Speech Emotion Recognition using Time-frequency Random Circular Shift and Deep Neural Networks

Sylvain Xia^1*, Dominique Fourer*^{1}*, Liliana Audin*^{2}*, Jean-Luc Rouas*^{2} *and Takaaki Shochi*^{2}

^{1} IBISC (EA4526) - Univ. Évry/Paris-Saclay
Évry-Courcouronnes, France
^{2} IMS (UMR5218) / LaBRI (UMR5800)
CNRS, University of Bordeaux, Talence, France

dominique.fourer@univ-evry.fr

Abstract

This paper addresses the problem of emotion recognition from a speech signal. Thus, we investigate a data augmentation technique based on circular shift of the input time-frequency representation which significantly enhances the emotion prediction results using a deep convolutional neural network method. After an investigation of the best combination of the method parameters, we comparatively assess several neural network architectures (Alexnet, Resnet and Inception) using our approach applied on two publicly available datasets: eNTERFACE05 and EMO-DB. Our results reveal an improvement of the prediction accuracy in comparison to a more complicated technique of the state of the art based on Discriminant Temporal Pyramid Matching (DCNN-DTPM).

Index Terms: Speech Emotion Recognition (SER), Deep Convolutional Neural Networks, Time-frequency, Random Circular Shift (RCS)

1. Introduction

Speech is one of the most used medium for human communication. It conveys not only semantic and linguistic information but also more subtle para-linguistic information such as emotions which play a major role for a better understanding of a natural interaction. Speech Emotion Recognition (SER) gained interest during the last three decades [1] since it can find a large number of applications for robots [2], human-machine interfaces [3] or transcription systems which aim at recognizing the emotional state of a speaker from recorded utterances. From a computational point of view, SER can be addressed as a a continuous state activation problem [4] or a category prediction problem as considered in the present work.

Classical methods for SER involve two steps which are respectively the feature extraction and the classification [5]. Traditional techniques also include a preprocessing step for enhancing the speech signal of interest and for computing a suitable signal representation (e.g. time-frequency representation) which is used for computing relevant signal features [6, 7]. Recently, deep learning methods propose to address simultaneously both the feature extraction and the classification steps through a unique neural architecture. Deep learning techniques for SER have several advantages over traditional methods such as detecting the complex structure and features without the need for manual parameter estimation and tuning. They tend to extract low-level features from the given raw data. Commonly used neural architectures for SER are Long-Short Term Memory (LSTM) [8], Auto-Encoder (AE) [9], Deep Neural Network (DNN) [10, 11], Deep Convolutional Neural Network (DCNN) [12] and also attention mechanisms [13, 14]. Inspired by the promising results obtained using DCNN, we aim at proposing a simpler but efficient method achieving comparable results than those obtained by a more complicated approach. Hence, we investigate several DCNN architectures combined with a data augmentation technique based on time-frequency random circular shift transformations which have been shown relevant for improving the training of SER prediction models. Our paper is organized as follows. We formulate the problem of SER and describe materials in Section 2. The investigated method is presented in Section 3 and our numerical results are presented in Section 4. Conclusion and future work direction are discussed in Section 5.

2. Speech Emotion Recognition

2.1. Problem Formulation

The present work aims at predicting the emotion label y from the observation of a speech signal x. We consider a discrete-time signal resulting from the sampling process at rate Fs. The speech signal is real-valued and is denoted $x[n]$ where n is the sample index related to the considered time instant. SER method aims at computing the emotion label $\hat{y}$ which maximizes the posterior probability $p(y|x)$. We consider here a supervised learning scenario configuration where the probability model is constructed from a training dataset using a neural network archicture. Hence, the trained model is able to estimate from an arbitrary input signal x, its probability of belonging to a class of emotion y. The highest probability for a given class corresponds to the recognized emotion, implying that we are considering a finite discrete space of emotions. Knowing the acoustic properties of a speech signal, the analyzed frequency range of this study is limited to 0 Hz to 10 kHz.

