# EM algorithm based on GMM to extract voice features and identify male and female voices

## 1.Theory

### 1.1.EM algorithm

For the sample $x \sim F(x,\theta)$, there are$x = (x^{(1)}, x^{(2)}, \ldots, x^{(n)})$, if the$MLE$method is used to find the unknown parameter$\theta$, then

$\hat{\theta} = \underset{\theta}{argmax} L(\theta) = \underset{\theta}{argmax} \sum_{i=1}^{n} lnp(x^{(i)}, \theta)$, but if the samples come from several same distributions with different parameters, it cannot be obtained by the $MLE$method, and each

A sample $x^{(i)}$has a probability of $z^{(k)}$obeying a certain distribution, $z^{(k)}$is also a random variable, obeying a multinomial distribution,

There are $z = (z^{(1)}, z^{(2)}, \ldots, z^{(k)})$, set $P(z^{(i)}) = Q_i(z^{(i)})$,

Then there is $\sum_{z^{(i)}} Q_i(z^{(i)}) = 1$, and for solving $\theta$, $z$ is a hidden variable, so how to solve the unknown parameter $\theta$?

$$\hat{\theta} = \underset{\theta}{argmax} \sum_{i=1}^{n} lnp(x^{(i)}, \theta) = \underset{\theta}{argmax} \sum_{i=1}^{n} ln[\sum_{z^{(i)}} P(z^{(i)}) p(x^{(i)}, \theta | z^{(i)})]$$

$$= \underset{\theta}{argmax} \sum_{i=1}^{n} ln[\sum_{z^{(i)}} p(x^{(i)}, z^{(i)}, \theta)]$$

Then transform $\hat{\theta}$, $\hat{\theta} = \underset{\theta}{argmax} \sum_{i=1}^{n} ln[\sum_{z^{(i)}} Q_i(z^{(i)}) \frac{p(x^{(i)}, z^{(i)}, \theta)}{Q_i(z^{(i)})}]$, from $Jensen$ inequality, if $f''(x) \geq 0$, there is $Ef(x) \geq f(Ex), (lnx)'' = -\frac{1}{x^2} < 0, Elnx \leq ln(Ex)$, it can be seen that $\sum_{z^{(i)}} Q_i(z^{(i)}) \frac{p(x^{(i)}, z^{(i)}, \theta)}{Q_i(z^{(i)})}$ is the expectation of $\frac{p(x^{(i)}, z^{(i)}, \theta)}{Q_i(z^{(i)})}$ about $z^{(i)}$.

$$\sum_{i=1}^{n} ln[\sum_{z^{(i)}} Q_i(z^{(i)}) \frac{p(x^{(i)}, z^{(i)}, \theta)}{Q_i(z^{(i)})}] = \sum_{i=1}^{n} ln[E \frac{p(x^{(i)}, z^{(i)}, \theta)}{Q_i(z^{(i)})}] \geq \sum_{i=1}^{n} Eln \frac{p(x^{(i)}, z^{(i)}, \theta)}{Q_i(z^{(i)})}$$

$$= \sum_{i=1}^{n} \sum_{z^{(i)}} Q_i(z^{(i)}) ln \frac{p(x^{(i)}, z^{(i)}, \theta)}{Q_i(z^{(i)})}$$

$$= \sum_{i=1}^{n} \sum_{z^{(i)}} Q_i(z^{(i)}) lnp(x^{(i)}, z^{(i)}, \theta) - \sum_{i=1}^{n} \sum_{z^{(i)}} Q_i(z^{(i)}) lnQ_i(z^{(i)})$$

$\hat{\theta} = \underset{\theta}{argmax} \left[ \sum_{i=1}^{n} \sum_{z^{(i)}} Q_i(z^{(i)}) lnp(x^{(i)}, z^{(i)}, \theta) - \sum_{i=1}^{n} \sum_{z^{(i)}} Q_i(z^{(i)}) lnQ_i(z^{(i)}) \right]$, It can be seen that the second half of the above formula has nothing to do with $\theta$.

