Proceedings of the 2007 IEEE Symposium on Computational Intelligence in Image and Signal Processing (CIISP 2007)

Improving Human Computer Interaction through Spoken Natural Language

Omar Florez-Choque
National University of San Agustin, Department of Computer Science
Peruvian Computer Society
Arequipa, Peru
Email: omarflorez19@gmail.com

Ernesto Cuadros-Vargas
San Pablo Catholic University, Arequipa-Peru
Peruvian Computer Society
Arequipa, Peru
Email: ecuadros@spc.org.pe

Abstract— **The fastest and most straightforward way of communication for mankind is the voice. Therefore, the best way to interact with computers should be the voice too. That is why at the moment men are searching new ways to interact with computers. This interaction is improved if the words spoken by the speaker are organized in Natural Language.**

In this article, it is proposed a model to recover information from databases through queries in Spanish Natural Language using the voice as the way of communication. This model incorporates a Hybrid Intelligent System based on *Genetic Algorithms* and a Kohonen Self-Organizing Map (SOM) to recognize the present phonemes in a word through time. This approach allows us to remake up a word with speaker independence. Furthermore, it is proposed the use of a compiler with type 2 grammar according to the Chomsky Hierarchy to support the syntactic and semantic structure in Spanish language. Our experiments suggest that the Spoken Natural Language improves notably the Human-Computer interaction when compared with traditional input methods such as: mouse or keybord.

I. Introduction

According to Wickens [1] the information in the Human-Computer interaction is carried out by means of two modalities, sounds and images. The user performs several tasks at the same time. If these tasks require the similar resources from the user, it will take place effects of *interference* which increase the user mental load. An interface based on natural language to recover information from a database allows us to reduce the user mental load, since it requires the input of data by means of voice and it presents the processed information in a visual interface. This model allows a high interaction degree between the user and the computer. Laurel [2] defines this interaction degree in proportion to the assignment of anthropomorphic skills to the computer from user during the interaction.

In that sense, the Isolated Word Recognition systems were some of the first applications in the voice treatment, due to the easiness of recognizing words directly of the voice spectrum. However these systems depended on the speaker and involved the storing and training of the whole word, which affected the scalability of the system in very extensive domains as the Natural Language. For that reason, now a days words are recognized through the identification of phonemes, with which it is possible to remake up a word. This task implies several phases such as: identifying the limits where words begin and end inside the voice spectrum, normalizing the data, applying slicing windows to separate in segments the signal, each segment can store a phoneme, and finally extracting the main features of each segment.

Atal demonstrated that short-time power spectrum of speech in different frequencies, presented a high correlation degree and thus they may be very useful for speech recognition with speaker independence [3]. One of the contributions of our work lies in proposing the use of a sequence of 12 Mel-frequency Cepstrum Coefficients (MFCC) with a short period of time (5 miliseconds) as training vectors of a Hybrid Intelligent System based on neural networks and genetic algorithms for the words recognition in real time with speaker independence. This new approach arises from the observation that according to the selected segments, some coefficients of MFCC showed a high correlation in different words, this shows the presence of similar phonemes in different words. The neural network SOM is very useful to reduce the dimensions of each feature vector and visualize in a map of 2 dimensions the clusters of phonemes.

Voice2SQL is particularly useful to attend users with difficulties in the use of conventional input devices as the keyboards or pointer devices. This kind of users, generally with dissabilities, will be able to carry out queries on a database through simple spoken user interface. Across the following pages, we propose a system for the generation of SQL sentences based on Spoken Natural Language.

The rest of this paper is organized as follows. In Section II briefly describes related works to voice recognition. Section III describes the software architecture of the model by discussing the hierarchical relationships between each one of the components. Section IV describes the steps performed to obtain a vector from the voice signal. Section V shows one of our contribution, that is linguistic criteria which is used in the convertion of Natural Language to SQL sentences. In the Section VI we discuss the obtained results. And lastly, section VII provides our conclusions.

