IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Journal Européen des Systèmes Automatisés
Vol. 54, No. 2, April, 2021, pp. 283-287
Journal homepage: http://iieta.org/journals/jesa

Recognition Using DNN with Bacterial Foraging Optimization Using MFCC Coefficients

Gottumukkala HimaBindu[1*], Gondi Lakshmeeswari[2], Giddaluru Lalitha[1], Pedalanka P.S. Subhashini[3]

[1] Department of CSE, School of Technology, GITAM (Deemed to be University), Hyderabad, Telangana 502329, India
[2] Department of CSE, GITAM Institute of Technology, GITAM (Deemed to be University), Visakhapatnam 530045, Andhra Pradesh, India
[3] Electronics and Communication Engineering, RVR & JC College of Engineering, Guntur 522019, Andhra Pradesh, India

Corresponding Author Email: ghimabindu.anu@gmail.com

https://doi.org/10.18280/jesa.540210

Received: 10 January 2021
Accepted: 29 March 2021

Keywords:
bacterial foraging optimization, deep neural network, speech recognition, segmentation, noise removal

ABSTRACT

Speech is an important mode of communication for people. For a long time, researchers have been working hard to develop conversational machines which will communicate with speech technology. Voice recognition is a part of a science called signal processing. Speech recognition is becoming more successful for providing user authentication. The process of user recognition is becoming more popular now a days for providing security by authenticating the users. With the rising importance of automated information processing and telecommunications, the usefulness of recognizing an individual from the features of user voice is increasing. In this paper, the three stages of speech recognition processing are defined as pre-processing, feature extraction and decoding. Speech comprehension has been significantly enhanced by using foreign languages. Automatic Speech Recognition (ASR) aims to translate text to speech. Speaker recognition is the method of recognizing an individual through his/her voice signals. The new speaker initially privileges identity for speaker authentication, and then the stated model is used for identification. The identity argument is approved when the match is above a predefined threshold. The speech used for these tasks may be either text-dependent or text-independent. The article uses Bacterial Foraging Optimization Algorithm (BFO) for accurate speech recognition through Mel Frequency Cepstral Coefficients (MFCC) model using DNN. Speech recognition efficiency is compared to that of the conventional system.

1. INTRODUCTION

Speech is one of the most effective ways of communication. Speech recognition is done by human beings every day [1]. It means to be able to recognize a foreign language, and be able to pronounce the common sounds [2]. Speech recognition technology refers to a system capable of recognizing the "miracle" of conversational speech. This is so because it is difficult to construct a computer that can understand spoken discourse on any topic in any situation because of the difficulty of the process [3].

Speech recognition is the mechanism by which a machine recognizes spoken word. Signal processing converts input speech into identifiable speech form [4]. Speaker (or lecturers or teachers) recognition method can be applied by studying the voiced/unvoiced components or by evaluating the speech energy distribution [5]. Since the model is modified to best reflect observed expression, it is called model-based system. Speech recognition has to be conducted in various ambient environments, thus, the features derived have to be resilient to background noise and system mismatch [6]. Speech emotion recognition technology is also desirable for applications which require natural man-machine interaction such as web movies and computer tutorial programs where the response of these systems depends on the detected emotion [7].

Speaker recognition is essential to the field of communication and surveillance [8]. The recognition process was tested by matching training and test results. Several experiments were performed to develop this kind of recognition process. Some disturbing and distressing shortcomings of the approaches are linear channel distortion, reverberation, and additive noise [9]. Here, the feature extraction of audio is regarded as a difficult task due to the existence of non-stationary noise and reverberation [10]. The characteristics which are used by the speech recognition system are fundamental and spectrum frequency histograms, linear prediction cepstral coefficients (LPCC) [11], instantaneous spectra covariance matrix, averaged auto-correlation, and MFCC (Multi-Frequency Cepstral Coefficients). LPCC and MFCC has a major role in identifying speakers [12].

The best among the means of identification of speakers is MFCC [13]. The local spectral properties of the speech signal can be analyzed using MFCC. The learned features are measured for the full dataset samples and the learned features are saved for speaker recognition [14]. Both training and research set samples are taken into account in MASS [15].

