Indonesian-English Code-Switching Speech Recognition using the Machine Speech Chain based Semi-Supervised Learning

Rais Vaza Man Tazakka[1*], Dessi Lestari[1], Ayu Purwarianti[1], Dipta Tanaya[2], Kurniawati Azizah[2], Sakriani Sakti[3,4]

[1]Institut Teknologi Bandung, Indonesia
[2]University of Indonesia, Indonesia
[3]Japan Advanced Institute of Science and Technology, Japan
[4]Nara Institute of Science and Technology, Japan
13519060@std.stei.itb.ac.id, {dessipuji,ayu}@itb.ac.id,
{diptatanaya,kurniawati.azizah}@cs.ui.ac.id, ssakti@jaist.ac.jp

Abstract

Indonesia is home to a diverse linguistic landscape, where individuals seamlessly transition between Indonesian, English, and local dialects in their everyday conversations—a phenomenon known as code-switching. Understanding and accommodating this linguistic fluidity is essential, particularly in the development of accurate speech recognition systems. However, tackling Indonesian-English code-switching poses a challenge due to the scarcity of paired code-switching data. Thus, this study endeavors to address Indonesian-English code-switching in speech recognition, leveraging unlabeled data and employing a semi-supervised technique known as the machine speech chain. Our findings demonstrate that the machine speech chain method effectively enhances automatic speech recognition (ASR) performance in recognizing code-switching between Indonesian and English, utilizing previously untapped resources of unlabeled data.

Keywords: code-switching, speech recognition systems, machine speech chain

1. Introduction

The advancement in speech processing technology has enabled machines to process and respond to human speech, such as automatic speech recognition (ASR) systems, which can transcribe spoken audio into a corresponding sequence of words (Keshet and Bengio, 2009). There are also text-to-speech (TTS) systems that can generate synthetic speech for a given text input.

Several approaches can be used to develop a speech recognition system. However, with the emergence of deep learning, many state-of-the-art speech recognition models are built using neural network-based approaches (Tjandra et al., 2020).

In most cases, a speech recognition model is trained for one language only. For example, a speech recognition model trained exclusively for the Indonesian language can only recognize Indonesian. It cannot recognize a speech comprising more than one language such as a code-switching speech.

Code-switching is a phenomenon of alternating between two or more languages in a conversation (Nakayama et al., 2019). This phenomenon can be found in the communication of the Indonesian community, as observed in Margana (2013), which documented the phenomenon of Indonesian-English code-switching in several educational institutions in the Special Region of Yogyakarta Province. Code-switching is a very common phenomenon in Indonesia since many Indonesians use several different languages in their daily conversations involving Indonesian, English, and local languages.

Phonetic-wise, the Indonesian and English languages have different sets of phonemes which can be seen in Table 1 (Andi-Pallawa and Alam, 2013). The English language has æ, ʌ, ɜ, v, θ, and ð which are not present in the Indonesian phonological system. There are also several important things to note as explained in Andi-Pallawa and Alam (2013): (1) Phonetic features b, d, g, z, s, ʧ, ʤ do not exist in the final position of Indonesian words; (2) p, t, k are never aspirated in Indonesian words; and (3) r is pronounced clearly in Indonesian, unlike in English.

Handling code-switching Indonesian-English speech is important since several words have the same pronunciation in both languages while referring to completely different meanings. Examples of Indonesian and English words that have the same pronunciation but have different meanings are given in Table 2. Failing to handle code-switching speech may result in a wrong speech recognition.

Despite the importance of handling code-switching in a speech recognition system, there are not much labeled code-switching Indonesian-English data. Therefore, this study aims to handle the code-switching phenomenon in a speech

*This work was conducted while the first author was doing internship at HA3CI Laboratory, JAIST, Japan under JST Sakura Science Program.

SIGUL2024 Workshop, pages 143–148
21-22 May, 2024.

