Jointly Predicting Emotion, Age, and Country Using Pre-Trained Acoustic Embedding

Bagus Tris Atmaja  
AIST  
Tsukuba, Japan  
b-atmaja@aist.go.jp

Zanjabila  
ITS  
Surabaya, Indonesia  
zanjabilaabil@gmail.com

Akira Sasou  
AIST  
Tsukuba, Japan  
a-sasou@aist.go.jp

Abstract — In this paper, we demonstrated the benefit of using a pre-trained model to extract acoustic embedding to jointly predict (multitask learning) three tasks: emotion, age, and native country. The pre-trained model was trained with wav2vec 2.0 large and robust model on the speech emotion corpus. The emotion and age tasks were regression problems, while country prediction was a classification task. A single harmonic mean from three metrics was used to evaluate the performance of multitask learning. The classifier was a linear network with two independent layers and shared layers connected to the output layers. This study explores multitask learning on different acoustic features (including the acoustic embedding extracted from a model trained on an affective speech dataset), seed numbers, batch sizes, and waveform normalizations for predicting paralinguistic information from speech.

Index Terms — speech emotion recognition, affective computing, acoustic embedding, multitask learning, age prediction, country prediction

I. INTRODUCTION

Research on computational paralinguistics advances with the advent of artificial intelligence, big data, and speed up internet access. Among many issues, computational paralinguistics research still relies on limited small data for experiments [1]. The need to conduct research on big data is needed for generalization and performance improvements.

A large-scale dataset was made for accelerating research on computational paralinguistics, the HUME-VB dataset [2]. The dataset contains more than 36 hours of audio data with a 48 kHz of sampling rate. The samples in the dataset were labeled with three pieces of information: intensity rating for ten different expressed emotions (float in ranges [0, 1]), age (integer), and country (string). This dataset was created to support the following tasks: multitask learning, emotion generation, and few-shot learning. Multitask learning includes emotion recognition from vocalizations.

Speech emotion recognition is a branch of computational paralinguistics that deals with the accurate prediction of emotional score/class from speech. Since the expression of speakers’ emotions can be perceived by the human ears of listeners, it is also possible for the computer to have the same ability: recognize emotion from the sound. Instead of a speech, which has semantic or linguistic meanings, the burst of brief vocalizations is an interesting source for predicting expressed emotion in human social life since it contains rich information on emotion [3].

Humans communicate emotion through two different kinds of vocalizations, prosody and vocal burst [4]. Prosody interacts with the words (linguistic) to convey feelings and attitudes like in speech, while vocal burst just occurs without linguistic meanings. Examples of vocal bursts are laughs, cries, sighs, shrieks, growls, roars, and oohs. While the study of the prosody of speech for speech emotion recognition is widely conducted (indicated by the number of available datasets), the study of the vocal burst emotion recognition is currently undergoing with support of new vocalization datasets, e.g., [2], [5].

It is not the only emotion that can be recognized in the human’s voice. Age, gender, and nationality could also be detected from voices. The change of voice in age is recognizable by both humans and computers, although it is the hardest among emotion and gender [6]. By transferring information about age and gender, the recognition of dimensional emotion could be improved [7]. The task of combining several tasks together based on the same or different inputs is known as multitask learning.

At the ICML Expressive Vocalizations (ExVo) Workshop and Competition, a multitask learning task is held to predict the average intensity of each of 10 emotions perceived in vocal bursts, the speaker’s age, and the speaker’s Native-Country. The challenge utilized a large vocalization dataset [3] which was shared across three different tasks. The emotion and predictions are regression problems; native-country prediction is a classification problem. The participants were evaluated by a single score of the harmonic mean from these three problems.

This paper contributes to the research on affective vocal bursts by evaluating the pre-trained speech embeddings extractor, which was trained specifically on an affective speech dataset, to jointly predict emotion, age, and the country as in the ExVo challenge. We hypothesize that using this kind of acoustic feature extractor will lead to better results than traditional feature extractors, which extract physical information of audio signals. The rest of this paper explains related work, methods, results and discussion, and conclusion.

