PrimaDNN':A Characteristics-aware DNN Customization for Singing Technique Detection

Yuya Yamamoto
Doctoral Program in Informatics
University of Tsukuba,
Ibaraki, Japan
s2130507@s.tsukuba.ac.jp

Juhan Nam
Graduate School of Culture Technology
KAIST
Daejeon, South Korea
juhan.nam@kaist.ac.kr,

Hiroko Terasawa
Doctoral Program in Informatics
University of Tsukuba,
Ibaraki, Japan
terasawa@slis.tsukuba.ac.jp

***Abstract*—Professional vocalists modulate their voice timbre or pitch to make their vocal performance more expressive. Such fluctuations are called singing techniques. Automatic detection of singing techniques from audio tracks can be beneficial to understand how each singer expresses the performance, yet it can also be difficult due to the wide variety of the singing techniques. A deep neural network (DNN) model can handle such variety; however, there might be a possibility that considering the characteristics of the data improves the performance of singing technique detection. In this paper, we propose PrimaDNN, a CRNN model with a characteristics-oriented improvement. The features of the model are: 1) input feature representation based on auxiliary pitch information and multi-resolution mel spectrograms, 2) Convolution module based on the Squeeze-and-excitation (SENet) and the Instance normalization. In the results of J-POP singing technique detection, PrimaDNN achieved the best results of 44.9% at the overall macro-F measure, compared to conventional works. We also found that the contribution of each component varies depending on the type of singing technique.**

***Index Terms*—singing techniques, audio feature extraction, deep neural network**

I. Introduction

A singing voice is one of the most essential elements of music, providing impactful emotional expressions through melody and lyrics. In particular, in popular music, the role of the singing voice is even more critical, as the vocal quality and unique style of singers are crucial in attracting people. Vocals mainly consist of singer individuality (i.e., vocal fold vibration and vocal tract resonance) and singing expressions (i.e., fine control of pitch, timbre, and loudness). We define the latter as singing techniques, which are the singing voice versions of extended playing techniques. Automatic identification of singing techniques from sung voice tracks can contribute to the understanding of different singing styles, and can have applications in music discovery, vocal training, and user-generated content. It can also help to reduce the laborious and specialized process of analyzing singing techniques for billions of songs. In our previous work [1], we tackled singing technique detection from real-world repertoires in J-POP, where the demand for singing technique analysis is high. Figure 1 shows a quick overview of singing technique detection. It is a multi-class, multi-label classification in each analysis frame, where the input is an audio track of a singing voice, and the output is a timeline of the appearance of singing techniques. The difficulties of the task lie in localization and classification, where a wide variety of noise and fluctuation exists.

Recently, deep neural networks have achieved high performance on identification tasks, even in challenging conditions. Therefore, in our previous work [1], we adopted the CRNN model, which is one of the succeeding DNN architectures in many audio and music identification tasks [2]–[4] in a series of experiments. We found that DNN models considering the characteristics of data have the potential to improve identification results. Figure 2 shows the spectrogram of singing techniques that we analyzed in this paper. Each technique displays different pitch modulation or spectral patterns. For instance, Vibrato' shows a sinusoidal-shaped periodic pitch modulation, whereas Scooping' shows an S-shaped continuous pitch change. In terms of timbral techniques, Vocal fry' exhibits fast pulsive patterns, while Rasp' shows sub-harmonics. Therefore, the model must capture such a wide variety of acoustic characteristics to improve the detection performance.

In this paper, we reconsider the architecture of DNN by considering the characteristics of singing techniques. To achieve the aforementioned improvement, our model focuses on the following two aspects: 1) feature representation that captures the wide variety and 2) the mechanism that suppresses the effect of features that have nothing to do with the desired targets. For the first aspect, we adopt two approaches: multi-resolution mel spectrograms to capture various types of modulation, and mel-band pitchgram that explicitly informs the sung pitch heights. For the second aspect, we also adopt two approaches: Squeeze-and-Excitation network to dynamically select the important feature map on the convolutional layer, and Instance normalization to prevent instance-specific mean and co-variance shift that may impede the capturing of target features. We named the DNN model PrimaDNN' (pronounced prima-don-na)[1].

II. Architecture

Fig. 3 shows our proposed PrimaDNN' model. It is following a CRNN model that has four convolutional layers, 1 Bi-directional LSTM layer, 1 Fully-connected (FC) layer and

[1]We provide the detail at https://yamathcy.github.io/eusipco23primadnn/.

