Classification of Stuttering Events Using I-Vector

Samah A. Ghonem[1], Sherif Abdou[2], Mahmoud E. Shoman[3], Nivin Ghamry[4]

Information Technology Department, Faculty of Computers and Information, Cairo University

[1]samah.a.ghoname@gmail.com

[3]m.essmael@fci-cu.edu.eg

[2]sh.ma.abdou@gmail.com

[4]nivin@fci-cu.edu.eg

Abstract: ***Stuttering represents the main speech disfluency problem with the most two common stuttering disfluencies events are repetitions and prolongations. It is most desired to classify these disfluencies automatically rather than manually classification, which is a subjective, time-consuming task, and depends on speech language pathologists experience. In the proposed work, a new automatic classification approach is presented which depends on using the i-vector methodology that was usually used only in speaker verification/recognition applications, a sufficient accuracy relative to the amount of data used resulted as 52.43% ,69.56%,40%,50% for normal, repetition, prolongation, rep-pro[1] classes respectively and 64.75%,71.63% for normal, disfluent classes. Best accuracies for classifying the rep. and pro. classes with equal number of samples in each class resulted from the i-vector approach with 77.5%, 82.5% for rep., pro respectively compared to the Mel-Frequency Cepstrum Coefficients/Linear Prediction Cepstrum Coefficients (MFCC/LPCC)- K-Nearest Neighbour/Linear Discriminant Analysis (KNN/LDA) approaches tested on the same data set.***

Key words: *Stuttering, Disfluencies events, I-vector, Equal size classes.*

1 INTRODUCTION

Speech is the most natural way of communication between humans. It consists of articulation, voice, and fluency. Lungs are the source of energy for the sound, vocal cords affect the airflow from lungs to produce quasi-periodic excitation and the vocal tract shape controls the produced sound unit [19],[9].

Speech fluency is the main feature that affects transferring of information via speech. It is defined by the simplicity in connecting syllables, sounds, phrases and words together to form a message.

Stuttering is defined as a disorder in the flow of speech through involuntary syllable/word and phrase repetitions, interjections, silent pauses, hesitations, blocks and prolongations (Table 1 explains different stuttering events with an example on each)[1],[20]. There is 1% of the population suffers from speech stuttering, it affects both male and female but with ratio 3-4:1 times, which makes it one of the main problems in the speech and language pathology field. Speech pathology treatments help people who stutter to produce a fluent speech. Repetition and prolongation are the most frequent disfluencies in stuttered speech, so they have been used in stuttering assessment which is important before and after speech therapy to evaluate the stuttering performance, Speech language pathologists SLP usually do assessment through manually counting and classifying the number of occurrences of disfluencies after transcribing the recorded speech, but this method is time-consuming, prone to error, subjective and inconsistent also it depends on the experience of SLPs, so it is better to automate the classification of disfluencies using speech recognition technologies and computational intelligence.

The rest of the paper is organized as follows: In section 2, the past solutions for the predetermined problem via different classification and feature extraction techniques are presented. In section 3, the used dataset has been defined. In section 4, the proposed work is introduced with a brief description of each step in the work presented in a separate subsection. Section 5, shows the experimental results of the work, and section 6 present the paper conclusion with the future work.

TABLE 1: STUTTERING EVENTS

Stuttering disfluency	Description	Example
Syllable repetition	Repeating a syllable	w w where is she going?
Phrase repetition	Repeating a phrase	Where is where is she going?
Word repetition	Repeating a whole word	Where where is she going?
Prolongation	Extending a sound for a long duration	Where is shshshshshe going?
Interjections	Adding irrelevant meaningless words	Where is um she going?
Silent pauses	Adding pause within a word	Where is she go-(pause)-ing?
Blocks	Adding a long silence between words	Where is she (block) going?

[1] The category that includes segments of both repetition and prolongation disfluencies appear at the same time.

