JOURNAL OF APPLIED INTELLIGENT SYSTEMS & INFORMATION SCIENCES

VOL. 4, ISSUE 1, PP. 43-49, OCTOBER 2023.
Available at: https://journal.research.fanap.com/
DOI: https://doi.org/10.22034/JAISIS.2023.409480.1065

INVESTIGATING THE EFFECT OF GENDER ON SPEECH EMOTION RECOGNITION USING THE FIRST 13 MEL-FREQUENCY CEPSTRAL COEFFICIENTS

Mahsa Ravanbakhsh [1,*], Mohammad Ravanbakhsh[2]

[1] *Department of Cognitive Linguistics, Institute for Cognitive Sciences Studies (ICSS), Tehran, Iran,*
[2] *Independent Researcher, Houston, TX.*

ABSTRACT

Speech is the most significant form of human communication for exchanging different types of information, in which words and grammar content are just one part of the message, and other types of information like age, gender, and emotional state of the speaker are also exchanged and influence the context and meaning of the message. Speech Emotion Recognition (SER) is a qualitative study of non-verbal emotion in speech's intonation. SER has an important role in human-machine interfaces and automatic service systems. This study investigates the effect of gender on SER. In this investigation, the first thirteen Mel-frequency cepstral coefficients extracted from the audio signal of emotional speech are used along with different classification algorithms. The proposed SER algorithm is trained by 85% of samples from women and men. For the testing, we used the remaining 15%. The results show slightly better accuracy in recognizing the emotions for women compared to men i.e., anger was better recognized in men, while boredom, disgust, and sadness were better recognized in women.

***KEYWORDS*:** Classification, Gender, Mel-frequency cepstral coefficients, Speech, Speech Emotional Recognition

1. INTRODUCTION

Speech is the fundamental form of human interaction for exchanging information through an audio signal. Speech can be considered as a package of integrated information; in addition to the speaker's message other types of information such as the speaker's age, gender, emotional state, and level of education are also conveyed. The study of the information content in speech has a special place among the research of speech processing, artificial intelligence, linguistics, and many other research fields. It also has many applications in various fields such as human-machine interaction.

Speech as the only verbal form of communication and information exchange is based on the syntactic combination of words carried over a speech signal (Deller et al., 1999). Speech can be studied from two directions; one is the processes of speech production and the other is the speech understanding (Luria, 1982). The motivation of speech production is the idea or message, the speaker is intended to share with his/her

* *Corresponding Author, Email: mahsa.ravanbakhsh@gmail.com*

RECEIVED: 02 AUGUST 2023; **ACCEPTED**: 26 SEPTEMBER, 2023; **PUBLISHED ONLINE**: 18 OCTOBER, 2023

audience. The speaker conveys the intended idea, message, or content through sound waves that are produced by a series of neural processes and muscle movements. The vocal tract plays an essential role in speech production and uses the mechanisms inherited from the human ancestral species (Standring, 2005). The vocal tract in humans consists of a tube at the top of the lungs with two flaps and diaphragm muscles that blow air out (Baars & Gage, 2013). This path is a tube with a turbulent air source that is regulated by the vocal cords and two flap-like tissues in the larynx, and the quality of the produced vocal sound depends on the resonance between the resonator spaces and different levels of the head and upper half of the trunk (Baars & Gage, 2013). Sound waves are transmitted to the audience's auditory system through a medium that is usually air (Deller et al., 1999; Holliday & Resnick, 1978). The process of speech perception begins when the audience collects acoustic pressure waves in the outer ear, then in the middle ear and inner ear convert them into nerve pulses, after these pulses are interpreted in the brain's auditory cortex to understand the speaker's idea (Deller et al., 1999).

The larynx and vocal cords are the source of sound production and they produce different kinds of sounds (Raphael et al., 2011). However, factors such as age, gender, and height affect the length and mass of the vocal cords, which in turn affect the pitch and loudness of the voice (Raphael et al., 2011). The length of the vocal cords in an adult woman is 13 to 17 mm and in an adult man in maximum is 17 to 24 mm (Raphael et al., 2011). These variations cause some differences in the speech signals produced between women and men.

Various environmental stimuli can affect the emotional state of a person. In addition to body gestures, facial expressions, heart rate, and breathing, changes in emotional state also appear in speech. Changes in emotional states are represented in the speech in both verbal and non-verbal forms. Verbal representation is through the vocabulary and syntax used in speech expression, and non-verbal expression is through the intonation and tone of speech. The intonation of speech has a paralinguistic role and can change the meaning of the speech (Alinezhad, 2010). The quantitative study of emotional speech is known as Speech Emotion Recognition (SER).

