

Chapter

Speech Recognition Based on Statistical Features

Jabbar Hussein

Abstract

The requisition of intelligent devices that might classify a vocalized utterance have been skipping utterance research. The challenging task with utterance recognition models given for the language nature whereby there're no apparent limits among words, an acoustic start with ending are impacted through the neighboring words, also, with various talkers utterance: female/male, senior/young, low/loud utterance, read/spontaneous, fast/slow vocalizing proportion and the utterance sign could be influenced by ambient noise. Accordingly, utterance recognition was exceeding abound of such challenges. To avert particular problems, information steered statistical curriculum built on considerable amounts of vocalized data has been utilized. With this itemize, the aim is to reconnoiter creativity that has making these implements plausible. Utterance recognition and language comprehension have been two important reconnoitering antes thereupon has normally been logged nearer as matters with indicatively and audio vocal, whereby the domain for audio vocal data have stayed introduced as robust impact to the matter thru drib accomplishment. Hence, we amid about determinate methods to utterances and language manipulating, whereby a data around a talking sign and a language that it converses, adjoining thru valuable utilized of information, is established come from inherent recognition of utterance data thru an understandable math-statistical formality.

Keywords: speech recognition, statistical features, language model and automatic speech recognition (ASR)

1. Introduction

The objective of getting a machine to see fluidly spoken talk and react in a characteristic voice has been driving taking research for over 50 years. We are as yet not yet where machines dependably comprehend familiar speech, spoken by anybody, and in any acoustic climate. Disregarding the excess specialized issues that should be tackled, the fields of automatic speech recognition (ASR) and comprehension have made enormous advances and the innovation is presently promptly accessible and utilized on an everyday premise in various applications and administrations [1, 2]. This chapter targets exploring the innovation that has made these applications conceivable. Talking recognition and language understanding are two significant exploration pushes that have customarily been drawn closer as issues in semantics and acoustic phonetics, where

scope of acoustic-phonetic information has been presented as a powerful influence for the issue with astoundingly little achievement. Here, in any case, we center around measurable techniques for speeches and language handling, where the information about a talking signal and the language that it communicates, along with useful usage of the information, is created from genuine knowledge of speech information through an obvious mathematical-statistical formalism.

2. Language Modeling (LM)

With this part, we will reflect on the issue of building a semantic model from a bunch of model words and sentences within an etymological. Semantic models were at first settled for the issue of ‘Speech Recognition’ (SR); they stay assume a predominant part in current (SR) frameworks. They are additionally regularly utilized in other (NLP) utilizes. The element assessment techniques that were initially settled for etymological demonstration, as characterized in this section, are significant in numerous different conditions, like the tagging and analysis problems [3].

Our occupation is as per the following. Expect that we take a body, which is a gathering of the sentences in one linguistic. A few years we could have of composition from the ‘Washington Post’, or we could own an exceptionally huge quantity of original copies by using the web. Accepted this corpus, we might want to estimate the elements of an etymological model. A semantic model is a clear cut as follows. To begin with, we will depict (V) to stand the gathering for entirely words within the language. For instance, once structure the phonetic system concerning the English language, we could say:

$$V = \{ \text{that, cat, funs, maxim, bays, man,} \} \quad (1)$$

For all intents and purposes, (V) can be very large: it could have nearly thousands of words and we expect (V) to be a restricted set. Where a language sentence is a preparation of words, as:

$$x_1, x_2, \dots, x_n.$$

Here (n) is the number with the end goal that ($n \geq 1$), where we consume: $x_i \in V$ for $i \in \{1 \dots (n - 1)\}$, and where we expect to be (x_n) is a particular symbol, HALT (we accept that HALT is certainly not a partner of V). We’ll in no time see the reason why it is appropriate to expect that each sentence decorations in the HALT symbol.

So (V) will depict to become the gathering of entirely sentences within the language V : here, this is a non-limitless group, since the sentences can be of different dimensions.