II. RELATED WORK

Research on multitask learning, predicting several tasks simultaneously using unified architecture, progressively in-
increased due to its effectiveness in predicting several outputs with the same or similar inputs. Speech processing is an ideal test bed for multitask learning. Several pieces of information could be extracted from the same speech input, whether it is text (automatic speech recognition, ASR), gender, age, nationality, emotion, language, and disease. The following is a summary of the work of multitask learning done in the past with key differences from this study presented at the end of this section.

Parthasarathy and Busso [8] proposed two schemes of multitask learning architectures for evaluating valence, arousal, and dominance simultaneously. The first architecture is without an independent layer for each task (shared layers only), while the second architecture is with independent layers for each task. They found the second architecture with independent layers performed better than the first one without independent layers. The scores (measured in concordance correlation coefficient) also showed improvements from the baseline architecture with single-task learnings.

Lee [9] also evaluated two architectures of multitask learning optimized for emotion recognition. The first architecture is composed of three task-specific softmax layers to predict gender, emotion, and language. The second architecture is composed of one softmax layer containing two tasks, language and emotion. The results revealed a better generalization of the second architecture to predict emotion categories.

Kim et al. [10] proposed to use gender and naturalness information to minimize the large mismatch between the training and test data in speech emotion recognition in the wild. The method employed traditional acoustic features \( f_o \), voice probability, zero-crossing-rate, MFCC with their energies and first-time derivatives) extracted on frame-level and calculated the high-level features on top of it. The high-level features were then fed into an extreme learning machine to predict the categories of emotion. They obtain significant improvement over the baseline with single-task learning. The method was evaluated on five different datasets for generalization.

Atmaja and Akagi [11] evaluated different loss functions – CCC, MSE, and MAE – for multitask learning dimensional emotion recognition. The traditional approach of deep learning commonly employed MSE loss to minimize the error and training stage. Since the goal is to maximize CCC, they proposed to use CCC loss as the loss function to replace MSE. The results on two datasets (IEMOCAP and MSP-IMPROV) and two acoustic features (GeMAPS and python Audio Analysis) showed the consistency that CCC loss is superior to MSE and MAE.

Li et al. [12] proposed additional information on age, gender, and emotion for speaker verification using multitask learning and domain adversarial training. The multitask learning part minimizes losses of three variables (speaker, gender, and nationality). The domain adversarial training, which also employs multitask learning, minimizes losses of speaker and emotion. The results showed that multitask improved the performance from the baseline by about 16% while the domain adversarial training improved the performance from the baseline by about 22%. The baseline used ResNet networks.

Cai et al. [13] employed multitask learning by predicting text characters and predicting emotion in the training phase. The model is trained to minimize the loss of categorical emotion (cross-entropy) and loss of character recognition (connectionist temporal classification). The inference phase removes the character recognition path to predicting emotion categories only. The proposed multitask learning achieved an accuracy of 78% compared to 72% of the baseline method with capsule networks.

Atmaja et al. [14] evaluated multitask learning of emotion recognition (dimensional) naturalness scores from speech. They evaluated two different architectures with and without independent layers. The architecture without independent layers (shared layers only) exhibits the best performance in predicting valence, arousal, and dominance scores. The shared layers have been built using three layers of fully connected networks with nodes of 512, 256, and 128. However, the scores for naturalness recognition in multitask learning is lower than in single-task learning.

Instead of focusing on the single task evaluation on multitask learning (e.g., only predicting emotion in multitask emotion and transcription [13], or emotion and language [9]), this study focused on all the tasks evaluated on the multitask learning. We employed the harmonic mean evaluation from three metrics for three tasks and used this harmonic mean as the final evaluation. This multitask learning evaluation using all tasks was not evaluated on the previous tasks, where the authors only focused on emotion recognition or speaker verification.

