EasyChair Preprint

№ 15151

Machine Speech Chain with Emotion Recognition

Akeyla Pradia Naufal, Dessi Puji Lestari, Ayu Purwarianti, Kurniawati Azizah, Dipta Tanaya and Sakriani Sakti

September 28, 2024

Machine Speech Chain with Emotion Recognition

Akeyla Pradia Naufal
School of Electrical Engineering and Informatics
Bandung Institute of Technology
Bandung, Indonesia
13519178@std.stei.itb.ac.id

Dessi Puji Lestari
School of Electrical Engineering and Informatics
Bandung Institute of Technology
Bandung, Indonesia
dessipuji@std.stei.itb.ac.id

Ayu Purwarianti
School of Electrical Engineering and Informatics
Bandung Institute of Technology
Bandung, Indonesia
ayu@std.stei.itb.ac.id

Kurniawati Azizah
Faculty of Computer Science
University of Indonesia
Depok, Indonesia
kurniawati.azizah@cs.ui.ac.id

Dipta Tanaya
Faculty of Computer Science
University of Indonesia
Depok, Indonesia
diptatanaya@cs.ui.ac.id

Sakriani Sakti
Graduate School of Information Science
Nara Institute of Science and Technology
Nara, Japan
ssakti@is.naist.ac.jp

***Abstract*— Developing natural speech recognition and speech synthesis systems requires speech data that authentically represents real emotions. However, this type of data is often challenging to obtain. Machine speech chain offers a solution to this challenge by using unpaired data to continue training models initially trained with paired data. Given the relative abundance of unpaired data compared to paired data, machine speech chain can be instrumental in recognizing emotions in speech where training data is limited. This study investigates the application of machine speech chain in speech emotion recognition and speech recognition of emotional speech. Our findings indicate that a model trained with 50% paired neutral emotion speech data and 22% paired non-neutral emotional speech data shows a reduction in Character Error Rate (CER) from 37.55% to 34.52% when further trained with unpaired neutral emotion speech data. The CER further decreases to 33.75% when additionally trained with combined unpaired speech data. The accuracy of recognizing non-neutral emotions ranged from 2.18% to 53.51%, though the F1 score fluctuated, increasing by up to 20.6% and decreasing by up to 23.4%. These results suggest that the model demonstrates a bias towards the majority class, as reflected by the values of the two metrics.**

Keywords—speech recognition; speech emotion recognition; machine speech chain; unpaired data

I. INTRODUCTION

Speech is a natural communication method for humans. Automatic Speech Recognition (ASR) and Text-to-Speech Synthesis (TTS) are technologies that can facilitate human-machine interaction. Research in these areas aims to enhance the naturalness of human-computer interaction.

Natural interpersonal communication incorporates paralinguistic information, such as emotional content in speech. Thus, automatic speech emotion recognition (SER) by machines could significantly enhance the naturalness of human-machine communication. However, a major obstacle in developing SER systems is the method of data collection.

Speech emotion datasets can be classified into three types based on their collection methods: acted, elicited, or natural [1]. Acted emotion datasets consist of utterances acted by actors. While these datasets are relatively easy to obtain, they often fail to accurately represent subtle and genuine emotions. Elicited emotion datasets involve utterances spoken by individuals in simulated environments. These datasets provide a more natural representation of emotions than the acted emotion dataset. However, they are more challenging to set up. The best dataset to represent emotion is the natural emotion dataset. The utterances are collected from spontaneous conversations, such as those found in talk shows or radio programs. However, these datasets are costly to acquire due to ethical and privacy concerns.

Training ASR, TTS, and SER systems requires paired data, such as speech and transcription for ASR and TTS, and speech and emotion annotations for SER. This supervised training approach necessitates professional emotion annotations. These annotations can substantially increase the cost, especially when dealing with large datasets.

The machine speech chain [2] offers a solution as a closed-loop training system integrating ASR and TTS. It employs both paired and unpaired data to train the models. Given the greater availability of unpaired data compared to paired data, the machine speech chain can address the issue of limited data availability for training ASR and SER systems.