2 LITERATURE REVIEW

Many research works have been done on the problem of classification of stuttering events in past years, most of these research works have focused especially on the recognition of repetition and prolongation stuttering events as the most common events appears in stutterer's speech. **In [1]**, the authors used Mel-frequency cepstral coefficients (MFCC), perceptual linear predictive (PLP) features to represent speech signals, then with calculating a similarity matrix and using some morphological image processing tools to process the similarity matrix after converting it to an image, the disfluent parts are detected with the best accuracy of 99.84% for prolongation detection, 98.07%, 99.87% for syllable/word and phrase repetition detection respectively. **In [7]**, speech signals were represented using six acoustic features: volume, zero-crossing rate, spectral entropy, high order derivatives, VH and VE curve, and speech segments were detected using end-point detection technique according to the VH curve threshold, dynamic time warping (DTW) was used for repetition recognition with an accuracy of 83%. GMM and Gaussian super vector support vector machine (SVM) have been used for repetition and prolongation disfluencies classification **in [6]** with MFCC and its derivatives as a speech parameterization technique, with a resulting accuracy of 93.79% from GMM classifier and of 98.24% from GMM-SVM one. Prosodic features that represent several human speech characteristics like speaking rate, pitch, loudness, energy, and duration have been used with cepstral features, MFCC, delta Mel-frequency cepstral coefficients (DMFCC), delta-delta Mel-frequency cepstral coefficients (DDMFCC), in classification of repetition and prolongation stuttering disfluencies **in [4]** with the results of classification have been tested using SVM classifier, Using cepstral features alone with SVM obtained an accuracy of 84.73% unlike using the combination of prosodic features and cepstral ones, which obtained a 96.85% accuracy with an improvement in performance of classification between 2 and 3% than using cepstral features alone. **In [2]**, three types of features were used to detect repetition disfluency in stuttered speech, these features were MFCC, formants, shimmer, with a total feature vector of 17 features (13 MFCC + 3 formants + 1 shimmer), and after extracting these features from each isolated unit of the data used, they are compared to the features of the subsequent units and using the dynamic time warping DTW technique with a suitable threshold, each unit is classified to be either repetition or no repetition one with a resulting classification accuracy of 94%. **In [5]**, classification of four types of disfluencies have been proposed using MFCC as feature extraction technique used to estimate the Gaussian mixture model (GMM) model parameters, with a resulting performance of 96.40%, 95.0%, 95.70%, 98.60% for syllable repetition, word repetition, prolongation, interjection disfluencies respectively with the best accuracy resulted from using a 64 GMM mixture components. Artificial Neural Network (ANN) has also been used as a classifier **in [3]** for detection of repetition and prolongation stuttering disfluencies with different combinations of acoustic and pitch related features (MFCC, pitch, energy, zero crossing rate and formants) used as a representation of the speech signals, the best accuracy had been achieved using MFCC alone with a 88.29% average accuracy while the combination of formants, pitch, energy features achieved best accuracy for repetition detection with 94.52% and the best accuracy for prolongation disfluencies detection was 96.71% from using MFCC and formants features. Authors **in [11]** have addressed the difference between using MFCC and linear prediction cepstral coefficients (LPCC) in recognizing repetition and prolongation stuttering events, the K-nearest neighbor(KNN) and linear discriminant analysis(LDA) classifiers have been used to test these two feature extraction techniques with a resulting accuracy of 92.55% for using MFCC and 94.51% for LPCC, with a conclusion that LPCC outperforms MFCC in such a problem.

3 EXPERIMENTAL DATA

The part of data used in this work have been chosen from the College London Archive of Stuttered Speech (UCLASS) database [16] that is designed to be used for research and clinical purposes to detect the language and speech behavior of stuttered speakers. In the proposed work, the subset used from the UCLASS database contains 39 reading records, consists of 18 speakers (males, and females) with a wide range of age (between 8 and 20 years) and stuttering events.

4 METHODOLOGY

I-Vector methodology has been recently used in speaker recognition/verification applications and proved significant results in such applications. In the proposed work, I-Vector has been used for classification of the two most common stuttering disfluencies which are repetitions and prolongations. The process of applying the I-Vector methodology in stuttering disfluencies classification contains four main steps which are: Data Segmentation, Feature Extraction, I-Vector algorithm, and Classification. **Figure 1** shows these main steps, with the sub-steps of each. In the following subsections, each of these steps will be explained concisely.

A. Feature Extraction

It is the process of extracting parameters from the speech signal that represent it to be used in modeling and pattern matching techniques of speech processing applications.

