

Article

Effects of Data Augmentations on Speech Emotion Recognition

Bagus Tris Atmaja^{1,2,*}  and Akira Sasou¹ ¹ National Institute of Advanced Industrial Science and Technology, Tsukuba 305-8560, Japan² Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia (on leave)

* Correspondence: b-atmaja@aist.go.jp

Abstract: Data augmentation techniques recently gained more adoption in speech processing, including speech emotion recognition. Although more data tends to be more effective, there may be a trade-off in which more data will not provide a better model. This paper reports experiments on investigating the effects of data augmentation in speech emotion recognition. The investigation aims at finding the most useful type of data augmentation and the number of data augmentations for speech emotion recognition. The experiments are conducted on the Japanese Twitter-based emotional speech corpus. The results show that for speaker-independent data, two data augmentations with glottal source extraction and silence removal exhibited the best performance among others, even with more data augmentation techniques. For the text-independent data (including speaker and text-independent), more data augmentations tend to improve speech emotion recognition performances. The results highlight the trade-off between the number of data augmentation and the performance of speech emotion recognition showing the necessity to choose a proper data augmentation technique for a specific application.

Keywords: speech emotion recognition, affective computing, data augmentations, wav2vec 2.0, SVM

1. Introduction

The development of speech emotion recognition (SER) has been directed in either finding the new features correlated to emotion or building classifiers suited for emotion recognition problems. The initial survey on SER addressed features and classification schemes as the most important aspects of designing SER besides the database or dataset [1]. A recent survey on SER [2] also has shown that features (including pre-processing step) and classifiers are the most studied areas among others in SER.

Instead of proposing a new method, investigating the impact of data augmentation is worth studying. For example, reference [3] found that the key finding of their research building self-supervised learning automatic speech recognition is the extremely large and diverse datasets. However, for speech emotion recognition, the availability of the dataset is not as large as speech recognition datasets. Therefore, the research on data augmentation is a good opportunity to investigate the impact of data augmentation on SER performance.

Data augmentation has been found useful in the other areas outside speech emotion recognition. Ko et al. [4] evaluated a low-cost data augmentation technique by speed perturbation and found it improved the word error rate (WER) over other methods. Casanova et al. [5] employed both transfer learning and data augmentation for improving COVID-19 detection from cough sounds. Similarly, data augmentation could improve SER performances in specific ways. However, no study has been found regarding these impacts of data augmentation on SER performance.

In view of the above assumption based on findings in the other research areas, this paper proposes a systematic investigation of the effect of data augmentation on speech emotion recognition. This paper concretely contributes mainly to the following investigations:

- the data augmentation types that contribute to the performance of speech emotion recognition;
- the number of data augmentations that optimally improves the performance of speech emotion recognition.



Citation: Lastname, F.; Lastname, F.; Lastname, F. Effects of Data Augmentations. *Preprints* 2022, 1, 0. <https://doi.org/>

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

The experiments are conducted on the Japanese Twitter-based emotional speech corpus. We examine four types of data augmentation, including glottal source extraction, silence removal, impulse response convolution, and noise addition. The number of data augmentations is the number combination of four types of data augmentation above: one augmentation, two augmentations, three augmentations, and four augmentations.

2. Previous Work

Research on speech processing has been focused on developing new methods instead of evaluating existing methods with different evaluation conditions. Nevertheless, evaluating the training conditions for such speech processing goals is a challenging task in addition to developing the new methods. In [6], the authors evaluated training data selection between neutral and emotional speech for speech recognition. The results showed that training a sufficient amount of spontaneous data was more beneficial than a small amount of emotional speech. In other words, emotional speech is harder to be recognized than neutral speech for speech recognition.

In [7], the authors investigated the splitting criteria for training speech emotion recognition on the JTES dataset, whether splitting the training data by speakers (speaker-independent), by texts (text-independent), or both (speaker+text-independent). The results showed that splitting by speakers and text was the most difficult among others. It is difficult for the SER model to infer new data with different linguistic information from training data. The current model needs to be trained on the same linguistic information to obtain sufficient performance.

In [8], the authors highlighted that the key to their SER performance improvements was data augmentation. However, no direct comparison of the same method with and without data augmentation was found. This gap is left; however, it is important to study the contribution of data augmentation on SER to direct future work.

This study aims to fill that gap. We wanted to compare the performance of emotion recognition among different data augmentation types and the number of data augmentations. As an additional analysis, we perform a benchmark on the JTES and IEMOCAP datasets, which can be used as a comparison tool for future studies.

3. Methods

3.1. Dataset

JTES

Japanese Twitter-based emotional speech (JTES) corpus ([9]) is evaluated in this research. The dataset contains 20000 utterances from 100 speakers, four emotion categories, and 50 sentences. From all samples in the dataset, we choose speaker-independent (SI), text-independent (TI), and speaker and text-independent (STI) criteria following the previous research [7,8]. The SI and TI criteria result in 16000 utterances for training and the rest 4000 utterances for the test (80/20 split). The STI criterion results in 14400 samples for training and 400 samples for the test. We excluded the speaker-dependent (SD) criterion from the previous study [7] since SI is more challenging than SD [10] and is a standard evaluation in most speech processing tasks, including SER. For training data, we performed data augmentations in addition to the original data. For the test data, we did not add any data augmentation technique. The data augmentation techniques evaluated in the training phase are described in the following section. Figure 1 shows the flow of training data selection to investigate the data augmentation on SER.

