Italian Speech Emotion Recognition

Irene Mantegazza and Stavros Ntalampiras
Department of Computer Science, University of Milan
irene.mantegazza@studenti.unimi.it, stavros.ntalampiras@unimi.it

2023 24th International Conference on Digital Signal Processing (DSP) | 979-8-3503-3959-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/DSP58604.2023.10167766

***Abstract*—Affective computing is gaining increased interest by the scientific community in the last decades with the acoustic modality playing a central role. This paper presents an extensive computational analysis of emotional speech focusing on the Italian language. More precisely, we propose a novel classification algorithm based on a suitable data augmentation scheme. The aim is to classify the seven emotions (*anger*, *disgust*, *fear*, *joy*, *neutral*, *sadness*, and *surprise*) included in the only publicly available database of Italian emotional speech, i.e. EMOVO. To this end, we employed two feature sets, Mel Frequency Cepstral Coefficients and log-Mel spectrogram, each one combined with a suitable classifier, i.e. Mutilayer perceptron and Convolutional neural network respectively. The implementation and evaluation of the proposed SER pipeline can be accessed through the following link: https://github.com/irenemante/ser_emovo**

***Index Terms*—Affective computing, Convolutional neural network, Multilayer perceptron, data augmentation, MFCCs, log-Mel spectrogram.**

I. INTRODUCTION

Speech comprises one of the most natural ways to express ourselves [1]. As such, researchers have extensively considered speech as a fast and efficient method in human computer interaction. Indeed, during the last decades, there has been a considerable amount of research on speech processing concentrated on paralinguistic aspects [2]. Despite the great progress made in speech recognition, we are still far from having natural interactions between humans and machine due to the poor automated understanding of the speaker's emotional state. The field of Speech Emotion Recognition (SER), which is defined as the process of extracting the emotional state of a speaker from her/his speech. SER can be applied to a wide range of applications such as computer tutorials applications, car assistants, where it can be adopted to recognize the mental status of the driver, automatic translation systems, dialogue systems for spoken languages such as call center conversations and medical applications, in which is employed as a diagnostic tool for therapists [3]. SER is a particularly challenging task for different reasons: 1) diversity in sentences, speakers, languages, speaking style and rate in the datasets complicates the identification of suitable feature sets 2) the expression of emotions depends on the cultural background of the speaker and this is why it is preferred to use datasets in which speech is recorded by people with the same cultural background.

An important prerequisite of SER is to establish the set of emotions to be classified. Existing research has determined the following list of distinct emotions: Anger, Disgust, Fear, Joy, Sadness, and Surprise [4]. They are called Archetypal emotions and the vast majority of SER datasets are based on them [5]. Indeed, a large amount of datasets has been created to serve emotion classification tasks covering covers the most common languages in the world, as English, Chinese, Arabic, Japanese, Spanish, Portuguese, French, Italian, Greek, Russian, etc. [1] [6].

This article focuses on the Italian language which has received limited attention by the scientific community. To this end, we selected the database EMOVO which includes speech coming from 3 actors and 3 actresses speaking 14 sentences simulating the 6 Archetypal emotions plus the neutral state [7]. The implemented audio pattern recognition pipeline includes feature extraction and pattern recognition components, while we experimented with both traditional and deep learning approaches. First, log-Mel spectrograms are extracted from the audio files of EMOVO and fed to a Convolutional neural network. The second approach uses Mel Frequency Cepstral Coefficients (MFCCS) along with their deltas as inputs for a Multilayer Perceptron. At the same time, we experimented we two data augmentation schemes, namely pitch shifting and noise addition, which were able to improve the generalization capabilities of the deep model. Importantly, to the best of our knowledge, this constitutes the best performance reported in the literature regarding Italian SER.

The rest of this work is organized as follows: section II explains the feature extraction and pattern classification modules, while III presents the dataset and the adopted experimental protocol. Section IV analyzes the obtained experimental results and we draw our conclusions in section V.

