Automatic Detection Technique for Speech Recognition based on Neural Networks Inter-Disciplinary

Mohamad A. A. Al- Rababah, Abdusamad Al-Marghilani, Akram Aref Hamarshi
Northern Border University, KSA

***Abstract*—Automatic speech recognition allows the machine to understand and process information provided orally by a human user. It consists of using matching techniques to compare a sound wave to a set of samples, usually composed of words but also of phonemes. This field uses the knowledge of several sciences: anatomy, phonetics, signal processing, linguistics, computer science, artificial intelligence and statistics. The latest acoustic modeling methods provide deep neural networks for speech recognition. In particular, recurrent neural networks (RNNs) have several characteristics that make them a model of choice for automatic speech processing. They can keep and take into account in their decisions past and future contextual information. This paper specifically studies the behavior of Long Short-Term Memory (LSTM)-based neural networks on a specific task of automatic speech processing: speech detection. LSTM model were compared to two neural models: Multi-Layer Perceptron (MLP) and Elman's Recurrent Neural Network (RNN). Tests on five speech detection tasks show the efficiency of the Long Short-Term Memory (LSTM) model.**

Keywords*—Speech recognition; automatic detection; recurrent neural network (RNN); LSTM

I. INTRODUCTION

Machine learning is a form of Artificial Intelligence (AI) that gives a machine the ability to evolve by acquiring new knowledge. Understanding speech is not an easy task for a machine. A machine, just like the human brain, must first recognize speech before understanding it. The formalism of generative grammars introduced by Noham Chomsky [1] is part of a process of theorizing language and allows a formalization that was possible to teach a machine. In the 1980s, statistical approaches very different from the initial linguistic formalism were born and quickly gained popularity because of their ease of implementation: rather than calling upon experts to formalize a given language, was trying to create a probabilistic model from a representative sample of the language to be modeled automatically. From there, and thanks to the increase of the computing power and the storage capacity of the machines, it became possible to perform many tasks of Automatic Natural Language Processing (ANLP) such as machine translation, automatic summarization, data mining, speech recognition and comprehension [2].

The scope of the ANLP on which this paper is based is the understanding of speech. It is usually done through a number of stages. The first, optional, can be to transform the initial expression of the language into a format that maximizes the performance of machine learning. For an oral message, for example, a decoder will be used in an Automatic Speech Recognition (ASR) step in order to obtain the message in textual form or in the form of a lattice of textual hypotheses. Indeed the modeling of semantic content is considered easier by working on the text than on the acoustic signal, because the latter has a very high variability.

The second step towards understanding is the learning phase: building a model that will serve as a support for understanding. This model is produced using prior knowledge. This knowledge can come from experts in the field or, for statistical models, from a certain number of data representatives of the phenomenon to be modeled, and it is often necessary to call upon humans to annotate or even transcribe these data.

The third step is the use of this model to offer an understanding to a given statement. It can use several models and cross their results to refine understanding.

In deployed systems based on automatic processing of the natural language, and more specifically with regard to speech recognition systems, another step comes into play: the adaptation of the models. Indeed, such systems must necessarily adapt to follow the evolution of habits and language of users, as well as that of the services offered. Here again, the system needs experts who regularly update the models, classically providing a new body of learning adapted to the observed evolution of users or services.

These four major phases or stages (transformation, learning, understanding and adaptation) are not necessarily sequential and can be combined to maximize their effectiveness. Generally, a model resulting from learning is used many times in an understanding stage, and the adaptation phase often only occurs when the model gives signs of weakness.

The speech recognition consists in giving meaning to an oral utterance. In this it approaches a problem of classification if one considers that one chooses a sense among N possible senses, the sense being the class allotted to the statement. The block scheme of speech recognition system is presented in Fig. 1.

