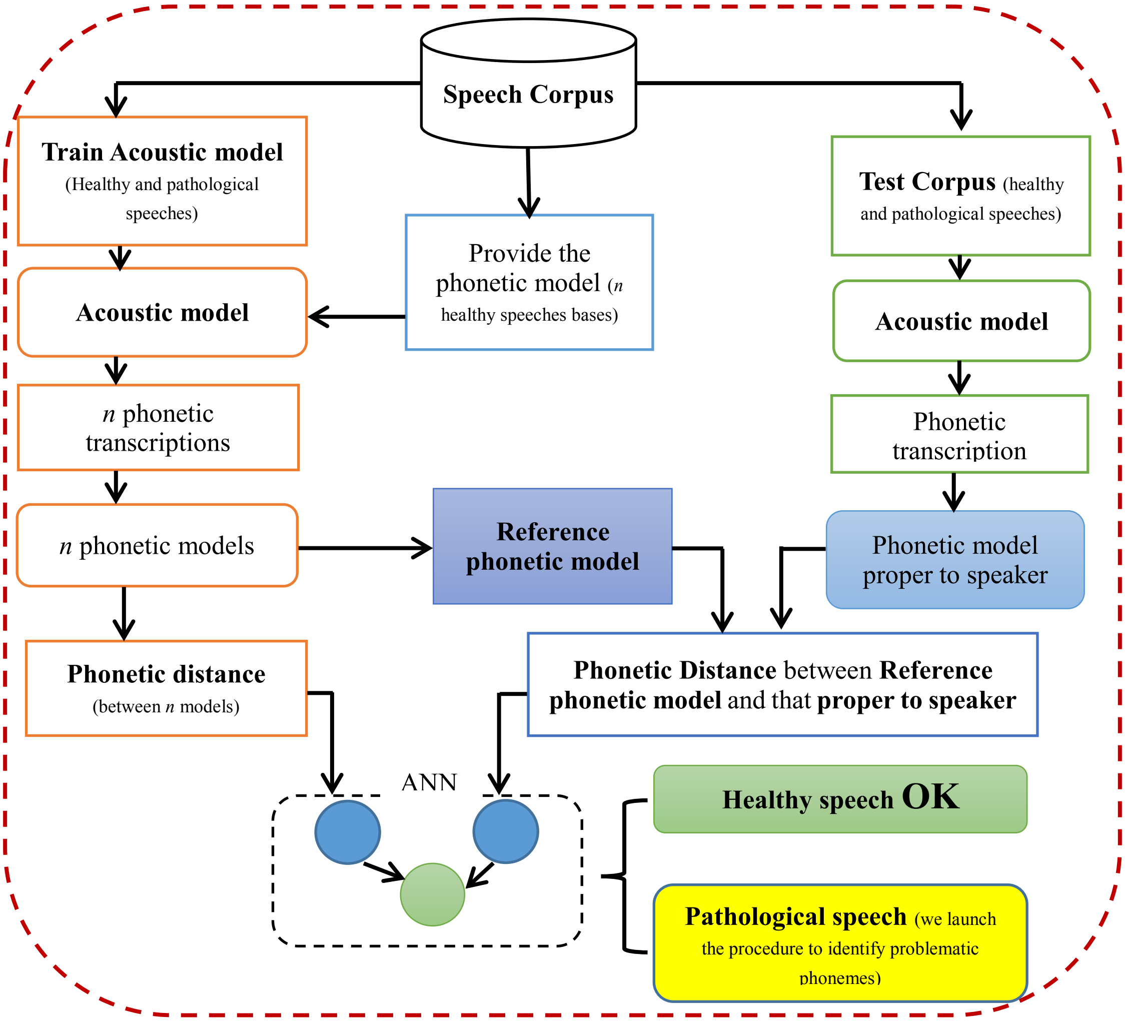
Speech Corpus
Train Acoustic model (Healthy and pathological speeches)
Provide the phonetic model (n healthy speeches bases)
Test Corpus (healthy and pathological speeches)
Acoustic model
Acoustic model
n phonetic transcriptions
Phonetic transcription
n phonetic models
Reference phonetic model
Phonetic model proper to speaker
Phonetic distance (between n models)
Phonetic Distance between Reference phonetic model and that proper to speaker
ANN
Healthy speech OK
Pathological speech (we launch the procedure to identify problematic phonemes)

Figure 1: Our proposed approach

3.1 Phonetic Model Generation

First, to generate an Arabic phonetic model we need an acoustic model of Arabic language and a large speech corpus. The Arabic speech base must be recorded by native speakers and containing just healthy speeches. In the second step, we use the Sphinx_align tool and our acoustic model to generate the phonetic transcription which corresponds to our speech corpus. In the last step and from the resulting phonetic transcription file (output of sphinx_align tool), we calculate probabilities of occurrence for each bi-phoneme in our corpus.

