Mimic Human Speech in Bahasa Indonesia Using Speech Recognition and Speech Synthesis

Valens Prabagita Ivan Susilo
Department of Computer Science
President University
Cikarang, Bekasi, 17550, Indonesia
prabagita12@gmail.com

Abstract

Everyday people use speech recognition and speech synthesis unconsciously. These technologies help them with their activities. With each technology can produce any kinds of software related to speech. Combine both of technologies can produce many more. One of the combinations is mimic human speech. This research will discuss about Speech Recognition that use Convolutional Neural Network as machine learning model and Speech Synthesis that use Concatenative Synthesis with syllables as speech unit. The purpose of this research is to develop application to collect, train, and mimic speech in Bahasa Indonesia. User can participate to record their speech. The collected speech will be trained to be used in the application to recognize the speech. After the collected speech is trained, User can mimic their speech by identify the speech and generate the speech. The application to collect and mimic speech developed as website application and command prompt.

1 Introduction

Speech recognition is the process to get data by analysing speech. The opposite of speech recognition is speech synthesis, the process to produce artificial speech. Therefore, speech recognition is known as speech-to-text and speech synthesis is known as text-to-speech. With each technology can produce any kinds of software related to speech. Combine both of technologies can produce many more. One of the combinations is mimic human speech. The most known usage of mimic human speech is creating a digital speech that will be used as the artificial assistance's speech vocal. Making the artificial assistance more private or personal to the user.

This research aims to develop application which can be used to collect speech data, train machine learning model with collected data and mimic speech in Bahasa Indonesia.

2 Limitation

The limitations of this application are as following:

- There are 9 selected syllables to be used in the application, a, i, na, ma, mu, di, ri, ku, and kan. The syllables are used as speech unit.
- Recorded speech in 1 second, with sample rate 16000 and mono sound.

3 Method

The approach used to achieve this research objectives are using the following techniques:

3.1 Concatenative Synthesis

Concatenative synthesis connecting pre-recorded natural utterances is probably the easiest way to produce intelligible and natural sounding synthetic speech. One of the most important aspects in concatenative synthesis is to find correct unit length. In present systems units used are usually words, syllables, demisyllables, phonemes, diphones, and sometimes even triphones [Hande, 2014].

3.2 MFCC

MFCC is one of the most commonly used feature extraction method in speech recognition introduced by Davis and Mermelstein in the 1980's [practicalcrypthography.com, 2012].

3.3 Convolutional Neural Network

Convolutional neural network (CNN) is one of known variants neural network model to recognized image. The model is designed to recognize an image object no matter what surface the object is on. The model doesn't have to re-learn the idea of child for every possible surface it could appear on [Geitgey, 2016].

4 Experimental Result

In order to evaluate the effectiveness of the proposed methods in the previous section within the application, experiments are done to ensure the application runs well.

The dataset during the speech recognition evaluation is 10000 male and female speech data, each 500 on each syllable, on not noisy background and each 500 unknown sound. Corrected result shows from the more than 75% of model accuracy. The table also fill with accuracy with the syllable result along with it. Random or unknown condition is tested with silent condition, o, and mi syllables.

There are 4 testers. Male user that the records is trained by the application is called Tester 1. Female user that the records is trained by the application is called Tester 2. Male user that the record hasn't trained by the application called Tester 3. Female user that the record hasn't trained by the application called Tester 4. Every tester is test in each the following environment:

- Noisy background (Loud music or people chit-chat).
- Semi noisy background (Rain noise or sound from the other rooms).
- Not noisy background.
-

Table 1: Tester 1 results on noisy background.

No	Spoken Syllable	Noisy Background			Correct Result
		1	2	3	
1	a	52 *(a)*	100 *(a)*	100 *(unknown)*	1/3
2	i	62 *(na)*	85 *(ri)*	96 *(ma)*	0/3
3	na	100 *(unknown)*	86 *(kan)*	100 *(unknown)*	0/3
4	ma	100 *(a)*	84 *(unknown)*	100 *(a)*	0/3
5	mu	99 *(kan)*	71 *(ku)*	88 *(unknown)*	0/3
6	di	59 *(mu)*	100 *(ri)*	97 *(ri)*	0/3
7	ri	50 *(na)*	100 *(ri)*	92 *(a)*	1/3
8	ku	100 *(ku)*	85 *(a)*	100 *(kan)*	1/3
9	kan	72 *(unknown)*	61 *(kan)*	79 *(kan)*	3/3
10	Unknown 1 *(silent)*	99 *(unknown)*	100 *(a)*	99 *(unknown)*	2/3
11	Unknown 2 *(o)*	100 *(na)*	100 *(a)*	100 *(ku)*	0/3
12	Unknown 3 *(mi)*	100 *(a)*	51 *(kan)*	96 *(ma)*	1/3

Table 2: Tester 1 results on semi noisy background.

