JOURNAL OF MODERN SCIENCE
SPECIAL ISSUE
3/57/2024
www.jomswsge.com

DOI: doi.org/10.13166/jms/191163

MAGDALENA HAŁAS
WSEI University in Lublin, Poland
ORCID iD: orcid.org/0009-0007-4813-7507

MICHAŁ MAJ
WSEI University in Lublin, Poland
ORCID iD: orcid.org/0000-0002-7604-8559

EWA GUZ
WSEI University in Lublin, Poland
ORCID iD: orcid.org/0000-0002-0507-2172

MARCIN STENCEL
WSEI University in Lublin, Poland
ORCID iD: orcid.org/0000-0001-5807-9717

TOMASZ CIEPLAK
Lublin University of Technology, Poland
ORCID iD: orcid.org/0000-0002-2712-6098

ADVANCED EMOTION ANALYSIS: HARNESSING FACIAL IMAGE PROCESSING AND SPEECH RECOGNITION THROUGH DEEP LEARNING

ZAAWANSOWANA ANALIZA EMOCJI: WYKORZYSTANIE PRZETWARZANIA OBRAZU TWARZY I ROZPOZNAWANIA MOWY POPRZEZ GŁĘBOKIE UCZENIE

Abstract

The human face hides many secrets and is one of the most expressive human features. Human faces even contain hidden information about a person's personality. Considering the fundamental role of the human face, it is necessary to prepare appropriate deep-learning solutions that analyze human face data. This technology is becoming increasingly common in many industries, such as online retail, advertising testing, virtual makeovers, etc. For example, facial analysis technology now allows online shoppers to virtually apply makeup and try on jewelry or new glasses to get an accurate picture of what these products will look like.

The human sense of hearing is a treasure trove of information about the current environment and the location and properties of sound-producing objects. For instance, we effortlessly absorb the sounds of birds singing outside the window, traffic passing in the distance, or the lyrics of a song on the radio. The human auditory system can process the intricate mix of sounds reaching our ears and create high-level abstractions of the environment by analyzing and grouping measured sensory signals. The process of obtaining segregation and identifying sources of a received complex acoustic signal, known as sound scene analysis, is a domain where the power of deep learning shines. The machine implementation of this functionality (separation and classification of sound sources) is pivotal in applications such as speech recognition in noise, automatic music transcription, searching and retrieving multimedia data, or recognizing emotions in statements.

Streszczenie

Ludzka twarz skrywa wiele tajemnic i jest jedną z najbardziej wyrazistych cech ludzkich. Ludzkie twarze zawierają nawet ukryte informacje o osobowości człowieka. Biorąc pod uwagę fundamentalną rolę ludzkiej twarzy, należy przygotować odpowiednie rozwiązania oparte na głębokim uczeniu się, które analizują dane dotyczące ludzkiej twarzy. Dana technologia staje się coraz bardziej powszechna w wielu branżach, takich jak sprzedaż detaliczna przez Internet, testowanie reklam, wirtualne metamorfozy etc. Na przykład technologia analizy twarzy pozwala obecnie kupującym online wirtualnie nałożyć makijaż, przymierzyć biżuterię lub nowe okulary, aby uzyskać dokładny obraz tego, jak te produkty będą wyglądać w rzeczywistości.

Zmysł słuchu człowieka dostarcza wielu bogatych informacji o obecnym otoczeniu w odniesieniu do lokalizacji i właściwości obiektów wytwarzających dźwięk. Możemy na przykład bez trudu przyswoić odgłosy ptaków śpiewających za oknem, ruch uliczny

odbywający się w oddali czy też słysząc słowa piosenki w radio. Układ słuchowy człowieka jest w stanie przetwarzać złożoną mieszankę dźwiękową docierającą do naszych uszu i tworzyć abstrakcje otoczenia na wysokim poziomie poprzez analizę i grupowanie zmierzonych sygnałów sensorycznych. Proces uzyskiwania segregacji i identyfikacji źródeł odebranego złożonego sygnału akustycznego jest znany jako analiza sceny dźwiękowej. Łatwo sobie wyobrazić, że maszynowa realizacja tej funkcjonalności (separacja i klasyfikacja źródeł dźwięku) jest bardzo przydatna w zastosowaniach takich jak rozpoznawanie mowy w hałasie, automatyczna transkrypcja muzyki, wyszukiwanie i odzyskiwanie danych multimedialnych czy też rozpoznawanie emocji w wypowiedziach.

