Journal of Software Engineering, Information and Communication Technology (SEICT)

Journal homepage: https://ejournal.upi.edu/index.php/SEICT

# Exploration of Spontaneous Speech Corpus Development in Urban Agriculture Instructional Videos

*Trisna Gelar*[*]*, Aprianti Nanda*

Department Informatics and Computer Engineering, Politeknik Negeri Bandung, Indonesia
Correspondence: E-mail: trisna.gelar@polban.ac.id[*]

## ABSTRACT

Video transcription can be obtained automatically based on the original language translation of the video maker's speech, but the quality of the transcription depends on the quality of the audio signal and the natural voice of the speaker. In this study, Deep Speech is used to predict letters based on acoustic recognition without understanding language rules. The Common Voice multilingual corpus helps Deep Seech to transcribe Indonesian. However, this corpus does not accommodate the special topic of urban agriculture, so an additional corpus is needed to build acoustic and language models with the urban agriculture domain. A total of 15 popular videos with closed captions and nine E-Books with the theme of Horticulture (fruit, vegetables and medicinal plants) were curated. The video data were extracted into audio and transcription according to specifications as training data, while the agricultural text data were transformed into language models, which were used to predict recognition results. The evaluation results show that the number of epochs has an effect on improving the transcription performance. The language model score used during prediction improved WER performance as it interpreted words with agricultural terms. Another finding was that the model was unable to predict short words with informal varieties and located at the end of the sentence.

## ARTICLE INFO

***Article History:***
*Submitted/Received 25 Mar 2021*
*First Revised 02 Apr 2021*
*Accepted 05 Apr 2021*
*First Available online 10 May 2022*
*Publication Date 01 Jun 2022*

***Keyword:***
*Acoustic models,*
*Corpus exploration,*
*Language models,*
*Spontaneous speech,*
*Urban agriculture.*

## 1. INTRODUCTION

People do urban farming activities to utilise their free time, or to find a new hobby. One of the learning media used is instructional videos that are widely available and free. In this case, YouTube channels are generally used to learn new things and/or find solutions to problems related to work, school or hobbies (Vingilis et al., 2017).

The quality of instructional video content on Youtube varies, depending on uploads from users or user generated content, whether beginners, professionals, individuals or communities. The availability of transcriptions or subtitles on videos is one of the physical components that determine whether the video is liked or not (Shoufan, 2019). Transcriptions on videos can make it easier to find content by search engines or improve access for people with hearing impairments. Transcriptions can help viewers understand a process or activity.

Video transcriptions can be obtained automatically based on the original language translation of the video creator's speech, but the quality of the transcription is highly dependent on the audio quality of the video. The process of audio-to-text transformation or Automatic Speech Recognition (ASR) depends on whether the audio signal has variations or noise. The ASR problem becomes more difficult when the speech is continuous and recorded under diverse noise conditions from various sources (Huang et al., 2016). This condition is found in every urban agriculture instructional video, where the speaker speaks spontaneously or informally outdoors.

ASR research activities for Indonesian focus on planned and clearly spoken speech (Budiman and Widjaja., 2020) and general themes (Benkerzaz et al., 2019). However, there is a lack of research on spontaneous or informal speech recognition. The main obstacle is the availability of training data in Bahasa Indonesia which is a language with limited resources (Prakoso et al., 2016). In addition, for domain-specific conversations such as urban agriculture, additional vocabulary or specialised terms are required.

One of the opensource ASR tools is Mozilla DeepSpeech (Batternberg et al., 2017), which can predict letters based on sound without the need to understand language rules. The DeepSpeech model along with the Common Voice multilingual dataset (Zhang et al., 2019) can be used to build ASR models for other languages using transfer learning methods from English. However, if implemented directly into Indonesian, this can cause spelling errors or produce meaningless word sequences in the transcription results. To correct this, it is necessary to develop acoustic and language models related to agricultural topics.

This research aims to develop a corpus of this study explores the procedure of converting Youtube video data into audio and transcribing text into acoustic models and agricultural text into language models. The spontaneous speech corpus is taken from Youtube videos of Agriculture city which is a curated playlist while the language model is sourced from books with agricultural themes (Watanabe et al., 2017).

As the public's interest in urban agriculture increases, recommendations for good quality instructional videos should be improved. One way is to add video transcription to the video. In addition, research can be continued to obtain a summary of the context of the video and can be saved into a video description (Aafaq et al., 2020).

DOI: https://doi.org/10.17509/seict.v3i1.44548
p-ISSN:2744-1656 e-ISSN:2774-1699

## 2. METHODS

Quantitative methods are used in this research, the performance of the urban agriculture instructional video corpus model will be measured empirically. **Figure 1** is an overview of the research stages.

