Multi-Level Improvement for a Transcription Generated by Automatic Speech Recognition System for Arabic

Heithem Amich, Mohamed Ben Mohamed, and Mounir Zrigui
LaTICE Laboratory, Monastir Faculty of Sciences, Tunisia

Abstract: *In this paper we will propose a novel approach to improving an automatic speech recognition system. The proposed method constructs a search space based on the relations of semantic dependence of the output of a recognition system. Then, it applies syntactic and phonetic filters so as to choose the most probable hypotheses. To achieve this objective, different techniques are deployed, such as the word2vec or the language model Recurrent Neural Networks Language Models (RNNLM) or ever the language model tagged in addition to a phonetic pruning system. The obtained results showed that the proposed approach allowed to improve the accuracy of the system especially for the recognition of mispronounced words and irrelevant words.*

Keywords: *Automatic speech recognition, multi-level improvement, language model, semantic similarity, phonetic pruning.*

Received July 12, 2016; accepted March 26, 2017

1. Introduction

The automatic recognition of speech is a pioneering task of artificial intelligence which was always had a great appeal to researchers [6]. The particular importance of speech processing is explained by the privileged status of speech as a vector of information in our human society. This research area involves different disciplines including signal processing, information theory, statistics, algorithm, linguistics, phonetics, acoustics etc.

The automatic recognition of speech is a process which allows moving from an acoustic signal of speech to the transcription of the signal in a written version. This message could then be used by version treatments. Indeed, how does a transcription system work? From a recording, the system starts by calculating a transformation of the signal in acoustic parameters adapted to a recognition engine [2]. This latter makes use of acoustic and linguistic knowledge to produce the transcription [9]. It all depends on well formalized theories like spectral analysis, information theory and dynamic programming. Although an ideal transcription system remains always non-existent, several research efforts have recently been made to come up with robust systems [1]. Automatic speech processing still has a few shortcomings. In fact, the main limitations which hinder the development of efficient systems are generally linked to the great deal of variability in speech [30]. On this respect, we remind of the intra-speaker variability, due to the elocution (singing voice, shouting, whispering, hoarse, husky, under stress), inter speaker variability (male voice, female voice, or child voice) as well as the variability caused by the signal acquisition device (type of microphone), or by the environment (noise, crass talk) [12, 30]. Moreover, the degradation of performance is generally due to the lack of precise rules to formalize knowledge to different decoding levels (including, syntax, semantics, and pragmatics). Besides, these different levels seen to be closely inter twined. Nowadays most large vocabulary transcription systems are based on statistical methods with learning techniques from oral corpora where the correct transcription is known in advance. A statistical ASR is made up of several components following the acoustic and linguistic modeling of speech signal with a view to its recognition. Many techniques have been developed to improve each component of the system so as take account of or reduce the problems related to speech variability. Never the less, each technique has certain weaknesses [18].

This leads us to develop an approach which takes account neither of the recognition modules adopted by a ASR, now its search algorithms, or its smoothing techniques, which is the strong point of this approach. As a matter of fact, we considered the Automatic Speech Recognition (ASR) as a black box devoid of any power of decision. Its role is limited to providing the transcription which will trigger our correction process. Finally, our approach in the only one responsible for correcting mis-recognized hypotheses and irrelevant word. After a brief state of the art on the technique of improving transcriptions, we describe our approach in section 3. In section 4, we present, compare, and discuss different evaluation results. In the last section, we present a few perspectives of our work.

2. State of the Art

Improving the performance of ASR caught the attention of specialists in many languages. A lot of works were carried out to improve the competency of the various components of the system such as the linguistic and acoustic models and to significantly improve the decoding quality and the transcription quality a priori. In this framework, Lecouteux [15, 16] presents a combinational method allowing to exploit a priori manual transcriptions and to integrate then directly into the heart of a SARP. This method allows to effectively guiding the recognition system with the help of auxiliary information. He also combined SRALs based on guided decoding [18]. With reference to previous research works, Salim [27] proposed a fusion system between an original sentence containing an error and sentence of clarification.

Thus, he proposed many alignments of levenshtein variants [12] and a reranker to select the best hypothesis. Antoine Laurent [14] came up with a method allowing to help the user in the step of correcting ASR outputs and to correctly transcribe proper names to facilitate the automatic indexing of transcribed reunions [22, 23].