IEMOCAP

The interactive emotional dyadic motion capture (IEMOCAP) database is one of the most common datasets for speech emotion recognition tasks. The dataset focuses on understanding expressive human communication through a combination of verbal and non-verbal channels from both speech and gestures. The original corpus contains 10039 utterances/sentences; this research utilized a subset of that total utterance commonly evaluated

for the four categories of emotion [11,12]. The four emotion categories are neutral, happy, anger, and sadness. The excitement emotion category is merged with the happy emotion. From five sessions in the dataset, Sessions 1-4 are allocated for training, while Session 5 is for a test. This leave-one-session-out (LOSO) evaluation is speaker-independent since the speakers for each session are different (two speakers per session). As in JTES, the augmentations are performed for training but not for the test set.

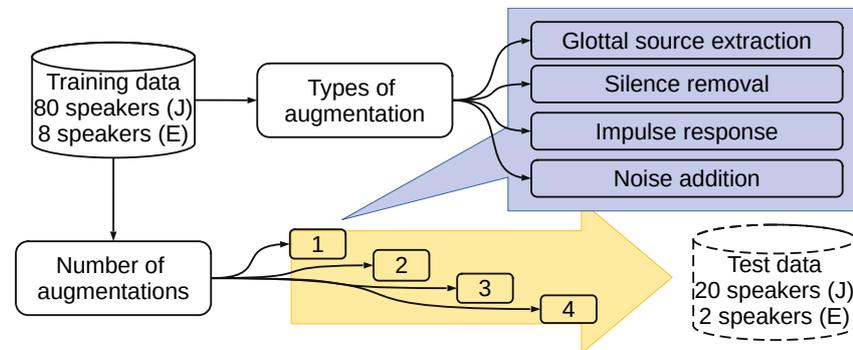


Figure 1. Flow of data selection for training speaker-independent SER in the experiments of data augmentation; J:Japanese (JTES); E: English (IEMOCAP); For text-independent, the split of training/test is based on sentences instead of speakers.

3.2. Data Augmentations

Four data augmentations are applied to the dataset. The choice of different data augmentations is to determine the type of data augmentation that contributes to the training of speech emotion recognition. Furthermore, these four data augmentations can be combined to evaluate the optimal number of data augmentations for SER in this study. The data augmentation techniques are described in the following section.

3.2.1. Glottal inverse filtering (Glottal source extraction, glt)

Several studies [13–17] have investigated the characteristics of glottal flow regarding different emotional states and have indicated that using glottal inverse filtering (GIF)-based speech analysis, emotional cues can be obtained from speech signals. In this data augmentation, we adopt a GIF method based on a constrained auto-regressive hidden Markov model (CAR-HMM) [18]. The CAR-HMM extracts the glottal flow derivative from a voiced speech by introducing some gain constraints to the AR filter. The glottal flow derivative is defined by a glottal flow that is combined with lip radiation characteristics. The gain constraints to the AR filter are introduced to avoid some characteristics of the glottal flow derivative remaining in the estimated AR filter.

The CAR-HMM, as depicted in Figure 2, is a kind of source-filter model, where the AR filter and the HMM, as depicted in (b) and (f), represent a vocal tract and a generative model of a glottal flow derivative, respectively. The states in the HMM are concatenated in a ring form so that the state transition circulates to represent the periodicity of the voiced speech. Each state has an output probability distribution (PDF) of a single Gaussian distribution defined by an expectation and a variance. Given a state transition sequence, the expectations and variances of the output PDFs can align (for example) as depicted in (d) and (e), respectively. The glottal flow derivative, as depicted in (c), is then defined by the realized values of the aligned output PDFs. Finally, the CAR-HMM is assumed to generate the voiced speech as depicted in (a) by filtering the glottal flow derivative with the AR filter.

In the CAR-HMM-based speech analysis, the AR coefficients, the glottal flow derivatives, and the corresponding HMM parameters must be estimated using a given voiced speech. The iteratively-parameter estimation algorithm of the CAR-HMM is adopted to obtain these values. The details have been described in [6]. In this augmentation, the

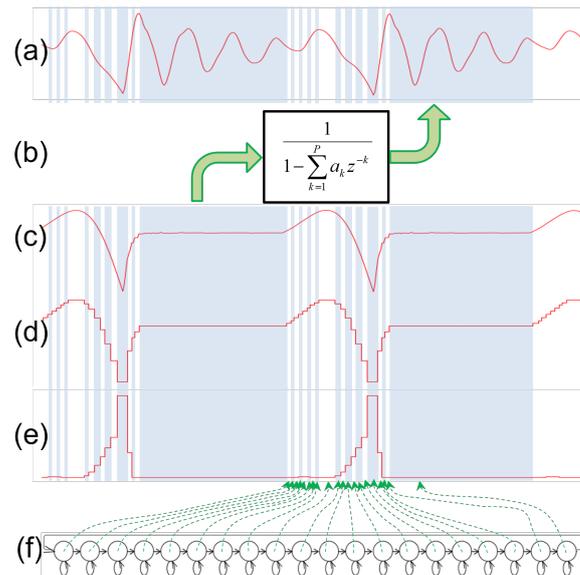


Figure 2. Source filter model based on CAR-HMM

prediction order of the AR filter and the number of HMM states are set to be 16 and 20, respectively. The AR filter's gains at both the DC and the Nyquist frequency are constrained to be 1.