No	Spoken Syllable	Semi Noisy Background			Correct Result
		1	2	3	
1	a	100 *(a)*	100 *(a)*	100 *(a)*	3/3
2	i	100 *(i)*	36 *(ri)*	100 *(i)*	2/3
3	na	97 *(mu)*	48 *(ma)*	99 *(na)*	1/3
4	ma	49 *(i)*	98 *(ri)*	79 *(mu)*	0/3
5	mu	55 *(ri)*	99 *(ku)*	50 *(ri)*	0/3
6	di	57 *(i)*	100 *(di)*	100 *(i)*	1/3
7	ri	100 *(ri)*	75 *(ri)*	100 *(ri)*	3/3
8	ku	92 *(ku)*	68 *(ku)*	91 *(ku)*	2/3
9	kan	100 *(kan)*	94 *(na)*	100 *(kan)*	2/3
10	Unknown 1 *(silent)*	100 *(a)*	99 *(unknown)*	100 *(ma)*	2/3
11	Unknown 2 *(o)*	100 *(ku)*	89 *(ku)*	72 *(ku)*	1/3
12	Unknown 3 *(mi)*	100 *(ri)*	94 *(ri)*	100 *(ri)*	0/3

Table 3: Tester 1 results on not noisy background.

No	Spoken Syllable	Not Noisy Background			Correct Result
		1	2	3	
1	a	100 (a)	100 (a)	100 (a)	3/3
2	i	94 (i)	48 (kan)	94 (i)	2/3
3	na	80 (na)	99 (na)	100 (na)	3/3
4	ma	96 (di)	90 (i)	57 (di)	0/3
5	mu	100 (i)	86 (ku)	95 (i)	0/3
6	di	99 *(ri)*	100 *(i)*	80 *(ri)*	0/3
7	ri	100 *(ri)*	96 *(ri)*	91 *(ri)*	3/3
8	ku	85 *(ku)*	96 *(ku)*	100 *(ku)*	3/3
9	kan	92 *(kan)*	100 *(kan)*	100 *(kan)*	3/3
10	Unknown 1 *(silent)*	98 *(unknown)*	98 *(unknown)*	98 *(unknown)*	3/3
11	Unknown 2 *(o)*	100 *(ku)*	98 *(ku)*	65 *(ku)*	1/3
12	Unknown 3 *(mi)*	54 *(mu)*	100 *(ri)*	67 *(mu)*	2/3

Table 4: Tester 2 results on noisy background.

No	Spoken Syllable	Noisy Background			Correct Result
		1	2	3	
1	a	99 *(unknown)*	99 *(a)*	99 *(a)*	2/3
2	i	82 *(unknown)*	98 *(ma)*	99 *(ri)*	0/3
3	na	52 *(ri)*	99 *(ma)*	61 *(ma)*	0/3
4	ma	50 *(unknown)*	37 *(unknown)*	95 *(ri)*	0/3
5	mu	99 *(unknown)*	52 *(di)*	99 *(unknown)*	0/3
6	di	99 *(i)*	99 *(unknown)*	99 *(unknown)*	0/3
7	ri	86 *(i)*	20 *(kan)*	57 *(di)*	0/3
8	ku	99 *(ku)*	68 *(kan)*	90 *(unknown)*	1/3
9	kan	84 *(kan)*	99 *(kan)*	78 *(a)*	2/3
10	Unknown 1 *(silent)*	100 *(unknown)*	100 *(unknown)*	100 *(unknown)*	3/3
11	Unknown 2 *(o)*	99 *(ma)*	99 *(ku)*	99 *(ku)*	0/3
12	Unknown 3 *(mi)*	36 *(di)*	96 *(mu)*	99 *(di)*	1/3

Table 5: Tester 2 results on semi noisy background.

