Speech Phoneme Classification by Intelligent Decision-Level Fusion

Fréjus A.A. Laleye, Eugène C. Ezin and Cina Motamed

Abstract This paper explores the decision fusion for the phoneme recognition problem through intelligent combination of Naive Bayes and Learning Vector Quantization (LVQ) classifiers and feature fusion using Mel-frequency Cepstral Coefficients (MFCC), Relative Spectral Transform—Perceptual Linear Prediction (Rasta-PLP) and Perceptual Linear Prediction (PLP). This work emphasizes optimal decision making from decisions of classifiers which are trained on different features. The proposed architecture consists of three decision fusion approaches which are weighted mean, deep belief networks (DBN) and fuzzy logic. We proposed a performance comparison on a dataset of an African language phoneme, Fongbe, for experiments. The latter produced the overall decision fusion performance with the proposed approach using fuzzy logic whose classification accuracies are 95.54 % for consonants and 83.97 % for vowels despite the lower execution time of Deep Belief Networks.

1 Introduction

Phoneme classification is an integrated process to phoneme recognition and an important step in automatic speech recognition. Since the 60s, very significant research progress related to the development of statistical methods and artificial intelligence techniques, have tried to overcome the problems of analysis and characterization of the speech signal. Among the problems, there is still the acoustic and linguistic

F.A.A. Laleye · E.C. Ezin
Unité de Recherche en Informatique et Sciences Appliquées,
Institut de Mathématiques et de Sciences Physiques, Université d'Abomey-Calavi,
BP 613 Porto-Novo, Abomey Calavi, Benin
e-mail: frejus.laleye@imsp-uac.org

F.A.A. Laleye · C. Motamed (✉)
Laboratoire d'Informatique Signal et Image de la Côte d'Opale,
Université du Littoral Côte d'Opale, 50 rue F. Buisson, BP 719,
62228 Calais Cedex, France
e-mail: motamed@lisic.univ-littoral.fr

J. Filipe et al. (eds.), *Informatics in Control, Automation and Robotics 12th International Conference, ICINCO 2015 Colmar, France, July 21-23, 2015 Revised Selected Papers*,
Lecture Notes in Electrical Engineering 383, DOI 10.1007/978-3-319-31898-1_4

specificity of each language. Considering the number of languages that exists, there were some good reasons for addressing phoneme recognition problems.

The aim of speech recognition is to convert acoustic signal to generate a set of words from a phonemic or syllabic segmentation of the sentence contained in the signal. Phoneme classification is the process of finding the phonetic identity of a short section of a spoken signal [11]. To obtain good recognition performance, the phoneme classification step must be well achieved in order to provide phoneme acoustic knowledge of a given language. Phoneme classification is applied in various applications such as speech and speaker recognition, speaker indexing, synthesis etc. and it is a difficult and challenging problem.

In this paper, we placed the phoneme recognition problems in a classification con- text from multiple classifiers. We dealt with the decision-level fusion from two different classifiers namely Naive Bayes and Learning Vector Quantization (LVQ). Since the 90s, the combining classifiers has been one of the most sustained research directions in the area of pattern recognition. Methods of decision-level fusion have been successfully applied in various areas such as the recognition and verification of signatures, the identification and face recognition or the medical image analysis. In automatic speech recognition, decision-level fusion was introduced to recognize phoneme, speech, speaker age and gender and to identify language with the best performance. The work we present in this paper deals with the phoneme recognition of Fongbe language which is an unressourced language. Fongbe is an African language spoken especially in Benin, Togo and Nigeria countries. It is a poorly endowed language which is characterized by a series of vowels (oral and nasal) and consonants (oral and nasal). Its recent written form consists of a number of Latin characters and the International Phoneoutic Alphabet. Scientific studies on the Fongbe started in 1963. In 2010, there was the first publication of Fongbe-French dictionary [3]. Since 1976, several linguists have worked on the language and many papers have been published on the linguistic aspects of Fongbe. Until today, these works have been aimed at the linguistic description of Fongbe, but very few works have addressed automatic processing with a computing perspective.

The idea behind this work is to propose a robust discriminatory system of consonants and vowels thanks to intelligent classifier combination based on decision-level fusion. To achieve this goal, we investigated on both methods of decision fusion namely the non-parametric method using weighted combination and parametric method using deep neural networks and a proposed adaptive approach based on fuzzy logic. The intelligent decision-level fusion used in this work to perform classification is carried out after the feature-level fusion of MFCC, Rasta-PLP, PLP coefficients applied to classifiers to represent the phonetic identity of each phoneme of the chosen language. In other words, the features were initially combined to produce coefficients as input variables to the classifiers. Experiments were performed on our Fongbe phoneme dataset and showed better performance with the proposed fuzzy logic approach. The rest of the paper is organized as follows. In Sect. 2, we briefly present the related works on phoneme recognition and decision fusion. Section 2.1 presents an overview on the proposed classification system. In Sect. 3, we describe the classifier methods and their algorithms. In Sect. 4, the proposed Fongbe phoneme classification is detailed and explained. Experimental results are reported in Sect. 5. In the same section we present a detailed analysis of the used performance parameters to evaluate the decision fusion methods. We finally conclude this paper in Sect. 6.

