Proc. of the Fourth International Conference on Advances in Computing, Electronics and Communication - ACEC 2016.

ISBN: 978-1-63248-113-9 doi: 10.15224/ 978-1-63248-113-9-53

Enhancing Arabic Phoneme Recognizer using Duration Modeling Techniques

Mohamed O.M. Khelifa[1], Yousfi Abdellah[2], Yahya O.M. Elhadj[3], Mostafa Belkasmi[1]

Abstract*— in some languages like Classical Arabic (The language of the Holy Quran), phoneme duration is considered as a distinguishing cue in Quranic phonology. Phonological variation of phonemes occurrences contributes to an inaccurate pronunciation of phonemes and therefore inaccurate ASR system. Thus a good phonemes duration modeling can be an essential issue. Currently, the most effective models used in automatic speech recognition (ASR) systems are based on statistical approaches namely Hidden Markov Model (HMM). In standard HMM speech recognition framework, the duration information is poorly employed. However, previous studies have demonstrated that using an HMM with explicit duration modeling techniques have improved the recognition performances in many targeted languages. This paper presents an important phase of our ongoing work which aims to build an accurate Arabic recognizer for teaching and learning purposes. It presents an implementation of an HsMM model (Hidden semi-Markov Model) whose main role is enhancing the classical HMM duration behavior. In this model, both Gamma and Gaussian distributions are used for modeling state durations and compared with the standard geometrical distribution. Experiments have been conducted on a particular database of ten speakers and more than eight hours of speech collected from recitations of the Holy Quran in which all classical Arabic sounds are covered. Results show an accuracy improvement of about 1% over the baseline HMM-based recognizer, which proves the suitability of Gamma distribution in state duration modeling. *(Abstract)

Keywords*— Hidden Markov Model (HMM); Hidden semi-Markov Mode (HsMM); Gamma Distribution; Gaussian Distribution. *(key words)

I. Introduction

Automatic Speech Recognition (ASR) is the field of research which aims to extract the information contained in a speech signal by means of computers. It consists of using matching techniques to compare a sound wave to a set of samples usually compounds of words but recently of phonemes (a basic sound unit in such languages) [1], [2].

[1] TES Research Team
ENSIAS College of Engineering / Mohammed V University in RABAT
Rabat-Morocco

[2] FSJES-souissi
Mohammed V University in RABAT
Rabat-Morocco

[3] Doha Institute, Doha, Qatar; SAMoVA Research Team, IRIT
Paul Sabatier University
Toulouse, France

it is a branch of artificial intelligence (AI) and is linked with many fields of knowledge such as linguistics, acoustics, and pattern recognition. Research in the ASR area has captivated the public and many scholars around the world. In its infancy, anticipation on its applications was very optimistic: what more natural than talking to a computer without having troubles caused by keyboard manipulation? Unfortunately, despite the incredible evolution of computers and knowledge, ASR system does remain a topic of an active research and results still far from the ideal [3], [4], [5]. However, if an ideal ASR system does not yet exist, concrete applications are emerging gradually. ASR system starts to equip mobile phones or GPS by identifying certain keywords to perform the required tasks. Thus various IT applications and IT-solutions appear as automatic translation systems, handicapped people's help, speakers and languages identification, authentication and information retrieval [6].

In spite of wider utilization of ASR systems in foreign languages, the Arabic language still suffers from the rarity of efficient ASR applications, especially for language teaching and evaluation.

One Distinguished application of Arabic ASR is the teaching of the Classical Arabic sound system. Yet classical Arabic is not utilized in everyday communication, it is needed when we consider learning the Holy Quran (the Holy Book of Muslims) and classical heritage of Arabic poetry. Moreover, it can open the door for various classes of Islamic applications.

This paper presents the second phase of our ongoing work whose ultimate goal is building a high-performance Arabic ASR-based system for teaching and learning purposes. The first phase of this work was the development of a baseline HMM-based recognizer for basic Arabic phonemes. This second one aims to enhance the performance of this baseline recognizer by implementing an explicit duration model instead of the implicit one of the ordinary HMM.