$\hat{\theta} = \underset{\theta}{argmax} \sum_{i=1}^{n} \sum_{z^{(i)}} Q_i(z^{(i)}) lnp(x^{(i)}, z^{(i)}, \theta)$

But note that we are here if and only if the $Jensen$ inequality is true, that is, $P(X = EX) = 1$, that is, when $X$ is considered to be a constant, put it here, that is, $\frac{p(x^{(i)}, z^{(i)}, \theta)}{Q_i(z^{(i)})} = c$ ($c$ is a constant),
$L(\theta) = \sum_{i=1}^{n} ln[E \frac{p(x^{(i)}, z^{(i)}, \theta)}{Q_i(z^{(i)})}] \geq \sum_{i=1}^{n} Eln \frac{p(x^{(i)}, z^{(i)}, \theta)}{Q_i(z^{(i)})}$, if we maximize $L(\theta)$, by maximizing the right half of the formula That is to maximize the lower bound of $L(\theta)$, it may not be possible to find the maximum value of $L(\theta)$, unless the equal sign is established, that is, $\frac{p(x^{(i)}, z^{(i)}, \theta)}{Q_i(z^{(i)})} = c$, which is actually the $E$ step in the $EM$ algorithm.

If you define $\frac{p(x^{(i)}, z^{(i)}, \theta)}{Q_i(z^{(i)})} = c$, $p(x^{(i)}, z^{(i)}, \theta) = c * Q_i(z^{(i)})$, sum $z^{(i)}$ on both sides,
$\therefore \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}, \theta) = c * \sum_{z^{(i)}} Q_i(z^{(i)}) = c$

$\frac{p(x^{(i)}, z^{(i)}, \theta)}{Q_i(z^{(i)})} = \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}, \theta), Q_i(z^{(i)}) = \frac{p(x^{(i)}, z^{(i)}, \theta)}{\sum_{z^{(i)}} p(x^{(i)}, z^{(i)}, \theta)} = \frac{p(x^{(i)}, z^{(i)}, \theta)}{\sum_{z^{(i)}} p(z^{(i)}) p(x^{(i)}, \theta | z^{(i)})} = p(z^{(i)} | x^{(i)}, \theta)$

if and only if $\frac{p(x^{(i)}, z^{(i)}, \theta)}{Q_i(z^{(i)})} = c$ is a fixed value, there are:

$$\hat{\theta} = \underset{\theta}{argmax} \sum_{i=1}^{n} \sum_{z^{(i)}} p(z^{(i)}|x^{(i)}, \theta) * lnp(x^{(i)}, z^{(i)}, \theta)$$

> It can be seen that $\sum_{i=1}^{n} \sum_{z^{(i)}} p(z^{(i)}|x^{(i)}, \theta) * lnp(x^{(i)}, z^{(i)}, \theta)$ is the conditional expectation of $lnp(x^{(i)}, z^{(i)}, \theta)$ about $z|x, \theta$, that is, when the sample and When the unknown parameter $\theta$ is given, since we don't know the value of the hidden variable $z$, we replace it with expectation. Of course, under the conditions of sample $x$ and parameter $\theta$, after calculating the conditional expectation, we go to maximize $L(\theta)$ about $\theta$. This is the basic idea of $EM$ algorithm.
>
> Therefore, our $E$ step is to randomly set $\theta^j$ as the initial value, of course, this initial value also includes the hidden variable $z$, and then use the sample $x$ to find the conditional probability $Q_i(z^{(i)}) = p(z^{(i)}|x^{(i)}, \theta^j), \theta^j), j = 1, 2, \ldots N$ is the number of iterations, $\theta^j$ is different from $\theta$, $\theta$ It is the optimal value that God knows. $\theta^j$ is the value of $\theta$ that is constantly iteratively approaching $\theta$. After finding out $Q_i(z^{(i)})$, it is substituted into $L(\theta) = \sum_{i=1}^{n} \sum_{z^{(i)}} Q_i(z^{(i)}) * lnp(x^{(i)}, z^{(i)}, \theta)$, and then $\theta$ Find partial derivatives, maximize $L(\theta)$, get $\theta^{j+1}$, this is $M$ step. Then replace $\theta^{j+1}$ with the original $\theta^j$, and then re-execute $E$ step and iterate continuously until $L(\theta)$ converges.

