Robust Experimental Design for Speech Analysis Applications

by

Aaquila Arul Arzela Mariajohn

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved April 2020 by the
Graduate Supervisory Committee:

Visar Berisha, Chair
Andreas Spanias
Julie Liss

ARIZONA STATE UNIVERSITY

May 2020

ABSTRACT

In many biological research studies, including speech analysis, clinical research, and prediction studies, the validity of the study is dependent on the effectiveness of the training data set to represent the target population. For example, in speech analysis, if one is performing emotion classification based on speech, the performance of the classifier is mainly dependent on the number and quality of the training data set. For small sample sizes and unbalanced data, classifiers developed in this context may be focusing on the differences in the training data set rather than emotion (e.g., focusing on gender, age, and dialect).

This thesis evaluates several sampling methods and a non-parametric approach to sample sizes required to minimize the effect of these nuisance variables on classification performance. This work specifically focused on speech analysis applications, and hence the work was done with speech features like Mel-Frequency Cepstral Coefficients (MFCC) and Filter Bank Cepstral Coefficients (FBCC). The non-parametric divergence (D_p divergence) measure was used to study the difference between different sampling schemes (Stratified and Multistage sampling) and the changes due to the sentence types in the sampling set for the process.

ACKNOWLEDGMENTS

I thank everyone who helped in the completion of this thesis, mainly my thesis advisor, Dr. Visar Berisha, for being patient with me and supporting me throughout this thesis. I thank my thesis committee members Dr. Andreas Spanias and Dr. Julie Liss, for providing feedback on this thesis. I thank ASU for providing me with the necessary resources for the simulations. I also thank all my friends and family for supporting me in finishing this thesis.

TABLE OF CONTENTS

Page

LIST OF TABLES....vi

LIST OF FIGURES....vii

CHAPTER

1. INTRODUCTION....1

1.1. Sampling Methods....2

1.2. D_p Divergence....3

1.3. Features....5

2.EXPERIMENTAL DESIGN....7

2.1. Dataset Used....8

2.2. Generating Features....9

2.3. Simulation Environments....9

2.4. Forming Strata....10

2.5. Sampling Methods....13

2.6. D_p Divergence Measurement....14

2.7. Analysis....14

2.8. Extrapolation....18

2.9. Choosing the Threshold Crossings....21

3.SIMULATION RESULTS....22

4.CONCLUSION....31

Page

REFERENCES ……..…...33

APPENDIX

APPENDIX A……..…….35

LIST OF TABLES

Table Page

1 Limits Used to Create Different Groups Based on Age 12

2 Simulation Ranges for MFCC Coefficient Analysis 17

3 Comparison of Curve Fitting Models 19

4 Coefficient Bounds 21

5 Threshold Points Obtained..... 22

6 Performance Results for Age Classification 30

LIST OF FIGURES

Figure Page

1 Histogram Plot for the Division of Dataset into 3 Age Groups 11

2 Histogram Plot for the Division of Dataset into 6 Age Groups 11

3 Block Diagram of Simulations Done with MFCC Coefficients. 16

4 Dp Divergence Data Points Obtained 18

5 Curve Fitting Graphs 20

6 Comparison of Thresholds Obtained for Levels of Strata 24

7 Comparison of Threshold Obtained for Different Sentence Types 24

8 Comparison of Thresholds Obtained for Different Features 25

9 Comparison of Thresholds Obtained for Different Sampling Methods 25

10 Comparison of Thresholds Obtained for Different Age Groups 26

11 Average of Thresholds for Each Level of Strata 26

12 Threshold Points Averaged for Features 27

13 Threshold Points Averaged for Sampling Sentence Type 27

14 Threshold Points Averaged for Sampling Methods 28

15 Threshold Points Averaged for Different Age Groups 28

16 Divergence Graph Obtained for the New Training Data with New Features 29

1. INTRODUCTION

The accuracy of a research study heavily depends on the effectiveness of the training data set to represent the target population. Broadly speaking, studies show that a more significant accuracy can be obtained with a larger sample size [1]. For clinical research, it is not possible to collect a large target population for a study. The usual approach to solve this is to include a part of the target population called the sample population in the training data [2]. The availability of training data for the target population can be challenging. In addition to this challenge of availability, the training data may be unbalanced, resulting in biased results. In the case of a binary emotion classifier, if we have a training data which is unbalanced in gender, the classifier may be giving more weight to the difference in the training data due to gender difference rather than the emotion. Similar is the case with any machine learning or deep learning-based studies as these analyze the results based on the training set. This thesis deals with this problem of unbalanced and insufficient training data and focuses explicitly on sample size estimation or power estimation for unbalanced data for speech analysis applications.

Sample size estimation is the process of predicting a sample size that gives adequate credibility to reject the null hypothesis put forward in the study [3]. To do sample size estimation, one usually needs to know the distribution of the data being analyzed; based on

this distribution, power estimation can be done in a parameterized way [4]. However, the distribution of the data is not always known.

In this thesis, we evaluate the effects of various experimental design methods to avoid these challenges using the D_p divergence measure. This process gives a minimum number of samples required in a training set to make it balanced by detecting the possible differences in the training set. The following sections detail some of the background knowledge used in this thesis.

1.1.Sampling methods

The process of selecting the sample population can be done in multiple ways depending on the sampling methods. There are probability sampling methods and non-probability sampling methods. Probability sampling methods ensure equal representation for all members of the target population in the sample population, whereas non-sampling methods do not [2] [3]. In probability sampling methods, there are four different methods: simple random, stratified random, systematic, and clustered random.

In simple random, the required number of samples are drawn randomly from the entire training set. In stratified random sampling, the entire available data set is classified into subgroups (called strata) based on demographic factors that can influence the study. An equal number of samples are randomly drawn from each of these strata, giving equal representation for each of the strata in the sampling set over the simple random case. In

systematic random sampling, the samples are selected using a fixed rule (e.g., one in every five). In cluster sampling (also called multistage sampling), a random number of strata are selected, and from each of these, a random number of samples are chosen. In this thesis, we have analyzed and implemented stratified and multistage probability sampling methods.

1.2.D_p divergence

D_p divergence measure belongs to the class of f-divergences and provides a non-parametric way of estimating the classification bounds. D_p divergence has the advantage that it can be used when the conditional distributions of data for the problem are not known [5]. For the problem of finding the D_p divergence measure between two different sets of data with probability distribution functions of f_0 and f_1 and for a parameter $p \in (0,1)$, $q = 1 - p$, from [6] it is given by:

$$D_p(f_0, f_1) = \frac{1}{4pq}\left[\int \frac{(pf_0(\mathbf{x}) - qf_1(\mathbf{x}))^2}{pf_0(\mathbf{x}) + qf_1(\mathbf{x})}\, d\mathbf{x} - (p-q)^2\right]$$

To find the $D_p(f_0, f_1)$ value using the above equation, f_0 and f_1 functions should be known. But using the method given in [6], the $D_p(f_0, f_1)$ value can be found by the extension of the Friedman-Rafsky (FR) multi-variate two-sample test statistic [7]. A two-sample test is used to verify the null hypothesis that both p and q are not the same and can give a probability that these are not the same. This computation can be done using the empirical

data of the two distributions with N_p and N_q samples from p and q respectively. Let $\mathbf{X_p} \in \mathbb{R}^{N_p \times K}$ denote a sample from the distribution of p and $\mathbf{X_q} \in \mathbb{R}^{N_q \times K}$ denote a sample from the distribution of q and each of these samples are vectors of length K. And by [6], the above equation reduces to:

$$D_p(f_0, f_1) = 1 - \mathrm{C}(\mathbf{X_p}, \mathbf{X_q}) \frac{N_p + N_q}{2N_pN_q}$$

Here, $\mathrm{C}(\mathbf{X_p}, \mathbf{X_q})$ is the FR test statistic. This value is found by [7] the following steps:

1. Construct a minimum spanning tree by combining both the data points of f_0 and f_1. A minimum spanning tree is a tree with a unique path between each node.
2. Remove all the edges connecting different samples (edges connecting $\mathbf{X_p}$ to $\mathbf{X_q}$).
3. $\mathrm{C}(\mathbf{X_p}, \mathbf{X_q})$ is the number of the resulting disjoint subtrees.

This D_p divergence value is always between zero and one. Also, it becomes zero when both the distribution functions are the same and gives a value close to one when they are different. Hence, using this divergence measure a null hypothesis that $\mathbf{X_p}$ and $\mathbf{X_q}$ are not the same can be rejected for a D_p divergence value of zero. This divergence measure, along with its properties, was used to implement this thesis.