Recognizing or classifying the emotional content of speech means that the speaker's emotional state can be determined using the analysis performed on the speech signal [independent of the message]. The SER process consists of several steps as shown in Fig. 1. As seen in Fig. 1, speech signals as input to this model in the first step are preprocessed to be prepared for feature extraction. In the next step, the desired features are extracted from the input, and then by considering the components relevant to the goals of the research, feature selection is done to form a feature vector with appropriate dimensions. This vector is then introduced to the classifier. Finally, SER is performed. Many studies about SER show that the expression of emotional states has a complex nature that makes it hard for quantitative studies of emotion.

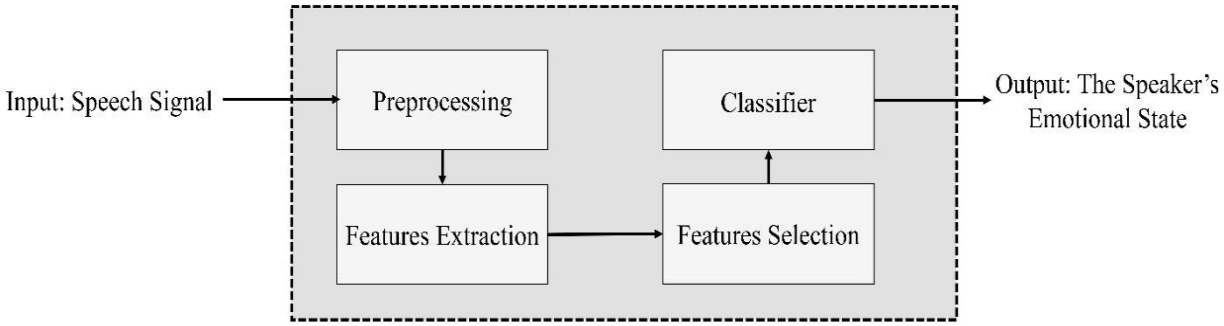

Fig. 1. This diagram shows the steps of the SER process.

Most studies about SER rely on feature extraction algorithms. Among the features used for SER, we can mention some features such as Pitch and Formant frequencies (Vergin et al., 1996; Slomka & Sridharan, 1997), Autocorrelation Coefficients, Reflection Coefficients, Linear Predictive Coefficients, Mel-frequency Cepstral Coefficients (MFCCs), and the Log Area Ratio of the cross-sectional of the vocal cords (Childers, 1991) or a combination of these characteristics (Parris and Carey, 1996). To SER, models such as Artificial Neural Networks (Yusnita et al., 2018), Hidden Markov Model (Shafran et al., 2003), Support Vector Machine (Li et

al., 2013; Kotti & Kotropoulos, 2008), Gaussian Mixed Model (Harb & Chen, 2003; Zeng et al., 2006) and Convolutional Neural Networks (Tursunov et al., 2021) are used. Another group of studies focused on improving and optimizing the algorithms presented for SER, features extracted from the speech signal, or both. The effect of gender on SER is one of the topics studied with this approach; For example, (Xu et al., 2018) investigated the effect of gender on the performance of the SER system in the phases of training and testing in a gender-dependent or gender-independent manner.

This study aims to study the effect of gender on SER. In this research, the first 13 MFCCs extracted from the audio signal of emotional speech samples of the Berlin database of emotional speech (Emo-DB) (Burkhardt et al., 2005) are used for feature extraction, and by using different classification models and algorithms SER is performed to investigate the effect of gender on SER. In the following, first, the research method and then the results obtained from the software simulation are presented and, in the end, the conclusion of the research is discussed.

2. RESEARCH METHOD

In this part, the speech database used to evaluate SER models and algorithms is introduced. The following describes how to calculate MFCCs and how to form the feature vector. In the end, the models and algorithms used to recognize and classify the emotion of speech are explained.

2.1. The Emotional Speech Database

In this study, the Emo-DB speech samples were used as input data (Burkhardt et al., 2005). This database is based on the German language and it contains 800 samples. These samples include seven emotional states: anger, boredom, disgust, fear, happiness, sadness, and neutral. More than 500 samples in this dataset have a recognition rate of over 80% and a naturalness of over 60% as reported by the creators of the database (Burkhardt et al., 2005). The speech samples were collected and prepared from 10 participants, including 5 women and 5 men. This combination of participants makes the Emo-DB a suitable and valuable option for evaluating the effect of gender on SER.

