Automatic Fongbe Phoneme Recognition From Spoken Speech Signal

Fréjus A. A. Laleye[1,2], Eugène C. Ezin[2] and Cina Motamed[1]
[1]*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*
[2]*Unité de Recherche en Informatique et Sciences Appliquées, Institut de Mathématiques et de Sciences Physiques, Université d'Abomey-Calavi, BP 613 Porto-Novo, Benin*

Keywords: Formant Analysis, Fuzzy Logic, Deep Belief Networks, Phoneme Recognition, Continuous Speech Segmentation, Fongbe Language.

Abstract: This paper reports our efforts toward an automatic phoneme recognition for an under-resourced language, Fongbe. We propose a complete recipe of algorithms from speech segmentation to phoneme recognition in a continuous speech signal. We investigated a strictly fuzzy approach for simultaneous speech segmentation and phoneme classification. The implemented automatic phoneme recognition system integrates an acoustic analysis based on calculation of the formants for vowel phonemes and calculation of pitch and intensity of consonant phonemes. Vowel and consonant phonemes are obtained at classification. Experiments were performed on Fongbe language (an African tonal language spoken especially in Benin, Togo and Nigeria) and results of phoneme error rate are reported.

1 INTRODUCTION

Phoneme recognition is an integrated process in an Automatic Speech Recognition (ASR). Most large vocabulary automatic speech recognition systems use several interconnected layers of recognition for optimum performance. Phonemes are perhaps the most common sub-word modules used in ASR systems (Baghdasaryan and Beex, 2011). Many methods exist for phoneme recognition which are introduced for generating of phonetic symbols from the extracted feature vectors and calculated on speech signals. These are methods based on approaches such as Artificial Neural Network (ANN) (Palaz et al., 2013; Yousafzai et al., 2009; chwarz et al., 2006), Hidden Markov Models (HMM) (Young, 2008; marani et al., 2009), Support Vector Machine (SVM) (Solera-Urena et al., 2007) or some hybrid methods (Trentin and Gori, 2007; Anapathy et al., 2009).

This paper focuses on the development of an automatic phoneme recognition from continuous speech using several interconnected layers starting from the continuous speech segmentation to phoneme recognition. The overall system is the fusion of the recognition layer investigated in this work, the segmentation and the classification layers investigated respectively in (Laleye et al., 2015b) and (Laleye et al., 2015a) which were previous works on Fongbe language.

Figure 1 presents an overview of the phoneme recognition system. It includes the segmentation (i), classification (ii) and recognition (iii) layers as explained above and the processing order. To realize (i), we used the time domain approaches to generate features for detecting phoneme segments and fuzzy logic approach for a matching phase performed through supervised learning. Thus, (i) provides the phoneme segments with an accurate in the definition of boundaries. The Layer (ii) consists of a discriminatory system of consonants and vowels using an intelligent classifier combination based on decision-level fusion. It produces phonemes that are either consonants or vowels whose phonetic identity is recognized at (iii). Layer (iii) is based on acoustic analysis of Fongbe phonemes for identifying with more precision the phonetic content of the phoneme to its input. Time-domain features and frequency-domain features are respectively extracted from the speech signal to perform the segmentation and classification tasks. Frequency-domain features are computed on phoneme segments in step 2 order to constitute the classification system inputs. In step 5, we calculated the formant frequencies to facilitate the recognition of consonants or vowels.

Experiments were performed on a read speech corpus which contains 3117 utterances in Fongbe. we calculated the phone error rate score according to

Laleye, F., Ezin, E. and Motamed, C.
Automatic Fongbe Phoneme Recognition From Spoken Speech Signal.
DOI: 10.5220/0006004101020109
In *Proceedings of the 13th International Conference on Informatics in Control, Automation and Robotics (ICINCO 2016) - Volume 1*, pages 102-109
ISBN: 978-989-758-198-4

