ISSN 1870-4069

An Overview of Gender Recognition Systems Using Acoustic Voice Analysis

Enrique Díaz-Ocampo, Andrea Magadán-Salazar,
Raúl Pinto-Elías, Máximo López-Sánchez

Tecnológico Nacional de México CENIDET,
Cuernavaca, Morelos,
Mexico

{m22ce002,andrea.ms,raul.pe,maximo.ls}@cenidet.tecnm.mx

Abstract. Intelligent learning environments are a category of academic software that implements artificial intelligence techniques to assist a student in learning. However, to have a greater impact on their learning, it is required to personalize the learning process. For this it is necessary to determine certain characteristics and traits of the student. One example of these characteristics that can be inferred from the learner's speech is gender. Gender differences in voice presents the essentialist and the constructivist view. The essentialist view explains the differences between male and female voices from anatomical differences, and the second one from socially learned behavior. Voice gender recognition systems is a term that refers the automatization of gender detection by an acoustic signal of voice. Most voice gender recognition systems only employ the essentialist approach of gender. The present article is an overview of the voice gender recognition system's development. From some classic models developed in the nineties, up to those presented in recent years. These systems developed five characteristics over the years, that will be exposed in this work and could be implemented as auxiliary tools in intelligent learning environments in order to extract features from students.

Keywords: Voice gender recognition, voice activity detector, ensembled methods, pitch, fundamental frequency.

1 Introduction

Humans developed a biological and complex system, that allow them to obtain paralinguistic information (emotion, gender, age, health, pitch, etc.) from voice. However, the process of automating this task has required several years of biological, mathematical, and computational research.

From the earliest works [1, 2] in the 1990s to the present year, there have been multiple designs of speech-based gender recognition systems that implement artificial intelligences. Such is the case that intelligent assistants personalize themselves based on the tone of the speaker. But, there have been few overviews of these systems (see [3, 4]).

pp. 7–15; rec. 2022-09-16; acc. 2022-09-30

In particular, overviews of these systems are required that expose the type of features and computational tools that can be used to extract them and thus can be used in different environments like Intelligent learning environments.

In this paper, an overview of gender recognition systems using acoustic voice analysis was carried out. The methodology for summarizing is based on [5]. The relevance of this topic lies in the fact that the extraction of paralinguistic information from the user is necessary to be able to design systems that achieve a higher degree of affinity with the user such as those used in intelligent learning environments. In this way, new academic proposals can be explored and those already in use can be evaluated according to the characteristics of the learner.

This paper is organized as follows. In Section 2 the description of the research protocol will be presented, where 4 questions relevant to the research of voice gender recognition systems will be presented. Subsequently, the selection and extraction of documents in two search engines: Google Scholar and Science Direct is presented. In Section 3 the results obtained are discussed and answers to each of the questions are provided by means of a detailed explanation based on the articles found. Finally, in section 4 will provide the most relevant aspects of this work and future work.

2 Description of the Research Protocol

The research protocol is based on the search for information to answer the following guiding questions:

- What are the biological features used in gender recognition systems by voice?
- What characteristics determine the design of a voice gender recognition system?
- What phases are implemented voice gender recognition systems?
- What are the current trends in voice gender recognition systems and what aspects characterize them?

The first was conceived because there are various combinations of characteristics for gender detection. However, if biological characteristics are used, certain traits of a person can be determined beyond their gender. The second question is about recent architectures in gender recognition systems. Finally, the third question is about the breakdown of the phases implemented in this type of systems.

2.1 Document Selection and Data Extraction Process

This section describes the search criteria and tools used in the research. Two search engines were used: Google Academic and Science Direct. The breakdown of the class words used is presented below:

- (“Acoustic analysis” OR “análisis acústico”) AND (“voice analysis” OR “análisis de voz”) AND (“gender” OR “género”).
- ((“Acoustic analysis” OR “análisis acústico”) OR (“voice analysis” OR “análisis de voz”)) AND (“gender” OR “género ”).
- “Deep Learning Gender Classification” OR “voice gender classification”.

- ("voice" AND "gender" AND "score") AND "fuzzy".
- ("voice" AND "gender" AND "score") AND "machine learning".
- "Voice" AND "background noise" AND "gender recognition" AND "analysis".
- "Voice" AND "background noise" AND "speaker gender recognition".
- "Voice" AND "noise" AND "gender recognition" AND "uncontrolled".
- "Voice" AND "noise" AND "speaker gender recognition".

