Feature extraction module for word recognition on a mobile device based on discrete wavelet transform

Blanca E. Carvajal-Gámez[1], Erika Hernández Rubio[2], Francisco J. Hernández-Castañeda[2], Amilcar Meneses Viveros[3]

[1]Instituto Politécnico Nacional, Unidad Profesional Interdisciplinaria y Tecnología Avanzada, México D.F. México
[2]Instituto Politécnico Nacional, SEPI-ESCOM, México D.F. México
[3]CINVESTAV-IPN, Departamento de Computación, México D.F. México
[1,2]{ehernandezru, becarvajal}@ipn.mx, [2]hcasfj@gmail.com
[3]ameneses@cs.cinvestav.mx

Abstract. The voice is a natural means of interaction between users and mobile devices. Currently, most of the applications based on speech recognition for mobile devices performs searches on large databases through web services without strict voice recognition. These web services require too many resources of storage and processing, which creates a dependency between the mobile device and the communications network. In this paper a word recognition algorithm is presented to perform speech recognition. The processing algorithm is performed in the mobile device. The main contribution is the recognition of words, not just limited to syllables. Word recognition is performed by extracting characteristics through discrete wavelet transform- Haar, to calculate the estimate of the variation of the field of sampling and finally we applied fuzzy logic.

Keywords.- speech recognition, mobile devices, discrete wavelet transform, fuzzy logic.

1. Introduction

The ability to recognize human speech has been an area of interest mainly to the large variety of applications that can be performed. One of the priorities in the development of mobile device technology is to improve the quality and capacity for speech recognition. The speech recognition allows mobile devices adapt the voice information in a comprehensible way, which means the identification and understanding of the information received[1].

One objective of speech recognition systems is to facilitate communication between the user and the device, through the development of applications. There are applications for control systems, through spoken commands, utilizing a voice user interface (VUI), useful for controlling multiple devices, especially for people with certain physical disabilities or working with multimodal interfaces [2]. Other applications allow the user iteration with lexical recognition systems, predefined or online. In this paper, an adaptive VUI [3] mobile device is proposed, to identify words in a noisy environment.

Advances in Computer Science
Research in computing Science 81, 2014 pp. 145-155

Notwithstanding the recent development of methods for speech recognition based in a data base (Google speech recognition API) [4] and synthesis [4], implementations have been performed on personal computers and through remote servers, still have some unresolved needs of embedded systems for mobile devices, autonomous systems, offline systems, control systems, interactive voice, computer equipment for people with physical disabilities. The recognition of large vocabulary of the human speech requires complicated algorithms with big amounts of computations and large memory spaces [1-4]. This results in high energy consumption, greater recognition time and increased error rate. In this paper we use of the discrete wavelet transform (DWT), for extraction of the main characteristics of the voice signal, is proposed in addition to providing compression of the voice signal, resulting in reduced energy consumption, computational complexity and eliminates dependence on the network with an Internet connection. The proposed algorithm with the DWT, decomposes the signals in different frequency components, it possible to locate a specific vibration signal [5]. The DWT can be used for a wide variety of signal processing tasks, such as compression, noise elimination and enhancement of recorded speech [5]. DWT Haar is considered due to the facility in performing its calculations and characteristics relating to the preservation, compaction and redistribution of energy, all without altering the original signal [5]. After applying the compression and obtaining the spectral content of the audio signal with the DWT; continues for extracting characteristics, such as: the variance, standard deviation and the average, and the recognition result is obtained by means of a logic fuzzy that takes as input the result of the application of the DWT-Haar. A method of speech recognition for mobile devices offline for uncontrolled environments, as well as a system to reduce dependence on the internet is proposed. This system of speech recognition is installed on a smartphone with Android 4.0 operating system, with a RAM of 512 MB and a processor Qualcomm MSM7225A @ 800 MHz-1256 DMIPS, with which superior results were obtained at 70% for noisy environments these results are shown in the fourth section. This paper is organized as follows: the second section presents some recent works, the third section presents the architecture and the speech recognition system proposed, the fourth section presents the speech recognition tests performed, and an analysis of the results and finally, conclusions are presented.