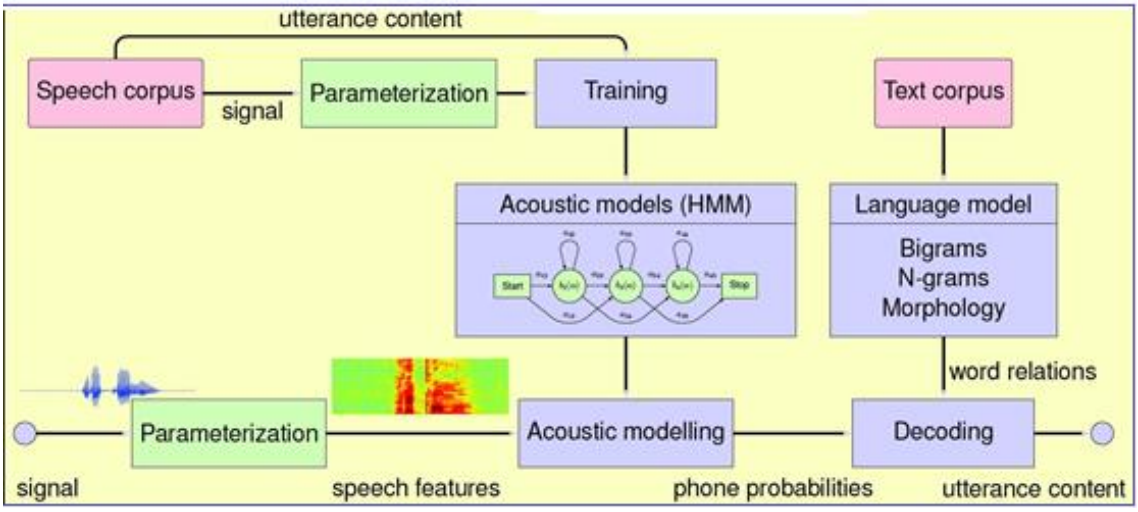

Fig. 1. Block scheme of speech recognition system.

The use of artificial neural networks for automatic speech processing is not recent. As early as the 1980s, systems using neural networks appeared to recognize vowels and then to recognize phonemes. But the results obtained at that time do not make it possible to improve the state of the art. During the next two decades some progress is made [30] but it is necessary to wait until the beginning of 2010 and the emergence of deep neural networks (DNN) with new methods of learning and specific computing resources (GPU), for that they become the state-of-the-art solution for wide vocabulary speech recognition systems [27], [28].

In recent years, a type of recurrent neural network has become the norm thanks to its excellent performance on many and varied tasks: Long Short-Term Memory (LSTM) -based neural networks.

The remainder of this paper is divided into five sections. After introducing, related works on speech recognition systems are presented in Section 2. Section 3 presents the Long Short-Term Memory (LSTM) -based neural networks. Section 4, experiments and results are detailed. Finally, this paper is concluded in Section 5.

II. Related Works

During the 1950s, speech recognition research focused on the acoustic component of speech. With the help of a tool called a spectrograph, which displays the image of speech spectra, they were able to define the main articulatory characteristics in order to be able to distinguish the different sounds of speech. Based on this visual recognition of sounds, the electrical device created in the Bell lab in 1952 could recognize the ten numbers spoken by a single speaker, comparing the acoustic parameters of the audio signal with reference models [3].

However, the success of the experiments is based on very strict conditions: reduced vocabulary, phonemes/isolated words, few speakers, recordings in laboratory conditions, etc. Acoustic methods alone are therefore insufficient for continuous speech and multi-speaker. As a result, the linguistic information begins to be taken into account in the recognition systems, to add context to the phonemes/words to be recognized and thus to improve the recognition performance [4].

In 1971, the United States Defense Advanced Research Projects Agency (DARPA) launched a five-year project to test the feasibility of automatically understanding continuous speech, which favors the creation of three new systems [5]. The "Hearsay-II" systems of CMU (Carnegie Mellon University) and "HWIM" (Hear What I Mean) of BBN (Bolt Beranek and Newman Inc.) are based on artificial intelligence: the recognition of speech is formulated as a heuristic research problem among multiple sources of knowledge.

From the 1980s researchers focused on the recognition of connected words. The biggest change of the time is defined by the transition from rule-based systems to systems based on statistical models [6]. The speech signal starts to be represented in terms of probabilities thanks to the HMM (Hidden Markov Models), which makes it possible to combine linguistic information with acoustic temporal realizations of speech sounds. This also motivates the emergence of statistical models of language, called n-grams. The innovation in acoustic signal analysis consists in the combination of the cepstral coefficients and their first and second-order temporal derivatives. These methods have been predominant in subsequent research and continue to be used even nowadays, with constant additional improvements.