3.2.2. Silence removal (speech cleaned, spc)

In this data augmentation, we remove silence durations before the beginning point and after the endpoint of each utterance (start-end silence removal).

3.2.3. Applying impulse response (ir)

In this data augmentation, we convolve original JTES audio data with a random impulse response. Two impulse response datasets are evaluated [19,20]. The amplitude of the impulse response is set to be 0.5 from the maximum amplitude of the corresponding audio signal.

3.2.4. Noise addition (noi)

In noise addition data augmentation, we mix original audio files in the JTES dataset with noises from the Environmental Sound Classification dataset with 50 classes (ESC-50) [21]. The ESC-50 consists of 2000 noises in five major categories: animal, natural, human, interior, and exterior sounds.

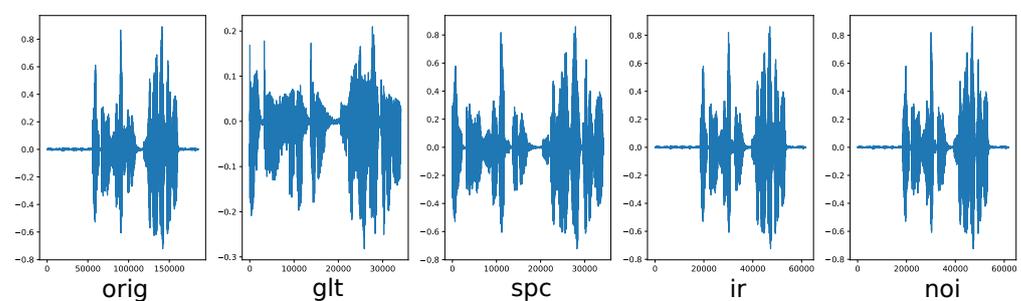


Figure 3. Plots of the original and augmented data from a sample in JTES dataset

Figure 3 shows the example of data augmentations in the JTES dataset with data augmentation techniques above. The most left is the original sample from the dataset. The rest are augmented data by glottal source extraction (*glt*), silence removal (*spc*), applying impulse response (*ir*), and adding noise (*noi*). It can be seen clearly the difference among

data. For instance, in *glt* and *spc*, the duration of the signal becomes shorter than in others, while in *ir* and *noi* the shape of the signals is changed due to impulse response convolution and noise addition.

In addition to the four types of data augmentation techniques, we also evaluated the performance of combinations of the four data augmentations above. The results of the combinations are six arrangements for two data augmentations, four arrangements for three augmentations, and one arrangement for four data augmentations. The number of training data on each data augmentation is shown in Table 1.

Table 1. Number of data (utterances) in each number of data augmentation

Number of data augmentations	Number of training data			
	JTES-SI	JTES-TI	JTES-STI	IEMOCAP
Without augmentation	16000	16000	14400	4290
With one augmentation	32000	32000	28800	8580
With two augmentations	48000	48000	43200	12870
With three augmentations	64000	64000	57600	17160
With four augmentations	80000	80000	72000	21450
Test data	2000	2000	400	1241

3.3. Acoustic features

We employed a model for dimensional speech emotion recognition [22,23] based on wav2vec 2.0 [24] to extract speech embeddings as acoustic features. This model is trained on MSP-IMPROV dataset [25]. The size of the embedding is a vector of 1024-dimensional, which is the number of hidden states in the model.

3.4. Classifier

A support vector machine (SVM) for classification (SVC) is chosen as the classifier. SVM is one the most common classifiers for SER tasks [26]. Other SER research employed MLP [10], LSTM [8,27,28], or combinations of CNN with LSTM [12,29,30]. This study chose SVM since it is the simplest classifier among them. This study focuses on investigating the effect of different data augmentations instead of proposing a new classifier architecture. The acoustic feature described in the previous subsection is standardized by removing the mean and scaling to unit variance before it is fed into SVC. The SVC is implemented through the scikit-learn library [31]. We left the configuration of SVC as default except for the kernel coefficient. The kernel coefficient is set to $1/1024$ ('auto').

3.5. Evaluation metric

We evaluated the performances of data augmentations on SER with a single metric, namely unweighted average recall (UAR). This metric is an average recall of predictions over the true labels for each emotion category. This metric is also known as balanced accuracy or unweighted accuracy (UA).

