ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

GENDER RECOGNITION BY VOICE

Project for the course

MATH-412 Statistical Machine Learning,
Ecole Polytechnique Federale de Lausanne,
Fall 2017

Authors: Adrien Besson, Hippolyte Lefebvre, Greg
Professor: Dr. Emeric Thibaud

December 19, 2017

Contents

List of Figures

List of Tables

1 Introduction

In the last decades, automatic gender recognition (AGD) from speech has grown many interest thanks to the digitization of an extensive number of applications and the development of mobile platforms [1]–[8].

The applications of AGR have increased consequently. Indeed, in general, the accuracy of gender-dependent systems is higher than the one of gender-independent systems [4]. Thus, AGR improves the prediction of other speaker traits such as age [9] and emotional state [10], [11]. It can also facilitate speech recognition by gender-based normalization [12] and is a key feature for more natural and personalized dialog systems such as Siri.

The AGR techniques are based on statistical features extracted from the speech signals such as maximum, minimum and average frequency measured in a time span. These features translate physiological differences between males and females like the length of the vocal chords or the glottal shape [13]. Among all the features, it appears that the fundamental frequency plays a crucial role in gender classification as described in many studies [1], [14], [15]. In recent works, the use of the fundamental frequency coupled with spectral components such as Mel-frequency spectral components [16] or relative spectral perceptual linear predictive coefficients [5] have demonstrated best AGR performances even in noisy environments.

In this project, we study different state-of-the-art classification methods applied to the task of gender recognition by voice. The study is based on a dataset of features extracted from 3168 subjects available on Kaggle[1] and described in details in Chapter 2. A preliminary exploratory data analysis is performed in Chapter 3 which leads us to a first intuitive classification technique described in Chapter ??. Starting from the conclusions of this intuitive approach, the exhaustive comparison of the methods is achieved in Chapter 4 and the best model is selected. Eventually, the best model is tested on 4 voices recorded by the authors in Chapter 5.

[1] `https://www.kaggle.com/primaryobjects/voicegender`

2 The dataset

2.1 General considerations

The voice gender dataset[1] consists of features extracted from 3168 recorded voice samples, collected from male and female speakers. The features have been computed using tuneR[2] and seewave[3], two acoustic analysis packages of R.

The dataset takes the form of a csv files where each row is composed of the following acoustical features of each voice:

- **meanfreq:** mean frequency (in kHz)
- **sd:** standard deviation of frequency
- **median:** median frequency (in kHz)
- **Q25:** first quantile (in kHz)
- **Q75:** third quantile (in kHz)
- **IQR:** interquantile range (in kHz)
- **skew:** skewness of the spectrum
- **kurt:** kurtosis
- **sp.ent:** spectral entropy
- **sfm:** spectral flatness
- **mode:** mode frequency
- **centroid:** frequency centroid
- **peakf:** peak frequency (frequency with highest energy)
- **meanfun:** average of fundamental frequency measured across acoustic signal
- **minfun:** minimum fundamental frequency measured across acoustic signal
- **maxfun:** maximum fundamental frequency measured across acoustic signal
- **meandom:** average of dominant frequency measured across acoustic signal
- **mindom:** minimum of dominant frequency measured across acoustic signal
- **maxdom:** maximum of dominant frequency measured across acoustic signal
- **dfrange:** range of dominant frequency measured across acoustic signal
- **modindx:** modulation index. Calculated as the accumulated absolute difference between adjacent measurements of fundamental frequencies divided by the frequency range
- **label:** male or female

The features are all quantitative and represents frequency characteristics of the voices.

[1] `https://www.kaggle.com/primaryobjects/voicegender`
[2] `https://cran.r-project.org/web/packages/tuneR/tuneR.pdf`
[3] `https://cran.r-project.org/web/packages/seewave/seewave.pdf`

2.2 Description of the features

Before starting the data analysis, it is important to perfectly understand the features involved in the exercise. This will be very useful in a preprocessing step, since it will allow us to remove collinear features. It will also be a great asset when it will come to the analysis of the most important features in the gender recognition.