From the 1990s researchers focused on minimizing recognition errors. The DARPA program continues with a keen interest in natural language. His biggest challenge is associated with the "Switchboard" corpus which focuses on spontaneous and conversational speech. The University of Cambridge has created and released a set of tools called the Hidden Markov Model Tool Kit (HTK) [7] which is one of the most widely adopted software programs for automatic speech recognition.

In the 2000s, the DARPA program focuses on the detection of sentence boundaries, noises or disfluences, obtaining abstracts or translations in a context of spontaneous speech and multi-languages. Methods to evaluate the

confidence (reliability) of the recognition hypotheses were also studied during this period [8].

Neural networks first appeared in the 1950s, but could not be used because of practical problems. They were reintroduced in the late 1980s [9], but could not provide sufficient improvement over HMM systems. It is only since 2010 that context-dependent neural networks have surpassed HMM-GMM systems [10]. This improvement is due to the use of the many hidden layers (Deep Neural Network), made possible by an efficient unsupervised pre-training algorithm [11]. In addition, the calculation architecture using graphics processors (GPU) can efficiently parallel the learning and decoding of speech [26].

Neural networks are also increasingly used in lexical modeling [12], [13]; recurring models provide significant improvements over traditional n-gram back-off models [14]. A new set of tools called Kaldi [15] makes it possible to use state-of-the-art techniques for speech recognition.

Nowadays, researchers are increasingly interested in making systems capable of meeting all types of needs: machine translation, foreign language learning, assistance for the disabled or elderly, etc. Some examples of common research concern the detection of sentence boundaries [16], speech recognition in noisy environments [17], detection of distress phrases [18], commands [19] or keywords [20], etc. Multimodal communication, which takes into account additional information on the face, the movement of the lips and/or the articulation, also begins to be taken into account [21]-[23].

III. Long Short-Term Memory (LSTM)-based Neural Networks

The interest of RNNs lies in their ability to exploit contextual information to move from an input sequence to an output sequence that is as close as possible to the target sequence. Unfortunately, for standard RNNs learning can be difficult and the context really exploited very local [29]. The problem lies in the fact that an input vector can only influence future decisions through recursive links (and thus via the repeated multiplication by the matrix V_j and the repeated application of the activation function) and therefore this influence decreases or increases exponentially as one moves forward in the sequence (Fig. 2). This phenomenon is often called a vanishing gradient problem in the literature because it impacts the retro-propagation of the gradient.

In the 1990s, many neural architectures and learning methods were tested to try to counter this phenomenon. Among these methods are heuristic optimization methods that do not use the gradient such as simulated annealing or discrete error propagation methods, the introduction of explicit delays in the recurrent architecture or the hierarchical compression of sequences. But the approach that has proven most effective and has now become the standard for dealing with sequences is the LSTM model. In fact, this model introduces multiplicative logic gates that make it possible to preserve and access relevant information over long intervals, thus reducing the impact of the evanescent gradient problem (Fig. 3). It is to this type of neural model that this paper is mainly interested and the LSTM model is described in this section.

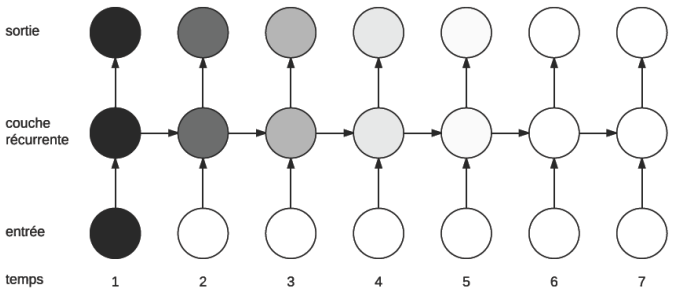

Fig. 2. Evanescent gradient problem in a standard RNN.

A layer of LSTM $L_j^{"}$ is composed of four RNN layers and which interact with each other Three of these RNN layers have the function of transfer of the logistic function (and therefore of the outputs between 0 and 1) and act as logical gates controlling:

-The amount of information that enters the LSTM cells of the layer *j*,

-The amount of information that is retained by the internal state of the cells from one step to the other,

-The amount of information coming out of the LSTM cells of the layer *j*.

