International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:8, No:9, 2014

Recognition of Noisy Words Using the Time Delay Neural Networks Approach

Khenfer-Koummich Fatima, Mesbahi Larbi, Hendel Fatiha

***Abstract*—**This paper presents a recognition system for isolated words like robot commands. It's carried out by Time Delay Neural Networks; TDNN. To teleoperate a robot for specific tasks as turn, close, etc… In industrial environment and taking into account the noise coming from the machine. The choice of TDNN is based on its generalization in terms of accuracy, in more it acts as a filter that allows the passage of certain desirable frequency characteristics of speech; the goal is to determine the parameters of this filter for making an adaptable system to the variability of speech signal and to noise especially, for this the back propagation technique was used in learning phase. The approach was applied on commands pronounced in two languages separately: The French and Arabic. The results for two test bases of 300 spoken words for each one are 87%, 97.6% in neutral environment and 77.67%, 92.67% when the white Gaussian noisy was added with a SNR of 35 dB.

***Keywords*—**Neural networks, Noise, Speech Recognition.

I. Introduction

THE human operator can communicate with the robot through conventional control inputs, such as a mouse [1]. A keyboard [2]. Or others ways such gloves [3]. And muscular activity sensors [4]. However, verbal communication is the primary and most natural way of human communication [5]. Nowadays it is possible to interact with robots via speech. That necessity to make an interface, to realize it, the speech recognition system will comprise two parts; an encoder and a decoder. Encoder analyses the signal to extract a number of relevant parameters. Decoder uses these parameters to reconstruct a synthetic speech signal. Given very important development that can make it in security systems, human-machine interfaces. But this technology could yet hide variability problems as inter-and intra -speaker inter- language and noise that still to be resolved in this area of research.

Several studies have focused on different ways of speech signal representation, comparative studies between different techniques of parameterization show that coding by Mel Frequency Cepstral Coefficient (MFCC) of 39 coefficients gives better representation of speech signal in Article [6]. Especially when this signal contains noise. Others were interested in decoding problem to find the most appropriate language model for an effective speech recognition system, there were researchers in charge of a growing number of studies and whose mission is to decode the information carried by the speech signal; it can use either Markov approaches [7]. And noisy words have been solved by the use of filters before coding [8]. Either neuronal approaches such as TDNN specially used to treat temporal forms, and in particular include the relationship of the components in its architecture; time characteristic. This approach has been applied to the classification of three phonemes (B / D / G) in [9]. And it has been tested to recognize noisy words from learning neutral words in [10].Comparative studies have also shown that TDNNs are more efficient for phonemes identification and continuous speech recognition like Japanese language [11]. And it also adapts to the speaker recognition [12].

Our studies focus in this work on the achievement of a bilingual system of the speech recognition (French-Arabic), to execute tasks of robot; for example turn: Dawarane in Arab and tourne in French …. We took as an example, the two languages for their differences in voicing of consonants and vowels content. For this purpose taking into account the temporal aspect is very important, the choice of TDNNs can find these needs, given its characteristics and learning to adapt to variability of the speech signal. An interesting faculties of TDNN is the adaptation to noise, it will act as a filter with its coefficients also called shared weight will be determined in learning phase.

II. Isolated Word Recognition

The pronunciation of a word can give a multitude of potential waveforms, which makes its structure complex and variable in time. This signal can be pseudo periodic as voiced sounds or random as fricatives, or pulse as plosives. And good system should be able to detect these types of patterns and correlations sufficient to return a representation of word.

Currently, no system is able to solve the problem in the context of large vocabularies with speaker independence, continuity of the speech signal, co articulation, variability (inter-and intra-speaker inter-language, noise).

Many techniques have been proposed to increase the robustness of systems, in particular as regards their resistance to noises. These techniques can be classified into three main approaches are summarized in the following paragraphs:

1) Noisy signal preprocessing to reduce the influence of noise.
2) Conversion of a system to enable it to process the noisy speech.
3) Parameterization of speech using robust methods.