Phoneme	Indonesian	English
Consonant Phonemes		
p, b, t, d, k, g, f, s, z, ʃ , ʒ, h, ʧ, ʤ , m, n, ŋ, l, r, j, w	✓	✓
v, θ, ð	×	✓
Vowel Phonemes		
i, ɪ, u, ʊ, ɛ, ə, e, a, ɑ, o, ɔ	✓	✓
æ, ɜ, ʌ	×	✓

Table 1: List of Indonesian and English phonemes

Indonesian	English
"Asing" (*Foreign*)	*I sing*
"Demam" (*Fever*)	*The Mom*
"Es" (*Ice*)	*As*
"Kol" (*Cabbage*)	*Call*
"Kos" (*Boarding House*)	*Cost*
"Tang" (*Pliers*)	*Tongue*

Table 2: Examples of Indonesian and English words that have the same or similar pronunciation but are of different meanings

recognition system leveraging unlabeled data and utilizing a semi-supervised approach.

2. Related Study

Research on addressing Indonesian-English code-switching in speech recognition systems is indeed limited. One study by Hartanto (2019) focused on this topic. However, it utilized statistical methods, specifically Hidden Markov Models and Gaussian Mixture Models, instead of a deep learning approach. It is noteworthy that this method solely relied on labeled data and did not incorporate unlabeled data.

The Wav2Vec model, as presented in Schneider et al. (2019), utilizes unlabeled data for speech recognition through a self-supervised approach. In the pre-training phase, it learns to predict one part of unlabeled audio from another, capturing crucial audio features. Utilizing Convolutional Neural Networks (CNN) for feature extraction and recurrent layers or transformers for contextualization, the model transforms audio into contextual representations. Fine-tuning aligns these representations with corresponding text, making Wav2Vec suitable for converting audio to text. It is essential to note that Wav2Vec is purpose-built for speech recognition tasks.

3. Machine Speech Chain

3.1. Basic Machine Speech Chain

The Machine Speech Chain, developed by Tjandra et al. (2020), is a semi-supervised method connecting speech recognition and speech synthesis models through deep learning. This sequence-to-sequence model enables training with both labeled and unlabeled data.

In its learning process, three distinct stages are involved:

1. **Paired speech-text training for ASR and TTS**: Utilizing labeled data with pairs of speech-text, both ASR and TTS models are independently trained by minimizing the loss between predicted label sequences and ground truth sequences.

2. **Unpaired speech data only (ASR → TTS)**: With unlabeled speech features, ASR transcribes unlabeled speech input, and TTS reconstructs the original speech signal based on the text generated by ASR. TTS training involves minimizing the loss between the synthesized speech signal and the ground truth speech signal.

3. **Unpaired text data only (TTS → ASR)**: Given only text input, TTS generates speech signals, while ASR reconstructs the original transcription text based on the speech generated by TTS. Training for ASR is done by minimizing the loss between the transcription generated by ASR and the ground truth transcription.

The training process is carried out in a sequential order from the supervised stage to the unsupervised one. It begins with the supervised stage utilizing the paired speech-text data. Subsequently, the resulting ASR and TTS models from the supervised stage are trained further in the unsupervised stage utilizing the unpaired speech and the unpaired text data. The aforementioned stage 2 and stage 3 are done repeatedly after one another until a specified number of training.

In the standard machine speech chain, an issue arises when training data involves multiple speakers. When using unlabeled speech data for training, the synthesized speech characteristics from the speech synthesis model may differ from the ground truth speech characteristics. This discrepancy, such as generating speech with the voice of speaker B while the ground truth is from speaker A, leads to substantial loss function calculations, disrupting the unsupervised training phase.

