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

Sylvain Xia[1], Dominique Fourer[1], Liliana Audin-Garcia[2], Jean-Luc Rouas[3] and Takaaki Shochi[3]
[1]IBISC (EA 4526), Univ. Évry/Paris-Saclay, Courcouronnes, France
[2]IMS, (CNRS UMR 5218). Cognitique Team. Bordeaux INP-ENSC. Talence, France
[3]LABRI (UMR 5800), Univ. Bordeaux, Talence, France
[1]dominique.fourer@univ-evry.fr

Abstract

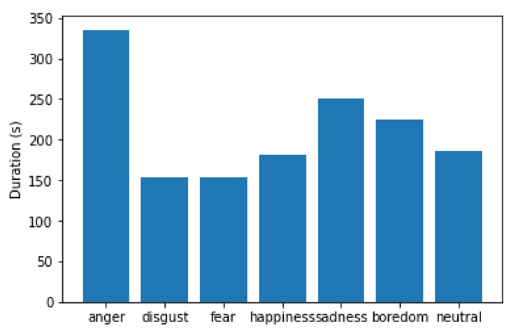
This paper addresses the problem of emotion recognition from a speech signal. Thus, we investigate a data augmentation technique based on circular shift of the input time-frequency representation which significantly enhances the emotion prediction results using a deep convolutional neural network method. After an investigation of the best combination of the method parameters, we comparatively assess several 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).

Main Contributions

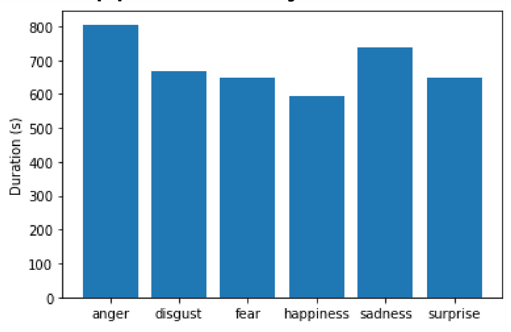
- Investigation of time-frequency representations (STFT and CQT) used as the input of advanced deep neural networks (Alexnet, Resnet and Inception) for speech emotion recognition (SER).
- Introduction of a new scalable data augmentation technique called random circular shift (RCS).
- Comparative evaluation with a recent state-of-the-art deep learning method (DCNN-DTPM) applied on two public datasets.
- Python codes available : `https://github.com/llnanis/SER-RCS`

Materials

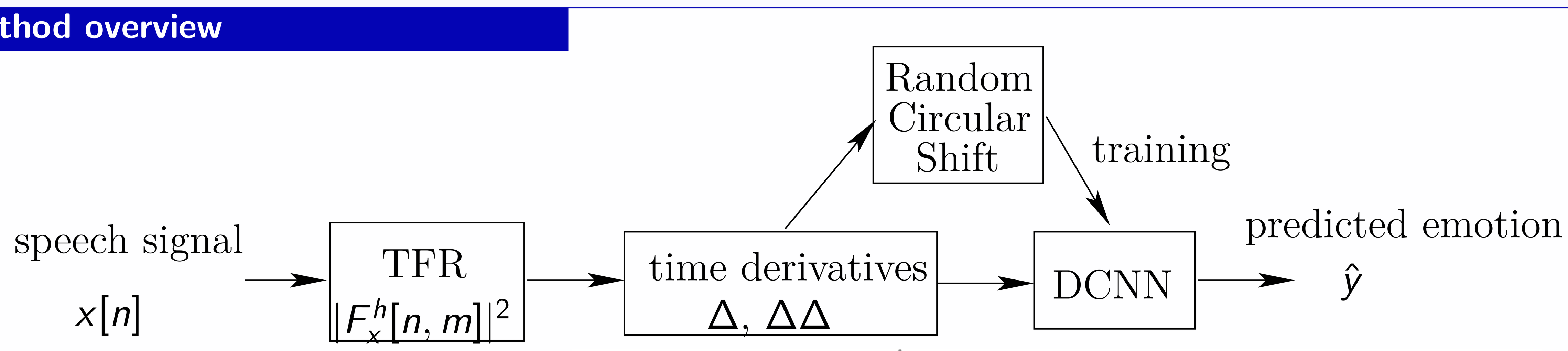
eNTERFACE05 [MKMP06] is an audiovisual dataset recorded at a sampling rate of $F_s = 44.1$ kHz by 44 speakers of different nationalities. This dataset contains 1,293 English utterances pronounced by actors corresponding to a total of approximately 68 minutes of speech. Each speaker is recorded for multiple sentences with 6 different emotions : anger, disgust, fear, happiness, sadness and surprise. All emotions are equally represented in the dataset.