Bongares [7] studied the methods of combining transcription systems of large vocabulary speech. His study focuses on the coupling of heterogeneous transcription systems with the aim of improving the transcription quality. Combining different transcription systems is based on the idea of exploiting the strengths of each system in order to obtain a final improved transcription [31]. In the literature, we find many works which made use clarification systems [32]. These systems may require the user's intervention to disambiguate the homophones, spell out of vocabulary word or reformulate part of their original sentence. This is done in the aim of correcting errors. The Dragon Naturally Speaking system [25] allows consulting, through an interface, the words of the transcription and correcting them with specific commands. In the same context, Hoste proposes a system which identifies incorrect words on the basis of an estimation of words previously checked by the user [13].

Merhbene *et al*. [21] and Favre *et al*. [10] propose approaches allowing to locate error segment and to detect out of vocabulary word in order to initiate a dialogue of clarification so as to improve the final transcript.

3. The Proposed Approach

In this section, we will present our system in details. The process of automatic correction of mis-spelt words from Arabic will be done in two main phases, as shown in Figure1.

The steps of the left block scheme's yellow represent the first phase. It is particularly appropriate for extending the search space for the word to correct. The second stage is it at the right scheme. This phase is responsible for selecting the most likely word.

3.1. Creation of Search Space

Having received a w_n word from ASR and given its context $w_0, w_1, \ldots, w_{n-1}$, this part in essential to develop an expanded search space including the words to be treated, later by the system bused, on the one hand, on a language model and on the semantic similarity on the other.

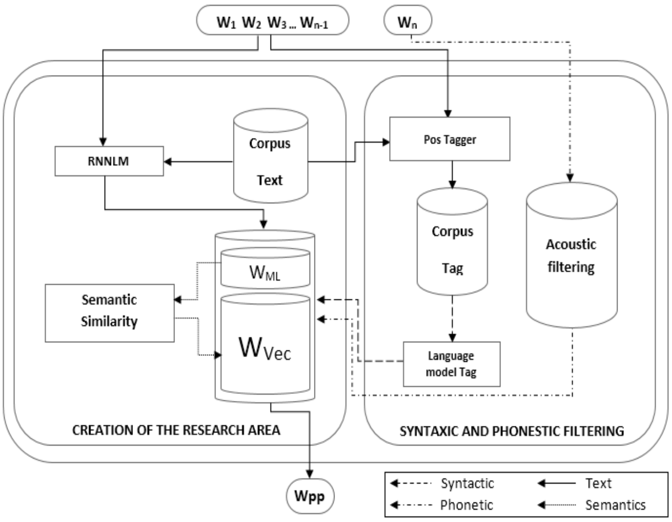

Figure 1. General architecture the transcription enhancement system (Symat).

3.1.1. Language Model RNNLM

Let $S= w_0, w_1, \ldots, w_{n-1}$ be the context at a given instant our approach aims to estimate all of the mast likely hypotheses w_n by using an RNNLM language model. This preliminary phase consists of passing the set of observations S to a language model in order to retrieve the set of the most likely words which could complete *S*. The RNNLM model is based on the association of neural networks at word level. In what follows, we briefly remind of the mathematical strategies relative to this model. Recently, deep neural networks have made a great success in the fields of image processing, acoustic modelling [8, 28], language modeling [3, 4], etc., Language models based on neural networks do better than standard back off n-gram models [5]. Words are projected into low dimensional space similar words are grouped together. RNNLM could be a deep neural network LM due to its recurrent connection between input layer and hidden layer. The network has an input layer *x*, a hidden layer S and an output layer y. We denote input to the network in time *t* as *x* (*t*) and output as y (*t*). *S* (*t*) refers to the state of the network (hidden layer). In put vector (*x*) is formed by concatenating vector w (*t*) which represents current word. Output is made from neurons in context layer S at time *t*-1 [24]. The architecture of the neuronal network used to calculate conditional probabilities is organized in three layers. The input layer reads a word *w* (t-1) and a continuous *S* (t-1). The hidden layer compresses the information of these two inputs and calculates a new representation *S* (*t*) for the input of the next propagation. The value is then passed on to the output layer, which provides the conditional probabilities P (*w* (*t*) | *w* (*t -1*), s (*t - 1*)). RNNLM can be expressed as follows

$$x\ (t) = w\ (t-1) + s\ (t-1) \qquad (1)$$

$$S_j\ (t) = f\ (\textstyle\sum_i \quad Ui(t)uij) \qquad (2)$$

$$Y_k\ (t) = g(\textstyle\sum_i \quad Sj(t)kj) \qquad (3)$$

Where *f* (*z*) is a function of sigmoid activation:

$$f(z) = \frac{1}{1+e^{-z}} \qquad (4)$$

And *g* (*z*) is a softmax function:

$$g(zm) = \frac{e^{zm}}{\sum K e^{zk}} \qquad (5)$$

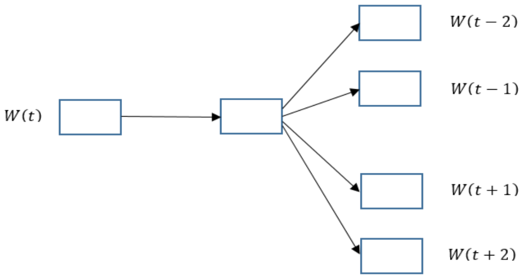

Figure 2. Frame of the drive words vector features.

A continuous skip-gram model shown in Figure 2 is used to train a high quality word vector. It tries to maximize classification of a word based on the context words in the same sentence instead of predict the next word based on the history word.

3.1.2. Semantic Similarity

Identifying the similarity between words is an important TAL task regarding the domains where this technique could be useful, such as the search for information, automatic translation or even the automatic generation of text. The ability to correctly identify the semantic similarity between words is essential for our system. This is because of its contribution to the reconstruction of research space. The search for similarity is based on the word 2vec techniques [26]. Word 2vec is a neuronal network with two layers having as an input a text corpus and as an output a set of vectors representing the characteristics of the input word in this corpus. Work is then taken to measuring the cosines similarity where an angle of 0 degree expresses a total similarity, whereas an angle of 90 degrees expresses no similarity. The following table present a list of words associated with the word «July» rising word2vec, in order of proximity.

Table 1. A list of words associated with the word "July" using word2vec.

Mot	Distance Cosinus
June	0.9557317
April	0.9386088
May	0.9324805
August	0.9314448
March	0.9097166

3.2. Selection of the Most Probable Word

Having collected a well-defined number of lexicons constituting the search space, we highlighted the techniques allowing filtering, classifying and finding the most appropriate hypothesis. We adopted two filtering methods: the syntactic filtering and the phonetic, filtering.

3.2.1. Tagged Language Model

At this stage, we have accumulated a search space containing a set of lexicon originating from two sources operating at two different levels: a syntactic level and a semantic level. Our objective is to identify the most probable final hypothesis. To this end, we applied a syntactic filter to classify the set of word W_{vec} + W_{ML} and assignment a higher probability to the hypothesis having the most likely label. We have deployed a language model based on labels. This model operates exclusively at the level of labels. The training of this model is realized on corpus of labels. This corpus is the result of syntactic analysis of the corpus used at the stage of creating the search space. The aim of this training is to guess the label while being given a history $E_1E_2..E_{N-1}$: P(E_n | $E_1E_2….E_{N-1}$).

Note that the following En is the syntactic result of the input S: $W_1…..W_{n-1}$ and the search of the most likely label En is provided by the RNNLM model. The syntactic analyzer is provided by the Stanford parse syntactic analyzer [11, 13].

3.2.2. Phonetic Comparison

Having obtained a set of word W_{vec} + W_{ML} classified by a syntactic confidence score, we introduced another filtering mechanism operating at a phonetic level. This tool compares the frequency spectrum of the word W_n coming from a ASR and the frequency spectra of the word W_{vec}+ W_{ML}. This method consists in aligning the signals of two words, then measuring the degree of similarity of two spectra. At the end of this phase, we estimate the word W_n having the most likely label and the highest degree of acoustic similarity. This example shows how to measure the similarities of signal. Whether they are correlated or not?

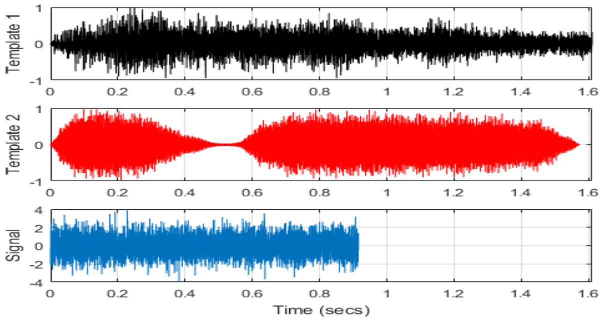

Figure 3. Comparing the similarity of two signals.

