International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 10, October-2015

Arabic Phonemes Recognition Engine: Building Recipe

Wael A. Sultan
Basic Engineering Sciences Dept.
Benha University
Benha, Egypt

M. Hesham Farouk
Engineering Math. & Physics Dept.
Cairo University
Giza, 12613, Egypt

***Abstract*— Arabic phonemes recognition is a very important step in most of Arabic speech recognition based applications. This work presents a recipe for building an efficient Arabic phonemes recognizer with HMMs trained by two databases for Modern Standard Arabic (MSA). HMM parameters such as number of states and number of GMMs per state are optimized. And a comparison between models trained with each database is given. HTK tool has been used in this work and 70.2% maximum recognition rate has been achieved which is very interesting compared with other researches.**

Keywords*— Statistical modeling; Arabic Speech Recognition; HMM; Gaussian Mixtures; Phoneme Model; Insertion penalty

I. INTRODUCTION

Arabic phonemes recognition engine is the first step toward building many important applications such as the phonetic search keyword spotting (PS-KWS) and any many other automatic speech recognition (ASR) based applications. Like the most other speech recognition systems, our Arabic phonemes recognition engine is built on Hidden Markov models (HMMs) which is the core of almost all systems that are using the data-driven statistical approach of speech recognition process.

Many researches present a recipe for building Arabic phonemes recognizer with different approaches as in [1] with 56.79% of recognition rate as maximum score, or [2] with 66.5% recognition rate, while other presented Arabic phonemes recognizer as an introductory step for many applications as in [3], [4] and [5]. Meanwhile, in this paper we investigate how an Arabic phoneme engine is built using HMMs with different Gaussian mixture models (GMMs) and trained with two different databases, particularly; a focus will be paid to the parameters that affect the recognition accuracy of phonemes.

II. PROBLEM STATEMENT

The speech recognition problem is the estimation of the most probable sentence $\widehat{W}$ out of all sentences in the language L given the input speech signal O [6]. This can be expressed as:

$$\widehat{W} = \arg\max_{W \in L} \log\big(P(O|W)P(W)\big) \quad (1)$$

But because of $P(O|W)$ and $P(W)$ comes from two different knowledge sources, particularly from acoustic and language models, so this combination needs to be balanced. The most common modification for balancing two probabilities is to use a language model weight LW and insertion penalty IP, i.e.

$$\widehat{W} = \arg\max_{W \in L} \log\big(P(O|W)p(W)\big) + LW \log\big(P(W)\big) + N * IP \quad (2)$$

Where N the size of W.

Although the systematic optimization of LW and IP is very necessary, very few works have been done in this field [7] [8] [9], since there is no clear physical meaning of these two parameters. In this work we find the optimal values for phoneme insertion penalty (PIP) suitable for Arabic phonemes recognition experimentally.

III. ACOUSTIC MODELING WITH HMM

HMMs are statistical models used to track the temporal changes of non-stationary time series. The HMM models speech as a two-part probabilistic process. The first part models the sequence of transitions of speech over time. The second part models the features in a given state as a probability density function over the space of features [10] [11]. This doubly stochastic nature of the HMM is well suited to the task of continuous speech recognition where the goal is to classify a sequence of phonemes as they proceed in time. A HMM is a Markov chain where the output observation is a random variable generated according to an output probabilistic function associated with each state.

In this work, a phoneme is modelled using a different-states HMM model (particularly 3-states and 5-states HMM) as shown in *"*Fig. 1*"*. A state is provided for each part of the phoneme in a left to right representation. In the 3-states model as shown in *"*Fig. 1-a*"*, a phoneme is represented by the middle state
(state 2) while the start and end states (states 1 and 3) are used to tie models of cascaded phonemes with each other.

