Fuzzy k-Nearest Neighbors Applied to Phoneme Recognition

Ines Ben Fredj and Kaïs Ouni
Research Unit Signals and Mechatronic Systems SMS, UR13ES49
National Engineering School of Carthage, Carthage University
Tunis, Tunisia
ines_benfredj@yahoo.fr , kais.ouni@enicarthage.rnu.tn

***Abstract*—In this work, the Fuzzy kNN (FkNN), an alternative of the standard kNN algorithm, is used for Timit phoneme recognition. Phoneme is the smallest unit that composes speech. For this reason, if phoneme recognition is performed, it can achieve a significant word and text recognition. Thus, the main idea consists on assigning phoneme membership to the data phonemes by measuring the distance to its kNN. FkNN compute the fuzzy distances between the data phonemes that define the cluster fuzziness. Mel Frequency Cepstral Coefficients (MFCC) associated with their first and second derivatives and energy coefficient were extracted from the speech signals as an input of the recognition system. A comparison of a crisp and fuzzy kNN was performed. Experiments show that FkNN algorithm not only can lead to significant recognition rates, but also may supersede in some ways Hidden Markov Model (HMM) the reference of speech recognition.**

Keywords—FkNN; Fuzzy; HMM; kNN; MFCC; Speech recognition.

I. INTRODUCTION

For many years, phoneme recognition has presented a main topic studied by researchers. It is the process of matching a sequence of acoustic vectors for a phoneme label [1]; Such process requires extensive use of language models based on mathematical approaches such as fuzzy logic, support vector machines (SVM), neural networks [2] [3], hidden Markov models (HMM), which aims to solve the problems of data modeling, classification and decoding. The relevant recognition pattern is the one that ensures the compromise between high accuracy and minimum execution time which is a difficult task [4].

For instance, the authors in [5] describe a novel phoneme recognition approach that use as input a raw speech signal to the artificial neural networks (ANN) and a phoneme class conditional probability as output. They applied on Timit phoneme recognition task, different ANN architectures to prove the advantage of a Convolutional Neural Networks (CNN). Authors compared their approach to conventional approach where spectral-based features Mel Frequency Cepstral Coefficients (MFCC) were extracted and modeled by a multilayer perceptron. They exhibited that the proposed approach can yield better phoneme recognition performance when compared to the conventional approach. Phoneme recognition accuracy varied from 38.91% to 71.80%.

In addition, other researchers recommend in [6] an analysis of a low-dimensional representation of speech for modeling speech dynamics, extracted using bottleneck neural networks. Experiments are operated on Timit corpus. They spend the bottleneck features in a conventional HMM for phoneme recognition. Recognition accuracy attains 70.6% with only 9-dimensional features.

Furthermore, authors in [7] illustrate that incorporating articulatory parameters can improve the rate of Timit phoneme recognition based on HMM. They exploit the voice signal by MFCC with articulatory parameters. The articulatory parameters provide pertinent information for phoneme recognition objective such as 69.26% was obtained as the highest precision rate. Authors note that the recognition rates of correct phonemes were at least as high as the precision rate.

For our study, an approach of Timit phoneme recognition using the Fuzzy k-Nearest Neighbor (FkNN) is implemented. This algorithm has been used in several areas such as speech emotion recognition [8], moulds detection [9], activity recognition [10], face recognition [11] and others. Despite that it is an old algorithm (since 1985); it shows until nowadays that it can be an effective way for recognition as already mentioned in the references.

FkNN is characterized by simple implementation, efficiency and speed of execution particularly with 'k' reduced. This presents a major motivation of our work. Therefore, FkNN is introduced for Timit phoneme recognition using a set of classical speech parameter as MFCC.

The remainder of the paper is prepared as follows. Section 2 presents the basic principle of FkNN algorithm. Section 3 includes the organization of Timit corpus and the architecture of the proposed phoneme recognition approach. Section 4 details the experimental setup and discusses the results obtained. And finally, section 5 concludes the paper and defines some futures works.