2.2. Materials

Our study is based on two datasets which are respectively eNTERFACE05 [15] and EMO-DB [16]. They are freely available for the sake of reproducible research. eNTERFACE05 [15] is an audiovisual dataset recorded at a sampling rate of $F_s = 44.1$ kHz by 44 speakers of different nationalities. This dataset contains 1,293 English utterances pronounced by actors corresponding to a total of approximately 68 minutes of speech. Each speaker is recorded for multiple sentences with 6 different emotions: anger, disgust, fear, happiness, sadness and surprise. All emotions are globally equally represented in the whole dataset. EMO-DB [16] is a pure-audio dataset recorded by 10 speakers containing 535 utterances which correspond to

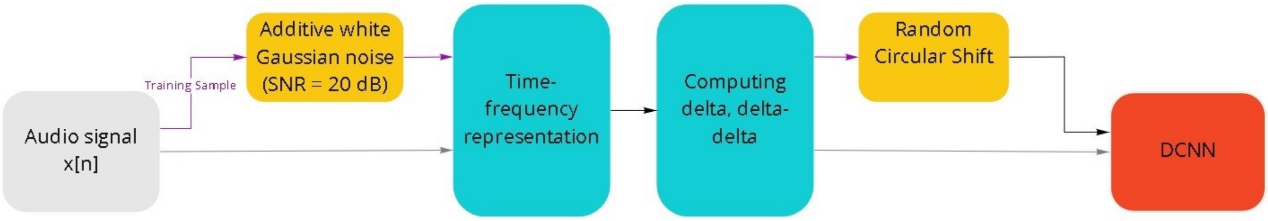

Additive white Gaussian noise (SNR = 20 dB)
Training Sample
Audio signal x[n]
Time-frequency representation
Computing delta, delta-delta
Random Circular Shift
DCNN

Figure 1: *Overview of the proposed approach.*

a total of 7 different emotions: anger, disgust, fear, happiness, sadness, boredom and neutral. All the utterances are expressed in German and recorded in an anechoic chamber at a sampling rate of $F_s = 16$ kHz with a 16-bit resolution. This dataset contains approximately a total of 25 minutes of speech.

3. Method

3.1. Global framework

The overall approach proposed in this study is depicted in Fig. 1 and considers as the input of a deep convolutional neural network (DCNN) a time-frequency representation and its delta features which correspond to the first and second-order derivative with respect to time (Δ and $\Delta\Delta$) [17]. We consider for the input, a 3D-tensor or a 3-channel image where each time-frequency coordinate is associated to a real-valued triplet. This enables the use of deep learning architectures designed for image processing which aims at recognizing specific patterns for addressing a classification problem. Our approach also considers a data augmentation process [18] during the training step, which artificially increases the number of training examples by applying signal transformations on the original ones to improve the robustness of the trained model. To this end, our strategy uses Random Circular Shift (RCS) transformations combined with an additive white Gaussian noise that is detailed below.

3.2. Input of the DCNN

In this study, we investigate two distinct time-frequency representations as the input of a DCNN which are respectively the short-time Fourier transform (STFT) and the constant-Q transform (CQT). Given a discrete-time finite-length signal $x[n]$, with time index $n \in \{0, 1, \dots, N-1\}$, and an analysis window h, the discrete short-time Fourier transform of x is computed as [19]:

$$F_x^h[n,m] = \sum_{k=-\infty}^{+\infty} x[k]h[n-k]^* e^{-j\frac{2\pi mk}{M}} \qquad (1)$$

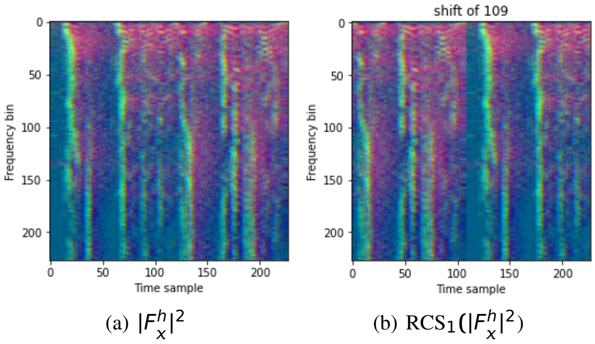
where $j^2 = -1$ and z^* is the complex conjugate of z. A real-valued representation also called spectrogram is simply computed as $|F_x^h[n,m]|^2$. The constant-Q transform (CQT) [20] is a modified version of the STFT commonly used in musical applications. It corresponds to a STFT where the length K of the window h now depends on the frequency bin $m > 0$ such as: $K_m = \frac{Q}{m}$ where Q, also called quality factor, is constant. Delta and delta-deltas representations are obtained from the considered time-frequency representation by computing finite differences along the time axis [17].