2 Related Works

This work deals with two different issues namely decision-level fusion from multiple classifiers and phoneme classification of a West Africa local language (Fongbe).

2.1 Overview on Phoneme Classification

Some of the recent research works related to phoneme classification applied to the world's languages are discussed as follows.

In [22], the authors proposed an approach of phoneme classification which performed better on TIMIT speech corpus, with warp factor value greater than 1. They have worked on compensating inter-speaker variability through Vocal tract length normalization multi-speaker frequency warping alternative approach. Finally, they compared each phoneme recognition results from warping factor between 0.74 and 1.54 with 0.02 increments on nine different ranges of frequency warping boundary. Their obtained results showed that performance in phoneme recognition and spoken word recognition was respectively improved by 0.7 % and 0.5% using warp factor of 1.40 on frequency range of 300–5000 Hz.

Phoneme classification is investigated for linear feature domains with the aim of improving robustness to additive noise [1]. In this paper, the authors performed their experiments on all phonemes from the TIMIT database in order to study some of the potential benefits of phoneme classification in linear feature domains directly related to acoustic waveform, with the aim of implementing exact noise adaptation of the resulting density model. Their conclusion was that they obtained the best practical classifiers paper by using the combination of acoustic waveforms with PLP $+ \triangle + \triangle\triangle$.

In [11], the authors integrated into phoneme classification a non-linear manifold learning technique, namely "Diffusion maps" that is to build a graph from the feature vectors and maps the connections in the graph to Euclidean distances, so using Euclidean distances for classification after the non-linear mapping is optimal. The experiments performed on more than 1100 isolated phonemes, excerpted from the TIMIT speech database, of both male and female speakers show that Diffusion maps allows dimensionality reduction and improves the classification results.

The work presented in [30] successfully investigates a convolutional neural network approach for raw speech signal with the experiments performed on the TIMIT and Wall Street Journal corpus datasets. Still on the TIMIT datasets, the authors in [37] focused their work on the robustness of phoneme classification to additive noise in the acoustic waveform domain using support vector machines (SVMs). The authors in [9] used a preprocessing technique based on a modified Rasta-PLP algorithm and a classification algorithm based on a simplified Time Delay Neural Network (TDNN) architecture to propose an automatic system for classifying the English stops [b, d, g, p, t, k]. And in [8], they proposed an artificial Neural Network architecture to detect and classify correctly the acoustic features in speech signals.

Several works have been achieved on the TIMIT dataset which is the reference speech dataset, but other works were performed on other languages than those included in the TIMIT dataset. We can cite, for example the following papers [19, 26, 28, 34], where the authors worked respectively on Vietnamese, Afrikaans, English, Xhosa, Hausa language and all American English phonemes.

A state of the art on the works related to Fongbe language stands out the works in the linguistic area. In [2], the authors studied how six Fon enunciative particles work: the six emphatic particles h...n "hence", sin "but", m "in", 1 "insist", lo "I am warning you", and n "there". Their work aimed at showing the variety and specificity of these enunciative particles. In these works [3, 20] listed in the Fongbe language processing, the authors introduced and studied grammar, syntax and lexicology of Fongbe.

In [18], the authors addressed the Fongbe automatic processing by proposing a classification system based on a weighted combination of two different classifiers. Because of the uncertainty of obtained opinions of each classifier due to the imbalance per class of training data, the authors used the weighted voting to recognize the consonants and vowels.

2.2 Decision-Level Fusion Methods

The second issue dealt with in this work is the decision fusion for optimal Fongbe phoneme classification. Combining decisions from classifiers to achieve an optimal decision and higher accuracy became an important research topic. In the literature, there are researchers who decided to combine multiple classifiers [6, 16, 33]. Other researchers worked on mixture of experts [14, 15].

In decision fusion methods, there are so-called non-parametric methods (classifiers outputs are combined in a scheme whose parameters are invariant) and the learning methods that seek to learn and adapt on the available data, the necessary parameters to the fusion. In speech recognition, several researchers successfully adopted the decision level fusion to recognize phoneme, speech, speaker age and gender and to identify language. For example, the authors in [24] performed decision level combination of multiple modalities for the recognition and the analysis of emotional expression. Some authors adopted non-parametric methods as weighted mean [13, 21, 27] and majority voting [7, 31]. Others adopted parametric methods as Bayesian inference [25, 32, 36] and Dempster-Shafer method [10].