1.3.Features

MFCC and FBCC are the two features derived here. MFCC is prevalently used for speech feature extraction in various systems [8]. MFCC is Mel Frequency Cepstral Coefficients. These represent the frequency components in different frequency bands and their energies. Mel frequency is a scale of frequency that represents how humans perceive frequency. Humans can discern lower frequencies better than the higher frequency ranges. So, this scale converts the linear frequency range in such a way that the higher frequency range is compressed, and the lower frequency range is expanded. This conversion is done using the log function given by:

$$M(f) = 1125 \ln (1 + \frac{f}{700})$$

Here, f is the linear frequency, and $M(f)$ is the frequency in the Mel scale. The converse of this can be found by:

$$f(m) = 700(10^{m/2595} - 1)$$

The following are the steps involved in deriving MFCC. Pre-emphasis is done to the signal before any analysis to boost the energy levels of higher frequencies. This boosting makes the information of higher frequencies more available as the lower frequencies. The signal is then divided into shorter frames assuming that the signal remains stationary in this interval by windowing. Windowing the frames helps in reducing the noise that shows up in the frequency domain due to the time domain chopping. The Discrete Fourier Transform of each frame $s_i(n)$ and a window function of $h(n)$ is done by:

$$S_i(k) = \sum_{n=1}^{N} s_i(n)h(n)e^{\frac{-j2\pi kn}{N}}, 1 \leq k \leq K$$

Here, N is the window length, and K is the DFT length. The power spectral estimate of each frame is found from its periodogram by:

$$P_i(k) = \frac{1}{N}|S_i(k)|^2$$

Next, this periodogram signal is passed through triangular bandpass filter banks in the Mel scale given by:

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \leq k \leq f(m) \\ 1 & k = f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)} & f(m) < k \leq f(m+1) \\ 0 & k > f(m+1) \end{cases}$$

Here, m is the frequency in the Mel scale, $f(m)$ is the frequency in linear scale, and k is the linear frequency variable of the filter bank. The output from these filter banks forms the FBCC features. The next step is to find the log of each of the filter band outputs. Typically, 26 filter bands are used. Finally, Discrete Cosine Transform (DCT) of these 26 signals is taken, of which only the lower 13 filter band values are retained, and these form the MFCC coefficients.

2. EXPERIMENTAL DESIGN

The primary objective here was to develop a process that could help in deriving the least number of samples needed in a training set that removes the effect of unbalanced data. We focused on speech analysis applications, hence the training data that we worked with were the features (section 2.2) derived from the speech samples in the training set (section 2.1). We chose to work with the demographic traits of gender, age, and dialect as those that can cause differences. Following the process of forming strata (section 2.4), we were able to create different strata from the training set, samples of each stratum being more similar to each other. This process helps in forming a balanced dataset based on the traits chosen. We formed two different classes, each with a specific number of samples (as given in Table 2) from the balanced data set using either of the sampling methods in section 2.5. D_p divergence value of these two classes was calculated using the steps in section 1.2 and 2.6. This value gives us a quantitative measure of how close these two classes are in their distribution. The same process was repeated for a higher number of samples in each class. We repeated this process for different types of analysis (section 2.7). Once we obtained the raw data points, we extrapolated the data as in section 2.8. We did this extrapolation to derive the threshold crossing point (section 2.9) of the D_p divergence graph versus the number of speakers for each of the analyses done. We compared different sampling methods, various features, and levels of strata by analyzing the convergence of the D_p

divergence versus the number of speaker's graph. We found the mean of all the threshold points obtained for each of MFCC and FBCC to conclude a performance analysis between them. We repeated the same averaging process for comparison between different sampling methods and different levels of strata as well. The comparison results are given in chapter 3. The following sections describe in detail the procedure and specifics followed for the design and implementation of the process explained above.

2.1.Dataset Used

The data set used for all the experiments was TIMIT Acoustic – Phonetic Continuous Speech Corpus. This data set has a complete recording of around 630 speakers at a sampling frequency of 16 kHz with each speaker reading ten sentences, of which two are universal among all speakers, and the rest eight are different. This data set includes recordings of speakers from 8 different dialects, different age groups (between 20 and 76 years old), and gender. We considered two different types of analysis, namely the same (using only the two universal sentences in the data set per person) and different (using all the ten sentences in the data set per person). The data set is divided into two separate groups one for training and one for testing. We worked with the training data set, which included 462 different speakers.

2.2.Generating features

We extracted two different sets of features, namely MFCC and FBCC, from the above data set using a window size of 25ms and an overlap length of 10 ms. This windowing results in a frame length of 400 samples, and so a Fast Fourier Transform (FFT) size of 400 or above gives the accurate FFT values. We chose an FFT size of 400(default) for MFCC and 512(default) for Mel-filter bank coefficients. For both the features, we extracted 14 different coefficients, where the first coefficient is the frame energy. Along with these, we also extracted the delta and double delta coefficients, resulting in 42 coefficients per frame. Delta coefficients are the change in the original 14 coefficients from one frame to another, and double delta is the change in the delta coefficients from one frame to another.

Each of the coefficients obtained had a different mean and range, resulting in invalid results when computing D_p divergence values due to the unequal weightage the coefficients with larger amplitudes had over the ones with smaller values. Hence, we performed the mean normalization of these coefficients by converting them to values between -1 and 1. We obtained this by using the mean and range of each of the coefficients across all the training set.

2.3.Simulation environments

We used MATLAB to generate the MFCC coefficients and python to generate the Mel-filter bank coefficients using the library python_speech_features on the anaconda

environment. Due to the large size of the feature matrices generated during the sampling experiments, we simulated these on the agave cluster to incorporate the large memory requirements.

2.4.Forming Strata

As described before, it is possible to divide the data set into different strata based on gender, dialect region, and age group. There are two different useful gender groups - male and female. There are eight different dialect groups possible for the TIMIT data set.

Moreover, we divided the data set into three and six different groups, based on age, forming two possible versions (3A and 6A, respectively, for naming conventions). We created the age groups by making sure the number of speakers in each group was nearly equal. This process ensures an equal probability of being chosen for all the speakers across different age groups. The histogram representation for each age group division and the corresponding limits used are given in Figure 1, Figure 2, and Table 1.

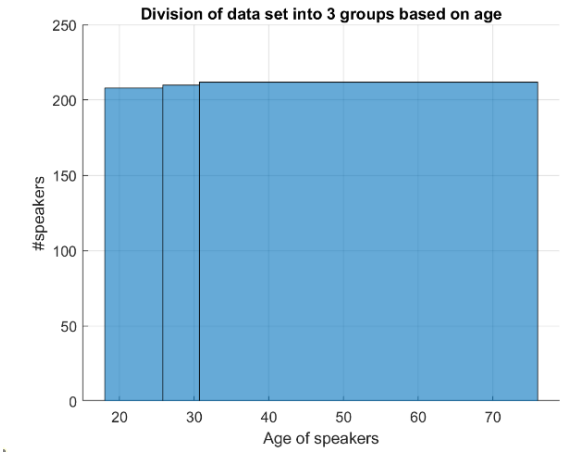


Figure 1: Histogram plot for the division of dataset into 3 age groups

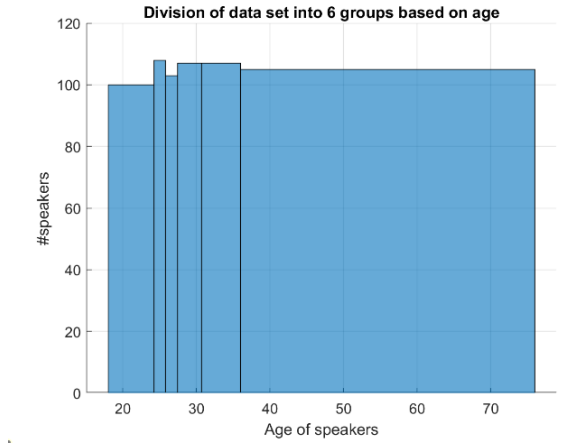


Figure 2: Histogram plot for the division of dataset into 6 age groups

Age limits for each group	
3 age groups (years)	**6 age groups** (years)
18 – 25.76 25.76 – 30.67 30.67 – 76	18 – 24.2 24.2 – 25.76 25.76 – 27.45 27.45 – 30.67 30.67 - 36 36 - 76

Table 1: Limits used to create different groups based on age

For the first level of strata (named T to represent training samples), there is only a single stratum with all the speakers in the dataset. For the second level of strata (named T + G for Training + Gender), we divided the stratum in the first level into two different strata based on gender. For the third level of strata, we simulated two different variations choosing either age-based (T + G + A) or dialect based (T + G + D). Adding the fourth level of strata, we simulated only T + G + D + A and not T + G + A + D as the second case results in empty strata due to the size of the training set.

2.5.Sampling methods

The D_p divergence measurements between two different classes (which is the subset of the data set chosen for analysis) frame the primary analysis of this thesis. The method of selecting the constituent strata in each class gives us different sampling schemes. We simulated two different sampling schemes: Stratified Sampling and Multistage Sampling. For stratified sampling (SS) scheme, there is equal representation for all the strata in the class, hence one speaker, randomly chosen from each stratum, constitutes the class for the starting point of D_p divergence measurements between two classes. The D_p divergence measurements found the next point by choosing two speakers randomly from each stratum. The repetition of this process for a higher number of speakers from each stratum gave more analysis points.

For the multistage sampling (MS) scheme, we chose only a fixed number of randomly chosen strata. For the selected strata, we repeat the same process as above. For this sampling scheme, two different types of sampling are possible, namely matched and unmatched multistage sampling schemes. In matched sampling, we chose the same strata for both the classes formed for the D_p divergence analysis. Whereas, in unmatched sampling, we chose different combinations of strata for each class.

2.6.D_p divergence measurement

We chose D_p divergence as a quantitative measurement to analyze the similarity between two classes. Each class here was formed by concatenating the feature vector of each sample in the class one after the other. We did the D_p divergence measurements for a range of speakers, as given in Table 2. We repeated each of these D_p divergence measurements for 400 Monte Carlo iterations to reduce errors due to the randomness involved in the sampling process. Using the Markov Chain Monte Carlo Sampler, the required measures can be found from the probability distribution of the 400 D_p divergence values obtained. Here we used the mean of these 400 samples to represent the final value.