II. Previous works

Previous works in the voice recognition implied the detection of features from a signal in the time-domain (*waveform*) [4], [5], [6], [7], these features included measures in the time-domain as the energy of the signal, the average rate of *zero-crossing*, and the correlation of features using a short-time power spectrum of speech in different frequencies. Two of the main problems in the voice recognition are the variability of the spectrum energy and the search of patterns to form possible words. The problem of variability can be partially solved by the coefficients of Cepstral. These coefficients attenuate the components of the spectrum. Some of the methods proposed in the past that fulfilles this task were the average subtraction of Cepstrals [8] and the *RASTA filtering* [9]. The use of a group of phonemes represented by *Hidden Markov models*, is also broadly used [10], [11], [12]. Also, for features extraction, the creators of Sphinx II [10] has shown that Cepstral coefficients provide a better recognition performance because these coefficients are evaluated in a logarithmic and equispaced scale. On the other hand, professor Teuvo Kohonen in 1988, introduced a new architecture of neural networs called Self-Organizing Map (SOM), Kohonen used SOM for the recognition of phonetic units as a speaker adaptative system with an unlimited vocabulary.

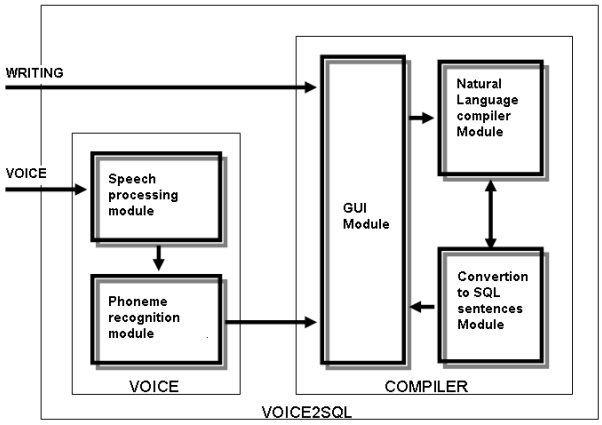

Fig. 1. Software architecture of Voice2SQL. Notice two types of input to the system: spoken natural language and written natural language.

Finally, along these years a kind of Intelligent Hybrid System denominated *evolutionary neural networks* have been used in the design of the architecture of a neural network to solve the classic XOR problem ([13], [14], [15]) with interesting theoretical results. Another application is the use of genetic algorithms to select the training data and to interpret the behavior of the outputs of the neural network. Schaffer, Whitley and Eshelman (1992) have extensively studied these areas.

III. Software architecture of the system

The software architecture of the proposed system can be seen in the Figure 1 where two main components can be distinguished :

- The component of voice precessing (*VOICE*) and
- The component of Natural Language compilation (*COMPILER*)

The *VOICE* component includes the voice processing module, which captures the voice from the microphone, identifies the beginning and end of each word inside the voice spectrum, segments the sign using slicing windows, calculates the Mel-frequency Cepstrum Coefficients (MFCC) for each segment and normalizes each segments to 256 features vectors. Each vector stores the first 12 Cepstrum Coefficients of each segment. The module of voice-to-text conversion uses the Self-Organizing Map of Kohonen for training with feature vectors as inputs. In view of each vector represents a segment, it is possible to find phonemes inside the segments. The classification of the vectors represents the cluster of similar phonemes on the SOM.

The *COMPILER* component includes the module of graphic interface, the module of natural language compilation, which verifies the correct syntactic structure of the sentence, based on the grammatical rules of the Royal Academy of Spanish Language [16] and the conversion of natural language to SQL sentences module, wich carries out the convertion based on syntactic approaches for the recognition of Entities.

IV. Feature extraction from the speech signal

It is possible to obtain information from the signal in the short-time power spectrum, so the number of extracted features will be smaller than the total number of points obtained based on the signal without processing. We propose the use of MFCC to extract the signal features and a later data normalization to obtain meaningful distances among vectors.