No	Spoken Syllable	Semi Noisy Background			Correct Result
		1	2	3	
1	a	100 *(a)*	100 *(unknown)*	99 *(a)*	2/3
2	i	92 *(ri)*	97 *(unknown)*	96 *(ri)*	0/3
3	na	83 *(ma)*	62 *(unknown)*	75 *(kan)*	0/3
4	ma	87 *(unknown)*	64 *(ma)*	90 *(na)*	2/3
5	mu	76 *(ma)*	78 *(ku)*	66 *(di)*	0/3
6	di	58 *(mu)*	48 *(mu)*	30 *(kan)*	0/3
7	ri	99 *(kan)*	89 *(kan)*	99 *(ri)*	1/3
8	ku	99 *(unknown)*	99 *(ku)*	88 *(unknown)*	1/3
9	kan	99 *(kan)*	65 *(kan)*	85 *(kan)*	2/3
10	Unknown 1 *(silent)*	100 *(unknown)*	79 *(a)*	99 *(ri)*	1/3
11	Unknown 2 *(o)*	99 *(a)*	82 *(ku)*	99 *(na)*	0/3
12	Unknown 3 *(mi)*	89 *(kan)*	35 *(ku)*	96 *(ri)*	1/3

Table 6: Tester 2 results on not noisy background.

No	Spoken Syllable	Not Noisy Background			Correct Result
		1	2	3	
1	a	88 *(a)*	99 *(a)*	99 *(a)*	3/3
2	i	99 *(i)*	99 *(i)*	52 *(i)*	2/3
3	na	97 *(a)*	54 *(kan)*	99 *(a)*	0/3
4	ma	99 *(na)*	99 *(na)*	98 *(na)*	0/3
5	mu	81 *(i)*	95 *(i)*	44 *(i)*	0/3
6	di	99 *(i)*	36 *(i)*	96 *(i)*	0/3
7	ri	47 *(ri)*	32 *(unknown)*	48 *(i)*	0/3
8	ku	98 *(ku)*	67 *(kan)*	99 *(ku)*	2/3
9	kan	84 *(na)*	85 *(na)*	87 *(na)*	0/3
10	Unknown 1 *(silent)*	99 *(unknown)*	99 *(unknown)*	100 *(unknown)*	3/3
11	Unknown 2 *(o)*	100 *(ku)*	99 *(ku)*	95 *(ku)*	0/3
12	Unknown 3 *(mi)*	67 *(ri)*	78 *(di)*	77 *(i)*	1/3

Table 7: Tester 3 results on noisy background.

No	Spoken Syllable	Noisy Background			Correct Result
		1	2	3	
1	a	100 *(a)*	99 *(a)*	100 *(a)*	3/3
2	i	71 *(ma)*	99 *(ri)*	99 *(ri)*	0/3
3	na	70 *(ku)*	99 *(a)*	57 *(unknown)*	0/3
4	ma	99 *(a)*	100 *(a)*	100 *(a)*	0/3
5	mu	74 *(mu)*	96 *(kan)*	74 *(mu)*	0/3
6	di	99 *(ri)*	96 *(ri)*	89 *(ri)*	0/3
7	ri	100 *(unknown)*	90 *(na)*	86 *(a)*	0/3
8	ku	98 *(kan)*	64 *(mu)*	99 *(a)*	0/3
9	kan	99 *(na)*	100 *(kan)*	98 *(kan)*	2/3
10	Unknown 1 *(silent)*	99 *(a)*	99 *(a)*	100 *(a)*	0/3
11	Unknown 2 *(o)*	92 *(na)*	84 *(na)*	61 *(mu)*	1/3
12	Unknown 3 *(mi)*	64 *(kan)*	53 *(na)*	99 *(ri)*	2/3

Table 8: Tester 3 results on semi noisy background.