The criteria for elimination were as follows:

- Papers published from 2018 to 2022.
- Only peer-reviewed articles.
- Only articles in Spanish and English.
- Available in pdf format.
- Articles that will show the architecture of their systems.
- Articles using public or accessible databases.
- Articles that will use free software or similar for its processing.

3 Results

Using the criteria established in the present work, 31 articles relevant to the present work were found. They were subsequently reviewed in terms of both their contributions and their references in order to provide information for the research questions posed in this study.

3.1 Biological Features in Voice Gender Recognition Systems

During the 1990s, the first investigations consisted of investigating the correlation of acoustic measures of laryngeal activity and the phonetic structure of healthy people. In [6], the latter was investigated by choosing various vowels of the English language pronounced by a group of 8 people (4 women and 4 men) between 20 and 45 years old.

Acoustic measurements: Fundamental Frecuency or F_0 (frequency of vibration of the vocal folds), Jitter (Frequency variation of F_0), Shimmmer (mean percent change in waveform amplitude amount period), Signal-to-noise ratio (decibel ratio of total energy in the acoustic speech signal to the energy in the aperiodic or noise component), Harmoninc-to-noise-rato (ratio between periodic and non-periodic components of a speech signal, articulate rate (number of syllabes per second), number and duration of speech pauses. The most representative results to distinguish the gender were F_0 and jitter.

Despite the studies focused on the fundamental frequency for the detection of gender, in general, it is not a decisive factor. This is due the characteristics of flexibility of the vocal cords and dimensions of the vocal tract. In the 2000s, several parameters were defined to be able to make the gender distinction between people of different ages and health conditions [7]. Here are some examples:

- Prosodic: accent, stress, rhythm, tone, pitch, and intonation.
- Acoustic: Fundamental frecuency, mean of frequency spectrum, standard deviation of frequency, duration, formant frequencies, etc.

– Cepstral: Coefficients of linear, Mel, Mel inverse and rectangular frequencies, among others.

In the 2010's, more sophisticated features like formant frequencies, aperiodicity, and spectrum level were used to study voice gender perception [8]. Nowadays, the acoustic voice analysis area is in charge of developing non-invasive techniques to study multiple acoustic parameters obtained by special microphones [9].

3.2 Systems of Gender Recognition through Voice Analysis

One way to understand the development of gender recognition systems through acoustic voice analysis is from the following points:

- Characteristics used in the classifier.
- Voice recording and multi-label environment.

The acoustic characteristics are analyzed with classical statistical techniques. For example, scatter plots, probability distributions and correlation matrices are used for each gender feature set [10]. Also, each acoustic parameter is given a weight w obtained by the implemented classifiers [10]. Futhermore, papers focused on acoustic features use multiple classifiers and ensemble techniques to solve the classification problem [11]. This is a heuristic method to avoid overfitting.

On the other hand, cepstral features are common in deep learning algorithms. In [12], multiple combinations were explored in a multilayer perceptron neural network. While in [13], multiple normalization techniques for cepstral coefficients were compared to improve classifier metrics. In addition, due to their vector form, they have been implemented with tensor analysis [14]. However, the trends in recent years have been to design more sophisticated networks (e.g., temporally convolutional) and using several thousand coefficients to improve the accuracy of gender recognition [15].

The voices used by a gender recognition system can be in controlled and uncontrolled environments. In the first case, specialized microphones are used, each audio lasts the same length and each voice reproduces a specific text [9, 16]. In the second case, the voices come from public voice sets, where each user can upload his own recorded audio. An example of such databases are Mozilla Common Voice Dataset [17].

3.3 Phases of a Voice Gender Recognition Systems

The phases of a classical voice gender recognition system consist of the choice of a dataset, a preprocessing phase, a data extraction phase, obtaining the feature vectors, the implementation of the classifier and the evaluation of the metrics. The first one process consists of removing as much noise as possible from our samples in order to obtain clean signal for further processing.

Then, feature extraction is performed. Among the most popular are the mel frequency cepstral coefficients (MFCC) [16] and acoustic parameters [18]. Since some features has more relevant information that the others, it is necessary to implement a feature selection [19]. The feature vectors feed the classifier or assembly techniques [20]. Finally, the metrics will depend on the condition of balanced or unbalanced dataset.