## 1.2. Convergence proof of EM algorithm

So how do you know that $L(\theta)$ will converge (maximize) according to this method?

That is to prove that $L(\theta, \theta^{j+1}) \geq L(\theta, \theta^j)$ can show that $L(\theta, \theta^j)$ is getting bigger every time.

$L(\theta, \theta^j) = \sum_{i=1}^{n} \sum_{z^{(i)}} p(z^{(i)}|x^{(i)}, \theta^j) * lnp(x^{(i)}, z^{(i)}, \theta)$, ①

定义$H(\theta, \theta^j) = \sum_{i=1}^{n} \sum_{z^{(i)}} p(z^{(i)}|x^{(i)}, \theta^j) * lnp(z^{(i)}|x^{(i)}, \theta)$, ②

①−②:

$$L(\theta, \theta^j) - H(\theta, \theta^j) = \sum_{i=1}^{n} \sum_{z^{(i)}} p(z^{(i)}|x^{(i)}, \theta^j) * ln\frac{p(x^{(i)}, z^{(i)}, \theta)}{p(z^{(i)}|x^{(i)}, \theta)} = \sum_{i=1}^{n} \sum_{z^{(i)}} p(z^{(i)}|x^{(i)}, \theta^j) * lnp(x^{(i)}, \theta)$$
$$= \sum_{i=1}^{n} lnp(x^{(i)}, \theta) \sum_{z^{(i)}} p(z^{(i)}|x^{(i)}, \theta^j) = \sum_{i=1}^{n} lnp(x^{(i)}, \theta)$$

即$\sum_{i=1}^{n} lnp(x^{(i)}, \theta) = L(\theta, \theta^j) - H(\theta, \theta^j)$

$$\sum_{i=1}^{n} lnp(x^{(i)}, \theta^{j+1}) - \sum_{i=1}^{n} lnp(x^{(i)}, \theta^j) = [L(\theta^{j+1}, \theta^j) - H(\theta^{j+1}, \theta^j)] - [L(\theta^j, \theta^j) - H(\theta^j, \theta^j)]$$
$$= [L(\theta^{j+1}, \theta^j) - L(\theta^j, \theta^j)] - [H(\theta^{j+1}, \theta^j) - H(\theta^j, \theta^j)]$$

The first half of the definition is $A$, $A = L(\theta^{j+1}, \theta^j) - L(\theta^j, \theta^j)$, from $\theta^j$ into $L(\theta)$, to $\theta^{j+1}$ into $L(\theta)$, it is constantly maximizing $L(\theta)$, which belongs to $M$ step, so $L(\ theta^{j+1}, \theta^j)$ must be greater than $L(\theta^j, \theta^j)$, $A \geq 0$

The second half of the definition is $B$,

$$B = H(\theta^{j+1}, \theta^j) - H(\theta^j, \theta^j) = \sum_{i=1}^{n} \sum_{z^{(i)}} p(z^{(i)}|x^{(i)}, \theta^j) * lnp(z^{(i)}|x^{(i)}, \theta^{j+1}) - \sum_{i=1}^{n} \sum_{z^{(i)}} p(z^{(i)}|x^{(i)}, \theta^j) * lnp(z^{(i)}|x^{(i)}, \theta^j)$$
$$= \sum_{i=1}^{n} \sum_{z^{(i)}} p(z^{(i)}|x^{(i)}, \theta^j) * ln\frac{p(z^{(i)}|x^{(i)}, \theta^{j+1})}{p(z^{(i)}|x^{(i)}, \theta^j)} = \sum_{i=1}^{n} Eln\frac{p(z^{(i)}|x^{(i)}, \theta^{j+1})}{p(z^{(i)}|x^{(i)}, \theta^j)} \leq \sum_{i=1}^{n} lnE\frac{p(z^{(i)}|x^{(i)}, \theta^{j+1})}{p(z^{(i)}|x^{(i)}, \theta^j)}$$
$$= \sum_{i=1}^{n} ln \sum_{z^{(i)}} p(z^{(i)}|x^{(i)}, \theta^j)\frac{p(z^{(i)}|x^{(i)}, \theta^{j+1})}{p(z^{(i)}|x^{(i)}, \theta^j)} = \sum_{i=1}^{n} ln \sum_{z^{(i)}} p(z^{(i)}|x^{(i)}, \theta^{j+1}) = \sum_{i=1}^{n} ln1 = 0$$

