Workshop Proceedings of the 19th International Conference on Intelligent Environments (IE2023)
G. Bekaroo et al. (Eds.)

doi:10.3233/AISE230012

Speech Emotion Recognition Using Deep Convolutional Neural Networks Improved by the Fast Continuous Wavelet Transform

Björn E. VAN ZWOL[a, 1], Mathijs A. LANGEZAAL[a,b, 1], Lukas P.A. ARTS[a], Albert GATT[a], and Egon L. VAN DEN BROEK[a]

[a] *Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands*

[b] *Population-Based Epidemiological Cohorts Unit UMS11, INSERM, Paris, France*

Abstract. The fast Continuous Wavelet Transform (fCWT) is used to improve Deep Convolutional Neural Networks (DCNN)'s Speech Emotion Recognition (SER). While being computationally efficient, the fCWT's time-frequency analysis overcomes traditional methods' resolution limitations (e.g., Short-Term Fourier Transform). fCWT-induced DCNNs are compared to state-of-the-art DCNN SER systems. Comparing different wavelet parameters, we also provide an empirical strategy for balancing temporal and spectral features in speech signals. We suggest that this strategy is of generic interest for non-stationary signal processing where large amounts of data are available. fCWT's potential for improving SER accuracy in real-time applications is confirmed. In parallel, the variance in the cross-validation folds confirmed deep learning's vulnerability on non-big data sets.

Keywords. Deep Learning, Deep Convolutional Neural Networks, Signal Processing, Continuous Wavelet Transform, fCWT, Speech Emotion Recognition

1. Introduction

When developing Intelligent Environments (IE), we are faced with many challenges, both from the engineering and the human side. IE even requires that the two can communicate bidirectionally. Consequently, IE's intelligence needs to have good quality signals as input [1], as without this no intelligence is possible.

Humans excel in communicating and interacting with both each other and machinery either tangible (e.g., robots [2]) or ambient (e.g., IE [3]). This relies on i) principles of (good) communication (e.g., Grice's Maxims), ii) linguistic information, and iii) paralinguistic information (e.g., emotions). Although each of these elements have been studied extensively, especially the latter remained challenging [4]. Here, we consider a recent technological advancement in signal processing to improve the analysis of paralinguistic information, more specifically emotions.

Speech is a Natural User Interface and likely, the most natural form of human communication. Consequently, automatic speech recognition has drawn a lot of attention

[1] Shared first and corresponding authors: b.e.vanzwol@uu.nl and m.a.langezaal@uu.nl.

from both academia and industry. In its slipstream, Speech Emotion Recognition (SER) received increasing attention [5], as was shown that also paralinguistic aspects of speech are important and sometimes even most important [6]. For example, the pitch, volume, and intonation in speech all carry cues of various degrees that unveil speakers' emotional state [7]. With Human-Computer Interaction, Human-Robot Interaction, and interaction of humans in IE it was shown to be valuable to understand human's emotions [2,3]. This enables natural and fluent interactions [8]. However, in all these cases, the speech signal needs to be processed in real-time with the highest possible accuracy.

Essentially, SER can be considered as a pattern recognition problem. Taken this perspective, feature extraction is a critical element of the processing pipeline, as it allows identifying parts of the signal that carry meaningful information, as opposed to noise. Simple features include pitch, loudness, and duration. Additionally, frequency content has been studied, which provided features like linear prediction cepstral coefficients and mel-frequency cepstral coefficients, which focuses on parts of a signal's spectrum that is connected to human hearing [9]. Another method for processing and analyzing speech data for SER is using Time-Frequency Representations (TFR), which extract both temporal and spectral features of the speech signal [9,10]. Several techniques exist for performing time-frequency analysis, such as the Short-Term Fourier Transform (STFT) and Continuous Wavelet Transform (CWT). Here, a trade-off between the quality of the analysis and the computational load must often be considered. However, in the case of real-time SER, both are equally important.