2. Related work

According to [1], speech recognition is the ability of a machine or other device to identify words or phrases and convert these in action.

There are two types of speech recognition systems:

I. recognition systems based on speech recognition methods operate directly on a voice terminal.
II. recognition systems based on methods of speech recognition voice operated from a server.

Those that use a server to transmit the sound signal or voice characteristics to the server, which runs on a search engine and returns the speech recognition to the device, it runs so online. The speech recognition systems operate on a server has a

search engine of the word possible matches found, this type of recognition systems is restricted to small vocabularies of certain areas or geographic areas[1].

Placing a speech recognition system on a mobile device and is connected to the server to continually search for the word causes the energy consumption is higher. Currently, the way speech recognition for mobile devices is working through services on the network. The speech recognition is essentially an integration of automatic speech recognition (ASR) and search for the words that match the voice in a database located on a server [6-7]. To improve search precision, some voice search systems also include a module for natural language understanding (NLU) to analyze the output of the ASR in significant segments [6]. Services of speech recognition in network resource-intensive storage and processing[8-10]. An example of how these resources are used in [11] where some of the important features for the development of the Voice Search application are listed by Google as it is the power of the database and the schema for the generation of the structure recognition of patterns of language you lie to the specific case of the creation of the modeling language together its relevant database is shown in [12], this for the Mandarin-Chinese language by Google.

Finally in [13] a system for speech recognition to the Japanese language is done using the Google database. The use of these resources creates a dependency between the mobile device and the communication network as well as has some other problems such as vulnerability to noise, high computational complexity, high memory consumption, high energy consumption and unfriendly interfaces to user [8-10].

3. Proposed method

The most significant problems in speech recognition systems are related to the individuality of the human voice (such as age or gender to name a few), dialect, speaking rate, the context of phonics, noise background, the characteristics of voice acquisition device (microphone), and the directional characteristics of the source speech signal, among others [1-3]. All these problems are considered in the proposed method, because the objective of this research is that the speech recognition is inherent to the environment and the people who use the application. The speech recognition module, based on the DWT-Haar, comprises three main blocks: encoding and compression, feature extraction and recognition, as shown in Figure 1.

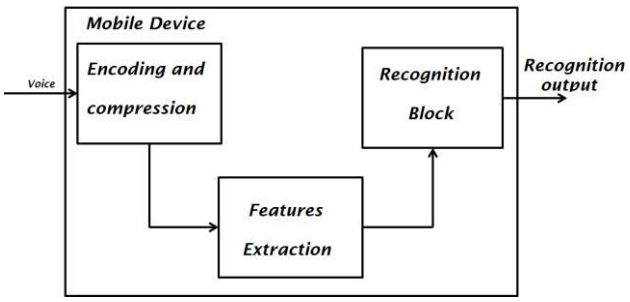


Fig. 1 – Block diagram of the proposed method.

The following subsections explain each of the blocks of the proposed module.

3.1 Block for enconding and compression

In this block, two vectors are obtained: The approximation vector (AV) and the fluctuations vector (FV).

These vectors are obtained when DWT-Haar is applied in the original speech signal. The size of the vectors AV and FV is half the size of the original speech vector. The AV vector contains the low frequency components of the voice signal. The vector FV contains high frequency components of the speech original signal. For this work the AV was chosen, because this vector has the largest amount of information about the speech original signal. To encode speech signal through native methods of recording, signal compression is obtained through the AV with WAV format. Compression has a rate of 22050 samples per second with 16 bits per sample without encoding pulse code modulation (PCM). In the Figure 2, the wavelet decomposition of the original signal is shown in the vector AV $y_{Lo\_D[w/2]}$, and the vector FV, $y_{Hi\_D[w/2]}$. Where $w/2$, is half of the speech signal original vector length.