This model also helps users to talk normally without having to pause. They use subword sounds as basic unit of recognition in speech recognition [16]. This is the sequence of production process which can produce bigger linguistic unit such as words/sentences thus recognition [17]. Since the number of subword units are small (typically 45 for Indian languages), collecting adequate data of subword is not difficult [18]. Adding a new word is easy because all that is required is a simple re-wording of an existing word.

2. RELATED WORK

Bacterial Foraging Optimization Algorithm is a recently-developed family of improved optimization techniques. The strategy of the algorithm is called Group foraging technique of a swarm of E.coli microorganisms in multi-optimal function optimization [19]. Germs seek out energy is a fashion to increase energy per model time. The microorganisms can interact between each other in a variety of ways [20]. A bacterium has to prepare for and consider two factors in advance. Chemotaxis is the mechanism where a bacterium travels using quick steps when searching for the nutrients [21]. The main concept of BAO is mimicking chemotactic movement of the microorganisms in the problem room [22]. This theory notes that animals spend about 65 seconds foraging for food and avoiding possible danger in this period.

Liu et al. [1] has successfully developed and trained a hidden Markov model using a neural network. A profound encoder was used to obtain specific acoustic characteristics. Then, a CNN is placed on a final image to extract pictorial topographies from raw aperture zone images. The system can produce nearly optimal phoneme tags because it is trained by using the prepared data. The book presents the integrated use of multi-scale HMMs to assimilate both acoustic and images. The maximum weighted stream posterior (MWSP) technique for video language recognition in various circumstances is very safe and expedient. Another big advantage of WMSP is that it can do away with need of some precise dimension of the sign in question. MWSP modalities can be extended with other modalities that have been proposed by the numerous authors in this paper.

Zhang et al. [3] propose coding methods that are substituted by a hunting device. The chief theme is to focus at making "reliable" divisions in the space of a result which draws on the train of the bit. Various methods are recognized and evaluated for their portrayals of elements and their clear effort organizations. Besides, two-division strategies are discovered and successfully introduced. The first technique is understanding the possibility feature for any challenging audio model specifically.

In the following information technologies, different models joined together, and the arrangement endeavors to strengthen the decision-making through a collective knotted arrangement [23]. The un-narrated truth has been put forward for recording audio exposition. During this time, we need to use GAN modelling to evaluate data cohort [24]. Furthermore, at this stage, we want to inform each epoch data cohort into GAN database [25]. With collection of countless specific circumstances facts into data, these provisional GANs may give factual tags out in public for use for audio demonstrate [26]. The messages of the audio recordings are edited along with other less challenging elements to create an enhanced version of this discourse [27].

3. PROPOSED MODEL

Speech recognition, popularly also known as Automatic Speech Recognition (ASR) is the method of translating speech signal to a sequence of words. Sample speech is initially gained from the user in a controlled way. The proposed model processes voice signals and extracts discriminatory information from the speaker. The proposed model extracts feature from the input voice by dividing the voice samples into small segments. Today, we can conveniently store the databases of speech recognition for different purposes thanks to the development of computer technology. The complete design of the work is seen in Figure 1.

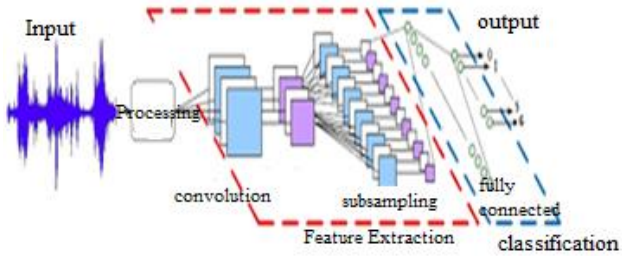

Figure 1. Speech recognition system model

In speech recognition positions, MFCC is outstanding among the best prototypes used by machine speech recognition and speaker recognition programs, and mesh store investigation is made. Logarithmic frequency bands are commonly used in speech recognition applications. Approximates the human auditory better than most strategies. In order to obtain the membership function coefficients from the filtered signal, remove discontinuities through the hamming window to get it correctly.