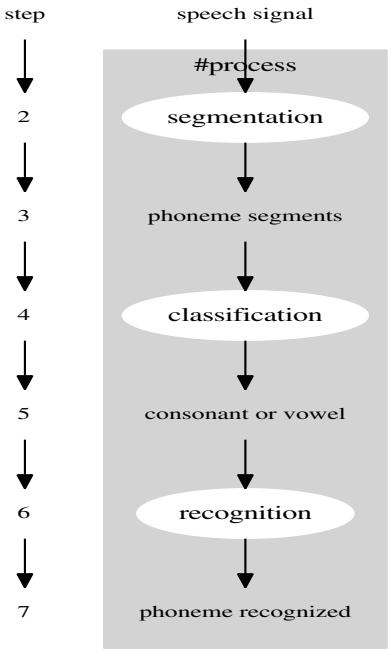


Figure 1: Overview of our automatic phoneme recognition.

the different DBN model used for recognizing each phoneme in its subclass. Experiments have shown a significant improvement in the results given by phone error rates (PER).

The remainder of this paper is organized as follows. The next section describes the automatic text-independent segmentation of Fongbe continuous speech. Section 3 describes how Fongbe phonemes are classified into consonant or vowel classes. Section 4 focuses on phoneme recognition system. Section 5 presents and comments the experimental results of PER that we obtained. Section 6 concludes this paper with results.

2 AUTOMATIC TEXT-INDEPENDENT SEGMENTATION OF FONGBE CONTINUOUS SPEECH

In this section, We present the phoneme segmentation method (preliminary work on Fongbe) used in this phoneme recognition work. we summarize this method in (Laleye et al., 2015b) to perform a text-independent segmentation of Fongbe continuous speech. The segmentation task was performed from non-linear speech analysis using fuzzy logic approach. It is based on the calculation of time domain features such as short-term energy, zero crossing rate

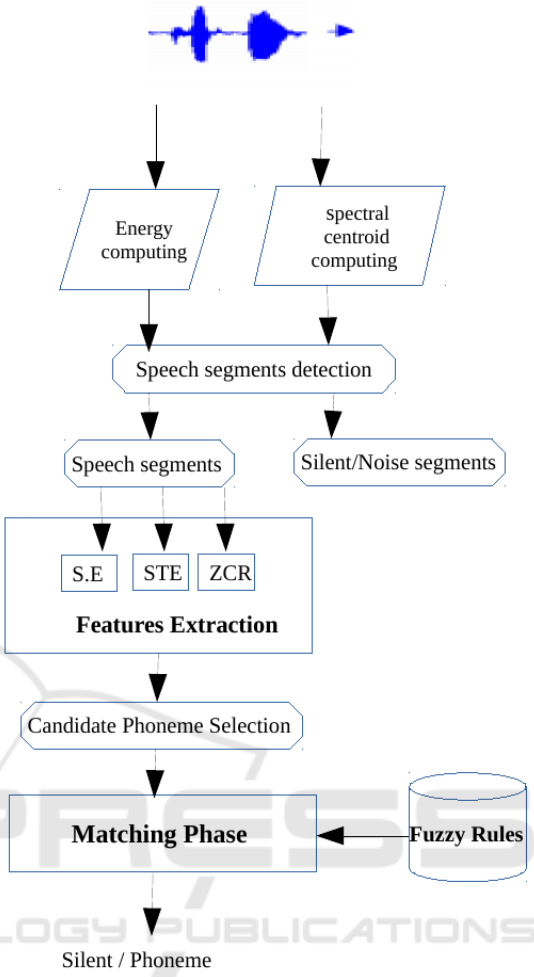


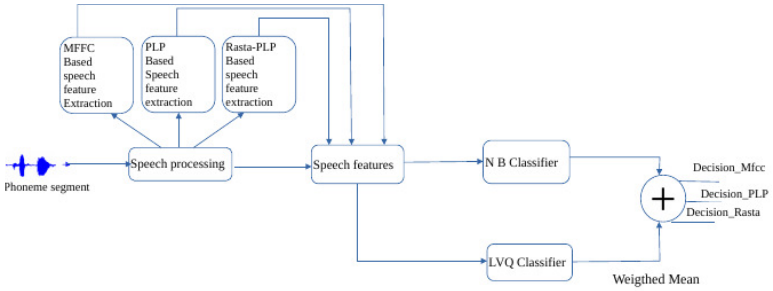
Figure 2: Flow diagram of the segmentation algorithm.