2.2. Cepstral Coefficients of Mel- Frequency

The Emo-DB samples are analyzed as input data in this study. The SER algorithm computes the first 13 MFCCs for the initial step. The MFCCs are inspired by models of sound perception and auditory properties of the human ear in receiving and understanding speech (Davis & Mermelstein, 1980). These coefficients are based on frequency analysis in the human ear (Ayat, 2008). The extraction of MFCCs includes several steps (Deller et al., 1999; Ayat, 2008). First, the speech signal is divided into overlapping windows or frames with 20ms in length. Then, using the fast Fourier transform (FFT), the Fourier spectrum of the window is obtained, and its amplitude is extracted. Next, a filter bank is placed logarithmically on the obtained spectrum as shown in Eq. (1). In the following, using these values, the MFCCs are obtained using Eq. (2), where F represents the number of filters, X_j is the output obtained from the jth filter, C_i is the ith obtained MFCCs, and N is the number of coefficients, respectively (Ayat, 2008). Here N is equal to 13 ($N < F$). These coefficients are extracted for all samples and then stored.

$$F_{mel} = 2595 log_{10}\left[1 + \frac{F_{Hz}}{700}\right] \quad (1)$$

$$C_i = \sum_{j=1}^{F} \log(X_j)\, cos\left[\frac{\pi i(j-0.55)}{F}\right] \qquad 1 \le i \le F \quad (2)$$

Fig. 2 illustrates the first 13 MFCCs extracted from two emotional speech samples of the Emo-DB which are expressed by a woman and a man, respectively. They are the same age and express the same utterance with a neutral state.

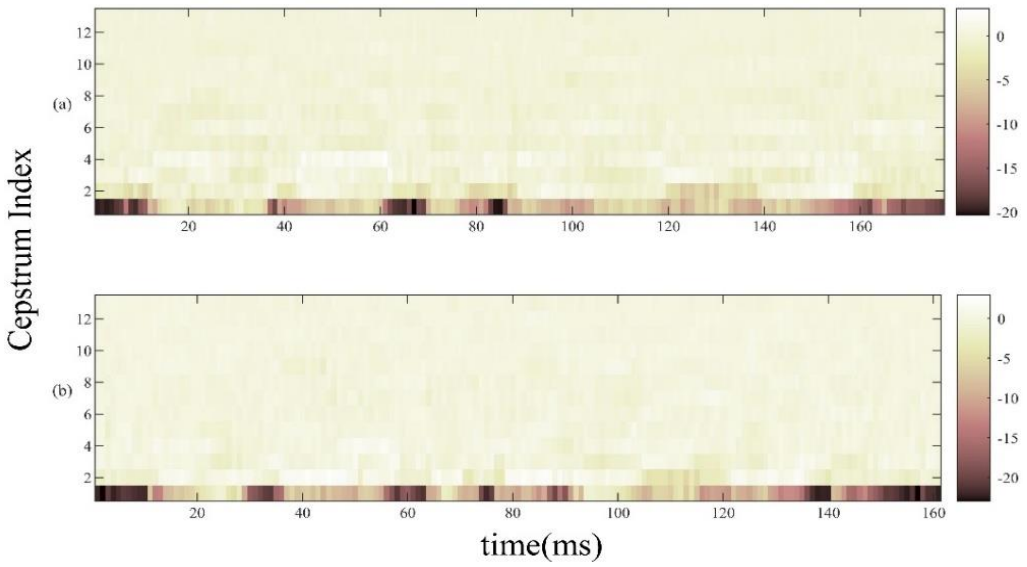

Fig. 2. The first 13 MFCCs from two emotional speech samples of the Emo-DB which were expressed by: (a) a woman and (b) a man.

2.3. Selection and Formation of Feature Vector

Due to the large dimensions size of the extracted MFCCs, they are not appropriate for use in Neural Networks. To prepare them for use in a neural network, some statistical values obtained from them are used instead. These values are *Minimum*, *Maximum*, *Mean*, *Standard Deviation*, *Median*, *Skewness*, and *Kurtosis* as also been used (Demircan & Kahramanli, 2014). The final vector obtained from the statistical computations is a feature vector with 91 values. Fig. 3 better illustrates how the process of reducing the size is done.