3.3. DCNN architectures

Our study investigates 3 well-known DCNN architectures which are respectively Alexnet [21], Resnet-152 [22] and Inception-v3 [23] belonging to the best state-of-the-art methods evaluated in image classification scenarios. The parameters of the convolutional layers of each architecture were pretrained on the Imagenet dataset[1] for which the considered input is now replaced by the 3D-tensor time-frequency representation previously described. We modified the final fully connected layers of each architecture to fit with the labels of the investigated emotion datasets where the output is a softmax value (in range $[0,1]$) corresponding to $p(y|x)$ over the set of considered emotions and for which argmax corresponds to the estimated emotion label $\hat{y}$.

3.4. Data augmentation strategy

Data augmentation consists in applying two transformations simultaneously on the training dataset to increase the number of examples. First, we can add a white Gaussian noise directly to the signal x in order to obtain a Signal Noise Ratio (SNR) of 20 dB. Second, we can apply directly on the time-frequency representation random circular shifts (RCS) along the time axis to obtain a new training example where the pronounced utterances are randomly merged. This process is obtained by randomly selecting (uniform law) a time instant at which the original image is circularly shifted as illustrated in Fig. 2. RCS can be applied an arbitrary number of times θ that is denoted RCS-θ such as each training example provides θ new examples. Given a number of shifts θ, each shift is performed with a random translation along time axis. The exceeding part of the image is then looped back at the beginning.



(a) $|F_x^h|^2$ (b) $\text{RCS}_1(|F_x^h|^2)$

Figure 2: *Example of RCS applied once ($\theta = 1$) to a 3-channel time-frequency representation.*

[1] https://image-net.org/

4. Numerical Results

4.1. Experimental setup

We split the audio signals of each analyzed dataset into 3-second-long segments without overlap which are considered as individuals respectively in the considered training and test datasets. The time-frequency representations are computed using the STFT or CQT using the Hann window to obtain a number of frequency bins $M = 1023$ for eNTERFACE05 and $M = 455$ for EMO-DB. The hop size is adapted to fulfill the required input size of each DCNN architecture which are different. In fact, the Alexnet and Resnet152 require inputs of size (227,227,3) and Inceptionv3 needs inputs of size (300,300,3). For the frequency axis, the number of bins corresponds to a range of frequencies up to 9785 Hz or to 7982 Hz respectively for eNTERFACE05 and EMO-DB that is finally cropped to obtain an integer number of bins. The values of the first channel of the considered time-frequency representation are then converted to the logarithmic scale (expressed in dB) before computing the second and third channel corresponding to delta and delta-delta. The resulting values are then mapped to the [0,255] range for each channel independently to obtain an input tensor of dimension $H \times W \times 3$ where H and W correspond to the height and width of the required input size of the corresponding DCNN. Each considered DCNN is initialized on ImageNet before being trained again on the considered dataset without freezing the weight of any layer. A data augmentation of 50% is first performed on each original dataset by the addition of a white Gaussian noise (SNR=20dB) for which the signals are chosen randomly. The implementation of the method is done in python using the pytorch and torchvision libraries for which the code is freely available for the sake of reproducible research[2]. The computations are completed using Cuda and two NVIDIA GPUs: a GeForce GTX 1080 Ti (GPU1) and a Tesla V100 PCIE 16GB (GPU2). In the remaining, we evaluate our proposed method in two experiments. The first experiment aims at tuning the parameters of each method in order to maximize the resulting accuracy. The second experiment is a comparative evaluation using the best tuned method with a state-of-the-art SER method [12] denoted DCNN-DTPM, where the experimental conditions are identical to those reported by the authors.

