Detecting Voice Registers from Opera using Deep Learning

Christos A. Makridis and Bramwel O. Omondi *

April 2, 2022

Abstract

Using a sample of XX classical opera singers, we use artificial intelligence (AI) to predict the voice type of singers. We obtain a [accuracy metrics here]

Keywords:

JEL Codes:

*Christos: Columbia Business School and Stanford University, cmakridi@stanford.edu. We would like to thank ... These views are our own and do not reflect those of any affiliated institutions.

1 Introduction

Fine arts, specifically classical opera, contributes XX in GDP each year. For example, the Metropolitan Opera, located in New York City, has an annual budget of $300 million.[1] Moreover, there are XX new graduates who enter the labor market as classical opera singers each year. However, survey evidence suggests that as many as 80% of singers do not know their voice type, which makes it challenging, especially for new graduates, to book paid gigs. Singers have to know their voice type in order to curate the right portfolio of work so that they can represent themselves properly to casting directors

Unfortunately, identifying one's voice type is not a trivial effort. Moreover, because of a lack of competition and transparency within the market for voice lessons, voice teachers exhibit significant variability in the way that they classify a student's voice type. The inability to obtain reliable voice type recommendations, therefore, behaves as a significant barrier in the arts, particularly for low income and minority students who do not have access to premium voice lessons.

The primary contribution of this paper is to build a model for predicting a classical opera singer's voice type using artificial intelligence (AI) and use it to pilot an app that can detect a singer's voice type after three minutes of singing. We find...

2 Data and Measurement

The data gathered for this study contained 7,765 songs of different lengths. It contained 1,211 baritone, 462 bass, 273 contraltos, 321 countertenor, 1,060 mezzosopranos, 2,851 sopranos and

[1] https://www.npr.org/sections/coronavirus-live-updates/2020/03/19/818378901/the-metropolitan-opera-tells-its-union-employees-they-will-not-be-paid-after-mar

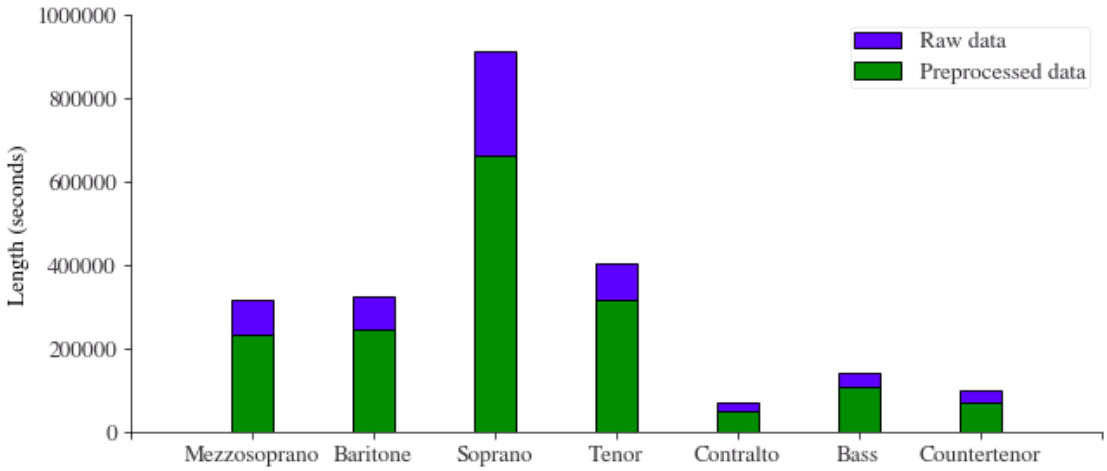
1,587 tenor registers. The bass register is further divided into bass cantante and bass dramatic. Baritones category is split between baritones bassbaritones, baritones dramatic and baritones lyric. Tenor is sub-categorized into tenor dramatic, tenor leggero, tenor spinto and tenor lyric. The mezzosoprano category comprises dramatic, lyric and coloratura. Sopranos get divided into five sub-categories namely; dramatic, lyric, coloratura, soubrette and spinto. Given the nature of these registers, all bass, baritone, tenor and countertenor registers were sung by male voices, and all contraltos, mezzosoprano and soprano registers were sung by female voices. In total, all tracks added up to more than 630 hours of recordings.