We next, at that point, provide the following description:

Definition: (LM) An etymological system includes of the restricted group V , also, $p(x_1, \dots, x_n)$ toward such an extent [4]:

- With any-value of $(x_1, \dots, x_n) \in V$, then, at that point, $p(x_1, \dots, x_n) \geq 0$
- Also,

$$\sum_{(x_1 \dots x_n) \in V^+} 1 + p(x_1, x_2, \dots, x_n)$$

Where $p(x_1, \dots, x_n)$ is a likelihood distribution for the (V) sentences. For example, a delineation of a terrible strategy to the instruction of a phonetic system from a preparation body, contemplate a succeeding characterize $c(x_1, \dots, x_n)$ to being the time amount, where the (x_1, \dots, x_n) is acknowledged in our preparation body, also (N) to stand for the complete amount from sentences during the preparation body. We might then characterize:

$$p(x_1 \dots x_n) = \frac{c(x_1 \dots x_n)}{N}$$

This is, anyway, an exceptionally unfortunate system: in explicit it shall dispense (0) likelihood to somewhat sentence that is not understood in a preparation body. Accordingly, it will neglect to rearrange sentences that poor person was acknowledged in a preparation data. A critical useful commitment of hereupon section shall be is adduce approaches that upon in all actuality carry out streamline for sentences that aren't understood in our preparation information. Firstly look, an etymological demonstrating issue appears similar to somewhat unusual work, thus, why it to be thought of? There is a pair of causes [5]:

1. Semantic systems are truly important for an expansive assortment to be uses, a clearest maybe SR plus machine transformation. Within numerous implementations, it's entirely important to own a decent "past" dissemination $p(x_1, x_2, \dots, x_n)$ above whichever sentences are/aren't possible for the language. For an instance, in SR the phonetic model is joint with an audio system that models the way to express different words: Certain strategy for consider it's that upon the audio system produces countless candidate sentences, created with probabilities; a semantic system is then used to rework these choices in light of the fact that they are so plausible to being the sentence within a language.
2. A strategies we are characterized to portraying a (p), then for speculating elements come from the resultant system for showing models, can stand for help with various settings all through the course; for instance, in Neural Network (NN) and Hidden Markov Models (HMM), that we shall acknowledge subsequently, and for systems to the standard language depicting.

3. Statistical features

Every talking signal equivalent for some word is placed in an individual file. Various talking features can be considered, deeming the vocalized words just as an acoustic sign, and come from that the acoustic features could be elicited and so, generally categorized built onto their semantic clarification just as cognitive and physical traits. Furthermore, statistical traits containing, RMS, absolute mean value (AMV), median absolute value (MAV), standard deviation (STD), variance value, covariance, maximum & minimum values and others, as follows [6, 7]:

3.1 AMV

It's come from the outright measure for the sign information. Quite possibly the most standard component could be utilized during the features elicited. It's established by:

$$\bar{P} = 1/R \sum_{r=1}^R P_r \quad (2)$$

Here (P_r) is the information vector, and (R) is the input vector size.

3.2 STD

The (STD) element can be accustomed to working out the value of mean-variation for every part in signal information. It is established by:

$$STD = \sqrt{\frac{1}{R-1} \sum_{r=1}^R (P_r - \bar{P})^2} \quad (3)$$

3.3 Variance

It is the square of (STD). It is established by:

$$VAR = \frac{1}{R-1} \sum_{r=1}^R (P_r - \bar{P})^2 \quad (4)$$

3.4 RMS

As a (MAV) of signals oftentimes will quite often way to be or nearly be zero, an RMS is a best gauging to the qualities of the signs. RMS will be predefined by means of a square-root for the sign mean-square. It will connect with (STD) and is characterized as:

$$RMS = \sqrt{\frac{1}{R} \sum_{r=1}^R P_r^2} \quad (5)$$

3.5 Maximum & minimum values

They could be deemed just as significant features for the sign. Where they could be founded through calculating the biggest and the tiniest amounts of these data, just as predefined in the subsequent:

$$P_{max} = \max (P_1, P_2, \dots, P_R) \quad (6)$$

$$P_{min} = \min (P_1, P_2, \dots, P_R) \quad (7)$$

3.6 MAV

MAV of signs could be founded by the calculating the medium amount in a group of progressives arranged absolute amounts. With two midpoint values, the medium shall is the mean of those amounts.