II. RELATED WORKS

Speech emotion recognition system for Indonesian, particularly those utilizing natural emotion datasets, have beendeveloped in studies such as [3] and [4]. In [3], an Indonesian emotional speech corpus was created using recordings from national television talk show. The corresponding emotion recognition model, utilizing Support Vector Machine (SVM) with combination acoustic and lexical features, achieved an average F-measure of 71.3%.

In [4], conversational Indonesian speech corpus was constructed from various Indonesian podcasts on YouTube. This corpus includes annotations for six emotions: happiness, anger, sadness, surprise, disgust, and fear. Employing Long Short-Term Memory (LSTM) algorithms, the model achieved an F-measure of 58.17% across all six emotions and 75.52% for the first four emotions.

Several studies have explored the application of machine speech chains in different scenarios. Novitasari et al. [4] utilized paired speech data from the Indonesian language to train models for Indonesian local languages, including Javanese, Sundanese, Balinese, and Bataknese, using exclusively unpaired data. Nakayama et al. [6] employed paired English and Japanese speech data to develop models for Japanese-English code-switching speech, relying solely on unpaired data. Similarly, Tazakka et al. [7] used this approach for Indonesian-English code-switching speech. Additionally, Yue et al. [8] demonstrated the use of machine speech chain for domain adaptation, specifically transitioning from audiobook data to presentation contexts.

Given the diverse applications of the machine speech chain, its potential for use in speech emotion recognition becomes a compelling question. Traditional supervised speech emotion recognition tasks require extensive emotional annotation, making a system that reduces or eliminates this need highly desirable. Since the machine speech chain performs well with unpaired data, this paper investigates its application in both speech emotion recognition and speech recognition in emotional speech.

III. Machine Speech Chain

Machine speech chain integrates ASR and TTS training within a closed-loop system. The architecture employed for the TTS and the ASR models are, respectively, MultiSpeech [9] and Speech-Transformer [10]. The overall architecture of the machine speech chain is illustrated in Fig. 1. The process consists of three distinct training phases:

1. Supervised training of ASR and TTS
   In this phase, paired speech and text data are utilized to train ASR and TTS models independently.
2. Semi-supervised training of TTS with unpaired speech data
   This phase involves unpaired speech data as input for the trained ASR model. The text output predicted by ASR is then used as input for the trained TTS model. The speech synthesized by TTS is compared with the original unpaired speech to refine the TTS model.
3. Semi-supervised training of ASR with unpaired text data
   During this phase, unpaired text is used as input for the trained TTS model. The synthesized speech produced by TTS serves as input for the trained ASR model. The text predicted by ASR from the synthesized speech is compared with the original text to enhance the ASR model.

The second and third phases are collectively referred to as the semi-supervised training phase or speech-chain phase, as these phases are alternated every epoch. In contrast, the supervised training phases are conducted independently at the beginning of the training process.

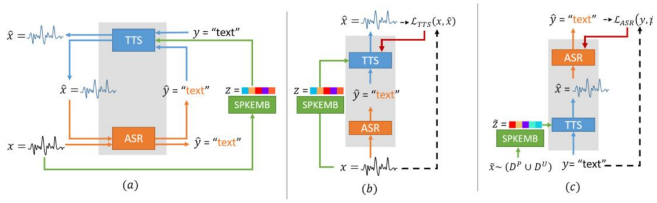

Fig. 1. (a) Overall machine speech chain architecture with speaker-adaption, (b) semi-supervised training of TTS, and (c) semi-supervised training of ASR. x and x̂ denote Mel-spectrogram while y and ŷ denote text. [2]