The black and red signals show the signals of two most likely words generated by search space. The third signal corresponds to the word signal generated by ASR. This figure shows that there is no phonetic similarity between the two candidates with the third signal. Just by looking at the time series, the signal seems not to correspond to one of both models. A closer look reveals that the signals did different lengths and sample rates.

4. The Case of the First Word of the Sentence

Concerning the previous steps of our approach, we recalled the different phases of the automatic correction of transcriptions provided by an automatic speech recognition system. We elaborated architecture capable of sending back the next most likely hypothesis W_n after taking the n-1 hypotheses produced by a ASR as input: it's worth mentioning that it's evident to find the words having indices between 2 and n given that there is data to manipulate. However, at the start of our procedure, we had w0 data to activate our approach so as to find the first word of the sentence. To overcome this limitation, we have partially changed our strategy. Indeed, we temporarily accepted the two most likely words generated by a ASR W_{11} and W_{12}. We remind that a speech recognition system uses these three pillars lexicon, the language model and the acoustic model to provide a text representing the transcription of a sound signal (the best one). It's also possible to retain several recognition hypotheses. The output world, then, be a list of best hypotheses N, a word graph or a confusion network. We limited ourselves to extracting the two most likely words among the retained N best hypotheses of a ASR of the first word of a sentence. This is simple due to the lack of data, which obliges us to accept W_{11} and W_{12}. However, the choice is not final. We have designed the method which reviews and verifies the first word of the sentence. The final result can accept W_{11} or rather W_{12} as well as a new lexicon retained by our approach based on a set of probabilities.

5. General Approach

In this section, we will present a detailed representation of our automatic correction system of the transcript provided from a speech recognition system. This procedure is carried out in 4 steps:

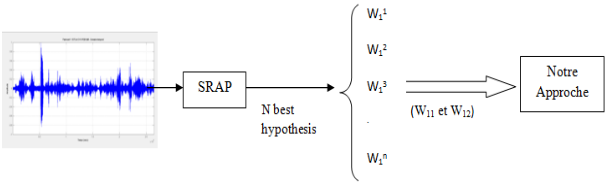

Figure 4. First phase of our approach.

- The first step consists in extracting the two best hypotheses of first word of the sentence 1 from a ASR.
- Having acquired the two hypotheses W_{11} and W_{12}, we accept W_{11}. Then, we pass W_{21} to our search approach.
- It's essential to indicate the origin of the word. That is to say, if it is the result of the language model W_2M_1 or rather the result of word2vec W_{2vec1}.
- Of the word comes from the language model, we pass W_{11} and W_{2ML1} to our approach in order to determine W_{3ML1} or W_{3vec1}. Otherwise, shift back to by using an inverse language model choose either W_{11} or W_{12} or even another word proposed by the language model. This back shift is done only when the word, retrieved by our approach, comes from the tool word 2 vec. Needless to remind that we could also define a sort of in versed language model whose words were generated in a reverse order (from right to left):

$$P_{\text{reversed}}\overrightarrow{(w)} \overset{\text{def}}{=} P(w_n) \,.\, P(w_{n-1} \mid w_n) \,.\, P(w_{n-2} \mid w_{n-1}w_n). \\ P(w_{n-3} \mid w_{n-2}w_{n-1}).\dots P(w_2 \mid w_3w_4).\, P(w_1 \mid w_2w_3). \quad (6)$$

Following each word generated by a ASR, it is susceptible to change the old word found by our approach during a back shift. The final choice is decided when we process the last word of the sentence, which can influence or substitute the previously executed hypotheses.

6. Experimentation

To construct the language model, we have used an Arabic text corpus of 100M words collected from corpus available on the used. This same corpus served to the construction of the model based on label. As for the testing of our system, we recorded a caustic corpus of 40 hours. We set up our Symat system at the exit of two known SPAP namely Sphinx [29] and HTK [20]. The following table details the results.

Table 2. Results of the system.

	Precision	Recall	F-mesure
Sphinx	51,38	56,41	53,78
Sphinx + SYMAT	56,52	62,05	59,16
HTK	46,24	50,77	48,40
HTK + SYMAT	52,72	57,88	55,18

The obtained results show that the proposed approach effectively contributed to improving ASR. We mayalso note that our method is more efficient for the HTK system than for the Sphinx system. This is justified by:

- The high clean error rate of the HTK system as compared to the sphinx system [19].
- The acoustic models trained by Sphinx were much better than that of HTK [17].