Fueled by available SER databases, deep learning showed to be increasingly successful in speech processing [11] The application of deep learning on SER showed two main advantages: i) combine the feature extraction and classification steps and ii) automate and/or simplify the feature extraction, especially in cases where large amounts of data are available. For more elaborate discussions on this topic, we refer to [12,13,14]. Zhang et al. [9] transformed their speech signals to spectrograms: two-dimensional TFR of the signal. Next, they used these as input for a Deep Convolutional Neural Networks (DCNN), which was pre-trained on image datasets. They performed transfer learning on AlexNet [15], extracting the output of the 'FC7' layer. They combined this output into a global utterance-level feature representations and fed this to a linear Support Vector Machine (SVM). Xia et al. [10] used a similar approach as [9]. Using a DCNN for both feature extraction, representation, and classification, they applied an end-to-end analysis. Using a STFT on the signal, they obtained their TFR. Subsequently, they applied a novel Data Augmentation (DA) technique dubbed Random Circular Shift (RCS), which takes part of the end of a signal and adds it to the front. Zhang et al. [9] and Xia et al. [10] obtained a similar performance.

TFR deals with a fundamental resolution limitation caused by the uncertainty principle of signal processing. TFR-based systems typically use STFT [10], which has a suboptimal solution to the resolution limitations [16]. In practice, this results in a loss of information not visible in STFT-based TFR. Given the DCNN's ability to detect abstract, low-level features of input data, a high-quality TFR is desired. The CWT overcomes this drawback by using different resolutions for different frequencies; however, until recently numerical implementations of CWT did not allow real-time computation on edge computing devices. fast Continuous Wavelet Transform (fCWT) is a recent advancement in time-frequency analysis that provides real-time CWT [16]. Here, using DCNN, we explore the potential of fCWT-based compared to STFT-based SER [9,10].

In the next section, we provide a concise introduction on time-frequency analysis, the core of the current work. In Section 3, we describe our methods and implementation. Section Section 4 provides our results and discussion. We conclude in Section 5.

2. Time-Frequency Representations (TFR): Theoretical background

Speech is a signal that varies over time. Via the Fourier transform its frequency content, or the spectrum, can be obtained [1,16,17]. However, this provides a blurred representation of the speech signal, as it is a non-stationary signal: its frequencies change over time. Such changes are difficult to see in time and frequency space separately [16,17]. TFR provide a solution for analyzing such signals, by providing an overcomplete representation of the signal. This enables simultaneous temporal and spectral signal analysis.

Several transformations exist to obtain a TFR. The discrete STFT is the one used most, including in SER. Intuitively, the STFT can be understood as follows: For each time step, the signal is multiplied with the window, creating a segment of the original signal, which then is Fourier transformed to obtain the frequencies present in that particular segment. After plotting this spectrum vertically for the present time index, this index is incremented step by step to repeat the process for all time values. This process creates a two-dimensional representation in which the frequency content (on the vertical axis) can be analyzed as a function of time (on the horizontal axis).

Formally, the STFT is defined as follows [18]. Given a signal $x[m]$ as a function of the sampling or time index m:

$$\text{STFT}_w\{x\}(m,f) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-i2\pi fn}, \tag{1}$$

where f is a frequency, which is represented to be continuous for notational clarity; however, to implement STFT computationally, both f and m are discretized. w is the window function, which is equal to zero on most of its range, except for a compact interval centered around zero, where we use the commonly chosen Hann window, which is a simple raised cosine.

Crucial for any time-frequency analysis is the observation that the spectral components of a single point in time are fundamentally unknowable [17]. This is because a frequency can only be defined with respect to a non-zero time interval: for a single point, frequency becomes a meaningless concept. In other words, time t and frequency f cannot be known simultaneously with arbitrary precision. Hence, the product of their standard deviations σ_t and σ_f is bounded from below [19]:

$$\sigma_t \cdot \sigma_f \geq \frac{1}{4\pi}, \tag{2}$$

which is an uncertainty principle, similar to the Heisenberg uncertainty principle found in quantum mechanics [20]. In other words, an ‘exact’ TFR does not exist: per definition, one has to deal with an unavoidable trade-off between time and frequency resolution.