$\big($The $Jensen$ inequality is used in the above formula, $Elnx \leq ln(Ex)$ $\big)$, $B \leq 0$

$A \geq 0, B \leq 0$, the number greater than or equal to 0 minus the number less than or equal to 0 must be greater than or equal to 0, $L(\theta, \theta^{j+1}) \geq L(\theta, \theta^j)$, $$likelihood function converges

## 1.3.EM算法的步骤

## 1.3. The steps of EM algorithm

There are samples $x=(x^{(1)},x^{(2)},\ldots,x^{(n)})$, hidden variables $z=(z^{(1)},z^{(2)},\ldots,z^{(n)})$, set the maximum number of iterations to $N$, unknown parameters $\theta=(\theta_1,\theta_2,\ldots,\theta_m,z)$, note that $z$ is also an unknown parameter

1. Take $\theta_{iter}=\theta^0$, the initial number of iterations $n_{iter}=0$
2. $while\, n_{iter}\le N$ :

$$Estep=\begin{cases}last\theta_{iter}=\theta_{iter}\\ Q_i(z^{(i)})=p(z^{(i)}|x^{(i)},\theta_{iter})\\ L(\theta,\theta_{iter})=\sum_{i=1}^{n}\sum_{z^{(i)}}Q_i(z^{(i)})lnp(x^{(i)},z^{(i)},\theta_{iter})\end{cases}$$

$$Mstep=\begin{cases}\theta_{iter}=\underset{\theta}{argmax}L(\theta,\theta_{iter})\\ if\quad |L(\theta,\theta_{iter})-L(\theta,last\theta_{iter})|<\epsilon\\ \quad\quad break\end{cases}$$

$n=n+1$

$return\quad \theta_{iter}$

## 1.4. The solution process of GMM mixed Gaussian distribution in EM algorithm

Existing samples $x=(x_1,x_2,\ldots,x_n)$, each individual sample is a two-dimensional vector, that is, $x_i=(x_i^{(1)},x_i^{(2)})$

However, two Gaussian distributions are mixed in the sample, and the mean variance and correlation coefficient of the two Gaussian distributions are unknown, that is, the unknown parameter is $\theta=(u,\sigma^2,\rho)$

And each sample has the probability of $Q_i(z^{(i)})$ obeying a two-dimensional Gaussian distribution, and the two-dimensional Gaussian density is:

$$\varphi(x^{(1)},x^{(2)})=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}exp\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x^{(1)}-u^{(1)})^2}{\sigma_1^2}-2\rho\frac{(x^{(1)}-u^{(1)})(x^{(2)}-u^{(2)})}{\sigma_1\sigma_2}+\frac{(x^{(2)}-u^{(2)})^2}{\sigma_2^2}\right]\}$$

The maximum likelihood process of $$ adding hidden variables is:

$$\hat{\theta}=\underset{\theta}{argmax}\sum_{i=1}^{n}\sum_{j=1}^{2}p(z_{ji}|x_i,\theta)*lnp(x_i,z_ji,\theta),\quad x_i=(x_i^{(1)},x_i^{(2)})$$

Where $p(z_{ji}|x_i,\theta)=\frac{p(z_ji)p(x_i,\theta|z_ji)}{p(z_{1i})p(x_i,\theta|z_1i)+p(z_{2i})p(x_i,\theta|z_2i)},j=1,2;i=1,2,\ldots,n$