and the singularity exponents in each point of signal. As part of our work, we modified the calculation of singularity exponents by reducing the frame sizes in the scale-dependent functional $h(t)$ in order to detect only the boundaries between the phoneme segments in speech signal. For more details please refer to (Laleye et al., 2015b). The used algorithm for automatic text-independent segmentation is summarized in Figure 2 and includes the following tasks:

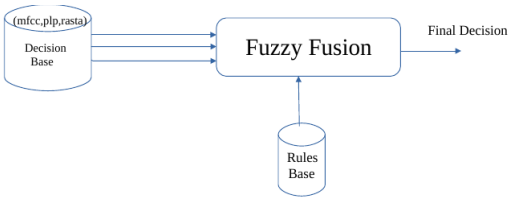
a - *Removal the silence from speech signal: a method based on the signal energy and the spectral centroid is used to remove the silence areas in the audio signal.*

b - *Singularity exponents Computation: local distribution of singularity exponents has been exploited for analyzing the temporal dynamics of speech segments previously obtained. This leads to the segment candidates.*

c - *Short-time energy and zero crossing rate calcu-*



(a) Classification and normalization.



(b) Decision fusion using fuzzy logic.

Figure 3: Paradigm of our classification system.

lation: the features are computed in each candidate segment generated by the local analysis of singularity exponents.

d - *Fuzzy rules application: fuzzy rules have been generated for the matching phase to improve the accuracy in the phoneme or syllable segments detection and the boundaries of beginning and the end.*

The phoneme segments are properly detected at the matching phase. The matching phase is performed by using fuzzy logic model which consists of a number of conditional *"if-then"* rules. The matching phase inputs are singularity exponents (SE), short-term energy (STE) and zero crossing rate (ZRC) and the output is the membership degree of silence and phoneme. The input variables are fuzzified into three complementary sets namely: *low, medium, high* and the output variable is fuzzified into two sets: silent and phoneme. The authors obtained for the different coefficients considering the features values:

- STE: low - medium - high
- ZCR: low - medium - high
- SE: low - medium - high

The fuzzy rules generated through supervised learning are presented in Table2. In this table, *x* is a variable that can take the value *low*, *medium* or *high*.

The different membership functions were obtained by examining the local distribution of each calculated feature. Local distribution has induced three subsets according to the variation of input data and output is obtained depending on the nature of data.

Table 1: Generated fuzzy rules.

Rules No	Input STE	ZCR	SE	Output
1	low	x	x	silent
2	x	low	x	silent
3	x	x	low	silent
4	medium	medium	medium	phoneme
5	medium	medium	high	phoneme
6	x	high	x	silent
7	high	medium	medium	phoneme
8	high	medium	high	phoneme

Because of the linearity of values in the subsets, a simple triangle curve (*trimf*) is used for the membership functions.

3 FONGBE PHONEME CLASSIFICATION

The phoneme classification system (Laleye et al., 2015a) consist of three modules which are each subdivided into submodules. The first module performs classification with Naive Bayes and Learning Vector Quantization (LVQ) classifier and produces outputs with the coefficients applied as input. It contains the submodules which are (i) feature extraction (MFCC, PLP, and Rasta-PLP), (ii) classification with Naive Bayes and LVQ. The second module performs weighted mean calculation of classifiers outputs and contains the submodule which is (iii) normalization

for classifiers decisions database. The last module performs the decision fusion with fuzzy approach. Figure 3 shows the various steps of classification.

Nature of the results obtained in the first step allows to apply fuzzy logic on four membership functions. The inputs to fuzzy logic system are MFCC, PLP and Rasta-PLP and the output obtained is the membership degree of a phoneme to 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 2 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. 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. 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.

This layer described in this section provides a classification in consonant or vowel of phoneme segments obtained in the first layer. The following section explain how our system recognize definitively the phonemes contained in a spoken speech.

