Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/87/20241591

Sound feature analysis and gender recognition based on deep learning: A review

Chenxu Zhu

School of, University of Liverpool, UK

sgczhu10@liverpool.ac.uk

Abstract. The application of deep learning in identifying sound features has become increasingly prevalent, greatly enhancing the performance of voice applications across various professional domains. This study focuses on deep learning techniques applied to sound feature analysis and gender recognition. It reviews methodologies, datasets, and case studies, emphasizing deep learning's crucial role in boosting efficiency and accuracy. Recent advancements highlight CNN-based architectures and novel models, demonstrating deep learning for voice systems for enhanced interaction and analysis. Challenges such as computational demands and limited data availability persist, but ongoing optimizations and multi-modal approaches promise future advancements in voice technology, enabling more intelligent and responsive interactions.

Keywords: Deep learning, Sound feature analysis, Gender recognition, Mel-Frequency Cepstral Coefficients (MFCC), Voice systems.

1. Introduction

The application of deep learning in identifying sound features has emerged as a mainstream trend, significantly enhancing the performance of various voice applications and user experiences across diverse professional fields [1] [2] [3] [4]. This advancement enables the processing and recognition of extensive sound data within shorter timeframes, thereby substantially improving the accuracy of related tasks.

Sound characteristics are primarily determined by the time domain, frequency domain, and time-frequency domain. Traditional methods, such as Mel-Frequency Cepstral Coefficients (MFCC), often face limitations in recognition effectiveness and accuracy [5] [6]. In contrast, deep learning can process larger datasets and optimize associated algorithms and outcomes, rendering it indispensable for tasks such as voice type and gender recognition. Deep learning excels in automatically learning complex features from data through the construction and training of deep neural networks.

Recent studies have extensively reviewed deep learning for sound classification. Some researchers critically evaluated recent research advancements concerning small data, specifically focusing on the use of data augmentation methods to increase the data available for deep learning classifiers in sound classification, including voice, speech, and related audio signals [2]. Besides, others presented a state-of-the-art review of various convolutional neural network (CNN) approaches in the audio domain, identifying challenges for sound classification systems [3]. Some experts also investigated the architecture and applications of deep learning in audio classification, providing an extensive review of existing research on audio-based techniques and discussing current limitations while proposing

DOI: 10.54254/2755-2721/87/20241591

directions for future research in audio-based deep learning methods [4]. These reviews consistently demonstrate the efficiency and accuracy of deep learning in sound classification. CNN and Recurrent Neural Networks (RNN) are two deep learning methods frequently employed to capture intricate patterns and time-dependent relationships within sound signals. These methods streamline the cumbersome processes of feature extraction and classification inherent in traditional methods, making them highly effective for sound classification tasks.

This review aims to identify the gender of the voice owner utilizing deep learning techniques, encompassing current development prospects, challenges, and future research directions. By examining the methodologies, datasets, and practical case studies, this review seeks to illustrate the pivotal role of deep learning in sound feature recognition. The focus will be on enhancing the efficiency and accuracy of voice recognition, improving the security and completeness of fields employing this technology, and elevating the overall quality of service.

2. Literature Survey

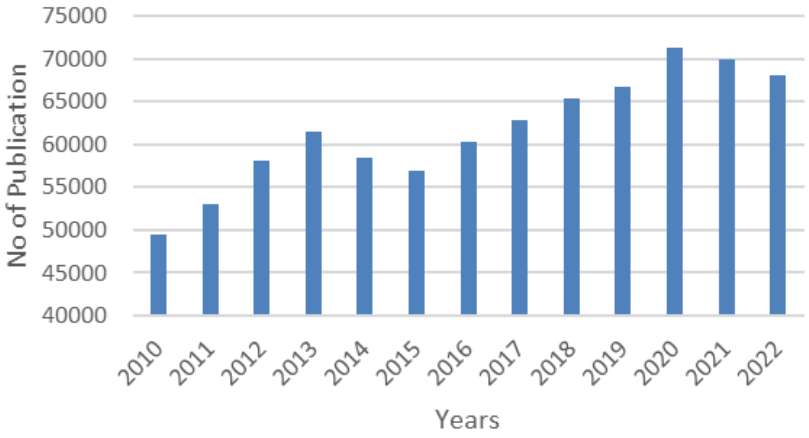

Figure 1. The number of papers searched using “deep learning” and “sound feature” per year.