As already pointed out in Section 2.1, the extracted features are all related to the spectrum.

Frequency-related features The mean frequency corresponds to a weighted average of the frequency by the amplitude of the spectral components:

$$\mu_f = \sum_{i=1}^{N} f_i y_i, \tag{2.1}$$

where N is the number of frequency components of the spectrum, f_i is the i-th frequency and y_i is the relative amplitude of the i-th component of the spectrum. As described in p.163 of the seewave documentation, it is equal to the feature 'centroid'. The standard deviation is calculated as:

$$\sigma_f = \sqrt{\sum_{i=1}^{N} y_i \left(f_i - \mu_f\right)^2} \tag{2.2}$$

The median frequency is calculated as the frequency where the spectrum is divided into frequency intervals of same energy. The calculation of the quartiles are based on the same criterion. The interquartile range is calculated as the difference between the third and the first quartile.

The feature "mode" characterizes the dominant frequency of the spectrum, *i.e.*the one with the highest amplitude. It is very similar to the peak frequency which corresponds to the frequency with the highest energy. The fundamental frequency is the lowest frequency of the spectrum.

The features "meanfun", "minfun", "maxfun", "meandom", "maxdom", "mindom", "dfrange" and "modindx" are based on short-time Fourier transform applied on segments of fixed durations, small compared to the duration of the whole signal. This permits to have features more localized in time.

In addition to the frequency-related features, we can find measures on the shape of the spectrum which may give very interesting additional information.

Skewness of the spectrum The skewness of the spectrum is a measure of its asymmetry around the mean frequency. It is calculated as follows:

$$S = \frac{1}{\sigma_f^3} \frac{\sum_{i=1}^{N} \left(f_i - \mu_f\right)^3}{N-1}. \tag{2.3}$$

From (2.3), it is clear that the sign of S gives information of the left or right asymmetry of the spectrum while the absolute value of S gives the strength of the asymmetry.

Kurtosis The Kurtosis is a measure of the "tailedness" of a probability distribution. It is calculated as the fourth order moment of the frequency distribution, described below:

$$K = \frac{1}{\sigma_f^4} \frac{\sum_{i=1}^{N} \left(f_i - \mu_f\right)^4}{N-1}. \tag{2.4}$$

When $K = 3$, the frequency distribution is normal. When $K < 3$, the frequency distribution is said to be *platikurtic*, it has fewer items around the means than in the tails, compared to a normal distribution. When $K > 3$, the distribution is said to be *leptokurtic* and has more frequency around the mean than in the tails, compared to a normal distribution.

Shannon spectral entropy The Shannon entropy is used to discriminate whether the voice signal is noisy or pure [17]. it is calculated as follows:

$$H = \frac{-\sum_{i=1}^{N} y_i \log_2 (y_i)}{\log_2 (N)} \tag{2.5}$$

If the signal is pure, then all the energy is concentrated in one frequency component, let us say the j-th component for which $y_j = 1$. In this case, $H = 0$. If the signal is a white noise, then $y_i = 1/N$, $\forall i \in \{1, ..., N\}$ and $H = 1$.

Spectral flatness The spectral flatness is rather similar to the spectral entropy. It is measured as the ratio between the geometric mean and the arithmetic mean:

$$F = N \frac{\sqrt[N]{\prod_{i=1}^{N} y_i}}{\sum_{i=1}^{N} y_i}. \tag{2.6}$$

In case of a white noise, the spectrum is flat and $H = 1$. In case of a pure tone, the geometrical mean is equal to zero and $H = 0$.

2.3 Cleaning the dataset

From the description of the features given in Section 2.2, a first cleaning of the dataset may be achieved before starting the analysis. Indeed, several features are exactly the same of collinear:

- The features “meanfreq” and “centroid” are exactly similar. So “centroid” has been removed;
- The following relationship holds: “IQR” = “Q75” − “Q25”. “IQR” has been removed.
- The following relationship holds: “dfrange” = “maxdom” − “mindom”. “dfrange” has been removed.