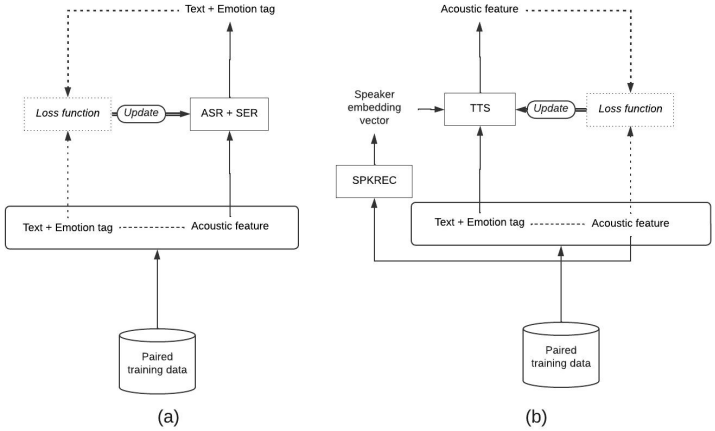

Fig. 2. Modified machine speech chain architecture with emotion recognition in supervised (a) ASR and (b) TTS training phase.

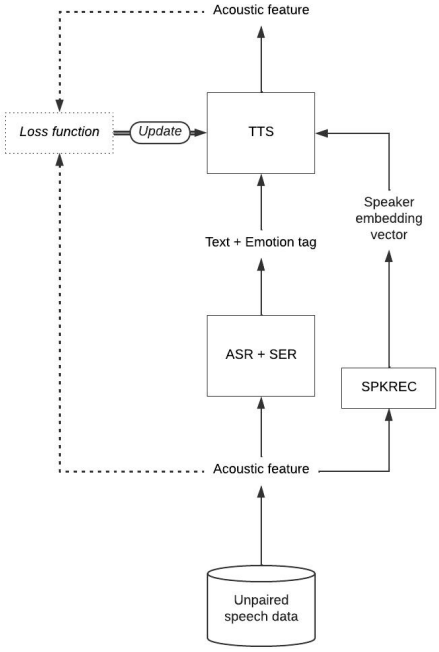

Fig. 3. Modified machine speech chain architecture with emotion recognition in semi-supervised ASR to TTS training phase.

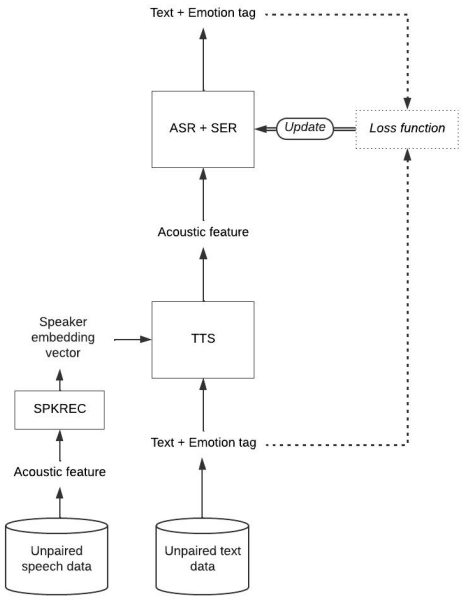

Fig. 4. Modified machine speech chain architecture with emotion recognition in semi-supervised TTS to ASR training phase.

When working with multi-speaker speech datasets, the TTS model must be capable of recognizing and synthesizing various speaker styles. To address this requirement, a speaker-embedding model known as DeepSpeaker [11] is utilized for multi-speaker speech training. DeepSpeaker can identify the speaker of a given speech and generate an embedding vector that encapsulates the speaker's characteristics.

The original machine speech chain architecture lacked modules for emotion recognition. To address this, we modified the architecture by incorporating speech emotion recognition. This modification involves adding an emotion tag to the text and adjusting the ASR module to output both the text and its corresponding emotion tag. The modified architecture is depicted in Fig. 2., Fig. 3, and Fig. 4.

IV. EXPERIMENTS

A. Datasets

For these experiments, two distinct speech corpora are utilized to serve different purposes. The first corpus is IndSpeech News LVCSR [12], referred to as IDSP in this paper. This corpus comprises 44,000 utterances from 400 speakers, with each utterance characterized by neutral emotion, similar to news reading. The primary objective of using this corpus is to establish a base for speech recognition in the ASR model, owing to its clarity in pronunciation.