The remainder of the paper is organized as follows: section 2 gives a brief overview of our baseline recognizer previously developed. In section 3 we will briefly introduce the HsMM, its assumptions, training and decoding algorithms, and also the implementation we made. Experiments and results are presented in sections 4 and 5. Finally, Section 6 concludes the paper.

Proc. of the Fourth International Conference on Advances in Computing, Electronics and Communication - ACEC 2016.

ISBN: 978-1-63248-113-9 doi: 10.15224/ 978-1-63248-113-9-53

II. HMM-based Recognizer

Our baseline recognizer [12] is based on Hidden Markov Model (HMM) technique. Especially, a left-to-right HMM topology and continuous Gaussian mixture density model are employed. The HMM is the fundamental model widely used in the development of speech recognizers [11],[12]. The internal structure of HMM doesn't come from any knowledge of speech. Therefore their utilization in speech recognition limited to compute quantities related to the speech (a computation model).

The recognizer is built for the purpose to be capable to accurately recognize the basic sounds of classical Arabic languages. it exploits a well-designed Arabic corpus [9] of ten reciters (speakers) and more than eight hours of speech manually segmented at three levels (phoneme, allophone, and words) with accurate time boundaries. The speech was collected from the recitation of the Holy Quran in which all classical Arabic sounds are covered. This corpus was developed in a previously project [8] by Al-Imam Muhammad ibn Saud Islamic University with the support of King Abed Al-Aziz City for Science and Technology (KACST) in Saudi Arabia. Hence we have adapted the corpus for the aim to be annotated in term of basic sounds. We mean by the basic sounds the basic phonemes (single phonemes) without any phonological variation and even considering neither the phonemes germinating nor others. Table 1 shows for each reader the number of sound files, their size, and duration. Table 2 lists the phonemes used and their labels. The Hidden Markov Model Toolkit (HTK) [13], software written in C language developed by Cambridge University which allows building and testing HMM-based recognizers, was used as development environment of our recognizer.

High recognition rates were achieved given an average of 98% for all speakers using 16 GMMs. However, an in-depth analysis of the obtained results shows that we still have a considerable confusion between some phonemes. In order to overcome this misrecognition, we decided to implement an HsMM model for the purpose to improve the durational behavior of the ordinary HMM that is well-known as an inadequate representation of real phonemes duration. It has been reported in several papers [7], [14], [15] that the use of an HSMM models enhances the ASR system performance in many targeted languages like English and Finish.

TABLE I. SOUND FILES AND THEIR DURATION BY READERS

Reader Number	Reader symbol	Sound Files Numbers	Duration (minutes)	Size (MB)
1	AAH	600	49.36	249
2	AAS	590	52.09	261
3	AMS	612	45.78	229
4	ANS	597	49.72	250
5	BAN	585	54.75	276
6	FFA	578	44.11	220
7	HSS	601	49.76	251
8	MAS	580	46.24	232
9	MAZ	608	51.47	258
10	SKG	584	44.29	220
Total		**5935**	**487.53 (8h, 8m)**	**2446**

III. HsMM

A. *HsMM's Durational Behavior*

Recently, ASR systems are based on modeling phonemes using HMMs with a left-to-right topology for each phoneme [11]. The successful applicability of HMM's to various aspects of speech modeling has been demonstrated in various experiments in recent years. These investigations are based on the assumption that speech signal is a quasi-stationary process whose static intervals can be described by the residence time of a single state of an ordinary HMM. The duration of a state in an ordinary HMM is an implicit random variable with an exponential probability density function (pdf). The probability distribution in remaining in a state i having spent a τ duration or sojourn time (i.e. probability of observing τ symbols in sate i) is given by: $p(\tau|\, i) = {a_{ii}}^{\tau-1}(1 - a_{ii})$ where a_{ii} denotes transitions from state i to itself. It is an exponential decreasing distribution. it gains its maximal value at the minimal duration $\tau = 1$ and decreases exponentially as τ increases. It has been found that the said distribution does not provide a correct representation of the statistical duration information of a state [16], [23]. as phoneme can be non-stationary e.g., stops, diphthongs, thus a single observation pdf can model only a stationary phoneme. This rather weak state duration modeling is considered as a main shortcoming of the ordinary HMM.