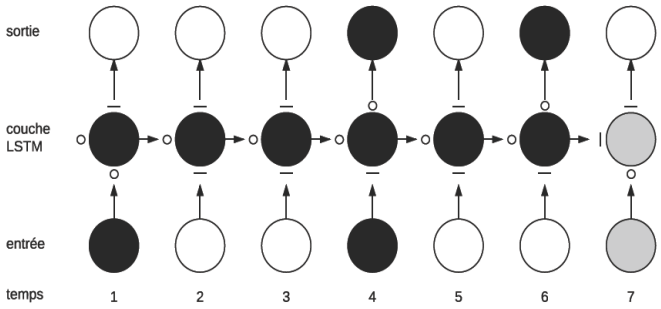

Fig. 3. Preservation of information (and gradient) in an LSTM layer.

The last of the four layers operates as a standard RNN layer and feeds the internal state of the LSTM cells via the gateways. Fig. 4 shows a synthetic description of information flows. Finally, to model the returns between the internal state of the cells and the 3 logic gates introduced by F. Three parameter vectors $ui, uf, uo \in R^{n_j}$ called "peepholes" were added. Fig. 5 shows the detailed operation of an LSTM layer.

As in the case of the standard RNN, the BPTT technique is used to compute the partial derivatives of the cost function *C*(z, c) that is minimized with respect to the different parameters of the LSTM layers. To do this, we go through the sequence by going back in time: *t*: $t_f \rightarrow 1$ and for each time step the gradients were determined with respect to the parameters. Then, global gradients are obtained by summing the contributions of each of the time steps. Fig. 6 describes the retro-propagation of the gradient in the layer $L_j^{"}$ at the time step *t* in the form of a computational graph.

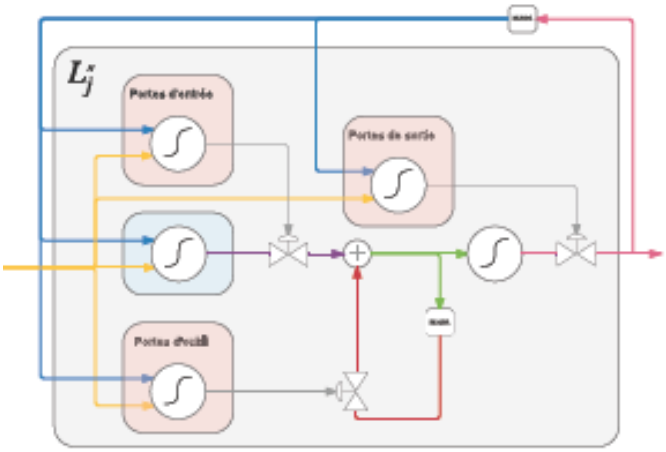

Fig. 4. Synthetic visualization of the propagation of information during the forward pass in an LSTM layer.

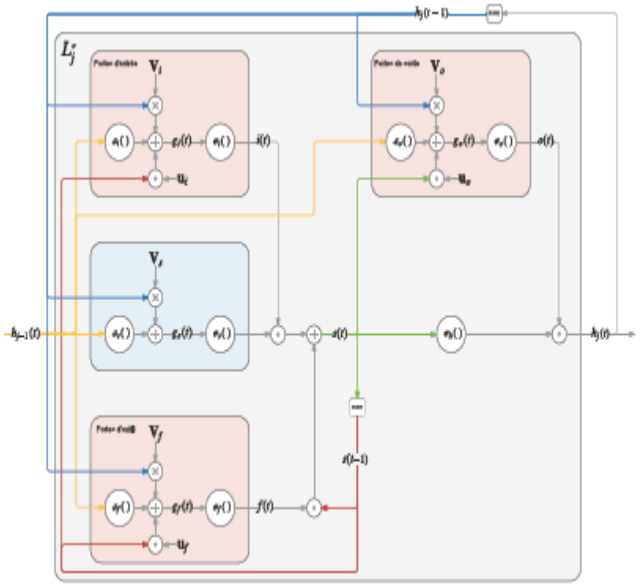

Fig. 5. Visualization of the propagation of the information during the forward pass in a layer of an LSTM.

In order to calculate the partial derivatives of C at the time step t with respect to the parameters of the layer $L_j^{"}$, the first step is to retro-propagate the gradient in the whole of the layer $L_j^{"}$, starting at the exit gates.