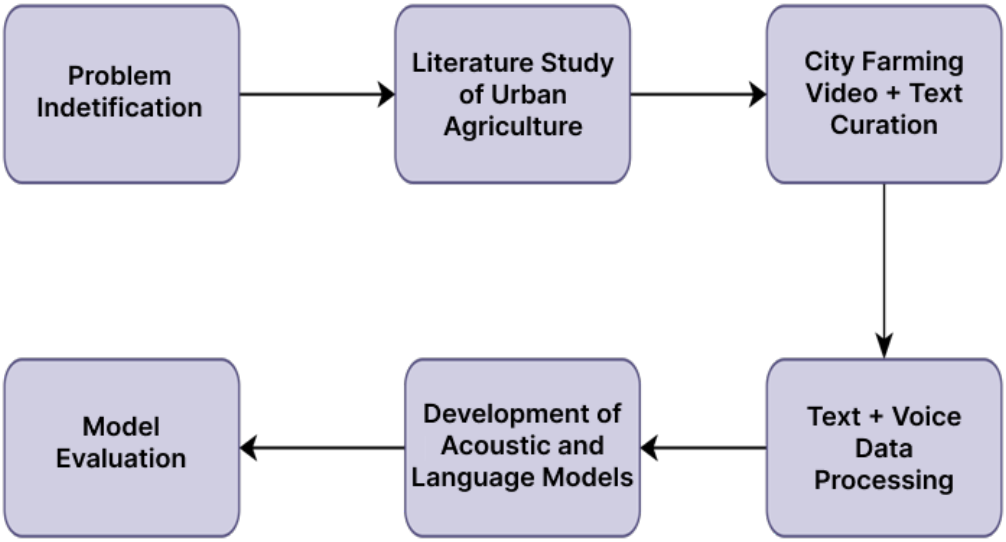


**Figure 1.** Research method of urban agriculture corpus development

### 2.1. Literature Study

#### 2.1.1. Instructional Video

There are two types of instructional video delivery: the first is practising the activity in real time/spontaneity. The second is delivered using a structured narrative, including a timeline of sub-tasks (Chang et al., 2017). The second type is reported to have a higher level of satisfaction as the learner can interact step-by-step. Unfortunately, not all video tutorials on YouTube are of this quality. Each video content on Youtube is uploaded by beginners and professionals, so the quality of the content delivered is different (Suryani, 2022).

#### 2.1.2. Spontaneous Conversation Corpus Development

Spontaneous speech is usually uttered by someone without planning what to say beforehand (Lapasau and Setiawati, 2020). In the context of instructional videos, spontaneous speech occurs when the producer performs live activities or one take of a video recording, informal vocabulary, reflective reactions, or local language.

The SMASH corpus (Lotfian and Busso, 2017) includes spontaneous speech from two Japanese male commentators making third-person audio commentary during a professional gaming competition. Each commentator made spontaneous comments while watching the game, not only about the fight, the commentators also entertained the audience. In addition

DOI: https://doi.org/10.17509/seict.v3i1.44548
p- ISSN 2774-1656 e- ISSN 2774-1699

to audio transcription, the researcher tagged the topics discussed. There are two annotation methods: automatic and manual (see **Figure 2**).

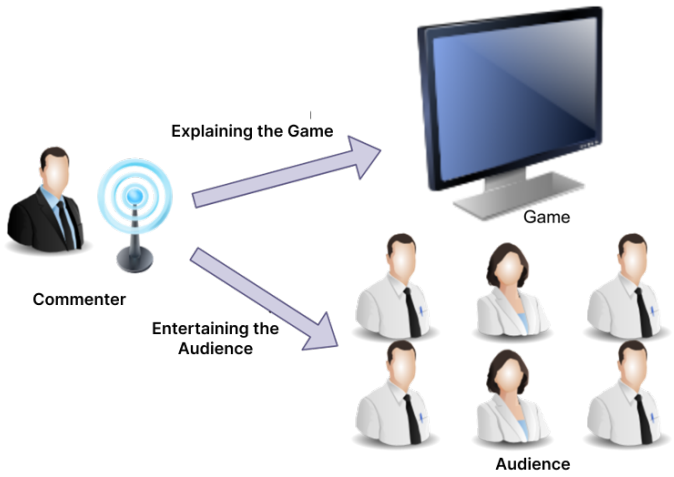


**Figure 2.** The process of commentators commenting on game competitions.

The SMASH corpus building process consists of four main steps, namely, 1) Data Creation, researchers curated the Legendary Pokemon game competition event commented by two professionals and divided the match scene into three rounds. 2) Transcription, performed automatically from audio to Japanese text using a cloud-based speech-to-text (STT) service (Google Cloud STT3). 3) Refine transcription is done by two volunteers to correct typos or word errors that are generated automatically. 4). The audio corpus and transcriptions were collected (see **Figure 3**).