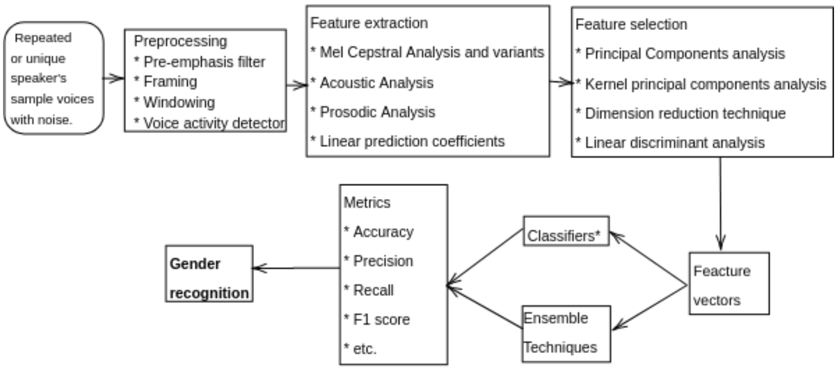

Fig. 1. General design of a gender recognition system.

3.4 General Trends in Gender Recognition Systems

A basic gender recognition system is compound by two elements [21]. The first element is a front-end system, that is designed for extract the relevant features in order to create a template model of voice's gender. The second element is a back-end system, which is the classifier algorithm implementation.

The first models for gender recognition in the 1990's, were based on vowel's analysis via statistical, probabilistic, and multilayer perceptrons models [22]. In 2000's, the trend was developing in systems with multiple probabilistic classifiers, mixed parameters and specific acoustic features. For the 2010s, parameters were built that allowed to distinguish the gender of a person but at the same time were not conditioned to language [23].

Current voice gender has multiple and complex modules. Figure 1 presents in general terms the current designs of voice gender recognition systems. In the following subsections it is discussed in depth and how to meet these requirements.

3.5 Five Main aspects in a Voice Gender Recognition System

Sample Voices. The datasets to be used for training and learning validation require handling the scenarios of repeated voices and unique voices. Since a database with multiple audios of a single person will present a bias towards the gender of that person, but a database of unique voices is more complicated to analyze since they are not common or free to use.

Existing ones generally have more voices of one gender than another (e.g. Mozilla Common Voice). So an analysis in these environments will better expose the strengths and weaknesses of both the features and the classifier algorithm. On the other hand, English language has been one of the most studied in terms of gender detection.

This is due to the availability of databases in that language. However, voice gender detection results have been obtained in multiple languages [24], and in dialects of the same language. Therefore, it is necessary to use databases with a variety of languages used in order to find patterns in them.

Preprocessing. The speech signal can be divided as the sum of a pseudo-periodic signal and a noise signal. It is usually assumed that such noise is stationary (i.e. statistical parameters are constant), so that various techniques can be employed to attenuate it [25]. One example is the use of different filters like pre-emphasis [26, 24]. Which is a time-domain finite impulse response filter with one free parameter.

Framing consists of dividing our signal into parts (intervals from 20 ms to 30 ms), so that the pseudo-periodic part of the speech signal and the stationary property of the noise signal are preserved. These segments are known as frames. For practical purposes, the pre-emphasized signal is blocked into 200 frames each 25 ms long with 10ms frame shift [26].

Windowing is the process of applying a window function to each frame obtained. A more sophisticated way to remove noise is by detecting audio segments where there is a human voice. This technique is known as voice activity detector [16, 27]. It consists of assuming that noise is a random variable that is added to our signal. Therefore, by means of hypothesis tests in each of the frames of our signal, the voice is detected.

Feature Extraction and Feature Selection. Feature extraction is associated with the type of voice emitted by the speakers. If the audios come from controlled environments and what was recorded consists of phonemes, prosodic features are often used. It is also possible to construct multiple statistics from countour pitch [28].

While in uncontrolled environments where there is a greater influence of noise, cepstral features are mostly used because of their robustness to noise [29]. Once the characteristics of interest have been obtained, a classic way of choosing the most relevant ones is from a correlation matrix. However, techniques such as principal component analysis and its variant with kernel modification are also used.

Classifiers and Metrics. Due to the flexibility of the vocal cords, as well as the dimensions of the vocal tract of each person, it is possible that there are voices of different genders but that share a similarity in their vector of characteristics. This causes the classifier to be wrong. This can be made worse if the classifier used assumes that the feature space is linearly separable when it is not.