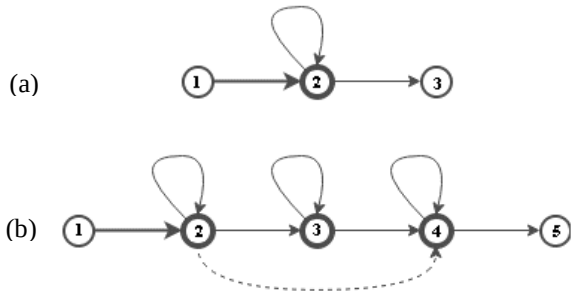


Fig. 1. Phoneme models with 3-states HMM (a) and 5-states HMM (b)

In the 5-states model as shown in *"*Fig. 1-b*"*, a phoneme is represented as follows, the left state (state 2) corresponds to the left part of the phoneme, the middle state (state 3) corresponds to the middle, and the right state (state 4) corresponds to the right part of the phoneme, where the first and last states (states 1 and 5) are entry and exit states respectively and are also used

to tie cascaded models with each other. The transitions are from left to right only i.e. left-right HMM, thus maintaining the causal nature of the speech signal. Transitions to the same state accounts for the natural variability in duration of different phonemes.

Hence for the initial HMM of a phoneme, the number of states and the transitions between these sates are needed to be defined. Next the effects of changing these parameters on the recognition of phonemes will be discussed.

IV. Training and Recognition process

Consider using HMMs to build a phoneme recognition engine, and assume we have a training database with assigned transcription consists of N different phonemes to be recognized, thus each phoneme have to be modeled with a distinct HMM. And also assume that for each phoneme we have k occurrences (observations) in our training database which provide us with the characteristics of that phoneme. In order to do a phoneme recognition, we must perform the following:

1- For each phoneme n in the language, we must build an HMM λ^n, and that by assuming an initial HMM for each phoneme and then training that model with our training database. This process is known as "Training phase" and shown in the following block diagram shown in *"*Fig. 2*"*;

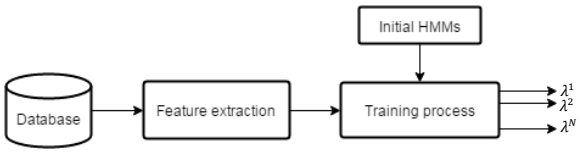


Fig. 2. Block diagram of phoneme models training process

2- For each unknown phoneme which is to be recognized, the process of detection that shown in *"*Fig. 3*"*, is carried by firstly obtaining the observation sequence O via a feature extraction phase and secondly calculating the model likelihoods for all N possible models,

$$P(O|\lambda^n) \qquad 1 \le n \le N \tag{3}$$

And thirdly select the phoneme whose model likelihood is highest,

$$n^* = \arg \max_{1 \le n \le N} [P(O|\lambda^n)] \tag{4}$$

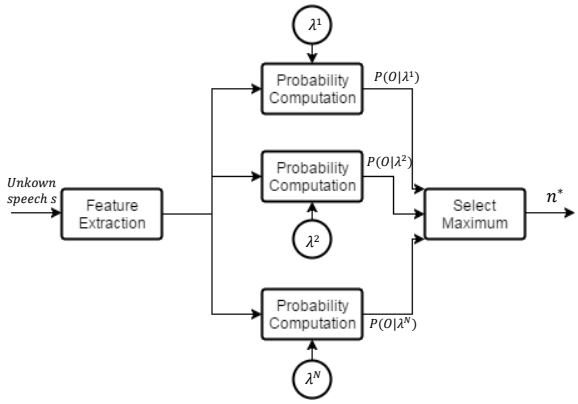


Fig. 3. Block diagram of phoneme model testing process

V. Experimental Setup

A. Database

We have the following Databases

1- Arabic Global phone database from European Language Resources Association (ELRA) [12]. It is a noise-free database composed of about 3165 speech utterances in the Arabic language by different speakers, 3115 of them are used for training and 50 are kept for test. Each utterance is associated with an Arabic transcription composed of 38 phonemes.

2- Data set of Voice of America (VOA) satellite radio news broadcasts in Arabic. The broadcasts were recorded by the Linguistic Data Consortium (LDC) [13]. It is also a noise-free database composed of about 5387 speech utterances in the Arabic language by different speakers, 4887 of them are used for training and 500 are kept for test. Each utterance is associated with an Arabic transcription composed of 48 phonemes as Buckwalter transliteration is used.