Keywords: *Multi-task learning, Deep Learning, Computer Vision, Person Classifier, Emotions Classifier*

Słowa kluczowe: *Uczenie się wielozadaniowe, głębokie uczenie, widzenie komputerowe, klasyfikator osób, klasyfikator emocji*

Introduction

Facial image processing focuses on extracting and analyzing information about human faces, which is essential in social interactions like recognition, emotion, and intention (Khaireddin & Chen, 2021a). In the past decade, it has emerged as a highly active research area dealing with face detection and tracking, facial feature detection, facial recognition, facial expression and emotion recognition, facial coding, and virtual face synthesis. Recent advances in machine learning, statistical classification methods, and complex deformable models have paved the way for numerous applications in fields such as image retrieval, surveillance and biometrics, visual speech understanding, virtual characters for e-learning, online marketing or entertainment, and intelligent human-computer interaction. However, substantial work is still required to develop more reliable systems, especially to handle pose and lighting variations in intricate real-world scenarios. While many methods predominantly focus on still images, new techniques can also process various types of inputs. For instance, video is increasingly becoming widespread and affordable, leading to a growing demand for human-centric, vision-based applications ranging from security to human-computer interaction and video annotation (Kehtarnavaz, 2008).

The basic assumption was to detect a person in the image and apply the same principle based on the person's facial expressions and voice recording. The trained model should detect a person based on a photo and then notify that it is a person from the video and voice recording. Investigation into the execution of four model directors and one multi-tasking model. The following data were used for the research: The Ryerson Audio-Visual Database of Emotional Speech and Song (. It contains audio-visual recordings of twenty-four actors with specially selected actions, providing additional facial expressions and voice timbre depending on the expected reaction. The database contains video recordings in mp4 format and audio in WAV format.

In the context of machine learning, particular learning, the method of data processing is essential for the learning results and the quality of the resulting model. Create utility models in DAT format directly from individual files saved in mp4 format. In other words, audio was turned into a spectrogram, i.e., sound in the form of reality. We are playing the lead role in a form that is friendly to machine learning algorithms. The sound we produce as a signal is the sound emitted by the air over time. Samples of characteristic air pressure over time can be taken and analyzed. The data is assumed to be sampled most often at 44.1 kHz or 44,100 seconds per second. The final captured waveform for the base can then be interpreted, modified, and analyzed.

Research Methodology

For decades, decoding emotional facial expressions has been an exciting research topic in psychology and, more recently, in human-computer interaction due to its significant commercial potential. Facial expressions intuitively reflect a person's mental state and are a common form of nonverbal communication. Most people can effectively express their personal feelings and communicate their intentions through facial expressions. Facial emotion recognition is mainly a technology designed to analyze sentiments captured by various forms of media, such as photos and videos. Built on artificial intelligence, facial emotion recognition belongs to the family of affective

processing technologies. Detects and analyzes various facial expressions to determine what emotions a person is showing.

The main idea was to detect a person in the image and determine the person's emotions based on facial expressions and the person's voice. This means the trained model should detect a person based on a photo and voice sample and the feelings accompanying a person from a video and voice recording. The study included the creation of four separate models and one multi-purpose model (Caruana, 1997).

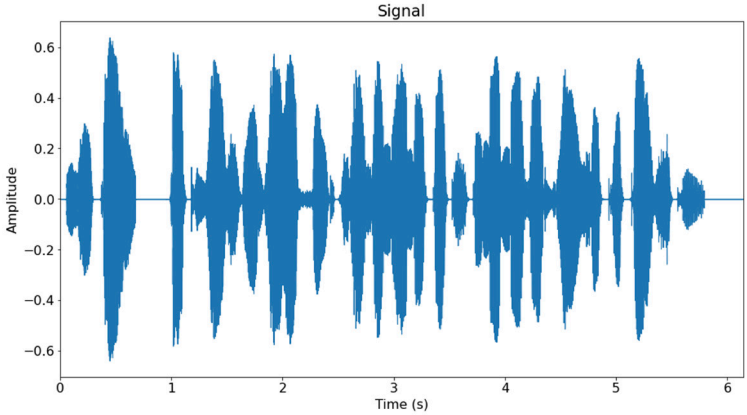
The following database was used for the research: The Ryerson Audio-Visual Database of Emotional Speech and Song. It contains audio-visual recordings of twenty-four actors who speak specially selected lines, presenting appropriate facial expressions and voice timbre, depending on the expected emotion. The database contains video recordings in mp4 format and audio recordings in WAV format (Misra et al., 2016).