The phonetic model is defined by the vector that its coefficients present the probability occurrences of each Arabic bi-phoneme. In the following figure, we present an extract form the standard Arabic phonetic model.

$$\begin{bmatrix} \boldsymbol{H\_D} \\ \boldsymbol{H\_DH} \\ \boldsymbol{H\_R} \\ \boldsymbol{H\_Z} \\ \boldsymbol{H\_S} \\ \boldsymbol{H\_SH} \\ \boldsymbol{H\_SS} \\ \boldsymbol{H\_DD} \\ \boldsymbol{H\_TT} \\ \boldsymbol{H\_DH2} \\ \boldsymbol{H\_AI} \\ \boldsymbol{H\_GH} \\ \boldsymbol{H\_F} \\ \boldsymbol{H\_Q} \\ \boldsymbol{H\_K} \end{bmatrix} = \begin{bmatrix} \mathbf{0.0015423348\%} \\ \mathbf{0.0046362397\%} \\ \mathbf{9.346364}\boldsymbol{E} - \mathbf{4\%} \\ \mathbf{1.4776867}\boldsymbol{E} - \mathbf{5\%} \\ \mathbf{0.0\%} \\ \mathbf{6.8342975}\boldsymbol{E} - \mathbf{5\%} \\ \mathbf{0.0\%} \\ \mathbf{3.140083}\boldsymbol{E} - \mathbf{5\%} \\ \mathbf{5.5413225}\boldsymbol{E} - \mathbf{6\%} \\ \mathbf{0.0\%} \\ \mathbf{0.0\%} \\ \mathbf{0.0\%} \\ \mathbf{3.6942151}\boldsymbol{E} - \mathbf{6\%} \\ \mathbf{3.8789258}\boldsymbol{E} - \mathbf{5\%} \\ \mathbf{1.4222728}\boldsymbol{E} - \mathbf{4\%} \end{bmatrix}$$

Figure 2: An extract from the reference phonetic model

3.2 Phonetic Distance

Phonetic distance is defined by the angle that separates two different phonetic models. The following pseudo-code summarizes the generation procedure of this distance:

1. We prepare *n* healthy speech corpus (C_i, *i*=1-*n*), and for each corpus, we generate the corresponding phonetic model M_i.

2. We define S={α_{ij}; 1≤*i*,*j*≤n and *i*≠*j*} a set of angles which separate M*i* and M*j* ($\alpha_{ij}=\alpha_{ji}$ and $\alpha_{ii}=0$).

3. We define the value *Max* by the maximum of the set {S}.

4. The value *δ* will be defined by the standard deviation of {S}.

5. We define the value *Avg* by the average of {S}.

6. We calculate the phonetic distance β=Max+|Avg-δ|.

To calculate elements of S, we follow these scalar product formulas:

$$M_i.M_j=\sum_{k=1}^{n} M_i[k]M_j[k] \qquad (1)$$

$$M_i.M_j=||M_i||.||M_j||.\cos(\alpha) \qquad (2)$$

We can deduce that:

$$\mathrm{Cos}(\alpha)=M_i.M_j/||M_i||.||M_j|| \qquad (3)$$

3.3 Speech Classification

For each new speaker, we use a speech sequence recorded by his voice and we follow the same previous procedure (in the section 3.1) to generate his proper phonetic model. By calculating the angle θ that separates between this model and that reference to the Arabic spoken language, we distinguish tow cases:

- If θ≤β, then the speech in input (pronounced by the concerned speaker) is heathy.
- Else (θ>β), the speech is classified as pathological and we launch the mispronounced phonemes identification procedure.

3.4 Artificial Neural Network

An artificial neural network (ANN) as a computing system is made up of a number of simple and highly interconnected elements, which processes information by its dynamic state response to external inputs. ANN models show a high potential to offer solutions to some problems which have hitherto been intractable by computers in the areas of computer science and artificial intelligence. Neural networks are better suited in achieving intelligent systems such as speech processing, image recognition, robotic control, etc.