In this work we adopted both methods to compare their performance in decision fusion of classifiers for an optimal phoneme classification of Fongbe language. First, we performed a weighted mean, which is a non-parametric method, to combine decisions. This method needs a threshold value chosen judiciously by experiment in the training stage. The second method we used is a parametric method with learning based on deep belief networks. Deep Belief Networks (DBNs) have recently shown impressive performance in decision fusion and classification problems [29]. In addition to these two methods we also used an adaptive approach based on fuzzy logic. Fuzzy logic is often used for classification problems and has recently shown a good performance in speech recognition [23]. Indeed, the limitations of the use of threshold value that requires weighted mean is that the value is fixed and does not provide flexibility to counter any variations in the input data. In order to overcome the limitations of the threshold based weighted mean which gives a hard output decision of which either "True" or "false" and the time that can be taken a training process of deep belief networks, we proposed a third approach based on fuzzy logic which can imitate the decision of humans by encoding their knowledge in the form of linguistic rules. Fuzzy logic requires the use of expert knowledge and is able to emulate human thinking capabilities in dealing with uncertainties.

3 Classification Methods and Algorithms

We detail in this section the algorithm implemented in the classification methods used for the discriminating system of consonants and vowels phonemes.

3.1 *Naive Bayes Classifier*

Naive Bayes is a probabilistic learning method based on the Bayes theorem of Thomas Bayes with independence assumptions between predictors. It appears in the speech recognition to solve the multi-class classification problems. It calculates explicitly the probabilities for hypothesis and it is robust to noise in input data. Despite its simplicity, the Naive Bayesian classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods. The Bayes classifier decides the class $c(x)$ of the input data x based on the Bayes rule:

$$p(c|x) = \frac{p(c,x)}{p(x)} \tag{1}$$

$$= \frac{p(c)p(x|c)}{\sum_{c'} p(c')p(x|c')} \tag{2}$$

where $p(c)$ is the prior probability of class c, and $p(x|c)$ is the class c-conditional probability of x.

Consider an example $X = \{x_1, x_2, \dots, x_n\}$
X is classified as the class $C = +$ if and only if,

$$F(X) = \frac{p(C = +|X)}{p(C = -|X)} \geq 1 \tag{3}$$

$F(X)$ is a Bayesian classifier.

Naive Bayes is the simplest form of Bayesian network, in which we assume that all attributes are independent given the class [38].

$$p(X|c) = p(x_1, x_2, \dots, x_n|c) = \prod_{i=1}^{n} p(x_i|c) \tag{4}$$

The naive Bayesian classifier is obtained by:

$$F_{nb}(X) = \frac{p(C = +|X)}{p(C = -|X)} \prod_{i=1}^{n} \frac{p(x_i|C = +)}{p(x_i|C = -)} \tag{5}$$

3.2 Learning Vector Quantization Classifier

Learning Vector Quantization (LVQ) is a supervised version of vector quantization. Networks LVQ were proposed by Kohonen [17] and are hybrid networks which use a partially supervised learning [5].

Algorithm LVQ method algorithm can be summarized as follows:

1. Initialize the weights $w_{ij}^{(1)}$ to random values between 0 and 1.
2. Adjust the learning coefficient $\eta(t)$
3. For each prototype p_i, find the neuron of the index i^* which has the weight vector $w_{i^*}^{(1)}$ closest to the p_i.
4. If the specified class at the network output for the neuron of the index i^* corresponds to the prototype of the index i, then do:

$$w_{i^*}^{(1)}(t+1) = w_{i^*}^{(1)}(t) + \eta(t)(p(t) - w_{i^*}^{(1)}(t)) \tag{6}$$

else

$$w_{i^*}^{(1)}(t+1) = w_{i^*}^{(1)}(t) - \eta(k)(p(t) - w_{i^*}^{(1)}(t)) \tag{7}$$

5. If the algorithm has converged with the desired accuracy, then stop otherwise go to step 2 by changing the prototype.

4 Proposed Phoneme Classification System

4.1 Overview of Classification System

Our intelligent fusion system is summarized in two modules which are each subdivided into submodules:

1. feature-level fusion and classification: the first module performs classification with Naive Bayes and LVQ classifier and produces outputs with the coefficients obtained after features fusion and which are applied as input. This module contains the submodules which are (i) signal denoising, (ii) feature extraction (MFCC, PLP, and Rasta-PLP), (iii) features fusion and classification with Naive Bayes and LVQ.
2. decision-level fusion and optimal decision making: the second module performs in parallel the decisions fusion with fuzzy approach that we proposed and the method with learning based on Deep Belief Networks.

Both modules are separated by an intermediate module which performs weighted mean calculation of classifiers outputs and contains the submodule which is (iv) standardization for classifiers decisions database. In this submodule, the outputs of the first module are combined to produce a single decision that is applied to the decision-level fusion module. The various steps are shown in Fig. 1.