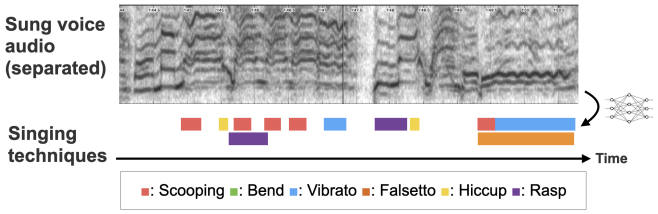

Fig. 1. The overview of singing technique detection. It is a multi-label and frame-wise classification that locates and identifies the singing techniques given a sung audio clip.

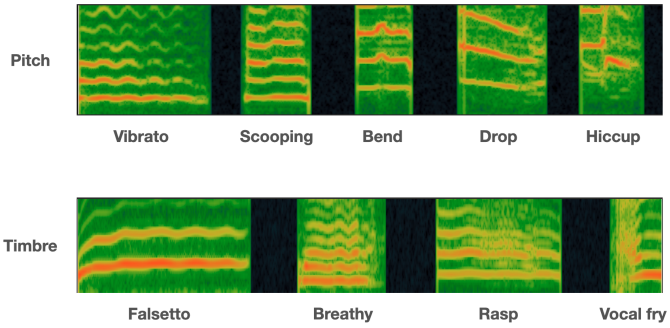

Fig. 2. The spectrograms of nine singing techniques that we treat in the experiment. The upper and the lower show pitchy and timbral techniques, respectively.

1 sigmoid activation layer. Only in the inference time, the output is binarized by thresholding with the value of 0.5.

A. Input feature

To overcome issues we use stacked multi-resolution mel spectrograms and 2D Mel band pitchgram for the input feature.

Multi-resolution mel spectrograms (MMelSpecs) are made by stacking three mel spectrograms which have different time-frequency resolutions with each other, in order to adapt wide modulation patterns both on time and frequency bands of singing techniques [5]. We adopt window sizes of (2048, 1024, and 512) for short-time Fourier Transform (STFT) with Hann-window, maintaining the same size for all mel spectrograms by zero-padding and applying fixed hop size. All of these mel spectrograms have a frequency dimension of 160 and each frame length of 10 ms.

In addition, we stacked **Mel-band pitchgram** [1], [6] on MMelSpecs. It has the same frequency dimensions as the input mel spectrograms and has one-hot where the pitch frequency exists. We use the pitch that is automatically estimated by CREPE [7], one of the state-of-the-art pitch extraction. Note that although the conventional work [1] that using ground-truth pitch shows the best performance, using CREPE also shows competitive results.

B. DNN architecture

We adopt **Squeeze-and-Excitation (SENet)** [8] and Instance normalization [9] for customization of the convolution layers of CRNN model. SENet is originally proposed in image domain, in order to enhance the representative power of a neural network by feature re-calibration that emphasizes informative features and suppresses useless ones. As the right side of Fig. 3 shows, SENet squeezes the input feature maps by Global average pooling, then reduces the channel dimension with a ratio of r on the first fully connected (FC) layer. Finally, the second FC layer rescales the channel dimension and outputs the importance of each feature map, which has a value range of $[0, 1]$. In all of the conditions that use SE, we empirically set r to 2 from the grid search on the range of $[16, 8, 4, 2]$.

For the normalization method, we use **instance normalization (IN)** instead of batch normalization (BN) everywhere in the network with the purpose of leading the model to focus on features relates to singing techniques. IN prevents instance-specific mean and covariance shift simplifying the learning process. IN is mainly used in style transfer to disentangle the content and style [10]. In the audio domain, it is used for speaker emotion recognition [11], speaker conversion [12] to suppress the effect of non-target attributes (e.g., speaker information, speech content, etc.) We expect that IN can get invariance of irrelevant attributes to singing techniques (e.g., singer identity, vocal mixing style, quality of vocal separation, vocal note density, etc.)