A DNN is a type of network with a fixed degree of intricacy and with dynamic boundaries. DNN utilizes a systematic technological process for managing data. A DNN network with abundant layers usually incorporates feature extraction and feature organization into a signal learning bank. These NN models have demonstrated eminent success in extensively recording contemporary designs. The layers used in DNN are represented as,

$$L_1 = T(N_1 x) + \log(aWi) \tag{1}$$

where, N_1 is weight of initial. Activation function determines whether a neuron will be activated or not. The activation function is designed to introduce non-linearity into the output of a neuron. The activation function determines what the output of the network will be. This is what an activation function of an ANN does too. It takes the output of the previous cell and transmits it into another cell which is its input. The logistics function special case can be defined as,

$$F(x) = e^{-x^2} \tag{2}$$

Here, x denotes the input of activation. The mapping between the current and next hidden layer after getting the first hidden layer is given as,

$$a_1 = F(W_1 a_l + b_1), l = 2, ..., L \tag{3}$$

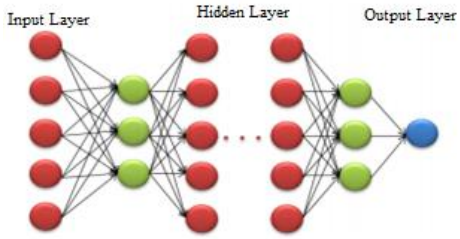

Figure 2. Hidden layer model

where, 'al' is all layers count considering BFO parameters and b is the parameters that are relevant for speech recognition. The hidden layer model using DNN is depicted in Figure 2.

Bacterial Foraging optimization algorithm

A BFO algorithm is based on three processes; Chemotaxis, Replication and Elimination Dispersion. BFO algorithm were discussed here.

BFO Algorithm

Input: Initialize Population and Parameters T,P, N_t, N_c, P(i) {i=1,2,3...θ}.

Step-1: Perform Preprocessing as:

$$e(v_i^{r'}) = p(v_i^{r'}) - \beta p(v_i^{r'} - 1) \quad (4)$$

$$e(v_i^{t'}) = p(v_i^{t'}) - \beta p(v_i^{t'} - 1) \quad (5)$$

where, 'e' is thge processing function, v is the vector for parameter calculation, β is the threshold limit for processing parameters.

Step-2: Create a sample iteration with same properties.

Step-3: Chemotaxis: The process of chemotaxis is the center of the algorithm, simulating E's foraging behavior. Shifting and tumbling coli. Bacteria tumble more often in poorer regions, while bacteria travel more often in areas where food is more plentiful.

The chemotaxis activity of the ith bacterium can be defined as:

$$\theta\, i\, (j+1, k, l) = \theta\, i\, (j, k, l) + C(i) * dcti \quad (6)$$

$$dcti = \Delta(i)\, \Delta T(i) \Delta(i) \quad (7)$$

where, the θ i (j, k, l) represents the ith bacterium at the jth chemotactic, kth reproductive, and lth elimination–dispersal steps.

Step-4: For i=1,2..S.

(i) Calculate fitness function FF as

$$FF(i,j,N(V)) = \theta\, i + dcti \quad (8)$$

(ii) A random variable is a vector between −1 and 1.

(iii) Tumble: generate a random vector $\Delta(i) \in R^n$ with each element $\Delta_m(i)$, m= 1,2,….,S,a random number on [-1,1].

(iv) Move: let (2.1). This results in a step of size C(i) in the direction of the tumble for bacteria i.

(v) Compute J(I,j+1,k,l)with $\theta^i(j+1,k, l)$.

(vi) Swim:

Step-5: let m=0 (counter for swim length).

while $m<N_s$ (if not climbed down too long).

(a) let m = m+1

(b) if $J(I,j+1,k, l) < J_{last}$,

let $J_{last} = J(I,j+1,k, l)$, Then, another step of size C(i) in this same direction will be taken as (2.1) and use the new generated $\theta^I(j+1,k, l)$ to compute the new J(I,j+1, k, *l*).

(c) else let m = N_s.

Go to next bacterium (i+1): if I ≠ S go to (b) to process the next bacteria.

Step-6: If $j<N_c$, go to Step 3. In this case, continue chemotaxis since the life of the bacteria is not over.

Step-7. Reproduction.

For the given k and l,
for each I = 1,2,…..,S, let

$$J_{health}^{i} = \sum_{j=1}^{N_c+1} J(i, j, l) \quad (9)$$

The health of the bacteria. Sort bacterium in order of ascending values (J_{health}).