4. Acoustic modeling and recognition methods

A worthy quality for (LM) is reflected to be a vital piece of a few frameworks for language information applications, like (SR), machine interpretation, and so on. The point of an LM is to characterize likely series of pre-defined language units, which are normally words. Syntactic and semantic and attributes of a language, coded through the LM, director these figures [8].

4.1 Neural network (NN)

The point of a semantic system is to rating the likelihood conveyance $p(w_1^T)$ for word- sequence ($w_1^{t-1} = w_1, \dots, w_T$). Through the 'chain norm', so, for this conveyance could be uttered as:

$$p(w_1^T) = \prod_{t=1}^T p(w_t | w_1^{t-1}) \quad (8)$$

Accompanying for a particular segment displays how Recurrent NN (RNN) and Feedforward NN (FNN) have been utilized for assess this particular likelihood conveyance [9].

4.1.1 FNN

Correspondingly with N-gram system, the FNN usages Markov-theory of tidiness: (N-1 to approximate 1) giving for:

$$p(w_1^T) \approx \prod_{t=1}^T p(w_t | w_{t-N+1}^{t-1}) \quad (9)$$

Consequently, each one with the terms convoluted within this creation, such as: $p(w_t | w_{t-N+1}^{t-1})$, was expected, distinctly, with one progressive estimation for a network depending on:

$$P_{t-j} = X_{t-j} \cdot U, j = N - 1, \dots, 1 \quad (10)$$

$$H_t = f \left(\sum_{j=1}^{N-1} P_{t-j} \cdot V_j \right) \quad (11)$$

$$O_t = g(H_t \cdot W) \quad (12)$$

Where (X_{t-i}) represents one coding for a word (w_{t-i}), while a (U) columns coding a continual word outline (i.e., embedding). Subsequently, (P_{t-i}) i represents a continual outlines for a (w_{t-i}) word. $V = [V_1, V_2, \dots, V_{N-1}]$ and (W) were the system connecting weighs, where they are educated all through preparing adding (U). Besides, the function $f(\cdot)$ is an initiation work, while the function $g(\cdot)$ is the softmax one. **Figure 1-** a displays a representation of a FNN through an extremely durable setting magnitude (N-1 = 3) thru a hidden stratum equal one.

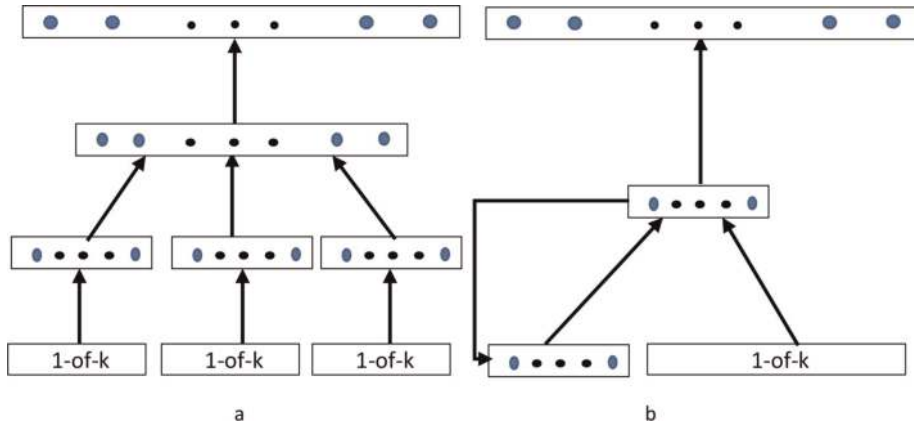


Figure 1
FNN vs. RNN architecture, a) FNN and b) RNN.

