Deep scattering network for speech emotion recognition

Premjeet Singh, Goutam Saha, Md Sahidullah

Deep scattering network for speech emotion recognition

Premjeet Singh[1], Goutam Saha[1], Md Sahidullah[2]
[1]*Dept of Electronics and ECE, Indian Institute of Technology Kharagpur, Kharagpur, India*
[2]*Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France*
premsingh@iitkgp.ac.in, gsaha@ece.iitkgp.ac.in, md.sahidullah@inria.fr

***Abstract*—This paper introduces scattering transform for speech emotion recognition (SER). Scattering transform generates feature representations which remain stable to deformations and shifting in time and frequency without much loss of information. In speech, the emotion cues are spread across time and localised in frequency. The time and frequency invariance characteristic of scattering coefficients provides a representation robust against emotion irrelevant variations e.g., different speakers, language, gender etc. while preserving the variations caused by emotion cues. Hence, such a representation captures the emotion information more efficiently from speech. We perform experiments to compare scattering coefficients with standard mel-frequency cepstral coefficients (MFCCs) over different databases. It is observed that frequency scattering performs better than time-domain scattering and MFCCs. We also investigate layer-wise scattering coefficients to analyse the importance of time shift and deformation stable scalogram and modulation spectrum coefficients for SER. We observe that layer-wise coefficients taken independently also perform better than MFCCs.**

***Index Terms*—Deep convolutional networks, Deep scattering transform, EmoDB, IEMOCAP, RAVDESS, Shift invariance, Speech emotion recognition**

I. Introduction

The introduction of deep learning has caused rapid development in different speech signal processing tasks. One such domain is *speech emotion recognition* (SER) which finds applications in health-care systems, sentiment analysis, and many other human-computer interaction applications [1]–[3]. Even after two decades of research, SER is still considered as a challenging task [2]. One of the major challenges is the interpersonal and intrapersonal variability in emotion expression, e.g., different speaking styles, cultural background, language, context, speaker's mood etc [1], [3]. This causes SER systems to face difficulty in generalization across different emotion speech samples, especially when the train and test data samples belong to different domains [4]. This downside motivates the search for a robust feature that can extract emotion-specific information irrespective of different variabilities.

The deep learning based SER approaches generally addresses the issue of variabilities by using deep convolutional neural networks (CNN), which are considered efficient feature extractors in terms of invariances learned against signal shifts [5], [6]. However, the complexity and *lack of explainability* of deep CNNs come as an added disadvantage to automatic feature extraction ability [7].

To obtain an end-to-end framework, some SER studies use 1D CNNs to learn the relevant features directly from raw speech [8]. Such studies observe that deep networks learn representations similar to handcrafted features [9]. However, due to inherent instability and requirement of large data samples for training, suitability of deep neural networks becomes uncertain [10].

Following similar lines, we propose the use of *scattering transform* for SER. Scattering transform was introduced in [11], [12] as a deep convolutional network involving convolution with *predefined kernel* instead of automatic kernel learning. It also introduces stability to both temporal shift and deformations in the feature representation of signals. As emotion cues spread temporally in speech, such characteristics of scattering transform should provide a representation robust to both time-shifting of cues and emotion irrelevant temporal variations. In [13], authors compute scattering coefficients across the log-frequency domain for frequency transposition invariance to obtain speaker-independent representation for speech recognition. Such characteristics may also help in learning speaker-independent emotion cues.

Scattering transform is used in both 1-D and 2-D data processing. Regarding 1-D signals, authors in [14] introduce joint time-frequency scattering that incorporates multiscale frequency energy distribution over time-invariant representation for various audio classification frameworks. Authors in [15] use two-layer scattering coefficients with CNN layers to obtain a stable descriptor of speaker information from raw speech. In [16], authors compute different moments of scattering coefficients layers and found that such moments contain enough information for decent voice synthesis quality. Authors in [17] used scattering transform for urban sound classification and report a marginal performance improvement but with reduced training data. They attribute their finding to better background information learning ability of temporal modulations. Along similar lines, [18] also applies to scatter coefficients for environmental sound classification. In [19] a joint time-frequency based scattering representation is used to analyze the timbral similarity between different acoustic instruments. Our work is the first to apply scattering transform for SER. We conduct experiments on three different SER datasets using time and frequency scattering coefficients and achieve noticeable improvement over standard mel-frequency cepstral coefficients (MFCCs). The main contributions of this

work are summarized below:

1) Optimization of scattering transform parameters and analysing its performance for SER,
2) Analysis of time-domain and frequency-domain scattering coefficients over different databases, and
3) Layer-wise analysis of scattering coefficients for SER.