3 Exploratory Data Analysis

As a preliminary step, we propose to perform an exploratory data analysis. This will give us some hints about the dataset, *e.g.*the most important features, their correlation etc.

In order to have a first overview of the features, a short description is summarized in Table 3.1. It can be noticed the frequencies have low values, which makes sense since they are expressed in kHz. The mean fundamental frequency is about 143 Hz which is coherent with the male and female fundamental frequencies [18].

Regarding the shape of the spectrum, the mean skewness indicates an average right-asymmetry of the spectrum. The mean kurtosis shows that the frequency distribution is leptokurtic. About the flatness of the spectrum, the features "sfm" and "sp.ent" seem to have inconsistent behaviour with respect to the their average value since one is above 0.5 and the other is below. However, the high standard deviation of "sfm" makes an analysis rather difficult.

Table 3.1 Description of the features of the dataset

	meanfreq	sd	median	Q25	Q75	skew	kurt	sp.ent	sfm
mean	0.181	0.0571	0.186	0.140	0.225	3.14	36.6	0.895	0.408
std	0.0299	0.0167	0.0364	0.0487	0.0236	4.24	134	0.0450	0.178
min	0.0394	0.0184	0.0110	0.000229	0.0429	0.142	2.07	0.739	0.0369
25 %	0.164	0.0420	0.170	0.111	0.209	1.65	5.67	0.863	0.258
50 %	0.185	0.0592	0.190	0.140	0.226	2.20	8.32	0.902	0.396
75 %	0.199	0.0670	0.211	0.176	0.244	2.93	13.7	0.929	0.534
max	0.251	0.115	0.261	0.247	0.273	34.7	1310	0.982	0.843

	mode	meanfun	minfun	maxfun	meandom	mindom	maxdom	modindx
mean	0.165	0.143	0.0368	0.259	0.829	0.0526	5.05	0.174
std	0.0772	0.0323	0.0192	0.0301	0.525	0.0633	3.52	0.119
min	0.00	0.0556	0.00977	0.103	0.00781	0.00488	0.00781	0.00
25 %	0.118	0.117	0.0182	0.254	0.420	0.00781	2.07	0.0998
50 %	0.187	0.140	0.0461	0.271	0.766	0.0234	4.99	0.139
75 %	0.221	0.170	0.0479	0.277	1.18	0.0703	7.01	0.210
max	0.280	0.238	0.204	0.279	2.96	0.459	21.87	0.932

Regarding the distribution of the samples, there are 1584 recording of male voices and 1584 recording of female voices. So the classes are perfectly balanced.

Let us have a look to the correlation between the features. The correlation matrix, displayed in Figure 3.1, exhibits high correlations between "skew" and "kurt" and between "sp.ent" and "sfm", which make sense since they quantify similar quantities. It can also be noticed that "meanfreq" and "median", "Q25", "Q75" are highly correlated which is self-evident given their definition. Thus, feature selection methods should be efficient in removing such redundancies in the dataset.