Therefore, for any machine learning algorithm or deep learning algorithm to be used, it is essential to assume a priori that any set of features derived or associated to F_0 will not be linearly separable. With respect to metrics,since is binary classification problem, precision is mainly used in balanced data and F1-measure is used for unbalanced dataset.

Capable of Recognizing Multi-labels. The voice is subject to change over time. This is visualized in the maturation of the voice in teenagers and its aging in older adults. Therefore gender detection is affected by these conditions. One way to solve the problem is the implementation of systems that detect age and gender independently [30].

A second way is by implementing multiple tags indicating gender and age range. In [31] a convolutional neural network was implemented for the detection of gender and age range and the labels young male, young female, adult male, adult female, senior male, and senior female were implemented.

In the case of [32], a convolutional neural network with a Specially Designed Multi-Attention Module through Speech Spectrograms was implemented for the detection of gender and age in English and Korean. The labels teens, twenties, thirties, fourties, fifties, and sixties were used. Each of these was associated with the male or female gender.

4 Conclusion

In this study, a brief discussion of the development of gender recognition systems was presented. The types of features for such recognition were presented. This work showed the methodology for the search and selection of articles, as well as an exposition of the different characteristics of these systems.

In addition, an overview of the general panorama was provided in the computational study of gender detection by voice. On the other hand, five characteristics of the current design trend of these systems to improve their performance in various contexts were exposed. Future work would include systematic reviews of multi-paralinguistic recognition systems for gender, age, accent and emotion. Based on these four features, it is posible to obtain a broader picture of the speaker interacting with the system.

It is expected that this work will facilitate the implementation of the recognizer in systems associated with intelligent learning environments with the objective of obtaining features that allow the personalization of student learning.

References

1. Childers, D. G., Wu, K.: Gender recognition from speech. Part I: Coarse analysis. The Journal of the Acoustical Society of America, vol. 90, no. 4, pp. 1828–1840 (1991)
2. Childers, D. G., Wu, K.: Gender recognition from speech. Part II: Fine analysis. The Journal of the Acoustical Society of America, vol. 90, no. 4, pp. 1841–1856 (1991)
3. Zhao, H., Wang, P.: A short review of age and gender recognition based on speech. In: 2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data an Security (IDS), pp. 183–185 (2019)
4. La Mura, M., Lamberti, P.: Human-machine interaction personalization: A review on gender and emotion recognition through speech analysis. In: 2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT, pp. 319–323 (2020)
5. Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., Khalil, M.: Lessons from applying the systematic literature review process within the software engineering domain. Journal of Systems and Software, vol. 80, no. 4, pp. 571–583 (2007)
6. Nittrouer, S., McGowan, R. S., Milenkovic, P. H., Beehler, D.: Acoustic measurements of men's and women's voices: A study of context effects and covariation. Journal of Speech, Language, and Hearing Research, vol. 30, no. 4, pp. 761–775 (1990)
7. Bellandese, M. H.: Fundamental frequency and gender identification in standard esophageal and tracheoesophageal speakers. Journal of Communication Disorders, vol. 42, no. 2, pp. 89–99 (2009)