The speech signals of both databases are sampled at 16 KHz (62.5 µs per time sample) with resolution of 16 bits. The frame size is 25 ms (400 samples) and the frames are calculated every 10 ms with overlapping of 15 ms between frames.

All utterances are converted into a set of feature vectors of 39 Mel-Frequency Cepstral Coefficients (MFCCs) which is the most widely used spectral representation for feature extraction of speech signals [14].

B. Software

HTK (Hidden Markov model Toolkit) is a toolkit used for building and testing continuous density HMM based recognizers with the selected database. HTK provides us with two evaluation parameters [15] described as follows;

1- Recognition rate/percentage "corr"

$$Corr = ((N - D - S)/N) * 100\% \tag{3}$$

2- Accuracy percentage "acc"

$$Acc = ((N - D - S - I)/N) * 100\% \tag{4}$$

Where I, S, and D are the total number of Insertion, Substitution, and Deletion errors respectively, while N is the number of words in the correct (reference) transcriptions.

C. Methodology

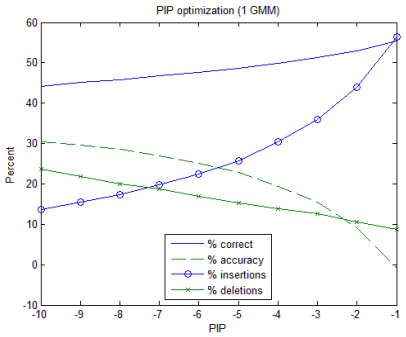
The performance of any phoneme recognizer depends on many parameters, some of them are related to the training process, particularly the suitability of the database, choose of the initial HMM parameters for each phoneme, and the number of GMMs, etc., and other parameters related to the testing/decoding process such as the likelihood estimation method, and the phoneme insertion penalty, etc.. In order to optimize most of these parameters, 3-states and 5-states initial HMMs are trained by two different databases and with different GMMs (1 to 256) and their performances are compered at their optimal PIP and finally conclusion and recommendations are given.

VI. Results

A. Three-states (1-emitting) Model

Two initial models of this type (3-states) is presented as one to ELRA database and the other to LDC database and then both these models are trained with number of GMMs from 1 to 256. After that a selected values of PIPs are been chosen intuitively and tested at each specific number of GMMs. And depending on the number of insertions and deletions errors percentage of the total used utterances in the testing process, the optimal values of PIPs are been chosen in each case of GMMs, e.g. in case of 1 GMM, as shown in "Fig. 2", the insertions and deletions errors are compensated around PIP = -7 in both presented models, hence this values has been suggested to be the optimal value for PIP at this case.

(a)



(b)

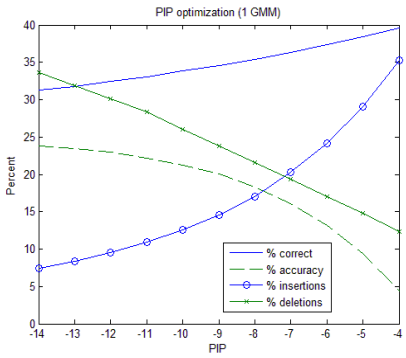


Fig. 4. PIP Optimization at GMMs =1 when 3-state HMM trained by (a) ELRA database (b) LDC database

In the same way, the optimal PIPs for other cases (at other numbers of GMMs) are been obtained. All 3-states models for both databases are tested at their optimal PIPs and their scores are given in the following table.