II. THE EXPLORED METHODS

Fig. 1 illustrates the pipeline of the proposed architecture. We start from data pre-processing, which includes DC-offset removal. Subsequently, we introduce a data augmentation block as the size of the available dataset is relatively limited. To this end we increase the dataset by injecting diverse modifications in the original data. The considered modifications include pitch shifting and addition of white noise. As such, the model may learn from multiple variants of the available emotional speech which may improve its generalization capabilities [8].

In the next step, the MFCCs along with their deltas and log-Mel spectrograms are extracted from the original and augmented signals. Zero-padding was employed to uniform

979-8-3503-3959-8/23/$31.00 ©2023 IEEE

[1] https://superkogito.github.io/SER-datasets/

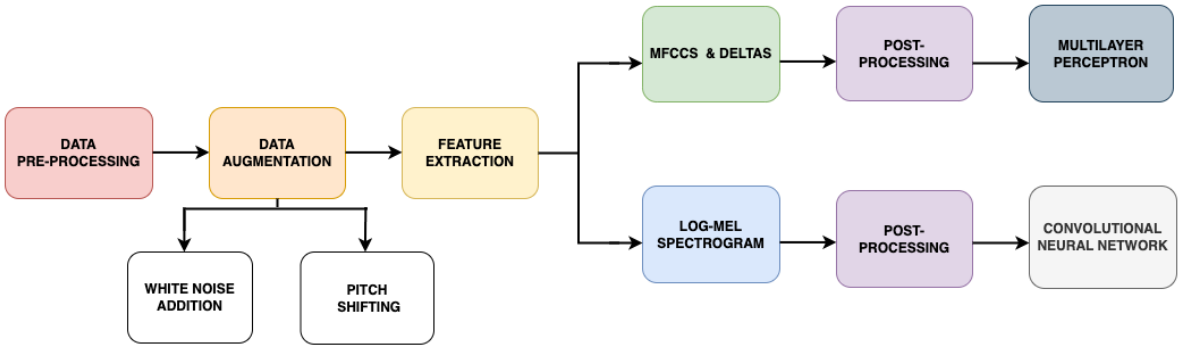


Fig. 1: The proposed pipeline starting from data pre-processing, augmentation, feature extraction and ending with the considered classification models.

the durations where necessary. After z-score normalization, the last phase consists in the creation of the machine learning models to carry out SER.

A. Feature extraction

This subsection describes the feature sets employed to characterize the available manifestations of emotional speech.

1) Mel-frequency Cepstral Coefficients: Mel-frequency cepstral coefficients (MFCCs) are a feature widely used in automatic speech and speaker recognition. For their derivation, the signal is split into overlapping frames from where the power of the Short Time Fourier Transform is computed. The obtained representation is filtered through a Mel scale filterbank to highlight the signal bands that are important for human perception. The last stages include log-scaling and decorrelation through the Discrete Cosine Transform (DCT) [9]. Finally the most significant 13 coefficients along with their deltas are kept.

As regards to the parameterization, sampling rate of the signals is 22.050 Hz, the number of coefficients to extract 13, the length of the FFT window 2048 and the hop length 512. The result is a multi-dimensional array composed of 26 elements (13 MFCCs coefficients and 13 deltas coefficients) whose length is 130, which corresponds to the resulting number of frames.

2) Log-Mel spectrogram: To compute the log-Mel spectrogram we used the same procedure described above for the MFCCs, with the only difference being the omission of the Discrete Cosine Transform. Compared to MFCCs, it does not destroy spatial relations and as such, it is more suitable to spatially local models, like CNN. The resulting feature set is a multi-dimensional array composed of 60 elements with dimension of 130, similarly to the MFCCs. Fig. 2 illustrates log-mel spectrograms representing all available emotional states.