For STFT, the resolution is determined by the size of the window function. A narrow window gives high time resolution with low frequency resolution and vice versa. Crucially for STFT, this resolution is *fixed* for all frequencies, since the same window is used for the entire signal. Since real-world signals contain different frequency components in distinct periods of time [16] this can limit its credibility. In particular, high frequencies often have a short duration, while low frequencies typically persist over longer time intervals. This triggered the creation of the wavelet transform and MultiResolution Analysis (MRA) [17]. MRA lets the resolution vary as a function of the frequency, which results in good time resolution for high frequencies, and good frequency resolution for low frequencies. As such, MRA is perfectly suited for real-world signals.

A common method for MRA is the CWT, defined as [16]:

$$\mathrm{CWT}_{\psi}\{x\}(a,b) = \frac{1}{|a|}\sum_{n=0}^{N-1} x[n]\,\bar{\psi}\left[\frac{n-b}{a}\right], \tag{3}$$

where x is a discrete signal, $a \propto 1/f$ is the inverse frequency (or: scale), b is the time translation (analogous to m in Eq. (1)), and ψ is the so-called mother wavelet, with the bar denoting the complex conjugate. It is called continuous because the wavelet is sampled continuously, in contrast with the discrete wavelet transform. The wavelet fulfills a similar role to the window in STFT, where its width changes depending on the frequency, as determined by s. This gives the CWT its multi-resolution nature. Note that the CWT does not 'solve' the resolution issue (see Eq. (2)). Depending on the frequency, it changes the size of the time and frequency intervals on which calculations are done. This is illustrated in Figure 1.

For the mother wavelet, we choose to employ the often used Morlet wavelet, which is the product of a complex exponential and a Gaussian window [21]:

$$\psi_{\sigma}(t) = \pi^{-1/4} e^{i2\pi t} e^{-\frac{t^2}{2\sigma}}, \tag{4}$$

where σ is the parameter that determines the width of the Gaussian. Changing this creates different representations, which interpolate between the original time series and the Fourier transform. A low and high σ provide respectively temporal and spectral features.

We hypothesize that MRA's relevance for SER is its faithful representation of speech's paralinguistic information. Until recently, CWT were too slow to perform the real-time computations needed for SER. However, the recent introduction of fCWT [16] has changed this, which enables the exploration of the potential of MRA for SER.

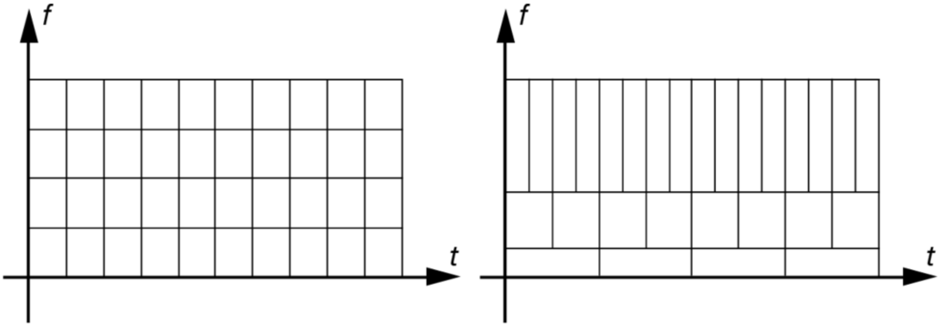

Figure 1. The fixed and frequency-dependent resolutions of the Short-Term Fourier Transform (STFT, left) and the Continuous Wavelet Transform (CWT, right), visualized on a schematic time (t)/frequency (f) plot. As determined by Eq. (2), all rectangles have an equal area. Image adapted from Letswave, https://letswave.cn/.