II. FUZZY K-NEAREST NEIGHBORS

Fuzzy k-Nearest Neighbor (FkNN) is based essentially on the standard k-NN algorithm. Both algorithms classify by the k nearest neighbors and make use of measure of similarity such as distance among samples to define classification decision [12]. The k-NN decision rule has often been used in

978-1-5090-4712-3/16/$31.00 ©2016 IEEE

pattern recognition problems. The major problem of this technique is about the equal importance given for the labeled samples when deciding about the class memberships of the pattern request. That's why the theory of fuzzy sets was introduced into the k-NN technique to improve a fuzzy version of the algorithm [13].

A. K-NN algorithm

K-NN algorithm consists on assigning a sample request to the class that contains more k-nearest neighbors. Classes are defined in advance and specified by sets of elements. The number of elements may vary among classes [14].

1) K-NN algorithm

K-NN algorithm is detailed as follow:

- Input: D the training set, 'k' training objects and test object $z = (x', y')$
- Compute $d(x', x)$ the distance between 'z' and all examples $(x, y) \in D$
- Select $D_z \subseteq D$ the set of 'k' closest training objects to z
- Output: $y' = \arg\max_{v} \sum_{(x_i, y_i) \in D_z} I(v = y_i)$

The object 'z' is assigned to the most common class among its 'k' nearest neighbors.

2) Distance measure

In general, the Euclidian distance is used for k-NN algorithm. However, other distance measures can be applied such as Minkowski, Tchebychev and Correlation distances [15].

Given two samples characterized by x_i, x_j vectors, Euclidian distance between these samples is given by:

$$d(x_i, x_j) = \sqrt{\sum_{p=1}^{N} (x_{ik} - x_{jk})^2} \quad (1)$$

FkNN employs fuzzy rules to outperform the kNN algorithm. Under kNN algorithm, once an input sample is fixed to a class, there is no indication of its strength of membership in that class. By this way, the basic plan of the FkNN is to assign membership as a function of the vector's distance from its k-nearest neighbors in the possible classes [16].

FkNN algorithm is presented as follow:

- Find k nearest neighbors $x_j, j = 1...k$ of the given set objects x using Euclidian distance
- Calculate the membership degrees for the N_C classes $(c_i, i = 1...N_C)$ as given by formula (2):

$$\mu_{c_i}(x) = \frac{\sum_{j=1}^{k} \mu_{c_i}(x_j) \mathrm{d}_j^{-2/(m-1)}}{\sum_{j=1}^{k} \mathrm{d}_j^{-2/(m-1)}} \quad (2)$$

where: $d_j = \left\| x - x_j \right\|$ is the Euclidian distance between x and x_j, $\mu_{c_i}(x_j)$ is the membership degree of x_j for class c_i and the parameter 'm' controls the effective magnitude of distance of the prototype neighbors from the query object and it is chosen during cross validation beside k.

- Finally, the class label c_x of the sample 'x' is selected by equation (3)

$$c_x(\mathrm{x}) = \arg\max_{i} (\mu_{c_i}(\mathrm{x})) \quad (3)$$

III. FkNN Phoneme Recognition Approach

A. Speech Corpus: Timit

Phoneme recognition experiments were performed using a data extracted from Timit database [17]. It consists of 630 speakers from 8 major dialect regions of the United States; each saying 10 sentences which gives 6300 sentences. All dialects are divided into a training set and a test set. For our study, the New England dialect (DR1) is used with its train and test sets. It contains 31 male speakers and 18 female speakers.

B. Feature extraction: MFCC

The analysis MFCC is widely applied in speech processing. It is based on the variation of the critical bands of the human ear with frequency; filters are spaced linearly at low frequencies and logarithmically at high frequencies. These filters are modeled by a non-linear scale based on knowledge of human perception that is the Mel scale.

To extract MFCC coefficients, the Hamming window is applied into the transformation from the time domain to the frequency domain. This transformation is assured by the Fourier transform [18]. A filter is assigned afterward by triangular filters spaced according to the Mel scale. This scale reproduces the selectivity of the human hearing. The log is calculated and a discrete cosine transform is done to return to the time domain. MFCC algorithm is detailed in [19] [20].