4. RESULT

A total of 419 items for group correction were chosen as the objects of study. These study items are primarily inmates with mild crimes, no subjective viciousness, like inmates sentenced to public monitoring, suspended, executed outside a prison with temporary sentence and parole staff ruling, etc.

The Equal Error rate EER is well-defined where the false acceptance rate (FAR) is equal to the false rejection rate (FRR). This measure is considered for predicting the accuracy of the speaker recognition. Parameters such as FAR and FRR are tested as follows.

$$FAR = \frac{No.of\ \ false\, accep\tan ce}{No.of\ identification\, attempt} \quad (10)$$

$$FRR = \frac{No.of\ \ false\, rejection}{No.of\ identification\, attempt} \quad (11)$$

Both these parameters evaluate the number of incorrect acceptance and number of incorrect rejections. The EER value of various features included in number of existing and proposed approaches are depicted in Table 1.

Table 1. Speaker identification approaches with EER comparison

Feature extraction methods	EER comparison levels for speech recognition			
	BFO	MBO	PSO	GA
MFCC	**0.0137**	0.0876	0.0865	0.0023
LPCC	0.0997	0.1652	0.0978	0.0675
LSF	0.0276	0.0768	0.0998	0.0085
DWT	0.0587	0.0979	0.1432	0.2867

The Figure 3 represents the Peak Signal to Noise Ratio. The proposed is compared to the traditional methods and the results show that he proposed model result is better than the traditional models.

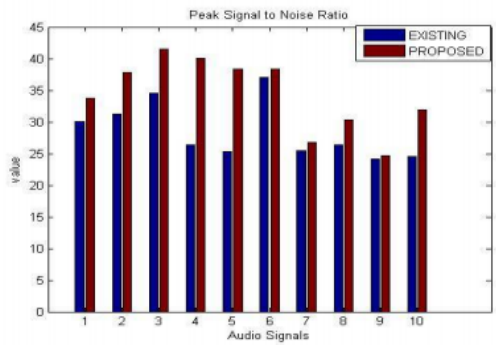

Figure 3. Peak signal to noise ratio comparison levels

The Mean Square Error model is depicted in Figure 4. The results show that the proposed model error rate is less and the model performance is better.

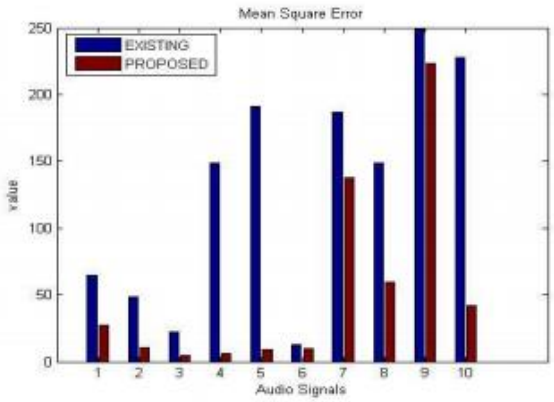

Figure 4. Mean square error

Table 2. Accuracy and execution time and MSE levels

Methods	Accuracy (%)	MSE
BFO	**97.34**	**32%**
MBO	91.23	47%
PSO	89.56	56%
GA	78.98	63%

The accuracy and execution time shown in Table 2, indicates, the proposed methods performance is found to be 10 times better than the existing method. The comparison graph for accuracy levels of this proposed and some other existing methods are highlighted in Figure 5.

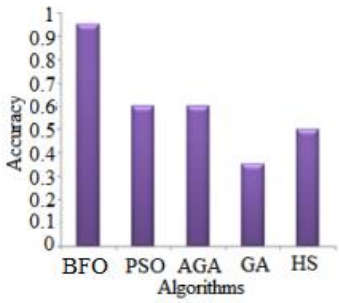

Figure 5. Accuracy levels

The fitness levels of the proposed and traditional models are indicated in Figure 6. The proposed model fitness levels are more when compared to the traditional methods.