B. Data augmentation

Data augmentation is a process of artificially increasing the amount of data by generating new data points from existing data. Typically, this introduces minor alterations to the data or employs machine learning models to generate new data points in the original space [10], [11]. In this work, the proposed data augmentation techniques are *white noise addition* and the *pitch shifting*. The first one injectes white noise fo the signal in improve the robustness and generalization, while during latter one, the pitch of the audio sample is raised or lowered, keeping the duration unchanged. Motivated by the encouraging results in different audio pattern recognition tasks [12] it was carried out in two diverse ways, i.e. using both a positive value (2) and a negative value (-2) regulating the pitch alteration. The positive value raises the pitch, while the negative one lowers it.

C. Classification Models

This section explains briefly the models considered in this work including both shallow and deep neural networks.

1) Multilayer Perceptron (MLP): It is an artificial neural network (ANN) encompassing multiple perceptrons. In this work, MLP used as inputs the MFCCs and the respective deltas. The proposed architecture for the MLP consists of one initial flatten layer reducing the three-dimensionality of the dataset followed by three dense layers with a decreasing number of neurons. Each dense layer is followed by a dropout one which sets input units to 0 with the frequency of 0.3 at each step of the training time in order to avoid overfitting. During early experimentations, it was evident that the presence of dropout layers did not limit significantly the overfitting phenomenon, early stopping was employed stopping the training process when the loss continues to increase for 10 subsequent epochs. At the same time, we applied L2 regularization to the weights of the dense layers with a coefficient equal to 0.001, thus adding penalties on layer parameters.

We used Rectified Linear Unit as activation function in order to take into account the non-linearity of the output of each layer, while the last dense layer employed the Softmax function, which produces probabilities with respect to all classes.

2) Convolutional neural network (CNN): This deep neural network is a stack of multi-layer neural networks including a group of convolutional layers, pooling layers and, typically, a limited number of fully connected layers. Since CNNs have shown exceptional abilities in various image processing tasks [13] log-Mel spectrogram representation was chosen as input

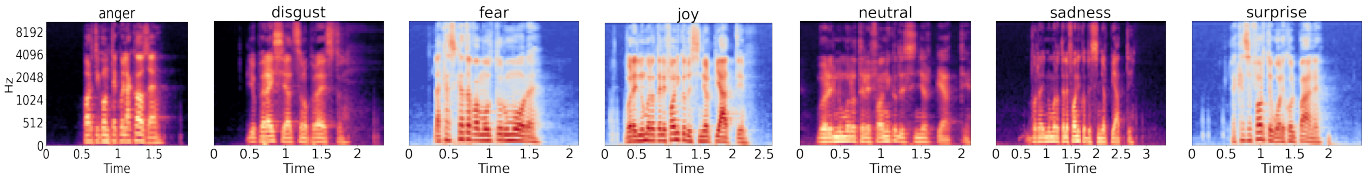
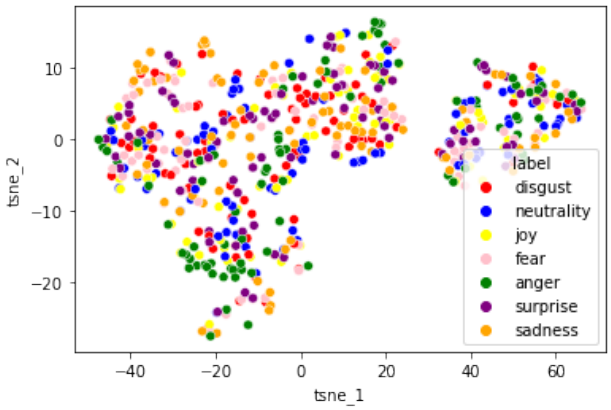
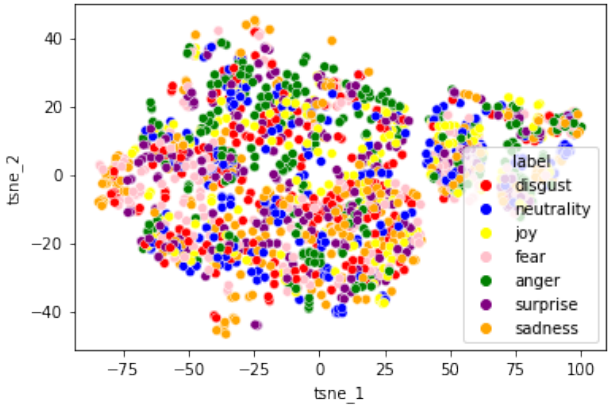


