IASJ المجلات الاكاديمية العراقية

Tikrit Journal of Pure Science

ISSN: 1813 – 1662 (Print) --- E-ISSN: 2415 – 1726 (Online)

Journal Homepage: http://tjps.tu.edu.iq/index.php/j

An overview for assessing a number of systems for estimating age and gender of speakers

Aalaa Ahmed Mohammed[1], Yusra Faisal Al-Irhayim[2]

[1] *College of Engineering, University of Tikrit, Tikrit, Iraq*

Dept.of computer sciences, College of Computer Sciences and Mathematics, University of Mosul, Mosul, Iraq

https://doi.org/10.25130/tjps.v26i1.105

ARTICLEINFO.

Article history:

-Received: 12 / 5 / 2020

-Accepted: 1 / 11 / 2020

-Available online: / / 2020

Keywords:

Corresponding Author:

Name: Aalaa Ahmed Mohammed

E-mail:

aalaa.alrashidy@tu.edu.iq

Tel:

ABSTRACT

The determination of the age and gender of the speaker of the speech signal is an interesting topic in the interaction between human-machine. Speech signal has a variety of applications ranging from speech analyses to allocate human-machine interactions. This paper aims to conduct a comparative study of age and gender classification algorithms applied to the speech signal. Comparison of experimental results of different sources of voices for speakers of different languages and methods of miscellaneous classification such as Bayes classifier, neural network, support vector machines, K-nearest neighbor, gaussien mixture model and hybrid method based on weighted analysis of a directed non-negative matrix and a neural network with a general recession as well as some deep learning methods, is done in order to show different results to classify the age and gender of the speaker when processing the speech signal. The study showed that methods and algorithms of deep learning have excelled in providing accuracy ratios higher than other methods, and it shows that the hybridization of two or more classification methods increases the accuracy level of the results.

1. Introduction

The process of estimating the speaker's age and gender is an important issues of speech processing because the human voice represents one of the biometric features that distinguish each person from the other. Nowadays, with the development of technology, recognizing age and gender of the speaker is essential in speech recognition and verification systems because age and gender information have a mutual effect on each other. The speaker's age and gender estimation process can be defined as the process of extracting the information of age and gender from the speaker's voice signal. There are two main stages in the process of defining a speaker's age and gender: The first: extracting and selecting effective and influential specifications that represent the characteristics of the speaker's voice. The second: classifier design that uses the properties obtained from the first stage to estimate age and gender [1].

Despite many studies that focused on improving the extraction of traits and designing works, the accuracy of classifications remained completely unfulfilled, so the aim of this paper was to study what is related to the speech signal through many useful information extracted in estimating the age and gender of the speaker and getting to know some of systems used for recognition. And know to what extent accurate and near-reality results were obtained.

1.1. The importance of estimating the speaker's age and gender

Estimating the age and gender of speakers has a great importance and has been the subject of several research studies because of entering the computer and modern technologies in many applications in our lives, including [2,3]: -

1- Strategies for effective marketing and advertising about commodities, as they are used in customer relationship management (CRM) systems.

2- Enhancing computer and human interaction systems, especially the conversation systems used by customer services in various companies.

3- Estimating the speaker's age is a useful tool in various applications, especially those that require knowledge of age to give permission to use them.

4- In forensic science, number of suspects can be reduced if there is evidence such as a telephone call.

1.2. The difficulties of estimating the speaker's age and gender

Distinguishing the speaker's gender does not carry significant difficulties, because there are clear differences between female and male voices in adulthood depending on physiological differences and the physical characteristics of the vocal cords in terms of length and degree of tension leading to the production of different sound signals in frequency and other distinctive characteristics, while an estimate The speaker's age is difficult in terms of different perspectives [4]: -

1- Usually there is a difference between the age of the perceived speaker, that is, what is perceived by the listener and the true age of the speaker.

2- It is difficult to develop a robust age estimation system because it requires a coded database with a wide and balanced age range.