This article describe the second approach to recognize isolated words in a noisy environment using the TDNN system

Khenfer-Koummich Fatima and Hendel Fatiha are with the Electronic Department, USTO-MB, Oran, 31000, Algeria (e-mail: Fatimakhanfar@yahoo.fr, fa_hendel@yahoo.fr).

Mesbahi Larbi is now with the IFSTTAR, Lille, France (e-mail: mesbahi_99@yahoo.com).

in order to perform tasks of robot in an industrial environment, where there will be a lot of noise coming from the machine. A Gaussian white noise is added to each neutral word, that its signal is perfectly stationary because it is generated from a Gaussian process.

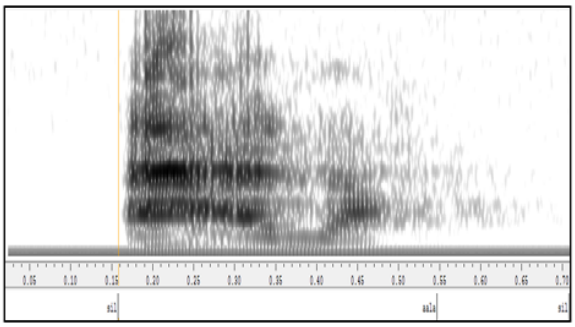

Fig. 1 (a) Spectrogram of neutral word "Aala"

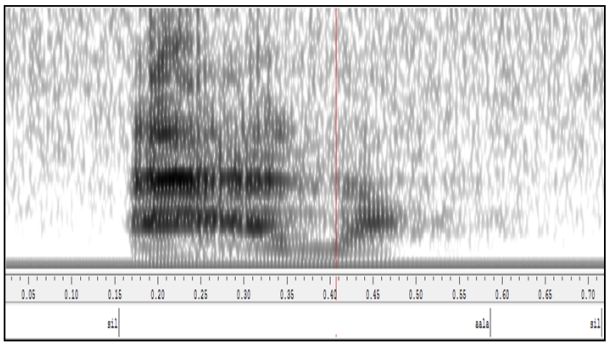

Fig. 1 (b) Spectrogram of noisy word "Aala" at 35dB

An analysis of spectrograms for the word Aala: up in English and haut in French is shown in Fig. 1 corresponds to a no-noisy signal at first (see Fig. 1 (a)) and a noisy signal at 35 decibels of SNR: Signal to Noise Ratio (see Fig. 1 (b)). The temporal evolution of the signal shows an increase in the strength of the noise as the SNR decreases, the speech signal will completely disappear. It could however still be visible in the lower part of the spectrogram for the sole reason that the Gaussian white noise is less aggressive in low than high frequencies.

An acoustic characteristic event must be recognized whatever its position in the word which setting of Event Filters or features detectors, invariant by translation temporal.

To realize this filter, we constraint a set of cells that share the same weight (filter coefficients); the different cells are connected to different parts of the signal, and thus detect the event characterized by the filter in different portions of the signal. Among the approaches of acoustic discrimination, there are neural networks, they are excellent tools for discrimination; they are able to extract relevant information of neutral words from noisy words, and making this system robust to noise, this is due to delays that are on connections. Moreover, the recognition of a short frame of 25ms, for example generally requires knowledge of neighboring frames, the effect of context is often determinant. Thereby determining elementary acoustic events, then detect these events in the signal, so word is formed by a specific sequence of events.

The TDNNs are variants of neural networks, thus effectively used to take into account the time dimension of the speech signal. The coefficients are learned using a modification of gradient back propagation algorithm, filter coefficients resulting from the learning will be optimized on the basis examples.

III. Description of TDNNS

The TDNN are constituted as the multilayer perceptions; MLP an input layer, a hidden layer or several hidden layers and an output layer, but they differ from the organization of the inter-layer connections in that it takes into account some notion of time, it will perform a scan time, so the input layer of the TDNN will take a spectrum window (footprint) signal which allows to recognize this signal.