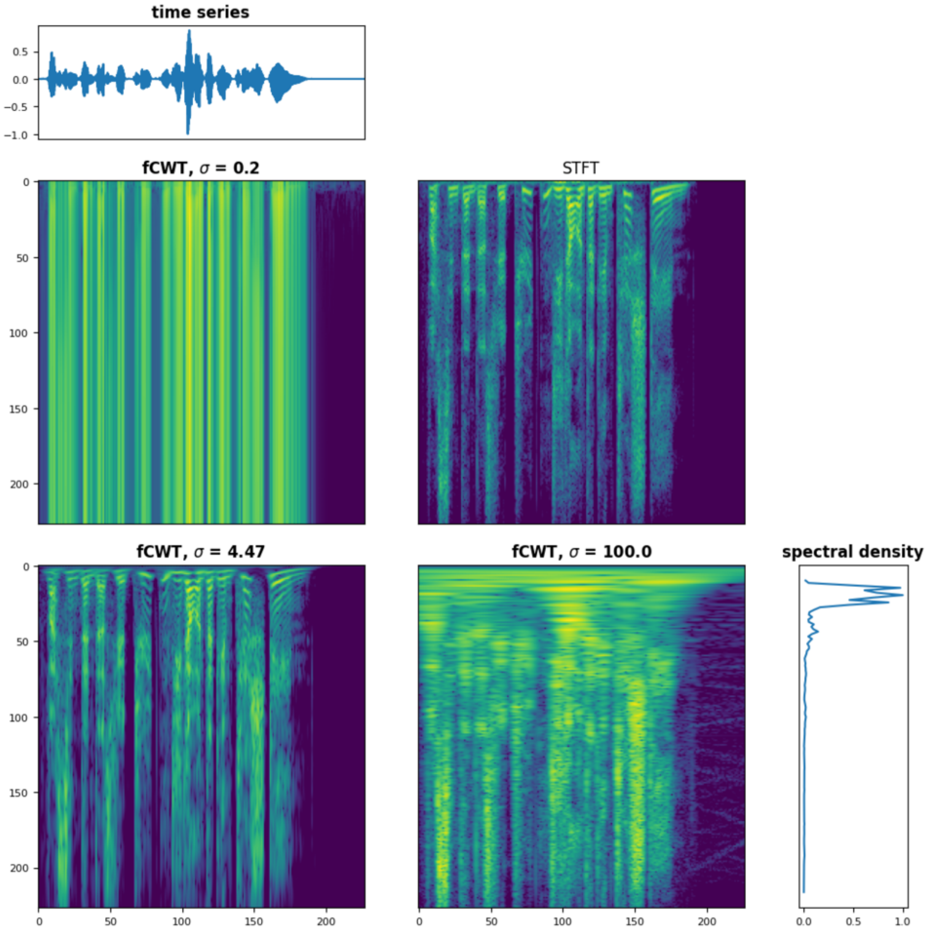

Figure 2. An arbitrary audio signal (top), along with several Time-Frequency Representations (TFRs, middle), and the normalized spectral density (bottom right). For the TFRs, three Continuous Wavelet Transforms (CWTs) with increasing values of σ are shown, as well as the Short-Time Fourier Transform (STFT) for comparison.

3. Methods

We hypothesize that fCWT's multi-resolution has advances over STFT (see Figure 2). To determine whether or not this is true, we replicated [10], using both STFT and fCWT.

3.1. Data

Like [10], we apply transfer learning on the DCNN architecture AlexNet. We also use the same databases: eNTERFACE05 [22] and EMO-DB [23]. eNTERFACE05 [22] is an audiovisual dataset ($F_s = 44.1kHz$) containing 1,293 English utterances from 44 actors of different nationalities, corresponding to a total of 68 minutes of speech. Of each actor an equal number of sentences with 6 different emotions: anger, disgust, fear, happiness, sadness and surprise are recorded. EMO-DB [23] is an audio-only dataset ($F_s = 16.0kHz$)

consisting of 535 German utterances by 10 actors, containing approximately 25 minutes of speech in total. The following 7 emotions are represented: anger, disgust, fear, happiness, sadness, boredom, and neutral.

With fCWT, the σ parameter determines the trade-off between the temporal and spectral features (see Section 2). This has been visualized in Figure 2, where examples of fCWT-TFRs for several values of σ have been shown, next to the TFR created with STFT. These show that varying σ highlights different features. Using fCWT-TFR with σ values in a range of $(0,100)$, with steps of $2:1,3,5,\ldots,97,99$, we developed, trained, and tested our models. We selected the fCWT-TFR with σ providing the best SER.