3- The patterns of sounds are affected by several factors such as weight, height, and emotional conditions.

2. Speech Signals Types

In the processing algorithms of speech signals, the speech sound classification is very critical and influential because it has been affected by the vocal tracts and the vocal cords that are the main parts of human-speech production systems [5,6,7].

a) Voiced Sounds: The voiced sounds are differentiated by the existence of acoustic waveform periodicity. The pitch frequency means that the resonant frequency, in which the vocal cords must vibrate, is mainly associated with the recognition of the pitch in the voiced speech. For male speakers, the pitch frequency is usually between 50 and 200Hz, while the female speakers have a high pitch that may reaches to twice more than that of male speakers. The effective recognition of voiced speech parts is essential for the efficiency of many processing algorithms of speech signals.

b) Unvoiced Sounds: The unvoiced sounds are not having any acoustic waveform periodicity, and so it has not distinguished relationship with the pitch. In the most of speech processing algorithms, the unvoiced speech is commonly appearing as a source of arbitrary white noises. In the unvoiced sounds there are no vocal cords vibrations and this is the main dissimilarity between voiced and unvoiced sounds.

c) Fricative Sounds: The fricatives are appearing as a noise resulting by a localized disorder in the vocal tracts that have incomplete constriction. Fricatives are could be either voiced or unvoiced; the main distinction in voiced fricatives is that the noisy attribute that is made by the constriction is also joined with vocal cords vibration, causing some of periodicity in the resulting sound.

d) Stop Sounds: There is voiced and unvoiced stop sounds, both are produced when some constriction occurs in the vocal tracts and leads to close it completely. The major variance is that when the vocal tract pressure has been increased by the constriction, the vocal cords vibrate in the voiced stops case, while they do not vibrate in the unvoiced stops case.

e) Nasal Sounds: There are many sounds that are nasal in their essential quality and there are produced when the velum is reduced so that oral cavity is minimized with staying connected with pharynx, and air pressure fluxes during the nasal tract. In this case, for some frequencies the mouth becomes a resonant cavity. Typically, the spectral reactions for the nasal sounds are wider than that of the voiced sounds because of the nasal and oral tracts coupling.

3. Speaker age and gender recognition

Speaker age and gender recognition is a process which is used to convert speaker voice into signal, by which the machine can be able to estimate the age and gender of the speaker. When this is achieved, the machine can be made to work, as desired. The machine could be a computer, a typewriter, or even a robot [8]. There are various voice recognition systems and various software and hardware devices, available which use various techniques to decode human speech [9]. For achieving the main two stages (features extracting and classification) some features extracting techniques must be used, such as [2][10][11]: mel-scaled power spectrogram, mel-frequency cepstral coefficient (MFCC), chorma-stft (short time fourier transform), spectral-contrast, tonnez, and linear predictive coding (LPC). As well as some classifiers such as [1][3][12]: deep learning methods (including convolutional neural network (CNN) and probabilistic neural network (PNN)), eigenvoice (I-vector), gaussian matrix models (GMM), hidden markov model (HMM), support vector regression (SVR), multilayer perceptrons (MLP), k-nearest neighbor (K-NN),decision tree (DT), and bayes classifier.

4. Stages of estimating the speaker's age and gender

The speaker age and gender estimation goes through several important stages, as shown in Figure 1:

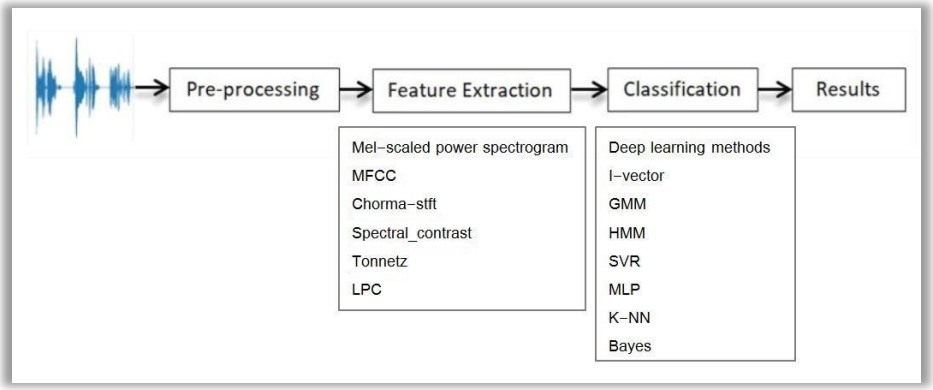

Fig. 1: stages of estimating the speaker's age and gender

1- Collecting data that is done by either recording sound or obtaining sound files from different databases. There are many voices datasets provided for researches purpose such as ITU-T (International Telegraph Union- Telecommunication), ETRI-VoiceDB2006, free spoken digit dataset (FSDD), Kannada dataset, LJ Speech corpus, etc.

2- Preprocessing of sound signals is considered a crucial step in the development of a robust and efficient speech or speaker recognition system. In which the sound signal is converted from analog to digital, and the signal is purified and noise is reduced [13].

3- Feature extraction process is very important technique that are used for speech and speaker recognition and also used for feature classification of voices. It is used for extracting the characteristics of the sound signal, which distinguishes each sound signal from the other. It does a great job in increasing the proficiency of the works. There are several types of feature extraction techniques such as Mel-scaled power spectrogram, MFCC, chorma-stft, spectral-contrast, Tonnetz, LPC, etc [14].

4- The classification process that is used to distinguish age and gender, is done by using different types of classifiers. The strength and efficiency of these classifiers are determined by the quality of the characteristics extracted. It is possible to use more than one level of classifiers to obtain more accurate results, in practice each classification technique is utilized for building a group of suggested systems and choosing the best model from them. There are many classification techniques such as deep learning methods, Gaussian mixture model (GMM), Support Vector Machine (SVM), Naïve Bays, Non-Negative Matrix Factorization, etc.

5. Classification Techniques

A. Deep learning

Deep learning is one of data mining processes that use deep neural network architectures. It is a branch of artificial intelligence and machine learning algorithms that has gained great importance in the past few years. Deep learning enables building neural networks composed of many layers of processing that are properly trained to gradually solve complex problems, and have the ability to learn high-level representations of data by exploiting multiple levels of abstraction as the information is processed in a parallel and distributed manner [15]. Deep learning algorithms are highly efficient and close to the human level in speech recognition, estimating the age and gender of the speaker compared to previous HMM and GMM-based techniques. The success of deep learning algorithms depends on the availability of a large amount of training data and highly efficient calculations [16]. There are several architectures for deep neural networks, including: Deep neural networks, Deep belief networks, Recurrent neural networks and Convolutional neural networks.

B. Gaussian mixture model (GMM)

Gaussian mixture model (GMM) is a powerful clustering algorithm used in the fields of speech and speaker recognition and can be defined as a parametric probability density function represented as a weighted sum of Gaussian component densities. In the training phase of GMM, parameters of probability density function for each class (gender or age) are estimated. Then, through the classification phase, by computing the maximum likelihood criterion, a decision is taken for each test utterance. The GMM is a combination of K Gaussian laws. Each law in the mixture is weighted and specified by two parameters: the mean and the covariance matrix ΣK [17].

C. Support Vector Machine (SVM)

One of the most robust prediction methods uses linear and nonlinear separating hyper-planes for data classification. However, since SVMs can only classify fixed length data vectors, this method cannot be readily applied to task involving variable length data classification. The variable length data has to be transformed to fixed length vectors before SVMs can be used. It is a generalized linear classifier with maximum-margin fitting functions. This fitting function provides regularization which helps the classifier generalized better. The classifier tends to ignore many of the features. Conventional statistical and Neural Network methods control model complexity by using a small number of features (the problem dimensionality or the number of hidden units) [18].