MFCC coefficients were extracted from the speech signal with 256 sample frames. They were Hamming windowed in segments of 25 ms length every 10 ms with a sampling frequency equal to 16 kHz. Twelve static MFCC added to energy coefficient were recovered by their first and second derivatives to represent temporal variations in the spectrum of the signal [21].

C. Data organization: phoneme distribution

To boost the phoneme recognition task, a particular phoneme class distribution according to pronunciation information [22] was employed as table 1 illustrates.

TABLE I. PHONEMES DISTRIBUTION

Macro class	Type	Phonemes
Vowels	Front	/iy/ /ih/ /ey/ /eh/ /ae/ /ix/ /ux/ /oy/
	Middle	/er/ /ax/ /ah/ /ay/
	Back	/uw/ /uh/ /ow/ /aw/ /ao/ /aa/
	Principal vowel	/axr/ /ax-h/
Affricates	----------	/jh/ /ch/
Stops	Voiced	/b//d/ /g
	Unvoiced	/p/ /t/ /k/
	Voiced Closure	/bcl/ /dcl/ /gcl/
	Unvoiced Closure	/pcl/ /tcl/ /kcl/
Fricatives	Voiced	/v/ /dh/ /z/ /zh/
	Unvoiced	/f/ /th/ /s/ /sh/
Nasals	Nasal	/m/ /n/ /ng/
	Syllabic nasal	/em/ /en/ /eng/
	Flap nasal	/nx/
Semi-vowels	Voiced	/l/ /r/ /y/ /w/ /el/
	Unvoiced	/hh//hv/
Silences	----------	/pau/ /epi/ /h#/

D. Choice of the parameter 'k'

The value of 'k' is really training-data dependent, altering the position of a few training data may lead to a major loss of performance. The value of optimum 'k' totally depends on the data used and may vary in different cases. In most schemas, a small 'k' value analytically adjusted can be more advantageous than a large value.

For our experiments, a range values for 'k' varying from 2 to 100 was chosen subjectively in order to survey the possibility of a maximum performance. Nevertheless, there are different tools for determining an optimal 'k' value such as the Cross-Validation and Bootstrap method [23].

E. Recognition process

Fig. 1 shows the architecture of recognition approach proposed.

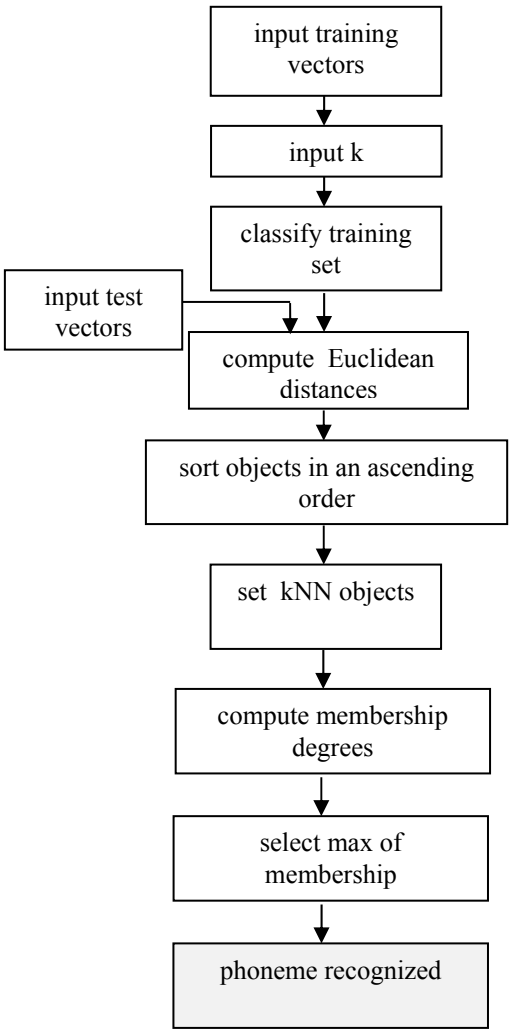

Fig. 1. Diagram of recognition process

After MFCC extraction and data organization, FkNN was evaluated on train and test sets.