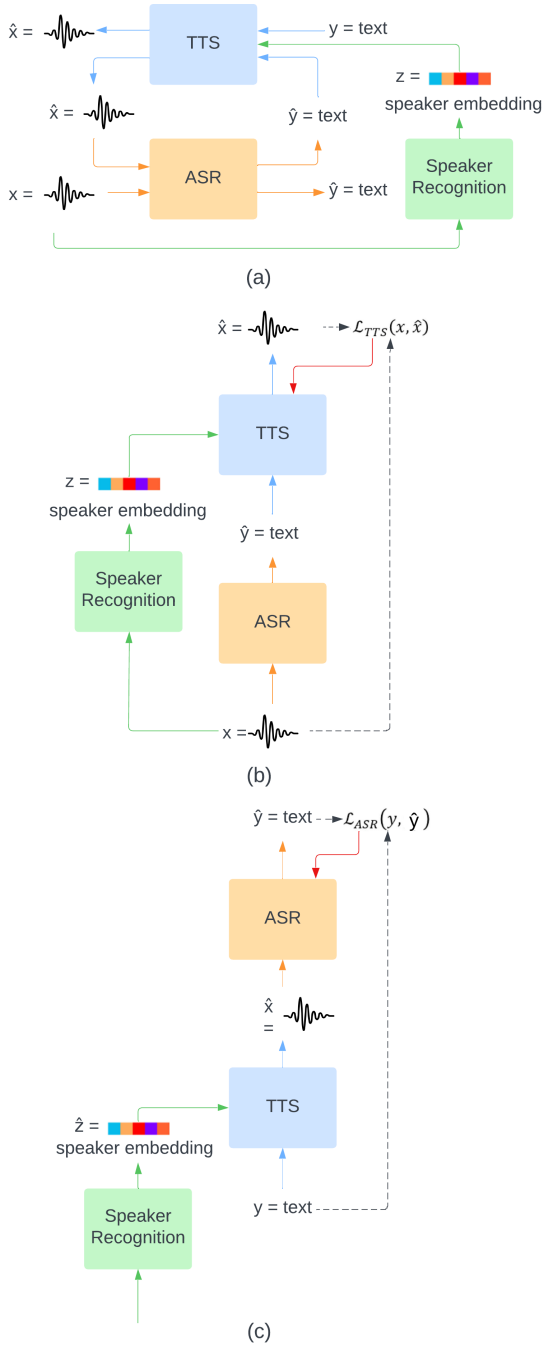

Figure 1: (a) Overview of a machine speech chain architecture with speaker recognition. Unrolled process of unsupervised training: (b) from ASR to TTS and (c) from TTS to ASR (Tjandra et al., 2020)

To tackle the challenge of differing speech characteristics between ground truth and synthesized speech during the unsupervised training phase, a speaker adaptation machine speech chain was introduced by Tjandra et al. (2020). This variation incorporates a speaker recognition model. This model takes speech as input and produces a speaker embedding representing the speaker's speech characteristics. The speaker embedding, combined with text input, is utilized by the speech synthesis model to generate speech with specific speaker characteristics. The training process of the speaker adaptation machine speech chain is akin to the basic machine speech chain, comprising a supervised stage and an unsupervised stage, as illustrated in Figure 1.

3.2. Machine Speech Chain for Code-Switching

There is also a machine speech chain architecture capable of handling code-switching (Nakayama et al., 2019). This model was developed for code-switching between English-Japanese and English-Chinese language pairs.

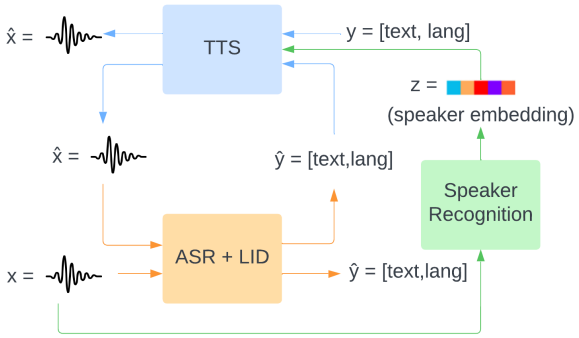

Figure 2: Overview of a multilingual machine speech chain architecture with speaker recognition (Nakayama et al., 2019)

At a high level, the architecture used is similar to the speaker adaptation machine speech chain architecture. However, there is a language identifier component within the ASR component to perform language recognition. ASR conducts multi-task learning for text transcription and language prediction using two softmax layers. Each character is provided with language information through language ID. An illustration of the machine speech chain architecture with a language identifier can be seen in Figure 2. Model training is conducted in two stages: (1) supervised training with *monolingual paired* text-speech data and (2) unsupervised training with *unpaired code-switching* data (text only or speech only).