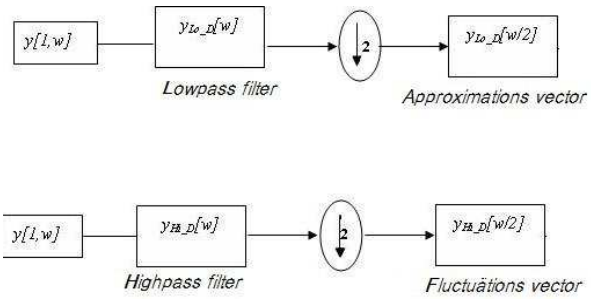


Fig. 2 – Block diagram of wavelet decomposition to the speech original signal.

With the wavelet decomposition of the speech signal and using the vectors AV-FV, we can see the spectral content of the signal, as shown in Figure 3.

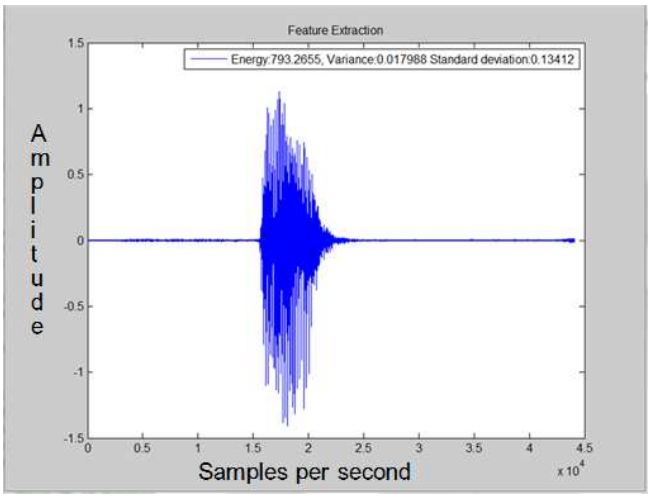


Fig. 3 – Approximations sub-vector obtained of the speech original signal.

In Figure 3, the plotting of the AV vector, obtained from DWT-Haar decomposition of the speech signal is displayed. This graph shows the reduction of 44,100 samples per second (standard WAV file) to 22,050 samples per second. In this vector, has the highest energy input speech signal therefore has the highest content of key features to extract the spectral content of the speech signal and is found in the word contained.

3.2 Block for extracting features

This block obtains the corresponding features of each input speech signal. This extraction is performed in the AV obtained in the previous block. Acquired characteristics are energy Eq. (1), the standard deviation Eq. (2), variance Eq. (3) and the center frequency of the speech signal in the AV.

$$E[y_{Lo\_D}[n]] = \sum_{m=1}^{n} \left| y_{Lo\_D}[m] \right|^2 \tag{1}$$

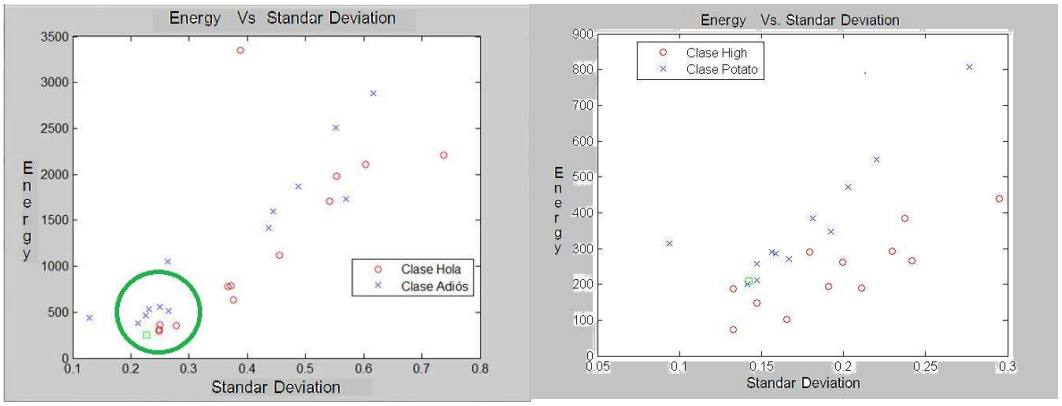
$$\sigma_{y_{Lo\_D}} = \sqrt{\sum_{m=1}^{n} \left(y_{(Lo\_D)_m} - y_{Lo\_D} \right)^2 / n} \tag{2}$$

$$\sigma^2_{y_{Lo\_D}} = \sum_{m=1}^{n} \left(y_{(Lo\_D)_m} - y_{Lo\_D} \right)^2 / n \tag{3}$$