4.2 Speech Feature Extraction

From phoneme signals we extracted MFCC, PLP and Rasta-PLP coefficients to perform the proposed adaptive decision fusion using Fuzzy approach and deep belief networks. The benefit of using these three types of coefficients is to expand the variation scale from input data of classification system. This enabled our system to learn more acoustic information of Fongbe phonemes. These three speech analysis techniques were initially allowed to train two classifiers and then put together to build the set of input variables to the decision fusion. Phoneme signals were split into frame segments of length 32 ms and the first 13 cepstral values were taken.

4.3 Decision Fusion Using Simple Weighted Mean

An intermediate step between the two steps was the normalization of output data of the first step. First, we calculated the weighted mean value of the two classifier outputs for each coefficient using the expression (8).

$$input_1 = \frac{S^{naivebayes} \times T^{naivebayes} + S^{lvq} \times T^{lvq}}{T^{naivebayes} + T^{lvq}} \tag{8}$$

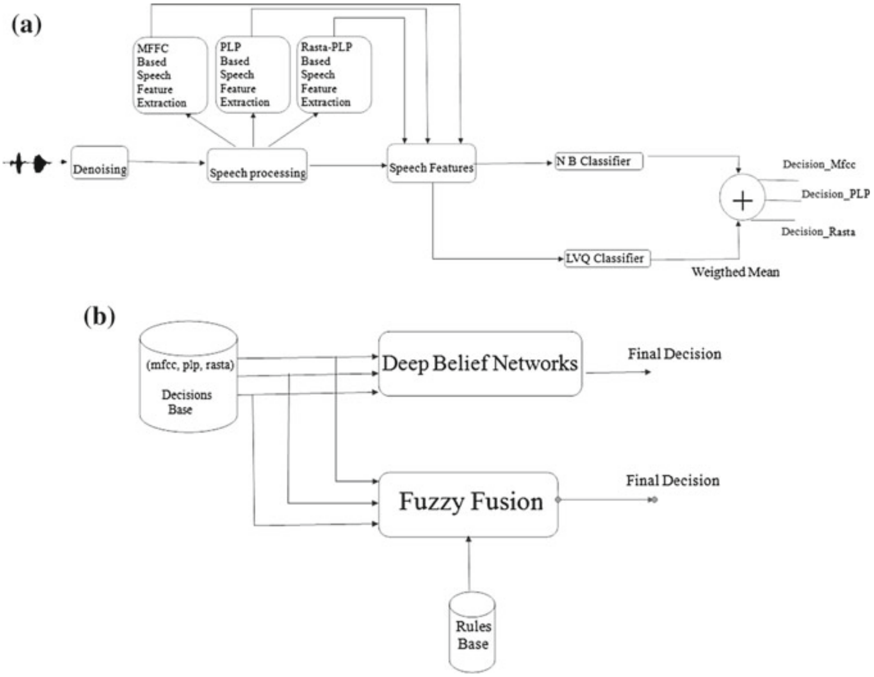

Fig. 1 Paradigm of our classification system. **a** Classification and standardization. **b** Decision fusion using fuzzy logic and deep belief networks

S^A represents the output of classifier A whereas τ^A represents the recognition rate of classifier A. Before applying fuzzy logic and neuronal technique to fuse the decisions of each classifier, we performed the output combination based on the simple weighted sums method using the threshold value obtained and given by Eq. 9.

$$\tau = -1.2 \sum_i C_i + 2.75(\sum_k w_k^1 \lambda_1 + \sum_k w_k^2 \lambda_2) \tag{9}$$

C_i: is the number of class i, w_k^1: weight of classifier k related to class 1, w_k^2: weight of classifier k related to the class 2, λ_1 and λ_2 are values that are 0 or 1 depending on the class. For example, for the consonant class: $\lambda_1 = 1$ and $\lambda_2 = 0$. The results are compared with fuzzy logic method and neuronal method to evaluate the performance of our phoneme classification system.

4.4 Fuzzy Logic Based Fusion

The Nature of the results obtained in the first step allows us to apply fuzzy logic on four membership functions. The inputs to our fuzzy logic system are MFCC, PLP

Table 1 Generated fuzzy rules

Rules no	Input			Output
	MFCC	Rasta	PLP	
1	Low	Low	Low	Consonant
2	Low	Low	Medium	Vowel
3	Low	Low	High	Consonant
4	Low	Medium	Low	Vowel
5	Low	High	Low	Consonant
6	Low	High	High	Consonant
7	Low	Very high	Low	Vowel
8	Low	Very high	Very high	Vowel
9	Medium	Low	Low	Vowel
10	Medium	Low	Very high	Vowel
11	Medium	Very high	Low	Vowel
12	Medium	Very high	Very high	Vowel
13	High	Low	Low	Consonant
14	High	Low	High	Consonant
15	High	High	Low	Consonant
16	High	High	High	Consonant
17	Very high	Low	Low	Vowel
18	Very high	Low	Medium	Vowel
19	Very high	Low	High	Consonant
20	Very high	Low	Very high	Vowel
21	Very high	Medium	Low	Vowel
22	Very high	Medium	Very high	Vowel
23	Very high	High	High	Consonant
24	Very high	Very high	Low	Vowel
25	Very high	Very high	Medium	Vowel
26	Very high	Very high	Very high	Vowel

and Rasta-PLP and the output obtained is the membership degree of a phoneme to a consonant or vowel class. The input variables are fuzzified into four complementary sets namely: *low, medium, high and very high* and the output variable is fuzzified into two sets namely: consonant and vowel. Table 1 shows the fuzzy rules which were generated after fuzzification.