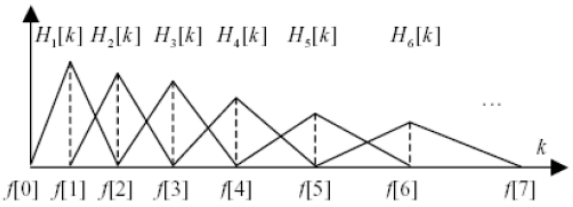

Fig. 2. Bank of triangular filters used to calculate the Cepstrum coefficients in the Mel scale. Notice that the area is the same for each filter

A. Mel-frequency Cepstrum Coefficients

The Mel-Frequency Cepstrum Coefficients (MFCC) characterizes the voice spectrum according to the number of chosen coefficients 12 or 13 are usually good approaches for speech recognition [12]. The MFCC are defined by the real part of the Cepstrum Coefficients of a short-time signal after a segmentation process, through slicing windows, as Hammin Windows, and Fourier and Cosine transforms on each segment of the signal.

The MFCC are an improvement to the conventional Cepstral coefficients. They take into account a logarithmic scale of frequency calles *Mel-frequency scale*. Mel-frequency scale was proposed by Stevens, Volkman and Newmann in 1937, and it is based on a musical perceptual scale, keeping in mind equispaced observers. Above 500 Hz, the equispaced exponentially intervals of frequency are perceived as if they were spaced lineally. This scale may be calculated by means of equation 1.

$$Mel(f_s) = 2595 \log_{10}(1 + \frac{f_s}{700}) \quad (1)$$

Then, it is possible to define a bank of M filters, where each filter is a triangular filter with the following transfer function:

$$H_m[k] = 0,\ k < f[m-1]$$

$$H_m[k] = \frac{2(k - f[m-1])}{(f[m+1] - f[m-1])(f[m] - f[m-1])},\ f[m-1] \leq k \leq f[m]$$

$$H_m[k] = \frac{2(f[m+1] - k)}{(f[m+1] - f[m-1])(f[m+1] - f[m])},\ f[m] \leq k \leq f[m+1]$$

$$H_m[k] = 0,\ k > f[m+1]$$

As depicted in Figure 2, each filter calculates the average sprectrum around each center of frequency with growing bandwidth. The area inside each filter is always the same.

Let $f_1$and f_h be the lowest and higher frequencies inside the bank of filters measured in Hertz. Let f_s be the sampling frequency measured in Hertz and be M the number of filters inside the bank of filters. The boundary points of the triangular areas of the Figure 2 are evenly spaced in the Mel-scale by means of

$$f[m] = (N/f_s) Mel^{-1}(Mel(f_1) + m\frac{Mel(f_h) - Mel(f_1)}{M+1}) \quad (2)$$

where the Mel scale is defined by

$$Mel(f) = \tilde{f} = 1125 \ln(1 + \frac{f}{700}) \quad (3)$$

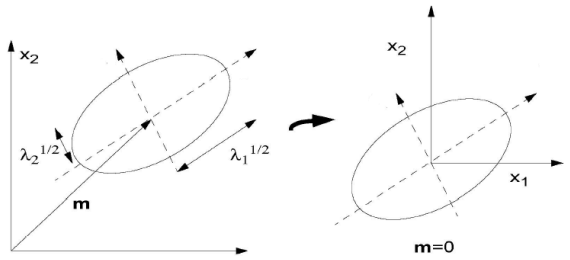

Fig. 3. Normalized features with mean zero and unitary variance.

and its inverse by

$$Mel^{-1}(\tilde{f}) = 700(e^{\frac{f}{1125}} - 1) \quad (4)$$

then, it is possible to calculate the energy based on a logarithmic scale at the output of each filter by means of

$$E[m] = \ln \sum_{k=0}^{N-1} |F[k]|^2 H_m[k] \quad (5)$$

Finally, the Cepstrum coefficients in Mel frequency scale are the Discrete Cosine Transform (DCT) of the output M filters .

$$MFCC[n] = \sum_{m=0}^{M-1} E[m] cos(\pi n(\frac{m+\frac{1}{2}}{M})), 0 \leq n \leq M \quad (6)$$

The work of Huang and Hong [17] demonstrates that the representation based on MFCC are specially useful for voice recognition.

1) Data normalization: The feature vector is formed by several extracted values from features extraction phase, in this case for the use of MFCC in each segment of the signal. Each coefficient of the MFCC is a dimension in the feature vector. Features with big values have a bigger influence on the dataset rather than features with small values, because bigger values will vary with more intensity the distance function.