Visualization of sound structure

In the context of machine learning intense learning, the data processing method is of fundamental importance for the learning efficiency and the quality of the obtained model (Rao, 2008). Considering the domain of audio data, before moving on to the model training stage itself, it is necessary to understand several steps that allow us to represent audio in a form that is friendly to machine learning algorithms. This short description will present the concept of a spectrogram, its improved version – mel spectrogram – and MFCC coefficients in the context of sound representation in images (Kehtarnavaz, 2008).

Initially, it is worth providing basic information about acquiring and processing sound. We treat the sound we hear as a signal, i.e., a change in air pressure over time. Samples of the air pressure occurring over time can be collected and analyzed by us. A specific data sampling rate is assumed, which may vary, but is most often 44.1 kHz or 44,100 samples per second. Ultimately, we captured a waveform for the signal, which we can then interpret, modify, and analyze using computer software. The figure below shows an example signal for a particular utterance (Rao, 2008).

Figure 1. *Example signal wave*

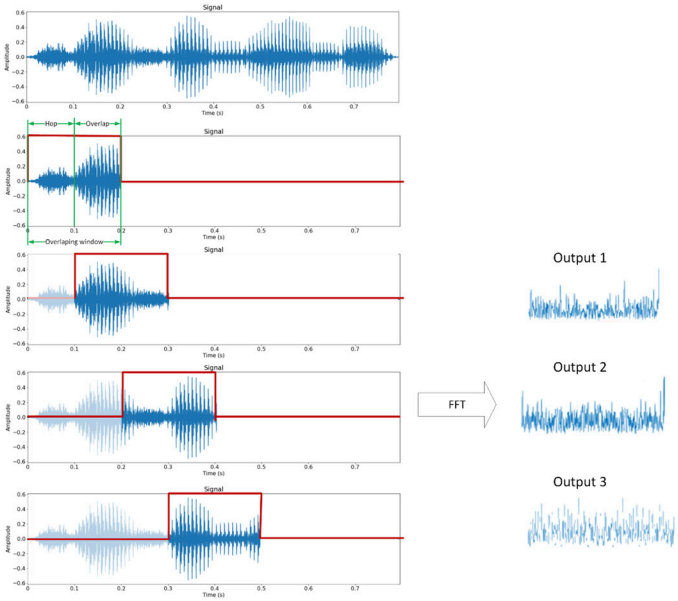


Spectrograms are a beneficial graphical representation of sound that allows you to extract additional information from audio files necessary from the point of view of machine learning. In a spectrogram, the horizontal axis represents time, the vertical axis represents frequency, and the color intensity represents the amplitude of the frequency at a specific point in time. This means that the lighter the color in the drawing, the more sound is concentrated around these particular frequencies. When the color becomes darker, we are closer to silence in the examined audio file. Therefore, the method of creating the spectrogram is an important issue (Gong et al., 2021). The next drawing shows the concept of creating a spectrogram.

This process consists of the following points:

1. Splitting the audio into overlapping windows
2. Performing a short-time Fourier transform in each window and taking the absolute value
3. Each received window has values expressed as a function of frequency – they need to be converted to decibels.
4. Rearrangement of designated windows in the duration domain of the original sound.