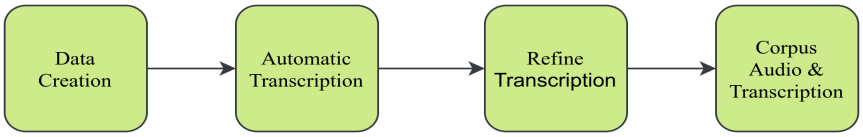


**Figure 3.** SMASH corpus creation process.

### 2.1.3. Automatic Speech Recognition

Automatic speech recognition (ASR) is a technology that enables interaction between humans and computers through voice (Dyarbirru and Hidayat, 2020). Google voice search applies an example of ASR technology that converts voice into text to perform everyday commands on mobile devices. In addition, in the education domain, ASR is used to transcribe teaching and learning activities into a set of indexed text so that it can be easily searched by students.

**Figure 4** is the ASR architecture (Huang et al, 2016), there are four processes, namely; 1). The process starts with the transformation of the audio signal using signal processing and feature extraction techniques so that it conforms to the acoustic model (AM) format. 2) The acoustic model combines knowledge about the sound can be acoustic or phonetic based on the features extracted in the previous stage and generates a score for the feature. 3) The language model measures the likelihood or score of combinations of word sequences. 4) The hypothesis generation process combines the AM and LM scores to produce a hypothesis of what word was spoken according to the audio signal.

DOI: https://doi.org/10.17509/seict.v3i1.44548
p-ISSN:2744-1656 e-ISSN:2774-1699

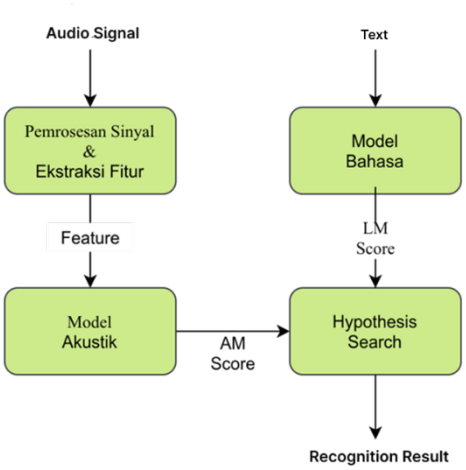


**Figure 4.** Automatic speech recognition architecture.

### 2.1.4. Audio and Language Models in Mozilla Deep Speech

Common Voice is a multilingual dataset of transcribed, community-based, Creative Commons Zero (CC0) licensed audio talks built by Mozilla (Handoko and Suyanto, 2019). The Common Voice Indonesian dataset consists of 54 unique voices with a total of 5 hours of speech and 4 hours of validation (Dyarbirru and Hidayat, 2020). The data obtained is the result of crowdsourcing, for several languages the Mozilla Deep Speech and Common Voice models produce an average CER improvement of 5.99 ± 5.48 (Tachbelie et al, 2022).

A common way of statistical language modelling is to measure the probability of an n-word sequence (bigrams, trigrams, etc.) occurring in a sentence. The probability measure is based on maximum likelihood estimation which depends on the available training data.

In Deepspeech, KenLM (Ruder et al, 2019) is used to process n-grams which is efficient in both time and memory resources used. PROBING and TRIE, a data structure designed to optimise memory and CPU usage. trie results from KenLM will be used as the basis for building the Language Model.

### 2.1.5. Automatic Speech Recognition Performance Calculation

Word error rate (WER) and Character error rate (CER) are performance measures in automatic spontaneous speech recognition (Bang et al., 2020; Besacier et al., 2014), WER measures the performance of correctly recognised word order predictions, while CER is based on phoneme order. Both metrics are derived from the Levenshtein distance formula and are useful for evaluating improvements to acoustic models.

Word error rate (WER) formula (Andrew C et al., 2004) is used to measure the ratio of prediction errors at the word level, a good WER value is close to zero, expressed in Equation.

$$WER/\ CER\ = \frac{S+D+1}{N} = \frac{S+D+1}{S+D+C}$$

S is the number of substituted words, D is the number of deleted words, I is the number of inserted words and C is the number of correct words, and N is the number of words in the tested sentence (N=S+D+C). Character error rate (CER) formula is used to measure the ratio/percentage of prediction errors at the phoneme level, a good CER value is close to zero.

DOI: https://doi.org/10.17509/seict.v3i1.44548
p- ISSN 2774-1656 e- ISSN 2774-1699

### 2.2. Acoustic Model Development

**Figure 5** illustrates the processing of curated agricultural instructional videos into acoustic models. There are four main stages, namely; 1). Agricultural Video Curation with 2). Data Processing, transcription using Google Speech to Text API automatically and corrected after listening to the audio equivalent. While the audio is downloaded in mp3 format and converted according to the Common Voice dataset specifications. 3). Once the audio corpus and transcriptions are collected, the file name, size and transcription text information are stored in a csv file. 4) Acoustic models of the audio corpus and transcriptions will be built using Mozilla Deepspeech.