The size of both data sets [22,23] is rather small. To increase their size, DA strategies can be used, which improve model robustness [24]. In [10], the DA strategy dubbed RCS was used. We compared this with a reference DA method: White Gaussian Noise (WGN) [25]. With RCS, a TFR is circularly shifted on the time axis to a randomly selected time instant. The exceeding part of the image is looped back at the beginning. With all data, this can be performed an arbitrary amount of times. Using RCS, [10] obtained high performance; however, it appears dubious from a signal processing perspective. When taking part of an utterance and adding it to the beginning of the signal, the signal's properties are changed (e.g., the intonation). High to low pitch might become low to high pitch. We hypothesize that RCS can 'scramble' the speech signal's paralinguistic properties. In line with [10], we applied RCS5, which resulted in 5 new training samples for each entry. WGN is directly applied to the speech signal before time-frequency analysis. To generate the same number of entries as with RCS5, we applied 5 levels of WGN, using signal-to-noise ratios 10, 15, 20, 25, and 30 dB [25].

Using respectively $M = 1{,}023$ eNTERFACE05 [22] and $M = 455$ EMO-DB [23] frequency bins, STFT was applied. As proposed in [10], the hop size is adapted to fulfill the DCNN required input size. For both fCWT and STFT, the speech signals are analyzed to a range of frequencies up to 9,785Hz (eNTERFACE05) and 7,982Hz (EMO-DB), respectively. The resulting TFR are converted to a logarithmic scale (in dB). Subsequently, the Δ and $\Delta\Delta$ features are taken, which correspond to estimates of the first and second time derivatives of the signal [9]. The resulting RGB-images are mapped to the [0,255] range and scaled to (227,227,3), DCNN's required input size. For fCWT, this is done using OpenCV-python package's 'INTER_AREA' interpolation.

3.2. Evaluation

We apply leave-one-speaker-out cross-validation and leave-one-speaker-group-out cross-validation on the EMO-DB, and eNTERFACE05 datasets [10]. For eNTERFACE05, 5 speakers are left out, resulting in 9 folds. For EMO-DB, 1 speaker is left out, resulting in 10 folds. For both the validation and test set, 1 fold is used. Using these cross-validation strategies, the two DA strategies and fCWT's optimal σ are performed and applied to the training folds only.

For each of the training sets, we develop a DCNN using the AlexNet architecture [15]. The parameters of the convolutional layers were pre-trained on the Imagenet dataset [26], where the architecture is adapted to allow for the aforementioned 3D-tensor TFR as input. The model's final layer is modified to fit the emotion labels y, where the output is a $softmax$ value corresponding to $p(y|x)$ over the considered emotions where the *argmax* corresponds to the estimated emotion label $\hat{y}$.

To allow for a fair comparison with [10], we used the same training strategy: a Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.001, momentum of 0.9, mini-batch size of 16, and 60 training epochs. During training, the validation set is used to calculate cross-entropy loss to reduce the chance of overfitting and allow for the best-performing model to be returned. Using the pre-trained Imagenet parameters, the DCNN is initialized and trained again on the training fold without freezing any layer. The implementation was done in Python, using PyTorch and the TorchVision packages.

4. Results

Using the average accuracy over all the cross validation test folds of the corresponding database, we evaluated the DCNN models. To test our hypothesis, for each database, we compared STFT and fCWT-trained DCNN models as well as the two DA strategies.

Table 1 provides the performance of the fCWT and STFT DCNN-models. We found optimal σ values of 7 and 3 for EMO-DB and eNTERFACE05, respectively. When RCS 5 was used as DA, with both EMO-DB [23] and eNTERFACE05 [22], fCWT achieved the highest performance. However, fCWT did not outperform STFT in all cases. In case of EMO-DB, we find that fCWT only outperforms STFT using RCS5, whereas fCWT is outperformed when no DA and WGN are used. For eNTERFACE05, fCWT outperforms STFT when RCS5 and WGN are used as DA, whereas STFT outperforms fCWT when no DA is applied. In all cases, the use of fCWT and STFT for TFR both resulted in the best model performance equally often: both 3×.

The model performances differ around 2%. Only with EMO-DB with RCS5 applied, we see a 5% difference between fCWT and STFT. Hence, our hypothesis that fCWT would outperform STFT-trained DCNNs on SER cannot be confirmed. It remains an open question whether the fixed resolution STFT or the CWT's MRA is best for SER. A possible reason that fCWT does not perform better is the fact that the required input size of the DCNN is fixed to (227,227,3) and, therefore, required a scaled down TFR. Image down scaling was done using interpolation, which might resulted in information loss. Full scale fCWT TFR might contain more paralinguistic information as it allows MRA. Hence, future studies could take advantage of fCWT's MRA by training DCNN that allow for a large input size, which do not require that much image down scaling and, hence, possibly improve SER classification.