Learning of the network is based on the following three properties:

1) Time window concept implies that each neuron in the current layer is connected to a subset of neurons in the preceding layer, the size of this window is the same between both layers. This time window allows each neuron having a local vision signal; it can be considered as a unit of detecting a local characteristic in this signal, Fig. 2.
2) Shared weight can reduce the number parameters of the neural network and thus induce a greater capacity generalization, for a given characteristic, therewith associated window will have the same weight in the temporal direction. Furthermore, this constraint results a capacity to extract the discrimination as the scanning signal. This concept shared weight also sounds like the intelligence human where more neurons compute the same function on different inputs.
3) Units in time; the delay units are basic units of TDNN model that have links with delays, a spatiotemporal summation is performed at each neuron.

A. Learning Phase

Filter settings based on TDNN will be determined in the learning phase, which is based on the optimization of neural networks for this; many learning techniques exist. For training network TDNN type, back propagation technique is adapted to this model, it consists to present the inputs of MFCCs vectors to the network and change its weighting so that we find the corresponding output. This algorithm includes in first phase to propagate forward inputs until a calculated output by (1). The second step, it compares between calculated output and known output, the weights are modified and error computed between is then minimized at the next iteration and it then back-propagates error to the front while changing the weighting layer. This process is repeated on all the examples until it gets an error output that is considered negligible [13].

$$y_j(t) = f\left(\sum_{i=1}^{N}\sum_{d=0}^{D} W_{dij}\, X_i(t-d)\right) \quad (1)$$

where: *Yj (t)* represents an obtained output of j-th neuron in upper layer by the network. *Xi (t)* is a state of i-th neuron in lower layer. *Wdij is* connection weight from i-th neuron in lower layer and j-th neuron in upper layer; these weights represent the filter coefficients.

B. Recognition Phase

In this phase it is sufficient to propagate forward the feature vectors representing the word to be recognized in each learned network until obtained outputs. After computing the cumulative of these obtained outputs (yj), finally the network which has the maximum accumulated over the duration of pronunciation represents a recognized word (Wr); this procedure is computed by (2) and viewed in Fig. 2.

$$w_r = \arg\max c_{Nk} = \arg\max\left(\sum_j y_j(t)\right)_{Nk} \quad (2)$$

C_{Nk} represents the Accumulation of k-th network, in our case; there are 10 networks or 10 words for each language.

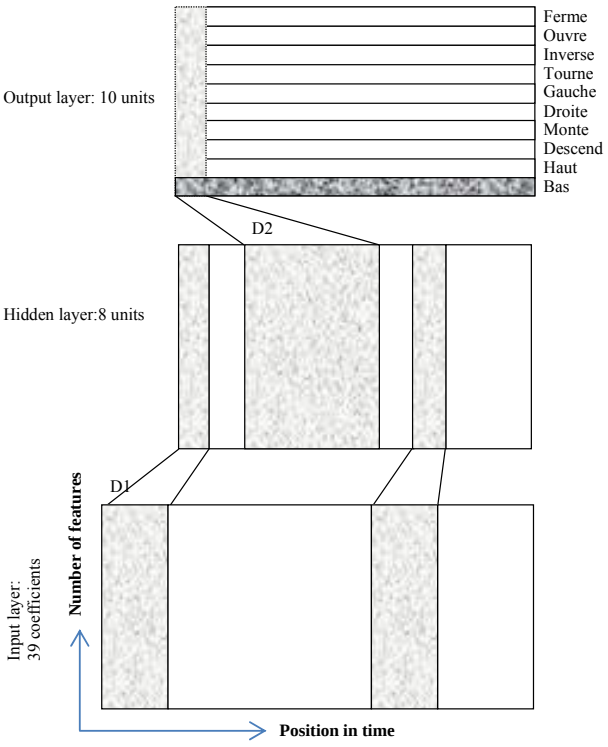

Fig. 2 Procedure of the word recognition ‘’Bas”