Processing elements in an ANN are also known as neurons. These neurons are interconnected by means of information channels called interconnections. Each neuron can have multiple inputs; while there can be only one output. Inputs to a neuron could be from external stimuli or outputs of the other neurons. Copies of the single output that comes from a neuron could be input to many other neurons in the network (Lee, 1992).

3.5 Mispronounced Phonemes Identification

During these two following sections (3.5 and 3.6), we note by:

- N is the phonetic model of Arabic spoken language.
- H is the phonetic model proper to speaker.
- M is the set that will contain mispronounced phonemes.
- R is the set that will contain replacement phonemes.

In this section, and for each speech sequence classified as pathological, we'll identify phonemes that pose mispronunciations for concerned speaker (native speaker suffering from language disabilities or non-native learner of Arabic spoken language). The main idea in this treatment is: "A mispronounced phoneme does never appear in the speech phonetic transcription, so we have a coefficient equal to zero in the phonetic model for all bi-phonemes containing a wrongly pronounced phoneme".

A simple comparison between N and H can lead to identify phonemes which pose mispronunciations. The following algorithm is used to identify mispronounced phonemes:

```
1. We propose these two set $M$ and $G$ with $M=G=\varnothing$.
2. for $i$=1 to length($N$)
    if H[i]=0 and N[i]≠0 then
$G=G\cup\{P, P'\}$ such as $N[i]$ equal to the probability of occurrence of bi-phoneme PP' in the reference phonetic model and $H[i]$ equal to the probability of the same bi-phoneme in the phonetic model proper to speaker.
3. for each phoneme $P$ in $G$
    if $\lceil nbr(P)/57 \rceil$=1 then $M=M\cup\{P\}$
```

We note by:

- nbr(P) is the repetition number of the phoneme P in the set G;
- 57 is the result of 29*2-1: all possibilities to combine an Arabic phoneme with all other phonemes of the Arabic alphabet including the phonemes itself (we have 29 consonants and vowels don't pose pronunciation problems in this case).

3.6 Replacement Phonemes Identification

This section is dedicated to identify replacement phonemes (phonemes pronounced instead of the mispronounced phonemes). The main idea is that the sum of probabilities of bi-phonemes containing a wrongly pronounced phoneme is distributed to bi-phonemes containing replacement phonemes.

For this treatment, we need two values: The standard deviation (δ) and the average (Avg) of all $\mathrm{N[i]}, 1 \leq \mathrm{i} \leq \mathrm{length(N)}$, such as $\mathrm{H[i]} = 0$ and $\mathrm{N[i]} \neq 0$:

- δ=Standard Deviation$\{\ \mathrm{N[i]}\ ,\ 1 \leq \mathrm{i} \leq \mathrm{length(N)}\ ,$ with $\mathrm{H[i]} = 0$ and $\mathrm{N[i]} \neq 0\}$.
- Avg=The Average$\{\ \mathrm{N[i]}, 1 \leq \mathrm{i} \leq \mathrm{length(N)},$ with $\mathrm{H[i]} = 0$ and $\mathrm{N[i]} \neq 0\}$.

The following pseudo-code is used to identify replacement phonemes:

```
1. We propose these two set $R$ and $B$ with $R=B=\varnothing$.
2. for $i$=1 to length($N$)
  if N[i]+(Avg-δ)< H[i] then
B=B∪{P, P'} such as N[i] equal to the probability of occurrence of bi-phoneme PP' in the reference phonetic model and H[i] equal to the probability of the same bi-phoneme (PP') in the phonetic model proper to speaker.
3. for each phoneme $P$ in $B$
  if $\lceil nbr(P)/57 \rceil$=1 then $R=R\cup\{P\}$
```

We note by:

- nbr(P) is the repetition number of the phoneme P in the set B;
- 57 is the result of 29*2-1: all possibilities to combine an Arabic phoneme with all other phonemes of the Arabic alphabet including the phonemes itself (we have 29 consonants and vowels don't pose pronunciation problems in this case).

4. Experiments and results

4.1 Test Conditions

The test is done in the following conditions:

- An Arabic corpus of six hours (healthy and pathological speech in *.wav format and mono speaker mode) has been prepared for training our acoustic model.
- To generate the reference phonetic model, a healthy Arabic speech base of eleven hours has been recorded.
- This healthy Arabic speech base is divided into five sub-corpuses, and for each one we determine its phonetic model.
- The test database was created with the help of speech therapists. It counts 100 Arabic speech sequences which 60 are pathological, 20 are healthy and 20 was been recorded by non-native speakers (French).