We trained the model using **Focal loss** [13]. Singing technique detection is difficult due to data imbalance, which can negatively affect detection performance. Focal loss addresses this by focusing training on hard examples (i.e., the frames where singing techniques appear in this case) and down-weighting the loss assigned to easy examples. The equation of Focal loss given the output activation p, is as follows:

$$L_{fl}(p_t) = -\alpha(1-p_t)^{\gamma}\log(p_t) \tag{1}$$

$$p_t = \begin{cases} p & label = 1 \\ 1-p & otherwise \end{cases} \tag{2}$$

$\alpha \in [0, 1]$ is a weighting factor for balancing the importance of positive and negative examples, and the term $(1-p_t)^{\gamma}$ is a modulating factor, with γ controlling the rate of dominant examples. We conducted a grid search on the range of $\alpha = [0.1, 0.13, 0.15, 0.2, 0.25]$ and $\gamma = [1, 1.33, 1.66, 2.0]$ and set α to 0.13 and γ to 1.33 for all conditions in the work that used focal loss.

III. Experiments

We conduct an experiment of nine-way singing technique detection.

A. dataset

We evaluated the proposed architecture using the COSIAN dataset [1], which includes 168 tracks of four famous hit songs sung by 42 solo singers of both genders. For the experiment, we selected the most common nine techniques. We processed the vocal tracks by sampling them to 44.1 kHz, separating them using Demucs v3 [14], and segmenting them into 10-second non-overlapping parts. We then obtained input features using MMelSpecs by applying short-time Fourier transform

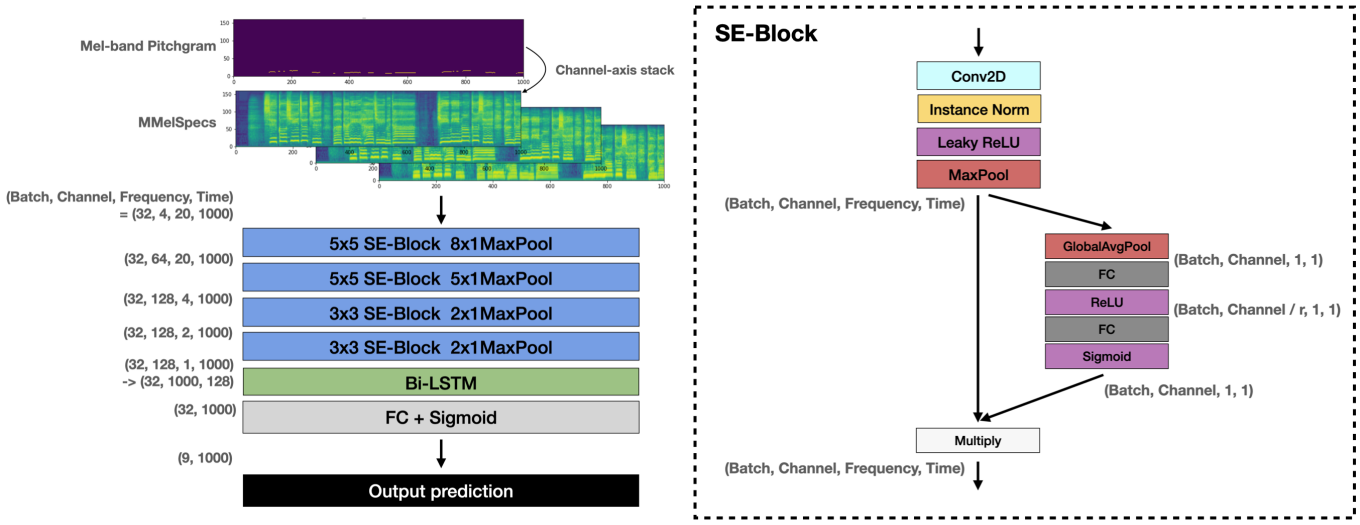

Fig. 3. The overview of PrimaDNN' model. (Left) the diagrams of architecture. (Right) the diagrams of SE-Block.

(STFT) with a 2048-sample Hann window and a hop size of 10 ms.

B. evaluation

To evaluate the performance of our proposed architecture, we conducted singer-wise seven-fold cross-validation, as in our previous work [1]. We divided the singers into seven groups and organized the dataset into training, validation, and test sets with a ratio of 5:1:1 for each set.

To account for label imbalances between singers, we used the nine most common singing techniques ('bend', 'breathy', 'drop', 'falsetto', 'hiccup', 'rasp', 'scooping', 'vibrato', and 'vocal fry'), which appeared in every fold of the cross-validation.

Our evaluation metrics included segment-based recall (**R**), precision (**P**), macro-F-measure (**Macro-F**), and micro-F-measure (**Micro-F**) [15], as well as the F-measure for each singing technique. We calculated these metrics using sed_eval[2]. The macro-F-measure represents the class-wise average of the F-measure, while the micro-F-measure represents the instance-wise average. We set the segment length to 100 ms for our evaluation.