II. Scattering Transform

For 1D signals, the idea behind scattering transform is to obtain a robust feature that remains invariant to the location of information cues (translation invariance) and time warping (diffeomorphism) of information across different instances of the signal.

A. Notion of Lipschitz continuity and information loss in MFCC

Let the signal $x(t)$ be translated by c. Then we have, $x_c(t) = x(t-c)$. Taking its Fourier transform, we obtain, $X_c(\omega) = e^{-j\omega c}X(\omega)$. Taking modulus on both sides, we obtain, $|X_c(\omega)| = |X(\omega)|$. In short time Fourier transform (STFT), translation c is localized to the time frame duration T. If $|c| \ll T$, the STFT representation is already invariant to such translation. However, a robust feature representation also requires stability to time deformations appearing in the signal. Let $x(t)$ is warped (deformed) in time by a factor τ, i.e., $x_\tau(t) = x(t-\tau(t))$. A transformation ϕ of $x(t)$ is said to be stable to deformation τ if, $||\phi(x) - \phi(x_\tau)|| \leq C sup_t|\tau^{'}(t)|||x||$, i.e., the distance between the non-deformed ($\phi(x)$) and deformed ($\phi(x_\tau)$) feature in transformation space ϕ should be less than a factor C of maximum amplitude/size of deformation ($sup_t|\tau^{'}(t)|$). This is also known as Lipschitz continuity condition. Applying this to Fourier transform, we have, $|||X(\omega)| - |X_\tau(\omega)||| \leq C\epsilon||x||$, where, time deformation $\tau(t) = \epsilon t$. Now the Fourier transform of $x_\tau(t) = x(t-\epsilon t)$ is given as $X_\tau(\omega) = \frac{1}{1-\epsilon}X(\frac{\omega}{1-\epsilon})$. Hence, the frequency components at ω are shifted to $\frac{1}{1-\omega}$. This effect is more prominent at high frequencies. Hence, modulus of Fourier transform is not Lipschitz continuous. In MFCC, filters with non-linearly varying bandwidths are applied over the STFT. As output of every mel-filter is averaged to get a single coefficient, filter application can be viewed as averaging performed over frequency domain which counters the effect of instability. The mel-filters applied around high frequency regions have higher bandwidth which reduces the effect of deformation instability. However, this averaging in frequency also leads to information loss, especially at higher frequencies. Now, consider $x_t(u)$ to be the signal frame at time t. Then $x_t(u) = x(u)\phi(u-t)$ where ϕ is a window of length T. The application of mel-filters over Fourier transform of $x_t(u)$, can be written in frequency as,

$$Mx(t,\lambda) = \frac{1}{2\pi}\int |x_t(\omega)|^2|\psi_\lambda(\omega)|^2 d\omega \tag{1}$$

where $x_t(\omega)$ is the frequency response of $x_t(u)$ and $\psi_\lambda(\omega)$ is the mel-filter with support λ. Since multiplication in frequency becomes convolution in time, Eq. 1 becomes,

$$\begin{aligned} Mx(t,\lambda) &= \int |x_t * \psi_\lambda(v)|^2 dv \\ &= \int \left| \int x(u)\phi(u-t)\psi_\lambda(v-u)du \right|^2 dv \end{aligned} \tag{2}$$

If T is much greater than the filter support λ in time, $\phi(t)$ is constant over λ. Hence we can write, $\phi(u-t)\psi_\lambda(v-u) \approx \phi(v-t)\psi_\lambda(v-u)$, which gives,

$$\begin{aligned} M_x(t,\lambda) &\approx \int \left| \int x(u)\psi_\lambda(v-u)du \right|^2 |\phi(v-t)|^2 dv \\ &= |x * \psi_\lambda|^2 * |\phi|^2(t) \end{aligned} \tag{3}$$

Eq. 3 shows that averaging in frequency domain is equivalent to time-domain averaging of $|x * \psi_\lambda|^2$ with window duration T and becomes the basis for scattering transform. Hence, $M_x(t,\lambda)$ remains invariant to time shifts smaller than T. In MFCC, T generally equals 20ms which limits the capturing of long time scale information. If T is increased, loss of high frequency information takes place due to averaging [13].