The following table summarizes these points:

Corpus	**Size**	**Prepared by**	**Speaker number**	**Age (years)**	**Objective**
1st Corpus	100 records	With the aid of Speech Therapist	100 Speakers	Between 13 and 49	Test
2ndCorpus	6 hours	Healthy and pathological peoples (native and non-native speakers)	6 Speakers	Between 17 and 47	Training acoustic model
3rd Corpus	11 hours	Native healthy peoples	15 speakers	Between 21 and 56	Generate the phonetic model

Table 1: The speech corpus used during this work

4.2 Experiment Results

The first step is consecrated to calculate the phonetic distance. We use for this task five phonetic models. The following table summarizes different distances between different phonetic models:

Model	M_1	M_2	M_3	M_4	M_5
M_1	0	0.1306207571312359°	0.16295310606493837°	0.11036318831226814°	0.14662384597458037°
M_2	0.1306207571312359°	0	0.1506320448535047°	0.1181644845403591°	0.1600254584725937°
M_3	0.16295310606493837°	0.1506320448535047°	0	0.09816562350665492°	0.13125479658442351°
M_4	0.11036318831226814°	0.1181644845403591°	0.09816562350665492°	0	0.11356749243562952°
M_5	0.14662384597458037°	0.1600254584725937°	0.13125479658442351°	0.11356749243562952°	0

Table 2: Phonetic distances between different phonetic models

Based on the previous table we can calculate:

S	Max	Avg.	δ
0.1306207571312359°, 0.16295310606493837°, 0.11036318831226814°, 0.1506320448535047°, 0.1181644845403591°, 0.09816562350665492°, 0.14662384597458037°, 0.1600254584725937°, 0.13125479658442351°, 0.11356749243562952°	0.16295310606493837°	0,13223708°	0,022237411°

So the phonetic distance:
β= 0.16295310606493837° +| 0.13223708°- 0.022237411°|= 0.272952775°

Table 3: The standard phonetic distance

The pathologies detection rate is summarized in this table.

Test Corpus	**Results**
60 pathological records (native speakers)	57 pathological records and 3 healthy records
20 healthy records (native speakers)	20 healthy records
20 speech sequences recorded by foreign speakers	20 pathological records

Table 4: Pathologies detection rate

The fourth table presents that three sequences from eighty pathological are falsely classified. To identify the reason of this false classification, we try to classify sequences that combine two sequences from these three falsely classified as shown in the following table:

#Sequences combination	**Classification**
Combine the 1st and the 2nd sequences	pathological
Combine the 1st and the 3rd sequences	pathological
Combine the 2nd and the 3rd sequences	pathological

Table 5: Ambiguous sequences classification

The following table summarizes the pathologies detection rate:

Speech Base	**Pathologies Detection Rate**
Healthy speeches	100%
Pathological speeches (native speakers)	95%
Speeches recorded by non-native speakers	100%
All records	97%
Combined records (after combining ambiguous records)	100%

Table 6: Pathology detection rate

In the case of pathological classification, we launch an algorithm to identify phonemes that pose mispronunciations. Following tables show phonemes posing pronunciation problems for each case:

Arabic phonemes	**Pronunciation disorders rate**
خ غ	36%
س ص	31%
ر	17%
ق ك	11%
ذ ض ظ	4%
Other	1%

Table 7: Problematic phonemes identification: case of native speakers

Arabic phonemes	**Pronunciation disorders rate**
ذ ض ظ	33%
ح	28%
ق	23%
خ ع غ	14%
Other	2%

Table 8: Problematic phonemes identification: case of non-native speakers

In the following table, we summarize results of replacement phonemes identification.

Speaker	**Mispronounced phonemes (M)**	**Replacement phonemes** (R)	**Corresponding between M and R**	
			Set M	**Set R**
Native speakers	ك س ص ر خ غ	غ ت ح ع ث	ك	ت
			ر	غ
			ص, س	ث
			غ	ع
			خ	ح
Foreign speakers	ح ذ ض ظ ع ق	ه د ء ك	ذ ض ظ	د
			ح	ه ء
			ق	ك
			ع	ء

Table 9: Replacement phonemes identification

4.3 Discussion

Results in the fifth table show the impact of the sequence size in classification procedure. Indeed, when we use a large sequence of speech we maximize the probability to have all possibilities of Arabic bi-phoneme combinations in such spoken sequence; then detection of vocal pathology becomes easy. Against, if we use a short speech sequence, highly probable we don't have all combination possibilities between Arabic phonemes in such record; so detection of vocal pathology becomes more difficult.