No	Spoken Syllable	Semi Noisy Background			Correct Result
		1	2	3	
1	a	100 *(a)*	99 *(a)*	99 *(a)*	3/3
2	i	99 *(ri)*	90 *(di)*	92 *(di)*	0/3
3	na	100 *(kan)*	99 *(kna)*	92 *(ma)*	0/3
4	ma	99 *(a)*	100 *(kan)*	100 *(kan)*	0/3
5	mu	99 *(mu)*	99 *(mu)*	65 *(ku)*	2/3
6	di	99 *(ri)*	100 *(ri)*	99 *(ri)*	0/3
7	ri	95 *(kan)*	76 *(i)*	78 *(di)*	0/3
8	ku	99 *(a)*	96 *(a)*	100 *(a)*	0/3
9	kan	99 *(a)*	96 *(a)*	99 *(a)*	0/3
10	Unknown 1 *(silent)*	99 *(ri)*	73 *(di)*	60 *(unknown)*	2/3
11	Unknown 2 *(o)*	100 *(a)*	99 *(ku)*	100 *(a)*	0/3
12	Unknown 3 *(mi)*	99 *(na)*	99 *(ri)*	99 *(i)*	0/3

Table 9: Tester 3 results on not noisy background.

No	Spoken Syllable	Not Noisy Background			Correct Result
		1	2	3	
1	a	100 *(a)*	100 *(a)*	100 *(a)*	3/3
2	i	92 *(ri)*	95 *(ri)*	41 *(ku)*	0/3
3	na	99 *(na)*	74 *(ku)*	80 *(a)*	1/3
4	ma	59 *(a)*	84 *(ku)*	48 *(kan)*	0/3
5	mu	99 *(mu)*	99 *(ri)*	99 *(ri)*	1/3
6	di	100 *(ri)*	99 *(i)*	99 *(i)*	0/3
7	ri	99 *(ri)*	96 *(ri)*	94 *(i)*	2/3
8	ku	73 *(di)*	100 *(mu)*	99 *(mu)*	0/3
9	kan	100 *(kan)*	99 *(kan)*	96 *(kan)*	3/3
10	Unknown 1 *(silent)*	98 *(unknown)*	98 *(unknown)*	98 *(unknown)*	3/3
11	Unknown 2 *(o)*	99 *(ku)*	68 *(mu)*	100 *(ku)*	1/3
12	Unknown 3 *(mi)*	99 *(ri)*	93 *(i)*	100 *(ri)*	0/3

Table 10: Tester 4 results on noisy background.

No	Spoken Syllable	Noisy Background			Correct Result
		1	2	3	
1	a	100 *(unknown)*	99 *(unknown)*	94 *(ma)*	0/3
2	i	97 *(unknown)*	99 *(ri)*	99 *(ri)*	0/3
3	na	100 *(na)*	99 *(ri)*	99 *(ma)*	1/3
4	ma	93 *(na)*	99 *(ma)*	100 *(na)*	1/3
5	mu	55 *(ri)*	99 *(unknown)*	43 *(unknown)*	0/3
6	di	99 *(ri)*	54 *(i)*	99 *(ma)*	0/3
7	ri	97 *(ri)*	74 *(ri)*	100 *(ri)*	2/3
8	ku	68 *(ma)*	92 *(ku)*	52 *(52)*	1/3
9	kan	48 *(ku)*	22 *(ku)*	99 *(a)*	0/3
10	Unknown 1 *(silent)*	100 *(unknown)*	100 *(unknown)*	100 *(unknown)*	3/3
11	Unknown 2 *(o)*	100 *(a)*	100 *(a)*	100 *(a)*	0/3
12	Unknown 3 *(mi)*	55 *(mu)*	99 *(ri)*	94 *(i)*	1/3

Table 11: Tester 4 results on semi noisy background.

No	Spoken Syllable	Semi Noisy Background			Correct Result
		1	2	3	
1	a	73 *(a)*	91 *(unknown)*	99 *(unknown)*	0/3
2	i	81 *(ri)*	99 *(i)*	100 *(ri)*	1/3
3	na	100 *(ma)*	99 *(unknown)*	99 *(unknown)*	0/3
4	ma	99 *(ma)*	99 *(na)*	78 *(ma)*	2/3
5	mu	81 *(unknown)*	99 *(unknown)*	99 *(unknown)*	0/3
6	di	100 *(ri)*	99 *(i)*	99 *(i)*	0/3
7	ri	85 *(ri)*	99 *(ri)*	98 *(ri)*	3/3
8	ku	96 *(ku)*	89 *(ku)*	99 *(a)*	2/3
9	kan	100 *(a)*	99 *(a)*	64 *(ma)*	0/3
10	Unknown 1 *(silent)*	99 *(unknown)*	99 *(ri)*	100 *(unknown)*	2/3
11	Unknown 2 *(o)*	99 *(a)*	100 *(a)*	99 *(a)*	0/3
12	Unknown 3 *(mi)*	43 *(a)*	59 *(ri)*	99 *(unknown)*	3/3

Table 12: Tester 4 results on not noisy background.