4. Experimental Setup

The workflow begins with data acquisition to collect the dataset used for model training. Two monolingual datasets were used: the English LJSpeech dataset (Ito and Johnson, 2017) which is 24 hours long and the 40 hours long monolingual Indonesian dataset (Sakti et al., 2008a). 3399 utterances of natural code-switching Indonesian-English from Har-

tanto (2019) were also used. On top of that, 3186 utterances of code-switching English-Indonesian were generated using GoogleTTS by selecting 3186 Indonesian text from Sakti et al. (2008b) and translating some of the words to English. The resulting code-switching Indonesian-English text is then fed to GoogleTTS to generate the code-switching speech.

It is crucial to highlight that, unlike the other three corpora, the natural code-switching speech from Hartanto (2019) exhibits distinct speech characteristics. The speeches are spontaneous, with speakers not reading a transcript but rather spontaneously uttering words. This leads to the presence of verbal fillers, labeled as '<filler>' in the transcript. An example featuring fillers in a speech is illustrated in Table 3. Despite being spontaneous, the sentences maintain a formal tone. Additionally, the speeches contain background noise beyond the speaker's voice.

Transcript without language ID
merupakan wearable device <filler>
Transcript with language ID
mID eID rID uID pID aID kID aID nID <spc> wEN eEN aEN rEN aEN bEN lEN eEN <spc> dEN eEN vEN iEN cEN eEN <spc> <filler>

Table 3: Example of the natural code-switching corpora

Every dataset consists of speech data and its corresponding transcriptions. The transcriptions are complemented with the language ID of the corresponding word embedded in every character. The character of an Indonesian word would be followed by the language identifier 'ID' while the English one would be followed by 'EN' as shown in Table 3. Each of monolingual (English and Indonesian combined), synthesized code-switching, and natural code-switching are divided into three sets: the training set, the validation set, and the test set, resulting in a total of 3 training sets, 3 validation sets, and 3 test sets.

All speech utterances are of single-channel and undergo a downsampling to a sample rate of 16kHz. 80-dimensional mel spectrogram features are extracted from the downsampled speech utterances.

The MultiSpeech (Chen et al., 2020), Speech-Transformer (Dong et al., 2018), and Deep Speaker (Li et al., 2017) are used as the architecture of the ASR, TTS, and speaker recognition models respectively. The speaker recognition model was trained on all datasets to generate the speaker embedding for every speech utterance. The resulting speaker embeddings are to be used by the TTS for training. During the supervised training stage, both the ASR and TTS models were trained using the **monolingual** dataset (LJSpeech and the Indonesian dataset). Subsequentially, there were two scenarios run during the unsupervised training stage: (1) one where both the ASR and TTS models are trained on the **synthesized code-switching** dataset and (2) one where both models are trained on the **natural code-switching** dataset. An evaluation is carried out to assess the performance of the ASR model.

5. Experiment Result

In Table 4 is the Character Error Rate (CER) evaluation of all developed ASR models on English, Indonesian, and code-switching Indonesian-English test set. The table compares the baseline ASR model that was only trained in a supervised manner using only labeled monolingual (English and Indonesian) data with an ASR model that is trained further using a machine speech chain mechanism.

Training Data	**En**	**Id**	**Syn CS**	**Nat CS**
Supervised training				
En+Id (paired)	2.43%	4.10%	37.57%	91.76%
Machine Speech Chain				
+EnId (synthesized CS) (unpaired)	2.73%	4.46%	18.56%	-
+EnId (natural CS) (unpaired)	2.729%	4.361%	-	82.62%

Table 4: CER of proposed machine speech chain

The three ASR models developed show great performance in recognizing monolingual English and Indonesian speech. The baseline model, which was trained on monolingual English and Indonesian data, obtained a CER of 2.430% for monolingual English and 4.103% for monolingual Indonesian while the machine speech chain obtained a score of around 2.7% for English and 4.4% for Indonesian. The slight performance decrease in recognizing monolingual speech by the machine speech chain model happened because the model generalized to the code-switching speech.