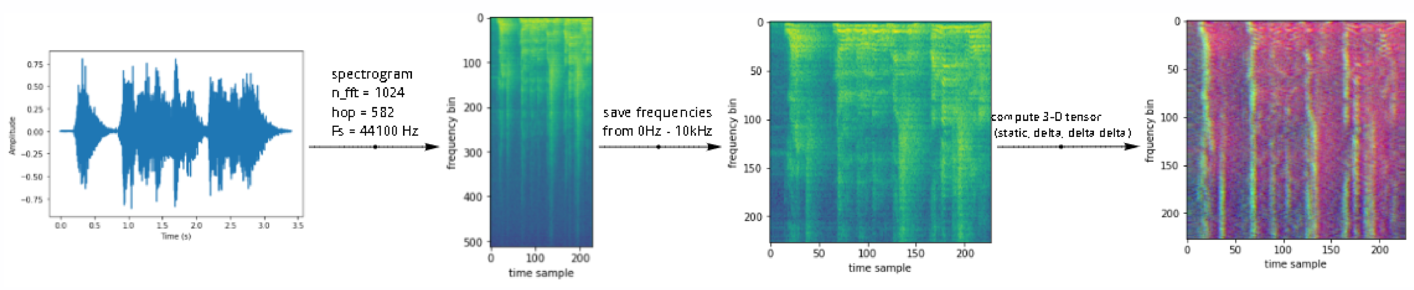
EMO-DB [BPR+05] is a pure-audio dataset recorded by 10 speakers containing 535 utterances which correspond to a total of 7 different emotions : anger, disgust, fear, happiness, sadness, boredom and neutral. All the utterances are expressed in German and recorded in an anechoic chamber at a sampling rate of $F_s = 16$ kHz. This dataset contains approximately 25 minutes of speech.



Method overview



- Time-Frequency Representation (TFR) :
- Time-Derivatives (Δ and $\Delta\Delta$)
- Random Circular Shift (RCS)
- Deep Convolutional Neural Network (DCNN)



DCNN input

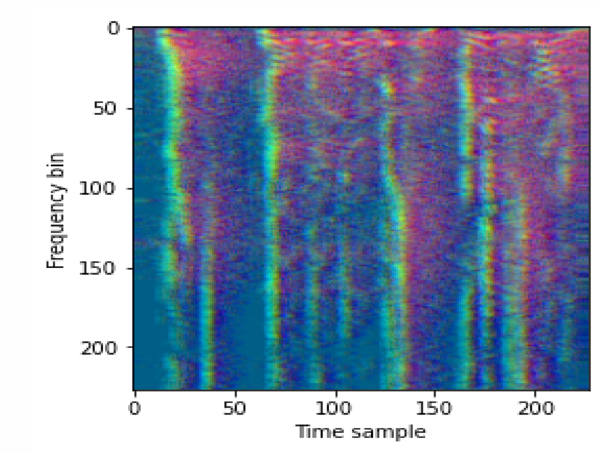
- Given a discrete-time finite-length signal $x[n]$, with time index $n \in \{0, 1, ..., N-1\}$, and an analysis window h, the discrete STFT of x is computed as [FHS+17] :