First, the input data is arranged in an interval as [Xmin … Xmax]. The different membership functions were obtained by examining the local distribution of samples of both classes (see Fig. 2). Local distribution has induced four subsets according to the variation of the input data and the output is obtained depending on the nature of the data. For example, if we give MFCC, PLP and Rasta as input to the system, the consonant or vowel output is obtained according to the subsets of the input data.

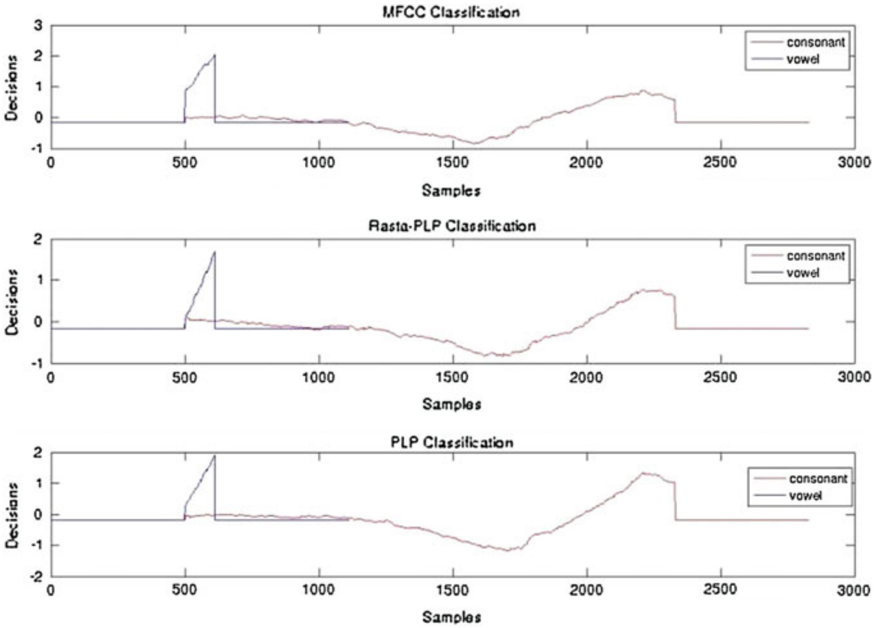

Fig. 2 *Top* Local distribution of decisions from MFCC coefficients classification, *Middle* local distribution of decisions from Rasta-PLP coefficients, *Bottom* local distribution of decisions from PLP coefficients

Because of the linearity of values in the subsets, a simple triangle curve (*trimf*) is used for low and medium membership functions and a trapeze curve (*trapmf*) is used for high and very high membership functions.

4.5 DBN Based Fusion

This method based on the use of deep belief networks (DBNs) requires a learning step for a good adaptation of the decisions to the system input. DBNs are multilayered probabilistic generative models which are constructed as hierarchies of recurrently connected simpler probabilistic graphical models, so called Restricted Boltzmann Machines (RBMs) [4, 12]. Every RBM consists of two layers of neurons, a hidden and a visible layer. Using unsupervised learning, each RBM is trained to encode in its weight matrix a probability distribution that predicts the activity of the visible layer from the activity of the hidden layer [29].

To perform the classifier for making of decision, we used the DBN parameters showed in Table 2

Table 2 DBN parameters

RBM layer 1	200 units
RBM layer 2	200 units
Learning rate	0.01
Training epochs	100
Batch size	8

Algorithms 1 and 2 summarize the different parts of our classifier implemented with Matlab. Function names give the idea about the operation they perform and sentences beginning with // represent comments. For example, final_decision_2 $\leftarrow dbnfusion(all\_input)$ means that the optimal decision given by DBN fusion is stored in final_decision_2.