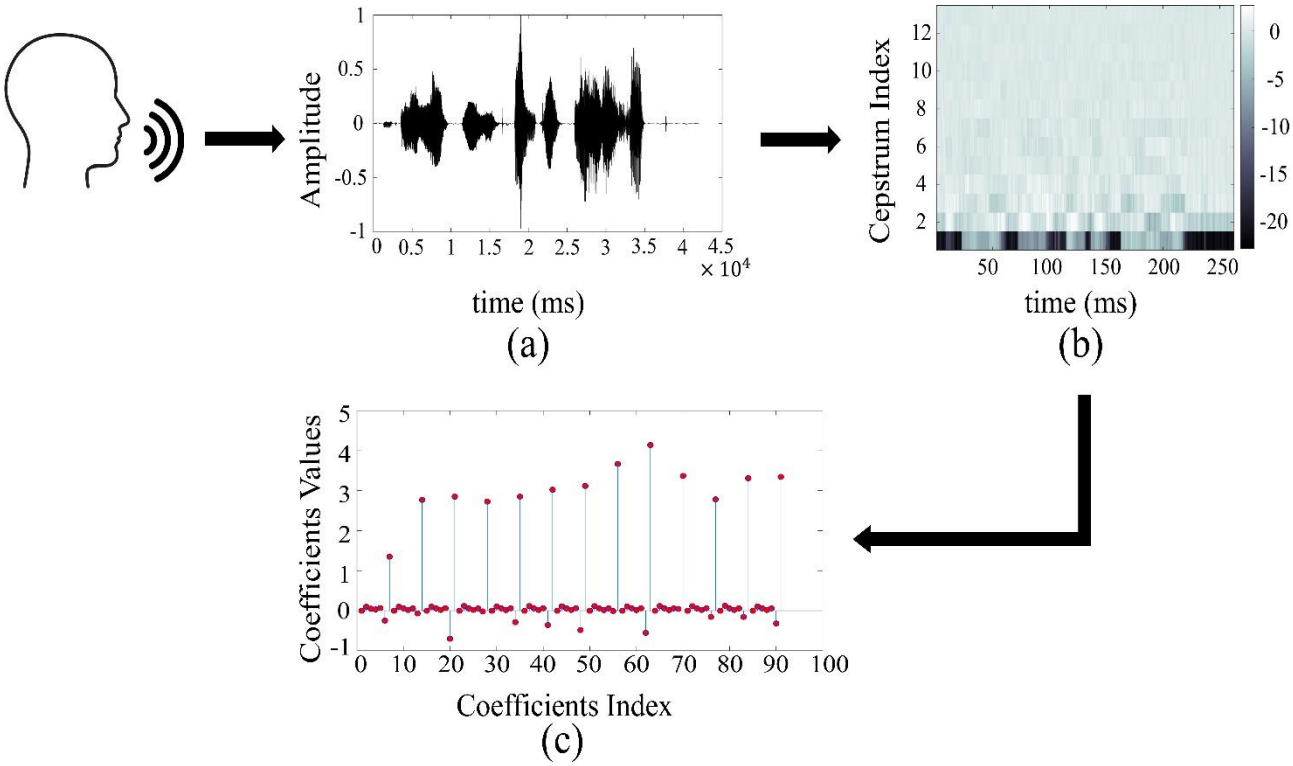

Fig. 3. The dimension reduction procedure that forms the feature vector from the first 13 MFCCs: (a) the Acoustic signal of the emotional speech sample, (b) The first 13 MFCCs of the acoustic signal of the emotional speech, and (c) The statistical values calculated from the first 13 MFCCs of the emotional speech acoustic signal.

2.4. SER Algorithms

In this study, five algorithms/models with supervised learning were applied to classify the emotional states. These models are Ensemble-Subspace Discriminant (E-SD), Linear SVM, SVM Kernel, Wide Neural Network, and Cubic KNN. The machine learning toolbox from MATLAB software (version 2021b) has been used for these classification models.

To train our model, we used 85% of the dataset. The remaining 15% was reserved for testing. We ensured that the test set contained an equal number of samples for both men and women.

3. RESULTS

The results obtained for the training and testing phases are presented in Tables 1, 2, and 3, respectively. Table 1 shows the results obtained from the training phase for different classification algorithms. Table 2 shows the results obtained from the test phase for different classification algorithms using the emotional speech datasets expressed by women. As seen in Table 2, Linear SVM has a better recognition accuracy rate than other used classification algorithms. Table 3 shows the results obtained from the test phase for different classification algorithms using the set of emotional speech data expressed by men. In Table 3, E-SD has a better recognition accuracy rate than other used classification algorithms. Based on the numbers from Tables 2 and 3, Linear SVM has a better recognition accuracy rate.

