Speech recognition on Indonesian language by using time delay neural network

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY
1990

Bagus Tris Atmaja*,
, Fandy Akhmad, Dhany
Arifianto, Masato Akagi
*bagus@jaist.ac.jp

AIS-Lab
School of Information Science
JAIST

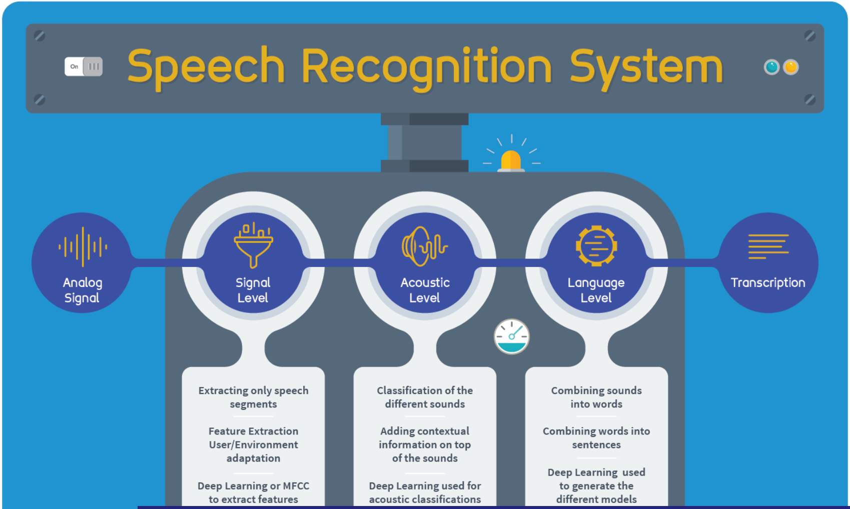
March 7, 2019

Motivation

- Ideally, speech should convey the correct message (intelligibility) while sounding like human speech (naturalness) with the right prosody (expressiveness).
- Most speech recognition engines which try to translate speech into correct message (text) require high computation load.
- By sampled the input as time series (time delay) in speech recognition (SR), the computation time can be reduced without having reduce the performance.
- Additionally, SR shows very good performance in well developed language, but lower performance on under resourced language.
- In this paper, we present our work on Indonesian speech recognition using time delay neural network (TDNN) to obtain text.

Speech Recognition:

The ability of a machine or program to identify words and phrases in spoken language and convert them to a machine-readable format



Speech Recognition System

On

Analog Signal → Signal Level → Acoustic Level → Language Level → Transcription

Signal Level

- Extracting only speech segments
- Feature Extraction User/Environment adaptation
- Deep Learning or MFCC to extract features

Acoustic Level

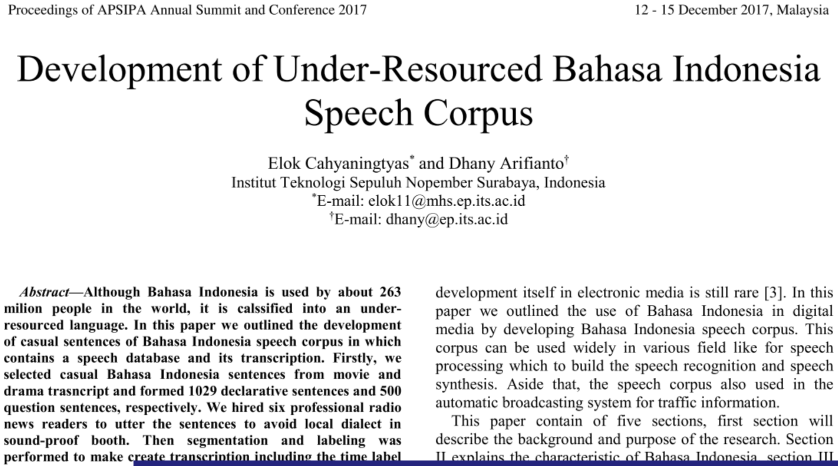
- Classification of the different sounds
- Adding contextual information on top of the sounds
- Deep Learning used for acoustic classifications

Language Level

- Combining sounds into words
- Combining words into sentences
- Deep Learning used to generate the different models

The Dataset

The dataset is obtained from the following paper. It is intended for speech synthesis that consist of clean speech only.



Proceedings of APSIPA Annual Summit and Conference 2017 — 12 - 15 December 2017, Malaysia

Development of Under-Resourced Bahasa Indonesia Speech Corpus

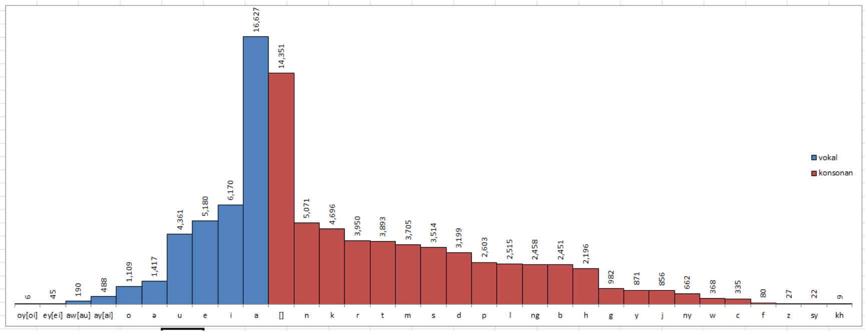
Elok Cahyaningtyas* and Dhany Arifianto†
Institut Teknologi Sepuluh Nopember Surabaya, Indonesia
*E-mail: elok11@mhs.ep.its.ac.id
†E-mail: dhany@ep.its.ac.id