IV. Implementation

Our application was implemented with MATLAB Neural Network Toolbox and the developed network is the TDNN type with only 3 layers, an input layer, a hidden layer and an output layer. For each layer; except for output layer, it is applied an analysis window, the size of this window is determined through several tested values on the training corpus. For input layer, an analysis window of two frames can give best results and each frame represents a MFCC vector of 39 coefficients, and therefore, the input layer consists of 78 neurons. Similarly, another analysis window of 4 frames, each frame consists of 8 units, where 24 neurons in the hidden layer. The last layer is the output layer, representing the words of the corpus which are 10 words; we shall have therefore 10 units activations receiving from the hidden layer, see Fig. 2.

In this architecture, the number of units is related to number of coefficients for each MFCC (39 units) vector in input layer and output layer to number of words to be recognized, there are 10 units in our case, but the number of units in hidden layer, there is not any assumptions, although it is a major factor influencing to behavior of learning and capacity of generalization and the choice is based on several tests: 8 units give a best rate of generalization.

Our database is constituted by ten (10) words (commands) pronounced separately in Arabic / French. Each word was pronounced thrice by 10 speakers of different age and sex (5 men and 5 women), there will be 300 pronunciations for each language, and these words are used for learning. Another basis is consisted by 300 words was utilized for testing.

To treat only the essential information, and reduce the number of data parameters, the silence that contains the word is eliminated by the Wave surfer software downloadable to address in [14].

The analysis phase of speech signal is used to extract the feature vectors; these are called MFCC "Mel Frequency Cestrum Coefficient", which are sent to the input layer in the network architecture. In order to make the best recognition system, with including noise, it is often added the speed Δ (primary derivatives) and Δ2 acceleration (the second derivative of the vector). Finally the number of coefficients is 39 for each vector MFCCs. It should be noted that the extraction of these features is performed by COPY tool of HTK [15].

Learning is a very important step towards satisfactory recognition results, this step needs time for learning, and it requires multiple attempts to adjust its parameters in order to improve results, because the adaptive learning is supervised, we must in first introduce the feature vectors (MFCC vectors) representing the words to be learned in the network. The learning is based on the technique of gradient descent. Each word is learned alone, the system requires 10 TDNN networks corresponding to 10 words for each language.

The principle of recognition is to pass the derived MFCCs vectors from the word to be recognized in obtained TDNN network in the learning phase, the accumulation of activity in the output layer determines the recognized word which represents the highest accumulation.

If the word is recognized by the network, it sends a command via the local area network (LAN) to the server to perform the tasks of mentor robot 5 axes; these commands are shown in Table I.

V. Discussion of Results

A. Recognition of Neutral Words

Our system is evaluated with the 30 examples basis for each word. Learning words in both languages are fully recognized; training rate equals 100 % as shown in Table I. The depicted test results in Table II show that the Arabic words are best recognized, these results can explain that Arabic contains more vowels that are easily recognizable, in addition the presence of diphthongs (double vowels) in Arabic as the case of the English language in forces the occupation of vowels in the word, as the case; "Aala" and "Dawarane", for the test recognition rate can reach up to 97.67 %, there are six words that are well recognized, learning certain classes over others; no-uniform learning. The presence of the complex consonants as occlusive / fricative or interdentals enriches up the phonetic space and also allows a good discrimination. After several tests with different values of delay ; D1 and D2, the best test rate obtained from French words is equal to 87% at D1 = 3 and D2 = 5 , this can be explained by the fact that the French word confusion is mainly due to the confusion in the consonants space. It should be noted that there are just two words; "Haut" and "Bas" that are incorrectly recognized, because the delay in input layer is in adapted for short words, in this case; it must choose the best delay for each word.