Table 2: Generated fuzzy rules. *x* is a variable that can take the value *low*, *medium* or *high*.

Rules No	Input			Output
	mfcc	rasta	plp	
1	low	low	low	consonant
2	low	low	medium	vowel
3	x	x	high	consonant
3	x	x	very high	vowel
4	low	medium	low	vowel
5	low	high	low	consonant
6	low	very high	low	vowel
7	medium	low	low	vowel
8	medium	very high	low	vowel
9	high	low	low	consonant
11	high	high	low	consonant
12	very high	low	low	vowel
13	very high	low	medium	vowel
14	very high	medium	low	vowel
15	very high	very high	low	vowel
16	very high	very high	medium	vowel

4 PHONEME RECOGNITION

This section describes the process of the third layer, which is to recognize each phoneme included in the speech on the basis of the formant frequencies. We established an acoustic analysis of the phonemes by calculating the formant F1, F2 and F3 from vowels and the fundamental frequency and intensity from consonants. The recognizer was therefore trained with a Deep Belief Network (DBN) and a fuzzy logic system with the features calculated for recognizing each of the consonants and vowels. The acoustic analysis by calculating the frequencies allowed us to represent the phonemes in a tree according to the values of the formant, pitch and intensity. Figure 4 shows the tree of the categorization of consonants and vowels

4.1 Formant Analysis of Vowels

Articulatory configuration of each vowel lead to specific formant values corresponding to the shape taken by the vocal tract for each vowel. The shape will induce a resonant frequency in production of vowels: higher its size is, lower its frequency is. The first three formant have a wide variation in the acoustic configuration of each vowel. The mandible aperture (**F1**) varies from $311Hz$ (front. / i/) to $681Hz$ (cent. / a/) for oral vowels and from $315Hz$ (front. /ĩ/) to $610Hz$ (cent. /ã/) for nasal vowels. The second formant (**F2**) which indicates the tongue movement in front or rear position takes values from $697Hz$ (back. / u/) to $2238Hz$ (front. / i/) for oral vowels and from $406Hz$ (back. /ũ/) to $1630Hz$ (back. /ĩ/). The third formant influenced by the rounding of the lips has values that vary between $2696Hz$ (back. / ɔ/) and $3174Hz$ (front. / ɔ/) for oral vowels and between $711Hz$ (back. /ɔ̃/) and $2870Hz$ (front. /ɛ̃/).

In **F1**-level, oral vowels (back. and front.) / i/, / u/, / e/ and / o/ have lower values while oral vowels (back. and front.) / ɛ/, / ɔ/ and the central oral vowels / a/ have the highest values. In each aperture level, a front vowel has a greater **F1** than a back vowel. Vowels / i/ and / a/ are respectively the low and high extremities in the frequency range of the mandible aperture (**F1**). These remarks are also observed with nasalized vowels. This indicates that front vowels are / i e ɛ i̱ ɛ̱/ and the listed back vowels are / u o ɔ u̱ ɔ̱/. In **F2**-level, front vowels have higher frequencies compared to back vowels that are realized with low frequencies. The only difference between the vowels from the same class (front and back) is the value of the third formant **F3**. **F3** of / u/ ($F3 = 3083Hz$) and /u̱/ ($F3 = 2720Hz$) has a lower value than the value of the vowels / i/ ($F3 = 3174Hz$) and /i̱/ ($F3 = 2738Hz$), thereby justifying the rounding of the lips during the production of vowels / u u̱/. Table 3 summarizes