### III. Methods

#### A. Datasets

This study relies on the HUME-VB dataset, which is used at the ICML Expressive Vocalization (ExVo) Competition 2022. The dataset is a large-scale emotional non-linguistic vocalization known as vocal burst. An example of this vocal burst is “argh!” to express distress emotion. There are ten emotions rated in continuous scores. These emotions are Amusement, Awe, Awkwardness, Distress, Excitement, Fear, Horror, Sadness, Surprise, and Triumph. The data were collected from 1702 speakers aged 20-39 years old. The collection locations are China, South Africa, Venezuela, and the US. The total duration of the dataset is almost 37 hours (36 hours 47 minutes). Although the data were split into train, validation, and test, the test set was closed by the organizer of the competition. Hence, we evaluated our methods mostly on the validation set (except in the last part, where test results are reported). More details about the dataset can be found in the [2].

#### B. Pre-trained Acoustic Embedding

We evaluated a pre-trained model finetuned on an affective speech dataset. The model [15], [16] is based on wav2vec2-large-robust model [17]. The model is trained on MSP-Podcasts dataset [18], a large affective speech corpus.
derived from YouTube with valence, arousal, and dominance scores. For this finetuning, the combined samples on the
MSP-Podcasts dataset have a combined length of roughly 21
hours. The model extracted the speech embedding from the
dataset (HUME-VB) with a size of 1024 dimensions for each
utterance. The output layers of the model, i.e., the logits, are
the scores of arousal, dominance, and valence, in ranges [0,
1]. We experimented with two variants of speech embedding
with this model. The first embedding is the hidden states of
the last layer before the output layer (1024-dim), and the
second embedding is the concatenation of hidden states with
the logits (1027-dim). We named the first embedding “w2v2-
R-er” (wav2vec 2.0 robust emotion recognition), and the
second embedding “w2v2-R-vad” (wav2vec 2.0 robust emotion
recognition with valence, arousal, and dominance). Notice the
term acoustic embedding (extracted from vocal bursts) here is
used instead of (traditional) acoustic features or speech
embedding (extracted from speech).

C. Classifier

We employed a model of multitask learning adopted from
[2]. The architecture of the model is shown in Figure 1. The
model accepts the input of acoustic features depending on the
size of the features. For instance, the dimension for w2v2-
R-er is 1024. Then, two linear layers are stacked as shared
layers. The number of nodes for each layer is 128 and 64,
respectively. We applied layer normalization [19] for each
linear layer, followed by LeakyReLU activation functions. A
group of layers builds up the independent layers for each
task. For emotion and country prediction, there is only a layer
followed by output layers. For age prediction, we used two
independent layers to bridge the gap in the big number of
nodes from the shared layer (64) to the single-node output
layer. The nodes for these independent layers are 32 and 16,
respectively.

The loss function minimizes losses of three tasks: MSE for
emotion recognition age prediction and cross-entropy (CE)
for country prediction. The total loss function \( \mathcal{L}_T \) is the
weighting sum of three losses given by the following formula,

\[
\mathcal{L}_T = \sum_{i=1}^{3} \left( \frac{\mathcal{L}_i}{2\sigma^2} + \frac{\alpha}{2} \right).
\]  

(1)

The coefficient of \( \alpha \) is set to be 0.34, 0.33, and 0.33 for
emotion, country, and age, respectively.

D. Evaluation Metrics

The evaluation of three paralinguistic tasks in this study
used a specific metric for each task. Emotion recognition is
evaluated in a concordance correlation coefficient (CCC), age
recognition is evaluated in mean absolute error (MAE), and
nationality prediction is evaluated in unweighted average recall
(UAR). These metrics are described below,

\[
CCC = \frac{2\hat{\rho}_{xy}^2}{\hat{\sigma}_x^2 + \hat{\sigma}_y^2 + (\mu_x - \mu_y)^2}. 
\]  

(2)

The average CCC scores for ten emotion categories is the
mean values,

\[
\hat{\mathcal{C}} = \sum_{i=1}^{10} CCC_i / 10. 
\]  

(3)

CCC is in range [-1, 1] with zero for perfect concordance/disconcordance, and 1 for perfect concordance.

Next is a metric to evaluate the performance of nationality/country prediction that is unweighted accuracy, also known
as unweighted average recall (UAR) and balanced accuracy.
Unweighted accuracy is formulated as,

\[
\hat{U} = \frac{1}{4} \sum_{i=1}^{4} \text{Recall}_i
\]  

(4)

where \( i \) is the corresponding country class, and 4 is the number
of countries (USA, China, South Africa, and Venezuela). UAR
ranges in 0-100 in % or 0-1 in normalized score.