D. Naïve Bays

Naive Bayes models have become popular for classification and clustering because of their simplicity, efficiency, and accuracy. It is a direct approach that finds the best hypothesis by using Bayes' theorem as a probability theorem for building rule-or graph-based classification models. Two well-known methods are used, which are the Bayesian network (BN) and naive Bayes (NB) models [2].

E. Non-Negative Matrix Factorization

Non-Negative Matrix Factorizations (NMF's) are popular for the problem of approximating nonnegative data. And moreover, the non-negative matrix factorization (NMF) based model has become one of the most popular collaborative filtering approaches in machine learning systems because of its high efficiency. NMF in its classical form is an unsupervised method but incorporating the classification labels into the NMF algorithms allows to specifically guiding them toward the extraction of data patterns relevant for discriminating the respective classes [19].

F. General Regression Neural Network

Generalized Regression Neural Networks (GRNNs) are single-pass associative memory feed-forward type Artificial Neural Networks (ANNs) and uses normalized Gaussian kernels in the hidden layer as activation functions. GRNN provides accurate and quick solution to regression, approximation, classification and fitting problems. GRNN can be used in system identification of dynamic systems as well as control of dynamic systems. Generalized Regression Neural Networks GRNN outperforms Back Propagation ANNs in the accuracy and training time; however, GRNN has some limitations such as the growth of the hidden layer. When training GRNN with a large dataset it is essential to reduce the data dimensionality using any of the data reduction techniques such as clustering or distance based algorithms [20].

6. Comparison Results of Several Researches

Estimating age and gender depending on voice was a difficult task for sound analysts but with the development of technology and with the increase in human-computer interaction (HCI), finding smart ways to estimate the age and gender of the speaker became more important. Table (1) provides a summary and comparison of several researches in terms of data, classification methods and accuracy ratios of each research as results.

Table 1: Comparing several researches results

Reference number	Author Name	Research title	Data set	The used method	Results
[2]	Rami S. Alkhawaldeh	DGR: Deep Gender Recognition of Human Speech	ITU-T(International Telegraph Union-Telecommunication)	Deep Convolution Neural Network, SMO(Sequential Minimal Optimization for SVM)	The accuracy ratio for the neural network is 99.97%, and for the SMO method is 99.7%.
[4]	Mohamad Hasan Bahari, Hugo Van hamme	Speaker Age Estimation and Gender Detection Based on Supervised Non-Negative Matrix Factorization	555 speakers	Hybrid architecture of Weighted Supervised Non-Negative Matrix Factorization (WSNMF) and General Regression Neural Network (GRNN)	The accuracy rate in the estimation of gender is 96%, and the age estimate is better at 15%, depending on the absolute error rate that is taken to estimate the age
[12]	Fatima K. Faek	Objective Gender and Age Recognition from Speech Sentences	114 speakers	SVM,K-NN	The accuracy rate in the estimation of gender is 96%, and the age estimate rate is 81.44%
[21]	David Doukhan, Jean Carrive, Félicien Vallet, Anthony Larcher, Sylvain Meignier	An Open-source Speaker Gender Detection Framework For Monitoring Gender Equality	2284 French speakers	GMM, CNN	The CNN method is best according to the F-measure that used as an efficiency criterion where the accuracy rate is 96.52%, while the accuracy rate for the GMM method is 95.74%, and the accuracy rate for the I-vector method is 95.51%.
[22]	Zimeng Hong	Speaker Gender Recognition System	400 speakers for 16 language	Naïve Bays	The efficiency of estimating the age and gender in this method is good compared to other methods where the accuracy rate is 92.75%, and the ECOC method accuracy rate is 91.25%, and for the KNN method was 55.75% accuracy, and the accuracy rate for the Tree method is 44%, while the accuracy rate of the SVM method is 93%.
[23]	Ming Li, Kyu J.Han, Shrikhanth Navayanan	Automatic speaker age And gender recognition using acoustic and prosodic level information fusion	772 speakers	GMM, SVM	The research shows that the use of several partial systems greatly increased the strength and efficiency of classification of different age and gender groups.
[24]	Hye-Jin Kim, Kyungsuk Bae, Ho-Sub Yoon	Age and gender classification for a home-robot service	ETRI-VoiceDB2006	GMM	The accuracy rate in the estimation of gender is 94.9%, and the age estimate rate is 94.6%
[25]	Leo Kristopher Piel	Speech-based identification of children's gender and age with neural networks	309 speakers	Feed forward deep neural networks and recurrent neural networks with convolutional layers	The accuracy rate in the estimation of gender is 92.8%, and the age estimate rate is 76.3%