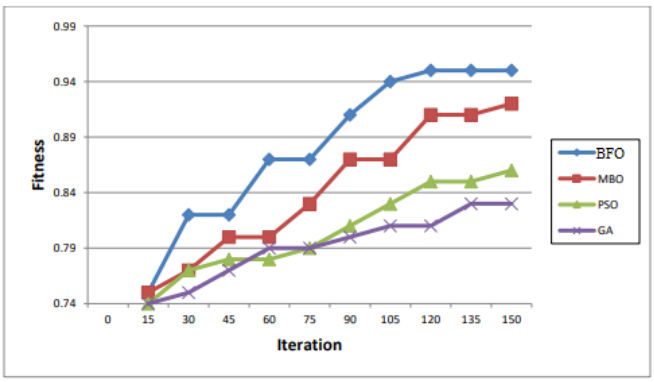

Figure 6. Fitness levels

5. CONCLUSION

Speech recognition is the process by which spoken words are heard by a computer. Input speech is transformed into a recognizable speech type through signal processing. The recognition method of speakers can be applied by observing the voiced components or analyzing the distribution of speech energy. In interactions with human machines, automatic speech recognition (ASR) can play a critical role. Computers which can recognize speech in native language can help to reap the benefit of information technology for a common man. New level crossing threshold allocation schemes based on non-uniform sampling that dynamically distribute the number of quantization levels based on the value of the given input signal amplitude spectrum. In the proposed model BFO model is used to perform speech recognition and the model is compared to the MBO, PSO and GA models. The proposed BFO model attains 97% in speech recognition and the performance of the model is high. In future the MFCC coefficients considered can be reduced and the BFO optimization model can be enhanced to get more accuracy levels.

REFERENCES

[1] Liu, Z., Wu, Z., Li, T., Li, J., Shen, C. (2018). GMM and CNN hybrid method for short utterance speaker recognition. IEEE Transactions on Industrial Informatics, 14(7): 3244-3252. https://doi.org/10.1109/tii.2018.2799928

[2] Cai, Z., Gu, J., Wen, C., Zhao, D., Huang, C., Huang, H., Chen, H. (2018). An intelligent Parkinson's disease diagnostic system based on a chaotic bacterial foraging optimization enhanced fuzzy KNN approach. Computational and Mathematical Methods in Medicine. https://doi.org/10.1155/2018/2396952

[3] Zhang, C., Koishida, K., Hansen, J.H. (2018). Text-independent speaker verification based on triplet convolutional neural network embeddings. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(9): 1633-1644. https://doi.org/10.1109/taslp.2018.2831456

[4] Zeinali, H., Sameti, H., Burget, L. (2017). Text-dependent speaker verification based on i-vectors, neural networks and hidden Markov models. Computer Speech & Language, 46: 53-71. https://doi.org/10.1016/j.csl.2017.04.005

[5] Fayek, H.M., Lech, M., Cavedon, L. (2017). Evaluating deep learning architectures for speech emotion recognition. Neural Networks, 92: 60-68. https://doi.org/10.1016/j.neunet.2017.02.013

[6] Ghahabi, O., Hernando, J. (2017). Deep learning backend for single and multisession i-vector speaker recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(4): 807-817. https://doi.org/10.1109/taslp.2017.2661705

[7] Cumani, S., Laface, P. (2017). Nonlinear i-vector transformations for PLDA-based speaker recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(4): 908-919. https://doi.org/10.1109/taslp.2017.2674966

[8] Mavrovouniotis, M., Li, C., Yang, S. (2017). A survey of swarm intelligence for dynamic optimization:

Algorithms and applications. Swarm and Evolutionary Computation, 33: 1-17. https://doi.org/10.1016/j.swevo.2016.12.005

[9] Wang, J.C., Wang, C.Y., Chin, Y.H., Liu, Y.T., Chen, E.T., Chang, P.C. (2017). Spectral-temporal receptive fields and MFCC balanced feature extraction for robust speaker recognition. Multimedia Tools and Applications, 76(3): 4055-4068. https://doi.org/10.1007/s11042-016-3335-0

[10] Vincent, E., Watanabe, S., Nugraha, A.A., Barker, J., Marxer, R. (2017). An analysis of environment, microphone and data simulation mismatches in robust speech recognition. Computer Speech & Language, 46: 535-557. https://doi.org/10.1016/j.csl.2016.11.005

[11] Wang, K., An, N., Li, B. N., Zhang, Y., Li, L. (2015). Speech emotion recognition using Fourier parameters. IEEE Transactions on Affective Computing, 6(1): 69-75. https://doi.org/10.1109/taffc.2015.2392101

