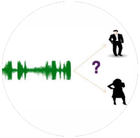
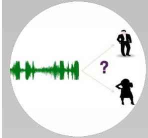


Final NLA Project
Voice Gender Detection using Tensor Power Method
MMODA Team:

Maxim Brazhnikov & Maxim Kuznetsov & Oleg Desheulin & Denis Rakitin & Anita Soloveva

December 16, 2020



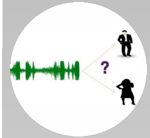
Final NLA Project Voice Gender Detection using Tensor Power Method **MMODA Team:**

Maxim Brazhnikov & Maxim Kuznetsov & Oleg Desheulin & Denis Rakitin & Anita Soloveva

Skoltech
Skolkovo Institute of Science and Technology

Problem Description

- Gender detection as one of the tasks of Speech Recognition, which has recently gained much popularity with the development of voice-based systems like Alexa, Siri and etc.
- Classical method for gender detection task is Gaussian Mixture Model (GMM) on Mel Frequency Cepstral Coefficients (MFCCs) (Neti & Roukos, 1997). Here we try another approach proposed by (Roy, Bhagath, & Das, 2020) which applies tensor power method to a tensor formed from MFCCs and compare this approach to GMM.

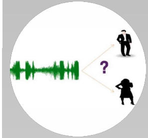


Our goals

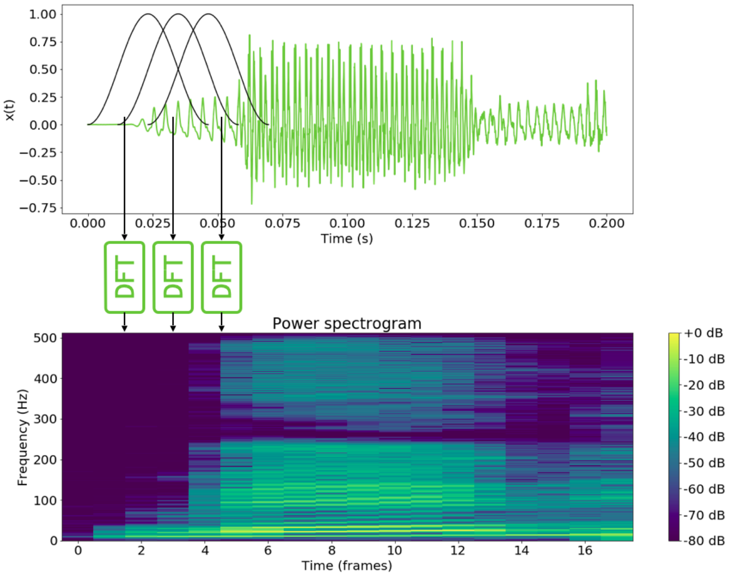
- Evaluate the accuracy of gender detection, perfomed by the approach proposed in (Roy et al., 2020), using different sizes of feature vectors, different number of eigenvectors, testing on two datasets:
  - TIMIT DR1 (New England dialect data)
  - SHRUTI (Bengali (minor Indo-Aryan language) data)

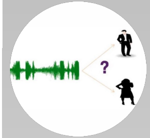
Additionally:

- Compare the performance with other approaches
- Test on our own Khanty (minor Finno-Ugric language) corpus



Mel-frequency cepstral coefficients





Final NLA Project
Voice Gender Detection using Tensor Power Method
MMODA Team:

Maxim Brazhnikov & Maxim Kuznetsov & Oleg Desheulin & Denis Rakitin & Anita Soloveva

Skoltech
Skolkovo Institute of Science and Technology

Tensor Formation

In order to build 3^{rd} order tensor we apply moment method.

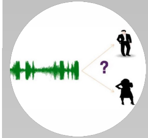
$$m_1 = \mathrm{E}[x] = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$M_2 = \mathrm{E}[x \odot x] - \sigma^2 I$$

where σ^2 is the smallest eigenvalue of covariance matrix $\Sigma = \mathrm{E}[x \odot x] - m_1 \odot m_1$ and M_2 can be decomposed as

$$M_2 = \sum_{i=1}^{r} w_i a_i \odot a_i \tag{1}$$

Naive approach is to calculate vectors a_i by eigendecomposition.