4.2. First experiment: tuning

Tables 1 and 2 show the effect of different parameter choices using our model trained with 60 epochs. We use a Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.001 and momentum of 0.9. We randomly split the dataset to have 240 samples for validation and test, which corresponds to all the sentences for the last 8 subjects (around 18%), and the remaining is used as training data. Our results show a clear advantage of STFT in comparison to CQT. We also assess the ImageNet Normalization (INN) effect which consists in normalizing the input signal through z-score ($\frac{x-\mu}{\sigma}$) using the mean μ and standard deviation σ of the ImageNet dataset. Our results show that INN provides poorer results. A higher batch size slightly improves the training speed for which the best results are obtained with 16. Finally, we show that RCS significantly improves accuracy for both STFT and CQT with the tested value $\theta = 5$. To further investigate the effect of data augmentation, we plot in Fig. 3 the resulting accuracy of the STFT+Alexnet method combined with RCS with a varying θ parameter. We show that we obtain a maximum accuracy of 91.25% for $\theta = 41$ for eNTERFACE05 and an accuracy above 86.92% for $\theta = 19$. The addition of a white Gaussian noise does not significantly improve the accuracy in our experiments.

Table 1: *Accuracy results obtained with STFT + Alexnet applied on the eNTERFACE05 dataset.*

Data aug.	mini-batch size	INN	train. time (min)	Acc. (%)
-	16	-	2	74.58
-	16	yes	2	73.33
-	32	-	2	70.41
-	32	yes	2	68.33
RCS5	**16**	**-**	**7**	**84.17**
RCS5	16	yes	7	82.91

Table 2: *Accuracy results obtained with CQT + Alexnet applied on the eNTERFACE05 dataset.*

Data aug.	mini-batch size	INN	train. time (min)	Acc. (%)
-	16	-	2	66.6
-	16	yes	2	59.58
-	32	-	2	62.08
-	32	yes	2	63.33
RCS5	**16**	**-**	**7**	**71.67**
RCS5	16	yes	7	68.75

Finally we compare in Table 3 the different investigated DCNN architectures applied on eNTERFACE05. Hence, we show that Alexnet provides the best accuracy results (maximal accuracy of 91.25% with RCS-41) and a better computational complexity with a training time of about 30 minutes for RCS41 using GPU2 when Inception takes approximately more than 4 hours and Resnet more than 7 hours. Following these results, we decide to only use the Alexnet architecture combined with RCS with the suitable θ parameter in the remaining.

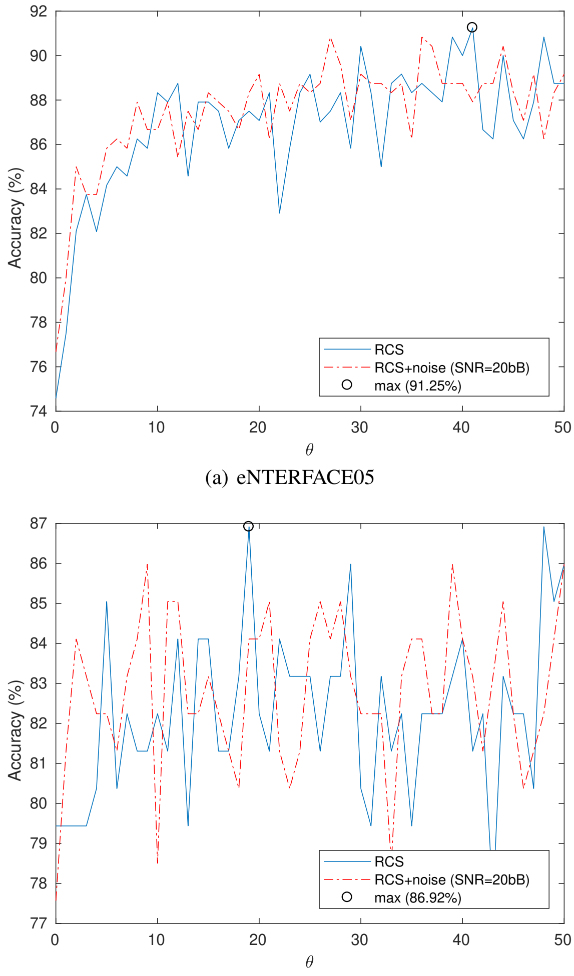
Table 3: *Comparison of the 3 investigated DCNN with different RCS applied on eNTERFACE05.*