Fig. 2: Log-Mel spectrograms extracted from samples belonging to all classes available in the EMOVO dataset.



(a) before augmentation



(b) after augmentation

Fig. 3: Feature space visualization before (top) and after (bottom) the data augmentation process.

of this model type [14]. The considered CNN architecture is composed of four 2-D convolutional layers, each of which is followed by a maxPooling layer, responsible for the reduction of the convolutional features maps dimension, and a dropout layer. The convolutional layers have a kernel size of 2x6 and an increasing number of filters, while the kernel size for maxPooling layers is 2x7. The resulting features maps are then flattened, through global average pooling layer, into a one-dimensional array, which comprises the input of a Fully connected neural network driving the final classification decision. The Fully connected neural network consists of two dense layers, followed by a dropout layer. As in Multilayer perceptron, the dropout layer, the regularizer and the early stopping (considering 10, 15 or 20 epochs) were employed in order to avoid overfitting.

III. Experimental Set-Up

This section presents the employed dataset including feature space visualization along with the experimental protocol.

A. Dataset and pre-processing

EMOVO dataset is composed of 588 audio files, the duration of which varies from 2 to 11 seconds. Each file contains a sentence manifested by an actor expressing a specific emotion. This dataset is characterized by a strong class balance because the considered emotions (*fear*, *sadness*, *joy*, *neutral*, *anger*, *disgust*, *surprise*) are represented with the same number of audio files. As mentioned in the feature extraction section, a multi-dimensional array is created and, for model learning and evaluation, it is important that the sequences representing an audio file have equal dimensionality. To this end, in order to ensure the same length of the arrays, it was necessary to reduce to 3 seconds the duration of the audio files that are longer than 3 seconds and to apply the zero-padding to the extracted features of the files that are shorter than 3 seconds. During the post-processing stage, the extracted features are standardized such that their distribution has a mean value of 0 and standard deviation of 1 [15].

B. Feature space visualization

In order to explore the feature space with the true distribution before and after the augmentation, we extracted from each audio file an audio embedding using OpenL3 [16], an open-source deep model based on the self-supervised L3-Net [17]. Among other tasks, OpenL3 has been employed for SER. The embedding, generated through OpenL3, is a multidimensional array whose dimensionality is directly proportional to the duration of the audio files. To guarantee the same dimensionality for all the embeddings, we applied the zero padding on the time series of the files shorter than 3 seconds and we reduced the length of longer files to 3 seconds. For the data visualization, t-SNE was adopted in order to reduce the embedding dimensionality to 2 features. Fig. 3 illustrates the feature space before and after data augmentation, where we can observe that the specific problem is highly non-linear with a substantial overlap between the available classes.

Table I: Recognition rates (in %) averaged across the considered emotional classes with respect to the proposed pipeline (ps: pitch shifting, na: noise addition). The highest rate is emboldened.

Dataset	CNN	MLP
Without augmentation	43.2±0.9	39.6±1.8
Augmented with ps	91.1±1.1	73.5±0.1.6
Augmented with ps&na	**96.5±0.6**	74.9±1.5

C. Experimental Protocol

For the evaluation of the proposed models, we adopted the stratified 10-fold cross-validation protocol. Stratified k-fold cross-validation is an extension of the k-fold cross-validation having the advantage of preserving the percentage of samples for each class [18]. Each experimental setting (both classifiers combined with both augmentation schemes) was iterated 10 times and we report average and standard deviation values.