IV. Experiments and Results

During this paper, speech detection tasks are varied in terms of difficulties, languages and acoustic environments. We have worked on pure detection tasks (that is, the goal of minimizing the number of badly signal windows classified speech/not speech) and speech detection tasks for speech recognition, that is to say, so as to minimize the Word Error Rate (WER) of an automatic speech recognition system used downstream of the speech detection system. Very different types of data were used, such as telephone conversations, working meeting recordings, and television series audio tapes.

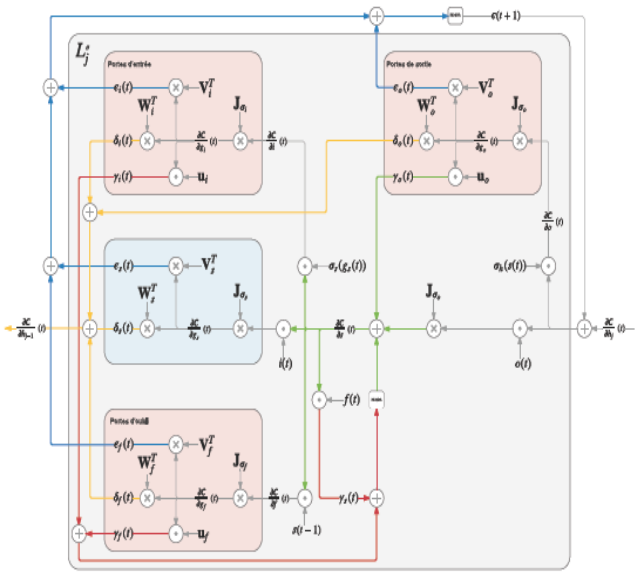

Fig. 6. Visualization of the propagation of the information during the back-pass in a layer of an LSTM.

In 2012, IARPA launched the Babel program with the aim of developing automatic speech recognition technologies that can be quickly applied to any spoken language. For all experiments, state-of-the-art automatic speech recognition systems were used. These systems are similar to those described in [24]. They use Multi-Layer Perceptron (MLP) as acoustic models and language models based on 4-gram models. They also respect the NIST constraint for system evaluation in the Babel program which states that no data other than that provided for the target language of the Babel program may be used. The official metric for the assessments performed is the Word Error Rate (WER) but the Frame Error Rate (FER) is also used in preliminary experiments.

The NIST regularly organizes open and international assessments of the different tasks of automatic speech processing. In 2015, The NIST organized the OpenSAD'15 evaluation to provide a framework for developers of speech detection systems to independently measure the performance of their systems on particular difficult audio data. Indeed, the audio signals collected for this evaluation come from the RATS program of DARPA which was mainly interested in highly distorted and/or noisy channels. These different channels are of type HF, VHF and UHF and are 7 in number (called A, B, C, E, F, G and H) with the particularity that two of these channels (A and C) are voluntarily absent from the learning and validation stages but present in the test stage to evaluate the generalization capacity of speech detection systems.

The official metric of the evaluation was the DCF defined as follows:

$$DCF = 0.75 \times P_{miss} + 0.25 \times P_{fa}$$

Where:

$$P_{miss} = \frac{\text{speech duration labeled as non} - \text{speech}}{\text{total speech time}}$$

And

$$P_{fa} = \frac{\text{non} - \text{speech duration labeled as speech}}{\text{total duration of non} - \text{speech}}$$

Three experiments were conducted on data collected in acoustic environments very different from the telephone conversations of the OpenSAD'15 evaluation and the Babel program. For these three experiments the ultimate goal was to segment into speakers. Segmentation into speakers consists of segmenting an audio signal into homogeneous speech turns, that is to say containing only one speaker. The metric of choice for this task is the Diarization Error Rate (DER) which is broken down into two parts: the FER, to which is added an error term corresponding to the confusions between speakers. Therefore, to optimize a speaker segmentation system it is preferable to minimize the FER of the speech detection system.

We worked on a task of segmentation in speakers in the audio streams of television programs collected for the LNE Audiovisual Emissions Recognition evaluation campaign (REPERE).

We also worked on the data of the AMI project which was a multidisciplinary consortium of 15 members whose mission was the research and development of technologies to improve interactions within a working group.

Finally, we worked on audio data from television series. The TVD corpus were used [25] and focused on the first season of the Game of Thrones (GoT) television series, which offers a variety of acoustic environments (indoor, outdoor, and battle scenes). This corpus is thus composed of ten episodes of 55 minutes approximately with the annotations of turns of speech by speaker.