4. Experiments

We conducted two experiments on the different data augmentations and apparatus (computing systems) to generalize the performances of evaluated systems on different machines. The main aim of using two apparatus is to investigate the effect of adding random impulse responses and noises to the original data aside from checking the consistency of results for adding glottal source filtering and removing start-end silences.

4.1. Experiment 1 (Exp. #1)

In experiment 1, we evaluated all data augmentation techniques described previously on a PC with Ubuntu 18.04 OS, AMD EPYC 7702P 64-Core CPU, 504GB RAM, and two NVIDIA GeForce RTX 3090 cards (24GB/card). For impulse response addition, we used audio files

from the MIT Acoustical Reverberation Scene Statistics Survey [19]. For noise addition, we used noise files from ESC-50: Dataset for Environmental Sound Classification [21].

4.2. Experiment 2 (Exp. #2)

In experiment 2, we evaluated all data augmentation techniques similarly to Experiment 1 on an HPC (abci.ai) with CentOS Linux 7, Intel Xeon Gold 6148 CPU, 360GB RAM, and NVIDIA Tesla V100 (16GB). For impulse response addition, we used audio files from the EchoThief Impulse Response Library [20]. For noise addition, we used the same noise files from ESC-50: Dataset for Environmental Sound Classification [21]. Note although the noise dataset is the same, the results of data augmentation are different since the audiomentation package version 0.19.0 [32], which was used to mix original data with noise, chose the noise files randomly to mix in.

5. Results and Discussion

Tables 2 – 5 show the results of experiments. The results of two investigations – the type and number of augmentations – can be inferred from each table. The results with one augmentation depict the investigation of the type of data augmentation that contributes to SER performances. The rest of the results depict the investigation of the effect of the number of data augmentations. We measured those performances of the data augmentation techniques on SER with the UA/UAR metric on 0-100 % scale. These results on two experiments are deterministic, meaning that the same scores are obtained for the same data for different runs on the same apparatus.

5.1. JTES-SI

The first evaluation is the speaker-independent criterion, a standard criterion in most speech processing tasks. The result shown in Table 2 exhibits higher performances than the previous study [33]. Although in that paper the author has proposed a method for multilingual SER, comparing monolingual SER with the same JTES dataset between that study and this study shows performance improvement, even for the results without augmentation in this study. The key ingredient of our study is the utilization of deep learning-based acoustic feature extraction, which used a pre-trained model from a specific emotional speech database using wav2vec 2.0 rather than neutral speech. The previous study [33] used INTERSPEECH 2010 acoustic feature set, which is non-deep learning-based feature extraction.

For this SI evaluation, the best improvement has been achieved by the largest number of data augmentations, i.e., four data augmentations with *glt*, *spc*, *ir*, and *noi*. The improvement is about 1% from the baseline data without augmentation. The individual augmentation (with one augmentation) shows degradation (*orig + glt*), and improvements with *orig + ir* obtained the highest score for an additional type of augmentation. The trend in this SI evaluation yields a presumption that more augmentation will improve SER performance.

5.2. JTES-TI

The text-independent evaluation is a new splitting criterion for SER [7,34]. Those previous studies have shown that text-independent evaluation is more difficult and challenging than traditional speaker-independent evaluation, and this study supports that finding. Compared to SI with 97.10% UA for baseline, TI obtained 75.08% of UA. In contrast to a previous study that evaluated the same number of test set for comparing SI, TI, and STI [7], this study evaluated a different number of test set between SI and TI to take advantage of the number of samples/utterances. By evaluating 2000 samples as a test set, this study revealed a similar pattern to the previous study on the same JTES dataset.

Different portions of the training/test split were also evaluated in this study (16000/4000) compared to the previous study [7] (19600/400 and 14400/400). However, the trends look similar in both studies. Both SI and TI improved about 10% from that previous study. The

Table 2. Unweighted average recall (UAR, %) of SER with different data augmentation techniques on JTES-SI. orig = original JTES dataset, glt = glottal source extraction, spc = speech cleaned, ir = impulse response, noi = noise addition.

Data	Exp. #1	Exp. #2
Without augmentation		
orig	97.10	97.10
With one augmentation		
orig + glt	96.45	96.45
orig + spc	97.20	97.20
orig + ir	97.73	97.43
orig + noi	97.55	97.60
With two augmentations		
orig + glt + spc	96.90	96.90
orig + spc + ir	97.78	97.58
orig + ir + noi	97.85	97.83
orig + noi + glt	97.23	97.50
orig + glt + ir	97.33	97.10
orig + spc + noi	97.65	97.65
With three augmentations		
orig + glt + spc + ir	97.53	97.40
orig + spc + ir + noi	97.98	97.95
orig + ir + noi + glt	97.78	97.78
orig + noi + glt + spc	97.60	97.68
With four augmentations		
orig + glt + spc + ir + noi	98.05	97.90

Table 3. Unweighted average recall (UAR, %) of SER with different data augmentation techniques on JTES-TI. orig = original JTES dataset, glt = glottal source extraction, spc = speech cleaned, ir = impulse response, noi = noise addition.