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

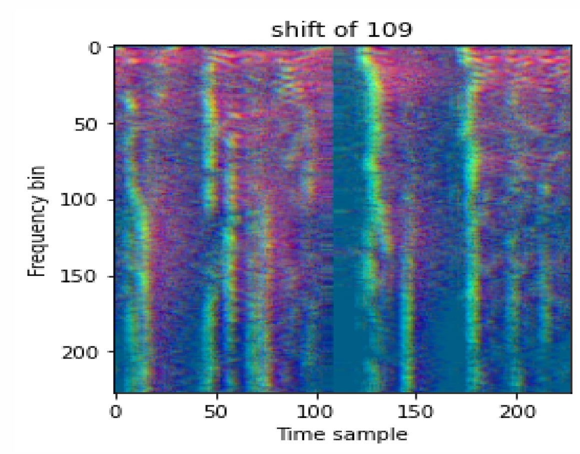
with $j^2 = -1$. A real-valued TFR is provided by $|F_x^h[n, m]|^2$.
- Constant-Q transform (CQT) is a modified STFT using a window h with a time-spread depending on the frequency bin $m > 0$ such as $K_m = \frac{Q}{m}$ where quality factor Q is constant.
- Delta and delta-deltas representations are obtained from the considered time-frequency representation by computing finite differences along the time axis.

Random Circular Shift (RCS)

RCS is a new data augmentation method applied to a TFR along the time axis to obtain new training examples with randomly merged utterances. We randomly select a time instant at which the original image is circularly shifted. RCS can be applied consecutively an arbitrary number of times θ.



(a) $|F_x^h|^2$



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

Numerical Results (Experiment 1 : Tuning)

STFT vs CQT on eNTERFACE05

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

(c) STFT + Alexnet

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

(d) CQT + Alexnet

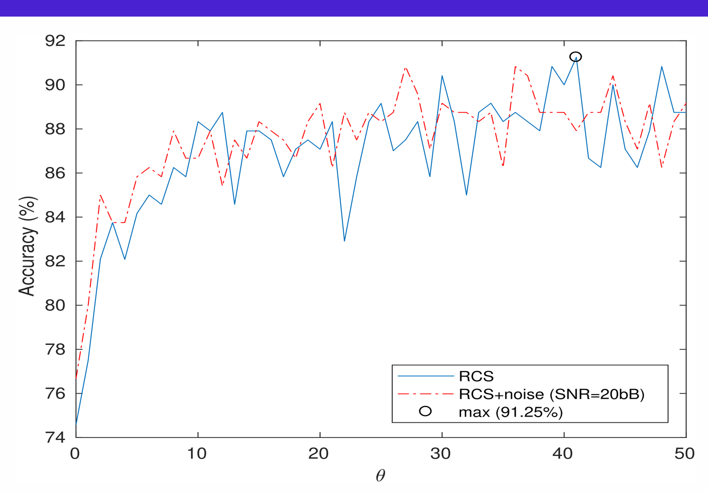
INN* : ImageNet Normalization

Alexnet vs Resnet vs Inception on eNTERFACE05

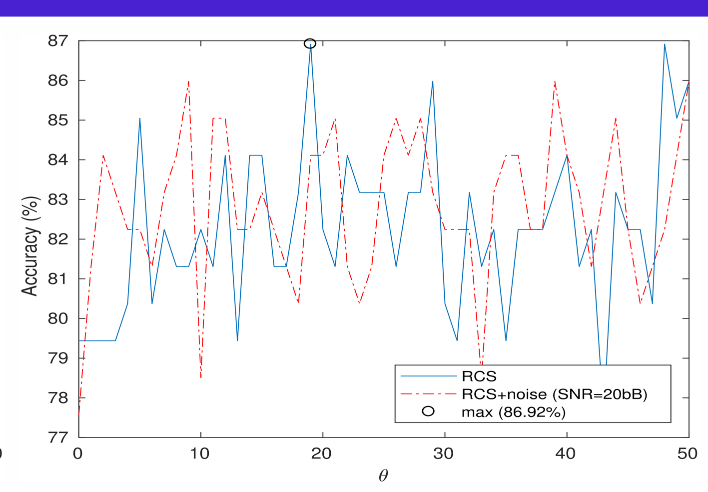
Computations use a NVIDIA GTX 1080 Ti (GPU1) and NVIDIA Tesla V100 (GPU2)

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)