No	Spoken Syllable	Not Noisy Background			Correct Result
		1	2	3	
1	a	99 *(unknown)*	75 *(a)*	100 *(a)*	2/3
2	i	88 *(i)*	99 *(i)*	99 *(ri)*	2/3
3	na	99 *(na)*	99 *(ma)*	99 *(ma)*	1/3
4	ma	55 *(ma)*	72 *(kan)*	99 *(na)*	0/3
5	mu	99 *(unknown)*	94 *(unknown)*	28 *(i)*	0/3
6	di	89 *(i)*	99 (ri)	99 *(ri)*	0/3
7	ri	99 *(ri)*	100 *(ri)*	99 *(ri)*	3/3
8	ku	39 *(kan)*	99 *(ku)*	58 *(mu)*	1/3
9	kan	100 *(a)*	99 *(a)*	100 *(a)*	3/3
10	Unknown 1 *(silent)*	40 *(unknown)*	24 *(unknown)*	60 *(unknown)*	3/3
11	Unknown 2 *(o)*	100 *(a)*	99 *(a)*	100 *(a)*	0/3
12	Unknown 3 *(mi)*	60 *(ri)*	100 *(i)*	99 *(ri)*	1/3

Although noisy and semi background on every tester not show a good result, it still can predict about 2 or 3 correct spoken syllables with high accuracy. Not noisy background isn't guaranteed all spoken syllables are correct even on tester 1 and 2. But, the result is better than noisy and semi background. This is because all the training data is recorded on not noisy background. Also, noise removal or reduction is not applied when extraction process is done before training begin.

Tester 1 and 2 at most moment have good result than tester 3 and 4. But, at some moment tester 3 and 4 have better result than tester 1 and 2. The most good result on tester 1 and 2 is due to the training data is all contain tester 1 and 2. When tester 3 and tester 4 have better result it can be cause by the background condition or the microphone ability to record the speech.

Syllable ma, mu, di, unknown o, and unknown mi have bad result on most time compare to others. The unknown o and mi are caused by the unknown training data is random and mostly background noise instead of focusing o and mi. The machine learning model can't predict syllable ma, mu, and di well. Most times they have low accuracy but correct prediction or they predict the relative syllable, in example ri or i is the result on di. This can be improved by adding more the training data as the training data is still relatively small. An optimum machine learning model can also improve the result on them and also other syllables as well.

The random text to test generating the speech are 'diriku', 'aku makan ikan', 'di mana mamamu', and 'halo namaku ivan'

'diriku', 'aku makan ikan', and 'di mana mamamu' can be generated. But, 'halo namaku ivan' cannot. It happens because 'h' is not found in the database as the text analysed from the beginning. The application alert user on the browser that 'h' is not found and listed registered syllables. Actually 'h' itself is not listed on available syllables. As the text contains syllable that not listed on available syllables or even contains available syllables but not identified yet will alert user and stop the generating process.

The following is the result from ‘diriku’, ‘aku makan ikan’, and ‘di mana mamamu’ in sequence:

Figure 1: Waveform of ‘diriku’.

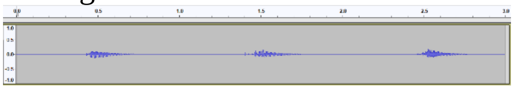


Figure 2: Waveform of 'aku makan ikan'.

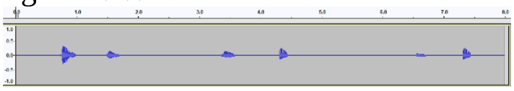
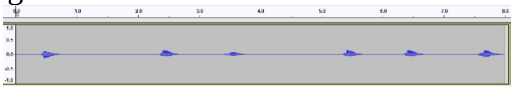
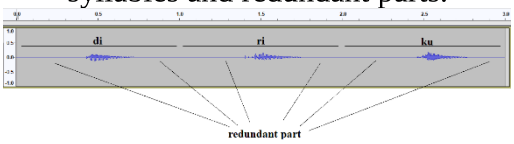


Figure 3: Waveform of 'di mana mamamu'.