The second corpus is the Indonesian Conversational Emotional Corpus [4], abbreviated as IDCEC. This corpus includes 10,413 utterances from 51 speakers, extracted from Indonesian-speaking YouTube podcasts. The utterances are notable for their spontaneity, casual nature, and the presence of mixed emotions. As outlined in the introduction, this corpus is categorized as a natural emotion dataset. It includes six emotions: happiness, anger, sadness, surprise, disgust, and fear. However, for this study, only the first four emotions are utilized, as the latter two emotions occur infrequently, as detailed in TABLE I.

TABLE I. NUMBERS OF SPEECHES FOR EACH EMOTION

Label	Number of Utterance
Happiness	3975
Anger	1451
Sadness	3418
Surprise	1606
Fear	345
Disgust	27

B. Experiment Data Split

The experiments utilize both speech corpora, varying the number of utterances in both supervised and semi-supervised training phases. TABLE II. provides a detailed overview of the different experimental scenarios. There are three types of training: supervised, speech chain with both datasets, and speech chain with the IDSP dataset only. For each of these, three different percentages of IDSP data are used for the supervised training: 10%, 30%, and 50%. This yields nine different experiments. Additionally, three baseline scenarios are included, which respectively utilize all the data from IDCEC, IDSP, and both datasets in supervised training.

TABLE II. NUMBER OF DATA USED IN ALL EXPERIMENT SCENARIOS

Supervised		Semi-Supervised ASR		Semi-Supervised TTS		% IDSP Super-vised
IDSP	*IDCEC*	*IDSP*	*IDCEC*	*IDSP*	*IDCEC*	
Supervised only						
5244	*2124*	*0*	*0*	*0*	*0*	*10*
13212	*2124*	*0*	*0*	*0*	*0*	*30*
21156	*2124*	*0*	*0*	*0*	*0*	*50*
Speech Chain with both datasets						
5244	*2124*	*17640*	*3621*	*17400*	*3692*	*10*
13212	*2124*	*13664*	*3621*	*13432*	*3692*	*30*
21156	*2124*	*9680*	*3621*	*9460*	*3692*	*50*
Speech Chain with IDSP only						
5244	*2124*	*17640*	*0*	*17400*	*0*	*10*
13212	*2124*	*13664*	*0*	*13432*	*0*	*30*
21156	*2124*	*9680*	*0*	*9460*	*0*	*50*
Baseline (Supervised with all data available)						
41000	*0*	*0*	*0*	*0*	*0*	*100*
0	*9437*	*0*	*0*	*0*	*0*	*0*
41000	*9437*	*0*	*0*	*0*	*0*	*100*

For all non-baseline experimental scenarios, 2124 IDCEC utterances are used for supervised training, representing approximately 22% of all the IDCEC utterances. Validation data consists of approximately 916 IDSP utterances and 478 IDCEC utterances, while test data includes 904 IDSP utterances and 477 IDCEC utterances. The remaining data are allocated for the semi-supervised training phase, either as

unpaired speech or unpaired text data. To avoid overlap between the supervised and semi-supervised training phases, some utterances, primarily from the IDSP corpus, are not utilized. Additionally, to ensure a balanced representation of different emotions, a specific data distribution is used, rather than randomly selected, as detailed in TABLE II.

C. *Input formats*

Standardizing the input format is essential when working with two different datasets. For speech, downsampling is applied to both datasets to ensure a uniform sample rate of 16kHz. Acoustic features are then extracted in the form of an 80-dimensional Mel spectrogram, which serves as the input for the ASR model.

Transcriptions of the speech are segmented into character tokens that correspond to Indonesian pronunciation. Given the spontaneous nature of the IDCEC speech, non-word tags are included to represent sounds such as <batuk> (coughing), <ketawa> (laughing), <noise>, and <unk> (unknown). Foreign words are tokenized according to the speaker's pronunciation. Additionally, an emotion tag is prefixed to the transcription in the modified machine speech chain architecture. These emotion tags range from <0> to <4>, representing neutral, happiness, anger, sadness, and surprise, respectively. As an example, refer to TABLE III. Notice that the English word "speechless" is tokenized as "s p i c l e s" to reflect its Indonesian pronunciation.