7. Conclusions

On this paper, we propose a multi layer approach with the aim of improving the transcription generated by an ASR for Arabic. This method exploits the syntactic, semantic and phonetic levels in order to evaluate the output of the ASR system and to propose the most likely word in case there’s an error. To enforce this approach, we resorted to the techniques of word similarity and to the RNNLM language model so as to establish a search space based on the history of a transcription $W_1...W_{n-1}$. After that, we carried out a phonetic and syntactic pruning to choose the most probable word. As a future work, we hope to promote our system from a model allowing taking account of the historic of applied corrections and assuring an adaptation of the correction process to a particular user.

References

[1] Aggarwal R. and Dave M., “Acoustic Modeling Problem for Automatic Speech Recognition System: Advances and Refinements Part (Part II),” *International Journal of Speech Technology*, pp. 309-320, 2011.

[2] Anusuya M. and Katti S., “Speech Recognition by Machine: A Review,” *International Journal of Computer Science and Information Security*, vol. 6, no. 3, pp. 181-205, 2009.

[3] Arisoy E., Sainath T., Kingsbury B., and Ramabhadran B., “Deep Neural Network Language Models,” *in Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, Montreal, pp. 20-28, 2012.

[4] Ben Mohamed M., Mallat S., Nahdi M., and Zrigui M., “Exploring The Potential Of Schemes In Building NLP Tools For Arabic Language,” *The International Arab Journal of Information Technology*, vol. 12, no. 6, pp. 566-573, 2015.

[5] Ben Mohamed M., Zrigui S., Anis Z., and Zrigui M., “N-Scheme Model: An Approach Towards Reducing Arabic Language Sparseness,” *in Proceedings of 5th International Conference on Information and Communication Technology and Accessibility*, Marrakech, pp. 1-5, 2015.

[6] Boehm B., “A Spiral Model of Software Development and Enhancement,” *IEEE Computers*, vol. 21, no. 5, pp. 61-72, 1988.

[7] Bougares F., Estève Y., Deléglise P., and Linarès G., “Bag Of N-Gram Driven Decoding For LVCSR System Harnessing,” *in Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, Waikoloa, pp. 278-282, 2011.

[8] Dahl G., Yu D., Deng L., and Acero A., “Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30-42, 2012.

[9] Dua M., Aggarwal R., Kadyan V., and Dua S., “Punjabi Automatic Speech Recognition Using HTK,” *International Journal of Computer Science Issues*, vol. 9, no. 1, pp. 359-364, 2012.

[10] Favre B., Rouvier M., and Béchet F., “Reranked Aligners for Interactive Transcript Correction,” *in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, pp. 146-150, 2014.

[11] Green S. and Manning C., “Better Arabic Parsing: Baselines, Evaluations, and Analysis,” *in Proceedings of the 23rd International Conference on Computational Linguistics*, Beijing, pp. 394-402, 2010.

[12] Helleseth T., Klove T., and Levenshtein V., “Error-Correction Capability of Binary Linear Codes,” *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1408-1423, 2005.

[13] Hoste L., Dumas B., and Signer B., “Speeg: A Multimodal Speech-And Gesture-Based Text Input Solution,” *in Proceedings of International Working Conference on Advanced Visual Interfaces*, Capri Island, pp. 156-163, 2012.

[14] Laurent A., Meignier S., Merlin T., and Deléglise P., “Computer-Assisted Transcription of Speech Based on Confusion Network Reordering,” *in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague, pp. 4884 4887, 2011.

[15] Lecouteux B., Nocera P., and Linarès G., “Décodageguidé Par Un Modèle Cache Sémantique ,” *Journées d ’Etude Sur la Parole*, Belgique, pp. 97-100, 2010.

[16] Lecouteux B., Linarès G., and Oger S., “Integrating Imperfect Transcripts in to Speech Recognition Systems for Building High-Quality

Corpora," *Computer Speech and Language*, vol. 26, no. 2, pp. 67-89, 2012.

[17] Ma G., Zhou W., Zheng J., and You X., "A Comparison between HTK and SPHINX on Chinese Mandarin," *in Proceedings of International Joint Conference on Artificial Intelligence*, Hainan Island, pp. 394-397, 2009.

[18] Mallat S., Ben Mohamed A., Hkiri E., Zouaghi A., and Zrigui M., "Semantic and Contextual Knowledge Representation for Lexical Disambiguation: Case of Arabic-French Query Translation," *Journal of Computing and Information Technology*, vol. 22, no. 3, pp. 191-215, 2014.