Abstract**—Although Bahasa Indonesia is used by about 263 milion people in the world, it is calssified into an under-resourced language. In this paper we outlined the development of casual sentences of Bahasa Indonesia speech corpus in which contains a speech database and its transcription. Firstly, we selected casual Bahasa Indonesia sentences from movie and drama trasncript and formed 1029 declarative sentences and 500 question sentences, respectively. We hired six professional radio news readers to utter the sentences to avoid local dialect in sound-proof booth. Then segmentation and labeling was performed to make create transcription including the time label**

development itself in electronic media is still rare [3]. In this paper we outlined the use of Bahasa Indonesia in digital media by developing Bahasa Indonesia speech corpus. This corpus can be used widely in various field like for speech processing which to build the speech recognition and speech synthesis. Aside that, the speech corpus also used in the automatic broadcasting system for traffic information.

This paper contain of five sections, first section will describe the background and purpose of the research. Section II explains the characteristic of Bahasa Indonesia, section III

The dataset: distribution of phoneme[1]



- Total phoneme is 33 including silence (32 without silence).
- Total utterances is 1529 sentences for each speaker (10h 39m 5s).
- The sentences are taken from movie and drama.

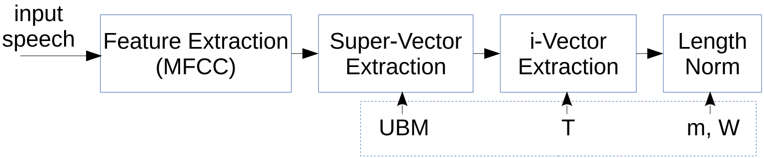
[1] E. Cahyaningtyas, D. Arifianto. Development of Under-Resourced Bahasa Indonesia Speech Corpus, in APSIPA-ASC, 2017.

The dataset: distribution of gender, train and dev

Gender	Train	Development
female 1	200	50
female 2	200	50
female 3	200	50
male 1	200	50
male 2	200	50
male 3	200	79
Total	1200	329

- Total of speaker is 6: 3 males and 3 females
- Form each speaker, we took 200 utterances for train, and 50 for development, except for the last one.

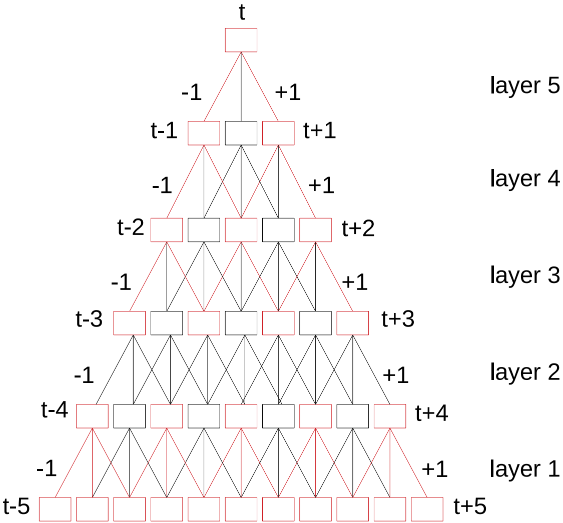
Feature Extraction[2]



- A number of 13 MFCCs are extracted including deltas and deltas-deltas.
- Channel supervector then is obtained from GMM/UBM.
- That channel supervector contains speaker (m) and session variabilities simultaneously (Tw).
- Given variability of matrix (T), obtain i-vectors (w) for each conversation side

[2] Richardson, Fred, Douglas Reynolds, and Najim Dehak. "A unified deep neural network for speaker and language recognition." arXiv preprint arXiv:1504.00923 (2015).

Architecture of TDNN



Result: Word Error Rate (WER)

Training approach	%WER
Monophone	37.17
Triphone 1	26.53
Triphone 2a	26.06
Triphone 2b	26.33
Triphone 3b	22.29
SGMM	18.22
TDNN	17.54

- Language model used: Iban language
- TDNN with 6 layers, each consist of 256 dimensions are trained to obtain the result.
- The result shows TDNN outperforms other training approaches.
- Total computation time: 26 minutes (10 minutes for SGMM + 6 minutes for ivector)

Result: Effect number of speakers

Number of speakers		Separation	%WER
Train	Dev		
4	2	Yes	33
6	6	No	17

- In the first result, we leave 2 speakers out (one for male and female).
- In the second result, we use the same speakers for training and development.

Conclusion

- Speech recognition based on TDNN is presented using Indonesian speech corpus.
- The result shows highest performance among other approaches
- Computation time is short (using small dataset), about 30 minutes using i9-7900X and GTX 1060.
- Number of speakers affect the result significantly, the more speakers the better result will be obtained as ivector learn speaker and environment specific information.