One of the famous approaches widely used for phonemes duration modeling in HMM-based speech recognition framework is adding explicit durational probability functions (pdf) to each single state, desiring to overcome the inappropriate geometrical state durational (pdf) of an ordinary HMM [17], [24], [25].

A hidden semi-Markov model (HsMM) is an extended version of the ordinary HMM in which a state duration distributions is explicitly defined. In contrast to the unique observation per state considered in an ordinary HMM, a sequence of observations can be emitted while in the state in the case of HsMM and each state has a variable duration. This is an integer variable and takes the value from the set $\{1, 2 \ldots\ldots D\}$ where D is the maximum duration allowed for a single state. Since this is an explicit duration model, there are no self-transition probabilities and the state duration distribution describes the state occupancy probability. The HsMM was firstly introduced by Ferguson [18] and refined by Levinson [19]. It has several forms that have their own assumptions and applications. However, all forms share the same idea about modeling explicitly the duration of a state.

Ferguson suggested an Estimation Maximization (EM) algorithm which can be used in estimating the (pdf) for the state duration. Levinson suggested an approach in which the

Proc. of the Fourth International Conference on Advances in Computing, Electronics and Communication - ACEC 2016.

ISBN: 978-1-63248-113-9 doi: 10.15224/ 978-1-63248-113-9-53

probability distribution of state duration is modeled by a continuous Gamma (pdf) to establish a continuously variable duration hidden Markov model (CVDHMM) [19]. Each of those two algorithms has their advantages and disadvantages. The Ferguson's algorithm requires a large amount of training data compared to that of Levinson. On the other hand, unlike Levinson's algorithm that uses Gamma distribution, that of Ferguson has no prior assumption on the parametric form of the state duration (pdf).
Furthermore, in regard to the computational requirements, Ferguson's algorithm only needs $OT(N^2 + ND)$ operations in the training process, in contrast to that of Levinson which needs $O(N^2T\,D)$ operations. where N is the number of states in the model, T is the period of observations used to estimate the model parameters, and D is the maximum duration allowed for a single state. by virtue of those advantages, Levinson's algorithm was chosen to be implemented in our recognizer.

B. *HsMM's Training and Decoding Algorithms*

By reason of space limitation, we cannot give here a full review of Levinson's algorithm. Hence we limit to the description of the implementation we made and its related issues.
The decoding algorithm of HsMM has a form that differs slightly from that of the classical Viterbi due to the nature of the semi-Markov chain.
By analogy with the forward-backward probabilities of an ordinary HMM, those of HsMM [22] can be expressed as:

$$\delta_t(i) = max_{S_{1:t-1}} P[S_{1:t-1}, S_t = i\,, O_{1:t}\,] \quad (1)$$

where O_t denotes the sequence of observation vector from 1 to time t and $S_t = i$ means phonetic state i which starts at time t . These probabilities may be calculated recursively as follows:

$$\delta_t(j) = \max_i max_\tau \delta_{t-\tau}(i) p(\tau|\, i) a_{ij} b(O_{t-\tau+1\ldots.}|\, O_{t|i}) \quad (2)$$

Where a_{ij} is the transition probability from phonetic state $s(t-1) = i$ to $s(t) = j$; $p(\tau|\, i)$ is the probability of duration τ that state i takes; $b(O_{t-\tau+1\ldots.}|\, O_{t|i})$ is the probability of observing the specified vectors during the sojourn time at state i.