DCNN	RCS	Acc. (%)	Train. time (min)
Alexnet	5	84.17	7 (GPU1)
Inceptionv3		85.83	60 (GPU1)
Resnet152		82.08	90 (GPU1)
Alexnet	27	90.83	34 (GPU1)
Inceptionv3		87.92	177 (GPU1)
Resnet152		86.25	300 (GPU1)
Alexnet	**41**	**91.25**	**30 (GPU2)**
Inceptionv3		87.92	267 (GPU2)
Resnet152		88.75	440 (GPU2)

4.3. Second experiment: comparative evaluation

We compare our proposed approach with a state-of-the-art technique based on a DCNN architecture based on a Discriminant Temporal Pyramid Matching (DTPM) strategy proposed in [12]. To obtain a fair comparison, now we use the same experimental setup as used in [12]. Hence, we evaluate our model with a Leave-One-Speaker-Group-Out (LOSGO) cross-validation strategy with five speaker for eNTERFACE05 dataset. As, the dataset contains 44 speakers, the last fold for validation contains the last remaining 4 speakers, giving us a total of 9 folds. A Leave-One-Speaker-Out (LOSO) strategy is used for EMO-DB dataset. Despite the authors of [12] trained

[2] https://github.com/llnanis/SER-RCS



(a) eNTERFACE05

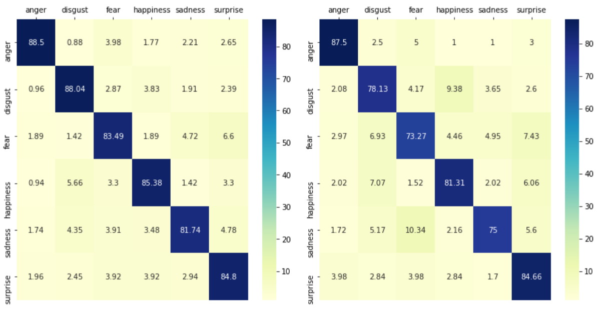
(b) EMO-DB

Figure 3: *Random Circular Shift (RCS) effect on the test accuracy using STFT+Alexnet on eNTERFACE05(a) and EMO-DB (b).*

their model with 300 epochs, we only used 60 epochs without early stopping with a mini-batch of size 30 and SGD with a learning rate of 0.001 and momentum of 0.9. Our method uses the STFT and RCS with a θ value which obtained the best results according to Fig. 3. Fig. 4 shows that for the eNTERFACE05 dataset, we achieve to recognize each emotion better than DCNN-DTPM, with a an average accuracy of 85.33% which is higher than 79.25% obtained with DCNN-DTPM. Fig. 5, shows that we obtain an average accuracy of 81.82% on the EMO-DB dataset. We achieve better results for recognizing the emotions sadness but all the other emotions are less accurately recognized since our method seems to make a confusion between happiness and fear. Thus, we obtain a slightly lower accuracy than DCNN-DTPM which has an accuracy of 87.31%. Our detailed results expressed in terms of Recall and F-score on each dataset are summarized in Table 4.

5. Conclusion

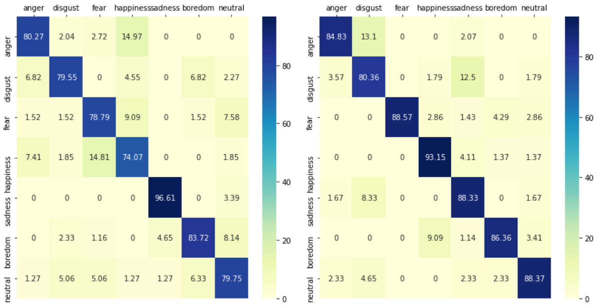
We have proposed and evaluated a simple but efficient method for emotion recognition from speech signals which uses a time-frequency representation as the input of a DCNN architecture combined with a data augmentation technique based random circular shift. Our results are mostly comparable (better on eNTERFACE05 but poorer on EMO-DB) to those obtained with a state-of-the-art method based on DTPM [12] that is more complicated to implement and train. We show that RCS can significantly improve our classification results with a reasonable increase of the training time and could be used with other CNN-based methods. We also compared several DCNN architecture and shown that Alexnet remains a suitable choice when applied to a time-frequency representation for audio classification. Future work will address the application of our proposed method in a real-world application scenario involving audio recordings from the French “Humavox” MSH project [24].