The metric to evaluate the last task, age prediction, is mean
absolute error (MAE). MAE is a common metric for evaluating
regression, and it is scale-dependent. The lower the scores, the
better the age prediction. The formulation of MAE is given by

\[
MAE = \frac{1}{n} \sum_{i=1}^{n} |x_i - y_i| 
\]  

(5)

where \( n \) is the number of samples in the evaluation or test
sets (which one is used to calculate the score).

Since MAE is scale-dependent, we inverted the MAE scores
for consistency with the previous two metrics,

\[
\hat{M} = 1/MAE. 
\]  

(6)

Now, for all metrics (CCC, UAR, 1/MAE), the higher scores,
the better predictions of emotion, country, and age. Finally,
for calculating overall performance, we used the harmonic
mean of three metrics above [2],

\[
S_{MTL} = \frac{3}{(1/\hat{C} + 1/\hat{M} + 1/\hat{U})}. 
\]  

(7)

\( S_{MTL} \) is our main metric to judge the performance of the
evaluated methods (hyperparameters, features, normalization).
Other previous metrics are used to determine the performance
of the corresponding method for individual tasks in multitask
learning.

IV. RESULTS AND DISCUSSION

We presented our results in different ablation studies:
choosing the right seed, comparing different acoustic features,
effects of normalization, and test results. For each study, we
run the experiment five times for each setting. For instance,
in choosing the right seed, we run the experiment five times on
seed “101”. The reported results are the average scores,
except for comparing acoustic features. The reported scores for
w2v2-R-er and w2v2-R-vad are chosen from the best from five
different runs, similar to the baseline. For the baseline results,
we quote the scores from the source [2].
A. Choosing the Right Seed

Seed initialization is an important step in deep learning methods [20], [21]. Here, as a first step, we choose the seven different seed values to choose the best one. We evaluated seed values 42, 101, 102, 103, 104, 105, 106. The reason for adding seed 42 to the other six seed values given in the baseline is its common use in the deep learning community.

Table I shows the results of using different seed values for multitask learning of emotion, age, and country on the validation set. The results are average scores of five trials with their standard deviation. Since the values of STD are similar in other ablation tests in this study, we only report these STD scores in Table I. The following reported scores, except stated, used the seed values of “106” which obtained the highest $S_{MTL}$ score from seed values evaluation.

B. Comparison of Different Features

The main focus of this study is to evaluate the acoustic embedding extracted using pre-trained models trained on the emotional speech dataset. Most speech processing tasks, especially speech emotion recognition, are trained using hand-crafted acoustic feature extractors (e.g., MFCC, spectrogram, or mel-spectrogram) [22]. Others are using pre-trained models but trained on neutral speech (e.g., wav2vec2 2.0, HuBERT, WavLM). In this study, we utilized a pre-trained model [16], [16] built using wav2vec 2.0 Robust on affective speech dataset.

Table II proves our presumption that our model will surpass the baseline scores. Both w2v2-R-er and w2v2-R-vad obtain higher $S_{MTL}$ scores than the baseline scores on the same configuration (batch size, seed, and other hyperparameters). Specifically, these scores obtained by two acoustic embeddings show the most remarkable improvement in emotion recognition score, in which the model to extract the acoustic embedding is trained. Not only for the emotion recognition task, but both scores for age and country predictions were also improved. The pre-trained model is shown to be helpful on other tasks probably due to the similarity of the task (paralinguistics and non-linguistic tasks), and the data trained to build the model contains the age and country information embedded on the extracted acoustic embedding. The data to train the model is MSP-Podcast in the English language. Although the dataset contains English only language, the pre-trained model may be able to discriminate between English with non-English language (related to country prediction) as in anomaly detection problems.

C. Evaluation of Different Batch Sizes

We evaluated five different batch sizes since there is evidence of the influence of batch size, particularly for emotion recognition task [23]. As found in the [23], we also found that the smallest evaluated batch size in this study resulted in the best performance. As the batch size increases, the performance decreases. For the best performance on the use of batch size $= 2$ (scores in Table III are average of 5 runs), we obtained the following scores: 0.534, 0.533, 0.267, and 0.401 for EmoCCC, Cou-UAR, and Age-1/MAE, and $S_{MTL}$, respectively. This is the highest $S_{MTL}$ score obtained in this study.