IV. Experimental Results

Different experiments were realized to compare the performances of the two models and to evaluate how the data augmentation affects the models' performance. The two models were tested on the following cases: a) original data, b) original data augmented with pitch scaling, and c) original data augmented with pitch scaling and white noise addition. As a figure of merits, we employed average recognition rate and confusion matrix. Table I includes the rates averaged across the considered emotional classes with respect to the proposed pipeline and augmentation schemes. Figures 5 and 4 demonstrate the confusions matrices for the best performing MLP and CNN models respectively.

In general, the CNN outperforms MLP in every conducted experiment; in fact, if we consider the data augmented through the two techniques, the level of accuracy of CNN reaches the 96.5%, while MLP reaches 74.9%. In addition, it emerged that the obtained outcomes without data augmentation are not satisfactory. The accuracy stops at 43.2% with CNN and 39.6% with MLP while in both the cases there is strong evidence of overfitting (when examining how the train and test accuracy and loss change over time during model learning) and considerable standard deviation values. This may be due to the limited size of EMOVO which might be insufficient to represent all seven emotional states during training.

On the contrary, when we consider the data augmented through the two *pitch scaling* techniques, the results are satisfying. It occurs because of the higher availability of training data across all considered classes. The subsequent application of *white noise addition* tends to eliminate overfitting in both models thanks to its ability to improve the generalization capability of the model. Also, the accuracy of the models tends to increase, especially for CNN as we can see in the Table I. Overall, *sadness* and *anger* resulted to be the most

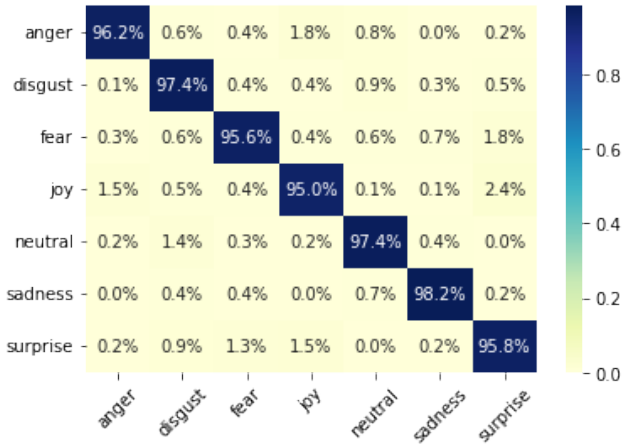


Fig. 4: Confusion matrix obtained using the CNN applied to the data augmented with pitch scaling and white noise addition. Rows: ground truth, columns: prediction.

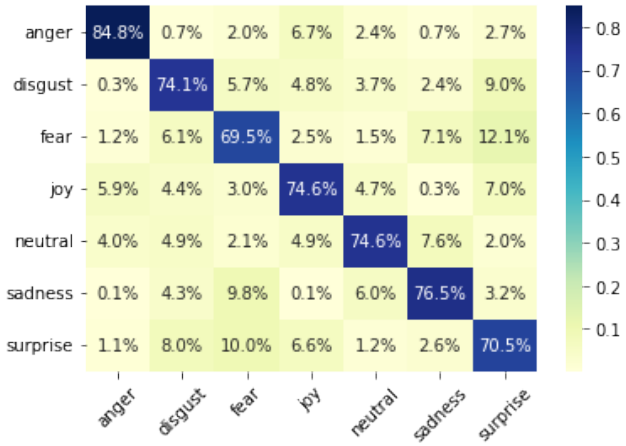


Fig. 5: Confusion matrix obtained using the MLP with input data augmented with pitch scaling and white noise addition. Rows: ground truth, columns: prediction.

correctly classified emotions respectively by CNN and MLP, as we can observe in Figures 5 and 4. We conclude that model learning based on the augmented dataset is able to provide almost excellent results in Italian SER.