[12] Borde, P., Varpe, A., Manza, R., Yannawar, P. (2015). Recognition of isolated words using Zernike and MFCC features for audio visual speech recognition. International Journal of Speech Technology, 18(2): 167-175. https://doi.org/10.1007/s10772-014-9257-1

[13] Singer, E., Reynolds, D.A. (2015). Domain mismatch compensation for speaker recognition using a library of whiteners. IEEE Signal Processing Letters, 22(11): 2000-2003. https://doi.org/10.1109/lsp.2015.2451591

[14] Gonzalez-Dominguez, J., Lopez-Moreno, I., Moreno, P. J., Gonzalez-Rodriguez, J. (2015). Frame-by-frame language identification in short utterances using deep neural networks. Neural Networks, 64: 49-58. https://doi.org/10.1016/j.neunet.2014.08.006

[15] Miao, Y., Zhang, H., Metze, F. (2015). Speaker adaptive training of deep neural network acoustic models using i-vectors. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(11): 1938-1949. https://doi.org/10.1109/taslp.2015.2457612

[16] Richardson, F., Reynolds, D., Dehak, N. (2015). Deep neural network approaches to speaker and language recognition. IEEE signal processing letters, 22(10): 1671-1675. https://doi.org/10.1109/lsp.2015.2420092

[17] Zhang, Z., Wang, L., Kai, A., Yamada, T., Li, W., Iwahashi, M. (2015). Deep neural network-based bottleneck feature and denoising autoencoder-based dereverberation for distant-talking speaker identification. EURASIP Journal on Audio, Speech, and Music Processing, 2015(1): 1-13. https://doi.org/10.1186/s13636-015-0056-7

[18] Chougule, S.V., Chavan, M.S. (2015). Robust spectral features for automatic speaker recognition in mismatch condition. Procedia Computer Science, 58: 272-279. https://doi.org/10.1016/j.procs.2015.08.021

[19] Das, S., Biswas, A., Dasgupta, S., Abraham, A. (2009). Bacterial foraging optimization algorithm: theoretical foundations, analysis, and applications. In Foundations of Computational Intelligence, 3: 23-55. https://doi.org/10.1007/978-3-642-01085-9_2

[20] Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H.G., Ogata, T. (2015). Audio-visual speech recognition using deep learning. Applied Intelligence, 42(4): 722-737. https://doi.org/10.1007/s10489-014-0629-7

[21] Vo, N.X., Van Ha, T. (2017). The quality of life-a systematic review orientation to establish utility score in Vietnam. Systematic Reviews in Pharmacy, 8(1): 92. https://doi.org/10.5530/srp.2017.1.16

[22] Najkar, N., Razzazi, F., Sameti, H. (2010). A novel approach to HMM-based speech recognition systems using particle swarm optimization. Mathematical and Computer Modelling, 52(11-12): 1910-1920. https://doi.org/10.1016/j.mcm.2010.03.041

[23] Qian, Y., Hu, H., Tan, T. (2019). Data augmentation using generative adversarial networks for robust speech recognition. Speech Communication, 114: 1-9. https://doi.org/10.1016/j.specom.2019.08.006

[24] Sun, L., Chen, S., Xu, J., Tian, Y. (2019). Improved monarch butterfly optimization algorithm based on opposition-based learning and random local perturbation. Complexity, 2019. https://doi.org/10.1155/2019/4182148

[25] Hassan, F., Mohammed, S.A.A., Philip, A., Hameed, A.A., Yousif, E. (2017). Gold (III) Complexes as Breast cancer drug. Systematic Reviews in Pharmacy, 8(1): 76. https://doi.org/10.5530/srp.2017.1.13

[26] Hu, H., Cai, Z., Hu, S., Cai, Y., Chen, J., Huang, S. (2018). Improving monarch butterfly optimization algorithm with self-adaptive population. Algorithms, 11(5): 71. https://doi.org/10.3390/a11050071

[27] Elavarasi, S., Suseendran, G. (2020). Automatic robot processing using speech recognition system. In Data Management, Analytics and Innovation, 185-195. https://doi.org/10.1007/978-981-32-9949-8_14