TABLE I
Rate Training for Neutral Words

Order number	Commands in French	Training Rate (%)	Commands in Arabic	Training Rate (%)	Description of commands
1	Bas	100	Assfel	100	the clamp moves to the down position
2	Haut	100	Aala	100	the clamp moves to the upper position
3	Descend	100	Ihbit	100	The arm goes down by one step
4	Monte	100	Issaad	100	The arm goes upper by one step
5	Droite	100	Yamine	100	The arm rotates from left to right
6	Gauche	100	Yassar	100	The arm rotates from right to left
7	Tourne	100	Dawarane	100	The clamp rotates
8	Inverse	100	Iklib	100	The clamp rotates in reverse
9	Ouvre	100	Ifteh	100	Opening of the clamp
10	Ferme	100	Aghlik	100	Closing of the clamp
	Rate	100	Rate	100	

TABLE II
Rate Test for Neutral Words

		Test rate for French words (%)				Test rate for Arabic words (%)			
D1-D2		2-4	3-5	4-6	10	2-4	3-5	4-6	10
Commands	1	50	63.33	60	56.67	90	93.33	93.33	96.67
	2	53.33	50	63.33	66.67	100	96.67	96.67	100
	3	100	100	96.67	100	100	100	83.33	86.67
	4	93.33	96.67	96.67	93.33	93.33	100	90	96.67
	5	100	96.67	96.67	100	96.67	96.67	96.67	93.33
	6	96.67	93.33	90	90	96.67	96.67	96.67	96.67
	7	86.67	100	90	83.33	100	93.33	100	100
	8	100	93.33	83.33	93.33	100	100	93.33	100
	9	83.33	93.33	86.67	73.33	100	100	100	93.33
	10	86.67	83.33	70	80	100	86.67	80	90
Total rate		85	**87.00**	83.33	83.67	**97.67**	96.33	93	95.33

B. Recognition of Noisy Words

Gaussian white noise with different levels of signal noise ratio: SNR is added to the learning and testing words of both languages; why the voice control systems are usually developed in a noisy environment, so it must also take into account the work environment. Fig. 3 represents the evolution of recognition rate according to signal noise ratio for the test basis words, which shows that the system based on TDNN begins to recognize noisy words from a threshold equals to 30 dB for the Arabic words and 40 dB for the French words at 35 dB; Rate test in Arabic words = 92.67% and Rate test in French words = 77.67%, so from that threshold the representatives frequents of the word will be retrieved by a filter based on the TDNN. The difference in threshold indicates that the noise affects more the portion of the consonants than vowels, because French is more consonant than Arabic in the articulation, and the obtained results in Fig. 4 confirm that, because before producing the noise, these words in the both languages were fully recognized; the rate was 100%.

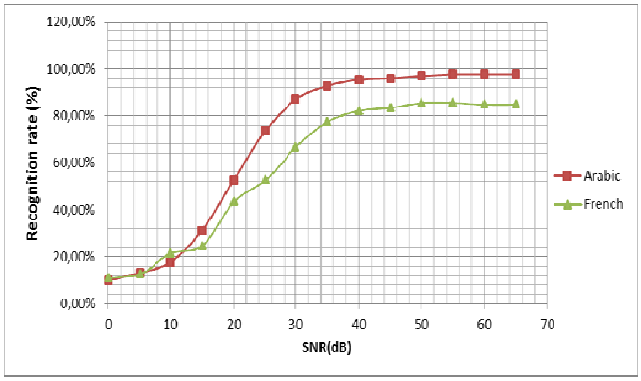

Fig. 3 Evolution of recognition rate according to signal noise ratio (SNR) for isolated words (test basis)

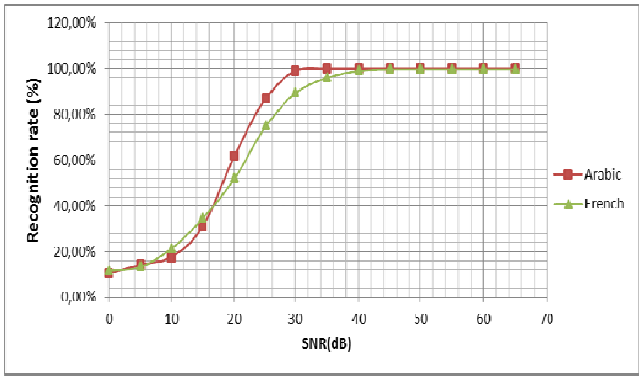

Fig. 4 Evolution of recognition rate according to signal noise ratio (SNR) for isolated words (training basis)