Best RCS θ value for each dataset



(e) eNTERFACE05

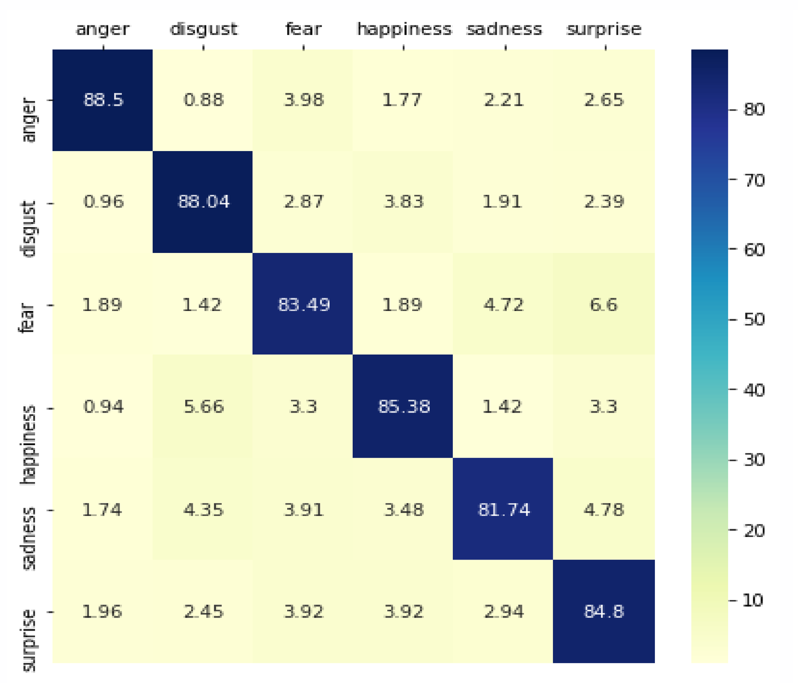


(f) EMO-DB

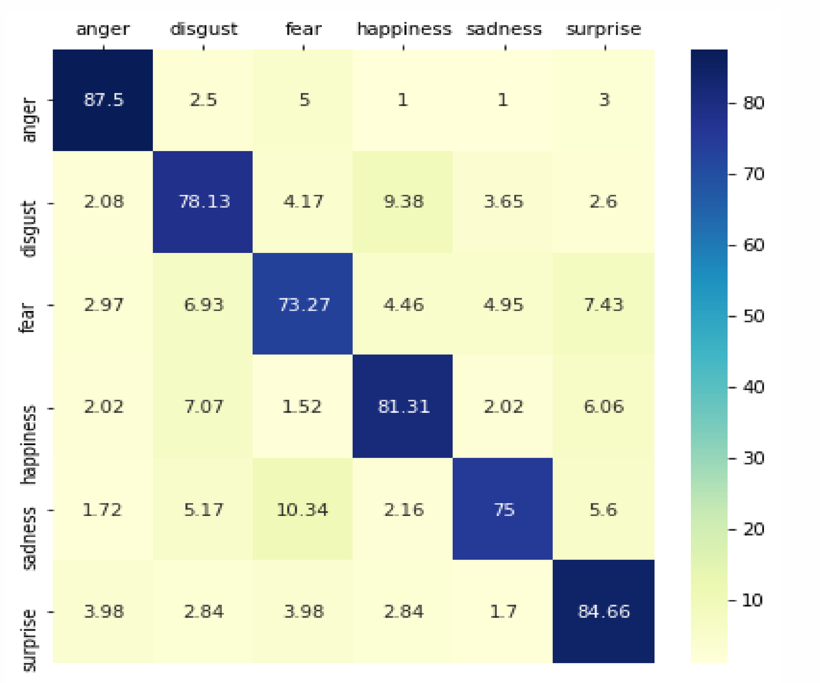
Numerical Results (Experiment 2 : Comparative evaluation)

Comparative evaluations use a n-fold cross-validation methodology.