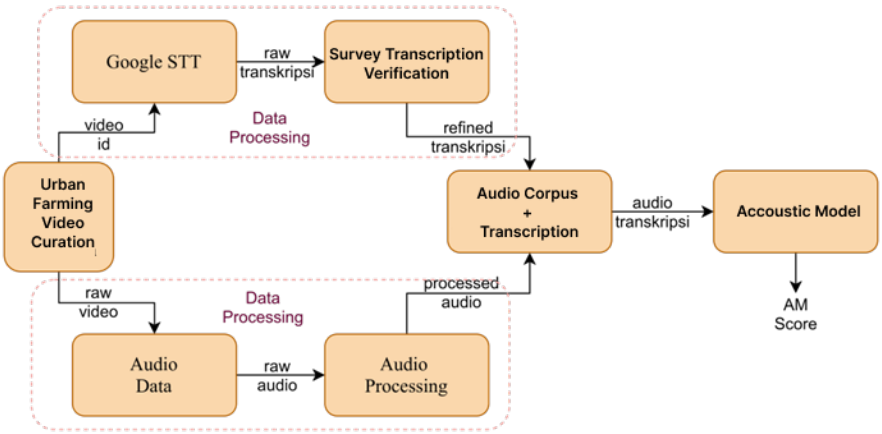


**Figure 5.** Agricultural video processing into acoustic models

### 2.3. Language Model Development

**Figure 6** illustrates the processing of agricultural book texts into language models. There are five stages, namely; 1). Curation of agricultural textbooks (PDF) 2). Extracting sentences on each page of the electronic book into text. 3). Performing text cleaning on the extraction results. 4) The text is collected into one agricultural text corpus document. 5) The agricultural text is processed into a language model using *KenLM*.

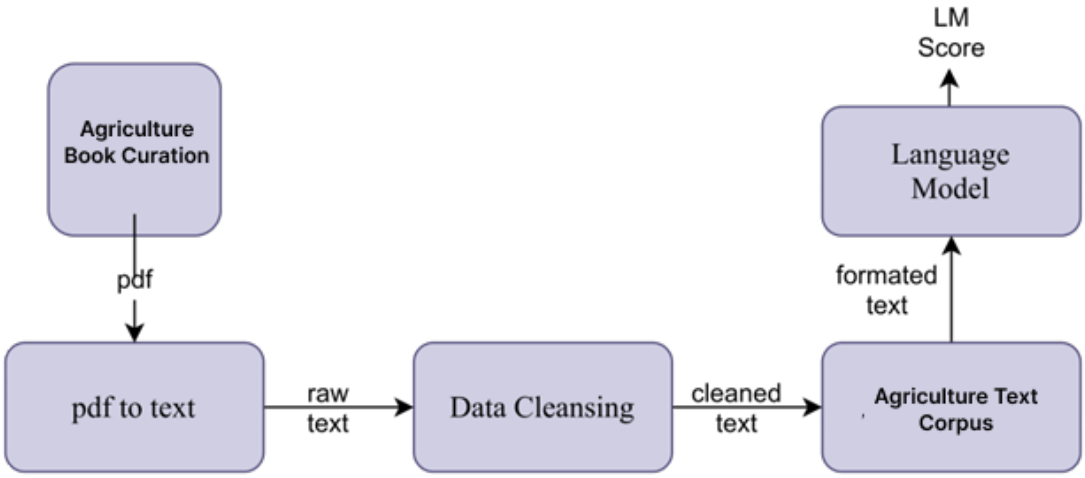


**Figure 6.** Processing of agricultural text into language models.

DOI: https://doi.org/10.17509/seict.v3i1.44548
p-ISSN:2744-1656 e-ISSN:2774-1699

## 3. RESULT AND DISCUSSION

### 3.1. Urban Agriculture Video and Text Data Curation

The acoustic corpus was obtained from audio and transcriptions of popular Youtube Instructional Videos, while the language corpus was obtained from Urban Agriculture-themed Books.

#### 3.1.1. Urban Agriculture Text Data Curation

Books with the theme of Horticulture (fruit, floriculture, vegetables and medicinal plants) **Table 1** were taken from the repository of the Ministry of Agriculture of the Republic of Indonesia (see http://repository.pertanian.go.id/handle/123456789/7076 ) as a source of text for the development of the language model. Each book has a PDF format so that further processing is needed so that text specific to agricultural terms can be extracted. In addition, the authors conducted manual preprocessing such as making bookmarks to separate less relevant parts of the book such as covers, preface, table of contents, tables that are difficult to extract directly by machine and bibliography (Permatasari and Linawati, 2021).