Each phoneme is represented by a vector of 39 MFCC coefficients which characterizes its three middle frames. As an input of FkNN, we calculate the mean of each phoneme vector in order to reduce data size and expedite compute of distances.

FkNN calculates the Euclidian distance between training data and each test phoneme vector and keeps the set of the k-nearest neighbors. After that, all objects are sorted in an ascending order to set the kNN objects.

As mentioned previously, several k-values were tested and adjusted from 2 to 100. Already for each 'k' selected, an output was gotten. The output that maximizes recognition rates was then fixed. It should be noted that the recognition performance was not be reached with the same 'k' for the test objects.

Therefore, the membership function is evaluated and a membership degree for all training phonemes is attributing for both test phonemes. To finish the recognition task, the

phoneme recognized is relative to the maximum membership degree.

IV. Results and Discussion

Table 2 shows the phoneme recognition rates according to the distribution data described above. The mean recognition rate is accomplished by weighting the class means by the number of phonemes in each class (using a weighted mean of the class means).

The data set size is variable among phoneme sets as for recognition rates. For the macro classes, we mention that the mean recognition rates have attained respectively 65.93%, 78.57%, 68.55%, 86.35%, 68.00%, 78.02% and 82.29% for vowels, affricates, stops, fricatives, nasals, semi-vowels and silences; thus a general recognition rate of 72.21% of phoneme recognition for a set of 4021 phonemes.

Moreover, we point out a relevant recognition rates especially for the voiced Stops (82.90%) and Fricatives (83.66%), the unvoiced Fricatives (88.03%), Nasals (91.66%) and Semi-vowels (92.30).

In addition, it can be observed that the recognition of vowels have lead to a moderate performance (under 60%). This can be justified through the difficulty already well-known for this task such the remarkable similarity of vowels. This point performs similar recognition decision justified by the Euclidian distance.

TABLE II. Phoneme Recognition Rates (PRR %)

Macro class	Phonemes	Number of Phonemes	PRR (%)
Vowels	/iy/ /ih/ /ey/ /eh/ /ae/ /ix/ /ux/ /oy/	509	65.61
	/er/ /ax/ /ah/ /ay/	520	66.15
	/uw/ /uh/ /ow/ /aw/ /ao/ /aa/	257	56.03
	/axr/ /ax-h/	82	97.56
Affricates	/jh/ /ch/	42	78.57
Stops	/b//d/ /g	193	82.90
	/p/ /t/ /k/	244	73.77
	/bcl/ /dcl/ /gcl/	206	62.13
	/pcl/ /tcl/ /kcl/	324	60.18
Fricatives	/v/ /dh/ /z/ /zh/	202	83.66
	/f/ /th/ /s/ /sh/	326	88.03
Nasals	/m/ /n/ /ng/	298	64.76
	/em/ /en/ /eng/	12	91.66
	/nx/	22	99.00
Semi-vowels	/l/ /r/ /y/ /w/ /el/	444	76.35
	/hh//hv/	52	92.30
Silences	/pau/ /epi/ /h#/	288	82.29
Total phonemes	**4021**	**Mean PRR**	**72.21**

Fig. 2 and Fig.3 illustrate a similar distribution of a different vowel samples such as /ay/ and /ah/ phonemes. Vectors in red color are training samples and vectors in blue color are a test sample.