eNTERFACE05

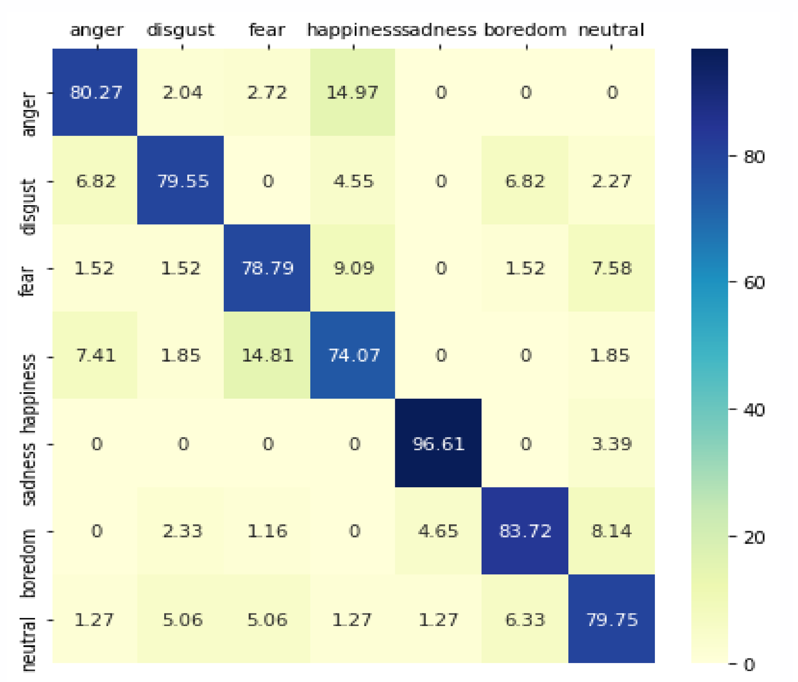


(g) proposed, STFT-Alex+RCS41 (Acc. 85.33%)

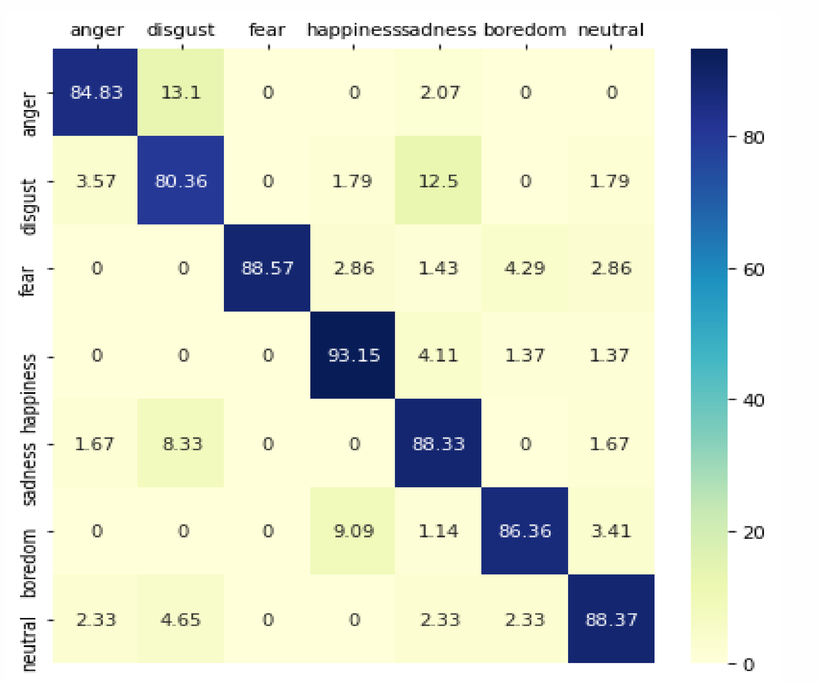


(h) DCNN-DTPM [ZDW18] (Acc. 79.25%)

EMO-DB



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



(j) DCNN-DTPM [ZDW18] (Acc. 87.31%)

Bibliography

[BPR+05] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. A database of german emotional speech. In *Ninth european conference on speech communication and technology*, 2005.

[FHS+17] 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*, pages 1170–1174, Kos Island, Greece, August 2017.

[MKMP06] Olivier Martin, Irene Kotsia, Benoit Macq, and Ioannis Pitas. The enterface'05 audio-visual emotion database. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pages 8–8. IEEE, 2006.

[ZDW18] Da Zhang, Xiyang Dai, and Yuan-Fang Wang. Dynamic temporal pyramid network : A closer look at multi-scale modeling for activity detection. In *Asian Conference on Computer Vision*, pages 712–728. Springer, 2018.

This project is supported by the French interMSH "Humavox" project involving MSHA and MSH Paris-Saclay.