For native speakers, phonemes that pose pronunciation disorders are due either from languages disabilities or from difficulty to master phoneme pronunciation.

For non-native speakers, phonemes that pose pronunciation problems are often similar to other phonemes in their native languages; speakers are often hampered by phonemes in native languages.

5. Conclusions and Future Works

Acoustic and phonetic analysis presents the proper method in spoken language diagnostics to detect vocal pathologies and detection of Arabic phonemes that pose pronunciation problems. Experiment results show that the proposed approach presents high classification accuracy; indeed we attain a classification rate of 97%, 100% after combining falsely classified sequences, and 100% in identifying phonemes that pose pronunciation problems. Thanks to previous results, computer scientists can benefit from our work for applications of processing of human speech.

It may be possible to benefit from this work to elaborate an automatic speech correction system for peoples suffering from language disabilities (Terbeh, 2013; Terbeh, 2014).

Also it may be possible to benefit from this work to elaborate a system of assistance for foreign speakers to learn the Arabic spoken language (Maraoui, 2012).

6. Acknowledgements

First, I would like to express my deepest regards to all members of the evaluation committee within LREC 2016 conference. I extend also my advance thanks to my supervisor Mr. Mounir ZRIGUI for his valuable advices and encouragements.

7. Bibliographical References

Zouaghi, A., Zrigui, M., Antoniadis, G. (2008) Automatic Understanding of Spontaneous Arabic Speech-A Numerical Model. TAL, 49(1), pp. 14--166.

Majidnezhad, V., Kheidorov, I. (2013). An ANN-based Method for Detecting Vocal Fold Pathology. International Journal of Computer Applications, 62(7).

Majidnezhad, V., Kheidorov, I. (2012). A HMM-Based Method for Vocal Fold Pathology Diagnosis. International Journal of Computer Science Issues, 9 (6-2).

Kukharchik, P., Martynov, D., Kheidorov, I., Kotov, O. (2007). Vocal Fold Pathology Detection using Modified Wavelet-like Features and Support Vector Machines. European Signal Processing Conference.

Terbeh, N., Labidi, M., Zrigui M. (2013). Automatic speech correction: A step to speech recognition for people with disabilities. ICTA.

Terbeh, N., Zrigui. M. (2014). Vers la Correction Automatique de la Parole Arabe. Citala.

Basheer, I.A., Hajmeer, M. (2000). Artificial neural networks: fundamentals, computing, design, and application. Journal of Microbiological Methods, 43, pp. 3--31.

Lee, K.Y., Cha, Y.T., Park, J.H. (1992). SHORT-TERM LOAD FORECASTING USING ANARTIFICIAL NEURAL NETWORK. Transactions on Power Systems, 7(1), pp. 124--132.

Little, M., McSharry, P., Moroz, I., Roberts, S. (2006). Nonlinear, Biophysically-Informed Speech Pathology Detection. ICASSP.

Ittichaichareon, C., Suksri, S., Yingthawornsuk. T. (2012). Speech Recognition using MFCC. International Conference on Computer Graphics, Simulation and Modeling.

Archaux, C., Laanaya, H., Martin, A., Khenchaf, A. (2004). An SVM based Churn Detector in Prepaid Mobile Telephony.

Alonso, J.B., Leon, J.D., Alonso, I., Ferrer, M.A. (2001). Automatic Detection of Pathologies in the Voice by HOS Based Parameters. EURASIP Journal on Applied Signal Processing, 4, pp. 275-284.

Adnene, C., Lamia, B. (2003). Analysis of Pathological Voices by Speech Processing. Signal Processing and Its Applications, 1(1), pp. 365-367.

Jolliffe, I.T. (2008). Principal Component Analysis. Book, 2nd Edition.

Bréhilin, L., Gascuel, O. Modèles de Markov caches et apprentissage de sequences.

Patane, G., Russo, M. (2001). The enhanced LBG Algorithm. Neural networks, 14 (9).

Bürki, A, Gendrot, C, Gravier, G, Linarès, G, Fougeron, C. (2008). Alignement automatique et analyse phonétique: comparaison de différents systèmes pour l'analyse du schwa. TAL, 49(3), pp. 1--33.

Maraoui, M., Zrigui, M., Antoniadis G. (2012). Use of NLP Tools in CALL System for Arabic. Int. J. Comput. Proc. Oriental Lang. 24(2), pp. 153—166.