Where $y_{Lo\_D} = \sum_{m=1}^{n} y_{Lo\_D} / n$ represent de mean value of AV, and $n = w/2$.

3.3 Block to recongnition

This block is determined by the spoken word. The words used are suggested: "Hola" and "Adios", and also "High" and "Potato". The words "Hola" and "Adios" in particular were chosen in order to show that the system works with high statistical dependence as shown in Figure 4 a) , contained in the green circle overlapping of these two words is observed.



a) b)

Fig. 4 – Graphic of energy vs. standard deviation, a) words “Hola” y “Adios”, b) "High" and "Potato”.

Unlike the words "High" and "Potato", which in Figure 4 b), the statistical independence is observed. This task is accomplished by using speech recognition. The entries in this block are the characteristics of the speech signal.

Fuzzy logic was used for word recognition in speech signal. This technique allows determining whether a spoken word in the set of speech signal features vector. Fuzzy logic is performed using the membership functions [14] for each of the features extracted from the equations (1), (2), (3) and (4). The membership function indicates the degree to which each element of a given universe belongs to a set. If the set is crisp, the membership function (characteristic function) take the values {0,1}, while if the set is blurred, it will take in the interval [0,1]. If the result of the membership function is equal to 0, then the element is not in the set. In contrast, if the result of the membership function is 1, then the element belongs to the set completely[14]. Gaussian membership, Eq. (4), is used for the purpose of this word recognizer in the speech signal by using fuzzy logic. The Gaussian membership is specified by two parameters $\{c,\sigma\}$ to determine the boundaries of the speech signal, and also to determine where the greatest amount of information is presented in the spectral content of the word to be identified.

$$Gaussian(x;c,\sigma) = e^{-\frac{1}{2}(\frac{x-c}{\sigma})^2} \quad (4)$$

The Gaussian function is determined by the values they take σ and *c*. Where*c* represents the center of the function and σ is the standard deviation (2). For this case, *c* is the mean value, and each value is the mean average of each standard sample, and σ is the standard deviation (2) of each test pattern.

4. Tests and results

The module for speech recognition embedded was implemented on a mobile device. This module identifies words: "Hola" and "Adiós", and also "High" and "Potato".
The speech recognition system was tested in a smartphone with Android 4.0 operating system, 512 MB RAM and Qualcomm MSM7225A @ 800 MHz-1256 DMIPS processor. A population with diverse gender and age were chosen for testing. Samples are shaped with a total population of 12 people, 6 men and 6 women. The ages are classified as follows. People over 50 years old: 3. People between 30 and 50 years old: 3. And people between 20 and 30 years old: 6.

To execute the embedded module, the audio is acquired through the microphone of the mobile device. The speech files that were used have different environmental conditions, because they want to test the robustness and accuracy of the identification of the word within the audio file. The module was also tested in different languages.

Tables 1 and 2 show the results obtained from the algorithm for the embedded module. In these tables you can see the quantitative results of this algorithm. Also a comparison between the signal acquired and processed audio signal is shown.