The matrix of conditional distribution of latent variables is:

| | $x_1$ | $x_2$ | ... | $x_n$ |
|---|---|---|---|---|
| $p(z_1)$ | $p(z_{11}\|x_1,\theta)$ | $p(z_{12}\|x_2,\theta)$ | ... | $p(z_{1n}\|x_n,\theta)$ |
| $p(z_2)$ | $p(z_{21}\|x_1,\theta)$ | $p(z_{22}\|x_2,\theta)$ | ... | $p(z_{2n}\|x_n,\theta)$ |
| $sum$ | 1 | 1 | ... | 1 |

For each $x_i=(x_i^{(1)},x_i^{(2)})\sim N(u^{(1)},u^{(2)};\sigma^{2(1)},\sigma^{2(2)};\rho)$, we first need to set the initial value:

$u^{(0)}=(u_{11},u_{12};u_{21},u_{22}),\quad \sigma^{2(0)}=(\sigma_{11}^2,\sigma_{12}^2;\sigma_{21}^2,\sigma_{22}^2),\quad \rho^{(0)}=(\rho_1,\rho_2),\quad p(z)^{(0)}=(p(z_1),p(z_2))$

By calculating the partial derivative method of the likelihood function, we can get:

$E$ step为:

$$p(z_1)^{(1)}=\frac{\sum_{i=1}^{n}p(z_1|x_i,\theta)}{n},\qquad p(z_2)^{(1)}=\frac{\sum_{i=1}^{n}p(z_2|x_i,\theta)}{n}$$

$M$ step为:

$$u_{11}^{(1)}=\frac{\sum_{i=1}^{n}p(z_1|x_i,\theta)x_i^{(1)}}{\sum_{i=1}^{n}p(z_1|x_i,\theta)},\quad u_{12}^{(1)}=\frac{\sum_{i=1}^{n}p(z_1|x_i,\theta)x_i^{(2)}}{\sum_{i=1}^{n}p(z_1|x_i,\theta)},\quad u_{21}^{(1)}=\frac{\sum_{i=1}^{n}p(z_2|x_i,\theta)x_i^{(1)}}{\sum_{i=1}^{n}p(z_2|x_i,\theta)},\quad u_{22}^{(1)}=\frac{\sum_{i=1}^{n}p(z_2|x_i,\theta)x_i^{(2)}}{\sum_{i=1}^{n}p(z_2|x_i,\theta)}$$

$$\sigma_{11}^{2(1)}=\frac{\sum_{i=1}^{n}p(z_1|x_i,\theta)(x_i^{(1)}-u_{11}^{(1)})^2}{\sum_{i=1}^{n}p(z_1|x_i,\theta)},\quad \sigma_{12}^{2(1)}=\frac{\sum_{i=1}^{n}p(z_1|x_i,\theta)(x_i^{(2)}-u_{12}^{(1)})^2}{\sum_{i=1}^{n}p(z_1|x_i,\theta)},\quad \sigma_{21}^{2(1)}=\frac{\sum_{i=1}^{n}p(z_2|x_i,\theta)(x_i^{(1)}-u_{21}^{(1)})^2}{\sum_{i=1}^{n}p(z_2|x_i,\theta)},\quad \sigma_{22}^{2(1)}=\frac{\sum_{i=1}^{n}p(z_2|x_i,\theta)(x_i^{(2)}-u_{22}^{(1)})^2}{\sum_{i=1}^{n}p(z_2|x_i,\theta)}$$

$$\rho_1^{(1)}=p(z_1)^{(1)}r,\quad \rho_2^{(1)}=p(z_2)^{(1)}r$$

iteration exit condition：$|L(\theta^{j+1})-L(\theta^j)|<\epsilon$

## 1.5. Evaluation method of EM algorithm

To measure the amount of information contained in the position parameter $\theta$ obtained by the EM algorithm is actually to find the $Fisher$ information amount of the sample, which is the variance of the likelihood equation. If the variance is larger, it means that the sample collects more information, which means that the $\hat{\theta}$ estimated by the sample is more representative of the population.