D. Effect of Waveform Normalization

Normalization may affect the performance of deep learning, particularly in speech processing. This effect is due to the model usually only working on the standard input and being prone to high deviation. It has been proved that such normalizations are effective for deep learning, e.g., batch normalization [24], group normalization [23], and layer normalization [19]. Aside from layer normalization, we also evaluated the model with waveform normalization. In this case, we utilize librosa toolkit [25] to normalize the audio array (the amplitude of waveform), i.e., the output is in the range [-1, 1]. Then, two
consistent with and without normalization, highlighting the two acoustic embeddings, w2v2-R-er and w2v2-R-vad, that discriminate these paralinguistics labels. The results for the normalization of the waveform will remove important information related to the loudness of the waveform. Hence, such normalization by normalizing the waveform of speech. This finding may be by normalizing the waveform of speech. This finding may be explained by the fact that emotion, age, and gender are related to the loudness of the waveform. Hence, such normalization of the waveform will remove important information that discriminates these paralinguistics labels. The results for two acoustic embeddings, w2v2-R-er and w2v2-R-vad, are consistent with and without normalization, highlighting the unnecessary processing of normalizing the speech waveform.

| Feature      | Dims. | CCC | UAR | I/MAE | S_{MTL} |
|--------------|-------|-----|-----|-------|---------|
| ComParE      | 6373  | 0.416 | 0.506 | 0.237 | 0.339   |
| eGeMAPS      | 88    | 0.353 | 0.423 | 0.249 | 0.324   |
| BoAW         | 125   | 0.335 | 0.417 | 0.234 | 0.311   |
| DeepSpec     | 4096  | 0.369 | 0.456 | 0.227 | 0.322   |
| w2v2-R-er    | 1024  | 0.533 | 0.523 | 0.252 | 0.386   |
| w2v2-R-vad   | 1027  | 0.534 | 0.525 | 0.253 | 0.388   |

E. Test Results

We presented the results of our predictions from the previous validation evaluation for the closed test set evaluation. In the submitted results, we evaluated three predictions from both w2v2-R-er (1st and 2nd) and w2v2-R-vad (3rd) feature sets with different seeds and batch sizes. The first prediction is with a batch size of 8 and a seed of 42. The second prediction is with a batch size of 4 and a seed of 106. The third prediction is with a batch size of 2 and a seed of 106.

Table V shows the score of harmonic mean (S_{MTL}) from our predictions (last three rows) and the baseline [2]. All of our submitted results obtained higher S_{MTL} scores than the highest baseline results. These scores reveal the effectiveness of the evaluated acoustic embedding, w2v2-R-er and w2v2-R-vad, which were extracted using the robust version of wav2vec 2.0 [26] on the MSP-Podcast emotional speech dataset [18]. There are no remarkable differences between the second and third scores. Also, the statistics (average and standard deviation) of non-deterministic predictions could not be calculated for the test set due to restrictions on the number of trials.

| Feature          | Batch size | Seed | S_{MTL} |
|------------------|------------|------|---------|
| ComParE          | 8          | best | 0.335   |
| eGeMAPS          | 8          | best | 0.314   |
| BoAW-125         | 8          | best | 0.299   |
| BoAW-250         | 8          | best | 0.305   |
| BoAW-500         | 8          | best | 0.302   |
| BoAW-1000        | 8          | best | 0.307   |
| BoAW-2000        | 8          | best | 0.303   |
| DeepSpec         | 8          | best | 0.305   |
| w2v2-R-er        | 8          | 42   | 0.358   |
| w2v2-R-er        | 4          | 106  | 0.378   |
| w2v2-R-vad       | 2          | 106  | 0.373   |

V. Conclusions

In this paper, we reported evaluations of multitask learning to jointly predict emotion, age, and country using acoustic embedding extracted from a pre-trained model. The model is finetuned on an affective speech dataset. The extracted acoustic embeddings were fed to an architecture consisting of shared layers and independent layers for these three tasks. The results showed improvements over the baseline methods with the common speech representations (ComParE, eGeMAPS,
BoAW, and DeepSpectrum). Two variants of acoustic embeddings are evaluated with original hidden states and concatenation of hidden states with logits. The latter performed best on the validation set. We conducted ablation studies on different seeds, batch sizes, and normalization. The study finds the optimum seed and batch size of the evaluated ranges and finds no improvement in performing waveform normalizations.