2.1 Data Preprocessing

One of the fundamental parts of our methodology is data preprocessing. Given that the tracks have different lengths, and noises during recording came from different origins - orchestra playing, secondary singers, and other noise sources - it is essential to retrieve only the vocals from the tracks and with as little noise as possible.

The first step is to extract only the vocals from the tracks, given that this is the only piece of information which distinguishes a register from another. The vocals were extracted using the open-source tool spleeter, which is a Tensorflow-based framework to preprocess data.Spleeter separates the vocals from the noise.

Given that there was a considerable amount of silence present in the data, these were removed using the open-source FFmpeg tool. Here, leading, intermediate and trailing zeros were removed from the tracks. After this step, the total length of the tracks was reduced of 26%, as shown in the Figure below.



Finally, we split the audios into same-length tracks, given that the algorithm will be trained with fixed-size labeled samples. The shortest track in the dataset has a length of 6s, so this was the length chosen to split all other tracks. In this sense, no data would be lost.

However, the six-second audio samples are not yet clean enough for the model. The next step entails listening to the samples and determining the cleanest ones to include in the final data-set. For example, in the case of multiple singers(duets) on the same audio track, the female voiced parts must be separated from male voiced parts. There are also samples that have more than three seconds of silence and must be removed. It is also noted that some few audio tracks such as 'Ave Maria' and 'Blow the Wind Southerly' are sung by different singers in different categories. Care has to be taken to not allow many of these samples in the final data-set because the model might learn to use the words sung as features.

In other cases, if it is discovered that a sample is clean, but it does not follow the general statistical/frequency distribution of its voice register, then it is also struck out. Ultimately, each sub-register has between 400 to 450 curated samples that are representative of the category they are placed in.

2.2 Feature Engineering

Here, MFFCs are used as features to train the model. The model will be trained with the feature images derived from the subtracks, and MFFCs are considered to be a good visual representation of audio as they contain information about the frequencies, dBs fpr each section of the tracks.

3 Statistical Strategy

Classifying voice registers, under a machine learning perspective, can be done following two different approaches. The first one is training a single model to discriminate between the registers. In this model, historically female and male voices could be misclassified into one or another. A second approach is training two models: one for the historically female registers, and another for the historically male registers. Here, a strict hypothesis is made, stating that all female voices belong to historically female registers, and that all male voices belong to historically male registers.

Here, historically female registers are Contralto, Mezzosoprano and Soprano. The historically male registers are Bass, Baritone, Tenor and Countertenor.

First, we create a baseline model which tries to classify each track following the lower and higher frequencies considered to represent each register. This is a rule-based approach, where the top 3 frequencies in each track will determine the register.

3.1 Baseline model

The split dataset has 40805 baritones, 17963 bass, 8579 contraltos, 11735 countertenor, 38300 mezzosopranos, 109316 sopranos and 51902 tenor subtracks. Each register has their intervals of