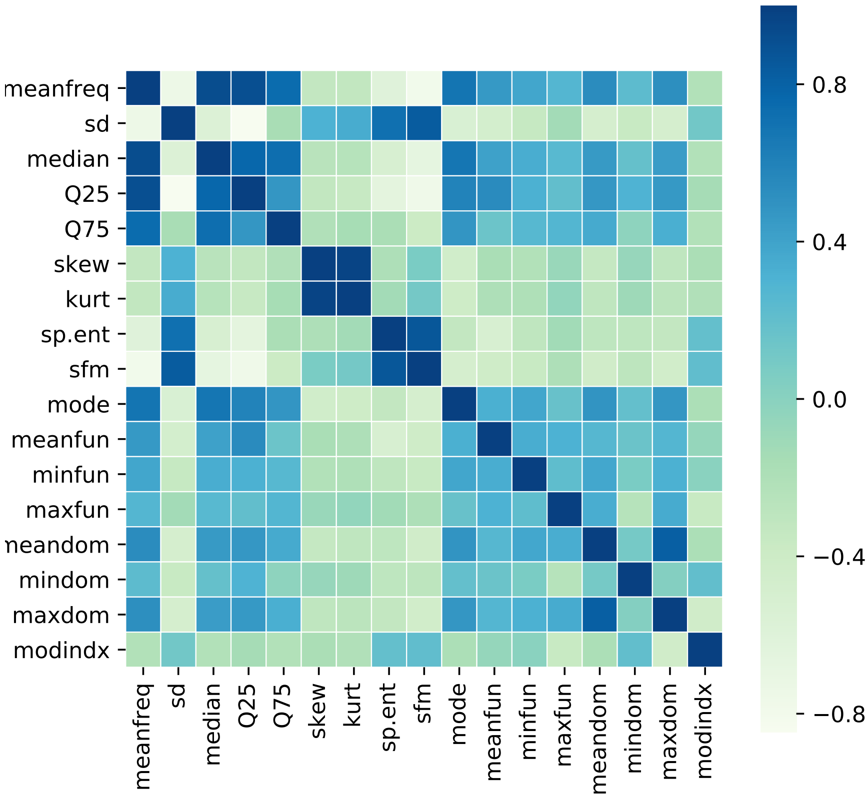
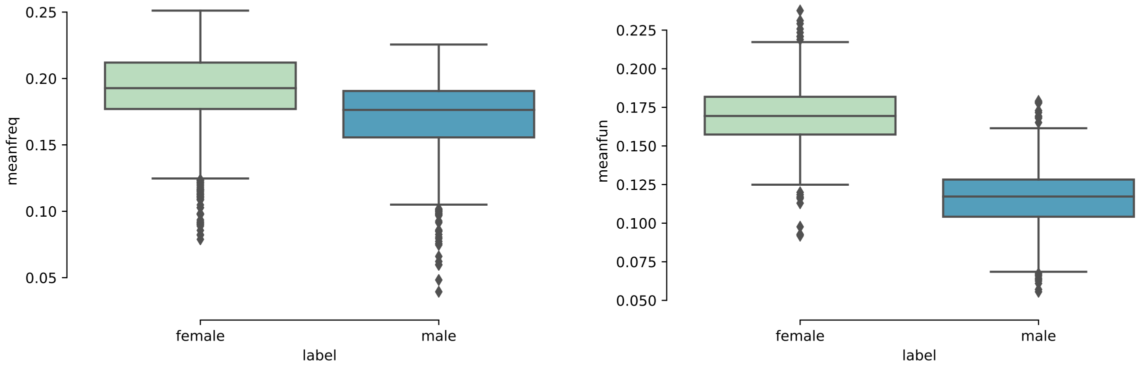


Figure 3.1 Correlation matrix of the dataset.

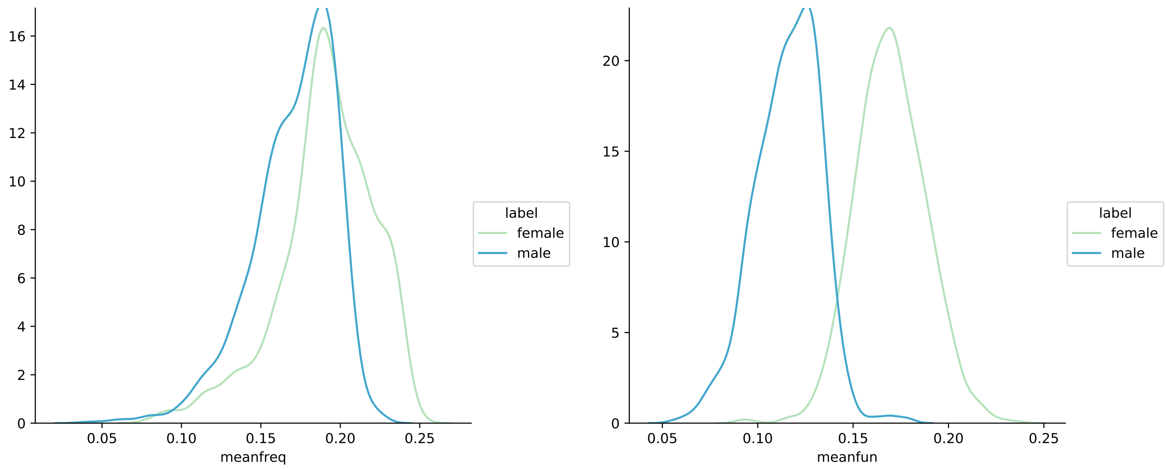
In the state-of-the-art, it appears that the fundamental frequency is a key feature for AGR, as stated in Chapter 1. Intuitively, we also think that the mean frequency should be a good classifier. In order to analyze this, Figs. 3.2a and 3.2b represent the box plots of "meanfreq" and "meanfun" respectively. It can be noticed that "meanfun" is indeed a key feature for classification since the overlap between male and female is very low. Regarding "meanfreq", the overlap is bigger than for "meanfun" but remains rather low.