Let m_d be the mean of the d-*th* feature. Let σ_d^2 be the variance of the d-*th* feature and let $\overset{x}{}_{nd}$ be the normalized feature. The mean, variance and normalized feature of the d-*th* feature of the feature vector with longitude N is defined respectively by means of

$$m_d = \frac{1}{N} \sum_{n=1}^{N} x_n \quad (7)$$

$$\sigma_d^2 = \frac{\sum_{n=1}^{N} (x_{nd} - m_d)^2}{N-1} \quad (8)$$

$$x_{nd} = \frac{x_{nd} - m_d}{\sigma_d} \quad (9)$$

After the normalization process, the features have mean zero and unitary variance, just as is observed in the Figure 3.

B. SOMs as phoneme recognizers

Genetic Algorithms (GAs) and Neural Networks (NNs) are two biologically motivated computational models. For that reason, it is not surprising that several researchers have explored the idea of *Evolutionary Neural Networks*. This approach, based on GAs, involves the code of possible solutions as strings with binary values of chromosomes, the initial population setting of chromosomes and the use of genetic operators such as selection, crossover and mutation

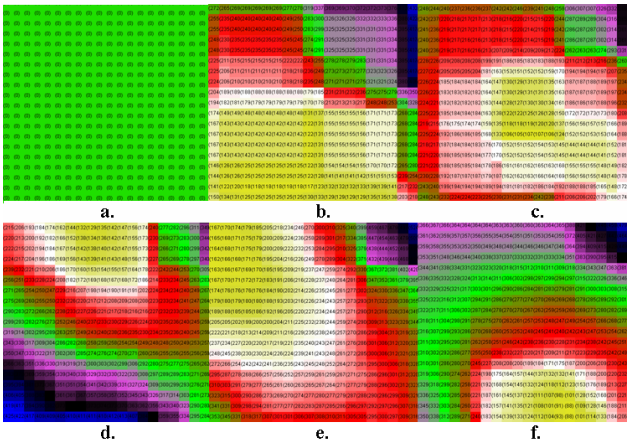

Fig. 4. Ilustration of the SOM with different iterations. a. 0 b. 100 c. 500 d. 1000 e. 3000 f. 5000. Notice that while the number of iterations increases, changes on the topology decrease, because learning rate and neighbourhood radius also decrease in function to the number of iterations.

allows us to derive in a SOM with better fitness. This fitness represents the quality in the topology of the Map, keeping in mind that each SOM finds different similarities among the same training vectors. Therefore, the GA finds the better topology, that is, the topology with smaller distance regarding the phonemes characterized as floating point vectors.

Let *N* be the size of the Kohonen Map and let *a* be the activity of each neuron, that is, the distance between the value of this neuron and the recent input vector. Therefore we have used the following equation to compute the *fitness* of each map:

$$fitness = \frac{1}{1 + \sum_{n=1}^{N} a} \quad (10)$$

According to the Figure 4 the evolution of the SOM is shown in different iterations. It is possible to observe how the SOM iterates to form regions of similar color based on the presence of common features, correlations, similarities and differences among the input patterns of the neural network. These patterns are formed by the extracted MFCC of the sentences spoken by the user. Therefore, regions with similar colors indicate the presence of similar phonemes. According to the Figure 5 we can see that some neurons have specialized to learn differents phonemes. Then, it is possible to generalize on the Kohonen Map to form words based on phonemes.

V. Natural Language Processing

Another of our contributions is the use of a linguistic criteria to identify the main Entities that participate in a sentence of Natural language, such as actions, tables, attributes, values and conditions.

First, words related with queries actions are recognized choose, calculate, enumerate, select, and so on) and their combinations in first, second, third person and passive voice. Then, indirect objects are recognized through the presence of prepositions (in, of, with, from, to, and so on) so the noun in the nominal sentence will identify the possible presence of an Entity, for instance in the following example **students** is reconized as an Entity:

- Enumerate the name *of* those **students**...