V. Conclusions

This work thoroughly presented a pipeline for efficient Italian SER. After comparing of two models, it resulted that the Convolutional neural network is more effective than the Multilayer Perceptron in classifying seven emotional states included in a publicly available dataset. It is important to underline the role that data augmentation has had in improving the classification of emotional states and ensuring improved recognition rates with respect to every class.

In the future, we are going to a) assess the way data augmentation influences SER in language-agnostic settings [19], b) experiment with few-shot learning approaches [20], and c) work with embeddings of diverse models at the same time [21].

REFERENCES

[1] Y. Li, P. Bell, and C. Lai, "Fusing asr outputs in joint training for speech emotion recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7362–7366.

[2] S. Ntalampiras, "Speech emotion recognition via learning analogies," *Pattern Recognition Letters*, vol. 144, pp. 21–26, Apr. 2021. [Online]. Available: https://doi.org/10.1016/j.patrec.2021.01.018

[3] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. W. Schuller, "Survey of deep representation learning for speech emotion recognition," *IEEE Transactions on Affective Computing*, pp. 1–1, 2021.

[4] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320310004619

[5] P. Chandrasekar, S. Chapaneri, and D. Jayaswal, "Automatic speech emotion recognition: A survey," in *2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA)*, 2014, pp. 341–346.

[6] M. Nicolini and S. Ntalampiras, "A hierarchical approach for multilingual speech emotion recognition," in *Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods*. SCITEPRESS - Science and Technology Publications, 2023. [Online]. Available: https://doi.org/10.5220/0011714800003411

[7] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco, "EMOVO corpus: an Italian emotional speech database," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014, pp. 3501–3504.

[8] D. Tompkins, K. Kumar, and J. Wu, "Maximizing audio event detection model performance on small datasets through knowledge transfer, data augmentation, and pretraining: an ablation study," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1016–1020.

[9] S. Ntalampiras, "Model ensemble for predicting heart and respiration rate from speech," *IEEE Internet Computing*, pp. 1–7, 2023.

[10] T. Iqbal, K. Helwani, A. Krishnaswamy, and W. Wang, "Enhancing audio augmentation methods with consistency learning," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 646–650.

[11] Q. Wang, J. Du, H.-X. Wu, J. Pan, F. Ma, and C.-H. Lee, "A four-stage data augmentation approach to resnet-conformer based acoustic modeling for sound event localization and detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–13, 2023.

[12] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[13] W. Zeng, W. Li, M. Zhang, H. Wang, M. Lv, Y. Yang, and R. Tao, "Microscopic hyperspectral image classification based on fusion transformer with parallel cnn," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–12, 2023.

[14] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206–219, 2019.

[15] S. Ntalampiras, "Deep learning of attitude in children's emotional speech," in *2020 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, 2020, pp. 1–5.

[16] A. L. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3852–3856.

[17] W. Lu, Y. Zhou, G. Wan, S. Hou, and S. Song, "L3-net: Towards learning based lidar localization for autonomous driving," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6382–6391.

[18] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013. [Online]. Available: https://faculty.marshall.usc.edu/gareth-james/ISL/

[19] S. Ntalampiras, "Toward language-agnostic speech emotion recognition," *Journal of the Audio Engineering Society*, vol. 68, no. 1/2, pp. 7–13, Feb. 2020. [Online]. Available: https://doi.org/10.17743/jaes.2019.0045

[20] ——, "One-shot learning for acoustic diagnosis of industrial machines," *Expert Systems with Applications*, vol. 178, p. 114984, Sep. 2021. [Online]. Available: https://doi.org/10.1016/j.eswa.2021.114984

[21] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Composing general audio representation by fusing multilayer features of a pre-trained model," in *2022 30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 200–204.