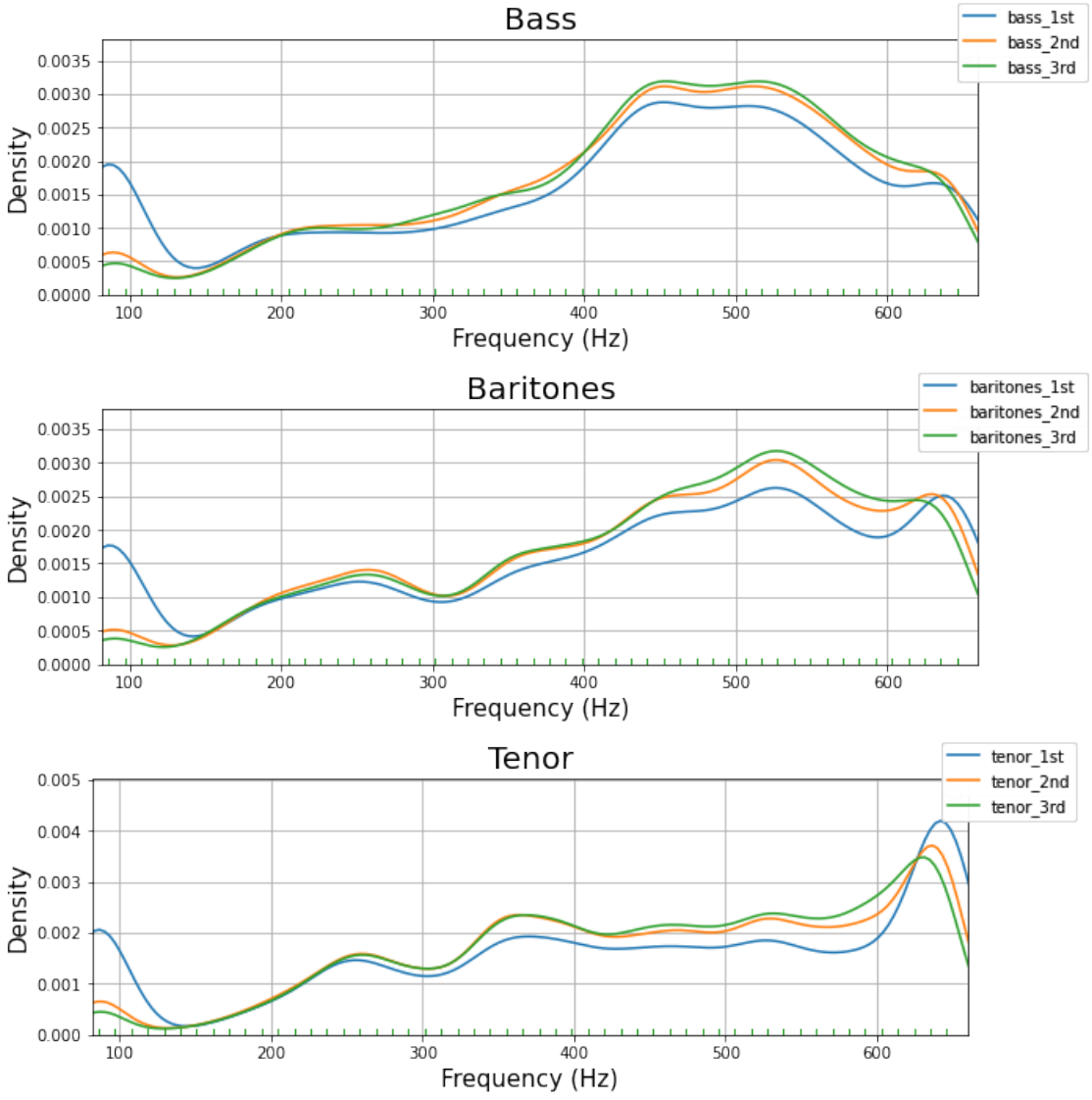
frequencies, noted as piano notes, as listed below.