For the estimation of state occupancy probability $p(\tau|\, i)$, we used the same methodology described in Levinson [19] which is to assume two kinds of continuous distributions for the state duration probability, namely, Gamma and Gaussian distributions. For the Gaussian distribution, the state occupancy probability distribution is defined as:

$$p(\tau|\, i) = \frac{1}{6_i(2\pi)^{1\backslash 2}} e^{\frac{-(\tau-m_i)^2}{26^2 i}}$$

where m_i and 6_i are the mean and variance of the Gaussian distribution in state i.

For the Gamma distribution, the state occupancy probability distribution is defined as:

$$p(\tau|\, i) = \frac{{\eta_i}^{v_i}\, \tau^{v_i-1} e^{-\eta_i \tau}}{\Gamma(v_i)}$$

Where η_i and v_i are parameters of the Gamma distribution having a mean of $\mu_i = v_i {\eta_i}^{-1}$ and a variance $6^2{}_i = v_i {\eta_i}^{-2}$.

We note here that for both examined distributions, the state duration mean and variance were estimated for each state in every model with the optimal state sequence for every training utterance. The parameter estimation was done via the Viterbi algorithm after the HMMs have been trained.

The algorithms mentioned above were implemented in the Hidden Markov Model Toolkit (HTK) by adjusting its BaumWelch and Viterbi library functions. The HTK tools: HInit, HRest, HERest and HVite were adjusted for the purpose to incorporate the explicit model into training and decoding procedures of the HMMs for both cases.

IV. Experimental Procedures

It is worth mentioning firstly that the methodology used to build this improved recognizer is similar to that of the baseline HMM-based recognizer with the exception of the use of explicit models instead of implicit one of the ordinary HMM. As we evoked earlier, the classical BaumWelch and Viterbi library functions in HTK were adjusted to integrate the explicit model for both cases, namely, Gamma and Gaussian distributions. Hence this improved recognizer is trained and tested using these new adjusted versions of HTK tools. Our corpus used consists of 5935 waveform files over its corresponding MFCC feature files, label files, TextGrids files and text files containing the corresponding text (Quranic ayah or part of it). In addition, it contains a list of 31 Arabic phonemes, an Arabic dictionary, a list of all unrepeated words included in the whole corpus and other useful files needed for recognizer development. Notice that a special phoneme is added to designate the silence regardless its occurring place. All phoneme models have three emitting states and for each state, a Gaussian Mixture Models (GMMs) are associated to identify the characteristics of the sound portion at this state. For each Hamming window of 10 ms, a vector of 39 acoustical coefficients is extracted. These coefficients are the first twelve MFCC plus their first and second derivatives to capture the static features of signal portion. Also, the energy plus its first and second derivatives are appended to represent the dynamic features. The conversion from the original waveform to a series of acoustical vectors is performed with the "HCopy" tool of HTK.

To initialize and train our HMMs, we used the following combination of HTK tools "HInit + HRest + HERest", which was chosen in the previous work as the best combination of HTK training tools. So, all the experimentations were

Proc. of the Fourth International Conference on Advances in Computing, Electronics and Communication - ACEC 2016.

ISBN: 978-1-63248-113-9 doi: 10.15224/ 978-1-63248-113-9-53

conducted using this combination. For the appropriate number of Gaussians in GMMs, we conducted various experiments varying from 1 to 16 GMMs on specific groups of training and testing sets defined as follows: As we have ten readers, we partitioned our corpus into ten groups of training and testing sets; they are all used in the experimentations, one at a time, and then a global average is computed. For each group, we consider a particular reader to construct the testing set by extracting the first ayah of each Surah from it; the remaining ayahs of this reader, as well as all ayahs of the other readers, are used for training. Hence about 93% of the corpus is used for training and 7% is used for testing. Noting that there is nothing in the literature indicating which number of GMMs is the best for a given context, and thus the optimal number has to be determined by experimentation.

For the recognition, we used the following flat language model (see "Fig. 3") to allow all pronunciation possibilities (any phoneme can appear after any other one).