TABLE III. AN EXAMPLE OF PROCESSING TEXT INPUT

Original	aku speechless dengarnya (with sad emotion)
Processed	<3> a k u <spc> s p i c l e s <spc> d e ng a r ny a

V. EXPERIMENT RESULT

A. *ASR*

The performance of the ASR system is evaluated using the Character Error Rate (CER). The results of the experiments are presented in TABLE IV.

TABLE IV. ASR EXPERIMENT RESULTS

% IDSP Supervised	**CER (%)**		
	Super-vised	**Speech Chain (both datasets)**	**Speech Chain (IDSP only)**
10	***76.95***	***127.06***	***71.86^***
30	***47.43***	***115.08***	***41.26^***
50	***37.55***	***33.75^***	***34.52^***
Baselines (supervised)			
All IDSP	***64.03***		
All IDCEC	***35.30***		
Both Dataset	***18.28***		

As anticipated from the results, increasing the amount of paired data used during the supervised training phase results in lower error rates. This trend is consistent across all training types: supervised-only, speech chain using both datasets and speech chain using IDSP data only. Supervised-only training exhibits higher error rates compared to scenarios where paired data from both datasets are utilized, or even when only paired IDCEC data are used. Notably, models trained with all paired data from the IDSP dataset exhibit higher error rates compared to the supervised-only experiments that utilize a combination of IDCEC data and 30% or 50% of IDSP data. This discrepancy may be attributed to the spontaneous nature of IDCEC speech, which contrasts with the clearer pronunciation of IDSP speech.

The speech chain reduces the error rates, but this is observed mostly in the experiments where the speech chainuses only unpaired IDSP speech. These experiments show a slight decrease in CER. Conversely, most experiments that involve the speech chain with both IDSP and IDCEC unpaired data, exhibit significantly higher error rates compared to the corresponding supervised-only experiment. The exception is the supervised-only experiment using 50% of IDSP data, which achieves slightly lower error rates compared to when the model is further trained using speech chain with unpaired data of both datasets. The comparison of CER across all the experiments is illustrated in Fig. 5.

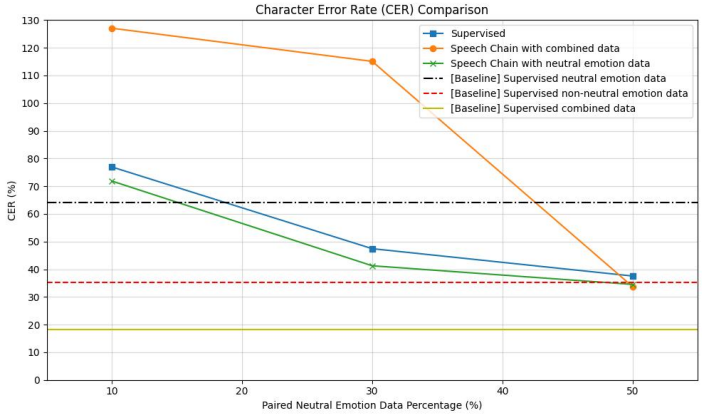

Fig. 5. Character error rate comparison across all experiments

B. *SER*

TABLE V. ASR EXPERIMENT RESULTS

% IDSP Super-vised	**Non-neutral Emotion Accuracy (%)**			**F1 Score**		
	Super-vised	**Speech Chain (both datasets)**	**Speech Chain (IDSP only)**	**Super-vised**	**Speech Chain (both datasets)**	**Speech Chain (IDSP only)**
10	***32.08***	***41.30^***	***39.62^***	***0.142***	***0.179^***	***0.128***
30	***27.04***	***41.51^***	***41.09^***	***0.162***	***0.124***	***0.165^***
50	***38.57***	***39.41^***	***41.30^***	***0.159***	***0.125***	***0.128***
Baselines (supervised)						
All IDSP	***0***			***0***		
All IDCEC	***31.45***			***0.248***		
Both Dataset	***33.96***			***0.181***		

The metrics used to evaluate SER are emotion accuracy and F1 score. Notably, only non-neutral emotion accuracy is

considered, as all neutral emotions are predicted with 100% accuracy. The results of the experiments are presented in TABLE V.