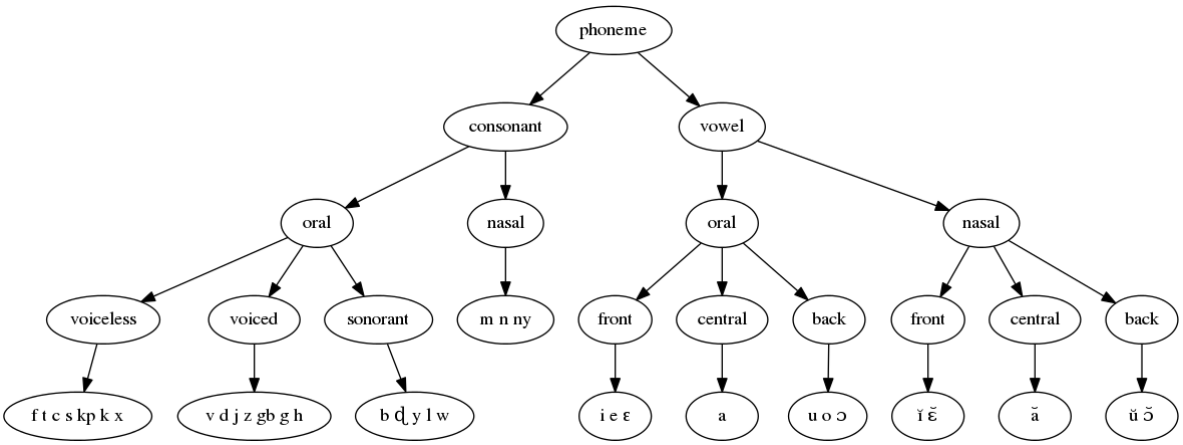


Figure 4: Tree of consonants and vowels.

Table 3: Mean values of formants F1, F2 et F3 for each phoneme of Fongbe vowel system.

			F1-*average*	F2-*average*	F3-*average*
Oral	front	i	311	2238	3174
		e	353	2042	3096
		ɛ	530	1255	3081
	back	u	326	697	3083
		o	380	822	2727
		ɔ	555	1051	2696
	cent.	a	681	1404	2789
Nasal	front	ḭ	315	1630	2738
		ɛ̰	405	1362	2870
	back	ṵ	325	406	2720
		ɔ̰	501	705	2711
	cent.	a̰	610	1015	2769

the averages of formant frequencies calculated on all phonemes in the data corpus.

4.2 Fongbe Consonant System

Fongbe consonant system is composed of twenty-two (22) phonemes grouped into two (2) classes (oral and nasal consonants). Unlike the acoustic configuration of vowels that is characterized by the presence of formant frequencies, the acoustic analysis of consonants consisted in calculating of two acoustic markers: fundamental frequency (pitch) that materializes the periodic vibration of the vocal cords and intensity that materializes the sub-glottis pressure. Usually, there are three main classes for acoustic analysis of consonants: plosives, fricatives and sonorants. The consonants from French are investigated in the next plosives, fricatives and consonant vocalic. In Fongbe, in addition to occlusive, fricative and consonant vocalic, there are two other classes that are nasal consonants and semi-consonants.

- The Occlusives from Fongbe have the same configuration than most other languages, including French. We distinguish sound and voiceless stops, which are mainly composed of voiced (/ d g gb j/) and unvoiced (/ b d/) oral consonants. They are characterized by the fundamental frequency (**f0**) which vary averagely between the voiced consonants (263Hz for / j/ and 279Hz for / gb/). The unvoiced consonants are realized with lower pitch than voiced consonants except / d/ and / j/. The fundamental frequencies (pitch) calculated over all speakers in data corpus are shown in Figure 5. The smallest average voice intensity (6) during the isolated pronunciation of voiced stops is obtained with the voiced consonant / d/ (56db) while the higher is obtained with the unvoiced / b/ (60db) voiceless Occlusives of Fongbe are composed of some voiceless oral consonants such as / c k kp t/ and also / p/ which is not considered a phonemic consonant but only found in some words borrowed from other languages. The voice intensity during the pronunciation of voiceless stop varies between 53db (for / c/) and 59db (for / p/).
- Articulatory configuration fricatives indicates the presence of noise throughout their pronunciation. As occlusives, there are sound fricatives (peri-

odic) and the voiceless fricatives (aperiodic) in Fongbe. The sound fricatives are composed of sound oral consonants / h v z/ and the voiceless fricatives are composed of voiceless oral consonants / f s x/. The fundamental frequency of / h/ (250 Hz) is lower than periodic fricatives / z/ (293 Hz) and / v/ (286 Hz). Voiceless fricatives are realized with low intensity values compared to the sound fricatives whose peaks are observed respectively with the consonants / x/ (54 db) and / z/ (58 db).