As shown in figure above each text can be distinguish very easy. As the recorded speech is take 1 second, the duration of the generated speech simply is the sums up of syllables in the text. Although the speech is generated well, can be heard and understood, there is still silent part or redundant part between the syllables except for space. It makes the speech become not fluently enough. This is because there is no processing to analysed and delete the redundant part in concatenative process or right after identifying the speech in the application.

Figure 4: Waveform of 'diriku' show syllables and redundant parts.



5 Discussion

In this section, there are some discussion why the proposed methods are used towards this research. The important points are:

- Concatenative Synthesis for Speech Synthesis
- Syllable for speech unit
- MFCC for Speech Recognition
- CNN for Machine Learning model

As in mimic speech, the speech is taken from recognized speech, concatenative synthesis can be the best approach other than the other approach. Besides it is the easiest way rather than the others, it also quick to develop. Articulatory and Formant synthesis are too complex because in need a lot of parameter to develop the vocal tract or set of rules that can fit to many speakers.

It is hard to find research regarding to exact amount of Bahasa Indonesia demisyllables, phonemes or smaller. Then, syllables are the best options for the speech unit as word need much more memory and less flexibility to generate speech in form of sentence.

MFCC has advantage over Linear Prediction Coefficients (LPC) [Dave & Pipalia, 2014]. It able to mimic human auditory system well. Although Perceptually Based Linear Predictive Analysis (PLP) also able to mimic human auditory system, MFCC is still be used due to its most common feature extraction. The technique is widely spread so that it easier to develop and debug.

Two main reason using CNN. First, the concatenative use speech unit. Each of the speech unit is small part of sentences or word. CNN have been considered an optimum way to small-footprint keyword, speech unit, spotting than other machine learning approach [Sainath & Parada, 2015]. Second, MFCC extraction process can be plot into spectogram. CNN is the most common way to solve or analyzed visual imagenery data, including spectogram.

6 Conclusion

There are several conclusions that can be obtained from this research:

- This application able to collect speech data through website.
- This application able to train machine learning model with collected data through command prompt.
- This application able to mimic speech in Bahasa Indonesia through website.

- This application able to recognize speech from record audio although the prediction and accuracy are not perfect. But, the machine learning able to predict well most of the times.
- This application able to generate speech based on inputted text although the result still has silent or redundant part. But, the generated speech can be heard and understood.

In the future, a further research in the speech recognition is improving machine learning model. When there is no right or wrong in modelling the machine learning model, there is always optimal model to get the best prediction and accuracy. In the speech synthesis by removing some of silence or unused part of the speech and also reducing background will make generated speech more fluently and good to hear. In Bahasa Indonesia, a research in determining Bahasa Indonesia phonemes can be a big improvement since the application use syllables as concatenative synthesis.

7 Acknowledgements

The author would like to thank Mr. Tjong Wan Sen as the thesis advisor for his advices and support throughout the development of this thesis. Also thank to other lecturer for their support, knowledge, and experience during the university life.

References

[Hande, 2014] Hande, S. S. (2014). A Review on Speech Synthesis an Artificial Voice Production. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, 8.

[practicalcrypthography.com, 2012] practicalcryptography.com. (2012). *Mel Frequency Cepstral Coefficient (MFCC) tutorial*. Retrieved from Practical Cryptography: http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/

[Geitgey, 2016] Geitgey, A. (2016, Juny 14). *Machine Learning is Fun! Part 3: Deep Learning and Convolutional Neural Networks*. Retrieved from Medium: https://medium.com/@ageitgey/machine-learning-is-fun-part-3-deep-learning-and-convolutional-neural-networks-f40359318721

[Sainath & Parada, 2015] Sainath, T. N., & Parada, C. (2015). *Convolutional Neural Networks for Small-footprint Keyword Spotting.* New York, NY, U.S.A: Google, Inc.

[Dave & Pipalia, 2014] Dave, B., & Pipalia, C. D. (2014). SPEECH RECOGNITION: A REVIEW. *International Journal of Advance Engineering and Research*, 7