Figs. 3.3a and 3.3b represent the distribution of male and female with respect to "meanfreq" and "meanfun" respectively. They confirm the analysis made with the box plot, *i.e.* that "meanfun" is a key component in AGR and is a far better classifier than "meanfreq".



(a) (b)

Figure 3.2 Box plots for (a)-“meanfreq” and (b)-“meanfun” features.



(a) (b)

Figure 3.3 Distribution of “male” and “female” for (a)-“meanfreq” and (b)-“meanfun” features.

4 Evaluation of the Best Classification Method

4.1 Considered classification methods

In our analysis we implement the following statistical learning methods to predict the class of $y \in \{female, male\}$.

- **Logistic Regression:** models the posterior probability of response $y \in \{female, male\}$, given the predictors $\mathbf{X}$, using the logistic function;
- **Regularized Logistic Regression:** The ℓ_1-penalty (LASSO) can be used for variable selection and shrinkage with logistic regression. It is a useful approach to examine which features are important in voice analysis;
- **Linear Discriminant Analysis (LDA):** LDA makes the assumption that the conditional densities $f(\mathbf{X}|y = male)$ and $f(\mathbf{X}|y = female)$ are both multivariate Gaussian with a common covariance matrix. It belongs to a family of techniques that use linear boundaries to separate classes in classification problems. If the assumption of normality is realistic, then LDA is expected to provide better results than logistic regression;
- **Quadratic Discriminant Analysis (QDA):** QDA is a similar method to LDA. The multivariate normality assumption remains, but unlike LDA, QDA assumes that each class has its own covariance matrix. If QDA has better predictive performance than LDA, it is an indication that a linear boundary is not the optimal to separate the 2 classes.
- **k-Nearest Neighbors (kNN):** kNN is a non-parametric approach which is highly flexible and uses non-linear decision boundaries. Before the implementation of kNN, it is crucial to scale the data, since this method relies on the euclidean distance between observations;
- **Classifications Trees:** Classification trees stratify the feature space recursively into simple regions and assign the label female or male using the majority vote. Bagging, random forests and gradient tree boosting are also implemented with the aim of improving predictive performance;
- **Support Vector Machines (SVM):** SVM perform well in classification problems where there is a clear margin of separation. We experiment with 2 kernels: the linear and the radial basis function (RBF).

4.2 Classification based on the fundamental frequency

Based on the exploratory data analysis described in Chapter 3, we propose to perform the classification based on the "meanfun" feature alone. This will give a baseline for further analysis described in the remaining of this Chapter.

To perform the analysis, we randomly split the dataset into a training set (80 %) and a test set (20 %). The training set is used to fit the models and for best parameter selection if needed. The test set is used to compute the classification error and to compare the models. The classification error considered in the study is the $0 - 1$ loss. The experiments have been performed on Python 3[1] with a single seed number for reproducibility of the results.