Table 1. For the EMO-DB [23] and eNTERFACE05 [22] databases, accuracies of the fast Continuous Wavelet Transform (fCWT) and Short-Term Fourier Transform (STFT)-trained Deep Convolutional Neural Network (DCNN) models per Data Augmentation (DA) strategy: Random Circular Shift (RCS) and White Gaussian Noise (WGN). See Section 3 for details.

		no DA	RCS5	WGN
EMO-DB	fCWT	65.2%	71.0%	66.0%
	STFT	65.6%	65.3%	67.9%
eNTERFACE05	fCWT	64.5%	76.6%	73.1%
	STFT	67.1%	74.1%	72.1%

Both RCS and WGN DA improves results compared to no DA (see Table 1). However, in some cases the improvement in results is marginal, in particular for STFT-WGN and fCWT-WGN trained on EMO-DB [23]. With EMO-DB [23], WGN did not improve the results. In contrast, results did improve with eNTERFACE05 [22] (see Table 1). STFT-RCS on EMO-DB is the odd one out, with <5% accuracy than for the other sets that use RCS (see Table 1). RCS showed to be the best DA strategy for the current setup. This suggests that performing RCS on TFR does not substantially distort the paralinguistic information. However, the performed experiments do not rule out this possibility. It could be that adding RCS works well for performance on this dataset but damages its generalization due to the hypothesized scrambling. Next to that, WGN is known to improve Out-of-Distribution (OoD) performance [25]. Additional evaluation of these models on other databases is needed.

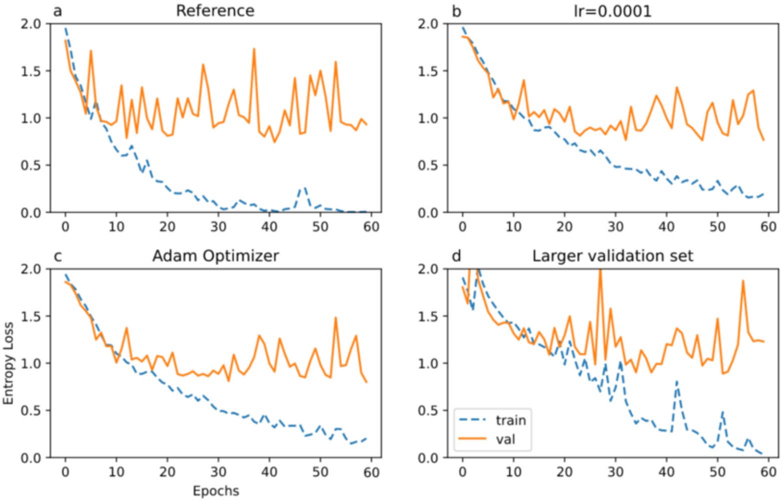

Figure 3. Entropy loss during Deep Convolutional Neural Network (DCNN) training and validation using different model training strategies. **A)** displays the original training strategy, **b)** a learning rate (lr) of 0.0001, **c)** the Adam optimizer, and **d)** a larger validation set. In all cases, the other hyperparameters are held constant using the original training strategy as reference. An arbitrary cross validation fold is plotted that used the Short Term Fourier Transform (STFT) without data augmentation on the EMO-DB database

The results presented in Table 1 are difficult to interpret, which may be attributed to the training strategy adopted from [10]. Figure 3a shows that the training loss decreases rapidly, while the validation loss reaches a plateau after a few epochs, which suggests overfitting. Additionally, the validation loss shows significant oscillation. These issues are inherent to the adopted training strategy due to three factors: 1) a high learning rate, 2) a suboptimal optimizer, and 3) an inadequately sized validation set. Consequently, This resulted in a large variance in the average model performance, and, hence, model comparison should be done with care. The obvious solution of confidence intervals around cross-validation estimates is difficult [27].