4.1.2 RNN

An RNN endeavor for catching entire histories within the setting parameter (h_t), whichever implies a condition for a system and also advances on schedule. Thus, it approaches (2) providing for [10]:

$$p(w_1^T) = \prod_{i=1}^T p(w_i | w_{i-1}, h_{i-1}) = \prod_{i=1}^T p(w_i | h_i) \quad (13)$$

RNN calculates this particular proration correspondingly for FNN. A major variance happens within Eqs. (10) and (11) which they are joined to:

$$H_i = f(X_{i-1} \cdot U + H_{i-1} \cdot V) \quad (14)$$

Figure 1-b shows an illustration of a typical RNN.

4.2 Hidden markov model HMM

We now turn to an important question: given a training body, in what way do we training the function (p)? With section we define HMM, a dominant idea from probability theory [11].

4.2.1 Sustained-length series markov models

Deem a series for arbitrary parameters, such as: (X_1, X_2, \dots, X_n) . Every arbitrary parameter could offtake whichever amount during a limited group (V). Until now we shalt adopt thereupon the dimension for the series (n), will be some permanent integer (such as: $n = 250$).

Our point is as per the following: we might want to demonstrate the series likelihood (x_1, x_2, \dots, x_n) , here ($n \geq 1$), also, $\{x_j \in V \text{ for } (j = 1 \dots n)\}$, thereupon for to say, and show a combined likelihood.

$P(X_1 = x_1, \dots, X_n = x_n)$. Where $|V|^n$ possible series of the form $x_1 \dots x_n$ are there, thus obviously, it's not possible to the reasonable amounts for $(|V| \& n)$ being only listing whole $(|V|^n)$ eventuality. Next, we shall to show the HMM for the same case with the applications of features extracting and recognition.

4.2.2 HMM and one-state method

HMM is a random structure utilized to forecast a greeter event reliant depending on the preceding data. A structure contains a group of statuses, whereby merely an output for the statuses could be observed, and so, whole the variations between the statuses are unidentified as shown in **Figure 2**. The HMM could be clustered into two categories just as shown through a knowing of an outputs: discrete HMM (DHMM) and continues HMM (CHMM).

With Discrete DHMM, this kind achieves (discrete-codes) which are moved through the states and the design (λ) is laid out by the 3-limits (π, A, B) .

While, with Continuous CHMM, "continuous" assigns the possibility of the result concentrations of the covered states. Comparable a Gaussian-limit, the results path the 'Probability Density Function (PDF)', here it's the symmetrical curve outlining the strategy looks like the ring. So, a discernment vector (O) , the PDF is found through a second proviso:

$$P(O) = \sum_{n=1}^k \frac{w_n}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{(O - \mu_n)^2}{\sigma_n^2} \right] \quad (15)$$

Here: w_n : the weight, σ_n : the standard deviation and μ_n : the mean for n^{th} -Gaussian mixture. It's significant thereupon vector covariance (Σ) is corresponding for a squared of the (σ_n) therefore, the CHMM is characterized as during the related group: Here: $(w_n, \mu_n$ and $\sigma_n)$ are: independently, the weight, mean and standard deviation of the n^{th} Gaussian mix. It's critical thereupon a (Σ) will be a comparing for squared of (σ_n) , so, in this way, the CHMM is described by means of the related group:

$$\lambda = (\mu, \Sigma, \pi, n, A) \quad (16)$$

Resulting focuses offer a synopsis of its design:

- N: Structure states number.
- M: Result code number.