8. Skuk, V. G., Schweinberger, S. R.: Influences of fundamental frequency, formant frequencies, aperiodicity, and spectrum level on the perception of voice gender (2014)
9. Bensoussan, Y., Pinto, J., Crowson, M., Walden, P. R., Rudzicz, F., Johns III, M.: Deep learning for voice gender identification: Proof-of-concept for gender-affirming voice care. The Laryngoscope, vol. 131, no. 5, pp. E1611–E1615 (2021)
10. Ertam, F.: An effective gender recognition approach using voice data via deeper LSTM networks. Applied Acoustics, vol. 156, pp. 351–358 (2019) doi: 10.1016/j.apacoust.2019.07.033
11. Sahar, R., Rao, T., Anuradha, S., Rao, B.: Performance analysis of ML algorithms to detect gender based on voice. Recent Trends in Intensive Computing, pp. 163–171 (2021) doi: 10.3233/APC210192
12. Kanani, I., Shah, H., Mankad, S. H.: On the performance of cepstral features for voice-based gender recognition. Information and Communication Technology for Intelligent Systems, pp. 327–333 (2018)
13. İleri, S., Karabina, A., Kılıç, E.: Comparison of different normalization techniques on speakers' gender detection. Mehmet Akif Ersoy Üniversitesi Uygulamalı Bilimler Dergisi, vol. 2, no. 2, pp. 1-12 (2018) doi: 10.31200/makuubd.410625
14. Roy, P., Bhagath, P., Das, P.: Gender detection from human voice using tensor analysis. In: Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL), pp. 211–217 (2020)
15. Nandan, V. G., Shivakumar, S., Sangeetha, J., Nayak, M. P., SK, N.: A comparative study of deep learning and machine learning approaches in speech emotion and gender recognition system. Natural Volatiles and Essential Oils, vol. 8, no. 5, pp. 12261–12273 (2021)
16. Alkhawaldeh, R. S.: DGR: Gender recognition of human speech using one-dimensional conventional neural network. Scientific Programming, vol. 2019 (2019)
17. Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., Weber, G.: Common voice: A massively-multilingual speech corpus. In: Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 4218–4222 (2020)
18. Kacamarga, M. F., Cenggoro, T. W., Budiarto, A., Rahutomo, R., Pardamean, B.: Analysis of acoustic features in gender identification model for english and bahasa indonesia telephone speeches. Procedia Computer Science, vol. 157, pp. 199–204 (2019)
19. Badr, A. A., Abdul-Hassan, A. K.: Speaker gender identification in matched and mismatched conditions based on stacking ensemble method. Journal of Engineering Science and Technology, vol. 17, no. 2, pp. 1119–1134 (2022)
20. Kushwah, S., Singh, S., Vats, K., Nemade, V.: Gender identification via voice analysis. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, pp. 746–753 (2019)
21. Kaur, S., Garg, D., Arora, D.: Comparative analysis of speech processing techniques for gender recognition. International Journal of Advances in Electrical and Electronics Engineering, pp. 278–283 (2012)
22. Konig, Y., Morgan, N., Chandra, C.: GDNN: A gender-dependent neural network for continuous speech recognition. International Computer Science Institute (1991)
23. Alsulaiman, M., Ali, Z., Muhammad, G.: Voice intensity based gender classification by using Simpson's rule with SVM. In: 2012 19th International Conference on Systems, Signals and Image Processing (IWSSIP), pp. 552–555 (2012)
24. Ponraj, A. S., Naveen M.: Speech recognition with gender identification and speaker diarization. In: 2020 IEEE International Conference for Innovation in Technology (INOCON), pp. 1–4 (2020)

25. Bawa, P., Kumar, V., Kadyan, V., Singh, A.: Noise-robust gender classification system through optimal selection of acoustic features. Deep Learning Approaches for Spoken and Natural Language Processing, pp. 147–159 (2021)
26. Guha, S., Das, A., Singh, P. K., Ahmadian, A., Senu, N., Sarkar, R.: Hybrid feature selection method based on harmony search and naked mole-rat algorithms for spoken language identification from audio signals. IEEE Access, vol. 8, pp. 182868–182887 (2020)
27. Ma, Y., Nishihara, A.: Efficient voice activity detection algorithm using long-term spectral flatness measure. EURASIP Journal on Audio, Speech, and Music Processing, vol. 2013, no. 1, pp. 1–18 (2013)
28. Shagi, G. U., Aji, S.: A machine learning approach for gender identification using statistical features of pitch in speeches. Applied Acoustics, vol. 185, pp. 108392 (2022)
29. Nasef, M. M., Sauber, A. M., Nabil, M. M.: Voice gender recognition under unconstrained environments using self-attention. Applied Acoustics, vol. 175, pp. 107823 (2021)
30. Sánchez-Hevia, H. A., Gil-Pita, R., Utrilla-Manso, M., Rosa-Zurera, M.: Convolutional-recurrent neural network for age and gender prediction from speech. In: 2019 Signal Processing Symposium (SPSympo), pp. 242–245 (2019)
31. Sánchez-Hevia, H. A., Gil-Pita, R., Utrilla-Manso, M., Rosa-Zurera, M.: Age group classification and gender recognition from speech with temporal convolutional neural networks. Multimedia Tools and Applications, vol. 81, no. 1, pp. 3535–3552 (2022)
32. Tursunov, A., Choeh, J. Y., Kwon, S.: Age and gender recognition using a convolutional neural network with a specially designed multi-attention module through speech spectrograms. Sensors, vol. 21, no. 17, pp. 5892 (2021)