Algorithm 1: Classification with Naive Bayes and LVQ

```
Data: Phoneme signals
Result: Decision of each classifier for each extraction technique.
signal denoising;
for $signal \in phoneme_database$ do
    signal←denoising(signal);
    base←put(signal)
end
Feature extraction;
for $signal \in base$ do
    m←mfcc_calculation(signal);
    p←plp_calculation(signal);
    r←rasta_calculation(signal);
    base_mfcc←put(m);
    base_plp←put(p);
    base_rasta←put(r);
end
training←put(m,p,r);
//Classification with Naive Bayes and LVQ;
for $i \leftarrow 1$ to $size(training)$ do
    if $i <= size(base\_mfcc)$ then
        $bayes\_mfcc\_decision \leftarrow bayes(training(i))$;
        lvq_mfcc_decision←lvq(training(i));
    end
    if $i > size(base\_mfcc)$ and $i <= size(base\_mfcc) + size(base\_plp)$ then
        $bayes\_plp\_decision \leftarrow bayes(training(i))$;
        lvq_plp_decision←lvq(training(i));
    end
    if $i > size(base\_mfcc) + size(base\_plp)$ and
    $i <= size(base\_mfcc) + size(base\_plp) + size(base\_rasta)$ then
        $bayes\_rasta\_decision \leftarrow bayes(training(i))$;
        lvq_rasta_decision←lvq(training(i));
    end
end
```

5 Experimental Results and Analysis

We present the different results obtained after training and testing with two classifiers and the results of decision fusion with fuzzy logic approach and deep belief networks. Experiments were performed on phonemes of the Fongbe language that we describe in the next section. Programming was done with Matlab in an environment which is Intel Core i7 CPU L 640 @ 2.13 GHz × 4 processor with 4 GB memory.

Algorithm 2: Decision fusion with Fuzzy logic and Deep belief networks

```
Data: Decision of each classifier for each extraction technique.
Result: Final Decision
//calculation of recognition rate;
for $j \leftarrow 1$ to $size(classes)$ and $k \leftarrow 1$ to $size(classifiers)$ do
    $\tau \leftarrow -1,2 \sum_i C_i + 2,75(\sum_k w_k^1 \lambda_1 + \sum_k w_k^2 \lambda_2)$;
end
//calculation of weighted mean values as input of fuzzy system;
for $l \leftarrow 1$ to 3 do
    $input_i \leftarrow \frac{S^{naivebayes} * \tau^{naivebayes} + S^{lvq} * \tau^{lvq}}{\tau^{naivebayes} + \tau^{lvq}}$;
    $all\_input \leftarrow put(input_i)$;
end
final_decision_1 ← $fuzzylogicsystem(all\_input)$;
final_decision_2 ← $dbnfusion(all\_input)$;
```

5.1 Speech Data Structure

The used speech dataset was obtained by recording different phonemes pronounced by foreigners and natives speakers with a recorder in various environments of real life. It contains 174 speakers whose ages are between 9 and 45 years, including 53 women (children and adults) and 119 men (children and adults). It is an audio corpus of around 4 h of pronounced phonemes which includes 4929 speech signals for all 32 phonemes. 80 % of speech signals in dataset is used to construct the training data and 20 % for the testing data.

5.2 Classification Results

LVQ parameters:

- number of hidden neurons: 60
- first class and second class percentage: 0.6 and 0.4
- learning rate: 0.005
- number of epochs: 750

Table 3 Training and testing results

Classifier	MFCC		Rasta-PLP		PLP	
	C_1	C_2	C_1	C_2	C_1	C_2
Training results						
Naive Bayes	88.66	51.53	90.43	59.17	88.2	68.25
LVQ	98.09	47.44	97.32	40.65	97.35	51.53
Testing results						
Naive Bayes	92.29	38.34	91.48	46.04	93.10	60.24
LVQ	98.78	24.95	98.58	21.70	97.97	20.89

Values are estimated in percentage

Normal distribution is used for Naive Bayes classification. Table 3 shows the training results and the testing recognition rate.

5.3 Decision Fusion Results of Classifiers

Table 4 presents the fusion results of the methods we used.

5.4 Performance Analysis

Several measures were developed to deal with the classification problem [35]. The values of True Positive (TP), True Negative (TN), False Positive and False Negative were calculated after decision fusion with the different used methods. These values were used to compute performance parameters like sensitivity (SE), specificity (SP), Likelihood Ratio Positive (LRP), Accuracy (Ac) and Precision (Pr). Three other important measures were used as evaluation metrics: F-measure, G-measure and execution time. F-mesure considers both the precision Pr and the sensitivity SE to compute the score which represents the weighted harmonic mean (precision and sensitivity). G-mean is defined by sensitivity and specificity and measures the balanced performance of learning between the positive class and the negative class.

Table 4 Results of decision fusion using fuzzy logic

Fusion methods	Consonant (%)	Vowel (%)
Weighted mean	99.73	54.02
Fuzzy logic	95.54	83.97
Deep belief networks	88.84	84.79

Table 5 Performance analysis

Parameters	Naive Bayes	LVQ	Using weighted mean	Using fuzzy logic	Using deep belief nets
SE	0.93	0.99	0.99	0.95	0.88
SP	0.60	0.25	0.38	0.84	0.86
LRP	2.36	1.32	1.60	5.94	6.28
LRN	0.12	0.04	0.03	0.06	0.14
Ac	0.77	0.62	**0.69**	**0.90**	**0.87**
Pr	0.70	0.57	0.62	0.86	0.88
F-measure	0.80	0.72	**0.76**	**0.90**	**0.88**
G-measure	0.75	0.50	**0.61**	**0.89**	**0.87**
Execution time (s)	–	–	**0.10**	**0.7**	**0.04**

Values in bold are emphasized for the performance comparison

Execution time measures the computation time of each fusion methods in the testing step.