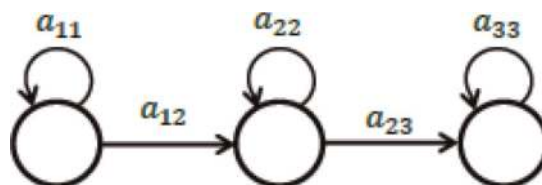


Figure 2.
 Status graph for 3-status L-R HMM.

- π : A major status likelihood element size ($N \times 1$).
- A : A variety likelihood design size ($N \times N$).
- B : The delivery likelihood design size ($N \times M$).

A distinction among a CHMMs & DHMMs, with respect to the HMM limits, is in dis-charge limit, where within CHMM; it's identified through a mean & covariance slightly from separate-codes.

4.2.3 Features elicited

During this deed, features elicited & classification were applied. Every speech sign equivalent for some word will be placed inside a particular file. Various talking features can be considered, furthermore, statistical features containing, RMS, AMV, MAV, STD, variance value, covariance, max. & min. Value and others.

During this effort, a statistical elicited: a mean value & covariance were the features utilized, since the statistical traits characterize a central for the sign and so decrease a necessary magnitude and the time of treating.

4.2.4 Recognition

During the recognition phase, the work is done through two portions:

- Training phase and.
- Testing phase, as follows:

4.2.4.1 Network training (NTr)

For each articulated word, and by joining all the series contracted from the (NTr) word, an array is formed. Whenever the array is outlined, it is given to the HMM to NT. Well applied in the work is an unmistakable framework thereupon comprises only one-status thru ceaseless result densities. No (π) and (A), happen during the one-status framework so, in the present circumstance, they are equal to one. Thusly, a framework (λ) is ordinarily established upon the (\sum and μ) for the advised vectors, just as displayed in **Figure 3**.

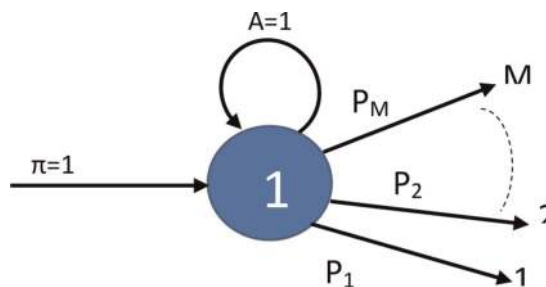


Figure 3.
State diagram of the continuous one-state model (COSM).

To prepare the word arrangement, a 'Baum-Welch' framework thru one-cycle was applied. Simply, by using single Gaussian-blend with (PDFs) were founded just as with Eq. (15), here $P(O) = [P_1, P_2, P_3, \dots, P_M]$. COSM status outline is shown in **Figure 2**.

4.2.4.2 Network testing (NTs)

For the network test, whole articulated words thereupon are not used during the HMM-NTr path, a comparable supra points, where every word will be individually handling. A (\sum and μ) for discernment vectors were ascertaining and also the Viterbi computation was applied to get their eventualities through the whole (PDFs) where they are contracted during a preparation technique. Then, at that point, the file of the best outrageous likelihood may be applied to separate the new word.

4.2.4.3 Example study

Tests are achieved on the work information bases: here with 5-people, 100 examples for everyone, NTr with 70 words & NTs with 30 words. It was a difficult advance in this work for information generation and assortment, on the grounds that the Arabic words sound extremely infrequent on the Internet and furthermore, the works and exploration about verbally expressed the Arabic words are exceptionally inadequate. Thus, recording the audio of the Arabic words from people's lives nearby us were the strategies utilized. The recording system is completed by utilizing (BOYA BY-M1) amplifier. Likewise, a Matlab (2017) utilized as the program that the greater part of the work done through it. Mono-sound with 16-digit coding, 1-channel, and (8000 Hz) sampling frequency. That requirement is picked on since that, the size of each recorded word is vital, as the size is lesser, the method of all tasks follows is quicker and less memory utilized.