Data	Exp. #1	Exp. #2
Without augmentation		
orig	75.08	75.05
With one augmentation		
orig + glt	76.43	76.45
orig + spc	76.23	76.23
orig + ir	75.30	74.70
orig + noi	75.38	75.20
With two augmentations		
orig + glt + spc	76.75	76.75
orig + spc + ir	75.38	74.68
orig + ir + noi	75.30	74.73
orig + noi + glt	76.05	75.98
orig + glt + ir	75.50	75.10
orig + spc + noi	75.80	75.45
With three augmentations		
orig + glt + spc + ir	75.70	74.75
orig + spc + ir + noi	75.48	74.25
orig + ir + noi + glt	75.43	75.00
orig + noi + glt + spc	75.93	75.00
With four augmentations		
orig + glt + spc + ir + noi	75.63	74.50

key factor, as explained above, is the different acoustic features used in this study with a less sophisticated classifier (SVM in this study vs. MLP in the previous study).

The best data augmentation in this TI evaluation is obtained by glottal source filtering. This approach leads to 1.35% of UA improvement. These results are opposite to the previous JTES-SI evaluation in which *glt* obtained the worst score among other types of audiomentation. In this TI evaluation, more data augmentation does not tend to increase the performance of SER. By using four data augmentations, the results are only improved by about 0.5% from the baseline. The highest UA for this TI evaluation is obtained by augmenting the original data with *glt* and *spc* (two augmentations), which yields the results of 76.75% UA.

5.3. JTE-STI

For JTES-STI evaluations, we found that cleaning original speech signals by removing silences is the most effective data augmentation technique by 2% UAR improvement. The other two data augmentations, the glottal source extraction and noise addition, slightly improved SER's performance. Augmenting SER with the impulse response in experiment 2 decreases the performance; the reverberation by room shapes degrades the recognition rate of emotional speech.

For the number of data augmentations in STI, we found that the number of data augmentations with two augmentations is the most effective data augmentation technique by 3% UAR improvement from the baseline without augmentation. Among six arrangements of two data augmentations, augmented original JTES dataset with glottal source extraction and silence removal consistently achieved the highest UAR on both experiment 1 and experiment 2. These results are also in line with the types of data augmentation

Table 4. Unweighted average recall (UAR, %) of SER with different data augmentation techniques on JTES-STI. orig = original JTES dataset, glt = glottal source extraction, spc = speech cleaned, ir = impulse response, noi = noise addition.

Data	Exp. #1	Exp. #2
Without augmentation		
orig	74.50	74.50
With one augmentation		
orig + glt	75.50	75.50
orig + spc	76.50	76.50
orig + ir	74.75	73.75
orig + noi	76.25	75.50
With two augmentations		
orig + glt + spc	77.25	77.25
orig + spc + ir	76.00	74.50
orig + ir + noi	75.50	74.50
orig + noi + glt	75.25	75.25
orig + glt + ir	75.00	73.75
orig + spc + noi	75.75	74.75
With three augmentations		
orig + glt + spc + ir	76.00	74.75
orig + spc + ir + noi	74.50	74.25
orig + ir + noi + glt	74.00	75.00
orig + noi + glt + spc	75.50	75.00
With four augmentations		
orig + glt + spc + ir + noi	74.75	74.50

techniques (with one augmentation), in which data augmentation with silence removal (speech cleaned, *spc*) topped the performance among other data augmentation types.

5.4. IEMOCAP

IEMOCAP with speaker-independent evaluation leads to the results shown in Table 5. As in the JTES evaluations, the improvement from the original data without augmentation to four data augmentations is about 1.5% of UA. In this IEMOCAP-SI evaluation, more data augmentations tend to improve SER performance as in JTES-SI evaluation.

By comparing individual data augmentations, it is not clear which type of data augmentations performs better than others. Augmenting original data with start-end silence removal and noise addition yields similar results to each other in experiment #1. These results are consistent (no one type of data augmentation is superior to the others) with previous results on JTES dataset evaluations in this study.

The highest result with four augmentations is achieved in experiment #1. The result of experiment #2 for this number of data augmentation is only lower than that of three data augmentations with *spc + ir + noi* (gap of 0.07%). By inspecting this number, we can presume that more data augmentations tend to yield performance improvement. In this study, this trend is observed for both JTES and IEMOCAP datasets with speaker-independent evaluations.

5.5. Comparison to the previous studies

As an additional analysis, we performed a literature study conducted similarly on the JTES dataset with a similar split scenario. We found that our results are competitive with the results of the literature studies for both with and without data augmentation.

Table 5. Unweighted average recall (UAR, %) of SER with different data augmentation techniques on IEMOCAP. orig = original IEMOCAP dataset, glt = glottal source extraction, spc = speech cleaned, ir = impulse response, noi = noise addition.