However, the general $Fisher$ information is relatively easy to find, and the likelihood equation here contains hidden variables, and its derivation is as follows:

First of all, the sample likelihood function that includes hidden variables, has been logarithmically taken, and is derived is $S(x,\theta)=\sum_{i=1}^{n}\frac{\partial\sum_z p(z|x_i,\theta^j)*lnp(x_i,z,\theta)}{\partial\theta}$, we call it the likelihood equation, Now we ask for the variance of $S(x,\theta)$, which can measure how much $EM$ algorithm contains $\theta$ information. Note here that $\theta^j$ in $p(z|x_i,\theta^j)$ is different from $\theta$ in $lnp(x_i,z,\theta)$.

$$E\big(S(x,\theta)\big)=\sum_{i=1}^{n}\sum_z p(z|x_i,\theta^j)\int_{-\infty}^{+\infty}\frac{\partial lnp(x_i,z,\theta)}{\partial\theta}p(x_i,z,\theta)dx=\sum_{i=1}^{n}\sum_z p(z|x_i,\theta^j)\int_{-\infty}^{+\infty}\frac{1}{p(x_i,z,\theta)}\frac{\partial p(x_i,\theta)}{\partial\theta}p(x_i,\theta)dx$$

$$=\sum_{i=1}^{n}\sum_z p(z|x_i,\theta^j)\frac{\partial}{\partial\theta}\int_{-\infty}^{+\infty}p(x_i,z,\theta)dx$$

(Derivation and integration are used in the above formula to exchange order)

$E\big(S(x,\theta)\big)=\sum_{i=1}^{n}\sum_z p(z|x_i,\theta^j)\frac{\partial}{\partial\theta}*1=0$, namely $E\big(S(x,\theta)\big)=0$

$$Var\big(S(x,\theta)\big)=E\big(S(x,\theta)\big)^2-[E\big(S(x,\theta)\big)]^2=E\big(S(x,\theta)\big)^2-0=E\big(S(x,\theta)\big)^2$$

$Var\big(S(x,\theta)\big)=E\big(S(x,\theta)\big)^2=\sum_{i=1}^{n}E\left(\frac{\partial\sum_z p(z|x_i,\theta^j)*lnp(x_i,z,\theta)}{\partial\theta}\right)^2$, remember $sum_z p(z|x_i,\theta^j)$ is $\alpha_i$, which represents the weight of each distribution, then:

$Var\big(S(x,\theta)\big)=nE\left(\alpha_1\frac{\partial lnp(x,z_1,\theta)}{\partial\theta}+\alpha_2\frac{\partial lnp(x,z_2,\theta)}{\partial\theta}+\ldots+\alpha_n\frac{\partial lnp(x,z_n,\theta)}{\partial\theta}\right)^2=nI(\theta)$ is the variance of the likelihood equation of the $EM$ algorithm, the larger the variance, that is, the larger the $I(\theta)$, the better the estimate.

# 2. The specific application of recognizing male and female voices

## 2.1. Extracting sound features

First download an open source Chinese voice data set ST-CMDS-20170001_1-OS. ST-CMDS is a Chinese voice data set released by an AI data company, which contains more than 100,000 voice files and about 100 hours of voice data. The data content is mainly based on the usual online voice chat and intelligent voice control sentences, with 855 different speakers, both male and female, suitable for use in various scenarios. The foreign mirror address is: https://link.ailemon.net/?target=http://www.openslr.org/resources/38/ST-CMDS-20170001_1-OS.tar.gz

This data set contains 102,600 pieces of voice data, each voice starts at about the first second and lasts for 3 seconds, mainly for life-related chats, but these 100,000 voices are not marked, we can imagine that the voice characteristics that are conducive to gender distinction are random variables that obey the multidimensional Gaussian distribution. Obviously, in this problem, we have two multidimensional Gaussian distributions mixed together. Our research purpose is to separate the two distributions and calculate the two multidimensional Mean vector and covariance matrix of a Gaussian distribution.

Then, give a voice at random, extract the features, and calculate the probability that it belongs to two distributions, whichever probability is greater, it is considered which distribution the feature belongs to, and the purpose of identifying male and female voices is achieved.