[19] Meena K., Subramaniam K., and Gomathy M., "Gender Classification In Speech Recognition Using Fuzzy Logic And Neural Network," *The International Arab Journal of Information Technology*, vol. 10, no. 5, pp. 477-485, 2013.

[20] Marin A., Kwiatkowski T., Ostendorf M., and Zettlemoyer L., "Using Syntactic and Confusion Network Structure for Out-of Vocabulary Word Detection," *in Proceedings of IEEE Spoken Language Technology Workshop*, Miami, pp. 159-164, 2012.

[21] Merhbene L., Zouaghi A., and Zrigui M., "A Semi-Supervised Method for Arabic Word Sense Disambiguation Using a Weighted Directed Graph," *in Proceedings of the 6th International Joint Conference on Natural Language Processing*, Nagoya, pp. 1027-1031, 2013.

[22] Merhbene L., Zouaghi A., and Zrigui M., "An Experimental Study for Some Supervised Lexical Disambiguation Methods of Arabic Language," *in Proceedings of 4th International Conference on Information and Communication Technology and Accessibility*, Hammamet, pp. 1-6, 2013.

[23] Mikolov T., Karafiat M., Burget L., Cernocky J., and Khudanpur S., "Recurrent Neural Network Based Language Model," *in Proceedings of INTERSPEECH*, Mukuhari, pp. 1045-1048, 2010.

[24] Prasad R., Kumar R., Ananthakrishnan S., Chen W., Hewavitharana S., Roy M., Choi F., Challenner A., Kan E., Neelakantan A., and Natarajan P., "Active Error Detection And Resolution For Speech-To-Speech Translation," *in Proceedings of International workshop on Spoken Language Translation*, Hong Kong, 2012.

[25] Pennington J., Socher R., and Manning C., "Glove: Global Vectors for Word Representation," *in Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Doha, pp. 1532-1543, 2014.

[26] Rouvier M., Favre B., and Béchet F., "Reranked Aligners for Interactive Transcript Correction," *in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, pp. 146-150, 2014.

[27] Salam M., Dzulkifli M., and Salleh S., "Malay Isolated Speech Recognition Using Neural Network: A Work In Finding Number of Hidden Nodes And Learning Parameters," *The International Arab Journal Information Technology*, vol. 8, no. 4, pp. 364-371, 2011.

[28] Satori H., Hiyassat H., Harti M., and Chenfour N., "Investigation Arabic Speech Recognition Using CMU Sphinx System," *The International Arab Journal of Information Technology*, vol. 6, no. 2, pp. 186-190, 2009.

[29] Siegler M. and Stern R., "on the Effect of Speech Rate in Large Vocabulary Speech Recognition System," *in Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, Detroit, pp. 612-615, 1995.

[30] Trigui A., Maraoui M., and Zrigui M., "Acoustic Study of the Gemination Effect in Standard Arabic Speech," *Idiopathic Polypoidal Choroidal Vasculopathy*, pp. 192-196, 2010.

[31] Trigui A., Terbeh N., Maraoui M., and Zrigui M., "Statistical Approach for Spontaneous ArabicSpeech Understanding Based on Stochastic Speech Recognition Module," *Research in Computing Science*, vol. 117, pp. 143-151, 2016.

[32] Zolnay A., Schluter R., and Ney H., "Robust Speech Recognition Using a Voiced-Unvoiced Feature," *in Proceedings of 7th International Conference on Spoken Language Processing*, vol. 2, Denver, pp. 1065-1068, 2002.

Heithem Amich received his BCs degree in computer science from the Faculty of Sciences of Monastir, Tunisia and his MSc degree from the Faculty of Mathematical, Physical and Natural Sciences of Tunis, Tunisia. He is member of LaTICE Laboratory, Monastir unit (Tunisia). His areas of interest include speech recognition system, natural language processing, machine learning.

Mohamed Ben Mohamed received his PhD from the Faculty of Economic Sciences and Management of Sfax, Tunisia. He is member of La TICE Laboratory, Monastir unit (Tunisia). His areas of interest include natural language processing, computer-assisted language learning and machine learning.

Mounir Zrigui is an associate professor at the University of Monastir, Tunisia. He received his PhD degree from the Paul Sabatier University, Toulouse, France in 1987 and his HDR in computer science from the Stendhal University, Grenoble, France in 2008. He is the head of Monastir unit of LaTICE laboratory. He has more than 25 years of experience including teaching and research in all aspects of automatic processing of natural language (written and oral).