TABLE I. Recognition Rate and Optimal PIP at a Specific No. of GMMs if Three-State Model

No. of GMMs		*1*	*2*	*4*	*8*	*16*	*32*	*64*	*128*	*256*
Optimal PIP	**ELRA**	-7	-7	-5	-5	-4	-4	-4	-4	-3
	LDC	-7	-7	-7	-6	-5	-5	-5	-4	-4
%Corr	**ELRA**	46.8	49.5	50.8	51.4	54.2	54.7	54.9	56.1	57.6
	LDC	36.4	38.9	40.8	42.9	45.2	46.9	49.2	52.6	55

Last table shows that the 3-states models trained by ELRA database score better than whose trained with LDC database, but a graphical comparison between the performance of these models shown in "Fig. 5" shows that, while increasing the number of GMMs the models trained by LDC database have better enhancement rate than those models trained with ELRA.

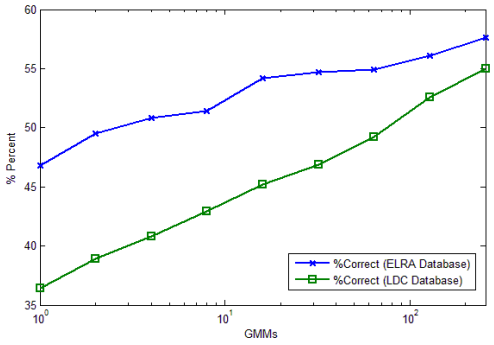


Fig. 5. Recognition rate of 3-state HMM at different GMMs with different Databases

B. Five-states (3-emitting) Model

As the same procedure we follow in the experiment of the 3-states model, two models of 5-states HMM have been created and one trained with ELRA database and the other one trained with LDC database with different number of GMMs from 1 to 256. Then both models are tested with their optimal PIPs and results are summarized in the following table.

TABLE II. Recognition Rate and Optimal PIP at a Specific No. of GMMs if Five-State Model

No. of GMMs		*1*	*2*	*4*	*8*	*16*	*32*	*64*	*128*	*256*
Optimal PIP	**ELRA**	-1	-1	-1	-1	-1	-1	-1	-1	-1
	LDC	-4	-4	-3	-3	-2	-2	-1	-1	0
%Corr	**ELRA**	50.3	52.1	54.1	55.5	57.3	58.1	59.4	61.4	62.9
	LDC	44.6	46.8	48.9	51.4	54.3	57	61.3	65.2	70.2

For the second time, while increasing the GMMs, the models trained with LDC database show better enhancement rate than those models trained with ELRA database. That was also the case when the 3-states HMMs are tested. And also this concept graphically demonstrated with *"*Fig. 6*"*.

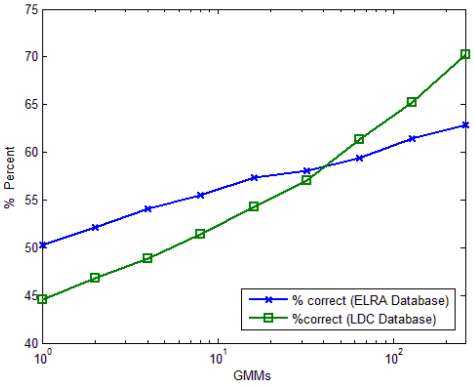


Fig. 6. Recognition rate of 5-state HMM at different GMMs with different Databases

As shown is *"*Fig. 6*"*, 5-states HMMs trained with LDC database beat whose trained with ELRA database particularly when number of training GMMs are greater than 32.

C. Analysis of the individual phonemes' models performance.

In the following we present a comparison between recognition rates of the HMMs of all individual phonemes that trained with our both databases with 256 GMMs.

An insight look in the following table (particularly with models trained with ELRA database) gives us an intuition that some phonemes are well recognized if they were modeled with 3-state HMM than if they were modeled with 5-state HMM and vise-versa, e.g. (/T/, /F/, /S/, /V/, and /Z/) have a better recognition rate with 3-state HMM, while (/A/, /i/, and /r/) are better to be modeled with 5-states HMMs. And the number of states doesn't affect a lot in the recognition rate of the other phonemes. On the other hand, in case of the models trained with LDC database, the 5-state models beat the 3-state models for all phonemes. We also notice that the silence model (/sil/) is recognized very well if it trained with ELRA database than if it trained with LDC database.