Table 4.1 Classification Error of the Methods for Classification With "meanfun" Feature

Type	Methods	
Max. Likelihood	Logistic reg.	0.0536
	Logistic reg. - Ridge	0.0505
	LDA	**0.0489**
	QDA	**0.0489**
Trees	Tree	0.0505
	Bagging	0.0505
	XGBoost	0.0505
SVM	Linear	0.0505
	Gaussian	**0.0489**
x	kNN	0.0520

The results, summarized in Table 4.1, show that the classification error is already very low when considering only the "meafun" feature. Indeed the average classsification error of the different classifiers is 0.0509. Thus, "meafun" is a very good feature for classification which is in accordance with the exploratory data analysis described in Chapter 3.

Regarding the classifiers, it can be seen that some of the ones described in Section PUTREF, are not mentioned in Tabel 4.1. Indeed, they do not make sense when considering only one feature for classification. About the relative performance of the different classifiers, the results are homogeneous around the average classification error and all the classifiers perform well. Zero-order methods such as tree-based methods (with one predictor, tree-based methods are nothing else than a threshold) already give a low classification error. More sophisticated methods, such as linear methods or kernel SVM, give results very similar to or slightly better than zero-order methods.

4.3 Classification based on a $80/20$ split of the dataset

4.3.1 Description

4.3.2 Results

4.4 Classification based on a $50/50$ split of the dataset

[1] `https://github.com/AdriBesson/Statistical_learning_course/tree/develop/project`

Table 4.2 Classification Error of the Methods for Different Seed Numbers With 80/20 Split

Type	Methods	Seed number 1	2	3	4	5
Max. Likelihood	Logistic reg.	0.0158	0.0347	0.0315	0.0237	0.0221
	Logistic reg. - Ridge	0.0158	0.0315	0.0315	0.0189	0.0284
	Logistic reg. - Lasso	0.0315	0.0363	0.0379	0.0284	0.0300
	LDA	0.0315	0.0410	0.0379	0.0284	0.0268
	QDA	0.0347	0.0347	0.0347	0.0268	0.0363
Trees	Tree	0.0379	0.0426	0.0315	0.0284	0.0300
	Pruned Tree	0.0394	0.0473	0.0347	0.0363	0.0300
	Bagging	0.0237	0.0410	0.0142	**0.0174**	0.0284
	Random Forest	0.0189	0.0347	**0.0126**	0.0205	0.0205
	XGBoost	0.0189	**0.0268**	**0.0126**	0.0205	**0.0189**
SVM	Linear	**0.0142**	0.0315	0.0284	0.0189	0.0252
	Gaussian	0.0205	0.0284	0.0126	0.0189	0.0221
x	kNN	0.0300	0.0347	0.0252	0.0379	0.0300

Table 4.3 Classification Error of the Methods for Different Seed Numbers With 50/50 Split

Type	Methods	Seed number 1	2	3	4	5
Max. Likelihood	Logistic reg.	0.0221	0.0262	0.0310	0.0261	0.0291
	Logistic reg. - Ridge	0.0196	0.0214	0.0305	0.0248	0.0278
	Logistic reg. - Lasso	0.0284	0.0294	0.0368	0.0348	0.0364
	LDA	0.0291	0.0269	0.0342	0.0309	0.0313
	QDA	0.0280	0.0257	0.0370	0.0362	0.0365
Trees	Tree	0.0284	0.0317	0.0387	0.0454	0.0468
	Pruned Tree	0.0394	0.0473	0.0347	0.0363	0.0300
	Bagging	0.0209	0.0244	0.0319	0.0253	0.0302
	Random Forest	0.0191	0.0206	0.0305	0.0248	0.0290
	XGBoost	**0.0170**	**0.0149**	**0.0259**	**0.0197**	**0.0227**
SVM	Linear	0.0209	0.0149	0.0259	0.0197	0.0227
	Gaussian	0.0184	0.0188	0.0272	0.0211	0.0234
x	kNN	0.0273	0.0312	0.0326	0.0410	0.0382

5 Application of the best classifier on an acquired dataset

Bibliography

[1] K. Wu and D. G. Childers, “Gender recognition from speech. Part I: Coarse analysis”, *J. Acoust. Soc. Am.*, vol. 90, pp. 1828–1840, 1991.