VI. Conclusion

In this work, recognition system is presented using neural network with time delay; TDNN and it is applied to isolated words recognition such robot commands pronounced in two different languages Arabic and French. The results show that recognition rate is good; this is why the TDNN behaves as a good filter to detect desirable frequencies whose parameters are determined in the learning phase. Eventually make the system adaptable to the variability of the speech signal, it can also show the difference between two languages, for example the recognition rate of Arabic is good compared to the French language especially considering the noise, it said that the articulation of the Arabic language is more vowelized than French. The choice of TDNNs approach works better in this type of word recognition and marks the difference in pronunciation of words in both languages. Our future work will focus on the introduction of new methods, increasing the size of the vocabulary, and to ensure the safety of the task to execute it, it will study the parameters that achieve the best results in the presence of noise, because the voice control systems are often developed in industrial environment.

References

[1] Barkana, D.E., Das J., Wang F., Groomes T. E., Sarkar N. "Incorporating verbal feedback into a robot-assisted rehabilitation system". Robotica, 2010.
[2] Courreges F., Edlie A., Poisson G., Vieyres P." Ergonomic mousse based interface for 3d orientation control of a tele-sonography robot". In Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2009), St. Louis, USA, PP.61-66, 2009.
[3] Drygajlo A., Prodanov P. J., Ramel G., Meisser M., Siegwart R." On developing a voice-enabled interface for interactive tour-guide robots". Advanced Robotics 17, 599-616, 2007.
[4] Elfes A. "Sonar-based real-world mapping and navigation." IEEE Journal of Robotics and Automation 3, 249-265, 1987.
[5] Ferre M., Macias-guarasa, J.,Aracil R., Barrientos A." Voice command generator for teleoperated robot systems." In Proceedings of the IEEE ROMAN 1998, Takamatsu, Japan 1998.
[6] Ben fredj I. and Ouni K. "Optimization of Features Parameters for HMM Phoneme Recognition of TIMIT Corpus". International Conference on Control, Engineering & Information Technology (CEIT'13). Vol.2, pp.90-94, 2013.
[7] L. R. Rabiner. "A Tutorial on Hidden Markov Models and selected applications in speech recognition", Proceedings of IEEE, Vol. 77, N°2, pp: 257-286, 1989.
[8] Wouter Gevaert, Georgi Tsenov, Valeri Mladenov, Senior Member, IEEE. "Neural Networks used for Speech Recognition". Journal of automatic control, university of Belgrade, vol. 20:1-7, 2010.
[9] A. Waibel "Modular construction of time delay neural networks for speech recognition, neural computation", Vol1, pp. 39-46, Massachusetts USA, 1989.
[10] Kevin J. Lang, A. Waibel. "A Time Delay Neural Network Architecture for Isolated Word Recognition"; Neural Networks, Vol. 3, pp. 23-43, 1990.
[11] Masahide Sugiyamat, Hidehumi Sazoait and Alexander H. Waibel. "Review of TDNNs (time delay neural network) architectures for speech recognition", in Japanese, 1991.
[12] Bennani Y. « Approches Connexionnistes pour la Reconnaissance Automatique du Locuteur » : Modélisation et Identification, Thèse de Doctorat en Sciences, ORSAY, 1992 (in french).
[13] Pierre J. « Techniques neuronales et applications », Les débuts de l'intelligence. 1935 (in french).
[14] http://sourceforge.net/projects/wavesurfer/files/latest/download.
[15] S. Young and al. "The HTK Book (for HTK version 3.4)". Cambridge University Engineering Department, December 2006. http://htk.eng.cam.ac.uk.