Through (HMM), the tests show that the strategy for utilizing the (μ and \sum) are the well one. Along these lines, this strategy is tried utilizing CHMM, and the accompanying particulars are worked:

- 1-Pre-processing: For the words information base: first phase of (DWT) give of 2002×1 vector size.
- 2- Covered Hamming window with 75% (overlapped) with ($n = 100$) of length.
- 3-Feature elicited: $C = [MV MN]$.
- 4- NTr.

Afterward an information assembly, so, we attempted our knowledge computation as assignments later:

- For arbitrary reasons choice (70)
- Test for the remainder (30)
- Playback phases (1 & 2) ordinarily

Here phase (c) will be changed up from the choice of the readiness set.

5. The results

The outcomes showed in **Table 1**, are laid out for 5-people every one has 100-words, preparing with 70 and the testing comes with 30.

Speakers	Training words	Test words	Recognition Rate %
1	70	30	100
2	70	30	100
3	70	30	100
4	70	30	100
5	70	30	100

Table 1
Recognition ratio for HMM.

Patterns for each individual are taken, as shown in **Figure 4**, furthermore launched thru 70 one to the NTr, with 30 for NTs, now, at that point, with the phase of 5-words, we will diminished preparation patterns with a test one expanded, our expectation for find the impact for a quantity of patterns thereto HMM calculation with a recognition ratio, just as displayed with **Figure 5**, a recognition ratio diminished according to diminishing the preparation patterns, so, thereupon is the standard outcome according to such calculation.

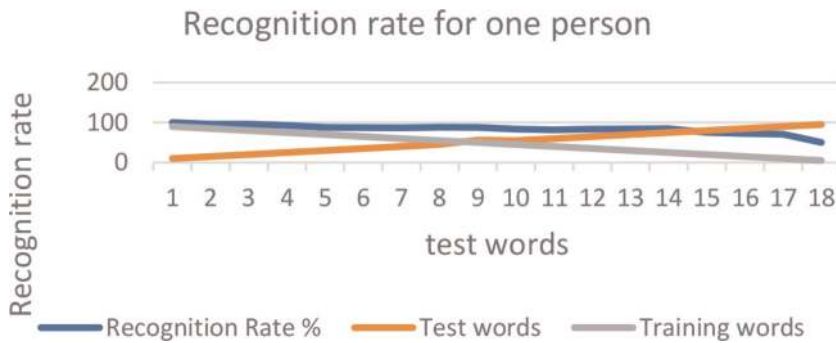


Figure 4.
One person recognition ratio with variables (NTr & NTs) words.

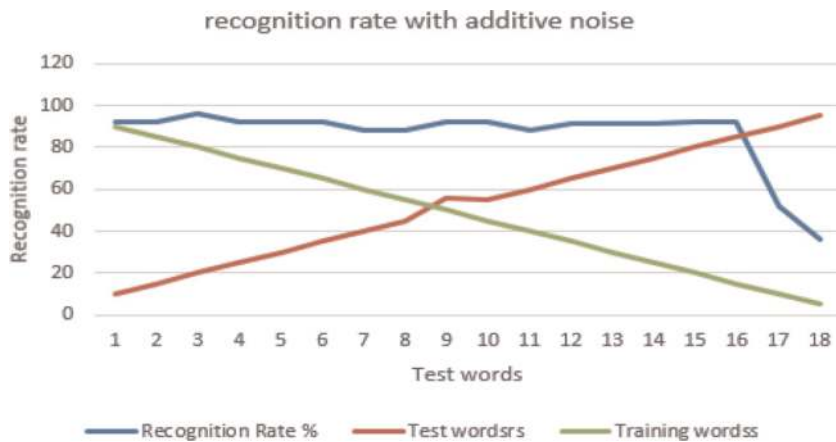


Figure 5.
One person recognition rate with additive noise.

Speakers	Training words	Test words	Recognition Rate %
1	70	30	91.6
2	70	30	83.3
3	70	30	91.6
4	70	30	83.3
5	70	30	83.3

Table 2
 Recognition rate for HMM with AWGN.