B. Layer-wise scattering coefficients

To prevent the information loss appearing in MFCC while, at the same time, capturing high frequency information, scattering transform computes second layer coefficients or the modulation spectrogram of signal. This includes computing the frequency response of the time-series of scalogram frequency bins. This is followed by averaging of both scalogram and modulation spectrum coefficients for stability against deformation. Scattering transform uses constant quality factor wavelet filters over complete frequency range to compute scalogram and modulation spectrogram. If Q defines the number of wavelets per octave, the center frequencies of wavelets becomes $\lambda = 2^{\frac{k}{Q}}$, where k is an integer and the bandwidth is proportional to $1/Q$. Every wavelet has a support of bandwidth λ/Q around center frequency λ. The same wavelet in time provide a spread of $\frac{2\pi Q}{\lambda}$. To make sure that this spread is less than T (because T is the time window over which time averaging will be computed for stable feature representation) wavelets are only defined for $\lambda \geq \frac{2\pi Q}{T}$. For frequencies less than $\frac{2\pi Q}{T}$ the frequency interval is covered with wavelets of equal bandwidth given by $\frac{2\pi}{T}$, hence covering complete frequency range. The mel-filter operation and filtering due to time window (T) in MFCC is replaced with wavelet-based filters in scattering transform.

An issue with using wavelets is that they "commute with translation" [20]. However, invariance to translation can generally be introduced by averaging. Following this, $\int x(t) * \psi_\lambda(t)dt$ should be translation invariant. But due to symmetric nature of wavelet, $\int x(t) * \psi_\lambda(t)dt = 0$. Hence, a modulus nonlinear operator is applied to prevent this vanishing of integral, i.e. $\int |x(t) * \psi_\lambda(t)|dt$. The reason behind choosing modulus

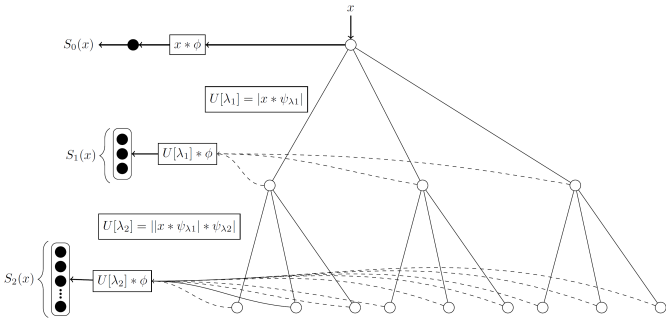

Fig. 1. Graphical representation of scattering transform. Here, x is the input signal and ϕ is the low-pass filter used for stability to deformations. ψ_1 and ψ_2 are wavelet filter banks corresponding to 1st and 2nd layer of scattering transform. The figure describes only 2 layer decomposition of signal x.

non-linearity is because modulus operation is contractive and preserves the norm of the vector to which it is applied [14]. The modulus of Fourier transform coefficients are then low-pass filtered with filter ϕ to obtain feature representation robust against time-warping induced deformation instability. These features are then invariant to translations smaller than 2^J, where J is the scale of the low-pass wavelet filter (ϕ). Therefore, any mth order scattering coefficients are given as,

$$S_{J_m}x(u) = U_m x * \phi_{2^J}(u) = \int U_m x(v)\phi_{2^J}(u-v)dv \quad (4)$$

where,

$$U_m x = U[\lambda_m]...U[\lambda_2]U[\lambda_1]x = |||x * \psi_{\lambda 1}| * \psi_{\lambda 2}|...| * \psi_{\lambda m}|$$

To obtain invariance to frequency transpositions, wavelet transform is first computed over the log-frequency axis (e.g., $log(\lambda_1)$) of time-domain scattering coefficients followed by wavelet averaging. This is similar to the DCT operation in MFCC [13]. Such averaging introduces invariance to shifts in frequency defined by the scale of the averaging wavelet. The frequency scattering coefficients are cascaded with time scattering coefficients to generate a feature representation of the signal.

III. Experimental Details

A. Dataset description

Three different speech corpora used in the experiments are described in Table I. The speech samples are downsampled at 16 kHz sampling frequency when required. EmoDB database is a German-language database whereas, both IEMOCAP and RAVDESS contain speech samples in the English language. For IEMOCAP, we use only four emotions (Happy, Angry, Sad, and Neutral) following other works with this dataset [4], [21].