Figure 2. *The process of creating a spectrogram*

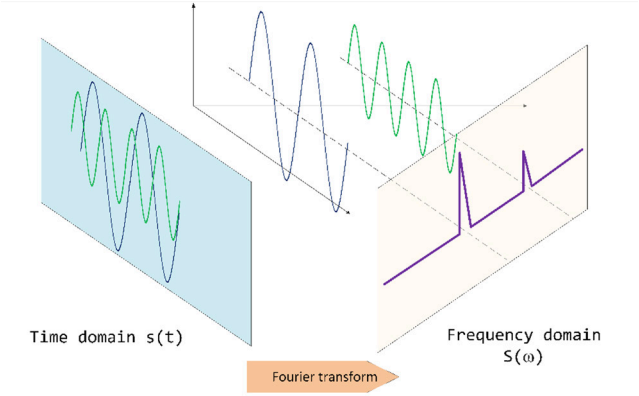


This is a simplified description of the spectrogram generation process. Of course, the most critical operation is the Fourier transform, which allows the signal to be divided into its frequencies and the amplitude of these frequencies. This allows us to move from the time domain to the frequency domain. Fourier's theorem suggests that any signal can be represented as a set of sine and cosine waves adding to the original signal. The short-time transform is a natural extension of the Fourier transform in addressing signal non-stationarity through windows in segment analysis. In the continuous case, the transform function is multiplied by a window function that is non-zero only for a short period. The resulting signal's Fourier transform (one-dimensional function) is taken, and the window is then moved along the time axis, resulting in a two-dimensional signal representation. Mathematically, this can be written as:

$$STFT\{x(t)\} = X(\tau, w) = \int_{-\infty}^{\infty} x(t)w(t-\tau)e^{-iwt}dt,$$

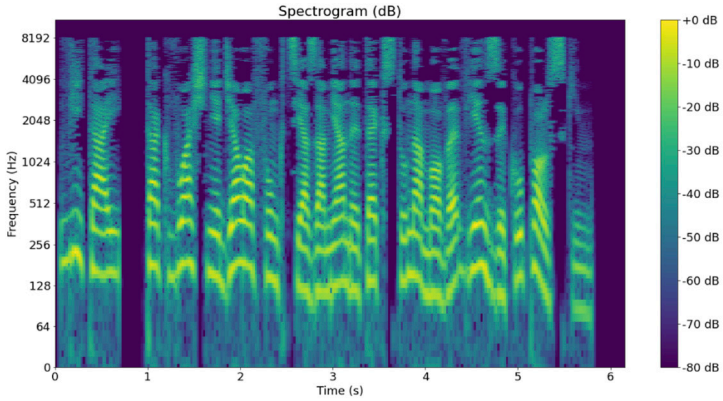
where $x(\tau)$ is the window function and is the signal to be transformed. The figure below shows the idea of transition between fields.

Figure 3. *Fourier transform*



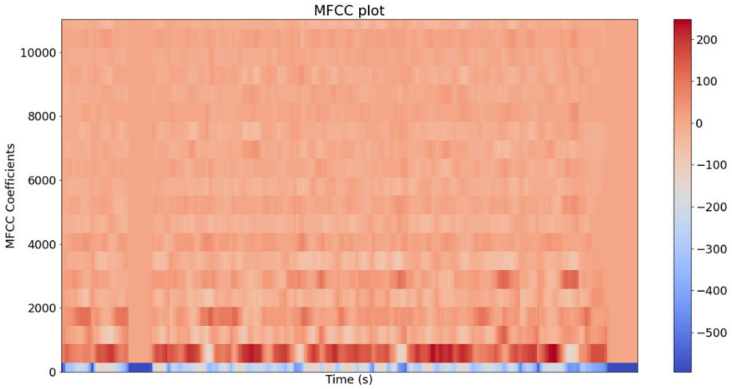
Thus, the spectrogram can be considered several FFTs stacked on each other. This is a way to visually represent the loudness or amplitude of a signal as it changes over time at different frequencies. As the spectrogram is calculated, additional details emerge behind the scenes. The y-axis is converted to a logarithmic scale, and the color dimension is converted to decibels (this can be thought of as a logarithmic scale of amplitude). Humans can only perceive a minimal and concentrated range of frequencies and amplitudes. The next drawing shows a spectrogram generated for the sound whose wave was presented in the previous drawing (Wyse, 2017).

Figure 4. *Example spectrogram*



In machine learning applications such as speech recognition or sound classification, the number of MFCC coefficients is a hyperparameter that allows you to improve the model. An example MFCC chart with 20 coefficients for the sound used in the previous sections is shown in the figure below (Muaidi et al., 2014).