(a) proposed, STFT-Alex+RCS41 (Acc. 85.33%)

(b) DCNN-DTPM [12] (Acc. 79.25%)

Figure 4: *eNTERFACE05 confusion matrices obtained using our proposed method STFT-Alexnet + RCS41 (a) and DCNN-DTPM [12].*



(a) proposed, STFT-Alex+RCS19 (Acc. 81.82%)

(b) DCNN-DTPM [12] (Acc. 87.31%)

Figure 5: *EMO-DB confusion matrices obtained using our proposed method STFT-Alexnet + RCS19 (a) and DCNN-DTPM [12].*

Table 4: *Detailed results on eNTERFACE05 and EMO-DB using our proposed method STFT-Alexnet and RCS with the best θ value.*

dataset	Accuracy (%)	average recall (%)	average F-score
eNTERFACE05	85.33	85.03	0.85
EMO-DB	81.82	80.18	0.80

6. Acknowledgements

This research was supported by the French MSH “Humavox” project.

7. References

[1] T. Dalgleish and M. Power, *Handbook of cognition and emotion*. John Wiley & Sons, 2000.

[2] C. Breazeal and R. Brooks, "Robot emotion: A functional perspective," *Who needs emotions*, pp. 271–310, 2005.

[3] T. Hashimoto, T. Yamaguchi, and J. Miyamichi, "Emotion-oriented man-machine interface for welfare intelligent robot," *Journal of the Robotics Society of Japan*, vol. 16, no. 7, pp. 993–1000, 1998.

[4] P. Buitelaar, I. D. Wood, S. Negi, M. Arcan, J. P. McCrae, A. Abele, C. Robin, V. Andryushechkin, H. Ziad, H. Sagha *et al.*, "Mixedemotions: An open-source toolbox for multimodal emotion analysis," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2454–2465, 2018.

[5] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117 327–117 345, 2019.

[6] T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *2005 IEEE International Conference on Multimedia and Expo*. IEEE, 2005, pp. 474–477.

[7] I. Anagnostopoulos, Christos-Nikolaos, G. Theodoros, and Ioannis, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011 jo - artificial intelligence review," *Artificial Intelligence Review*, vol. 43, 2015.

[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. IEEE ICASSP*, 2016, pp. 5200–5204.

[9] S. Latif, R. Rana, J. Qadir, and J. Epps, "Variational autoencoders for learning latent representations of speech emotion: A preliminary study," *arXiv preprint arXiv:1712.08708*, 2017.

[10] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *Proc. IEEE ICASSP)*, 2011, pp. 5688–5691.

[11] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Interspeech 2014*, Sep. 2014.

[12] D. Zhang, X. Dai, and Y.-F. Wang, "Dynamic temporal pyramid network: A closer look at multi-scale modeling for activity detection," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 712–728.

[13] M. Chen, X. He, J. Yang, and H. Zhang, "3-d convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, 2018.

[14] C. W. Lee, K. Y. Song, J. Jeong, and W. Y. Choi, "Convolutional attention networks for multimodal emotion recognition from speech and text data," *ACL 2018*, vol. 28, 2018.

[15] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *22nd International Conference on Data Engineering Workshops (ICDEW'06)*. IEEE, 2006, pp. 8–8.

[16] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Ninth european conference on speech communication and technology*, 2005.

[17] J. Picone, "Speech recognition using mel cepstrum, delta cepstrum and delta-delta," Ph.D. dissertation, Mississippi State University, 1998.

[18] D. A. Van Dyk and X.-L. Meng, "The art of data augmentation," *Journal of Computational and Graphical Statistics*, vol. 10, no. 1, pp. 1–50, 2001.

[19] D. Fourer, J. Harmouche, J. Schmitt, T. Oberlin, S. Meignen, F. Auger, and P. Flandrin, "The ASTRES toolbox for mode extraction of non-stationary multicomponent signals," in *Proc. EUSIPCO*, Kos Island, Greece, Aug. 2017, pp. 1170–1174.

[20] J. C. Brown, "Calculation of a constant q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., vol. 25. Curran Associates, Inc., 2012.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[24] L. Audin-Garcia, D. Fourer, J.-L. Rouas, and T. Shochi, "Humavox: Analyse acoustique et cognitive de la prosodie affective dans le soin gériatrique," in *Colloque Interdisciplinarité (s) du RnMSH*, 2021.