Data	Exp. #1	Exp. #2
Without augmentation		
orig	74.88	74.88
With one augmentation		
orig + glt	74.23	74.23
orig + spc	75.44	75.44
orig + ir	75.03	74.80
orig + noi	75.43	75.07
With two augmentations		
orig + glt + spc	75.66	75.66
orig + spc + ir	75.68	75.37
orig + ir + noi	75.15	75.46
orig + noi + glt	75.11	75.54
orig + glt + ir	75.01	74.5
orig + spc + noi	75.62	75.33
With three augmentations		
orig + glt + spc + ir	75.73	75.39
orig + spc + ir + noi	76.16	75.87
orig + ir + noi + glt	75.16	74.54
orig + noi + glt + spc	76.03	75.71
With four augmentations		
orig + glt + spc + ir + noi	76.39	75.80

Table 6. Comparison of JTES performances (Unweighted Accuracy, UA)

Reference	Split	Features	Augmentation	UA (%)
[7]	SI	emo_large	No	87.88
[35]	SI	Mel-cepstrum	No	71.31
[33]	SI	ComParE	No	81.44
This study	SI	wav2vec 2.0	Yes	97.95
[7]	TI	emo_large	No	64.36
This study	TI	wav2vec 2.0	Yes	76.75
[7]	STI	emo_large	No	69.56
[8]	STI	HSFs of 5 frames	Yes	73.40
[29]	STI	LLDs	No	64.40
This study	STI	wav2vec 2.0	No	74.50
This study	STI	wav2vec 2.0	Yes	77.25

Table 7. Comparison of IEMOCAP performances

Reference	Test set	Features	Aug.	WA (%)	UA (%)
[38]	CV	UniSpeech-SAT Large	No	70.78	-
[37]	CV	HuBERT large	No	67.56	-
[36]	CV	<i>Audio</i> ₂₅ + GloVe + BERT	No	77.51	78.41
[36]	Session 5	<i>Audio</i> ₂₅ + GloVe + BERT	No	83.08	83.22
[37]	Session 5	HuBERT large	No	63.90	64.54
This study	Session 5	wav2vec 2.0	No	74.13	74.88
This study	Session 5	wav2vec 2.0	Yes	75.50	76.39

Table 6 shows the benchmark of our results with the literature studies. It is also clear from Table 6 that our acoustic feature engineering contributes more than data augmentation techniques. Without data augmentation, wav2vec 2.0 pre-trained on MSP-IMPROV dataset [25] achieved higher UAR than other proposed methods. Note that reference [35] only reported speaker-independent only for their acoustic-only SER while reference [29] reported text-dependent SER on the JTES dataset. A split scenario with speaker+text-independent, as conducted in this study, is more difficult than speaker-independent or text-independent according to the previous study [7].

A potential future research direction based on the findings in this study is to tune wav2vec 2.0 on the Japanese language with emotional contents. wav2vec is sensitive to language and could extract linguistic information within acoustic features [23]. Hence, training speech embeddings in the Japanese spoken language could be a promising research direction. This study evaluated wav2vec 2.0 pre-trained on the English emotional speech dataset (MSP-IMPROV) and then used that model to extract acoustic features from the Japanese dataset.

As for IEMOCAP, the benchmark is shown in Table 7. It shows that our results obtained in this study are competitive with the state-of-the-art results for the IEMOCAP dataset. It is worth noting that in [36], the authors combined acoustic and linguistic features that yielded higher results. Compared to recent speech embeddings like HuBERT and UniSpeech-SAT, our results without augmentation shows remarkable higher scores, highlighting the benefit of training such models on specific affective speech database. The WA and UA scores of HuBERT Large [37] were computed manually from the provided toolkit in addition to the official scores for cross-validation (CV) evaluation [6].

6. Conclusions

This paper reported the results of SER with different data augmentation techniques. In conclusion, adding more data does not always improve SER performance. For speaker-independent, adding more data tends to improve SER performance. For text-independent

data (including spaker+text independent), the most accurate model was achieved by using only two data augmentations with glottal source filtering and start-end silence removal. For the types of data augmentations that contribute to the performance of speech emotion recognition, we found that no specific type of data augmentation performs superior to other types of data augmentations. One type of data augmentation performs better when it is combined with other types of data augmentations. We also conclude that text-independent is more difficult and challenging than speaker-independent, as previous research revealed. The result of a text-independent scenario that yields an assumption that smaller data obtains better SER performance than larger data shows the necessity to choose a proper data augmentation technique for text-independent SER.

Future research can be directed to enlarge the number of data augmentation techniques and to observe if the performance of SER can be improved with different data augmentation techniques reported in this paper. Training SER in a specific language arguably will lead to performance improvement due to the impact of the linguistic-dependency. Other augmentation techniques like SpecAugment could be explored for SER, which is not currently evaluated due to data processing differences (spectrogram vs. waveform).

7. Acknowledgements

This paper is based on results obtained from a project, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO), Japan.