Speakers	Training words	Test words	Recognition Rate %	Recognition Rate %
			One state HMM	MLFFNN
1	70	30	100	90
2	70	30	100	90
3	70	30	100	90
4	70	30	100	90
5	70	30	100	90

Table 3
 HMM comparison with MLFFNN.

To mimic the impacts of noise or fault with a presentation for a recognition framework, an ‘Additive White Gaussian Noise’ (AWGN) is strengthening for a patterns samples, preparing and test, since like the clamor covers whole range, an outcomes display great results, just as displayed with **Table 2**. Anyway, with **Figure 2**, it displays an AWGN impact regard one-individual recognition. While with less com-motion values, the results could be improved.

To make a comparison thru different methods, as NN, as Feed Forward NN (FFNN), as displayed in **Table 3**, one can see that the HMM has better outcomes.

6. Conclusions

With SR, it means the usage of an intelligent machine for recognizing spoken word. SR models could be utilized to recognize certain word or to verify a spoken word. Talking processing, talking production, features elicited and finally, patterns equivalent to the SR were presented. Our work has been led us to conclude that the statistical features of the signal are over-performing than the physical features of that signal. The preprocessing step is important for the classification goal.


Author details

Jabbar Hussein

Collage of Engineering, Kerbala University, Kerbala, Iraq

*Address all correspondence to: jabbar.salman@uofkerbala.edu.iq

IntechOpen

© 2022 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] Raut PC, Deoghare SU. Automatic speech recognition and its applications. *International Research Journal of Engineering and Technology (IRJET)*. 2016;**03**(05):2368
- [2] Wiqas G, Navdeep S. Literature review on automatic speech recognition. *International Journal of Computer Applications*. 2012;**41**(8):0975-8887
- [3] Rafal J, Oriol V, Mike S, Noam S, Yonghui W. Exploring the limits of language modeling, Google brain. arXiv: 1602.02410v2 [cs.CL]. 11 Feb 2016;2
- [4] Statistical Speech Recognition: A Tutorial, MC_He_Ch02.Indd. Achorn International; 2008
- [5] Michael C. Language Modeling, (Course Notes for NLP, Columbia University, Columbia). Im-spring; 2013. Available from: <http://www.cs.columbia.edu/~mcollins/lm-spring2013.pdf>
- [6] Othman OK, Khalid K, Aisha HA, Jamal ID. Statistical modeling for speech recognition. *World Applied Sciences Journal 20 (Mathematical Applications in Engineering)*. IDOSI Publications. 2012;**20**:115-122. DOI: 10.5829/idosi.wasj.2012.20.mae.99935. ISSN: 1818-4952
- [7] Husam A, Hala BAW, Abdul MJ, A. H. A new proposed statistical feature extraction method in speech emotion recognition. *Computers & Electrical Engineering*. 2021;**93**:107172
- [8] Youssef O, Dietrich K. A neural network approach for mixing language models. arXiv:1708.06989v1 [cs.CL]. 23 Aug, 2017;1
- [9] Sundermeyer M, Oparin I, Gauvain JL, Freiberg B, Schluter R, Ney H. Comparison of Feedforward and Recurrent Neural Network Language Models, 978-1-4799-0356-6/13. IEEE. 8430 ICASSP; 2013
- [10] Youssef O, Clayton G, Mittul S, Dietrich K. Sequential recurrent neural networks for language modeling. arXiv: 1703.08068v1 [cs.CL]. 23 Mar, 2017;1
- [11] Jabbar SH, Abdulkadhim AS, Thmer RS. Arabic speaker recognition using HMM. *Indonesian Journal of Electrical Engineering and Computer Science*. 2021;**23**(2):1212-1218. ISSN: 2502-4752, DOI: 10.11591/ijeecs.v23.i2.pp 1212-1218