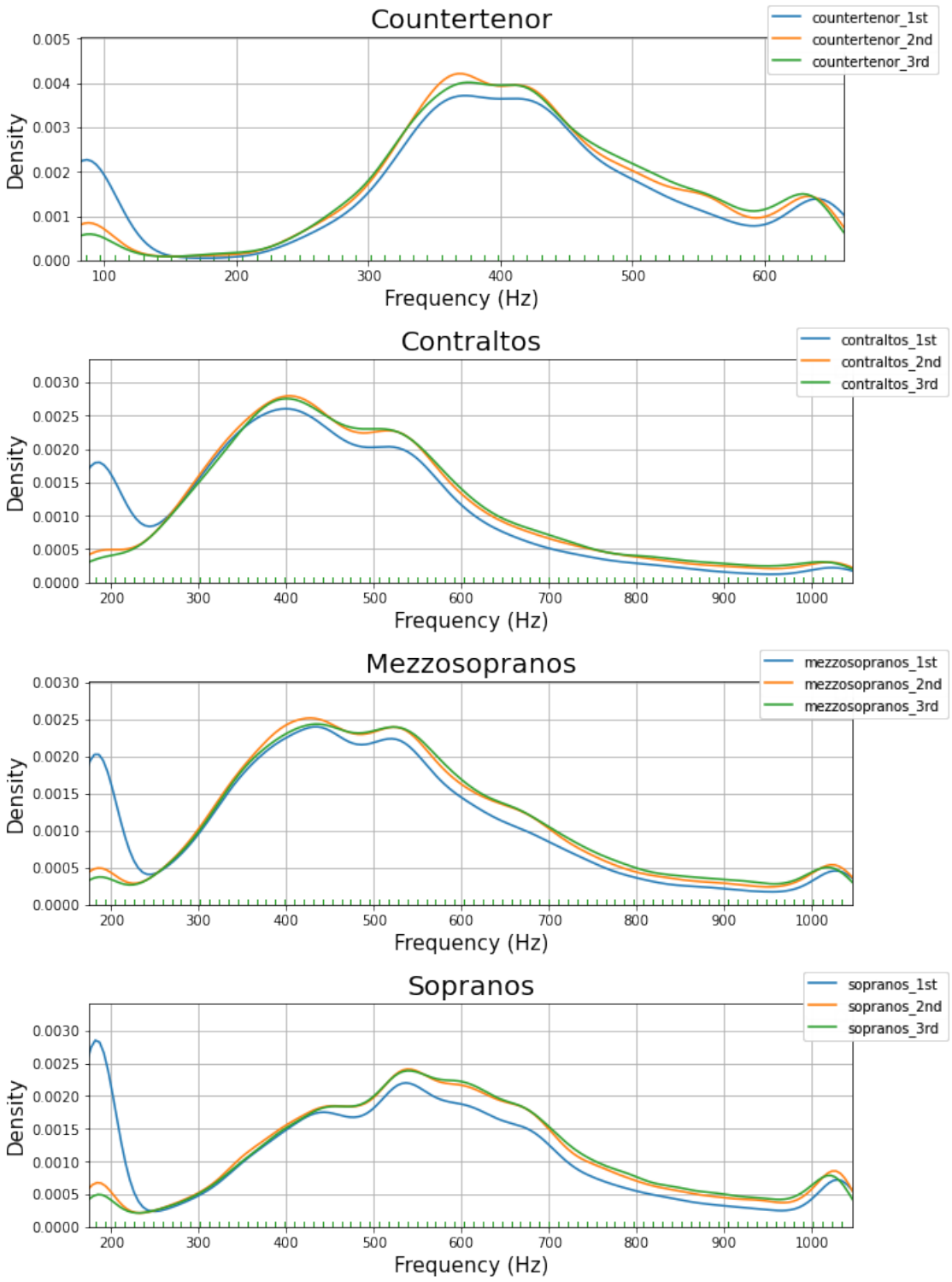
- Bass. This register is typically classified as having a vocal range extending from E2 to E4 (82.41-329.63Hz).
- Baritone. Composers typically write music for this voice in the range of F2 to F4 (87.31-349.23 Hz)in choral music, and from A2 to A4 (110.00-440.00 Hz) in operatic music.
- Tenor. The tenor's vocal range extends up to C5, and the low extreme is roughly A2 (110.00-523.25 Hz).
- Countertenor. A countertenor is a type of classical male singing voice whose vocal range is equivalent to that of the female contralto or mezzo-soprano voice types. The countertenor range is generally equivalent to an alto range, extending from approximately G3 to D5 or E5 (196.00 to 587.33 or 659.25 Hz).
- Contralto. The vocal range of this register is fairly rare, which is similar to the mezzo-soprano, and almost identical to that of a countertenor, typically between F3 to F5 (174.61 to 698.46 Hz).
- Mezzo-sopranos. In the lower and upper extremes, some mezzo-sopranos may extend down to F3 (174.61 Hz) and as high as C6 (1046.50 Hz).
- Soprano. The soprano's vocal range is from approximately C4 (261.63 Hz) to A5 (880.00 Hz) in choral music, or to C6 (1046.50 Hz) or higher in operatic music.