References

1. El Ayadi, M.; Kamel, M.S.; Karray, F. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognit.* **2011**, *44*, 572–587. <https://doi.org/10.1016/j.patcog.2010.09.020>.
2. Akçay, M.B.; Oguz, K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Commun.* **2020**, *116*, 56–76. <https://doi.org/10.1016/j.specom.2019.12.001>.
3. Zhang, Y.; Park, D.S.; Han, W.; Qin, J.; Gulati, A.; Shor, J.; Jansen, A.; Xu, Y.; Huang, Y.; Wang, S.; et al. BigSSL: Exploring the Frontier of Large-Scale Semi-Supervised Learning for Automatic Speech Recognition **2021**. pp. 1–14. [2109.13226].
4. Ko, T.; Peddinti, V.; Povey, D.; Khudanpur, S. Audio Augmentations for Speech Recognition. In Proceedings of the Interspeech 2015, 2015, Vol. 2, pp. 3–6.
5. Casanova, E.; Candido Jr., A.; Fernandes Jr., R.C.; Finger, M.; Gris, L.R.S.; Ponti, M.A.; Pinto da Silva, D.P. Transfer Learning and Data Augmentation Techniques to the COVID-19 Identification Tasks in ComParE 2021 **2021**. pp. 446–450. <https://doi.org/10.21437/interspeech.2021-1798>.
6. Yamada, Y.; Chiba, Y.; Nose, T.; Ito, A. Effect of Training Data Selection for Speech Recognition of Emotional Speech. *Int. J. Mach. Learn. Comput.* **2021**, *11*, 362–366. <https://doi.org/10.18178/ijmlc.2021.11.5.1062>.
7. Atmaja, B.T.; Sasou, A. Effect of different splitting criteria on the performance of speech emotion recognition. In Proceedings of the TENCON 2021 - 2021 IEEE Reg. 10 Conf. IEEE, 2021, pp. 760–764. <https://doi.org/10.1109/TENCON54134.2021.9707265>.
8. Chiba, Y.; Nose, T.; Ito, A. Multi-Stream Attention-Based BLSTM with Feature Segmentation for Speech Emotion Recognition. In Proceedings of the Interspeech 2020; ISCA: ISCA, 2020; Vol. 2020-October, pp. 3301–3305. <https://doi.org/10.21437/Interspeech.2020-1199>.
9. Takeishi, E.; Nose, T.; Chiba, Y.; Ito, A. Construction and analysis of phonetically and prosodically balanced emotional speech database. In Proceedings of the 2016 Conf. Orient. Chapter Int. Comm. Coord. Stand. Speech Databases Assess. Tech. O-COCOSDA 2016; IEEE: Bali, 2016; pp. 16–21. <https://doi.org/10.1109/ICSDA.2016.7918977>.
10. Atmaja, B.T.; Akagi, M. Deep Multilayer Perceptrons for Dimensional Speech Emotion Recognition. In Proceedings of the 2020 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. APSIPA ASC 2020 - Proc.; , 2020; pp. 325–331, [2004.02355].
11. Atmaja, B.T.; Akagi, M. Speech Emotion Recognition Based on Speech Segment Using LSTM with Attention Model. In Proceedings of the 2019 IEEE Int. Conf. Signals Syst. IEEE, 2019, pp. 40–44. <https://doi.org/10.1109/ICSIGSYS.2019.8811080>.
12. Singh, P.; Srivastava, R.; Rana, K.P.; Kumar, V. A multimodal hierarchical approach to speech emotion recognition from audio and text. *Knowledge-Based Syst.* **2021**, *229*, 107316. <https://doi.org/10.1016/j.knosys.2021.107316>.
13. Cummings, K.E.; Clements, M.A. Analysis of the glottal excitation of emotionally styled and stressed speech. *J. Acoust. Soc. Am.* **1995**, *98*, 88–98. <https://doi.org/10.1121/1.413664>.
14. Laukkanen, A.M.; Vilkmann, E.; Alku, P.; Oksanen, H. Physical variations related to stress and emotional state: A preliminary study. *J. Phon.* **1996**, *24*, 313–335. <https://doi.org/10.1006/jpho.1996.0017>.
15. Waaramaa, T.; Laukkanen, A.M.; Airas, M.; Alku, P. Perception of Emotional Valences and Activity Levels from Vowel Segments of Continuous Speech. *J. Voice* **2010**, *24*, 30–38. <https://doi.org/10.1016/j.jvoice.2008.04.004>.
16. Gobl, C.; Ní Chasaide, A. The role of voice quality in communicating emotion, mood and attitude. *Speech Commun.* **2003**, *40*, 189–212. [https://doi.org/10.1016/S0167-6393\(02\)00082-1](https://doi.org/10.1016/S0167-6393(02)00082-1).