But we have to solve two problems first:

**1. Traverse voice files**

**2. Extract sound features that are conducive to gender distinction**

The first point can be solved by the $Python\ OS$ library. The second point is the dataset of voice features marked with gender** on $kaggle$. The semantics of each voice feature are:

| feature | intepretation |
| --- | --- |
| meanfreq | mean frequency in kHz |

| feature | intepretation |
|---|---|
| sd | standard deviation of frequency |
| median | median frequency in kHz |
| Q25 | First quantile of frequency in kHz |
| Q75 | third quantile of frequency in kHz |
| IQR | Frequency Interquartile Range |
| skew | frequency skewness |
| kurt | frequency kurtosis |
| sp.ent | spectral entropy |
| sfm | Spectral Flatness |
| model | frequency model |
| centroid | frequency centroid |
| peakf | peak frequency |
| meanfun | the mean value of the fundamental frequency measured on an acoustic signal |
| minfun | minimum fundamental frequency measured on an acoustic signal |
| maxfun | maximum fundamental frequency measured on an acoustic signal |
| meandom | the mean value of the dominant frequency measured on an acoustic signal |
| mindom | The minimum value of the dominant frequency measured on an acoustic signal |
| maxdom | maximum value of the dominant frequency measured on the acoustic signal |
| dfrange | dominant frequency range measured on the acoustic signal |
| modindx | modulation index, calculated as the cumulative absolute difference between adjacent measurements of the fundamental frequency divided by the frequency range |

We can use $XGboost$ to train with the marked data set, and get the sound feature weight order of the ability to distinguish between male and female voices:

It can be seen that the sound features $meanfun$ and $IQR$ have a greater contribution to distinguishing gender. For other independent variables, it is necessary to observe whether there is multicollinearity:

Considering the contribution and correlation coefficient comprehensively, we choose $meanfun, IQR$ as independent variables of voice characteristics, $meanfun$ as the fundamental frequency value, the fundamental frequency value of boys is about $0 \sim 150hz$, and the fundamental frequency value of girls is about $100 \sim 500hz$, the fundamental frequency can well distinguish male and female voices; $IQR$ is the interquartile range of sound frequency, which reflects the frequency distribution of frequency extreme values.

The picture above is a **spectrum** of a randomly selected sound. It is an image based on **Short-Time Fourier Transform (STFT)** that brings together three types of information: **Frequency**, **Time** and **Amplitude**. The horizontal axis represents time, the vertical axis represents frequency, and the color of the frequency line represents energy (reflecting amplitude). The logarithmic scale is used here.

> There are also linear scale and Mel scale. The reason why there are different scales is that human ears have different recognition capabilities for sounds in different frequency bands. Human ears can distinguish sounds with low frequencies such as $500 \sim 1000hz$ very well, but they cannot hear the difference for sounds such as $20000 \sim 20500hz$. Therefore, the linear scale is simply to show the frequency change, and the logarithmic scale is not so steep, while the Mel scale is to transform the sound of each frequency band The changes are all scaled to the same scale that the human ear can perceive.

The **fundamental frequency** we want to extract here is the curve at the bottom of the picture above. It can be seen that the sound ranges from about 0.6 seconds to 3.3 seconds, and the fundamental frequency is about 128. It can be roughly guessed that it is a male voice.

Now it is necessary to traverse 102,600 voices, and extract the $meanfun$ and $IQR$ of each voice through the $librosa$ library. The calculation logic is as follows:

**1. The voice sampling rate in the file defaults to $22050hz$. In order to extract features more accurately, the sampling rate is converted to $44100hz$ when extracting the voice, and then the sound is converted to mono. Each voice only extracts the sound from the first second to the fourth second**

**2. $meanfun$ base frequency: each voice will convert 3 seconds of sound into 336 frames according to the sampling rate ($44100hz$), each frame has a base frequency value, and output the average base frequency**

**3.$IQR$ Interquartile range: Mel frequency cepstral coefficient matrix of each voice, extract the average value of each column, form an array of 336 elements, arrange from small to large, output interquartile range**