The results presented in Tables 2 and 3 show that in most cases, emotional states in emotional speech expressed by women are recognized more accurately than for men. But in total, it can be seen that the emotional state of anger in the emotional speech expressed by men has a better recognition rate than for women, and the three emotional states of Boredom, Disgust, and Sadness are better identified in women.

Table 1. The success rate in percentage of the results obtained from the training phase of SER algorithms using the data collection expressed by women and men.

Model Algorithm	Anger	Boredom	Disgust	Fear	Happiness	Sadness	Neutral	Acc.
E-SD	57.1	48.8	50.0	47.7	34.6	71.7	48.4	51.9
Linear SVM	54.9	51.4	57.7	49.1	32.8	75.0	45.6	51.6
SVM Kernel	51.4	49.1	40.0	37.3	28.9	54.2	30.6	43.5
Wide Neural Network	56.2	46.3	30.0	35.2	29.2	53.7	28.6	41.5
Cubic KNN	39.4	31.2	10.0	40.4	14.9	65.8	22.2	34.7

Table 2. The success rate in percentage of the results obtained from the test phase of the SER algorithms using the set of data expressed by women.

Model Algorithm	Anger	Boredom	Disgust	Fear	Happiness	Sadness	Neutral	Acc.
E-SD	50.0	46.2	66.7	66.7	28.6	100	66.7	53.3
Linear SVM	50.0	50.0	75.0	80.0	60.0	100	60.0	62.2
SVM Kernel	58.3	75.0	100	60.0	0.0	66.7	44.4	57.8
Wide Neural Network	54.5	41.7	60.0	66.7	50.0	100	40.0	53.3
Cubic KNN	46.2	33.3	100	75.0	0.0	100	42.9	44.4

Table 3. The success rate in percentage of the results obtained from the test phase of the SER algorithms using the set of data expressed by men.

Model Algorithm	Anger	Boredom	Disgust	Fear	Happiness	Sadness	Neutral	Acc.
E-SD	88.9	16.7	100	66.7	50.0	75.0	57.1	62.9
Linear SVM	66.7	20.0	0.0	71.4	0.0	75.0	60.0	57.1
SVM Kernel	64.3	20.0	0.0	80.0	0.0	100	50.0	57.1
Wide Neural Network	46.2	0.0	50.0	100	25.0	50.0	50.0	57.1
Cubic KNN	71.4	0.0	0.0	20.0	25.0	60.0	33.3	45.7

4. CONCLUSION

SER is one of the problems that has been studied by researchers from different aspects. In these studies, the choices of features extracted from the speech signal and the algorithms used for SER played an important role, and many studies are around these choices and algorithms. In this paper, we studied the effect of gender on SER. For this purpose, we used the emotional speech samples available in the Emo-DB to evaluate SER algorithms. In the first step, for all samples in the Emo-DB, the first 13 MFCCs were extracted, after those seven statistical values including Minimum, Maximum, Mean, Standard Deviation, Median, Skewness, and Kurtosis were calculated for each of the 13 available coefficients. Then, the feature vector for all examples of the Emo-DB was formed. For the testing, 15% of the samples for women in the database were randomly selected. The same was done for men. The remaining 85% of the data was used for the training phase of SER algorithms. In this article, we used five classification algorithms E-SD, Linear SVM, SVM Kernel, Wide Neural Network, and Cubic KNN.

The obtained results show that in most cases, emotional speech recognition is performed better for women. It was also observed that the recognition of anger was better for men. In the case of women, the recognition was better for boredom, disgust, and sadness. From these results, we concluded that gender plays a significant role in the effectiveness of non-verbal emotional expressions, and the accuracy of the recognition of algorithms was influenced by this parameter. In this study, a small sample was used and it would be interesting to see how these results scale and stay the same with larger datasets. Also, this study only considered the sample from a German language dataset, it would be interesting to expand the study across multiple languages and cultures.

APPENDIX

The Data Availability: The data used for this study is obtained from the Berlin database of emotional speech. This database is freely available to the public. More information about this database is available at: http://emodb.bilderbar.info/index-1280.html

REFERENCES

Alinezhad, B. (2010), A Study of the Relationship between Acoustic Features of "bæle" and the Paralinguistic Information, *Journal of Teaching Language Skills*. Shiraz, 2(1), 1–26.