Figure 1 illustrates the total number of papers retrieved using the search terms "deep learning" and "sound features" on Google Scholar from 2010 to 2022. The data reveals a gradual increase in the number of publications over this period, indicating growing research interest in this field. The number of papers rose from 49,500 in 2010 to a peak of 71,200 in 2020. This trend demonstrates that there has been consistent and increasing interest in the application of deep learning to sound features across the years.

3. Sound Features for Enhanced Gender Identification

Sound is made up of many features, each with different effects on audio processing and analysis. Mel-Frequency Cepstral Coefficients (MFCCs) are a commonly used feature in speech and audio processing, representing a scale of pitch that is perceived by the listener as aurally equal. MFCCs capture the power spectrum of a signal in a way that approximates the critical band structure of the human ear, making them highly effective in speech and speaker recognition tasks [7] [8].

Pitch, also known as the Fundamental Frequency, is another commonly used sound feature. It is a perceptual correlate of the fundamental frequency of a speech signal, conveying important information about the speaker such as intonation, stress, and emotional state. Pitch is often used in gender identification due to the different pitch ranges of men and women [9] [10].

Formants are the resonant frequencies of the vocal tract, and they differ between men and women due to variations in the vocal tract's physical structure [11] [12]. Other features, such as Spectral Features,

DOI: 10.54254/2755-2721/87/20241591

Zero-Crossing Rate (ZCR), and Linear Predictive Coding (LPC) Coefficients, also play significant roles in gender identification [13].

The process of computing MFCCs involves several steps: pre-emphasis, framing, windowing, Fast Fourier Transform (FFT), Mel filter bank processing, and Discrete Cosine Transform (DCT). Researchers can optimize MFCC performance by adjusting the number of Mel filter banks or selecting specific DCT coefficients to enhance the accuracy of gender identification [14].

Regarding Pitch and Formants, research can analyze the differences in the fundamental frequency range and formant frequencies between speakers of different genders. Spectral features, such as spectral centroid, bandwidth, and flatness, provide information about the shape of the sound spectrum. Comparing these spectral features for distribution differences is also a part of the processing. Additionally, the variation of ZCR in different speech segments, such as vowels and consonants, is an important indicator for evaluation.

4. Gender Recognition by Deep Learning

Deep learning plays a crucial role in sound feature analysis owing to its capability in processing complex data and extracting high-dimensional features. Within this domain, CNNs and RNNs are prominent architectures used for voice feature analysis and gender recognition.

CNNs, typically employed in image processing, are adept at analyzing sound features such as spectral graphs and MFCCs. Through convolutional layers, CNNs extract local features, while pooling layers effectively reduce the dimensionality of feature maps to capture essential time-frequency information [15], thereby achieving high-precision gender identification. On the other hand, RNNs excel in processing sequential data. In sound feature analysis, RNNs capture temporal sequences in speech signals, such as dynamic changes in fundamental frequency tracks, thereby enhancing gender recognition accuracy [16].

Preprocessing sound features is pivotal in deep learning pipelines as it enhances feature quality, consequently improving model accuracy and robustness. Key preprocessing techniques include:

1. Signal Denoising: Utilizing filtering and adaptive noise cancellation methods to enhance feature extraction quality and speech signal clarity.
2. Pre-emphasis: Applying a pre-emphasis filter to boost high-frequency components of voice signals [17].
3. Frame Segmentation and Windowing: Dividing speech signals into short-time frames to capture local time-frequency characteristics effectively.
4. Fourier Transform: Employing Fast Fourier Transform (FFT) to convert time-domain signals into frequency-domain representations, facilitating spectral analysis of each frame [18].
5. MFCC Extraction: Processing the spectrum through a Mel filter bank, computing the logarithmic energy output of filters, and applying Discrete Cosine Transform (DCT) to obtain MFCCs.
6. Spectrogram Generation: Using Short-Time Fourier Transform (STFT) to generate spectrograms of speech signals, facilitating CNN processing.
7. Normalization: Adjusting feature values to a consistent range to mitigate variations during model training, thereby enhancing convergence and performance [19] [20].

5. Advancements in Deep Learning for Voice Systems

AI-driven voice systems perform diverse tasks by accurately interpreting user commands, benefiting from deep learning's efficiency in feature extraction and command recognition. CNNs extract high-dimensional features, RNNs process time series features, and Transformer models with self-attention mechanisms efficiently recognize user commands and classify datasets derived from spectrograms. Self-attention mechanisms and Graph Convolutional Networks (GCNs) further enhance performance in speech recognition, music classification, and speaker recognition.