Then, it is detected the presence of preposition and comparators (greater than, lower than, equal to, different from) between a name (attribute name) and a identifier (attribute value)

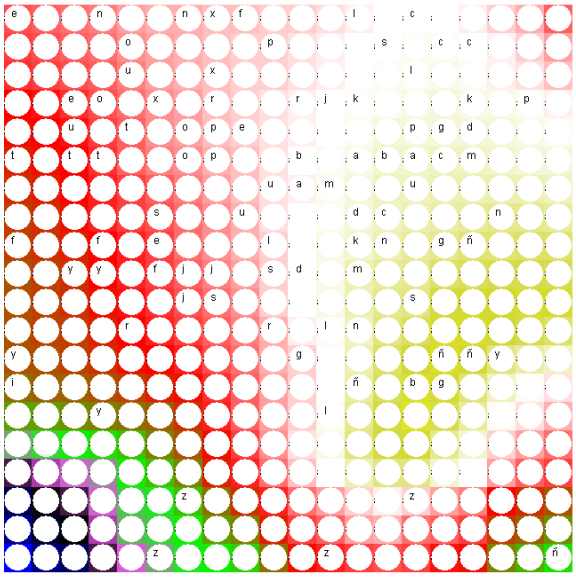

Fig. 5. Ilustration of the SOM with different learned phonemes. Note that some neurons have specialized in certain phonemes and *special phonemes* as 'ñ', 'y' and 'z' are located in isolated clusters.

- Enumerate the name of the students of the *course* **of** *AI*
- to Select those *marks* **grater than** *B*

therefore is possible to associate two syntactic elements in the following way

- *course* = *AI*
- *mark* >*B*

Then cardinal numbers are recognized (first, second, third, so on) in the following way

- Enumerate the **first** *place*...

to recognize structures of the type

- *place* = **1**

Finally, the presence of adjectives is detected. Adjetives represent the features and conditions of recognized Entities

- Enumerate the **best** *students*...
- Calculate the most **studious** *students*...

to recognize structures of this type

- **best**(*students*)
- **studious**(*students*)

In view of the fact that these adjectives can be subjectives to each person should be represented according to the domain of the problem. The Fuzzy Logic can help to model the grade of subjectivity of each adjective.

Finally some of the outputs of the system are shown based on queries in natural language:

- Voice entry: "*Calculate the average mark in the students of the course of AI* "
- Output:

```
Select AVG(Mark) From Students
Where Course = 'AI'
```

- Voice entry: "*Select the students of fifth cycle that are not registered in the course of AI* "
- Output:

```
Select * From Students Where
Course <> 'AI' and Cycle = 5
```

VI. Experimental results

The speech data was uttered for two males and two females Peruvian speakers. People of different countries have a different tone and speed in pronuntiation words, that is the reazons to highlight the nationality of the speakers. Each one read the Spanish alphabet three times in different order to train the SOM. Each dataset contained 63 phonemes and was divided in 5 codebooks (*abcde*, *fgijk*, *lmnño*, *pqrst* and *uvxyz*) each one with 256 feature vectors, each feature vector contained 12 MFCCs. Therefore each user training dataset contained 1280 training patterns.

The sound was digitized by 16 bits with segments of 5 miliseconds and sampled to 12.8 kHz. Our experiments showed that vowels and plosives are easily clustered, but it is more difficult with fricatives.

A. Results

The obtained results using three different classifiers are listed below. The classifiers were Self-Organizing Map (SOM), Learning Vector Quantization (LVQ) and Multi Layer Perceptron (MLP).

characteristic	classifier	accuracy
MFCC	SOM	78.7 %
MFCC	LVQ	92.3 %
MFCC	MLP	68.1 %

TABLE I
THE ACCURACY RATE IN THE PHONEMES RECOGNITION USING SOM, LVQ AND MLP AS CLASSIFIERS.

characteristic	classifier	error
MFCC	SOM	26.3 %
MFCC	LVQ	6.0 %
MFCC	MLP	28.2 %

TABLE II
THE ERROR RATE IN THE PHONEMES RECOGNITION USING SOM, LVQ AND MLP AS CLASSIFIERS.

VII. Conclusions

With this paper we have concluded that:

A foreseen advantage is related with spoken natural language. The use of voice and natural language allow a better experience with the computer. Furthermore, it allows to include users with disabilities such as: blind people and people with Parkinson.