When it comes to recognizing code-switching Indonesian-English speech, the baseline model showcased a poor performance with a CER score of 37.571% for synthesized code-switching speech and 91.76% for natural code-switching speech. However, an improvement is obtained when the

model is further trained with the machine speech chain mechanism on unlabeled code-switching speech with a CER score of 18.56% for the synthesized speech and 82.62% for the natural code-switching speech. The poor performance in recognizing natural code-switching speech was due to the noisy nature of the natural code-switching speech, which is different from the other three clean corpora. The machine speech chain ASR model trained on synthesized code-switching was not tested on natural code-switching data and vice versa since the two corpora have differing speech characteristics.

An example of the output made by the machine speech chain ASR is shown in Table 5. On the left side is an output generated by the machine speech chain ASR model trained on the synthetic code-switching data while on the right side is one generated by the ASR model trained on the natural synthetic code-switching data. The output examples say "verbal and economy to the wife" and "is wearable device" from left to right. As can be seen, the ASR model trained on the synthesized code-switching data generates a '<filler>' label.

Synthetic Code-Switching Data	**Natural Code-Switching Data**
verbal dan *economy* terhadap istrinya	merupakan *wearable device* <filler>

Table 5: ASR model output example on synthesis speech vs natural speech

6. Conclusion

In this study, ASR models were developed to handle code-switching Indonesian-English speech utilizing the semi-supervised machine speech chain method and leveraging unlabeled code-switching data. The method was able to improve the ASR performance in recognizing code-switching Indonesian-English speech by utilizing unlabeled data. However, the ASR model still shows a poor performance in recognizing natural code-switching speech because of its noisy nature. Future studies can be conducted by incorporating noise to the clean corpora (Ito and Johnson, 2017; Sakti et al., 2008a,b) to simulate noisy conditions before applying machine speech chain mechanism to the natural speech corpora from Hartanto (2019). Further study can also utilize clean and non-spontaneous speech corpora which are noise-free and clean from verbal filler.

7. Acknowledgements

Part of this work is supported by JSPS KAKENHI Grant Numbers JP21H05054 and JP23K21681, as well as JST Sakura Science Program.

8. Bibliographical References

Baso Andi-Pallawa and Andi Fiptar Abdi Alam. 2013. A comparative analysis between english and indonesian phonological systems. *International Journal of English Language Education*, 1.

Mingjian Chen, Xu Tan, Yi Ren, Jin Xu, Hao Sun, Sheng Zhao, Tao Qin, and Tie-Yan Liu. 2020. Multispeech: Multi-speaker text to speech with transformer.

Linhao Dong, Shuang Xu, and Bo Xu. 2018. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888.

Roland Hartanto. 2019. Penanganan alih kode indonesia-inggris pada sistem pengenal ucapan bahasa indonesia.

Joseph Keshet and Samy Bengio. 2009. *Automatic Speech and Speaker Recognition: Large Margin and Kernel Methods*. J. Wiley Sons.

Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan, and Zhenyao Zhu. 2017. Deep speaker: an end-to-end neural speaker embedding system.

Margana. 2013. Alih kode dalam proses pembelajaran bahasa inggris di sma. *Litera*, 12:39–52.

Sahoko Nakayama, Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2019. Zero-shot code-switching asr and tts with multilingual machine speech chain. pages 964–971. Institute of Electrical and Electronics Engineers Inc.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. Wav2vec: Unsupervised pre-training for speech recognition.

Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2020. Machine speech chain. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 28:976–989.

9. Language Resource References

Keith Ito and Linda Johnson. 2017. *The LJ Speech Dataset*.

Sakriani Sakti and Eka Kelana and Hammam Riza and Shinsuke Sakai and Konstantin Markov and Satoshi Nakamura. 2008a. *Development of Indonesian Large Vocabulary Continuous Speech Recognition System within A-STAR Project*.

Sakriani Sakti and Ranniery Maia and Shinsuke Sakai and Satoshi Nakamura. 2008b. *Development of HMM-based Indonesian Speech Synthesis*.