Recent research highlights various aspects of deep learning in voice systems. The researchers discussed the effectiveness of AI systems in handling simple and complex service requests, noting reduced customer complaints with prior user experience [21]. Some laboratories developed an intelligent

DOI: 10.54254/2755-2721/87/20241591

wheelchair controlled by CNN-based voice commands, aiding mobility for individuals with disabilities [22]. Besides, the expert introduced the Time-Frequency Capsule Neural Network (TFCap) for stable global information extraction from spectrograms, evaluated on the IEMOCAP database [23]. The researchers also reviewed convolutional feature extraction methods for deep neural network-based sound source localization [24]. In addition, a CNN-based approach integrating pre-processing, feature extraction, reduction, and classification stages for sound analysis had been processed by some scholars [25]. Another study demonstrated the enhancement of environmental sound classification using a convolutional RNN combined with a frame-level attention mechanism for discriminative feature learning [26]. These studies illustrate the broad application of deep learning in voice systems, encompassing device control, localization tasks, innovative methodologies, and refined environmental sound analysis.

6. Challenges and Future Directions

Deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), are complex and parameter-heavy, demanding significant computational resources and time for training [27] [28]. This poses challenges for researchers and institutions with limited computational capabilities. Moreover, these models require large volumes of labeled data, which can be particularly challenging in voice feature analysis and gender identification due to data scarcity and variations in data distribution across different application scenarios, hindering model generalization.

Despite these challenges, ongoing advancements in algorithm optimization and computing power are expected to enhance the accuracy and efficiency of deep learning models in voice feature analysis and gender recognition. Future developments may introduce more sophisticated neural network architectures capable of handling increasingly complex data, thereby pushing the field forward. Additionally, integrating multi-modal data (e.g., audio, text, and image) into deep learning models is poised to expand research horizons, aiming for more diverse and versatile applications [29].

The integration of voice feature analysis and gender recognition technologies with existing speech recognition systems holds promise for advancing voice applications [30]. Deep learning in feature analysis and the real-time processing strengths of speech recognition technology, future systems can deliver more intelligent and responsive voice interactions.

7. Summary

This review explores the pivotal role of sound features in gender identification. Recent advancements highlight CNN-based devices and innovative models, showcasing the integration of deep learning in voice systems for improved interaction and analysis. Challenges include computational demands and data scarcity, but ongoing optimizations and multi-modal approaches promise future advancements in voice technology and application integration, ensuring more intelligent and responsive voice interactions.

References

[1] Hu, H. C., Chang, S. Y., Wang, C. H., Li, K. J., Cho, H. Y., Chen, Y. T., ... & Lee, O. K. S. (2021). Deep learning application for vocal fold disease prediction through voice recognition: preliminary development study. Journal of medical Internet research, 23(6), e25247.

[2] Abayomi-Alli, O. O., Damaševičius, R., Qazi, A., Adedoyin-Olowe, M., & Misra, S. (2022). Data augmentation and deep learning methods in sound classification: A systematic review. Electronics, 11(22), 3795.

[3] Bhattacharya, S., Das, N., Sahu, S., Mondal, A., & Borah, S. (2021). Deep classification of sound: A concise review. In Proceeding of First Doctoral Symposium on Natural Computing Research: DSNCR 2020 (pp. 33-43). Springer Singapore.

[4] Wang, Y., Wei-Kocsis, J., Springer, J. A., & Matson, E. T. (2022, October). Deep learning in audio classification. In International Conference on Information and Software Technologies (pp. 64-77). Cham: Springer International Publishing.