2.7.Analysis

We simulated all the above D_p divergence measurements for two different types of data set based on the sentences spoken: 1. with only the two common sentences (same sentences) and 2. with all ten sentences included in the data set per speaker (different sentences). For each of these types, we measured the D_p divergence values in four different levels of strata. Moreover, for each of these levels except the first (T) and the second one (T + G), we implemented both the sampling schemes. T has only one stratum; hence we implemented it by picking the number of speakers needed directly from the training data set. For T + G,

the number of strata is only two, hence MS scheme implementation will be the same as level 1 T implementation.

For the T + G + D + A level, we simulated only the MS scheme. Furthermore, for this scheme, we implemented two different versions: randomly choosing eight strata (8g) and 16 strata (16g). For all other MS schemes, we randomly chose half of the number of strata in the specific level for measurements. Given in Figure 3 is a flow diagram showing all the simulations done on the training data set and given in Table 2 is the range of the speakers, starting point, number of speakers, ending point for each of these simulations. We chose the range of speakers in each case, such that there are enough points needed to extrapolate the graph further. This extrapolation helps in finding the point of zero crossings in the D_p divergence versus the number of speaker's graph.

We simulated all these analyses for two different feature sets – MFCC and FBCC. So, the coefficients chosen for each speaker to form the class are either of these feature coefficients.

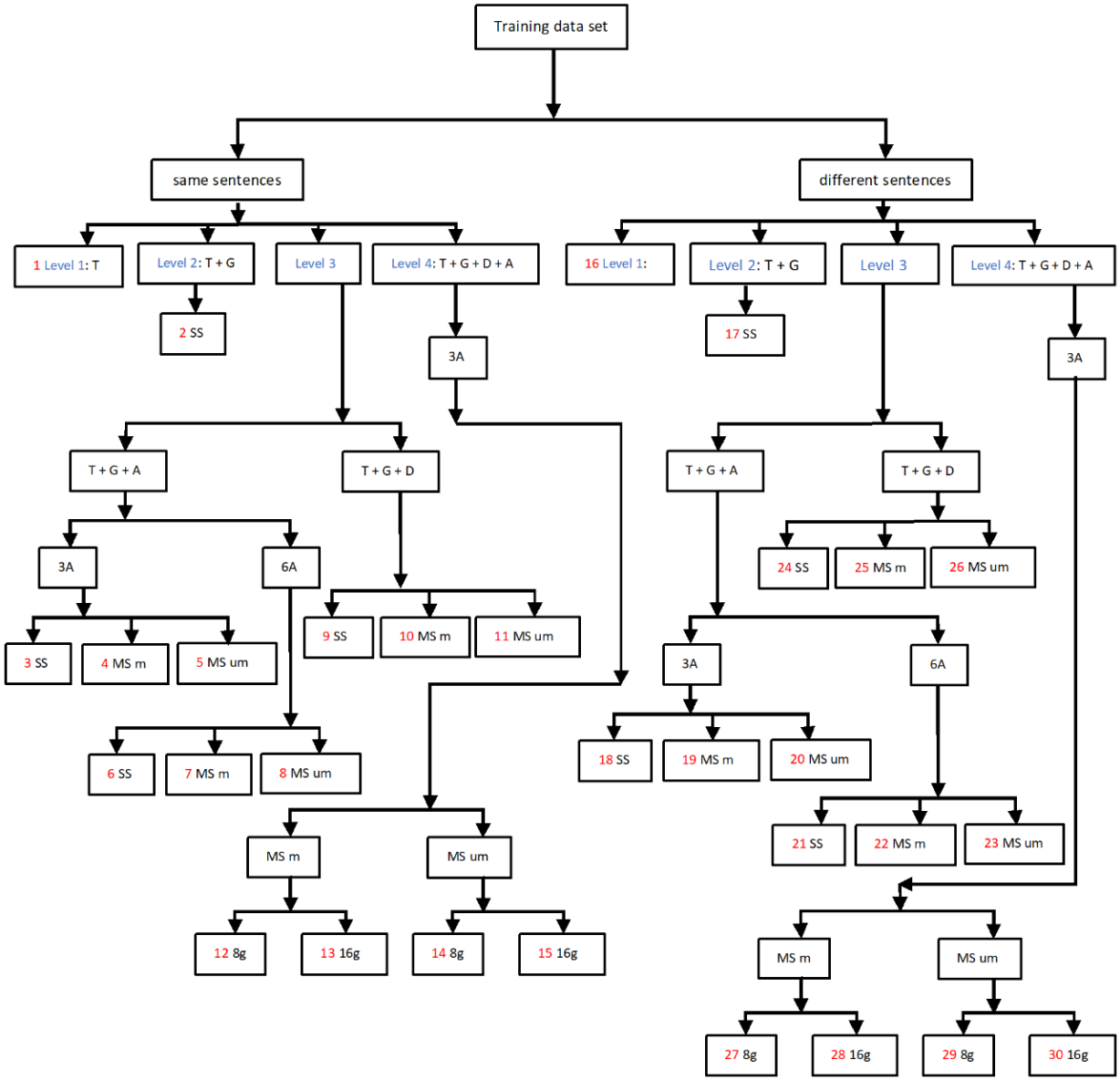


Figure 3: Block diagram of simulations done with MFCC coefficients. The number in red represents the corresponding cases for later reference. Numbers from 31 to 60 represents the same analysis done with FBCC coefficients.

No:	Case	#strata	#strata chosen for MS	Starting #speakers	Step size of simulation	Ending #speakers
1	MF-S1T1g	1	-	1	1	90
2	MF-S2TG2gSS	2	-	2	2	90
3	MF-S3TGA3gSS	6	-	6	6	192
4	MF-S3TGA3gMSm	6	3	3	3	120
5	MF-S3TGA3gMSum	6	3	3	3	120
6	MF-S3TGA6gSS	12	-	12	12	192
7	MF-S3TGA6gMSm	12	6	6	6	120
8	MF-S3TGA6gMSum	12	6	6	6	120
9	MF-S3TGD8gSS	16	-	16	16	224
10	MF-S3TGD8gMSm	16	8	8	8	144
11	MF-S3TGD8gMSum	16	8	8	8	144
12	MF-S4TGDA8gMSm	48	8	8	8	144
13	MF-S4TGDA16gMSm	48	16	16	16	144
14	MF-S4TGDA8gMSum	48	8	8	8	144
15	MF-S4TGDA16gMSum	48	16	16	16	144
16	MF-D1T1g	1	-	1	1	90
17	MF-D2TG2gSS	2	-	2	2	90
18	MF-D3TGA3gSS	6	-	6	6	192
19	MF-D3TGA3gMSm	6	3	3	3	120
20	MF-D3TGA3gMSum	6	3	3	3	120
21	MF-D3TGA6gSS	12	-	12	12	192
22	MF-D3TGA6gMSm	12	6	6	6	120
23	MF-D3TGA6gMSum	12	6	6	6	120
24	MF-D3TGD8gSS	16	-	16	16	224
25	MF-D3TGD8gMSm	16	8	8	8	144
26	MF-D3TGD8gMSum	16	8	8	8	144
27	MF-D4TGDA8gMSm	48	8	8	8	144
28	MF-D4TGDA16gMSm	48	16	16	16	144
29	MF-D4TGDA8gMSum	48	8	8	8	144
30	MF-D4TGDA16gMSum	48	16	16	16	144

Table 2: Simulation ranges for MFCC coefficient analysis. The same holds for FBCC coefficient analysis from 31 to 60 with ‘FB ‘in the case naming instead of ‘MF. ‘

2.8.Extrapolation

The D_p divergence values obtained during the simulations for an increasing number of speakers in a class gives the data, as shown in Figure *4*, which was observed to be decreasing exponentially. To generalize the process of finding the threshold crossing, it was necessary to find a curve-fitting model that could accurately represent the decreasing exponential curve for all the simulation cases.

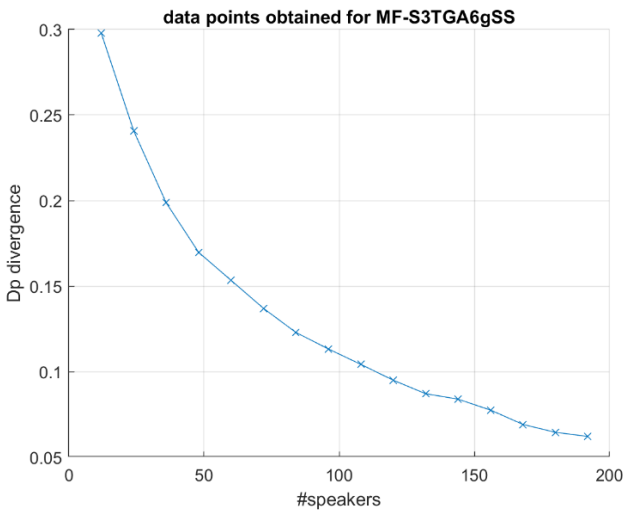


Figure 4: D_p divergence data points obtained

Given below in Table 3 is a comparison of different curve fitting models that we experimented with and the resulting graphs in Figure 5.