Figure 3b, c, and d show the impact of modifying the training strategy by using a smaller learning rate, the Adam optimizer, and a larger validation set, respectively. In each instance, the models display a slower overfitting rate on the training set, whereas Figure 3b and c show reduced validation loss oscillations. Figure 3d still exhibits heavy oscillation in addition to a decrease in model performance. This is likely due to the decreased training set size, resulting in a decrease in the generalizability of the model. These findings suggest that alternative training strategies and hyperparameter tuning may result in enhanced stability and superior model performance.

5. Conclusion

The fCWT was used to improve DCNN's SER. While being computationally efficient, fCWT's time-frequency analysis has been shown to overcome traditional methods' resolution limitations (e.g., STFT) [16]. To optimize model performance, different wavelet parameters were compared. This provided a strategy for balancing temporal and spectral features in non-stationary signals such as speech. Compared to a state-of-the-art DCNN SER system, fCWT-induced DCNN showed its potential to improve SER. However, oscillations in the cross validation folds show signed instable results due to a lack of data.

Other SER databases could also be considered to test the OoD performance of such models. Furthermore, future studies should take advantage of fCWT's MRA by developing and training DCNN that allow for a larger image input size to gain more insight into the effects of fCWT TFR on SER performance using DCNN. Nevertheless, once more one of deep learning's vulnerabilities is shown: a lack of data. A situation where IE often are in, as it is inevitable that its users and context vary. Securing reliable, high-quality signals that can be processed in real-time could enable use of models not addicted to big data (cf. [1]). Tuning of fCWT's hyperparameters might provide just this.

Acknowledgements

B.E.v.Z. and E.L.v.d.B. thank our partners, members from the HONDA project between the Honda Research Institute in Japan and Utrecht University in the Netherlands, for supporting this research. M.A.L. and E.L.v.d.B. have received funding for this work from ANSES under grant agreement no. PNR EST-2018/1/106. L.P.A.A. and E.L.v.d.B. have received funding for this work from the European Union's Horizon 2020 research and innovation program under grant agreement No. 952095, the IM-TWIN project. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

References

[1] Beerends JG, Neumann NMP, van den Broek EL, Llagostera A, Vranic J, Schmidmer C, et al. Subjective and objective assessment of full bandwidth speech quality. IEEE/ACM Transactions on Audio, Speech and Language Processing. 2020;28(1):440-9. doi:10.1109/TASLP.2019.2957871.

[2] van den Broek EL. Robot nannies: Future or fiction? Interaction Studies. 2010;11(2):274-82. doi:10.1075/is.11.2.16van.

[3] van den Broek EL, Schut MH, Westerink JHDM, Tuinenbreijer K. Unobtrusive Sensing of Emotions (USE). Journal of Ambient Intelligence and Smart Environments. 2009;1(3):287-99. doi:10.3233/AIS-2009-0034.

[4] van den Broek EL. Affective Signal Processing (ASP): Unraveling the mystery of emotions. Human Media Interaction (HMI), Faculty of Electrical Engineering, Mathematics, and Computer Science, University of Twente, Enschede, The Netherlands; 2011. doi:10.3990/1.9789036532433.

[5] Wani TM, Gunawan TS, Qadri SAA, Kartiwi M, Ambikairajah E. A Comprehensive Review of Speech Emotion Recognition Systems. IEEE Access. 2021;9:47795-814. doi:10.1109/ACCESS.2021.3068045.

[6] van den Broek EL, van der Sluis F, Dijkstra T. Cross-validation of bi-modal health-related stress assessment. Personal and Ubiquitous Computing. 2013;17(2):215-27. doi:10.1007/s00779-011-0468-z.

[7] Swain M, Routray A, Kabisatpathy P. Databases, features and classifiers for speech emotion recognition: A review. International Journal of Speech Technology. 2018;21(1):93-120. doi:10.1007/s10772-018-9491-z.

[8] Johnson DO, Cuijpers RH, Juola JF, Torta E, Simonov M, Frisiello A, et al. Socially Assistive Robots: A Comprehensive Approach to Extending Independent Living. International Journal of Social Robotics. 2014;6(2):195-211. doi:10.1007/s12369-013-0217-8.