Ayat S. (2008). *Fundamentals of speech signal processing*, Tehran: Payame Noor University Press. (Persian)

Baars, B., & Gage, N. (2013). *Fundamentals of Cognitive Neuroscience: A Beginner's Guide*. Academic Press.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B., Berlin, T. U., … Berlin, H. U. (2005). A Database of German Emotional Speech. *Conference of the International Speech Communication Association (INTERSPEECH)*, 1517–1520. https://doi.org/10.21437/interspeech.2005-446

Childers, D. G. (1991). Gender recognition from speech. Part I: Coarse analysis. *Journal of the Acoustical Society of America*, *90*(4), 1828–1840. https://doi.org/10.1121/1.401663

Davis, S. B., & Mermelstein, P. (1980). Comparison of Parametric Representations for. *Transactions on Acoustics, Speech, and Signal Processing*, *28*(4), 357–366.

Deller, Jr. John R., Hansen, John H.L. and Proakis, John G. (1999). *Discrete-Time Processing of Speech Signals*, Wiley-IEEE Press, Classic Reissue.

Demircan, S., & Kahramanli, H. (2014). Feature Extraction from Speech Data for Emotion Recognition. *Journal of Advances in Computer Networks*, *2*(1), 28–30. https://doi.org/10.7763/jacn.2014.v2.76

Halliday, D. and Resnick, R. (1978). *Physics*, John Wiley & Sons.

Harb, H., & Chen, L. (2003). Gender identification using a general audio classifier. *Proceedings - IEEE International Conference on Multimedia and Expo*, *2*, II733–II736. https://doi.org/10.1109/ICME.2003.1221721

Kotti, M., & Kotropoulos, C. (2008). Gender classification in two Emotional Speech databases. *Proceedings - International Conference on Pattern Recognition*. https://doi.org/10.1109/icpr.2008.4761624

Li, M., Han, K. J., & Narayanan, S. (2013). Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. *Computer Speech and Language*, 27(1), 151–167. https://doi.org/10.1016/j.csl.2012.01.008

Luria, A. R. (1982). *Language and Cognition*. Wiley.

Parris, E. S., & Carey, M. J. (1996). Language independent gender identification. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2*, 685–688. https://doi.org/10.1109/icassp.1996.543213

Raphael, L. J., Borden, G. J., & Harris, K. S. (2011) S*peech Science Primer: Physiology, Acoustics, and Perception of Speech*, Lippincott Williams & Wilkins.

Shafran, I., Riley, M., & Mohri, M. (2003). Voice signatures. *2003 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2003*, 31–36. https://doi.org/10.1109/ASRU.2003.1318399

Slomka, S., & Sridharan, S. (1997). Automatic gender identification optimized for language independence. *IEEE Region 10 Annual International Conference, Proceedings/TENCON*, *1*, 145–148. https://doi.org/10.1109/tencon.1997.647278

Standring, S. (Ed.), (2005). *Gray's anatomy: The anatomical basis of clinical practice*. (39th ed.). Edinburgh: Churchill Livingstone.

Tursunov, A., Mustaqeem, Choeh, J. Y., & Kwon, S. (2021). Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms. *Sensors*, *21*(17). https://doi.org/10.3390/s21175892

Vergin, R., Farhat, A., & O'Shaughnessy, D. (1996). Robust gender-dependent acoustic-phonetic modeling in continuous speech recognition based on a new automatic male/female classification. *International Conference on Spoken Language Processing, ICSLP, Proceedings*, *2*, 1081–1084. https://doi.org/10.1109/icslp.1996.607793

Xu, Z., Meyer, P., & Fingscheidt, T. (2018). On the effects of speaker gender in emotion recognition training data. *Speech Communication - 13th ITG- Symposium*, 61–65.

Yusnita, M. A., Hafiz, A. M., Fadzilah, M. N., Zulhanip, A. Z., & Idris, M. (2018). Automatic gender recognition using linear prediction coefficients and artificial neural network on the speech signal. *Proceedings - 7th IEEE International Conference on Control System, Computing and Engineering, ICCSCE 2017*, *2017-November* (November), 372–377. https://doi.org/10.1109/ICCSCE.2017.8284437

Zeng, Y. M., Wu, Z. Y., Falk, T., & Chan, W. Y. (2006). Robust GMM-based gender classification using pitch and RASTA-PLP parameters of speech. *Proceedings of the 2006 International Conference on Machine Learning and Cybernetics*, *2006*(August), 3376–3379. https://doi.org/10.1109/ICMLC.2006.258497