- There are three nasals consonants in Fongbe such as / m/, / n/ and / ŋ/. They are comparable to voiced stops in an articulatory viewpoint. They are realized with the lowest fundamental frequencies compared to voiced stops. Compared to other consonants from Fongbe, the nasal consonants are produced with the lowest intensity of voice and form, with the semi-consonants (/ w/ and / y/) the briefest in the pronunciation. The realization of semi-consonants are performed with almost the same periodic vibration rhythms of the vocal cords (pitch values) and the same pressure subglottic (intensity) than the voiced stops.

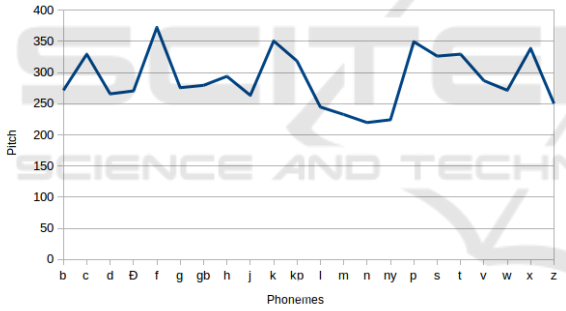


Figure 5: Pitch per consonfront.

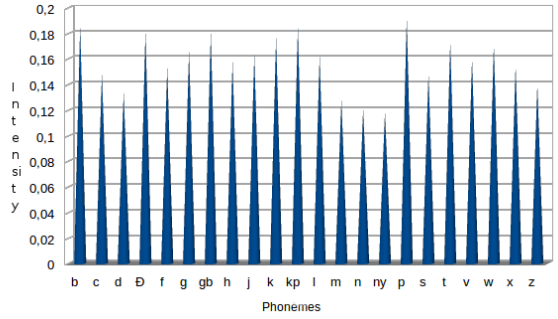


Figure 6: Intensity of each consonfront.

4.3 Baseline Phoneme Recognition System

We used two different approaches to perform the process of the last layer of our recognition system: fuzzy logic and DBN approaches. The Computation of the vowel formants, the fundamental frequency and intensity of consonants provide additional expert knowledge which can allow us to apply a fuzzy logic system to effectively recognize consonant or vowel phonemes (the inputs of this layer). Then, we used several DBNs in a hierarchical architecture based on the phoneme classification (Figure 4) to achieve recognition. Our recognition system consists of a mixture of fuzzy rules applied for recognizing subclasses of phonemes and DBN structures for recognizing the phonemes contained in each subclass. Thus, initially we used the fuzzy logic model which consist of a number of conditional "*if-then*" rules. When the classification layer provides a consonant to the recognition layer, the inputs of the fuzzy logic system are fundamental frequency and intensity and the output is the membership degree of voiceless, voiced, sonorant and nasal. In the case of a vowel as input to the third layer, the inputs are the formants **F1**, **F2** and **F3** and the output is the membership degree of oral front, oral central, oral back, nasal front, nasal central and nasal back. For a consonant as an input, the fundamental frequency (**f0**) is fuzzified into four complementary sets namely: *low, medium, high, very high* and intensity is fuzzified into three complementary sets namely: *low, medium, high*. For a vowel, each formant is fuzzified into three complementary sets namely: *low, medium, high*. The linearity of computed features induced the use of a simple triangle curve (trimf) for low, medium, high and very high membership functions. The fuzzy rules generated through supervised learning are presented in Table4 and Table5.

Table 4: The generated fuzzy rules applied to vowel for identifying its subclass.