[9] Zhang S, Zhang S, Huang T, Gao W. Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching. IEEE Transactions on Multimedia. 2018 Jun;20(6):1576-90. doi:10.1109/TMM.2017.2766843.

[10] Xia S, Fourer D, Audin L, Rouas JL, Shochi T. Speech Emotion Recognition using Time-frequency Random Circular Shift and Deep Neural Networks. In: Frota S, Vigário M, editors. Speech Prosody 2022. Lisbon, Portugal: Baixas, France: International Speech Communication Association (ISCA); 2022. p. 585-9. doi:10.21437/SpeechProsody.2022-119.

[11] Bhangale KB, Kothandaraman M. Survey of deep learning paradigms for speech processing. Wireless Personal Communications. 2022;125:1913-49. doi:10.1007/s11277-022-09640-y.

[12] Khalil RA, Jones E, Babar MI, Jan T, Zafar MH, Alhussain T. Speech emotion recognition using deep learning techniques: A review. IEEE Access. 2019;7:117327-45. doi:10.1109/ACCESS.2019.2936124.

[13] Koromilas P, Giannakopoulos T. Deep multimodal emotion recognition on human speech: A review. Applied Sciences. 2021;11(17):#7962. doi:10.3390/app11177962.

[14] Lieskovská E, Jakubec M, Jarina R, Chmulík M. A review on speech emotion recognition using deep learning and attention mechanism. Electronics. 2021;10(10):#1163. doi:10.3390/electronics10101163.

[15] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Communications of the ACM. 2017;60(6):84-90. doi:10.1145/3065386.

[16] Arts LPA, van den Broek EL. The fast Continuous Wavelet Transformation (fCWT) for real-time, high-quality, and noise-resistant time-frequency analysis. Nature Computational Science. 2022;2(1):47-58. doi:10.1038/s43588-021-00183-z.

[17] Mallat SG. A Wavelet Tour of Signal Processing: The Sparse Way. 3rd ed. Burlington, MA,USA: Academic Press / Elsevier Inc.; 2009.

[18] Garrido M. The Feedforward Short-Time Fourier Transform. IEEE Transactions on Circuits and Systems II – Express Briefs. 2016;63(9):868-72. doi:10.1109/TCSII.2016.2534838.

[19] Cohen L. Time-frequency analysis. Upper Saddle River, NJ, USA: Prentice Hall PTR; 1995.

[20] Griffiths DJ, Schroeter DF. Introduction to Quantum Mechanics. 3rd ed. New York, NY, USA: Cambridge University Press; 2018.

[21] Torrence C, Compo GP. A practical guide to wavelet analysis. Bulletin of the American Meteorological society. 1998;79(1):61-78. doi:10.1175/1520-0477(1998)079¡0061:APGTWA¿2.0.CO;2.

[22] Martin O, Kotsia I, Macq B, Pitas I. The eNTERFACE'05 Audio-Visual Emotion Database. In: 22nd International Conference on Data Engineering Workshops (ICDEW'06). Los Alamitos, CA, USA: IEEE Computer Society; 2006. p. #8. doi:10.1109/ICDEW.2006.145.

[23] Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B. A database of German emotional speech. In: Trancoso I, editor. Proceedings of Interspeech 2005. Lisbon, Portugal: Baixas, France: International Speech Communication Association (ISCA); 2005. p. 1517-20. doi:10.21437/Interspeech.2005-446.

[24] Shorten C, Khoshgoftaar TM. A survey on Image Data Augmentation for Deep Learning. Journal of Big Data. 2019;6(1):1-48. doi:10.1186/s40537-019-0197-0.

[25] Yin S, Liu C, Zhang Z, Lin Y, Wang D, Tejedor J, et al. Noisy training for deep neural networks in speech recognition. EURASIP Journal on Audio, Speech, and Music Processing. 2015;2015(1):1-14. doi:10.1186/s13636-014-0047-0.

[26] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision. 2015;115:211-52. doi:10.1007/s11263-015-0816-y.

[27] Efron B, Tibshirani R. Improvements on cross-validation: the 632+ bootstrap method. Journal of the American Statistical Association. 1997;92(438):548-60. doi:10.1080/01621459.1997.10474007.