TABLE III. COMPARSION BETWEEN RECOGNITION RATE BETWEEN ALL PHONEMES

Database from ELRA				*Database from LDC*			
phoneme	**Arabic Letter**	**% correct**		**phoneme**	**Arabic Letter**	**% correct**	
		3-state HMM	*5-state HMM*			*3-state HMM*	*5-state HMM*
a	◌َ	***63.7***	***74.7***	***A***	ا	***31.5***	***75.7***
A	ا	***53.5***	***60.4***	***b***	ب	***85.6***	***90.6***
b	ب	***85.9***	***86.6***	***t***	ت	***74.6***	***82.1***
c	ع	***81.2***	***78.7***	***v***	ث	***87.5***	***93.2***
C	ء	***67.3***	***72.3***	***j***	ج	***85.4***	***91.7***
d	د	***86.5***	***82.9***	***H***	ح	***96.6***	***99.4***
D	ض	***73.0***	***84.6***	***x***	خ	***95.9***	***98.9***
E	ى	***48.9***	***52.9***	***d***	د	***81.3***	***86.1***
f	ف	***87.0***	***87.5***	*******	ذ	***82.4***	***90.2***
F	ث	***60.9***	***36.4***	***r***	ر	***83.4***	***92.4***
G	غ	***84.6***	***71.4***	***z***	ز	***91.8***	***96.8***
h	ه	***68.6***	***66.7***	***s***	س	***89.6***	***93.5***
H	ح	***99.1***	***94.4***	***$***	ش	***95.6***	***99.2***
i	◌ِ	***44.4***	***57.8***	***S***	ص	***91.6***	***95.8***
j	ج	***89.2***	***85.3***	***D***	ض	***88.6***	***90.5***
k	ك	***89.4***	***86.7***	***T***	ط	***84.4***	***91.3***
l	ل	***77.1***	***83.2***	***Z***	ظ	***92.7***	***100***
m	م	***86.3***	***88.0***	***E***	ع	***86.4***	***91.6***
n	ن	***70.0***	***72.3***	***g***	غ	***89.4***	***93.0***
o	◌ُ	***46.1***	***40.6***	***f***	ف	***88.5***	***93.1***
Q	ق	***84.2***	***86.0***	***q***	ق	***92.7***	***97.1***
r	ر	***78.3***	***87.2***	***k***	ك	***91.1***	***93.7***
s	س	***80.0***	***81.4***	***l***	ل	***72.7***	***83.8***
S	ص	***76.8***	***68.2***	***m***	م	***84.3***	***90.1***
sh	ش	***95.5***	***92.6***	***n***	ن	***77.8***	***89.7***
sil		***93.6***	***95.8***	***h***	ه	***71.3***	***85.0***
t	ت	***70.2***	***75.3***	***w***	و	***87.1***	***92.1***
T	ط	***77.8***	***63.1***	***y***	ي	***80.4***	***87.3***
u	◌ُ	***61.5***	***63.9***	***'***	ء	***86.0***	***95.7***
U	و	***58.3***	***54.7***	***>***	أ	***31.4***	***82.2***
V	ذ	***91.2***	***71.9***	***<***	إ	***67.4***	***87.2***
w	و	***79.2***	***76.5***	***&***	ؤ	***97.6***	***97.7***
x	خ	***98.0***	***97.7***	***}***	ئ	***88.0***	***97.4***
y	ي	***64.6***	***64.0***	***|***	آ	***83.3***	***94.6***
z	ز	***89.6***	***89.6***	***Y***	إ	***60.5***	***98.3***
Z	ظ	***50.0***	***0.0***	***F***	◌ً	***77.0***	***87.8***
				K	◌ٍ	***98.9***	***97.7***
				p	ة	***72.2***	***74.8***
				sil		***34.0***	***49.8***

D. Discussion

Our experiments show the following;

- The suitability of the database is a very important factor in building any phoneme recognizer. Choosing the suitable database depends on the application of the recognizer itself, so it's important to choose the training database carefully.
- The initial HMM parameters such as the number of states is deeply related to the nature of the phoneme. As discussed before, it's better to model some phoneme with 3-states HMM than 5-states HMM and some other is the opposite with that. Inferring the optimal number of states in each phoneme model before training is impossible, hence it's recommended to perform some experiments with a small size dataset and try to optimize initial HMM parameters before start training with a big dataset.
- The number of GMMs is very important parameters in the model training process, increasing the number of GMMs increases the recognition rate, but because there is nothing without cost, we found that increasing the number of GMMs will also increases the processing time of both training and testing of the models. Hence the idea is to compromise between these parameters to build some models that fit well with both trained data and application.