**Table 1**. Curation results of urban agriculture textbooks

| No | Title | File Type | Bookmark | Page |
|---|---|---|---|---|
| 1 | Mango Cultivation and Postharvest | Image PDF | 7 | 33 |
| 2 | Citrus Cultivation | Image PDF | 5 | 13 |
| 3 | Litchi Cultivation Fieldbook | Image PDF | 17 | 67 |
| 4 | Good Handling Practices for Horticultural Commodities | Image PDF | 5 | 22 |
| 5 | Citrus Nutrient Deficiency | Image PDF | 4 | 41 |
| 6 | Dragon Fruit Cultivation | Image PDF | 5 | 43 |
| 7 | Providing Planting Media for Ornamental Leaf Plants | Image PDF | 5 | 19 |
| 8 | A Practical Guide to Durian Cultivation | Image PDF | 17 | 58 |
| 9 | Technical Guidelines for Passion Fruit Cultivation | Image PDF | 7 | 37 |

#### 3.1.2. Video Data Curation and Transcription

There are three data curation activities for urban agriculture instructional videos adopted from research (Gelar and Nanda, 2020), namely determining urban agriculture keywords, collecting videos on Youtube playlists, selecting fifteen videos based on the order of the highest views and likes using the Youtube API (Novendri et al, 2020), another requirement is that each video has a close caption feature so that transcription can be extracted automatically using the Google Speech to Text (Google STT) API. **Table 2** is the result of curation of agricultural instructional videos taken from the playlist, originating from the Youtube channel Info Ragam Pertanian, uploaded in 2021, (see https://www.youtube.com/playlist?list=PLcCat4rDFj3oB34O816CLu9pDSUwBZIK7).

DOI: https://doi.org/10.17509/seict.v3i1.44548
p- ISSN 2774-1656 e- ISSN 2774-1699

**Table 2.** Writing the table of results of the Agricultural Instructional Video Curation

| No | ID | Title | View | Like | Comment |
|---|---|---|---|---|---|
| 1 | x2c9ztok3MI | FARM TIPS How to Prune Tomato Leaves for Large, Heavy Fruits Disease-Free & Longevity | 3892 | 81 | 32 |
| 2 | LR7xT81DekU | Growing Semi-Organic Watermelon in No-till Soil Cost-effectively Abundant Results #PART 1 | 5299 | 75 | 16 |
| 3 | yxzYq4UHOEQ | How to prune chillies and large, curly chillies so that they are abundant and long-lived | 32207 | 542 | 69 |
| 4 | Yanzq_oglDk | Rice water POC fertiliser for chillies tomatoes & aubergines in the field to be fertile and free of pests and diseases | 4082 | 81 | 27 |
| 5 | F0tRmeqGZbw | The impact of nutrient deficiencies on yellow chilli leaves causes fruit & leaves to fall easily | 5045 | 85 | 35 |
| 6 | 0Auvzs07pTI | THE SECRET TO GROWING Lush Fruits & Anti-curly Leaves without Chemical Pesticides | 11332 | 241 | 47 |
| 7 | yAqHO6iXaPs | How to overcome the stunted growth of chilli leaves healed in 1 week plants back healthy & fertile | 112037 | 1289 | 218 |
| 8 | kzmXyugNhgA | High potassium organic fruit fertiliser for chilli peppers to produce abundant fruits that are resistant to falling off and free from mildew. | 15550 | 256 | 46 |
| 9 | OXZL0H8mf9g | Dissolving PETROGANIC & AMINO Fertilisers For Instant POC On Stunted Chilli Plants See The Results | 5900 | 119 | 68 |
| 10 | kG1YgTv5DV0 | How to Plant Organic Cayenne Pepper in Polybags Directly Fruit ANTI Leaf Curl #PART.1 | 3893 | 85 | 23 |
| 11 | gBGs_G6jgIQ | Tips for Success from Zero in Growing Organic Raw Chillies in a High-Fruiting Field Free of Leaf Curl & Pathogens | 8664 | 213 | 66 |
| 12 | NQ7Agi_LVsQ | CABE Land Preparation Without Tillage & Basic Fertiliser Still Fertile Full Organic | 5135 | 157 | 51 |
| 13 | EMSNXI9Yhc8 | The first organic fertiliser to fertilise cayenne pepper in polybags, homemade POC material PART.2 | 5645 | 112 | 35 |
| 14 | oiL_UNz2Lo8 | Tutorial for Growing Cayenne Pepper in the Field FULL ORGANIC TOT Save Cost Abundant Results #PART 2 | 5896 | 151 | 65 |
| 15 | SDHJVYSyzhc | Successful Harvesting of Large Chillies & CMK 4000 Trees 2 x a week in the Rainy Season Heavy Fruit Anti Falling off | 4871 | 103 | 34 |

### 3.2. Data Processing

To build a corpus, acoustic model and language model, agricultural data suitable for machine learning, such as converting video into audio and transcription and dividing data into training, test and validation data. As well as converting pdf-formatted electronic books into agricultural texts that can be processed into language models.