TABLE II. FLAT LANGUAGE MODEL FOR BASIC PHONEMES

```
$Phon = as10 | bs10 | cs10 | db10 | ds10 | fs10 | gs10 | hb10 | hs10 |
hz10 |is10 | jb10 | js10 | ks10 | ls10 | ms10 |ns10 | qs10 | rs10 | sb10 |
sil | ss10 | tb10 | ts10 | us10 | vb10 | vs10 | ws10 | xs10 | ys10 | zb10 |
zs10 ;
( < $Phon > )
```

This format is converted to an internal HTK representation by the tool "HParse".

The experimentation results are reported in Table 3. The recognition results obtained using Gamma distribution is depicted in "Fig. 1" to be more readable and analyzable. For the purpose of comparison, the results of both examined HsMMs and standard HMM models are reported in the same table.

in the results shown in the table below, HMM describes the standard HMM model used in the baseline system, HsMM1 describes the HsMM model where Gamma distribution was used for the state duration modeling and HsMM2 describes the HsMM model where Gaussian distribution was used for the state duration modeling.

TABLE III. AVERAGE RECOGNITION RATES FOR 1 TO 16 GMM (%)

GMMs	Model	AAH	AAS	AMS	ANS	BAN	FFA	HSS	MAS	MAZ	SKG
1	HMM	89.23	89.69	81.12	87.23	85.90	84.04	90.08	86.97	82.85	81.25
	HsMM1	91.40	90.82	82.18	88.63	86.94	85.88	90.21	88.30	84.64	81.91
	HsMM2	91.13	90.24	81.90	87.10	86.15	85.20	90.11	87.27	83.93	81.15
2	HMM	93.88	92.04	88.43	90.16	93.35	86.97	94.39	94.28	92.42	90.03
	HsMM1	94.15	92.69	88.56	90.82	93.95	88.18	95.86	90.96	93.12	90.82
	HSMM2	93.90	92.10	88.16	90.11	92.12	87.22	95.34	90.64	92.87	90.31
4	HMM	97.07	97.78	90.29	96.01	96.14	95.33	97.65	95.21	95.35	95.21
	HsMM1	97.61	97.65	90.16	96.48	97.22	95.81	98.33	96.30	94.28	96.37
	HsMM2	97.21	97.52	90.43	96.87	96.89	95.70	98.21	96.65	94.35	96.60
6	HMM	97.87	98.43	94.55	97.07	96.54	95.48	98.17	96.41	96.81	97.07
	HsMM1	97.47	98.86	95.08	97.47	97.68	95.54	98.14	97.44	97.43	97.98
	HsMM2	96.97	98.19	95.23	97.14	97.42	95.11	98.37	97.03	97.20	97.04
8	HMM	98.01	98.83	95.37	97.61	98.14	96.14	98.04	96.01	96.41	96.94
	HsMM1	98.20	98.70	96.11	98.31	98.73	97.76	98.17	97.81	97.72	97.28
	HsMM2	97.83	98.29	95.91	97.98	98.68	97.94	97.67	97.39	97.49	97.08
10	HMM	97.74	99.09	95.35	97.87	97.87	95.88	98.96	96.94	97.21	97.74
	HsMM1	97.89	99.09	96.94	98.67	97.61	96.56	99.15	97.20	98.13	97.21
	HsMM2	97.32	98.76	95.72	97.59	97.93	96.18	98.28	96.36	97.73	97.72
16	HMM	98.54	99.48	96.81	98.67	97.61	97.47	98.43	97.07	98.27	96.94
	HsMM1	99.61	99.65	98.12	99.17	98.11	99.27	99.29	98.22	98.79	98.14
	HsMM2	99.46	99.16	97.74	98.49	98.36	98.73	98.48	97.36	97.66	97.35