Figure 5. *Example MFCC plot*



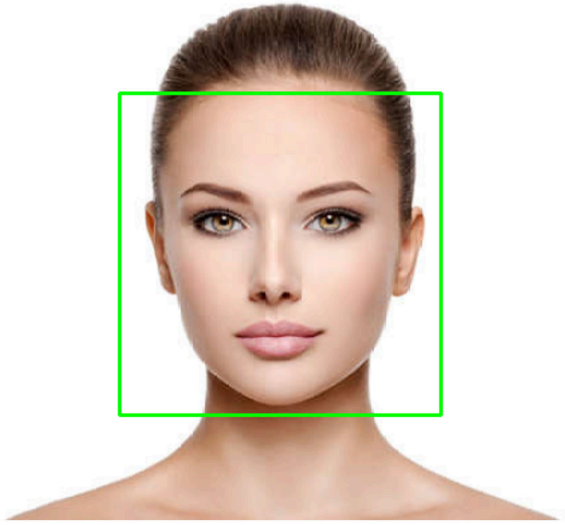
Visual emotion recognition

Facial image processing is a research field centered around extracting and analyzing data from human faces, which is crucial for social interactions such as recognition, emotions, and intentions (Wang & Deng, 2018). In the past decade, it has become a thriving research area focused on face detection and tracking, facial feature detection, facial recognition, facial expression and emotion recognition, facial coding, and virtual face synthesis (Maj et al., 2022). New machine learning techniques, statistical classification methods, and complex deformable models have recently driven advances that enable applications in image retrieval, surveillance, biometrics, visual speech understanding, virtual characters for e-learning, online marketing, entertainment, and intelligent human-computer interaction. Nonetheless, more work is needed to develop systems that are resilient to pose and lighting variations in real-world environments. While most methods focus on still image processing, newer techniques can handle various inputs. For instance,

video is becoming increasingly common and accessible, prompting a rising demand for human-centric, vision-based applications ranging from security to human-computer interaction and video annotation (Khaireddin & Chen, 2021b). Acquiring 3D data is also becoming more affordable, and processing this data can lead to better systems that are more resistant to lighting variations and enable easier extraction of discriminatory information.

Facial landmarks have been successfully applied to face alignment, head pose estimation, face swapping, blink detection, and much more. Facial landmark detection is a subset of the shape prediction problem. Given an input image (and usually an ROI that defines the object of interest), the shape predictor tries to locate critical points of interest along the shape. In the context of facial landmarks, we aim to detect important facial structures in the face using shape prediction methods.

Figure 6. *Detected face with haar cascade*



Results

Multi-Task Learning (Long et al., 2015) is one of the newest types of deep learning. It allows you to use one model for many tasks, such as segmentation, object detection, natural language processing, etc. (Liu et al., 2016). The main advantages mentioned include increased learning speed and increased inference speed. This is related to performing only one forward propagation and one back propagation. It is worth adding that this approach reduces the number of parameters. Therefore, the built model will take up less memory space. Such advantages can be obtained because the tasks are related, and learning two or more tasks can improve training. However, it is essential to remember which topics need to be taught together and which need to be taught separately (Maj et al., 2023).

Another essential feature of multi-task learning is that it shares weights. Similar to transfer learning, which is intended to help save time when the model is trained, multi-task training can shorten training time but also speed up inference (Maj et al., 2022). For example, training fifty layers three times to solve three tasks is unnecessary. Instead, you can run one neural network and change the heads. Thanks to this, a given network benefits from what other tasks learn (Ruder et al., 2017).

In this work, four models were trained separately. Each was separately responsible for classifying the person in the image, identifying the person using speech, determining the person's emotional state from the video recording, and indicating emotions in the voice based on the spectrogram. The division into the training set and the test set was 80:20. It is also worth noting that the ResNet50 network (Sarker, 2021), trained using ImageNet (Deng et al., 2010), was used to build the backbone model. Additionally, it is worth noting that the PyTorch library was used (Paszke et al., 2019). A single multi-task network model was also created to detect these four features for the experiment. Notably, a similar percentage prediction was obtained in both cases, which was 93%. The finally developed models were launched on test platforms (He et al., 2016).