DOI: https://doi.org/10.17509/seict.v3i1.44548
p-ISSN:2744-1656 e-ISSN:2774-1699

### 3.2.1. Agricultural Video to Audio Processing and Transcription

There are two main processes in this phase, namely:

1) Extracting audio and text transcriptions from curated videos and manually verifying the transcriptions using Google Forms. There are six procedures that have been implemented as follows:
   a. Downloading mp3 audio files from curated videos and retrieving transcription files with the help of the YouTube-dl library and YouTube Transcript API.
   b. Convert audio files according to common voice dataset library specifications (mono channel, sample rate 16 Khz and wav file format).
   c. Cutting audio files according to the timestamp on the transcription which consists of start, duration, and stop.
   d. Filtering audio with a speech duration of 2-4 seconds, because the observation results, other than this duration, the speech is short (1-2 words) or there is a long background music.
   e. Cleaning the transcription text data including case folding and removing punctuation marks.
2) Create a Google Form (List video id, sequence, timestamp, file size, splitted audio, sentence, file path) verifying transcription on each audio file to improve the results of automatic transcription (converting numbers/symbols/units into words; for example 1 kg into one kilogram). Unify the audio and transcription into a file structure that meets the Deep Speech library specification.
   There are three procedures that have been implemented as follows:
   a. Collect audio files into 1 folder with the name clips.
   *b.* Combined the final results of the transcription review into one csv format file consisting of file path, file size and text (transcription).
   **c.** Separating the csv file into 3 files that are prepared for training, validation and testing.

### 3.2.2. Agricultural Book Text Processing

There are two main processes in this phase, namely:

1) Perform text extraction from PDF formatted agricultural books, perform image to text extraction to get raw text with the Indonesian language tesseract-ocr library found in easyocr (with bounding box text parameters, namely: min_size: 0, slope_ths: 0.1, y_center_ths: 0.5, height_ths: 0.5, widh_ths: 0.7, decoder: beam search). The extraction process is performed for each image in the bookmark, and then aggregated back into a text collection or agricultural corpus.
2) Text data cleaning such as case folding, removing punctuation is done automatically and converting numbers, symbols and units into words and removing other irrelevant words (residual text that does not contain meaning) is done manually. The statistics of the processed agricultural text corpus are 6,257 sentences, 35,243 words and 252,366 characters. However, the extraction results cannot detect paragraphs (Table 3), so there are truncated sentences that affect the formation of n-grams.

DOI: https://doi.org/10.17509/seict.v3i1.44548
p- ISSN 2774-1656 e- ISSN 2774-1699

**Table 1.** Example sentence of agriculture corpus.

| No | Sentence |
|---|---|
| 1 | mangifera indica mango fruit is one of the fruits that |
| 2 | popular in Indonesia in the period of mango harvest area tends to be |
| 3 | up with production ranging between thousand tonnes when compared with |

### 3.3. Language Data Modelling

There are three main processes in this phase, namely:

1) Language modelling using KenLM tools with parameters implemented as follows: (top-k: 5000, arpa_order 5 max_arpa_memory "85%" --arpa_prune "0|0|1") resulting in three files namely (Trie, Arpa and lm.binary), the results of n-gram combinations as follows: ngram 1=4896, ngram 2=23168, ngram 3=2537, ngram 4=834 and ngram 5=314. Table 4 is an example of an n-gram word combination.
2) The Language model optimisation process is used to obtain the optimal lm_alpha and lm_beta parameters, used when creating the scorer for 10 trials.
3) The scorers of vocab, lm.binary and parameters lm_alpha and lm_beta obtained in the previous process are used during evaluation to improve / give meaning to the character prediction.

**Table 2.** Arpa's excerpt from KenLM's urban farming text.

| N-gram | Word |
|---|---|
| 1-gram | Introduction, fruit, mango |
| 2-gram | Fruit addressed,<br>chicken eggs,<br>free-range chicken |
| 3-gram | passion fruit production,<br>passion fruit juice,<br>and passion fruit konyal |
| 4-gram | raw husk roasted humus,<br>durian fruit borer seeds,<br>suitable for durian orchards,<br>practical guide to durian cultivation |
| 5-gram | and adults will damage the plant,<br>physical and physiological characteristics of the plant,<br>which are commonly used for crops,<br>fertiliser rate table for crops |

### 3.4. Acoustic Data Modelling

The Acoustic model training process was carried out with 2 main scenarios, namely using the Indonesian Common Voice dataset and a combination of Common Voice (CV) and City

DOI: https://doi.org/10.17509/seict.v3i1.44548
p-ISSN:2744-1656 e-ISSN:2774-1699

Farm Audio Data (PT) with training parameters on DeepSpeech adopted from the documentation. (see https://deepspeech.readthedocs.io/en/v0.9.3/TRAINING.html and playbook (see https://mozilla.github.io/deepspeech-playbook/) (see **Table 5**).