Rules No	Input F1	F2	F3	Output
1	low	low	low	oral front
2	low	low	medium	nasal front
3	low	low	high	oral front
4	low	high	medium	nasal back
5	low	high	high	oral back
6	medium	low	low	nasal front
7	medium	medium	high	nasal back
8	high	low	low	oral front
9	high	low	medium	nasal central
10	high	medium	medium	oral central
11	high	medium	high	oral back

After having detected the subclasses, we perform a learning step based on the use of Deep Belief Network. We trained 10 DBN models whose the neuron number in the output layer varies according on

Table 5: The generated fuzzy rules applied to consonant for identifying its subclass.

Rules No	Input **f0**	Input **intensity**	Output
1	low	medium	sonorant
2	low	high	nasal
3	medium	medium	voiced
4	medium	high	sonorant
5	high	low	voiced
6	high	medium	voiced
7	very high	low	voiceless
8	very high	medium	voiced

Table 6: DBN parameters.

	Model-256 units	Model-512 units	Model-1024 units
RBM Layer 1	256 units	512 units	1024 units
RBM Layer 2	256 units	512 units	1024 units
RBM Layer 3	256 units	512 units	1024 units
RBM Layer 4	256 units	512 units	1024 units
Learning rate	0.01	0.01	0.01
Training Epochs	160	160	160
Batch size	8	8	8

the number of elements in a subclass. We used three different architectures to compare the recognition performance and impact of the layer number of DBN models on the performance of our recognition system. This is also motivated by differences in the change of phoneme subclasses. Table 6 summarizes the parameters used for each DBN model;

4.4 Experiments Results

Experiments were performed with Matlab in an environment which is Intel Core i7 CPU L 640 @ 2.13GHz $\times$ 4 processor with 4GB memory.

4.5 Experimental Data

The complete recipe of algorithms proposed in this work has been built for an automatic recognition of Fongbe phonemes contained in a continuous speech. Fongbe language is the majority language of Benin, which is spoken by more than 50% of Benin's population, including 8 million speakers and also spoken in Nigeria and Togo. Fongbe is an under-resourced language which is characterized by a series of vowels (oral and nasal) and consonants (unvoiced, fricatives). By excluding compound words and derived words, the words of Fongbe language can be grouped into monosyllabic (V and CV), into bisyllabic (VCV; CVV; CVCV and VV) and trisyllable (VCVCV and CVCVCV). It is written officially in Benin with an alphabet derived from the Latin writting since 1975. It has a complex tonal system, with two lexical tones, high and low, which may be modified by means of tonal processes to drive three further phonetic tones: rising low-high, falling high-low and mid (Lefebvre and Brousseau., 2001). The Fongbe's vowel system is well suited to the vocalic timbre as it was designed by the first Phoneticians. It includes twelve timbres: 7 oral vowels with 4 degrees of aperture and 5 nasal vowels with 3 degrees of aperture. Its consonant system includes 22 phonemes in 7 orders and 4 series. The Fongbe speech corpus is divided into a train and test sets which respectively contain 2307 and 810 sentences uttered by 56 speakers whose ages are between 9 and 45 years. It contains for the full database approximately 12 thousand words and more than 96 thousand phonemes.

4.6 RESULTS

The best performing setting of DBN models (four hidden layer architectures) was used to compute the phone error rate (PER) for the core test set. PER is the standard evaluation metric for phoneme recognition systems. It measures the difference between the phoneme string returned by the recognizer and the correct reference transcription. The distance between the two phoneme strings is computed by the classical minimum edit distance algorithm (Huang et al., 2001). Its formula (Equation 1) is based on the number of phoneme substitutions (SD), insertions (IE) and deletions (DE) which are necessary to map between the correct and hypothesized strings.

$$PER = 100.\frac{SD+IE+DE}{N} \quad (1)$$

where N represents the number of phonemes in the correct transcription.

After calculating the PER metric, we also calculate Acc (recognition accuracy) by using:

$$Acc = 100 - PER \quad (2)$$

PER for four-layer architectures with different hidden layer sizes are presented in table 7.

Table 7: Performance metrics per model.