IV. Results and Discussion

A. Comparison with baseline

First, we compare our proposed model with baseline models. As baselines, we prepared four conventional models. 1) *eGeMaps LSTM* [16]: eGeMaps [17] is a feature set used in speech emotion recognition tasks. It consists of 25 low-level descriptors for each frame. In this model, we used eGeMaps as an input feature and fed it to an LSTM model. 2) *CRNN* [18] A simple CRNN model whose input is a 64-dimensional Mel spectrogram and has three convolutional layers, one Bi-GRU layer and one FC layer. 3) *CNN Self-Attention* [18], [19] Instead of Bi-GRU layer, multi-head attention is applied. This model achieved the best performance on sound event detection with data imbalance situation [18]. In addition, we also compared with *CRNN+PitchFocal*, a CRNN that is fed the Mel-band pitchgram and applied Focal loss, both of which improved the performance of singing technique detection [1]. All models were trained using the RAdam optimizer [20] with a learning rate of 1e-3. Training stopped if the value of the loss function on the validation set did not improve for 20 epochs.

We used binary cross entropy (BCE) as the loss function for *eGeMaPS*, *CRNN*, *CNN Self-Attention* and Focal loss [13] for *CRNN+PitchFocal* as in the original work.

TABLE I
The results of singing technique detection.

	Macro-F	Micro-F	P	R
eGeMaps LSTM	9.2%	6.3%	11.3%	1.6%
CRNN	37.7%	56.3%	42.2%	39.2%
CRNN+PitchFocal	40.2%	55.1%	37.7%	48.0%
CNN Self-Attention	42.0%	59.3%	43.4%	47.7%
PrimaDNN' (ours)	**44.9%**	**60.6%**	**43.8%**	**48.3%**

Table. I displays the results of the experiment. PrimaDNN' achieved 44.9% at *Macro-F*, 60.6 % at *Micro-F*, 43.8% at *Precision* and 48.3% at *Recall*, respectively, as shown in the bottom of the table. These results indicate that PrimaDNN' outperformed the conventional models in all of the metrics.

B. Ablation study

In order to understand the contribution of each component in our model, we conducted an ablation study by comparing our full model with several modified versions, as outlined below:

- **Single resolution**: Uses only a single resolution mel spectrogram that was processed by STFT with window length of 2048.
- **No SE**: Remove the SE blocks from each convolution layer.

[2]https://tut-arg.github.io/sed_eval/index.html

TABLE II
THE RESULTS OF SINGING TECHNIQUE DETECTION.

	Macro-F	Micro-F	P	R
PrimaDNN'(Full)	**44.9%**	**60.6%**	43.8%	48.3%
No pitch	39.0%	54.8%	36.6%	47.3%
Single resolution	42.9%	60.2%	44.1%	46.6%
No SE	43.8%	60.3%	43.0%	48.1%
BN	43.9%	59.6%	**44.6%**	48.1%
3x3	44.3%	60.0%	43.2%	**48.8%**

- **BN**: Replace IN with Batch Normalization (BN).
- **No pitch**: Removes mel band pitchgram from input.
- **3x3**: Adopt 3x3 for the kernel size of all convolution layer. (i.e., instead of 5x5 for the first and the second convolution layer.)

The experiments showed that the *full* model outperformed all the modified versions in terms of both *Macro-F* and *Micro-F*.

We further examine the class-wise F-measure and compare it with our previous best model (CRNN-PitchFocal) [1]. As shown in Fig. 4, our model outperforms the previous one in most techniquesd. The main difference between our model and the previous one is the frequency dimension of the input feature, where we adopted a higher resolution of 160. This improvement led to better performance in detecting pitchy techniques such as vibrato', bend', drop', and scooping', indicating that higher frequency resolution better represents fine pitch fluctuation.

We also found that the multi-resolution spectrogram improved the detection of vocal fry' compared to using a single resolution (i.e., with a window length of 2048 only). Vocal fry' has a pulsive modulation pattern as shown at the bottom of Fig. 2. Combining spectrograms with fine temporal resolution helps capture its characteristics. Additionally, instance normalization helped with the detection of 'falsetto'.

The *3x3* condition performed similarly to the *full* model. However, it showed better performance on techniques with shorter duration (e.g., drop' and vocal fry'), but worse performance on techniques with longer duration (e.g., falsetto', rasp', and 'vibrato'), compared to the *full* model. This indicates that the size of the receptive fields affects the detection performance of different techniques.