B. Experimental evaluation & Methodology

We first optimize the scattering transform parameters over the EmoDB database. The optimized parameter set is then used for SER evaluation over other databases. We perform only two-level decomposition as higher level coefficients contain a very small fraction of signal energy [14]. We optimize the number of *wavelets per octave* (Q) and *maximal wavelet length* or *averaging scale* (T) while using an empirically chosen fixed value of input signal duration (N). For frequency scattering, we select the maximal wavelet length as 32. We choose this to obtain invariance over five octaves, found optimum for EmoDB database. Also, following [13], for the first layer time-scattering coefficients, we compute only the frequency-domain wavelet decomposition. The required optimum frequency averaging is computed and applied by the classifier itself [13]. We use Morlet wavelet for both time and frequency scattering coefficients computation [13]. For implementation, we use the ScatNet toolkit[1].

TABLE I
Summary of the speech corpora used in the experiments.
(F=Female, M=Male)

Databases	Speakers	Emotions
Berlin Emotion Database (EmoDB) [22]	10 (5 F, 5 M)	7 (Anger, Sad, Boredom, Fear, Happy, Disgust and Neutral)
Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [23]	24 (12 F, 12 M)	8 (Calm, Happy, Sad, Angry, Neutral, Fearful, Surprise, and Disgust)
Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) [24]	10 (5 F, 5 M)	4 (Happy, Angry, Sad and Neutral)

For classification purposes, we use a radial basis function (RBF) kernel-based *support vector machine (SVM)*, as the feature representations obtained are utterance-level representations. We use the leave-one-speaker-out (LOSO) cross-validation strategy and report average performance obtained over different train-test pairs. For every experiment, we keep speech samples of one speaker in validation, one in test and remaining in training set. We optimize the hyperparameters of the SVM classifier on the validation set.

For comparison, we also compute MFCCs for speech utterances of duration N and evaluate the performance with the same LOSO cross-validation strategy. We choose the MFCC feature because of its characteristic of mimicking human sound perception and its versatility in various speech processing domains. For performance analysis, we use both *accuracy* and *unweighted average recall* (UAR) performance metrics. The UAR is defined as,

$$\text{UAR} = \frac{1}{K}\sum_{i=1}^{K}\frac{A_{ii}}{\sum_{j=1}^{K}A_{ij}}. \quad (5)$$

Here A is the contingency matrix, A_{ij} is the number of samples in class i classified into class j and K is the number of classes. UAR is better suited as a performance metric in class imbalance situations as compared to standard accuracy [25].

IV. Results

A. Parameter optimization over EmoDB

Figure 2 shows the results obtained by varying Q and T parameters of scattering transform over EmoDB database.

[1]https://github.com/scatnet/scatnet

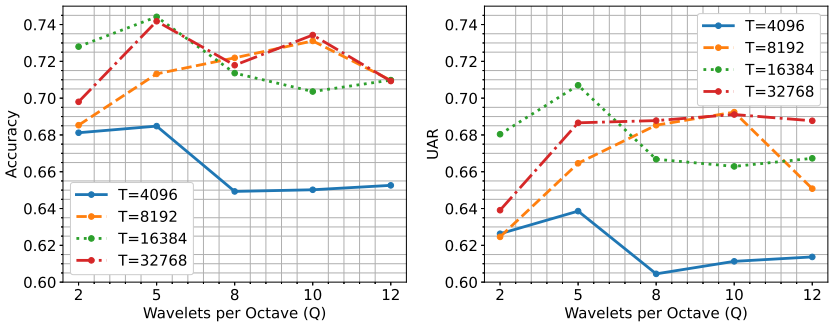

Fig. 2. Comparison of various time-domain scattering transform parameters over EmoDB database. Q is number of wavelets per octaves and T is the averaging scale in samples. Improved performance over lower values of Q shows the importance of coarse level frequency-domain information for SER.

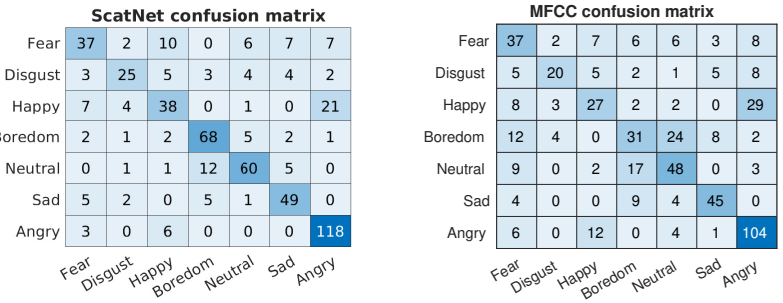

Fig. 3. Confusion Matrices of time-domain scattering coefficients and MFCC features over EmoDB database. Although scattering coefficients show higher classification rate for every emotion class as compared to MFCC, confusion across emotion classes of similar arousal characteristics, such as, Happy & Angry, Neutral & Boredom and Fear & Happy can be observed.