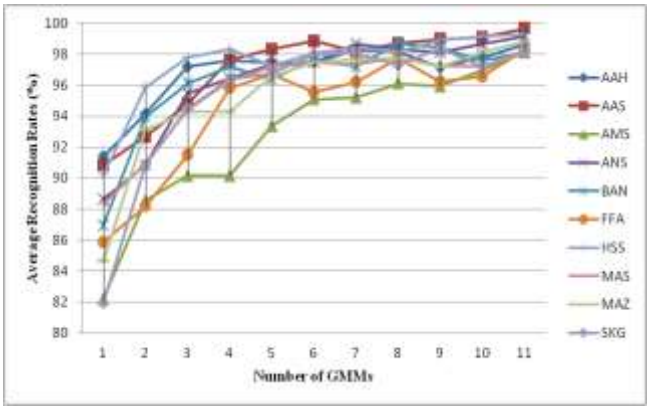

Figure 1. Average Recognition Rates for GMMs

V. Results and Discussion

From the results showed above, we noticed that an improvement in performance of about 1% was achieved for both Gamma and Gaussian distributions. Pointing out that the Gamma distribution best outperforms both the standard (Geometrical) and Gaussian distributions in almost all cases, while the Gaussian distribution outperforms the Geometrical one. The said improvement covers all the readers in the whole corpus. For example, using one GMM, for the Reader AMS, the lowest recognition rate passed from 81.12% in standard HMM to 82.18% in HsMM1 and 81.90% in HsMM2. While in the case of AAH, this percentage passed from 89.23 in standard HMM to 91.40 in HsMM1 and 91.13% in HsMM2. These improvements were respectively 1% and 2%. we noted that the rate of improvement varies from reader to another, in some readers, it seems to be approaching to 2% in only two readers while remaining in 1% in all others. These observed variations in improvement may be explained by the fact that some readers have a speaking rate higher or lower than others and thus the influence of HsMM models cannot be identical for each reader separately. Using 16 GMMs, the lowest recognition rate is passed from 96.81% in standard HMM to 98.14% in our improved HsMM-based recognizer. While the global average recognition rate is increased from 98% to 99%. Despite its importance role, the HsMM model cannot enhance performance as much as expected. Furthermore, the results reported here show clearly that it is needed to take into account the whole-phoneme duration modeling. It also deserves to be remarked that the use of GMM models leads to a significant improvement due to their capability to neutralize and separate phoneme characteristics. Their optimal numbers may depend on the model parameters and the amount of training data used. In this application, between 8 and 10 GMMs seem to be enough.

VI. Conclusion

We presented in this paper the results of an improved Arabic recognizer by implementing an HsMM model instead of the ordinary one used in HMM-based speech recognizers. Two kinds of models which model the state duration explicitly have

ISBN: 978-1-63248-113-9 doi: 10.15224/ 978-1-63248-113-9-53

been constructed and incorporated into the speech recognition process. An improvement of about 1% was achieved. The suitability of Gamma distribution in state duration modeling has been proven compared to both Geometric and Gaussian distributions. The superiority of Gamma distribution in state duration modeling can be assigned to its statistical properties and to the data used in estimating its parameters. As this implementation of HsMM that we have tested gives slight improvements. It cannot improve recognition performance so much as expected. This makes sense because the level in which the durational behavior should be matched is the whole phoneme segments, not states. Our future steps will focus on enhancing the performance of the implemented recognizer by investigating the incorporation of additional knowledge sources into the recognizer. Phonemes duration and energy seem to be the most relevant knowledge sources leading to an ASR system improvement. Unlike the ordinary HMM features which are frame-based, the new ones cover the whole phoneme segments. we are currently looking for the best way to incorporating them into the recognizer hoping that this incorporation leads to a better improvement.

Acknowledgment

The presented work exploits the results (Classical Arabic Sound Database) of a project previously funded by King Abed Al-Aziz City for Science and Technology (KACST) in Saudi Arabia under grant number “AT – 25 – 113”.

References

[1] Daniel Jurafsky and James H. Martin, “Speech and Language Processing”. Pearson Prentice Hall, 2nd Edition, 2009.

[2] L. Rabiner, B.H. Juang. (1993). Fundamentals of Speech Recognition. Prentice Hall.

[3] S. V. Gerven, F. Xie, “A Coparative Study Detection Methods,” in Proc. Eurospeech, vol. 3. 1095-1098, 1997.