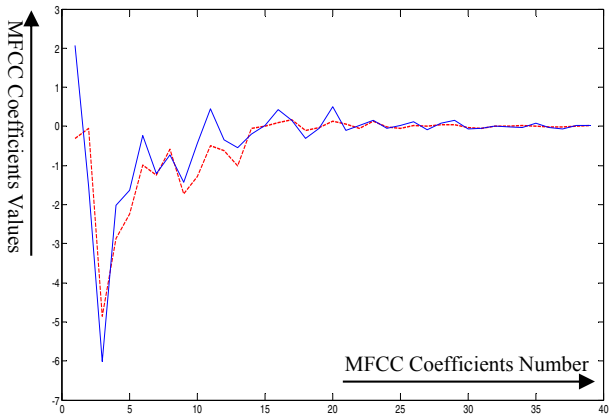

Fig. 2. Example of / ay/ distribution (red color for a training /ay/, blue color for a test /ay/)

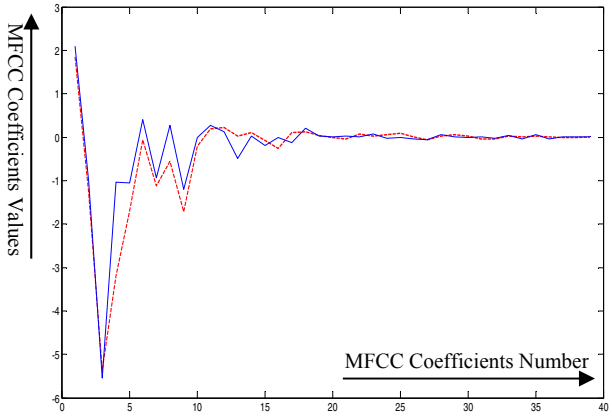

Fig. 3. Example of /ah/ distribution (red color for a training /ah/, blue color for a test /ah/)

In addition, a comparative study of the standard knn algorithm and FkNN algorithm induced a significant difference due to the fuzzy membership principally in training phase (see table 3).

TABLE III. Comparison of Crisp and Fuzzy KNN in Timit Phoneme Recognition

	Training: Mean Recognition Rate (%) for 13904 phonemes	Test: Mean Recognition Rate (%) for 4021 phonemes
KNN	83.84	72.01
FKNN	98.42	72.21

To conclude, in spite of some modest phoneme recognition rates, we can assert that there is an advantageous use of FkNN in the recognition approach proposed; this is mainly stated in term of a simple implementation and a rapid execution. Obviously, the mean recognition rate indicates that the approach is a reliable way and it can be compared with others recognition systems in particular with our previous work based on HMM published in [24].

V. Conclusions and Futures Works

We have achieved the objectives of this work via the application of the FkNN algorithm for Timit phoneme recognition to obtain a significant recognition performance.

The recognition approach consists of extracting mean MFCC vectors that characterize features in both training and test set. The FkNN is then applied by measuring Euclidian distance, fixing a 'k' nearest neighbors and attributing a membership degree. The maximum membership was ultimately used for making decision.

Results prove a promising approach that can play a significant role for the construction of a phoneme recognizer. As well, a considerable recognition rates were retained for several phonemes such as Stops, Fricatives, Nasals and Semi-vowels. A comparison of a crisp and fuzzy kNN reveals the profit of the fuzzy concept in our recognition approach.

Moreover, further experiment need to be done on the optimization the FkNN parameters; we plan to test different measures distance other than the Euclidian distance. Additionally, we will discover an optimal tool for electing the parameter 'k'. Another chief perspective is about the comparison of the standard kNN and the FkNN in phoneme recognition.

Also, we will assay further signal features such as LPCC and PLP coefficients and we will accomplish the phoneme recognition of all Timit dialects.

References

[1] L. R. Rabiner and B. H. Juang, "Speech Recognition: Statistical Methods," in Encyclopedia of language & linguistics (2nd ed.), pp. 1-18, 2006.

[2] L. Chergui, M. Kef and S. Chikhi, "New Hybrid Arabic Handwriting Recognizer," 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), pp. 319-325. Tunisia, 2012.

[3] M. Kef, L. Chergui and S. Chikhi, "Comparative Study of the Use of Geometrical Moments for Arabic Handwriting Recognition," 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), pp. 303-308, Tunisia, 2012.

[4] K. Samudravijaya, "Introduction to Automatic Speech Recognition," AU-KBC Research Centre, Chennai, 2011.

[5] D. Palaz, R. Collobert and M. D. Magimai, "Estimating Phoneme Class Conditional Probabilities from Raw Speech Signal using Convolutional Neural Networks," Interspeech, pp. 1766-1770, France, 2013.