Models	DE	IE	SD	PER	Acc
256 units	7.15%	4.60%	16.30%	**28.05%**	71.95%
512 units	9.02%	3.04%	12.5%	**24.56%**	75.44%
1024 units	10.03%	1.30%	13.4%	**24.73%**	75.27%

The best PER is obtained with 512 units as hidden layer sizes and is around 24%. This shows that by adding more hidden layer sizes, from 256 units to 512 units, the recognition system performance has significantly improved and has given the best PER.

5 CONCLUSION

In this work, we have achieved a first automatic recognition system of Fongbe phonemes starting from the continuous speech segmentation to the phonetic identity recognition of the units contained in the speech signal. We offer a complete recipe algorithms using fuzzy logic for segmentation, classification, identity recognition. The output of this complete recipe is a set of DBN models trained individually on 2307 sentences of the training set and according to the configuration of each subclass. Unlike most phoneme recognition systems, we first performed a recognition of plosives, fricatives and nasal consonants for consonant phonemes and recognition of oral and nasal for vowel phonemes. This yields a better recognition accuracy, decreasing especially the phone error rate. In this way working with the subclasses by reducing phoneme set increased approximately with 4% the recognition accuracy. Another important finding of this work is that we achieved very good Fongbe phoneme recognition accuracy with 512 units as hidden layer sizes in our DBN models.

REFERENCES

Anapathy, S., Thomas, S., and Hermansky, H. (2009). Modulation frequency features for phoneme recognition in noisy speech. *J. Acoust. Soc. Am*, 125(1):EL8–EL1.

Baghdasaryan, A. G. and Beex, A. A. (2011). Automatic phoneme recognition with segmental hidden markov models. In *Signals, Systems and Computers (ASILOMAR), 2011 Conference Record of the Forty Fifth Asilomar Conference on*, pages 569–574.

chwarz, P., Matejka, P., and Cernocky, J. (2006). Hierarchical structures of neural networks for phoneme recognition. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*.

Huang, X., Acero, A., and Hon, H.-W. (2001). Spoken language processing, a guide to theory, algorithm and system development. *Prentice Hall*.

Laleye, F. A. A., Ezin, E. C., and Motamed, C. (2015a). Adaptive decision-level fusion for fongbe phoneme classification using fuzzy logic and deep belief networks. In *Proceedings of the 12th International Conference on Informatics in Control, Automation and Robotics, Volume 1, Colmar, Alsace, France, 21-23 July*, pages 15–24.

Laleye, F. A. A., Ezin, E. C., and Motamed, C. (2015b). An algorithm based on fuzzy logic for text-independent fongbe speech segmentation. In *11th International Conference on Signal-Image Technology & Internet-Based Systems, SITIS 2015, Bangkok, Thailand, November 23-27*, pages 1–6.

Lefebvre, C. and Brousseau., A. (2001). A grammar of fonge, de gruyter mouton. page 608.

marani, S., Raviram, P., and Wahidabanu, R. (2009). Implementation of hmm and radial basis function for speech recognition. In *Int. Conf. on Intelligent Agent and Multi-Agent Systems, 2009 (IAMA 2009), Chennai*, pages 1–4.

Palaz, D., Collobert, R., and Magimai.-Doss, M. (2013). End-to-end phoneme sequence recognition using convolutional neural networks. *Idiap-RR*.

Solera-Urena, R., Martin-Iglesias, D., Gallardo-Antolin, A., Pelaez-Moreno, C., and Diaz-de Maria, F. (2007). Robust asr using support vector machines. *Speech Communication*, 49(4):253–267.

Trentin, E. and Gori, M. (2007). A survey of hybrid ann/hmm models for automatic speech recognition. *Neurocomputing*, 37(1):91–126.

Young, S. (2008). Hmms and related speech recognition technologies. *Springer Handbook of Speech Processing, Springer-Verlag Berlin Heidelberg*, pages 539–557.

Yousafzai, J., Cvetkovic, Z., and Sollich, P. (2009). Tuning support vector machines for robust phoneme classification with acoustic waveforms. In *10th Annual conference of the International Speech communication association*, pages 2359 – 2362, England. ISCA-INST SPEECH COMMUNICATION ASSOC.