While this study treated all emotions in the same weights, future studies may be directed to adjust these weights for optimum emotion recognition, as well as to improve overall harmonic mean evaluation for all tasks in multitask learning. Future work could also be directed to evaluate the effect of finetuning wav2vec 2.0 on the affective speech dataset (i.e., comparing wav2vec 2.0 with and without finetuning on that dataset). The results then can be compared against other state-of-the-art acoustic embeddings, e.g., End2You [27].

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

REFERENCES
[1] A. Batliner, S. Hantke, and B. W. Schuller, “Ethics and Good Practice in Computational Paralinguistics,” IEEE Trans. Affect. Comput., vol. 3045, 2020.
[2] J. Kim, G. Englebienne, K. P. Truong, and V. Evers, “Towards Speech Emotion and Naturalness Recognitions With Multitask and Single-Task Learnings,” IEEE Access, vol. 10, pp. 72 381–72 387, 2022.
[3] S.-w. Lee, “The Generalization Effect for Multilingual Speech Emotion Recognition across Heterogeneous Languages,” in ICASSP 2019 - 2019 IEEE Int. Conf. Acoust. Speech Signal Process. IEEE, may 2019, pp. 5881–5885.
[4] J. Kim, G. Englebienne, K. P. Truong, and V. Evers, “Towards Speech Emotion Recognition “in the Wild” Using Aggregated Corpora and Deep Multi-Task Learning,” in Interspeech 2017. ISCA: ISCA, aug 2017, pp. 1103–1107.
[5] Y. Wu and K. He, “Group Normalization,” 2018. [Online]. Available: https://github.com/facebookresearch/Detectron/
[6] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, “Dawn of the transformer era in speech emotion recognition: closing the valence gap,” mar 2022. [Online]. Available: https://arxiv.org/abs/2203.07378
[7] A. Weiss, “librosa/librosa: 0.7.2,” jan 2020. [Online]. Available: https://github.com/librosa/librosa/tree/v0.6.0.0
[8] H. Zhao, N. Ye, and R. Wang, “Transferring Age and Gender Attributes for Dimensional Emotion Prediction from Big Speech Data Using Hierarchical Deep Learning,” in 2018 IEEE 4th Int. Conf. Big Data Secur. Cloud (BigDataSecurity), IEEE Int. Conf. High Perform. Smart Comput. IEEE Int. Conf. Intell. Data Secur. IEEE, may 2018, pp. 20–24.
[9] P. Friesch, F.-R. Stoter, D. Herehno, T. Kim, M. Vollrath, and P. Friesch, F.-R. Stoter, D. Herehno, T. Kim, M. Vollrath, “A framework for self-supervised learning of speech representations,” Adv. Neural Inf. Process. Syst., vol. 2020-Decem, no. Figure 1, pp. 1–12, 2020.
[10] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kain, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, “Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training,” in Interspeech 2021, vol. 3. ISCA: ISCA, aug 2021, pp. 721–725.
[11] B. T. Atmaja, A. Sasou, and M. Akagi, “Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion,” Speech Commun., vol. 140, pp. 11–28, may 2022.
[12] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer Normalization,” arXiv:1607.06450v1, 2015.
[13] B. T. Atmaja, A. Sasou, and M. Akagi, “Speech Emotion and Naturalness Recognitions With Multitask and Single-Task Learnings,” in Interspeech 2021. ISCA: ISCA, aug 2021, pp. 4508–4512.
[14] B. T. Atmaja, A. Sasou, and M. Akagi, “Speech Emotion and Naturalness Recognitions With Multitask and Single-Task Learnings,” IEEE Access, vol. 10, pp. 72 381–72 387, 2022.