[5] Tiwari, V. (2010). MFCC and its applications in speaker recognition. International journal on. emerging technologies, 1(1), 19-22.
[6] Mohammed, R. A., Ali, A. E., & Hassan, N. F. (2019). Advantages and disadvantages of automatic speaker recognition systems. Journal of Al-Qadisiyah for computer science and mathematics, 11(3), Page-21.
[7] Abdul, Z. K., & Al-Talabani, A. K. (2022). Mel frequency cepstral coefficient and its applications: A review. IEEE Access, 10, 122136-122158.
[8] Muda, L., Begam, M., & Elamvazuthi, I. (2010). Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. arXiv preprint arXiv:1003.4083.
[9] Sharma, G., Umapathy, K., & Krishnan, S. (2020). Trends in audio signal feature extraction methods. Applied Acoustics, 158, 107020.
[10] Hirst, D. J., & de Looze, C. (2021). Measuring Speech. Fundamental frequency and pitch. Cambridge Handbook of Phonetics, (1), 336-361.
[11] Aalto, D., Malinen, J., & Vainio, M. (2018). Formants. In Oxford Research Encyclopedia of Linguistics.
[12] Schafer, R. W., & Rabiner, L. R. (1970). System for automatic formant analysis of voiced speech. The Journal of the Acoustical Society of America, 47(2B), 634-648.
[13] Chauhan, N., Isshiki, T., & Li, D. (2019, February). Speaker recognition using LPC, MFCC, ZCR features with ANN and SVM classifier for large input database. In 2019 IEEE 4th international conference on computer and communication systems (ICCCS) (pp. 130-133). IEEE.
[14] Sidhu, M. S., Latib, N. A. A., & Sidhu, K. K. (2024). MFCC in audio signal processing for voice disorder: a review. Multimedia Tools and Applications, 1-21.
[15] O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458.
[16] Son, G., Kwon, S., & Park, N. (2019). Gender classification based on the non-lexical cues of emergency calls with recurrent neural networks (RNN). Symmetry, 11(4), 525.
[17] Schnell, K., & Lacroix, A. (2007, August). Time-varying pre-emphasis and inverse filtering of speech. In INTERSPEECH (pp. 530-533).
[18] Heckbert, P. (1995). Fourier transforms and the fast Fourier transform (FFT) algorithm. Computer Graphics, 2(1995), 15-463.
[19] Serbes, G., Ulukaya, S., & Kahya, Y. P. (2018). An automated lung sound preprocessing and classification system based onspectral analysis methods. In Precision Medicine Powered by pHealth and Connected Health: ICBHI 2017, Thessaloniki, Greece, 18-21 November 2017 (pp. 45-49). Springer Singapore.
[20] Oh, W. (2020). Comparison of environmental sound classification performance of convolutional neural networks according to audio preprocessing methods. The Journal of the Acoustical Society of Korea, 39(3), 143-149.
[21] Wang, L., Huang, N., Hong, Y., Liu, L., Guo, X., & Chen, G. (2023). Voice-based AI in call center customer service: A natural field experiment. Production and Operations Management, 32(4), 1002-1018.
[22] Sharifuddin, M. S. I., Nordin, S., & Ali, A. M. (2019, September). Voice control intelligent wheelchair movement using CNNs. In 2019 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS) (pp. 40-43). IEEE.
[23] Liu, J., Song, Y., Wang, L., Dang, J., & Yu, R. (2021). Time-Frequency Representation Learning with Graph Convolutional Network for Dialogue-Level Speech Emotion Recognition. In Interspeech (pp. 4523-4527).
[24] Krause, D., Politis, A., & Kowalczyk, K. (2021, January). Comparison of convolution types in CNN-based feature extraction for sound source localization. In 2020 28th European Signal Processing Conference (EUSIPCO) (pp. 820-824). IEEE.

DOI: 10.54254/2755-2721/87/20241591

[25] Demir, F., Turkoglu, M., Aslan, M., & Sengur, A. (2020). A new pyramidal concatenated CNN approach for environmental sound classification. Applied Acoustics, 170, 107520.
[26] Zhang, Z., Xu, S., Zhang, S., Qiao, T., & Cao, S. (2021). Attention based convolutional recurrent neural network for environmental sound classification. Neurocomputing, 453, 896-903.
[27] Frikha, M., Taouil, K., Fakhfakh, A., & Derbel, F. (2022). Limitation of deep-learning algorithm for prediction of power consumption. Engineering Proceedings, 18(1), 26.
[28] Islam, M. S., Sultana, S., Kumar Roy, U., & Al Mahmud, J. (2020). A review on video classification with methods, findings, performance, challenges, limitations and future work. Jurnal Ilmiah Teknik Elektro Komputer dan Informatika, 6(2), 47-57.
[29] Amal, S., Safarnejad, L., Omiye, J. A., Ghanzouri, I., Cabot, J. H., & Ross, E. G. (2022). Use of multi-modal data and machine learning to improve cardiovascular disease care. Frontiers in cardiovascular medicine, 9, 840262.
[30] Deng, L. (2016). Deep learning: from speech recognition to language and multimodal processing. APSIPA Transactions on Signal and Information Processing, 5, e1.