An approach based on natural language is closely related with the chosen language. Others languages than Spanish with strongly different grammar (Cantonesse, German, English, and son on) will requiere a different approach to try verbs, prepositions, adjetives, etc. SOM allows to represent in a straighforward way a complex and stochastic phenomenon as the voice.

The approach based on phonemes does possible to scale the system to recognize any present word in the Spanish language. Therefore it is possible to work with unlimited dictionaries.

Future works will imply to use Semantic and Fuzzy Logic techniques to reduce the ambiguity and represent the subjectivity of adjectives. Furthemore, the use of Hidden Markov Models (HMM) will improve the accuracy to recognize words. In view of the fact that HMM takes into account the probabilities of transition among phonemes.

ACKNOWLEDGMENTS

The author wants to thank very especially the teachings and advices from teachers of the National University of San Agustín at Arequipa, specially Dr. Ernesto Cuadros-Vargas and Luis Alfaro Casas. Besides, I would like to express my deep gratitude to Dr. Rosa Alarcón Choque, University of Chile at Santiago, and Marco A. Alvarez, State University of Utah, for their invaluable encouragement.

REFERENCES

[1] C. D. Wickens and J. G. Hollands, *Engineering Psychology and Human Performance (3rd Edition)*. Prentice Hall, September 1999. [Online]. Available: http://www.amazon.de/exec/obidos/ASIN/0321047117

[2] B. Laurel, "Interface agents: Metaphors with character," 1999, pp. 355–365.

[3] B. Atal, "Automatic recognition of speakers from their voices," in *Proceedings of the IEEE*, vol. 64, April 1976, pp. 460–475.

[4] M. Hunt, "Spectral signal processing for asr," 1999.

[5] L. Gu and K. Rose, "Perceptual harmonic cepstral coefficients for speech recognition in noisy environment," 2001.

[6] M. Gales, "Model-based techniques for noise robust speech recognition," 1996. [Online]. Available: citeseer.ist.psu.edu/gales95modelbased.html

[7] R. Schlüter and H. Ney, "Using phase spectrum information for improved speech recognition performance," 1998.

[8] S. Johnson, P. Jourlin, G. Moore, K. S. Jones, and P. Woodland, "The cambridge university spoken document retrieval system," in *Proc ICASSP '99*, vol. 1, Phoenix, AZ, 1999, pp. 49–52. [Online]. Available: citeseer.ifi.unizh.ch/johnson99cambridge.html

[9] H. Hermansky and N. Morgan, "RASTA processing of speech," in *IEEE Transactions on Speech and Acoustics*, vol. 2, October 1994, pp. 587–589.

[10] X. Huang, F. Alleva, H.-W. Hon, M.-Y. Hwang, and R. Rosenfeld, "The SPHINX-II speech recognition system: an overview," *Computer Speech and Language*, vol. 7, no. 2, pp. 137–148, 1993. [Online]. Available: citeseer.ifi.unizh.ch/huang92sphinxii.html

[11] D. J. Kershaw, "Phonetic context-dependency in a hybrid ann/hmm speech recognition system," 1996. [Online]. Available: citeseer.ifi.unizh.ch/175909.html

[12] J. Neto, L. Almeida, M. Hochberg, C. Martins, L. Nunes, S. Renals, and A. Robinson, "Speakeradaptation for hybrid hmm-ann continuous speech recognition system," 1995. [Online]. Available: citeseer.ifi.unizh.ch/neto95speakeradaptation.html

[13] L. D. Whitley, S. Dominic, and R. Das, "Genetic reinforcement learning with multilayer neural networks." in *ICGA*, 1991, pp. 562–569.

[14] G. F. Miller, P. M. Todd, and S. U. Hegde, "Designing neural networks using genetic algorithms." in *ICGA*, 1989, pp. 379–384.

[15] S. G. Romaniuk, "Evolutionary growth perceptrons." in *ICGA*, 1993, pp. 334–341.

[16] F. R. Adrados, *Semantics and Syntax Studies of Spanish Language*. Planeta, January 1975.

[17] H. Huang, A. Acero, and H. Hon, *Spoken Language Processing - A Guide to Theory, Algorithms and Systems Development*. Prentice - Hall, 2001.