VII. Conclusion

An Engine of Arabic phoneme recognition has been built through this work via optimizing a lot of parameters. 5-states HMMs shows a good performance than 3-state HMMs with different number of states and the best recognition rate obtained with this model was 70.2 % when model trained with 256 GMMs.

References

[1] K. Nahar, W. Al-Khatib, M. Elshafei, H. Al-Muhtaseb and M. Alghamdi, "Data-driven Arabic phoneme recognition using varying number of HMM states," in Communications, Signal Processing, and their Applications (ICCSPA), 2013 1st International Conference on , 2013.

[2] N. Lotner, E. Tetariy, V. Silber-Varod, Y. Bar-Yosef, I. Opher and R. Aloni-Lavi, "Cross-Language Phoneme Recognition for Under-Resourced Languages," in Electrical & Electronics Engineers in Israel (IEEEI), 2012 IEEE 27th Convention of, 2012.

[3] E. Tetariy, Y. Bar-Yosef, V. Silber-Varod, M. Gishri, R. Alon-Lavi, V. Aharonson, I. Opher and A. Moyal, "Cross-language phoneme mapping for phonetic search keyword spotting in continuous speech of under- resourced languages," Artificial Intelligence Research, vol. 4, no. 2, p. p72, 2015.

[4] I. Szöke, P. Schwarz, P. Matějka, L. Burget, M. Karafiát and J. Černocký, "Phoneme based acoustics keyword spotting in informal continuous speech," in Text, Speech and Dialogue, Berlin Heidelberg, 2005.

[5] I.-F. Chen, C. Ni, B. P. Lim, N. F. Chen and C.-H. Lee, "A Keyword-Aware Language Modeling Approach to Spoken Keyword Search," Journal of Signal Processing Systems, pp. 1-10, 2015.

[6] M. Elmahdy, R. Gruhn and W. Minker, Novel Techniques for Dialectal Arabic Speech Recognition, Springer Science & Business Media, 2012.

[7] K. Takeda, A. Ogawa and F. Itakura, "Estimating entropy of a language from optimal word insertion penalty," in ICSLP, 1998.

[8] A. Ogawa, K. Takeda and F. Itakura, "Balancing acoustic and linguistic probabilities," in Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on, 1998.

[9] G. Donaj and Z. Kačič, "The Use of Several Language Models and Its Impact on Word Insertion Penalty in LVCSR," in Speech and Computer, Springer, 2013, pp. 354-361.

[10] X. Huang, A. Acero and H.-W. Hon, Spoken language processing: A guide to theory, algorithm, and system development, Prentice Hall PTR, 2001.

[11] L. R. Rabiner and B. H. Juang, Fundamentals of Speech Recognition, New Jersy: Prentice Hall PTR, 1993.

[12] ELRA, "ELRA-S0193, Global Phone Arabic," ELDA S.A., ELRA "European Language Resources Association" , 2014. [Online]. Available: http://www.elra.info/.

[13] LDC, "Arabic Broadcast News Transcripts," Linguistic Data Consortium (LDC), LDC2006S46, 2014. [Online]. Available: https://www.ldc.upenn.edu/.

[14] D. Jurafsky and H. James, "Speech and language processing an introduction to natural language processing, computational linguistics, and speech," Pearson Education, 2000.

[15] Y. S, K. D, O. J, O. D, V. V and W. P, The HTK Book (for HTK Version 3.4), Cambridge University Engineering Department, 2006.