The optimum value of signal duration (N) is empirically at 51000 samples (or 3.18 seconds at 16 kHz sampling rate). We start with the lowest value of T to be 4096 samples, which correspond to 256 ms at 16KHz, a duration considered optimum to convey sufficient emotion cues [26]. T is then increased by a factor of 2 until its value is below N. Interestingly, $T = 4096$ shows poor performance, whereas, $T = 32768$ shows better performance over different values of Q. In terms of parameter Q, we observed a general increase in performance for lower values of wavelets per octave. This indicates the importance of coarse-level frequency domain information for SER. This result is also comparable to the experiments done using constant-Q transform (CQT) for SER [27]. We select $T = 16384$ and $Q = 5$ as the optimum value of parameters for further experiments. Figure 3 shows the confusion matrices for scattering coefficients and MFCCs over EmoDB database. Scattering coefficients can better classify speech samples of different emotion classes in EmoDB as compared to MFCC.

B. Evaluation on other datasets

Table II shows the performance obtained with the optimised scattering parameters (ScatNet) over different databases and its comparison with standard MFCC. We compute 13 MFCCs over window length of 20 ms, 10 ms hop and 512 frequency bins over same utterance duration N (3.18 seconds). Mean and standard deviation of MFCCs is computed over all the frames to obtain a vector representation for every utterance. The scattering coefficients are observed to outperform MFCCs over every database. We also compare performance with frequency domain scattering transform (F-ScatNet) with similar optimized parameters (Q, T and N). The frequency transposition invariance obtained from frequency domain scattering introduces speaker invariance hence improving the performance. However, such improvement is not very prominent in the EmoDB database. One probable reason for such observation could be increased redundancy in frequency scattering coefficients because of small database size and low number of speakers.

TABLE II
PERFORMANCE COMPARISON BETWEEN FREQUENCY SCATTERING (F-SCATNET), TIME-DOMAIN SCATTERING (SCATNET) AND MFCC OVER DIFFERENT SER DATABASES. GIVEN VALUES ARE IN PERCENTAGES.

Database	F-ScatNet		ScatNet		MFCC	
	Acc.	UAR	Acc.	UAR	Acc.	UAR
EmoDB	74.59	70.27	74.40	71.30	58.39	54.03
RAVDESS	51.81	50.52	50.00	48.50	36.74	34.77
IEMOCAP	61.55	51.00	60.41	50.40	55.54	47.19

C. Layer-wise analysis

To analyse emotion relevance of different layers of scattering transform, we compare SER performance by using the first and second layer coefficients separately in Table III. The first layer coefficients are averaged scalogram coefficients similar to MFCC. However, the averaging provided by filter ϕ introduces time shift invariance which results in better performance than MFCC. The second layer coefficients further perform

better than first layer coefficients showing the relevance of modulation spectrum coefficients from SER perspective.

TABLE III
SER PERFORMANCE COMPARISON WITH FIRST AND SECOND LAYER TIME-DOMAIN SCATTERING COEFFICIENTS TAKEN INDEPENDENTLY. GIVEN VALUES ARE IN PERCENTAGES.

Database	First layer		Second layer	
	Accuracy	UAR	Accuracy	UAR
EmoDB	62.18	57.94	68.64	61.22
RAVDESS	38.06	36.72	48.61	47.20
IEMOCAP	56.48	47.30	58.93	48.80

V. DISCUSSION & CONCLUSION

We observe that the scattering transform coefficients, with optimized parameters, provide feature representations that are better suited for SER. The optimized wavelet averaging scale provides sufficient invariance and stability to irrelevant temporal variations to capture the emotion cues in the time domain. The low value of wavelets per octave generates coarse-level frequency domain information improving SER performance. Invariance to frequency transposition in frequency scattering further reduces the speaker-dependent variations enhancing the system performance. The performed layer-wise analysis shows the importance of time-domain averaging over the typical scalogram/mel-spectrogram coefficients. Results obtained with second layer scattering coefficients indicate the relevance of amplitude modulation of time series of different scalogram frequency bins.