We used the same dataset to evaluate the performance of Naive Bayes, LVQ and the decision fusion methods on consonants and vowels of Fongbe phonemes. Table 4 shows that by considering the balance of phoneme classes, decision fusion of classifiers based on fuzzy logic achieved better performance even if the approaches based on the weighted mean and deep belief networks classified respectively consonants and vowels better than fuzzy logic. We noticed that fuzzy logic approach combined efficiently the decisions and got the optimal decision, but with an execution time increased by sixty percent compared to DBN. The results in Table 5 show the highest performances of Fuzzy logic approach on Accuracy, F-measure and G-measure parameters which were the chosen metrics to evaluate the performance of the compared methods. The best performances obtained with fuzzy logic confirmed that adding extra expert knowledge improves decision making after decision combination made by multiple classifiers.

6 Conclusion

This paper evaluates the performance of three decision-level fusion methods by intelligent classifier combination in a speech phoneme classification problem. The performance evaluation was achieved with methods such as weighted mean, deep belief networks and fuzzy logic after combination of Naive Bayes and LVQ and feature-level fusion. The main idea is to make an optimal decision compared with the decisions obtained with each classifier. The results of the accuracy, F-measure and G-measure parameters achieved in Table 5, show the best performance with the proposed decision fusion using fuzzy logic which uses human reasoning. So, this paper highlights two main results: (i) the performance comparison of three decisions fusion methods in a phoneme classification problem with multiple classifiers and (ii) the proposal of a robust Fongbe phoneme classification system which incorporates a fusion of Naive Bayes and LVQ classifiers using fuzzy logic approach. This proposal builds on the performance achieved by our fuzzy logic-based approach compared to DBN-based approach and especially because of the limitations of the fixed threshold value in weighted combination. Future works include automatic speech segmentation into syllable units and an automatic continuous speech recognition based on speech phoneme classification.

References

1. Ager, M., Cvetkovic, Z., Sollich, P.: Phoneme classification in high-dimensional linear feature domains. Comput. Res. Repository (2013)
2. Agoli-Agbo, E.O., Bernard, C.: Les particules nonciatives du fon. Institut national des langues et civilisations orientales, Paris, 1st edition (2009)
3. Akoha, A.B.: Syntaxe et lexicologie du fon-gbe: Bénin. Ed. L'harmattan, p. 368 (2010)
4. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layer-wise training of deep networks. In: Advances in Neural Information Processing Systems (2006)
5. Borne, P., Benrejeb, M., Haggege, J.: Les rseaux de neurones, présentation et applications. TECHNIP Editions, p. 90 (2007)
6. Cho, S.-B., Kim, J.: Combining multiple neural networks by fuzzy integral and robust classification. IEEE Trans. Syst. Man Cybern. 380–384 (1995)
7. Corradini, A., Mehta, M., Bernsen, N., Martin, J., Abrilian, S.: Multimodal input fusion in humancomputer interaction. In: NATO-ASI Conference on Data Fusion for Situation Monitoring, Incident Detection, Alert and Response Management (2003)
8. Esposito, A., Ezin, E., Ceccarelli, M.: Preprocessing and neural classification of english stop consonants [b, d, g, p, t, k]. In: The 4th International Conference on Spoken Language Processing, pp. 1249–1252. Philadelphia (1996)
9. Esposito, A., Ezin, E., Ceccarelli, M.: Phoneme classification using a rasta-PLP preprocessing algorithm and a time delay neural network: performance studies. In: Proceedings of the 10th Italian Workshop on Neural Nets, pp. 207–217. Salerno (1998)
10. Foucher, S., Laliberte, F., Boulianne, G., Gagnon, L.: A dempster-shafer based fusion approach for audio-visual speech recognition with application to large vocabulary french speech. In: IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1 (2006)
11. Genussov, M., Lavner, Y., Cohen, I.: Classification of unvoiced fricative phonemes using geometric methods. In: 12th International Workshop on Acoustic Echo and Noise Control. Tel-Aviv, Israel (2010)
12. Hinton, G., Osindero, S., Teh, Y.: A fast learning algorithm for deep belief nets. Neural Comput. **18**, 1527–1554 (2006)
13. Iyengar, G., Nock, H., Neti., C.: Audio-visual synchrony for detection of monologue in video archives. In: IEEE International Conference on Multimedia and Expo, vol. 1, pp. 329–332 (2003)
14. Jacobs, R.: Methods for combining experts's probability assessments. Neural Comput. 867–888 (1995)
15. Jacobs, R., Jordan, M., Nowlan, S., Hinton, G.: Adaptive mixture of local experts. Neural Comput. 79–87 (1991)
16. Kittler, J., Hatef, M., Duin, R., Matas, J.: On combining classifiers. IEEE Trans. Patt. Anal. Mach. Intell. 226–239 (1998)
17. Kohonen, T.: An introduction to neural computing. Neural Netw. **1**, 3–16 (1988)