Model name	Pow1	Pow2	Exp1	Exp2
Equation	$a * x^b$	$a * x^b + c$	$a * e^{b*x}$	$a * e^{b*x} + c * e^{d*x}$
SSE	0.0023	8.0424e-05	0.0027	2.5741e-05
R-square	0.9669	0.9988	0.961	0.9996
Adjusted R-square	0.9645	0.9987	0.9583	0.9995
RMSE	0.0128	0.0025	0.0138	0.0015

Table 3: Comparison of curve fitting models

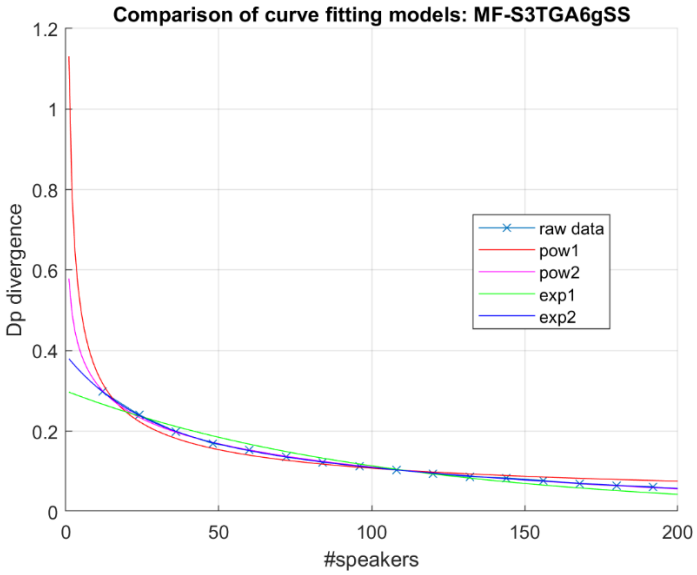


Figure 5: Curve fitting graphs

In Table 3, SSE stands for the sum of squared error, which predicts the deviation of the predicted values from the actual values. A lower value of SSE is preferred for higher accuracy in the fitting. R-square is the ratio of response variation to that of the total variation, and a higher value of R-square is preferred. Adjusted R-square is another version of R-square, and a higher value is preferred. RMSE is the abbreviated form of root mean square error, which is the standard deviation of the residuals, and a lower value of RMSE is preferred for higher accuracy.

Exp2 was chosen for all simulations after comparing the curve fitting models using the goodness of fit measures discussed above.

Hence, we implemented extrapolation using the curve fitting model ‘exp2’ given by:

$$val(x) = a * e^{b*x} + c * e^{d*x}$$

We bounded the coefficients for exp2 here, as given in Table 4, to force the curve to zero at infinity.

Coefficient	**Lower bound**	**Upper bound**
a	-∞	1
b	-∞	0
c	-∞	∞
d	-∞	0

Table 4: Coefficient bounds

2.9.Choosing the threshold crossings

To reject the null hypothesis that the distribution of the chosen classes is not the same, the D_p divergence value obtained should be zero, as given in section 1.2. However, the curve fitting model which was chosen in section 2.8 for the D_p divergence values for a different number of speakers approach zero at an infinite number of speakers. Hence, we chose a threshold level very close to zero to reject the null hypothesis. Moreover, the point at which the curve fitting model crosses this threshold level was chosen as the least point at which the null hypothesis can be rejected. At this point, it can be predicted that the distributions of both the classes are the same.

3. SIMULATION RESULTS

We derived the following threshold points from the analysis specified in Table 2 for a chosen threshold level of 0.025.

Sl. No:	Case	Threshold point
1	MF-S1T1g	321
2	MF-S2TG2gSS	324
3	MF-S3TGA3gSS	281
4	MF-S3TGA3gMSm	286
5	MF-S3TGA3gMSum	295
6	MF-S3TGA6gSS	330
7	MF-S3TGA6gMSm	287
8	MF-S3TGA6gMSum	307
9	MF-S3TGD8gSS	266
10	MF-S3TGD8gMSm	286
11	MF-S3TGD8gMSum	273
12	MF-S4TGDA8gMSm	49
13	MF-S4TGDA16gMSm	45
14	MF-S4TGDA8gMSum	92
15	MF-S4TGDA16gMSum	91
16	MF-D1T1g	295
17	MF-D2TG2gSS	307
18	MF-D3TGA3gSS	279
19	MF-D3TGA3gMSm	264
20	MF-D3TGA3gMSum	265
21	MF-D3TGA6gSS	270
22	MF-D3TGA6gMSm	275
23	MF-D3TGA6gMSum	274
24	MF-D3TGD8gSS	297
25	MF-D3TGD8gMSm	249
26	MF-D3TGD8gMSum	265
27	MF-D4TGDA8gMSm	38
28	MF-D4TGDA16gMSm	40

Sl. No:	Case	Threshold point
29	MF-D4TGDA8gMSum	91
30	MF-D4TGDA16gMSum	71
31	FB-S1T1g	240
32	FB-S2TG2gSS	245
33	FB-S3TGA3gSS	248
34	FB-S3TGA3gMSm	220
35	FB-S3TGA3gMSum	238
36	FB-S3TGA6gSS	258
37	FB-S3TGA6gMSm	220
38	FB-S3TGA6gMSum	234
39	FB-S3TGD8gSS	234
40	FB-S3TGD8gMSm	206
41	FB-S3TGD8gMSum	246
42	FB-S4TGDA8gMSm	35
43	FB-S4TGDA16gMSm	45
44	FB-S4TGDA8gMSum	89
45	FB-S4TGDA16gMSum	83
46	FB-D1T1g	206
47	FB-D2TG2gSS	205
48	FB-D3TGA3gSS	208
49	FB-D3TGA3gMSm	191
50	FB-D3TGA3gMSum	204
51	FB-D3TGA6gSS	225
52	FB-D3TGA6gMSm	185
53	FB-D3TGA6gMSum	207
54	FB-D3TGD8gSS	193
55	FB-D3TGD8gMSm	185
56	FB-D3TGD8gMSum	194
57	FB-D4TGDA8gMSm	29
58	FB-D4TGDA16gMSm	32
59	FB-D4TGDA8gMSum	74
60	FB-D4TGDA16gMSum	52

Table 5: Threshold points obtained

The D_p divergence graphs obtained for each of the analysis cases are given in the appendix A. The comparison graphs for each type of analysis is given below:

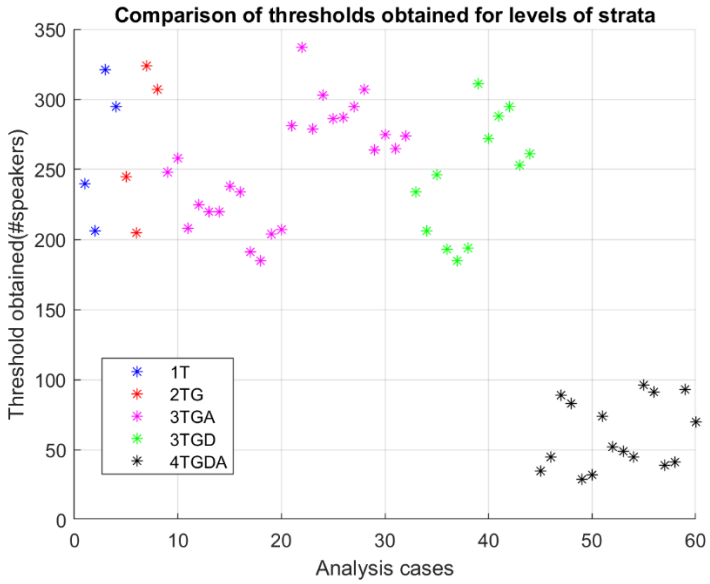


Figure 6: Comparison of thresholds obtained for levels of strata

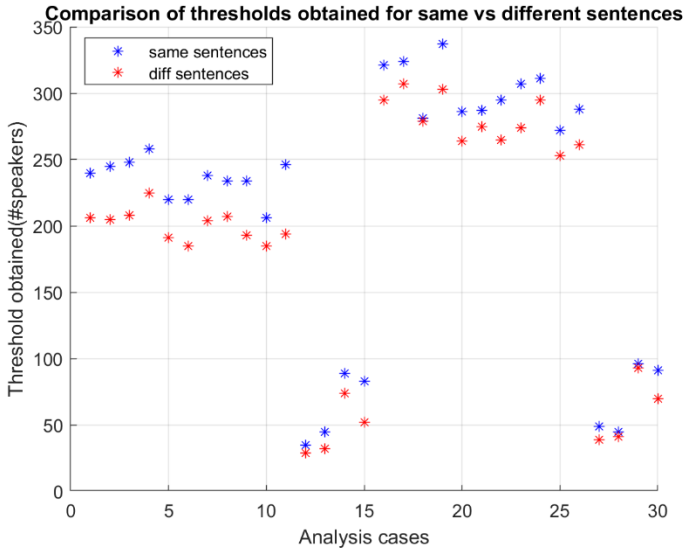


Figure 7: Comparison of threshold obtained for different sentence types

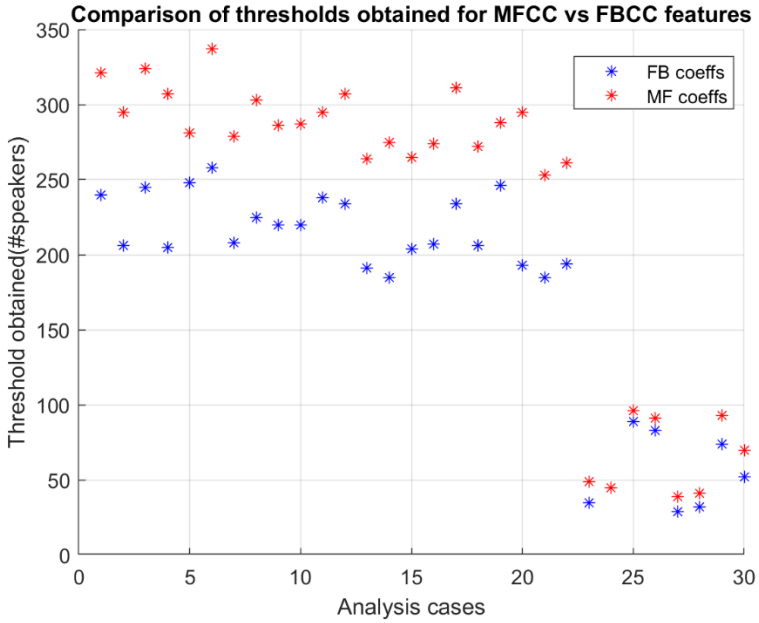


Figure 8: Comparison of thresholds obtained for different features

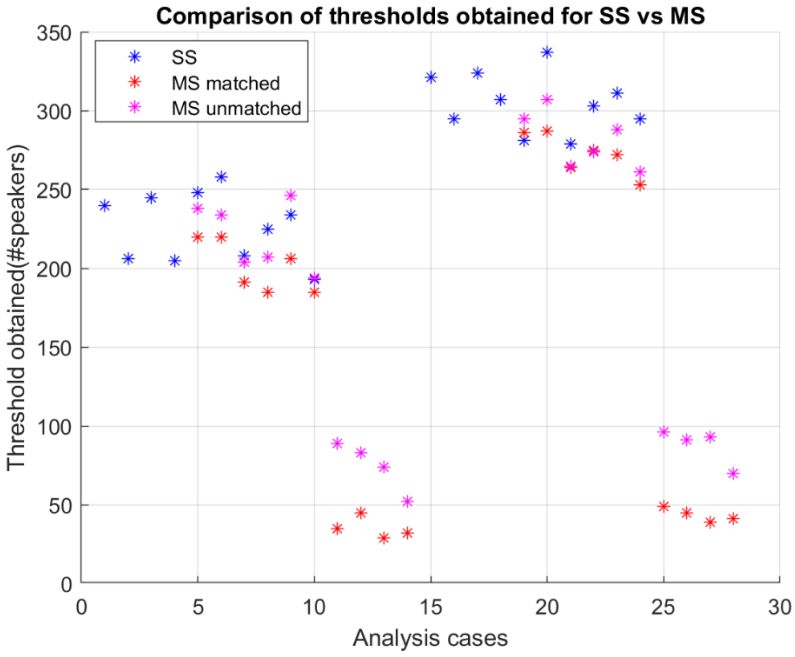


Figure 9: Comparison of thresholds obtained for different sampling methods

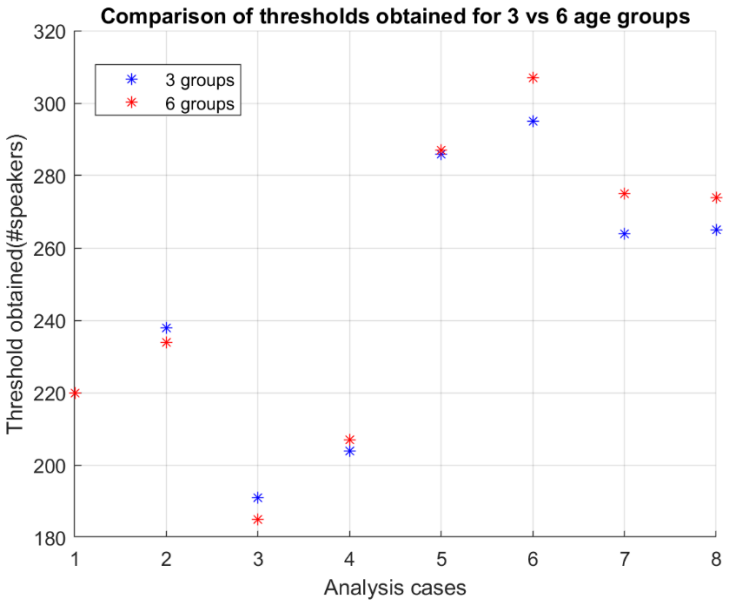


Figure 10: Comparison of thresholds obtained for different age groups

The results obtained for the comparison by averaging are given below:

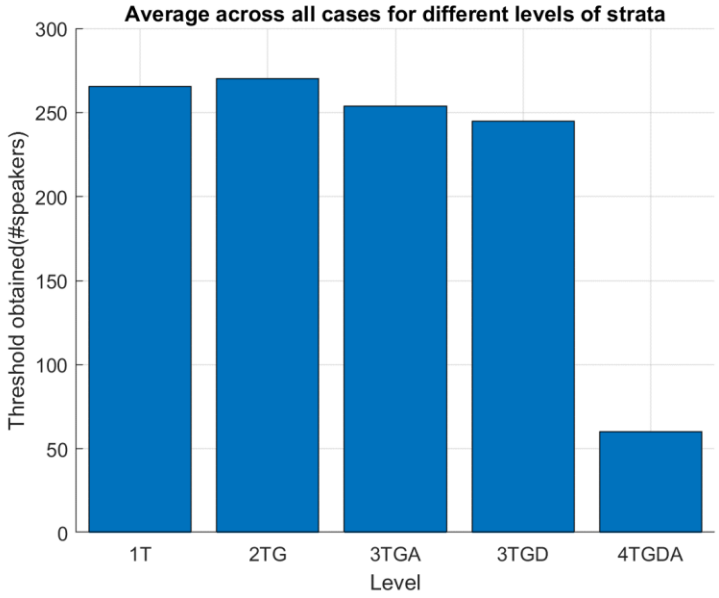


Figure 11: Average of thresholds for each level of strata

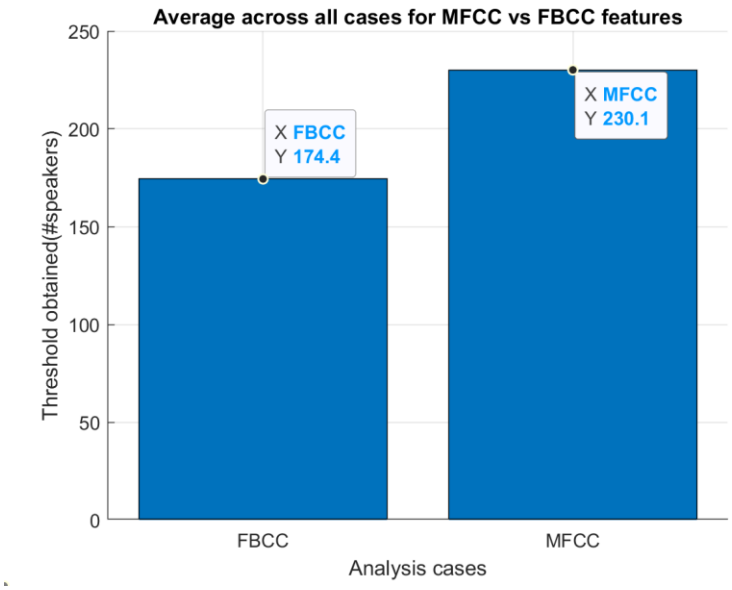


Figure 12: Threshold points averaged for features

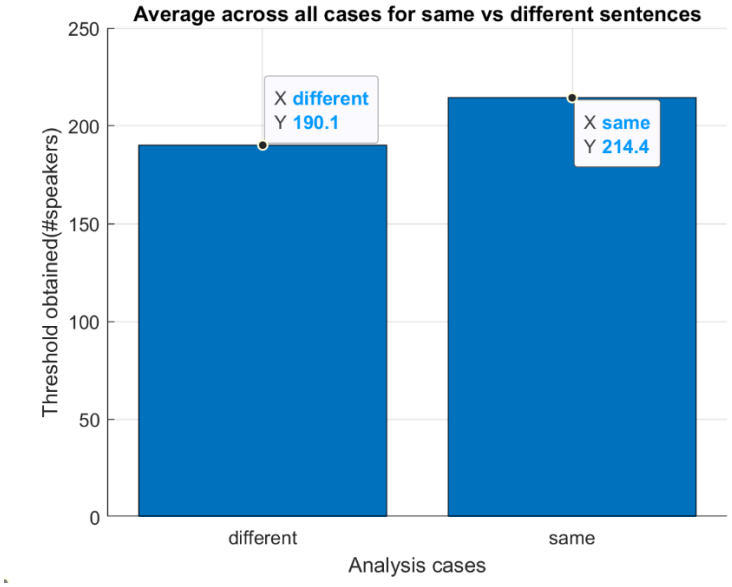


Figure 13: Threshold points averaged for sampling sentence type

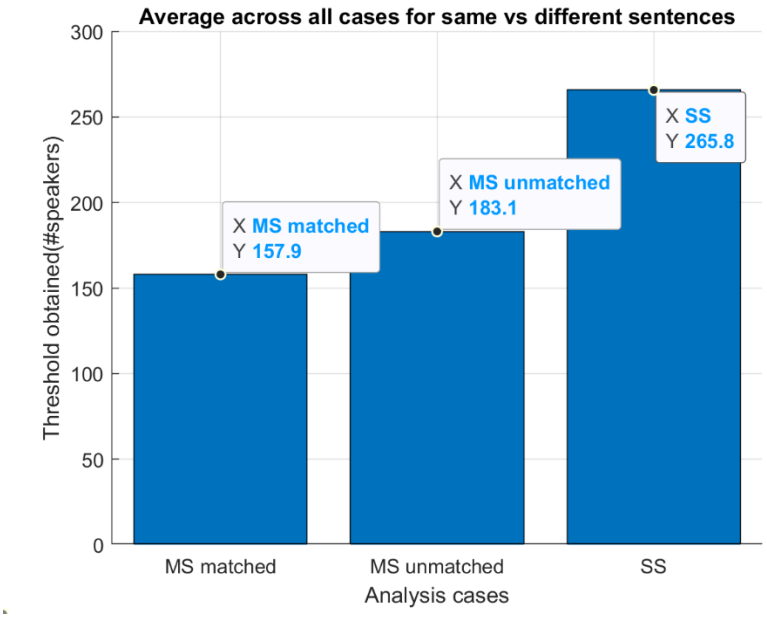


Figure 14: Threshold points averaged for sampling methods

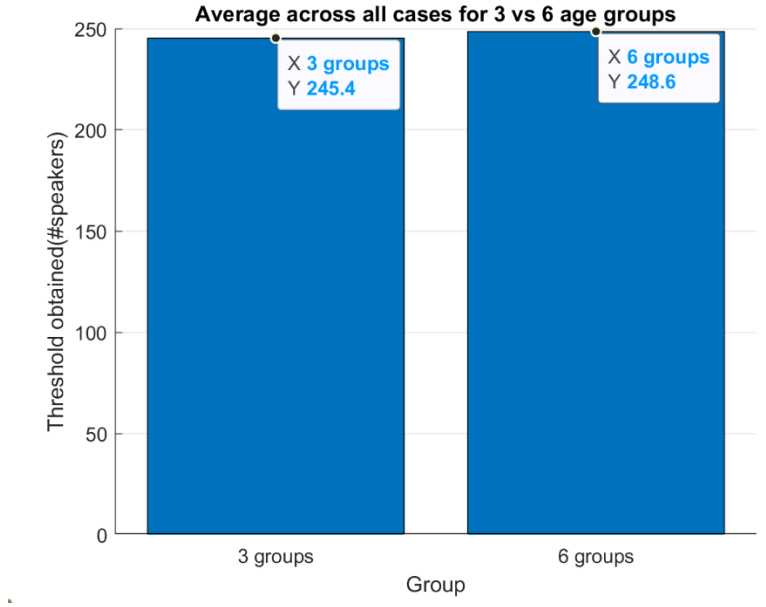


Figure 15: Threshold points averaged for different age groups

To show the improvement in accuracy with this design step, an age classifier was implemented using SVM with a gaussian kernel. The training data was a subset of the TIMIT training set derived randomly to form unbalanced data. Two different age groups were formed by dividing the data into adults (samples younger than 45 years) and elderly (samples older than or equal to 45 years). The training data consisted of 118 speakers with 102 adults (59 females and 43 males) and 16 elderly (5 females and 11 males). Thus, there was an imbalance in the gender in both groups. Besides, there were also eight dialect groups among the speakers. A new set of features (39 elements) appropriate for age classification was formed, and the divergence graph was obtained as below.

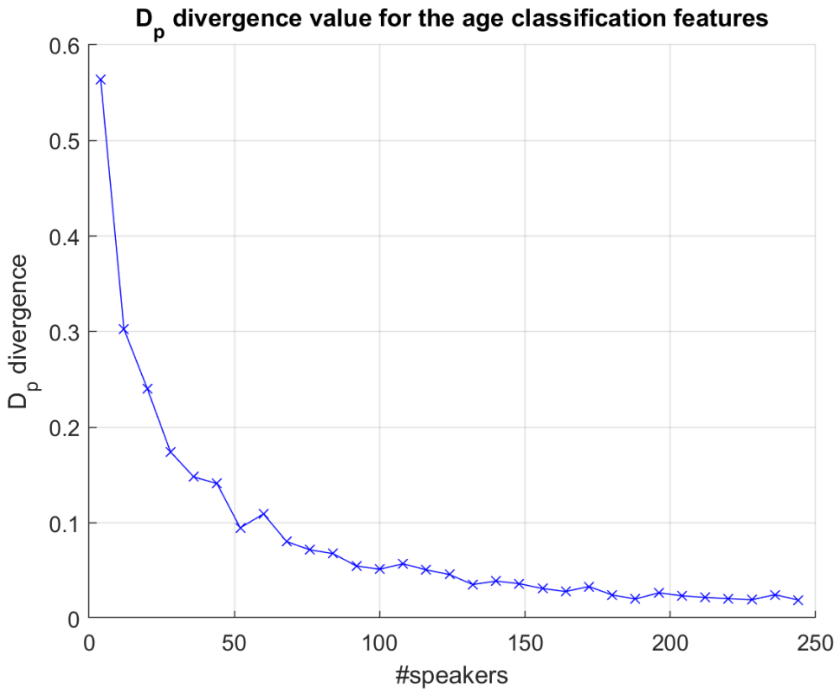


Figure 16: Divergence graph obtained for the new training data with new features

Testing the classifier with the unbalanced data gave lower accuracy rates. New sub-samples were formed from the unbalanced data using the stratified sampling schemes. This process of stratified sampling improved the performance of the classifier, as can be seen from the below table.

	Unbalanced data	Stratified (100 speakers)	Stratified (180 speakers)
Training	95.06%	100%	100%
Testing	77.81%	82.6875%	92.777%

Table 6: Performance results for age classification

From Figure 16, it can be observed that just by changing the sampling data to balanced, there is an improvement in accuracy. This improvement in accuracy is further increased by increasing the number of speakers in the group to around the threshold point (which is around 180 for a threshold level in divergence value of 0.025).

4. CONCLUSION

From Figure 8 and Figure 12, it can be observed that the analysis with FBCC features resulted in faster convergence or a smaller threshold crossing point than MFCC features. This result also indicates that FBCC is better in representing the speech samples than MFCC for the design process implemented here. The high correlation present in FBCC coefficients can be the reason for this faster convergence.

Figure 10 and Figure 15 gives a comparison between two different groupings of the training set based on age. The three-age group case converges better than the six age group.

Figure 9 and Figure 14 shows the comparison between the sampling methods SS, MS matched, and unmatched. MS matched converges faster than MS unmatched, which converges faster than SS. This can be because SS has a higher number of speakers than the corresponding case of MS in each class. The higher number of speakers contributes to more significant variability in the class, hence higher threshold points.

Figure 7 and Figure 13 shows that when there are different types or more number of sentences in the sampling set per speaker, the convergence is faster than having the same sentence types. Even with the variability resulting from the same speaker speaking different sentence types, the different sentence case has lower threshold crossing point.

Figure 6 and Figure 11 shows that as the levels of strata increases, the convergence becomes faster. For level 3, 3TGD performs slightly better than 3TGA, suggesting that dialect may be a better trait in forming balanced data than age. Also, there is a massive

drop in the threshold point for 4TGDA compared to other levels of strata, indicating that higher levels of strata are preferred to have a smaller threshold crossing point. A smaller threshold crossing point indicates the requirement for a smaller number of samples in the data set. This result shows that the experimental design implemented here helps in bringing down the number of training data required, usually (level 1 - 260), to a lower value of 60 (level 4).

Observing the improvement in the performance of the classifier as described in chapter 4, it shows that only by changing the sampling data to a balanced form, improvement in accuracy is possible.

Through this thesis, we were able to design a process to convert unbalanced training data into a balanced one. This process helps in deriving a minimum number of speakers for the training when there is a limitation in the training data. If one is working with a new set of features, following the design steps described in the project, one can obtain the divergence graphs. From this divergence graph, the appropriate feature can be selected that converges faster if the time complexity of training is a concern. This process can be used for other applications as well by changing the features required appropriately for the study.

REFERENCES

[1] T. v. d. Ploeg, P. C. Austin, and E. W. Steyerberg, "Modern modeling techniques are data-hungry: a simulation study for predicting dichotomous endpoints," BMC Medical Research Methodology, vol. 14, no. 137, 2014.

[2] M. Elfil and A. Negida, "Sampling methods in Clinical Research; an Educational Review," Emergency, vol. 5(1), no. e52, 2017.

[3] A. Shorten and C. Moorley, "Selecting the sample," Evid Based Nurs, vol. 17, no. 2, pp. 32-33, 2014.

[4] U. D. o. E. a. Biostatistics, "Power and sample size calculations" [Online]. Available: http://www.stat.ubc.ca/~rollin/ stats/ssize/index.html.

[5] V. Berisha, A. Wisler, A. O. Hero and A. Spanias, "Empirically Estimable Classification Bounds Based on a Nonparametric Divergence Measure," IEEE Transactions on Signal Processing, vol. 64, no. 3.

[6] V. Berisha and A. Hero, "Empirical non-parametric estimation of the fisher information," IEEE signal processing, vol. 22, no. 7, pp. 988 - 992, 2015.

[7] J. H. Friedman and L. C. Rafsky, "Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests," Ann Statist, pp. 697 - 717, 1979.

[8] M. A. Hossan, S. Memon and M. A. Gregory, "A novel approach for MFCC feature extraction," in International Conference on Signal Processing and Communication Systems, 2010.