**Table 3.** Acoustic model training parameters.

| No | Dataset | Epoch | Hidden Layers | Others Parameter |
|---|---|---|---|---|
| 1 | CV | 5 | 64 | |
| 2 | CV | 10 | 64 | --reduce_lr_on_plateau true \ |
| 3 | CV | 15 | 64 | --plateau_epochs 2 \ |
| 4 | CV+PT | 5 | 64 | --plateau_reduction 0.06 \ |
| 5 | CV+PT | 10 | 64 | --early_stop true \ |
| 6 | CV+PT | 15 | 64 | --es_epochs 5 \ |
| | | | | --es_min_delta 0.03 --dropout_rate 0.2 |

### 3.5. Data Evaluation

Based on **Table 6**, the performance comparison of acoustic and language models based on the number of epochs and the use of scorers (urban farming language model). The CER and WER values on both the CV dataset and the combined dataset decrease (get better) with each increase in the number of epochs (5, 10 and 15).

**Table 4.** Compilation of CER and WER comparison table of city agriculture corpus.

| No | Data | Epoch | No Scorer | | Scorer | |
|---|---|---|---|---|---|---|
| | | | CER | WER | CER | WER |
| 1 | CV | 5 | 0.811 | 1 | 0.938 | 0.986 |
| 2 | CV | 10 | 0.620 | 1.01 | 0.650 | 0.942 |
| 3 | CV | 15 | 0.598 | 1.02 | 0.617 | 0.925 |
| 4 | CV+PT | 5 | 0.508 | 0.99 | 0.615 | 0.895 |
| 5 | CV+PT | 10 | 0.490 | 0.99 | 0.593 | 0.870 |

DOI: https://doi.org/10.17509/seict.v3i1.44548
p- ISSN 2774-1656 e- ISSN 2774-1699

| | | | | | | |
|---|---|---|---|---|---|---|
| 6 | CV+PT | 15 | 0.483 | 0.99 | 0.590 | 0.869 |

The performance of WER with a scorer on the combined dataset is better because the language model can correctly predict meaningful words/sentences on the theme of urban agriculture, compared to the performance of CER measured by pronunciation accuracy, which is not better than the evaluation results without a scorer.

**Table 7** is a head-to-head snapshot of the prediction results of 5 sentences using the combined dataset model, epoch 15 with and without scorer. It can be seen that despite having a better CER value, the words/sentences produced by the model without scorer do not contain the correct meaning. The model with scorer cannot predict short informal speech words such as this, well and can't as well as short words such as two, this, one at the end of the sentence.

**Table 5.** Comparison of CER and WER of City Agriculture Corpus

| No | Original Sentence | Sentence Without Scorer | Sentence with Scorer |
|---|---|---|---|
| 1 | We can do fertilisation in two ways | sekapumetekan disatitaatukan aradua | fertilisation we can do there |
| 2 | first fertiliser nah fertiliser | pemukan pekama yakaepukan se | first fertiliser fertiliser |
| 3 | from sprout lodging disease then | daii apenya tidebakecamatemedian pe | of the disease then then |
| 4 | can't give as much as one | disatitamenberikan sebana saten | can we give as much as |
| 5 | with water on the main stem like this | lasekai paa batantitamasepatini | flavours on the main stem such as |

## 4. CONCLUSION

There are four main conclusions in the research related to the stages of data curation, data processing, model development and evaluation. First, audio data curation and transcription of the 15 most popular YouTube videos that have Indonesian closed caption attributes. Curation of agricultural text data from 9 Horticulture theme books in PDF format from the Ministry of Agriculture repository, each book is given bookmarking and deleting irrelevant information. Second, Audio data and transcriptions from YouTube videos have been processed and prepared according to the format that will be used as training, validation and test data for acoustic model development while agricultural book data is converted into clean text and has been processed for language model development. Third, the city agricultural audio data and the Indonesian Common Voice dataset have been used to build the acoustic model, a total of 6 training scenarios have been implemented based on data usage and number of epochs. The agricultural language data has been transformed into a language

DOI: https://doi.org/10.17509/seict.v3i1.44548
p-ISSN:2744-1656 e-ISSN:2774-1699

model score, used during evaluation, serving to give meaning to the predicted words. Fourth, the evaluation results show that the model with the highest epoch on each data is the best model, the WER and CER values always decrease with each epoch. CER performance on tests without language models is always better but the predicted sentences do not contain meaning. While the WER value in the test with the language model produces a better and meaningful value. Another finding is that the model cannot predict short informal words or sentence endings.

To improve the shortcomings, there are four possible solutions, namely; First, Adding short word spontaneous speech text such as (well, this, that and others) in addition to the text of agricultural terms. Second, Improving the procedure of automatically converting typical agricultural words/symbols/units using the Name Entity Recognition technique or agricultural dictionary lookup in a sentence, Third, Improving the word or paragraph beheading process to reduce truncated sentences when extracting pdf to text. Fourth, exploring fine tuning of acoustic models with hyperparameter optimisation, including the number of epochs, hidden layers, early stops, and others.