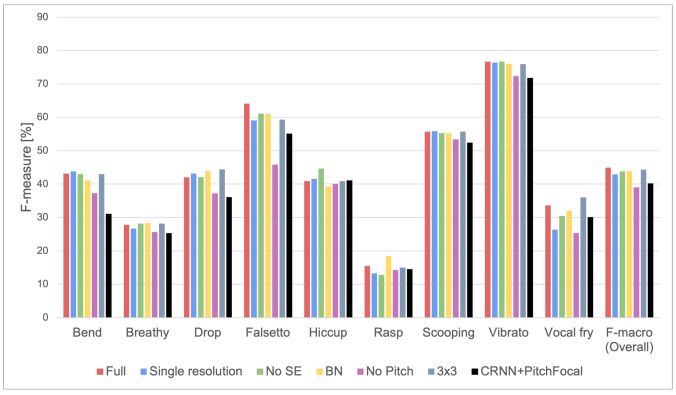

Fig. 4. The technique-wise F-measures for each method in ablation study.

C. Detection examples

In order to investigate the detailed detection performance, we present examples of detections made by CRNN+PitchFocal and PrimaDNN' with reference annotations in Figure 5. The example on the left side of the figure depicts a song with many fine fluctuations and note changes. CRNN+PitchFocal detected many false positives in the vibrato' category at the positions of note transition, whereas PrimaDNN' was able to suppress such false positives. The example on the right side of the figure depicts a song with a slow tempo and mellow mood sung by a female singer. As the figure shows, the section displayed does not have any falsetto', but CRNN+PitchFocal detected them as false positives. In contrast, PrimaDNN' did not detect any 'falsetto' sections as per the reference annotations, indicating that it may be more powerful and robust than CRNN+PitchFocal.

V. CONCLUSION

This paper introduces Prima-DNN', a DNN architecture that takes into account the specific characteristics of singing techniques. It employs multi-resolution mel spectrograms and Mel-band pitchgram for input features, Squeeze-and-Excitation network, and Instance normalization for convolutional layers. The proposed model achieves the best performance on the nine-way detection of singing techniques on the COSIAN dataset. Furthermore, it demonstrates an ability to reduce false negatives for difficult patterns such as those between fast passages and vibrato and non-falsetto singing at high pitch notes and falsetto.

The study [21] suggests that there are certain correlations between the appearance of singing techniques and musical context (e.g., note pitch and duration, phoneme of lyrics, the position of phrase, singer, etc.). Therefore, for future works, it is proposed to combine features related to other musical components such as musical notes, lyrics, and singer information. This could be done through the use of pre-trained features (e.g., Wav2Vec2.0 [22], ECAPA-TDNN speaker embedding [23]) or multi-task learning.

REFERENCES

[1] Y. Yamamoto, J. Nam, and H. Terasawa, "Analysis and detection of singing techniques in repertoires of j-pop solo singers," in *In Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*, 2022, pp. 384–391.

[2] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 34–48, 2018.

[3] R. Nishikimi, E. Nakamura, M. Goto, and K. Yoshii, "Audio-to-score singing transcription based on a crnn-hsmm hybrid model," *APSIPA Transactions on Signal and Information Processing*, vol. 10, p. e7, 2021.

[4] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 2392–2396.

[5] Y. Yamamoto, J. Nam, H. Terasawa, and Y. Hiraga, "Investigating time-frequency representations for audio feature extraction in singing technique classification," in *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021.