In contrast to the CER which decreases with an increase in paired data, neither the emotion accuracy nor the F1 score demonstrates a clear correlation with the amount of paired data used. A notable trend is that non-neutral emotion accuracy improves after speech-chain training, whether using combined data or only IDSP data. Among all experiments and baselines, the non-neutral emotion accuracy ranges from 27% to 42%. The supervised-only baseline trained exclusively on all IDCEC data is excluded from this analysis, as it is limited to neutral emotion data and is therefore unable to recognize other emotions. The comparison of non-neutral emotion F1 score and accuracy across all experiments, respectively, are illustrated in Fig. 6. and Fig. 7.

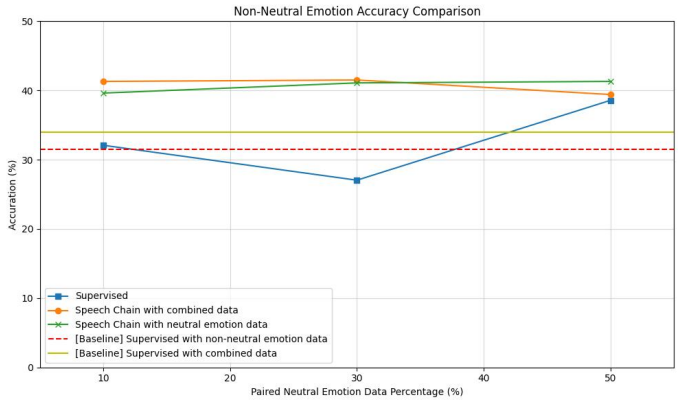

Fig. 6. Non-neutral emotion accuracy comparison across all experiments

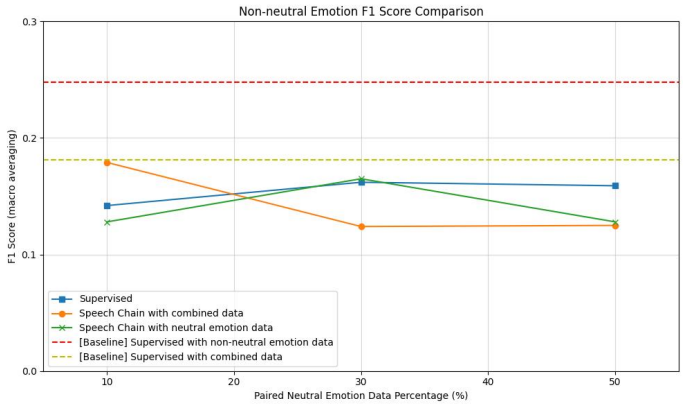

Fig. 7. Non-neutral emotion accuracy comparison across all experiments

However, the low F1 scores (generally below 0.2) indicate an imbalance in emotion prediction. There is no discernible trend in the F1 score relative to the amount of paired data used or the type of unpaired data. This is further illustrated by the emotion accuracy for each emotion, as shown in TABLE VI. The low accuracy for all non-neutral emotions, apart from happy, suggests that the SER model's performance is sub-optimal and may be biased towards the majority class. Another possible explanation for the low accuracy in recognizing anger and surprise could be that these emotions are not strongly expressed in the speech, making them more difficult to detect.

TABLE VI. EMOTION ACCURATION (%) WITH 50% IDSP SUPERVISED

Type	Neutral	Happiness	Anger	Sadness	Surprise
Supervised	100	65.78	0	24.16	0
Speech Chain (Both datasets)	100	73.78	0	14.09	1.54
Speech Chain (IDSP only)	100	79.11	0	12.08	1.54

Some examples of predictions made in the experiment, which was trained with 50% of the neutral emotion speech data in the supervised phase and further trained using the speech chain with combined speech data, are presented in Table VII. Since the model does not incorporate language model, the prediction could include generation of nonsensical words such as "adari" and "mikasih" or repetition of characters, as seen in "dalaaan". Furthermore, it could be seen that the second example predict the speech as having happy emotion whereas the real emotion is anger.