We conclude that deep convolutional scattering coefficients capture more relevant emotion-related information than the standard mel-filterbank based feature. However, the performance with optimized parameter set on EmoDB varies noticeably across databases, indicating the lack of generalization. We also observe that the system faces difficulty in classifying emotions which have similar arousal characteristics (e.g., Happy & Angry). This hints towards the requirement of further analysis on improving the scattering coefficient based feature representation for SER. Our work is a preliminary study that introduces scattering transform for SER. In future, we will explore other back-end classifiers like multi-layer perceptron, gaussian mixture models etc. with scattering coefficients. To further evaluate the generalisation across samples from different domains, cross-corpus analysis for SER can also be evaluated.

REFERENCES

[1] M. Shah Fahad, A. Ranjan, J. Yadav, and A. Deepak, "A survey of speech emotion recognition in natural environment," *Digital Signal Processing*, vol. 110, p. 102951, 2021.

[2] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, 2020.

[3] 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.

[4] J. Parry, D. Palaz, G. Clarke, P. Lecomte, R. Mead, M. Berger, and G. Hofer, "Analysis of deep learning architectures for cross-corpus speech emotion recognition," in *Proc. INTERSPEECH 2019*, 2019, pp. 1656–1660.

[5] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[6] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.

[7] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[8] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. ICASSP*, 2016, pp. 5200–5204.

[9] H. Muckenhirn, V. Abrol, M. Magimai-Doss, and S. Marcel, "Understanding and visualizing raw waveform-based CNNs," in *Proc. INTERSPEECH*, 2019, pp. 2345–2349.

[10] D. Rolnick, A. Veit, S. Belongie, and N. Shavit, "Deep learning is robust to massive label noise," *arXiv preprint arXiv:1705.10694*, 2017.

[11] S. Mallat, "Recursive interferometric representation," in *Proc. of EUSICO conference, Denmark*, 2010.

[12] ——, "Group invariant scattering," *Communications on Pure and Applied Mathematics*, vol. 65, no. 10, pp. 1331–1398, 2012.

[13] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, pp. 4114–4128, 2014.

[14] J. Andén, V. Lostanlen, and S. Mallat, "Joint time–frequency scattering," *IEEE Transactions on Signal Processing*, vol. 67, no. 14, pp. 3704–3718, 2019.

[15] W. Ghezaiel, L. Brun, and O. Lézoray, "Hybrid network for end-to-end text-independent speaker identification," in *International Conference on Pattern Recognition*, Milan (virtual), Italy, Jan. 2021.

[16] J. Bruna and S. Mallat, "Audio texture synthesis with scattering moments," *arXiv e-prints*, p. arXiv:1311.0407, Nov. 2013.

[17] J. Salamon and J. P. Bello, "Feature learning with deep scattering for urban sound analysis," in *Proc. EUSIPCO*, 2015, pp. 724–728.

[18] C. Baugé, M. Lagrange, J. Andén, and S. Mallat, "Representing environmental sounds using the separable scattering transform," in *Proc. ICASSP*, 2013, pp. 8667–8671.

[19] V. Lostanlen, C. El-Hajj, M. Rossignol, G. Lafay, J. Andén, and M. Lagrange, "Time–frequency scattering accurately models auditory similarities between instrumental playing techniques," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, pp. 1–21, 2021.

[20] J. Bruna and S. Mallat, "Invariant scattering convolution networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1872–1886, 2013.

[21] D. Luo, Y. Zou, and D. Huang, "Investigation on joint representation learning for robust feature extraction in speech emotion recognition," in *Proc. INTERSPEECH*, 2018, pp. 152–156.

[22] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. INTERSPEECH*, 2005, pp. 1517–1520.

[23] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLOS One*, vol. 13, no. 5, 2018.

[24] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, p. 335, 2008.

[25] A. Rosenberg, "Classifying skewed data: Importance weighting to optimize average recall," in *Proc. INTERSPEECH*, 2012, pp. 2242–2245.

[26] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576–1590, 2018.

[27] P. Singh, G. Saha, and M. Sahidullah, "Non-linear frequency warping using constant-q transformation for speech emotion recognition," in *2021 International Conference on Computer Communication and Informatics (ICCCI-2021)*, 2021.