[6] L. Bai, P. Jančovič, M. Russel and P. Weber, "Analysis of a low-dimensional bottleneck neural network representation of speech for modelling speech dynamics," Interspeech, Germany, 2015.

[7] A. Patiño-Saucedo, F .A. Sepulveda-Sepulveda and D .F. Gomez-Cajas, "Phoneme Recognition System Using Articulatory-Type Information," J. Tecciencia, vol 10, pp. 11-16, 2015.

[8] Y. Taoa, D. Chunhongb and S. Wangyangc, "Speech emotion recognition based on Fuzzy K-NN algorithm with fractionally spaced blind equalization," 2nd Workshop on Advanced Research and Technology in Industry Applications, Wartia, 2016.

[9] M. Kuskea, R. Rubiob, A. C. Romaina, J. Nicolasa and S. Marcob, "Fuzzy k-NN applied to moulds detection," J. Sensors & Actuators B: Chemical, pp. 52-60, 2005.

[10] V. Borges and W. Jeberson, "Fuzzy kNN Adaptation to Learning by Example in Activity Recognition Modeling," in Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics, vol. 1, pp. 29-36, India, 2015.

[11] P. Kasemsumran, S. Auephanwiriyakul and N. Theera-Umpon, "Face recognition using string grammar fuzzy K-nearest neighbor," 8th International Conference on Knowledge and Smart Technology, Thailand, 2016.

[12] F. A. El-Mouadib and A . F Abdalsaalam, "Comparison of crisp and fuzzy KNN classification algorithms," the international Arab Conference on Information Technology, Libya, 2010.

[13] J. M. Keller, M.R. Gray and J. A. Givens, "A Fuzzy K-Nearest Neighbor Algorithm," IEEE Transactions on Systems Man and Cybernetics. SMC-15, pp. 580- 585, 1985.

[14] D. T. Larose, "k-Nearest Neighbor Algorithm. Discovering Knowledge," in Data: An Introduction to Data Mining, 2005.

[15] G. Bhattacharya, K. Ghosh and A. Chowdhury, "An affinity-based new local distance function and similarity measure for kNN algorithm," Pattern Recognition Letters, vol. 33, pp. 356–363, 2012.

[16] J. Derrac, S. García and F. Herrera, "Fuzzy nearest neighbor algorithms: Taxonomy, experimental analysis and prospects," J. Information Sciences, vol. 260, pp. 98-119, 2014.

[17] The Linguistic Data Consortium https://catalog.ldc.upenn.edu/LDC93S1

[18] R. Latif, A. Dliou, S. Elouaham and M.L. Essi, "Experimental acoustic signal analyzed by the parametric and non parametric time-frequency representations," 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), pp. 706-711, Tunisia, 2012.

[19] I. Ben Fredj and K. Ouni, "A Novel Phonemes Classification Method Using Fuzzy Logic," J. Science Journal of Circuits, Systems and Signal Processing, pp. 1-5, 2013.

[20] A. Hmich, A. Badri, A. Sahel and M. Moughit, "Discriminating Coding applied to the Automatic Speaker Identification," 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), pp. 44-49. Tunisia, 2012.

[21] I. Ben Fredj and K. Ouni, "Effects of dynamic derivatives of speech signals on fuzzy phoneme recognition," 16ème édition du colloque COmpression et REprésentation des Signaux Audiovisuels, pp. 172-176, France, 2013.

[22] I. Thomas, I. Zukerman and B. Raskutti, "Extracting Phoneme Pronunciation Information from Corpora," Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning, pp. 175-183, 1998.

[23] A. K. Ghosh, "On optimum choice of k in nearest neighbor classification," J. Computational Statistics & Data Analysis, vol. 50, pp. 3113–3123, 2006.

[24] I. Ben Fredj and K. Ouni, "Phoneme Recognition using Hidden Markov Models: Evaluation with signal parameterization techniques," International Journal of Control, Energy and Electrical Engineering, vol. 1, pp. 57-61, 2014.