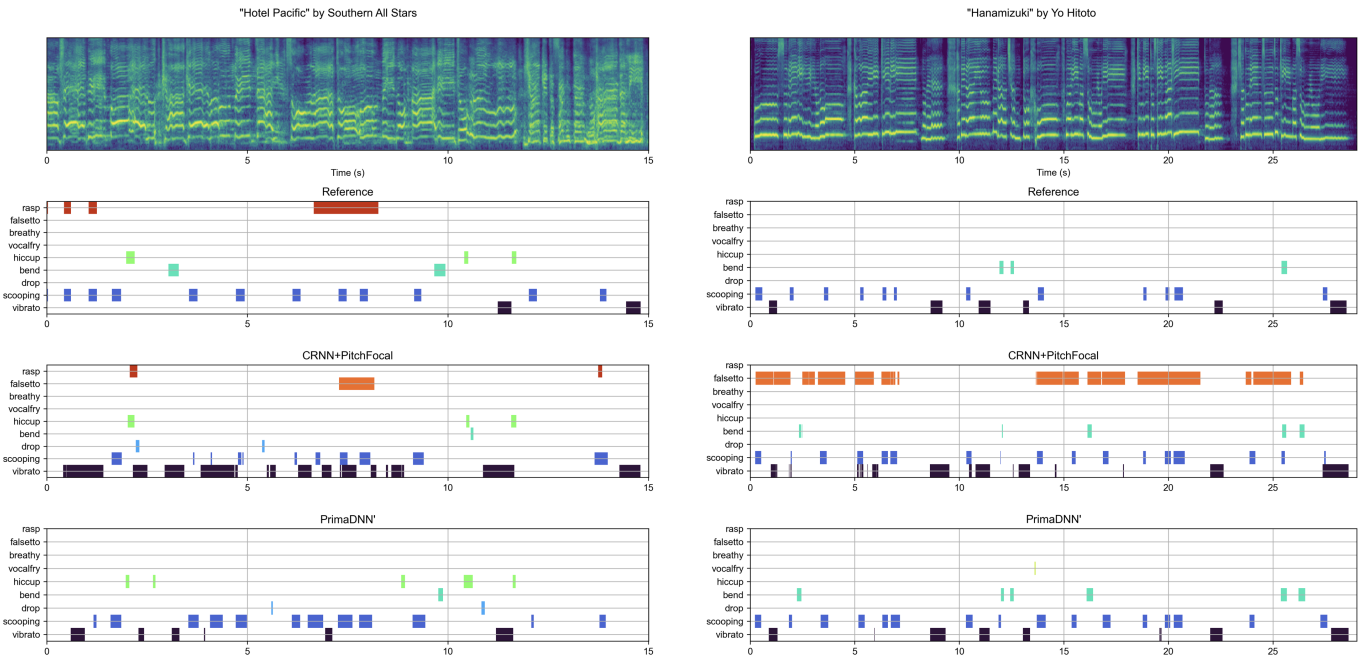

Fig. 5. Detection examples. From above, spectrogram, reference annotation, estimation of CRNN+PitchFocal [1], and estimation of PrimaDNN' in each row. (Left) Comparison on "Hotel Pacific" by Southern All Stars. (Right) Comparison on "Hanamizuki" by Yo Hitoto.

[6] T.-H. Hsieh, K.-H. Cheng, Z.-C. Fan, Y.-C. Yang, and Y.-H. Yang, "Addressing the confounds of accompaniments in singer identification," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1–5.

[7] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, pp. 161–165.

[8] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[9] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[10] X. Pan, P. Luo, J. Shi, and X. Tang, "Two at once: Enhancing learning and generalization capacities via ibn-net," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 464–479.

[11] C. Zhang and L. Xue, "Autoencoder with emotion embedding for speech emotion recognition," *IEEE access*, vol. 9, pp. 51 231–51 241, 2021.

[12] J. chieh Chou and H.-Y. Lee, "One-Shot Voice Conversion by Separating Speaker and Content Representations with Instance Normalization," in *Proc. Interspeech 2019*, 2019, pp. 664–668. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2663

[13] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2017, pp. 2980–2988.

[14] A. Défossez, "Hybrid spectrogram and waveform source separation," in *In Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)*, 2021.

[15] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.

[16] B. T. Atmaja and M. Akagi, "On the differences between song and speech emotion recognition: Effect of feature sets, feature types, and classifiers," in *2020 IEEE REGION 10 CONFERENCE (TENCON)*. IEEE, 2020, pp. 968–972.

[17] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE transactions on affective computing*, vol. 7, no. 2, pp. 190–202, 2015.

[18] K. Imoto, S. Mishima, Y. Arai, and R. Kondo, "Impact of sound duration and inactive frames on sound event detection performance," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2021, pp. 860–864.

[19] Q. Kong, Y. Xu, W. Wang, and M. D. Plumbley, "Sound event detection of weakly labelled data with cnn-transformer and automatic threshold optimization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2450–2460, 2020.

[20] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," in *International Conference on Learning Representations*, 2019.

[21] Y. Yamamoto, T. Nakano, M. Goto, H. Terasawa, and Y. Hiraga, "Analysis of frequency, acoustic characteristics, and occurrence location of singing techniques using imitated j-pop singing voice," *The Special Interest Group Technical Report of IPSJ (MUS)*, no. 20, pp. 1–8, 2021, (in Japanese).

[22] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[23] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.