7. Conclusion

This paper made a clear and simple overview of working of speaker's age and gender recognition by comparing several methods of estimating the age and gender of a speaker. Table (1) shows many types of classifiers that obtained high-accuracy ratios in their results for the detection of the speaker age and gender applied on multiple speech datasets with different numbers of speakers and different languages, with accuracy ratios exceeding 90% for classification methods such as deep learning methods, Vector

Support Machine (SVM), Gaussian Matrix Models (GMM), Supervised Non-Negative Matrix Factorization and Naïve Bayes Classifier, but deep learning methods outperformed in providing higher ratios than other methods, and it also shows that hybridization of two or more classification methods would Increase the level of accuracy of the results.

References

[1] Abu Mallouh, A. (2017). A framework for enhancing speaker age and gender classification by using a new feature set and deep neural network architectures. Ph.D. thesis, The School of Engineering, University of Bridgeport, United States: 95 pp.

[2] Alkhawaldeh, R. S. (2019). DGR: Gender recognition of human speech using one-dimensional conventional neural network. *Hindawi. Scientific Programming*, **(2019)7213717**:1-12.

[3] Sedaaghi, M. H. (2009). A comparative study of gender and age classification in speech signals. *Iranian Journal of Electrical & Electronic Engineering*, **(5) 1**:1-12.

[4] Bahari, M. H. and Hamme, H. V. (2011). Speaker age estimation and gender detection based on supervised non-negative matrix factorization. BIOMS: Milan: 1-6 pp.

[5] Sinha, P. (2010). Speech processing in embedded systems. Springer New York Dordrecht Heidelberg: London: 177 pp.

[6] Hernandez, M. J. (2016). A tutorial to extract the pitch in speech signals using autocorrelation. *Open Journal of Technology & Engineering Disciplines (OJTED)*, **(2)1**: 01-10.

[7] Kanabur,V.; Harakannanavar, S. and Torse, D. (2019). An extensive review of feature extraction techniques, challenges and trends in automatic speech recognition. *International Journal of Image, Graphics and Signal Processing*, **(11)5**: 1-12.

[8] Joshi, S.; Kumari, A.; Pai, P.; Sangaonkar, S. and D'Souza, M. (2017). Voice Recognition System. *Journal for Research*, **(03) 01**: 6-9.

[9] Kumar, P. M. (2016). A new human voice recognition system. *Asian Journal of Science and Applied Technology*, **(5)2**: 23-30.

[10] Chaudhari, S. and Kagalkar, R. (2014). A review of automatic speaker age classification, recognition and identifying speaker emotion using voice signal. *International Journal of Science and Research (IJSR)*, **(3)11**: 1307-1311.

[11] Mavaddati, S. (2018). Voice-based age and gender recognition based on learning generative sparse models. *International Journal of Engineering IJE Transactions*, **(31)9**: 1529-1535.

[12] Faek, F. K. (2015). Objective gender and age recognition from speech sentences. *ARO-The Scientific Journal of Koya University*, **(III)2**: 24-29.

[13] Keerio, A.; Mitra, B. K.; Birch, P.; Young, R.; and Chatwin, C. (2009). On preprocessing of speech signals. *International Journal of Signal Processing*, **(5)3**: 216-222.

[14] Ranjan, R. and Thakur, A. (2019). Analysis of feature extraction techniques for speech recognition system. *International Journal of Innovative Technology and Exploring Engineering*, **(8)7C2**: 197-200.