17. Yao, X.; Bai, W.; Ren, Y.; Liu, X.; Hui, Z. Exploration of glottal characteristics and the vocal folds behavior for the speech under emotion. *Neurocomputing* **2020**, *410*, 328–341. <https://doi.org/10.1016/j.neucom.2020.06.010>.
18. Sasou, A. Glottal inverse filtering by combining a constrained LP and an HMM-based generative model of glottal flow derivative. *Speech Commun.* **2018**, *104*, 113–128. <https://doi.org/10.1016/j.specom.2018.07.002>.
19. Traer, J.; McDermott, J.H. Statistics of natural reverberation enable perceptual separation of sound and space. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, E7856–E7865. <https://doi.org/10.1073/pnas.1612524113>.
20. EchoThief Impulse Response Library.
21. Piczak, K.J. ESC: Dataset for environmental sound classification. *MM 2015 - Proc. 2015 ACM Multimed. Conf.* **2015**, pp. 1015–1018. <https://doi.org/10.1145/2733373.2806390>.
22. Wagner, Johannes, Triantafyllopoulos, Andreas, Wierstorf, Hagen, Schmitt, Maximilian, Eyben, Florian, & Schuller, B.W. Model for Dimensional Speech Emotion Recognition based on Wav2vec 2.0 (1.1.0). <https://doi.org/10.5281/zenodo.6221126>.
23. Wagner, J.; Triantafyllopoulos, A.; Wierstorf, H.; Schmitt, M.; Burkhardt, F.; Eyben, F.; Schuller, B.W. Dawn of the transformer era in speech emotion recognition: closing the valence gap **2022**. [2203.07378].
24. Baevski, A.; Zhou, H.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. Neural Inf. Process. Syst.* **2020**, 2020-December, 1–12, [2006.11477].
25. Busso, C.; Parthasarathy, S.; Burmanian, A.; Abdelwahab, M.; Sadoughi, N.; Provost, E.M.; Member, S.S.; Parthasarathy, S.; Member, S.S.; Burmanian, A.; et al. MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception. *Trans. Affect. Comput.* **2017**, *8*, 67–80. <https://doi.org/10.1109/TAFFC.2016.2515617>.
26. Atmaja, B.T.; Sasou, A.; Akagi, M. Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion. *Speech Commun.* **2022**, *140*, 11–28. <https://doi.org/10.1016/j.specom.2022.03.002>.
27. Atmaja, B.T.; Akagi, M. Evaluation of Error and Correlation-Based Loss Functions For Multitask Learning Dimensional Speech Emotion Recognition. In Proceedings of the Annu. Conf. Acoust. Vib., 2020, [2003.10724].
28. Satt, A.; Rozenberg, S.; Hoory, R. Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms. In Proceedings of the Interspeech 2017; ISCA: ISCA, 2017; Vol. 2017-Augus, pp. 1089–1093. <https://doi.org/10.21437/Interspeech.2017-200>.
29. Horii, D.; Ito, A.; Nose, T. Analysis of Effectiveness of Feature Extraction by CNN for Speech Emotion Recognition. In Proceedings of the ASJ Autumn Meet., 2021, pp. 969–972.
30. Rintala, J. Speech Emotion Recognition from Raw Audio using Deep Learning. PhD thesis, Royal Institute of Technology (KTH), 2020.
31. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825—2830, [1201.0490].
32. Jordal, I. Audiomentations.
33. Lee, S.w. The Generalization Effect for Multilingual Speech Emotion Recognition across Heterogeneous Languages. In Proceedings of the ICASSP 2019 - 2019 IEEE Int. Conf. Acoust. Speech Signal Process. IEEE, 2019, pp. 5881–5885. <https://doi.org/10.1109/ICASSP2019.8683046>.
34. Pepino, L.; Riera, P.; Ferrer, L.; Gravano, A. Fusion Approaches for Emotion Recognition from Speech Using Acoustic and Text-Based Features. In Proceedings of the ICASSP 2020 - 2020 IEEE Int. Conf. Acoust. Speech Signal Process. IEEE, 2020, pp. 6484–6488. <https://doi.org/10.1109/ICASSP40776.2020.9054709>.
35. Nagase, R.; Fukumori, T.; Yamashita, Y. Speech Emotion Recognition with Fusion of Acoustic- and Linguistic-Feature-Based Decisions. In Proceedings of the APSIPA Annu. Summit Conf., 2021, number December, pp. 725–730.
36. Wu, W.; Zhang, C.; Woodland, P.C. Emotion Recognition by Fusing Time Synchronous and Time Asynchronous Representations. In Proceedings of the ICASSP 2021 - 2021 IEEE Int. Conf. Acoust. Speech Signal Process. IEEE, 2021, Vol. 2021-June, pp. 6269–6273, [2010.14102]. <https://doi.org/10.1109/ICASSP39728.2021.9414880>.
37. Yang, S.w.; Chi, P.H.; Chuang, Y.S.; Lai, C.I.J.; Lakhota, K.; Lin, Y.Y.; Liu, A.T.; Shi, J.; Chang, X.; Lin, G.T.; et al. SUPERB: Speech Processing Universal PERFORMANCE Benchmark. In Proceedings of the Interspeech 2021; ISCA: ISCA, 2021; pp. 1194–1198, [2105.01051]. <https://doi.org/10.21437/interspeech.2021-1775>.
38. Chen, S.; Wu, Y.; Wang, C.; Chen, Z.; Chen, Z.; Liu, S.; Wu, J.; Qian, Y.; Wei, F.; Li, J.; et al. Unispeech-Sat: Universal Speech Representation Learning With Speaker Aware Pre-Training. In Proceedings of the ICASSP 2022 - 2022 IEEE Int. Conf. Acoust. Speech Signal Process. IEEE, 2022, pp. 6152–6156, [2110.05752]. <https://doi.org/10.1109/ICASSP43922.2022.9747077>.