Here, an overview of the performance gain brought by the LSTM model is presented and compared to other neural models on the speech detection task. The three neuronal models tested are: Multi-Layer Perceptron (MLP), Elman's Recurrent Neural Network (RNN) and LSTM. In order to obtain a fair comparison of the modeling capacity of each of the artificial neural networks (ANN) tested, all the ANNs used are sized to have the same number of parameters: 6000. The different ANNs are compared on five speech detection tasks: minimization of the FER on the data REPERE, AMI, Game of Thrones and the Vietnamese corpus of the Babel program; and minimizing the DCF on the OpenSAD'15 evaluation data. The three ANNs were optimized for these tasks and the results obtained on the different corpus are detailed in Table I. It is important to note that for these tests the decision smoothing module is disabled in order to get a better insight into the raw capabilities of the different neural models.

The RNNs are particularly adapted to speech detection tasks because, by allowing to exploit the temporal context freely, these models are able to improve very significantly the decisions taken locally at each time step. Thus, there is a decrease in error rates of up to 42% relative between the MLP and the simplest of the recurrent models on the data of the AMI corpus.

The most successful recurring model is the LSTM model. In fact, this model, when compared to the standard RNN model, makes it possible to improve the error rates by 17% relative on the OpenSAD'15 and REPERE corpus, by 14% relative on the Game of Thrones corpus and allows obtaining an equivalent performance on the AMI and Babel corpus.

TABLE I. PERFORMANCE OF DIFFERENT SPEECH DETECTION SYSTEMS ON BABEL, REPERE, AMI, GAME OF THRONES AND OPENSAD'15 CORPUS DATA

Type of ANN	FER				DCF
	Babel	REPERE	AMI	Game of Thrones	OpenSAD'15
MLP	9.2	17.1	11.4	17.9	5.3
RNN Standard	6.4	16.4	6.6	12.7	4.1
LSTM RNN	5.9	13.5	6.2	11	3.4

V. CONCLUSIONS

In this paper a particular type of RNN called LSTM were studied and their use for an automatic speech processing task: speech detection. Comparisons with other neural models were presented on five speech detection tasks.

All tests show that the LSTM model is more efficient than Elman MLP and RNN neuron networks. With this model, the proposed method were ranked third in the NIST OpenSAD'15 evaluation campaign with a level of performance very close to the second ranked system while having ten to one hundred times fewer parameters. Future works include the use of LSTM for another task of automatic speech processing such as spoken language identification: separate one language from all others.

REFERENCES

[1] Chomsky, Noham. Syntactic Structures. Mouton & Co, 1957

[2] Mekyska J., Beitia B., Barroso N., Estanga A., Tainta M., Ecay-Torres M. Advances on automatic speech analysis for early detection of Alzheimer Disease: a non-literal multi-task approach. Curr. Alzheimer Res;15(2):139–148, 2018.

[3] DAVIS, K. H., R. BIDDULPH et S. BALASHEK. "Automatic Recognition of Spoken Digits". The Journal of the Acoustical Society of America 24.6, p. 637–642, 1952.

[4] REDDY, D. R. "Approach to computer speech recognition by direct analysis of the speech wave". The Journal of the Acoustical Society of America 40.5, 1966.

[5] MEDRESS, M. "Speech Understanding Systems: Report of a Steering Committee". In : SIGART Newsletter 62, p. 4–8, 1976.

[6] ARORA, S. J. and R. P. SINGH. "Automatic Speech Recognition: A Review". International Journal of Computer Applications 60.9, p. 34–44, 2012.

[7] YOUNG, S. J., D. KERSHAW, J. ODELL, D. OLLASON, V. VALTCHEV and P. WOODLAND. The HTK Book Version 3.4. Cambridge University Press, 2006.

[8] JIANG, H. "Confidence measures for speech recognition: A survey".Speech Communication 45.4, p. 455–470, 2005.

[9] KATAGIRI, S. Pattern Recognition in Speech and Language Processing. CRC Press 2003.

[10] DAHL, G. E., D. YU, L. DENG and A. ACERO. "Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition". IEEE Transactions on Audio, Speech and Language Processing, 2012.

[11] HINTON, G. E. and S. OSINDERO. “A fast learning algorithm for deep belief nets”. In : Neural Computation 18.7, p. 1527–1554, 2006.