Table 1. Results obtained by the embedded module to identify "*Hola*" and "*Adiós*"

	Embedded module			
Age range	**Processing time (seg)**	**Sp(%)**	**Se(%)**	**ACC(%)**
age>50	0.109	69.35	93.33	71.08
age>=30&&age<=50	0.116	71.87	96.66	66.66
age>=20&&age<30	0.106	69.49	100	70.55
Average	0.110	70.23	96.66	69.43

Table 2. Results obtained by the embedded module to identify "*High*" and "Potato"

	Embedded module			
Age range	**Processing time (seg)**	**Sp(%)**	**Se(%)**	**ACC(%)**
age>50	0.107	68.25	98.90	73.34
age>=30&&age<=50	0.120	64.82	97.56	72.22
age>=20&&age<30	0.120	63.22	100	73.33
Average	0.115	65.43	98.82	72.96

Performance results are calculated from the speech signal processing through the mobile device. To test the performance of the proposed method, we consider four cases: two for correct classifications and two for misclassification. The classifications

are: true positive (TP), false positive (FP), false negative (FN) and true negative (TN). By using these different measures of performance metrics as the following relation is obtained [15]:

$$Specificity = TN/(TN + FP) \tag{5}$$

$$Sensitivity = TP/(TP + FN) \tag{6}$$

$$Precision = (TP + TN)/samplesofspeechsignal \tag{7}$$

Specificity (Sp) is the ability to detect samples that do not correspond to the audio signal. The sensitivity (Se) reflects the ability of an algorithm to detect the sample audio signal. Accuracy (ACC) measures the ratio between the total number of correctly classified samples (sum of true positives and true negatives) by the number of samples of audio signal[15]. The positive predictive value, and accuracy rate, gives the proportion of samples of the audio signal identified which are true [12,16]. That is, the probability that a sample of the audio signal is identified as true positive. From the results of tests concentrated in Tables 1 and 2, we note the following. The module embedded on the mobile device detects up to 70% of search word (Sp) and up to 96.66% for detecting those who have the search word in the audio file (Se) for the Spanish language. For English language, the embedded module detects up to 65.43% the search word (Sp) and up to 98.82% for detecting those who have the search word in the audio file (Se). In the case of the Spanish language, an accuracy greater than 69.43% is obtained at a time of 0.110 seconds. And for English, an accuracy of 72.96% is obtained in a time of 0.115 seconds. The range of ages, regardless of gender, who presented the best accuracy for word recognition within the mobile device is above 50 years old, for the Spanish language, Table 1. In the case of English language, the same behavior was presented. The greatest accuracy was obtained with the elderly 50 years old, Table 2.

In the case of the Spanish words that begin or end with the same phoneme, such as cases that were used in this research as "Hola" and "Adiós" as well as the noisy environment of the cell itself and the age range of the people, causing the rate of Sp, and Acc is, decrease its performance.

In [4], a module for recognizing isolated words, in real time, on a mobile device is presented. In this study conducted with a sample population of 10 people in total. This population consists of 5 men and 5 women, and the age range of the members of the population is not mentioned. Replays of the audio signals on the mobile performed 3 times to create an average error. The tests in this work were attacked with white noise with SNR, Eq. (8), between 15 and 30 db[4].

$$SNR = \frac{\bar{y}_{Lo\_D}}{y_{Lo\_D}} \tag{8}$$

The SNR has a uniform distribution, so that altering the signal audio not affected and a substantial modification of this signal is taken. This is different to present audio signal to ambient noise, because in this case the signal is affected by impulsive noise. And this kind of noise no predictable distribution, generating in certain sections of

audio are altered significantly. The results presented in [4], are returned in a time of 11.61 ms and with an average of 61.4% in the worst case, when the signal is changed to white noise. And an average of 90.2% is obtained when the audio file is in a controlled environment.

5. Conclusion

The proposal presented in this research for isolated word recognition in a mobile device in uncontrolled environments gives yields higher than 70% in less than the time 0.120 seconds. This embedded word recognizer module requires no prior training or generation of a dictionary as those currently commercially. Also the so-embedded module works offline, ie; not require a connection to the network in order to perform their job recognition. This streamlines its use and management, adding portability and the generation of an App to be used as a tool in voice commands or support systems in any treatment such as Luria tests. We note also that being a working embedded system offline use so the battery is not as affected in the performance of this. The word recognition system achieves work for any genre and any age group not presenting any difficulty, to be altered or amended by voice acuity or severity of the tone of speech signal for the gender of the user, as well as the possible echo the voice generated by age.