TABLE VII. SOME EXAMPLES OF SPEECH PREDICTION VS ORIGINAL TEXT

Original	Predicted
<1> jadi dari mulai kita nikah semua kita tuh komunikasi dijaga banget	<1> jadi kayak adari mulai kita minta semuanya kita tuh kalau mikasih dan banget
<2> cuman ya gue omelin waktu itu udahlah gitu maksudnya enggak akan ada endingnya udah biar waktu yang jawab udah diam aja gitu	<1> makanya gue ngomelin aku itu dalaaaan itu maksudnya enggak anak ada yang dia udah dara kayaknya waktu yang enggak kayak
<3> karena emang sekrisis itu keadaannya	<3> karena maksudnya kayak sih itu kadangnya

VI. CONCLUSION

This paper demonstrates that machine speech chain can reduce character error rates in emotional speech recognition by further training a previously supervised ASR model with unpaired neutral emotion data. However, training with combined unpaired data may exacerbate the error rates. In the context of speech emotion recognition, the machine speech chain improves emotion accuracy by incorporating emotion tags into the transcription. Despite this, the low F1 scores across various emotions suggest that the current dataset may be insufficient in terms of numbers. The observed improvements in the F1 score in subsequent experiments highlight the potential of the speech chain method for emotion recognition. Consequently, future work will aim to integrate a larger amount of unlabeled emotional data to enhance the performance of the emotion recognition model using the machine speech chain approach.

ACKNOWLEDGMENT

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

REFERENCES

[1] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, Jan. 2020. doi:10.1016/j.specom.2019.12.001

[2] A. Tjandra, S. Sakti, and S. Nakamura, "Machine speech chain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 976–989, 2020. doi:10.1109/taslp.2020.2977776

[3] P. Kurniawati, D. P. Lestari and M. L. Khodra, "Speech emotion recognition from Indonesian spoken language using acoustic and lexical features," 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA), Seoul, Korea (South), 2017, pp. 1-7, doi: 10.1109/ICSDA.2017.8384467.

[4] A.N.I. Adma and D.P. Lestari, "Conversational speech emotion recognition from Indonesian spoken language using recurrent neural network-based model", *2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA),* pp. 1-6, 2021. doi: 10.1109/ICAICTA53211.2021.96

[5] S. Novitasari, A. Tjandra, S. Sakti, and S. Nakamura, "Cross-lingual machine speech chain for Javanese, Sundanese, Balinese, and Bataks speech recognition and synthesis," ACL Anthology, https://aclanthology.org/2020.sltu-1.18/ (accessed Jul. 4, 2024).

[6] S. Nakayama, A. Tjandra, S. Sakti, and S. Nakamura, "Zero-shot code-switching ASR and TTS with multilingual machine speech chain," *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Dec. 2019. doi:10.1109/asru46091.2019.9003926

[7] R.V.M. Tazakka, D.P. Lestari, A. Purwarianti, D. Tanaya, K. Azizah, and S. Sakti, "Indonesian-English code-switching speech recognition using the machine speech chain based semi-supervised learning," *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages@ LREC-COLING, 2024*, pp. 143-148.

[8] F. Yue, Y. Deng, L. He, and T. Ko, "Exploring machine speech chain for domain adaptation and few-shot speaker adaptation," 2021, *arXiv preprint arXiv:2104.03815*

[9] M. Chen, X. Tan, Y. Ren, J. Xu, H. Sun, S. Zhao, T. Qin, and T.-Y. Liu, "Multispeech: Multi-speaker text to speech with transformer," 2020, *arXiv preprint arXiv:2006.04664*

[10] L. Dong, S. Xu, and B. Xu, "Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition," *2018 IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP), 2018,* pp. 5884-5888

[11] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, "Deep speaker: an end-to-end neural speaker embedding system," 2017, *arXiv preprint arXiv:1705.02304*

[12] S. Sakti, E. Kelana, H. Riza, S. Sakai, K. Markov, and S. Nakamura, "Development of Indonesian large vocabulary continuous speech recognition system within A-STAR project", 2008, *Proceedings of the Workshop on Technologies and Corpora for Asia-Pacific Speech Translation (TCAST)*