18. Laleye, F.A.A., Ezin, E.C., Motamed, C.: Weighted combination of naive bayes and lvq classifier for fongbe phoneme classification. In: Tenth International Conference on Signal-Image Technology and Internet-Based Systems, pp. 7–13, Marrakech. IEEE (2014)
19. Le, V.-B., Besacier, L.: Automatic speech recognition for under-resourced languages: Application to vietnamese language. In: IEEE Transactions on Audio, Speech, and Language Processing, pp. 1471–1482. IEEE (2009)
20. Lefebvre, C., Brousseau, A.: A grammar of fonge, de gruyter mouton, p. 608 (2001)
21. Lewis, T.W., Powers., D.M.: Improved speech recognition using adaptive audio-visual fusion via a stochastic secondary classifier. Int. Symp. Intell. Multimedia Video Speech Process. **1**, 551–554 (2001)
22. Lung, J.W.J., Salam, M.S.H., Rehman, A., Rahim, M.S.M., Saba, T.: Fuzzy phoneme classification using multi-speaker vocal tract length normalization. IETE Technical Review, London, 2nd edn (2014)
23. Malcangi, M., Ouazzane, K., Patel, K.: Audio-visual fuzzy fusion for robust speech recognition. In: The 2013 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE, Dallas (2013)
24. Metallinou, A., Lee, S., Narayanan, S.: Decision level combination of multiple modalities for recognition and analysis of emotional expression. In: IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 2462–24665 (2010)
25. Meyer, G., Mulligan, J., Wuerger, S.: Continuous audio-visual digit recognition using n-best decision fusion. Inf. Fusion **5**, 91–101 (2004)
26. Mugler, E.M., Patton, J.L., Flint, R.D., Wright, Z.A., Schuele, S.U., Rosenow, J., Shih, J.J., Krusienski, D.J., Slutzky, M.W.: Direct classification of all american english phonemes using signals from functional speech motor cortex. J. Neural Eng. (2014)
27. Neti, C., Maison, B., Senior, A., Iyengar, G., Decuetos, P., Basu, S., Verma., A.: Joint processing of audio and visual information for multimedia indexing and human-computer interaction. In: Sixth International Conference RIAO, pp. 294–301. France, Paris (2000)
28. Niesler, T., Louw, P.H.: Comparative phonetic analysis and phoneme recognition for afrikaans, english and xhosa using the african speech technology telephone speech database. S. Afr. Comput. J. 3–12 (2004)
29. O'Connor, P., Neil, D., SC, L., Delbruck, T., Pfeiffer, M.: Real-time classification and sensor fusion with a spiking deep belief network. Front. Neurosci. (2013)
30. Palaz, D., Collobert, R., Magimai.-Doss, M.: End-to-end phoneme sequence recognition using convolutional neural networks. Idiap-RR (2013)
31. Pfleger, N.: Context based multimodal fusion. In: ACM International Conference on Multimodal Interfaces, pp. 265–272 (2004)
32. Pitsikalis, V., Katsamanis, A., Papandreou, G., Maragos, P.: Adaptive multimodal fusion by uncertainty compensation. In: Ninth International Conference on Spoken Language Processing, vol. 7, pp. 423–435. Pittsburgh (2006)
33. Rogova, G.: Combining the results of several neural networks classifiers. Neural Netw. 777–781 (1994)
34. Schlippe, T., Djomgang, E.G.K., Vu, N.T., Ochs, S., Schultz, T: Hausa large vocabulary continuous speech recognition. In: The third International Workshop on Spoken Languages Technologies for Under-resourced Languages. Cape-Town (2012)
35. Wang, S., Yao, X.: Diversity analysis on imbalanced data sets by using ensemble models. IEEE Symp. Comput. Intell. Data Min. 324–331 (2009)
36. Xu, H., Chua, T.: Fusion of av features and external information sources for event detection in team sports video. ACM Trans. Multimedia Comput. Commun. Appl. **2**, 44–67 (2006)
37. Yousafzai, J., Cvetkovic, Z., Sollich, P.: Tuning support vector machines for robust phoneme classification with acoustic waveforms. In: 10th Annual conference of the International Speech Communication Association, pp. 2359–2362. England (2009). ISCA-INST SPEECH COMMUNICATION ASSOC
38. Zhang, H.: Exploring conditions for the optimality of nave bayes. IJPRAI **19**, 183–198 (2005)