[15] Tebelskis, J. (1995). Speech recognition using neural networks. Ph.D thesis in computer science, The School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania:190 pp.

[16] Feng, L. (2004). Speaker recognition. M.Sc. thesis, The Intelligent Signal Processing group at Institute of Informatics and Mathematical Modelling, Technical University, Denmark: 112 pp.

[17] Sedaaghi, M. H. (2009). A comparative study of gender and age classification in speech signals. *Iranian Journal of Electrical & Electronic Engineering*, **(5)1**: 1-12.

[18] Anusuya, M. A. and Katti, S. K. (2009). Speech recognition by machine: A review. *International Journal of Computer Science and Information Security*, **(6)3**: 181-205.

[19] Leuschner, J. et.al. (2019). Supervised non-negative matrix factorization methods for MALDI imaging applications. *Bioinformatics*, **(35)11**: 1940-1947.

[20] Al-mahasneh, A. J.; Anavatti, S. G. and Garratt, M. A. (2018). Review of application of generalized regression neural networks in identification and control of dynamic systems. *arXiv*, **(abs/1805.11236)**: 5 pp.

[21] Doukhan, D.; Carrive, J.; Vallet, F.; Larcher, A. and Meignier, S. (2018). An open-source speaker gender detection framework for monitoring gender equality. IEEE International Conference on Acoustic Speech and Signal Processing, April 2018, Calgary, Canada: 5 pp.

[22] Zimeng, H. (2017). Speaker gender recognition system. M.Sc. thesis, University of Oulu, Oulu, Finland: 54 pp.

[23] Li, M.; Han, K. J. and Navayanan, S. (2013). Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. *Elsevier, Computer Speech and Language*, **(27)**:151–167.

[24] Kim, H.; Bae, K. and Yoon, H. (2007). Age and gender classification for a home-robot service. 16th IEEE International Conference on Robot & Human Interactive Communication, Jeju, Korea: p. 122-126.

[25] Piel, L. K. (2018). Speech-based identification of children's gender and age with neural networks. M.Sc. thesis, Tallinn University of Technology, Tallinn, Estonia: 85 pp.

نظرة عامة لتقييم عدد من أنظمة تقدير عمر وجنس المتحدثين

الاء احمد محمد احمد، يسرى فيصل الارحيم

[1]كلية الهندسة، جامعة تكريت، تكريت، العراق

[2]قسم علوم الحاسوب، كلية علوم الحاسوب والرياضيات، جامعة الموصل، الموصل، العراق

الملخص

إنّ تحديد عمر وجنس المتحدث من إشارة الكلام موضوعًا مثيرًا للاهتمام في التفاعل بين الإنسان والآلة. تحتوي إشارة الكلام على مجموعة متنوعة من التطبيقات تتراوح من تحليلات الكلام إلى تخصيص التفاعلات بين الإنسان والآلة. تهدف هذه الورقة إلى إجراء دراسة مقارنة لخوارزميات تصنيف العمر والجنس المطبقة على إشارة الكلام. تمت مقارنة النتائج التجريبية لمصادر مختلفة من الأصوات لمتحدثين بلغات مختلفة وطرق التصنيف المتنوعة مثل مصنف بايز والشبكة العصبية الالتفافية وآلات دعم المتجه والجار الأقرب ونموذج الخليط الغاوسي وطريقة هجينة بالاعتماد على التحليل الموزون لمصفوفة غير سالبة موجَهة وشبكة عصبية ذات انحسار عام بالإضافة إلى بعض طرق التعلم العميق، من أجل إظهار النتائج المختلفة لتصنيف العمر والجنس للمتحدث عند معالجة إشارة الكلام. أظهرت الدراسة أن طرق وخوارزميات التعلم العميق تفوقت في توفير نسب دقة أعلى من الطرق الأخرى، كما توضح أن تهجين طريقتين أو أكثر من طرق التصنيف يزيد من مستوى دقة النتائج.