The male registers range rougly between 82 and 660 Hz, while the female registers from 175 and 1047 Hz. Given that it is known what gender sings each track, we can concentrate only on the

rangers for that specific gender. Thus, for each track, we remove all frequencies lower and higher from the lower and upper bound for the gender in question, respectively.

In the next Figures, we show the density distribution for each register ranging within the respective gender intervals.





For each track, we obtain the top 3 frequencies and inspect to which register these frequencies have the highest densities.

Voice Register	Accuracy
Bass	0.3283
Baritones	0.1389
Tenor	0.2920
Countertenor	0.6683
Contraltos	0.4605
Mezzosopranos	0.1811
Sopranos	0.6438

Table 1: Baseline results for the literature-based intervals.

The baseline results show that determining the register based on the density of the frequencies performs better than random guess (which is 25 % for male registers and 33% for female registers) in most cases, except for baritones and mezzosopranos. Countertenor and sopranos were the registers with highest accuracy, which are the highest-pitch registers for males and females, respectively.

3.2 CNN

Convolutional Neural Networks are architectures known for dealing with images and being fast to train. Our first approach will be training a CNN with the MFCCs frequencies from the tracks. In sound processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. in this first experiment, we create a total of 20 MFCCs per track.

Our dataset has a unbalanced number of tracks per register. This implies a problem when training a model that should be balanced among the different registers (classes). Thus, given that contraltos is the register with the least number of tracks (8579), we take a *random subset* of 8579 tracks for each register in order to build a balanced model. 10% of the data is kept for validation and 90% for training. Here, a single CNN is trained to classify the MFCCs into one of the seven

registers (male and female undistinguished). The results are shown below:

Voice Register	Precision	Recall	F1-Score
Baritones	0.54	0.49	0.51
Bass	0.60	0.72	0.65
Contraltos	0.54	0.45	0.49
Countertenor	0.58	0.77	0.66
Mezzosopranos	0.43	0.31	0.36
Sopranos	0.55	0.63	0.59
Tenor	0.70	0.59	0.64

Table 2: Validation results for CNN trained on all registers (single-model approach).

These results show that CNN had a higher accuracy than the litterature-based intervals in all registers, except for contraltos and sopranos. Some registers are easier to be recognized than others. For instance, the male registers bass and countertenor had recall values above 70%, which are the lowest and highest registers within the male scale. On the other hand, contraltos and mezzosopranos had low recall values.

3.3 Gender-specific models

We now train two separate models: one for historically male registers, and one for historically female registers.

For each model, we consider all tracks of the registers in consideration. In this sense, we train a unbalanced model to classify the male and female registers separately.

Voice Register	Precision	Recall	F1-Score
Baritones	0.65	0.75	0.69
Bass	0.77	0.37	0.50
Countertenor	0.88	0.87	0.88
Tenor	0.80	0.85	0.82

Table 3: Validation results for CNN trained on male registers.

Results for the model trained with the male registers show that accuracies for baritones, coun-

tertenor and tenor were considerably higher; on the other hand, bass accuracy is much lower.

Voice Register	Precision	Recall	F1-Score
Contraltos	0.54	0.45	0.49
Mezzosopranos	0.67	0.32	0.44
Sopranos	0.79	0.94	0.86

Table 4: Validation results for CNN trained on female registers.

Results for the model trained with the female registers showed a very high accuracy for sopranos; however, contraltos did not have any improvement at all, while mezzosopranos improved of only 0.1%.