Tensor Formation

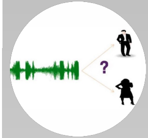
To avoid problems with highly correlated components authors suggests to calculate M_3 rather then M_2 using formulation (Hsu & Kakade, 2013):

$$M_3 = \mathrm{E}[x \odot x \odot x] - \sigma^2 \sum_{i=1}^{d} (m_1 \odot e_i \odot e_i + \cdots + e_i \odot e_i \odot m_1)$$

And like M_2 can be decomposed onto eigenvalue decomposition, for M_3 it also exists in following form:

$$M_3 = \sum_{i=1}^{r} w_i a_i \odot a_i \odot a_i$$

Where a_i is our target eigenvectors of M_2. Then eigenvectors of M_3 are also eigenvectors of M_2.



Eigenvectors Computation

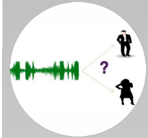
Similar to matrix eignevectors computation for tensors also exists Power Method:

$$a_{i,k+1} = \frac{A_i(I, a_{i,k}, a_{i,k})}{||A_i(I, a_{i,k}, a_{i,k})||_2}$$

And to project to next eigenvector space :

$$A_{i+1} = A_i - \lambda_i a_i \odot a_i \odot a_i$$

We apply this method to whitened M_3 and obtain our feature vectors.



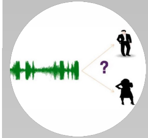
Evaluation

During training stage we found k dominant eigenvectors for both A_f (female) and A_m (male) dataset.
Then distances $D_{f,m}$ between unknown feature vector $x_i \in \mathcal{R}^n$ and eigenvectors $a_{\{f,m\};k}$ of each dataset A_f and A_m is calculated as:

$$D_{\{f,m\}} = \sum_{i=1}^{N} \min_k d(a_{\{f,m\};k}, x_i) \tag{2}$$

$$d(x,y) = 1 - \frac{(x,y)^2}{(x,x)(y,y)} \tag{3}$$

After than we compare D_m and D_f and choose the lowest distance to determine gender of the speaker.



Results with different sizes of feature vectors

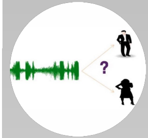
Size	Female accuracy	Male accuracy	Total accuracy
13	0.694	0.897	0.796
20	0.810	0.841	0.825
26	0.828	0.852	0.840

Table: Accuracy on Khanty dataset with 1 eigenvector.

Size	Female accuracy	Male accuracy	Total accuracy
13	0.992	0.777	0.884
20	0.989	0.919	0.954
26	0.989	0.953	0.971

Table: Accuracy on SHRUTI dataset with 1 eigenvector.

Tables are represent dependence of accuracy on size of feature vector.

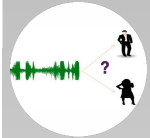


Results with different number of eigenvectors

#	Female accuracy	Male accuracy	Total accuracy
1	0.871	0.822	0.847
2	0.835	0.851	0.843
4	0.828	0.852	0.840

Table: Accuracy on Khanty dataset with 26-size feature vector.

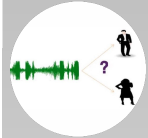
The tables is represent dependence of accuracy on the number of eigenvectors.



Comparison to other approaches

Method	Female accuracy	Male accuracy	Total accuracy
Article	0.871	0.822	0.847
Naive	0.802	0.856	0.829
GMM	0.968	0.914	0.945

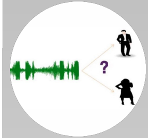
Table: Accuracy on Khanty dataset with the best choice of number of eigenvectors and size of feature vectors



Team Member contribution

- Maxim Brazhnikov: theoretical part, model building, presentation, report
- Maxim Kuznetsov: model testing, results analysis, presentation, report
- Oleg Desheulin: theoretical part, model building, presentation, report
- Denis Rakitin: model building, model testing, presentation, report
- Anita Soloveva: data preprocessing, model testing, presentation, report

Our GitHub repository



References

Hsu, D. J., & Kakade, S. M. (2013). Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Itcs '13*.

Neti, C., & Roukos, S. (1997). Phone-context specific gender-dependent acoustic-models for continuous speech recognition. *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, 192-198.

Roy, P., Bhagath, P., & Das, P. (2020). 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).