[12] SCHWENK, H. “Continuous space language models”. Computer Speech & Language 21.3, p. 492–518, 2007.

[13] MIKOLOV, T., M. KARAFIÁT, L. BURGET, J. CERNOCK`Y and S. KHUDANPUR. “Recurrent neural network based language model.” In Proc of Interspeech, p. 1045–1048, 2010.

[14] MIKOLOV, T., S. KOMBRINK, L. BURGET, J. CERNOCKÝ and S. KHUDANPUR. “Extensions of recurrent neural network language model”. In Procof the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),p. 5528–5531, 2011.

[15] POVEY, D. et al. “The Kaldi Speech Recognition Toolkit”. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, 2011.

[16] READ, J., R. DRIDAN, S. OEPEN and J. L. SOLBERG. “Sentence Boundary Detection: A Long Solved Problem?”, In Proc of COLING, p. 985–994, 2012.

[17] TU, C. and C. JUANG. “Recurrent type-2 fuzzy neural network using Haar wavelet energy and entropy features for speech detection in noisy environments”. Expert systems with applications 39.3, p. 2479–2488, 2012.

[18] AMAN, F., M. VACHER, S. ROSSATO and F. PORTET. “Speech Recognition of Aged Voices in the AAL Context: Detection of Distress Sentences”. In Proc of The 7th International Conference on Speech Technology and Human-Computer Dialogue , p. 177–184, 2013.

[19] VACHER, M., B. LECOUTEUX and F. PORTET. “Multichannel Automatic Recognition of Voice Command in a Multi-Room Smart Home : an Experiment involving Seniors and Users with Visual Impairment”. In Proc of Interspeech, p. 1008– 1012, 2014.

[20] JOTHILAKSHMI, S. “Spoken keyword detection using autoassociative neural networks”. International Journal of Speech Technology 17.1, p. 83–89, 2014.

[21] NGIAM, J., A. KHOSLA, M. KIM, J. NAM, H. LEE and A. Y. NG. “Multimodal deep learning”. In Proc of the 28th International Conference on Machine Learning (ICML), p. 689–696, 2011.

[22] GALATAS, G., G. POTAMIANOS and F. MAKEDON. “Audio-visual speech recognition incorporating facial depth information captured by the Kinect”. In Proc of the 20th European Signal Processing Conference (EUSIPCO), p. 2714–2717, 2012.

[23] REGENBOGEN, C., D. A. SCHNEIDER, R. E. GUR, F. SCHNEIDER, U. HABEL et T. KELLERMANN. “Multimodal human communication - Targeting facial expressions, speech content and prosody”. In : NeuroImage 60.4, p. 2346–2356, 2012.

[24] V.-B. Le, L. Lamel, A. Messaoudi, W. Hartmann, J.-L. Gauvain, C. Woehrling, J. Despres, and A. Roy, “Developing STT and KWS systems using limited language resources”, Interspeech, 2014. 27

[25] A. Roy, C. Guinaudeau, H. Bredin, and C. Barras, “Tvd : A reproducible and multiply aligned tv series dataset.” , LREC, 30, pp. 418–425, 2014.

[26] A.L. Maas, P. Qi, Z. Xie, A. Y. Hannun, C.T. Lengerich, D. Jurafsky, A.Y. Ng. "Building DNN Acoustic Models for Large Vocabulary Speech Recognition." arXiv preprint arXiv:1406.7806, 2014 Jun 30.

[27] Jun Ren, Mingzhe Liu. An Automatic Dysarthric Speech Recognition Approach using Deep Neural Networks. International Journal of Advanced Computer Science and Applications (IJACSA) Vol. 8, No. 12, 2017 .

[28] Choudhary, A. and Kshirsagar, R. Process Speech Recognition System Using Artificial Intelligence Technique. International Journal of Soft Computing and Engineering (IJSCE), 2, 2012.

[29] Zhiyan, Han, and Wang Jian. “Research on speech endpoint detection under low signal-to-noise ratios.” Control and Decision Conference (CCDC), 2015 27th Chinese. IEEE, 2015.

[30] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification : labelling unsegmented sequence data with recurrent neural networks,” in Proc of the 23rd international conference on Machine learning, pp: 369-376, ACM, 2006.