[2] D. G. Childers and K. Wu, “Gender recognition from speech. Part II: Fine analysis”, *J. Acoust. Soc. Am.*, vol. 90, pp. 1841–1856, 1991.

[3] D. Childers, Ke Wu, K. Bae, and D. Hicks, “Automatic recognition of gender by voice”, in *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, 1988, pp. 603–606.

[4] H. Harb and L. Chen, “Voice-Based Gender Identification in Multimedia Applications”, *J. Intell. Inf. Syst.*, vol. 24, pp. 179–198, 2005.

[5] Y.-m. Zeng, Z.-y. Wu, T. Falk, and W.-y. Chan, “Robust GMM Based Gender Classification using Pitch and RASTA-PLP Parameters of Speech”, in *2006 Int. Conf. Mach. Learn. Cybern.*, 2006, pp. 3376–3379.

[6] V. N. Sorokin and I. S. Makarov, “Gender recognition from vocal source”, *Acoust. Phys.*, vol. 54, pp. 571–578, 2008.

[7] Fl. Metze, J. Ajmera, R. Englert, U. Bub, F. Burkhardt, J. Stegmann, *et al.*, “Comparison of Four Approaches to Age and Gender Recognition for Telephone Applications”, in *2007 IEEE Int. Conf. Acoust. Speech Signal Process.*, 2007, pp. 1089–1092.

[8] T. Bocklet, A. Maier, J. G. Bauer, F. Burkhardt, and E. Noth, “Age and gender recognition for telephone applications based on GMM supervectors and support vector machines”, in *2008 IEEE Int. Conf. Acoust. Speech Signal Process.*, 2008, pp. 1605–1608.

[9] G. Levi and T. Hassncer, “Age and gender classification using convolutional neural networks”, in *2015 IEEE Conf. Comput. Vis. Pattern Recognit. Work.*, 2015, pp. 34–42.

[10] I. Bisio, A. Delfino, F. Lavagetto, M. Marchese, and A. Sciarrone, “Gender-Driven Emotion Recognition Through Speech Signals For Ambient Intelligence Applications”, *IEEE Trans. Emerg. Top. Comput.*, vol. 1, pp. 244–257, 2013.

[11] D. Ververidis and C. Kotropoulos, “Automatic speech classification to five emotional states based on gender information”, in *Eur. Signal Process. Conf.*, 2004.

[12] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, “Speaker normalization on conversational telephone speech”, in *1996 IEEE Int. Conf. Acoust. Speech, Signal Process. Conf. Proc.*, vol. 1, 1996, pp. 339–341.

[13] I. R. Titze, “Physiologic and acoustic differences between male and female voices”, *J. Acoust. Soc. Am.*, vol. 85, pp. 1699–1707, 1989.

[14] H. Hollien and E. Malcik, “Evaluation of gross-sectional studies of adolescent voice change in males”, *Speech Monogr.*, vol. 34, pp. 80–84, 1967.

[15] M. S. F. Poon and M. L. Ng, "The role of fundamental frequency and formants in voice gender identification", *Speech, Lang. Hear.*, vol. 18, pp. 161–165, 2015.

[16] M. Gupta, S. S. Bharti, and S. Agarwal, "Support vector machine based gender identification using voiced speech frames", in *2016 Fourth Int. Conf. Parallel, Distrib. Grid Comput.*, 2016, pp. 737–741.

[17] R. R. Nunes, M. P. de Almeida, and J. W. Sleigh, "Spectral entropy: A new method for anesthetic adequacy", *Rev. Bras. Anestesiol.*, vol. 54, 2004.

[18] H. Traunmüller, "The frequency range of the voice fundamental in the speech of male and female adults", Tech. Rep., 1994.