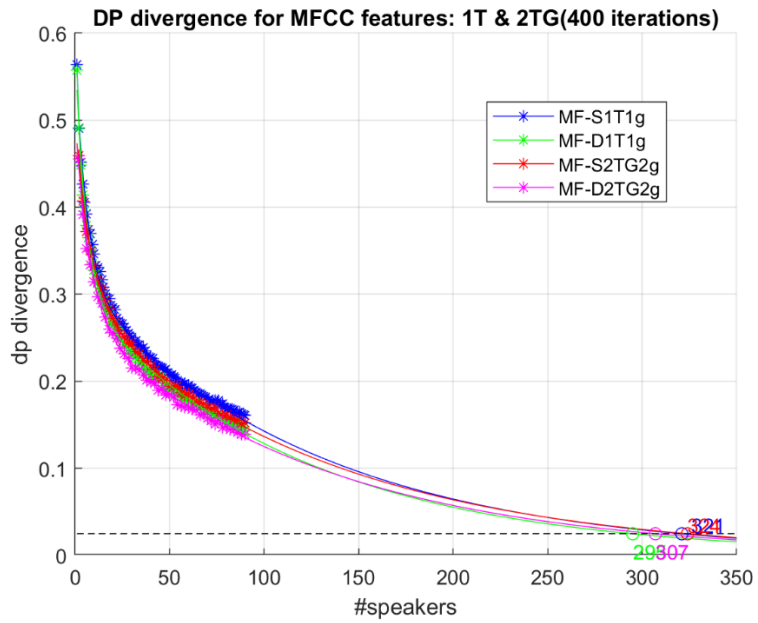
[9] R. L. Figueroa, Q. Zeng-Treitler, S. Kandula, and L. H. Ngo, "Predicting sample size required for classification performance," BMC Medical Informatics & Decision Making, vol. 12, no. 8, 2012.

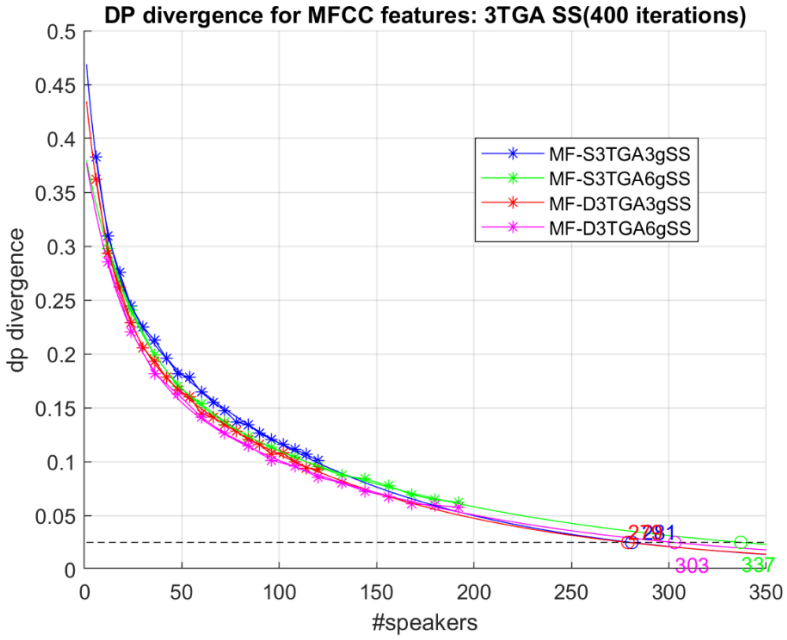
APPENDIX A

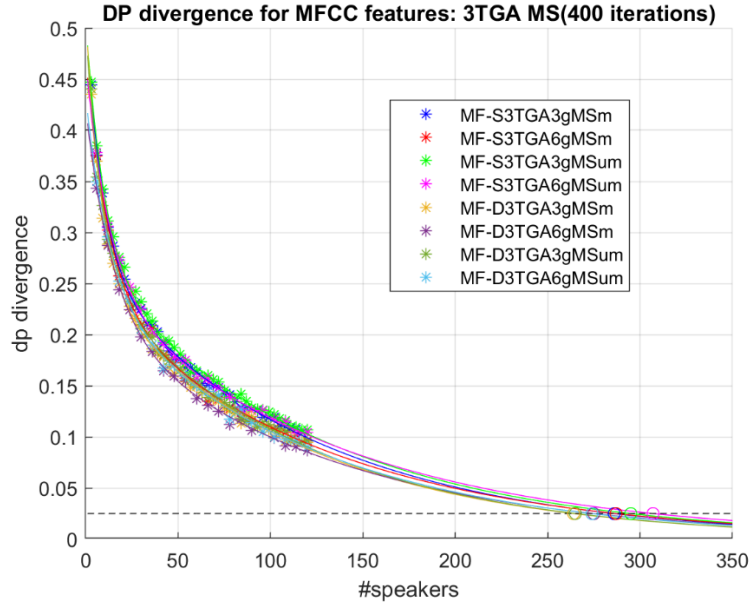
ADDITIONAL SIMULATION RESULTS OBTAINED

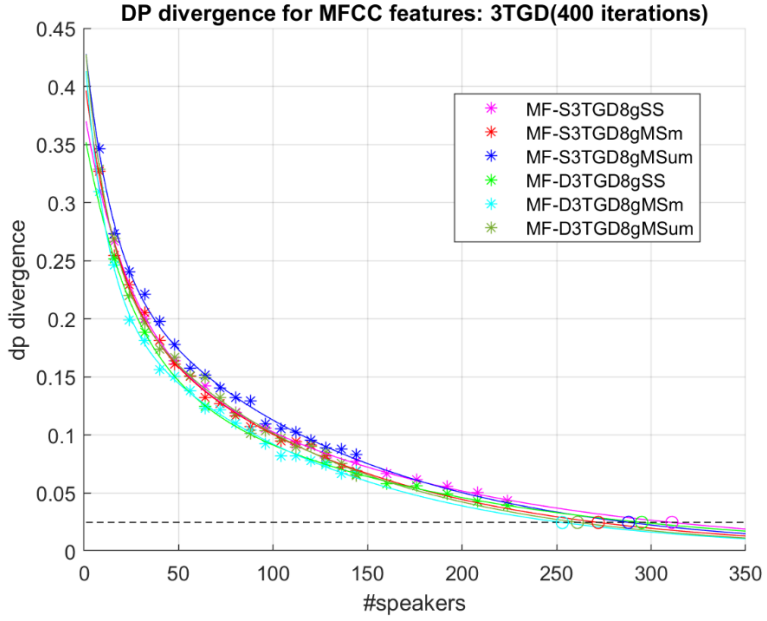
Given below are the D_p divergence graphs obtained for each of the analysis cases:

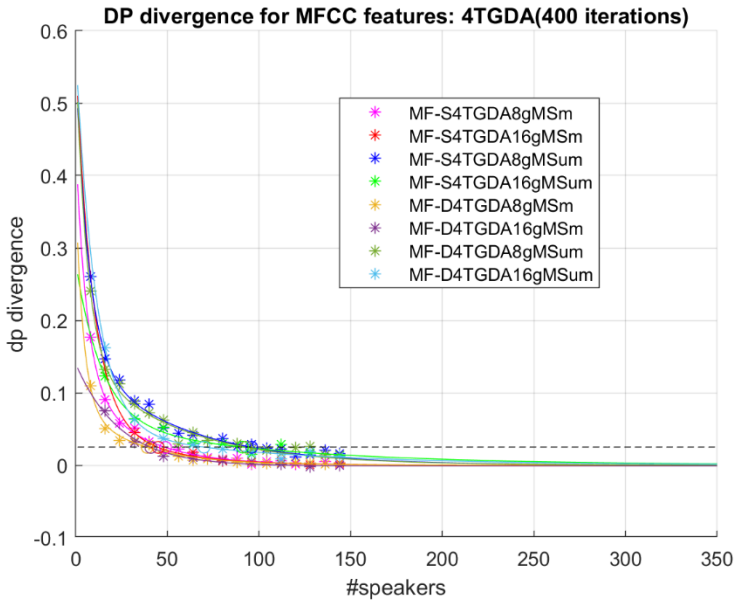
1. D_p divergence versus #speakers for MFCC features:



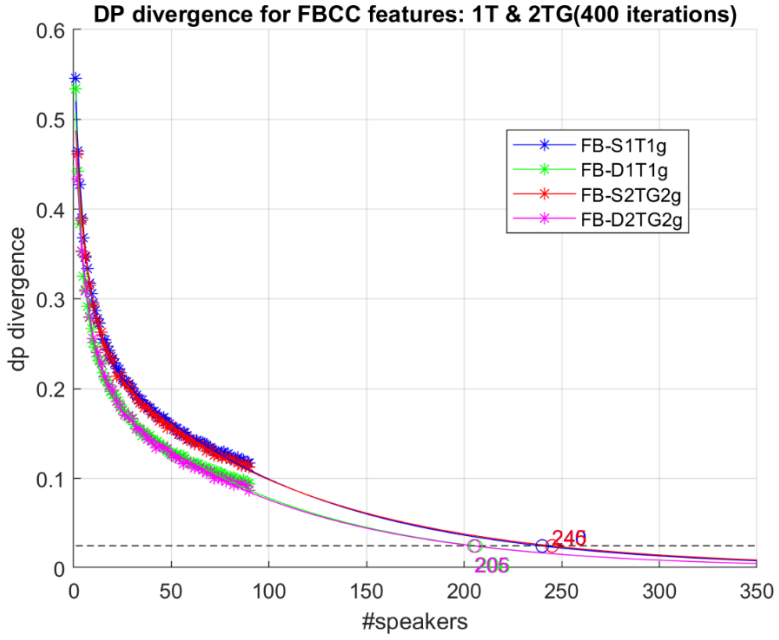


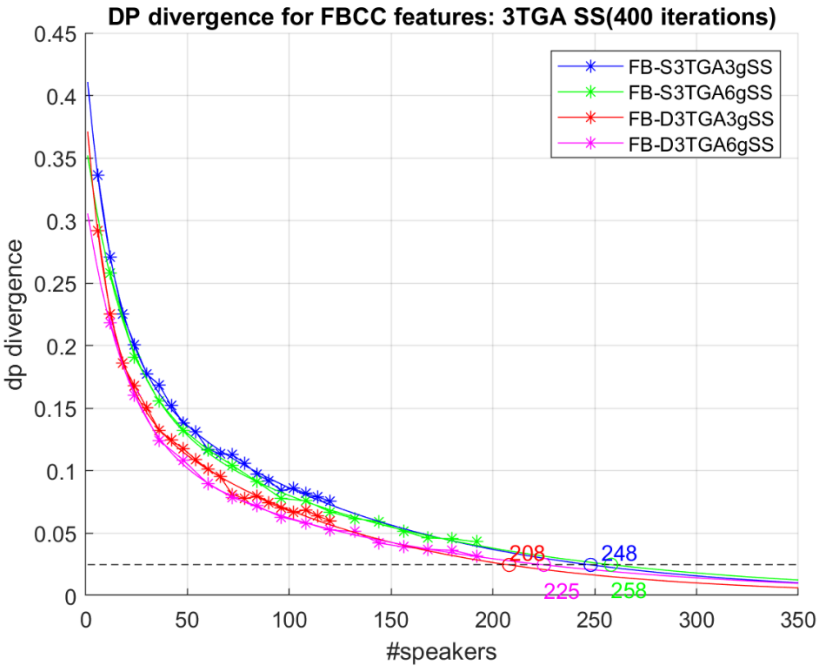

DP divergence for MFCC features: 3TGA MS(400 iterations)
dp divergence
#speakers
MF-S3TGA3gMSm
MF-S3TGA6gMSm
MF-S3TGA3gMSum
MF-S3TGA6gMSum
MF-D3TGA3gMSm
MF-D3TGA6gMSm
MF-D3TGA3gMSum
MF-D3TGA6gMSum

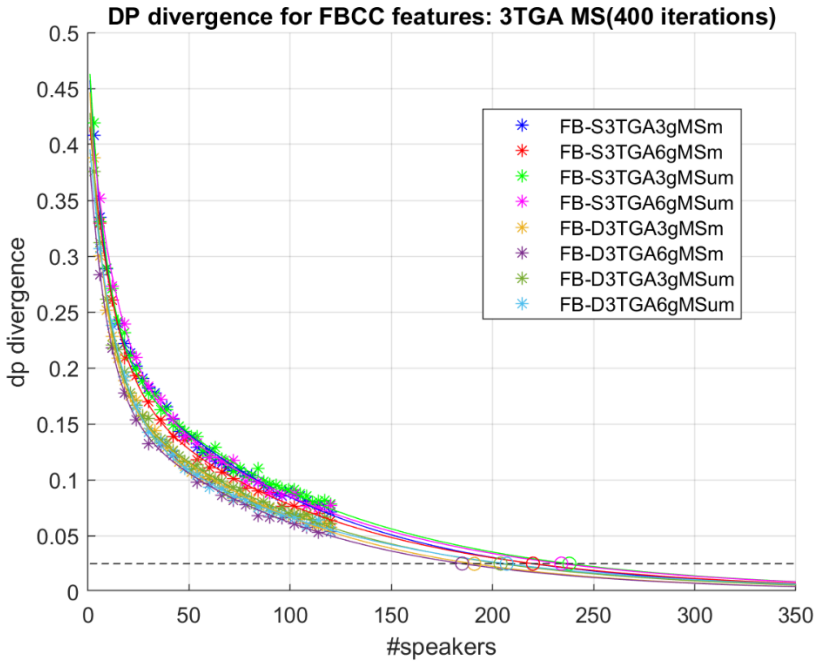

DP divergence for MFCC features: 3TGD(400 iterations)
dp divergence
#speakers
MF-S3TGD8gSS
MF-S3TGD8gMSm
MF-S3TGD8gMSum
MF-D3TGD8gSS
MF-D3TGD8gMSm
MF-D3TGD8gMSum

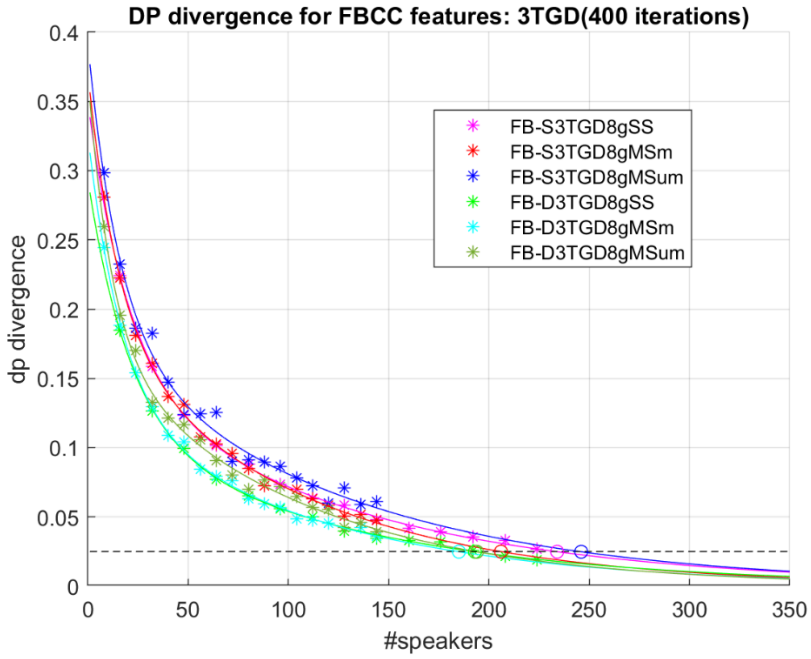


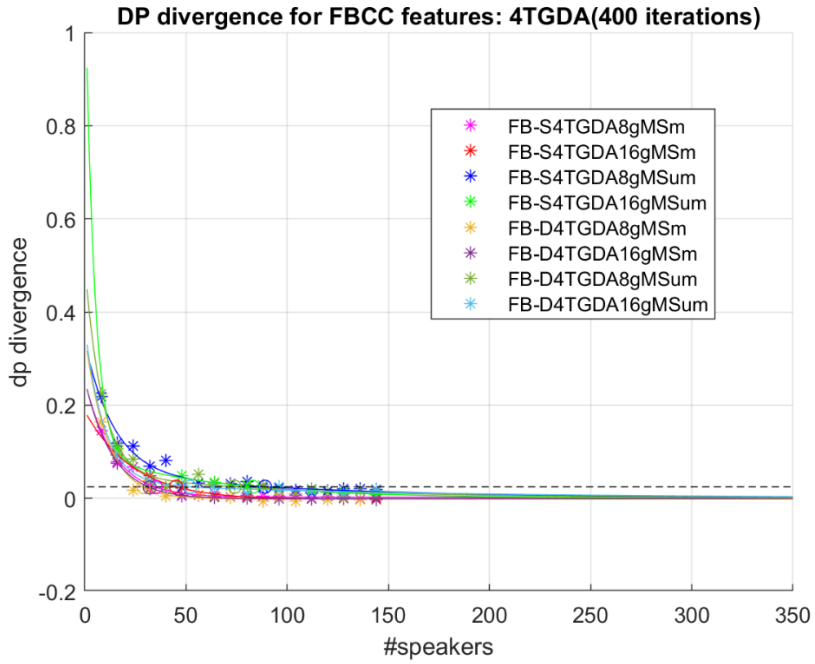
2. D_p divergence versus #speakers for FBCC features:




DP divergence for FBCC features: 3TGA SS(400 iterations)
FB-S3TGA3gSS
FB-S3TGA6gSS
FB-D3TGA3gSS
FB-D3TGA6gSS
dp divergence
#speakers
208
248
225
258


DP divergence for FBCC features: 3TGA MS(400 iterations)
FB-S3TGA3gMSm
FB-S3TGA6gMSm
FB-S3TGA3gMSum
FB-S3TGA6gMSum
FB-D3TGA3gMSm
FB-D3TGA6gMSm
FB-D3TGA3gMSum
FB-D3TGA6gMSum
dp divergence
#speakers


DP divergence for FBCC features: 3TGD(400 iterations)
dp divergence
#speakers
FB-S3TGD8gSS
FB-S3TGD8gMSm
FB-S3TGD8gMSum
FB-D3TGD8gSS
FB-D3TGD8gMSm
FB-D3TGD8gMSum


DP divergence for FBCC features: 4TGDA(400 iterations)
dp divergence
#speakers
FB-S4TGDA8gMSm
FB-S4TGDA16gMSm
FB-S4TGDA8gMSum
FB-S4TGDA16gMSum
FB-D4TGDA8gMSm
FB-D4TGDA16gMSm
FB-D4TGDA8gMSum
FB-D4TGDA16gMSum