## 5. ACKNOWLEDGEMENTS

Praise be to God Almighty, this Independent Research Programme (PM) can be carried out well. The PM Scheme funds are sourced from DIPA Politeknik Negeri Bandung with an activity implementation agreement letter Number: 105.78/PL1.R7/PG.00.03/2021. We thank you for the participation of all parties for the implementation of this research.

## 6. AUTHORS' NOTES

The authors declare that there is no conflict of interest regarding the publication of this article. The authors confirm that this paper is free from plagiarism.

## 7. REFERENCES

Aafaq, N., Mian, A., Liu, W., Gilani, S. Z., and Shah, M. (2019). Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys (CSUR), 52*(6), 1-37.

Bang, J. U., Yun, S., Kim, S. H., Choi, M. Y., Lee, M. K., Kim, Y. J., and Kim, S. H. (2020). Ksponspeech: Korean spontaneous speech corpus for automatic speech recognition. *Applied Sciences*, *10*(19), 1-17.

Benkerzaz, S., Elmir, Y., and Dennai, A. (2019). A study on automatic speech recognition. *Journal of Information Technology Review*, *10*(3), 77-85.

Besacier, L., Barnard, E., Karpov, A., and Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, *56*, 85-100.

Budiman, V. E., and Widjaja, A. (2020). Building acoustic and language model for continuous speech recognition in bahasa Indonesia. *Jurnal Teknik Informatika dan Sistem Informasi*, *6*(2), 301-310.

Chang, Y. J., Paruthi, G., Wu, H. Y., Lin, H. Y., and Newman, M. W. (2017). An investigation of using mobile and situated crowdsourcing to collect annotated travel activity data in real-word settings. *International Journal of Human-Computer Studies*, *102*(C), 81-102.

DOI: https://doi.org/10.17509/seict.v3i1.44548
p- ISSN 2774-1656 e- ISSN 2774-1699

Dyarbirru, Z., and Hidayat, S. (2020). Metode wavelet-MFCC dan korelasi dalam pengenalan suara digit. *JTIM: Jurnal Teknologi Informasi dan Multimedia*, *2*(2), 100-108.

Handoko, I. T., and Suyanto, S. (2019). Klasifikasi gender dan usia berdasarkan suara pembicara menggunakan hidden markov model. *Indonesia Journal on Computing (Indo-JC), 4*(3), 99-106.

Huang, Z., Siniscalchi, S. M., and Lee, C. H. (2016). A unified approach to transfer learning of deep neural networks with applications to speaker adaptation in automatic speech recognition. *Neurocomputing*, *218*, 448-459.

Gelar, T., and Nanda, A. (2020). Klasifikasi komentar video instruksional populer bertemakan pekarangan perkotaan menggunakan auto-keras. *Journal of Software Engineering, Information and Communicaton Technology (SEICT), 1*(1), 1-9.

Lapasau, M., and Setiawati, S. (2020). Slips of the tongue in Indonesia daily conversation: a psycholinguistic view. *Hortatori: Jurnal Pendidikan Bahasa dan Sastra Indonesia*, *4*(2), 127-132.

Lotfian, R., and Busso, C. (2017). Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, *10*(4), 471-483.

Novendri, R., Callista, A. S., Pratama, D. N., and Puspita, C. E. (2020). Sentiment analysis of YouTube movie trailer comments using naïve bayes. *Bulletin of Computer Science and Electrical Engineering*, *1*(1), 26-32.

Permatasari, P. A., and Linawati, L. J. (2021). Survei tentang analisis sentimen pada media sosial. *Maj. Ilm. Teknol. Elektro, 20*(2), 177-185.

Ruder, S., Vulić, I., and Søgaard, A. (2019). A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, *65*, 569-631.

Shoufan, A. (2019). Estimating the cognitive value of YouTube's educational videos: A learning analytics approach. *Computers in Human Behavior*, *92*(C), 450-458.

Suryani, N., Suryono, J., Gama, B., and Setyo, B. (2022). Faktor-faktor kebijakan redaksional youtube solopos dalam menentukan berita vaksin covid–19. *Media and Empowerment Communication Journal*, *1*(1), 15-24.

Tachbelie, M. Y., Abate, S. T., and Schultz, T. (2022). Multilingual speech recognition for GlobalPhone languages. *Speech Communication*, *140*, 71-86.

Vingilis, E., Yıldırım-Yenier, Z., Vingilis-Jaremko, L., Wickens, C., Seeley, J., Fleiter, J., and Grushka, D. H. (2017). Literature review on risky driving videos on YouTube: Unknown effects and areas for concern. *Traffic Injury Prevention*, *18*(6), 606-615.

Watanabe, S., Hori, T., Kim, S., Hershey, J. R., and Hayashi, T. (2017). Hybrid CTC/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, *11*(8), 1240-1253.

DOI: https://doi.org/10.17509/seict.v3i1.44548
p-ISSN:2744-1656 e-ISSN:2774-1699