[4] C. M. Bishop. (2006). Pattern Recognition and Machine Learning, Springer, 423 pages.

[5] X. Huang, A. Acero, and H. Hon. (2001). Spoken language processing, Prentice-Hall.

[6] Oparin, I.: Language Models for Automatic Speech Recognition Of inflectional Languages. PhD Thesis, University of West Bohemia, Plzen, Czech Republic (2009).

[7] M. J. Russell and A. E. Cook, “Experimental evaluation of duration modelling techniques for automatic speech recognition,” in Proc. ICASSP, 1987, pp. 2376–2379.

[8] Y.O.M. Elhadj, I.A. Alsughayeir, M. Alghamdi, M. Alkanhal, Y.M. Ohali, A.M. Alansari.(2012). Computerized teaching of the Holy Quran (in Arabic), Final Technical Report, King Abdulaziz City for Sciences and Technology (KACST), Riyadh, KSA.

[9] Y.O.M. Elhadj. (2009). Preparation of speech database with perfect reading of the last part of the Holly Quran (in Arabic). Proc. of the 3rd IEEE International Conference on Arabic Language Processing (CITAL'09), pp: 5-8, Rabat, Morocco, May 4-5, 2009.

[10] D. Manning Hardcover., "Foundations of Statistical Natural Language Processing" The MIT Press Cambridge, Massachusetts London, England, Second printing, (1999).

[11] Daniel Jurafsky, James H. Martin. (2008). Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition. 2nd Edition, Prentice Hall.

[12] Y.O.M. Elhadj, Mohamed .O.M. Khelifa, A. Yousfi and M. Belkasmi. 2016. An Accurate Recognizer for Basic Arabic Sounds.ARPN Journal of Engineering and Applied Sciences. VOL. 11, NO. 5, MARCH 2016, p:3239- 3243.

[13] S.Young, al. (2009). HTK Book (V.3.4). Cambridge University Engineering Dept, UK.

[14] A. Bonafonte, X. Ros, and J. B. Marino, “An efficient algorithm to find the best state sequence in HSMM,” in Proc. Eurospeech, 1993, pp. 1547–1550.

[15] P. Ramesh and J. G. Wilpon, “Modeling state durations in hidden Markov models for automatic speech recognition,” in Proc. Int. Conf. Acoust., Speech, Signal Process., 1992, pp. 381–384.

[16] Thomas H. C and Arthur. S. H, “Segmental durations in connected-speech signals: Current results,”The Journal of the Acoustical Society of America 83(4):1553-1573 · January 1988.

[17] Ronald, E. et al., “Probability & Statistics for Engineers & Scientists”. Boston : Upper Saddle River, NJ : Pearson Prentice Hall, cop. 2007.

[18] J. D. Ferguson. Variable duration models for speech. In Proc. Symp. on the Application of HMMs to Text and Speech, pages 143–179, 1980.

[19] S. E. Levinson. Continuously variable duration hidden Markov models for automatic speech recognition. Computer Speech and Language, 1(1):29–45, 1986.

[20] Christopher M. Bishop, "Mixture Models and the EM Algorithm" Microsoft Research, Cambridge, 2006 Advanced Tutorial,Lecture Series, CUED (2006).

[21] Dasu, Nagendra Abhinav, "Implementation of hidden semi-Markov models" (2011). UNLV Theses/Dissertations. Paper 997.

[22] Shun-Zheng Yu. 2010. Hidden semi-markov models. Artificial Intelligence, 174(2):215–243.

[23] Mari,O. et all "From HMM's to segment models: a unified view of stochastic modeling for speech recognition" IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL 4, NO 5, SEPTEMBER 1996.

[24] M. Tachibana, J. Yamagishi, T. Masuko, T. Kobayashi, Performance evaluation of style adaptation for hidden semi-Markov model based speech synthesis, in: INTERSPEECH-2005, 2005, pp. 2805–2808.

[25] Rabiner, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. In IEEE, volume 2, 257–286.