Conclusions

As assumed in the thesis, the model inference time can be accelerated thanks to multi-task learning, additionally ensuring an identical prediction rate. It has also been proven that a model built using multi-task learning techniques is a better solution regardless of the platform used. The classic approach, however, works well when only one thing is required to be detected. In this case, the single model will be faster than the multitasking model.

It can be determined that further work related to multi-task networks may improve the quality of the obtained results. It is expected that additional research will improve both prediction and inference time. The results' quality should also be enhanced by combining the multi-task learning presented above with other compression and optimization techniques. One such example would be knowledge distillation.

References

Caruana, R. (1997). Multitask Learning. Machine Learning, 28(1), 41–75. https://doi.org/10.1023/A:1007379606734

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, & Li Fei-Fei. (2010). *ImageNet: A large-scale hierarchical image database*. 248–255. https://doi.org/10.1109/cvpr.2009.5206848

Gong, Y., Chung, Y.-A., & Glass, J. (2021). AST: Audio Spectrogram Transformer. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *1*, 56–60. http://arxiv.org/abs/2104.01778

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-December*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Kehtarnavaz, N. (2008). Frequency Domain Processing. In *Digital Signal Processing System Design* (pp. 175–196). Elsevier. https://doi.org/10.1016/b978-0-12-374490-6.00007-6

Khaireddin, Y., & Chen, Z. (2021a). *Facial Emotion Recognition: State of the Art Performance on FER2013*. http://arxiv.org/abs/2105.03588

Khaireddin, Y., & Chen, Z. (2021b). *Facial Emotion Recognition: State of the Art Performance on FER2013*. http://arxiv.org/abs/2105.03588

Liu, S., Pan, S. J., & Ho, Q. (2016). Distributed Multi-Task Relationship Learning. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *Part F129685*, 937–946. http://arxiv.org/abs/1612.04022

Livingstone, S. R., & Russo, F. A. (2018). *The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)*. https://doi.org/10.5281/ZENODO.1188976

Long, M., Cao, Z., Wang, J., & Yu, P. S. (2015). Learning Multiple Tasks with Multilinear Relationship Networks. *Advances in Neural Information Processing Systems*, *2017-December*, 1595–1604. http://arxiv.org/abs/1506.02117

Maj, M., Rymarczyk, T., Cieplak, T., & Pliszczuk, D. (2022). Deep learning model optimization for faster inference using multi-task learning for embedded systems. *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, 892–893. https://doi.org/10.1145/3495243.3558274

Maj, M., Rymarczyk, T., Maciura, Ł., Cieplak, T., & Pliszczuk, D. (2023). Cross-Modal Perception for Customer Service. *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, 1–3. https://doi.org/10.1145/3570361.3615751

Misra, I., Shrivastava, A., Gupta, A., & Hebert, M. (2016). Cross-Stitch Networks for Multi-task Learning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-December*, 3994–4003. https://doi.org/10.1109/CVPR.2016.433

Muaidi, H., Al-Ahmad, A., Khdoor, T., Alqrainy, S., & Alkoffash, M. (2014). Arabic audio news retrieval system using dependent speaker mode, mel frequency cepstral coefficient and dynamic time warping techniques. *Research Journal of Applied Sciences, Engineering and Technology*, *7*(24), 5082–5097. https://doi.org/10.19026/rjaset.7.903

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, *32*. http://arxiv.org/abs/1912.01703

Rao, P. (2008). Audio signal processing. In *Studies in Computational Intelligence* (Vol. 83, pp. 169–189). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-75398-8_8

Ruder, S., Bingel, J., Augenstein, I., & Søgaard, A. (2017). Latent Multi-task Architecture Learning. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, 4822–4829. http://arxiv.org/abs/1705.08142

Sarker, I. H. (2021). Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. In *SN Computer Science* (Vol. 2, Issue 6, p. 420). Springer. https://doi.org/10.1007/s42979-021-00815-1

Wang, M., & Deng, W. (2018). Deep Face Recognition: A Survey. *Neurocomputing*, *429*, 215–244. https://doi.org/10.1016/j.neucom.2020.10.081

Wyse, L. (2017). *Audio Spectrogram Representations for Processing with Convolutional Neural Networks*. http://arxiv.org/abs/1706.09559