Concluding finally that although in some cases the performance is not expected, than the results shown in [4], embedded system mounted on an FPGA, where the processing is done faster and transparent manner.

Aknowledgment

The work is supported by Instituto Politecnico Nacional (IPN), Mexico.

References

1. Husnjak S.; Perakovic D.; Jovovic I.: Possibilities of using Speech Recognition Systems of Smart Terminal Devices in Traffic Environment; *Procedia Engineering*, vol. 69, p.p: 778 – 787, 2014.

2. Oviatt S., Julie A. Jacko, *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications*, CRC-Press, p.p: 286-304 , 2003.

3. Ons B.; GemmekeJort F.; Van Hamme H.: Fast vocabulary acquisition in an NMF-based self-learning vocal user interface, *Computer Speech and Language*, vol. 28, p.p: 997–1017, 2014.

4. Sledevic T.; Tamulevicius G.; Navakauskas D.: Upgrading FPGA Implementation of Isolated Word Recognition System for a Real-Time Operation, *ELEKTRONIKA IR ELEKTROTECHNIKA*, vol. 19, no.10, p.p: 1-6, 2013.

5. Walker,J. S., University of Wisconsin, *A Primer on WAVELETS and Their Scientific Applications*, Chapman and Hall/CRC, Second Edition. 2008.

6. Feng J.; Johnston M.; Bangalore, S., Speech and Multimodal Interaction in Mobile Search, *Signal Processing Magazine*, vol. 28, no. 4, p.p:40-49, 2011.

7. Jeong, H.D.J.; Sang-Kug Y.; Jiyoung L.; Ilsun Y.; Wooseok H.; Hee-Kyoung S., A Remote Computer Control System Using Speech Recognition Technologies of Mobile Devices, *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS)*, p.p: 595-600, 2013.

8. Cohen, J.: Embedded speech recognition applications in mobile phones: Status, trends, and challenges, *Acoustics, Speech and Signal Processing*, IEEE International Conference, p.p: 5352-5355, 2008.

9. Marshall J.; Tennent P.: Mobile interaction does not exist, *Extended Abstracts on Human Factors in Computing Systems*, p.p: 2069-2078, 2013.

10. Flynn R.; Jones E.: Speech Enhancement for Distributed Speech Recognition in Mobile Devices, *Consumer Electronics*, International Conference, p.p: 1-2, 2008.

11. Schalkwyk J.; Beeferman D.; Beaufays F.; Byrne B.; Chelba C.; Cohen M.; Kamvar M.; Strope B.; Google Search by Voice: A case study, Weinstein, A. Visions of Speech: Exploring New Voice Apps in Mobile Environments, *Springer* 2010.

12. Shan J.; Wu G.; Hu Z.; Tang X.; Jansche M.; Moreno P.: Search by Voice in Mandarin Chinese, *Interspeech*, p.p: 354-357, 2010.

13. Shimada T.; Nisimura R.; Tanaka M.; Kawahara H.; Irino T.: Developing a method to build Japanese speech recognition system based on 3-gram language model expansion with Google database, *Conference Anthology IEEE* , p.p: 1-6, 2013.

14. Cavus N.: The evaluation of Learning Management Systems using an artificial intelligence fuzzy logic algorithm, *Advances in engineering software*, vol. 41, no. 2, p.p: 248-254, 2010.

15. Muñoz Pérez C., Cabrera Padilla D., Carvajal Gámez B.E., Gallegos Funes F.J., Gendron D., Segmentación automática en imágenes RGB aplicando la técnica Fuzzy C-means morfología matemática para la ayuda de la foto-identificación de cetáceos, *COMIA 2014*, in press

16. Kumar A.; Tewari A.; Horrigan S.; Kam M.; Metze F., Canny J.: Rethinking Speech Recognition on Mobile Devices, *ACM*, p.p: 1–6, 2011.