3.4 Higher-MFCCs with PCA

In this experiment we increase the number of MFCCs to be generated per track. This increases the number of details present in each track, which can be critical while discriminating between two registers. For instance, there might be some pitches present in bass that are not present in baritones, or present in mezzosopranos but not in contraltos. Our results so far show that bass and mezzosopranos have lower accuracies than the registers in the extremities.

Nevertheless, training a model with a number of 128 MFCCs per track is very computationally expensive, so we build a PCA on top of the MFCCs to reduce data dimensionality.

Before each model is trained, the set of audio samples is tested to find the ideal number of components that will produce at least 99 percent of the variance in the data. The most used number if components was between 28-36.

The pca model, with the selected number of components, is then used to transform the data and reducing the dimensionality. The dataset is then split into training, validation and testing sets with the appropriate ratios that depend on the size of the dataset.

KerasTuner, a scalable hyperparameter optimization framework that solves the pain points of hyperparameter search, is used to quickly estimate the correct hyperparameters to be used in the final model. The metric used may be to minimize 'validation loss' or maximize 'validation accuracy' etc. The most used metric in this experiment was 'validation accuracy'.

A model is designed using the estimated optimal hyperparameters manually and trained on the training dataset. The model performance is analysed to identify and solve issues such as overfitting, underfitting, over-training, under-training etc. This is an iterative process whose major goal is to attain the most efficient model(a model with the least amount hyperparameters but most accurate).

Due to the large number of categories, this approach divides the subregisters into two genders: male and female. There are 10 female subregisters and 9 male subregisters. The number of classes per gender category is still considered large because there are few samples (450-550) per category. This prompts a further division of classes in the gender sub-registers, with the main task being to find two sets of categories that attain desired or acceptable metrics.

After several tests, it is discovered that the female sub-registers are best divided into two sets. Females Set A, consists of mezzosoprano dramatic, countertenor, contraltos and mezzosoprano lyric while Females Set B comprises soprano coloratura, mezzosoprano coloratura and soprano soubrette. The remaining female sub-registers - soprano lyric, soprano dramatic and soprano coloratura - were left out due to high similarity to the other sets, leading to low variance in the dataset, hence lower accuracy and F1 score.

The male sub-registers are best divided into two sets as well. Males Set A comprising tenor lyric, bass dramatic, baritone dramatic and tenor dramatic while Males Set B consisting of baritone bassbaritones, baritone lyric and tenor leggero. Bass cantante and tenor spinto sub-registers were

left out due to low variance.

	Precision	Recall	F1-Score	ROC AUC Score
Mezzosoprano Dramatic	0.66	0.69	0.67	0.88
Countertenor	0.85	0.68	0.75	0.93
Contraltos	0.75	0.84	0.79	0.93
Mezzosoprano Lyric	0.72	0.78	0.75	0.95

Table 5: Precision, Recall and F1-Score for Females Set A.

	Precision	Recall	F1-Score	ROC AUC Score
Mezzosoprano Coloratura	0.76	0.89	0.82	0.94
Soprano Coloratura	0.76	0.70	0.73	0.89
Soprano Soubrette	0.72	0.68	0.70	0.87

Table 6: Precision, Recall and F1-Score for Females Set B.

	Precision	Recall	F1-Score	ROC AUC Score
Tenor Lyric	0.84	0.83	0.84	0.98
Bass Dramatic	0.90	0.94	0.92	0.98
Baritone Dramatic	0.92	0.89	0.90	0.98
Tenor Dramatic	0.86	0.87	0.87	0.96

Table 7: Precision, Recall, F1-Score and ROC AUC Score for Males Set A.

	Precision	Recall	F1-Score	ROC AUC Score
Bass Cantante	0.85	0.76	0.80	0.97
Tenor Leggero	0.79	0.78	0.79	0.99
Baritone Lyric	0.75	0.87	0.80	0.94
Baritone Bassbaritones	0.90	0.89	0.90	0.96

Table 8: Precision, Recall F1-Score and ROC AUC Score for Males Set B.

4 Main Results

5 Conclusion

Tables and Figures

Online Appendix

References