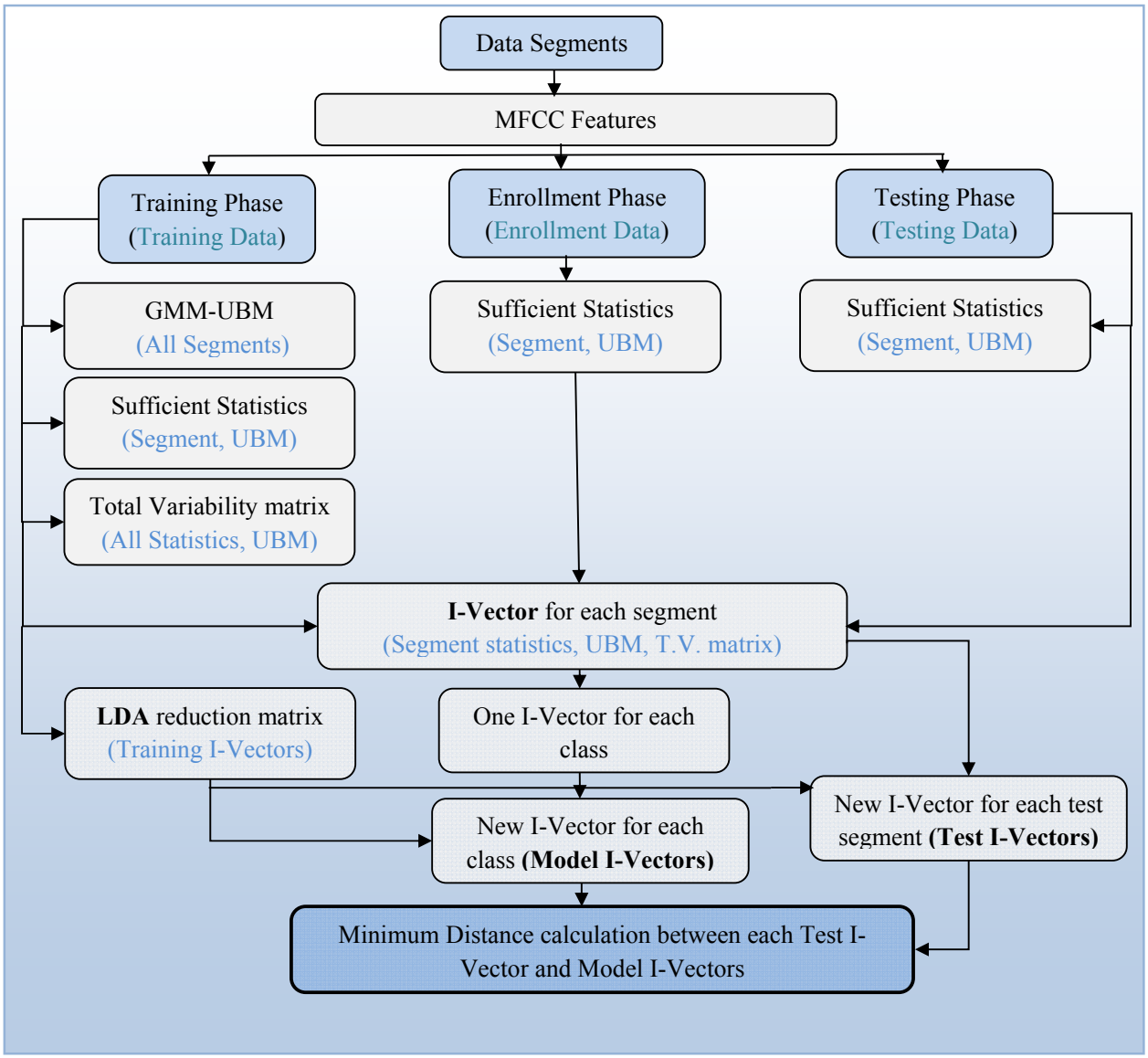

Figure 1: I-Vector Methodology

Mel Frequency Cepstral Coefficients (MFCC) is considered the most popular used feature extraction technique that characterizes speech signals in both speech and speaker recognition applications due to its robustness in recognition tasks related to the human voice. That is because they mostly represent phonetic and vocal cavity information and so represent the human auditory response more closely. It is also based on the variation known of the human perception to different frequencies, and so a **Mel-scale** is used such that it deals with the frequencies up to 1 KHZ linearly and logarithmically with frequencies above 1 KHZ, so emphasize more the lower frequency range [18].

The result of MFCC is a multidimensional feature vector for each speech frame, mostly 39 features are used (12-MFCC + Energy), their delta and their delta-delta coefficients.

The following is a brief description for steps needed to extract MFCCs from a speech signal:

1) Pre-emphasize: Passing the speech signal through a filter to emphasize higher frequencies:

$$y[n] = x[n] - ax[n-1] \quad , 0.9 \le a \le 1 \qquad (1)$$

where x[n] is the input speech signal, and y[n] is the output signal.

2) *Framing and Windowing:* To divide the speech signal into frames of size (10-msec:30-msec), with an overlap between each two adjacent frames ranging from 25% to 70% of the frame size, then using the window function (2) to minimize discontinuities between signal frames.

$$w[n] = 5.4 - 0.46\cos\left(\frac{2\pi n}{n-1}\right) \quad , 0 \le n \le N\text{-}1 \qquad (2)$$

3) *Signal Spectrum:* Fast Fourier transform (FFT) is applied on windowed frames to get the magnitude frequency response(spectrum) of these frames.

4) *Mel-Scale and Filter bank analysis:* Mel frequencies can be obtained from linear frequencies through this transformation:

$$Mel(f) = 2595\log_{10}\left(1 + \frac{f}{700}\right) \qquad (3)$$

After that, a set of triangular band pass filters (filter bank) which are equally spaced on the Mel-scale are applied on the resulting spectrum from Step-3 to get the energy of each filter. Then a logarithmic function is applied on the resulting energies to get the log energies of filters.

5) *Cepstrum Coefficients:* Finally, a discrete cosine transform (DCT) is performed on the log energies to transform from frequency domain analysis to time domain analysis and extract the Mel-frequency cepstrum coefficients through (4).

$$C_m = \sum_{k=1}^{N} \cos[m(k-0.5)]\, E_k \quad , m = 0,1,2,\dots L \qquad (4)$$

where **N** is the number of filters, **L** is the number of MFCCs, and $\mathbf{E_k}$ is the log energies obtained from filters. Dynamic behavior of the speech signal can be computed through equation (5) which computes either Delta (1st derivative) or Delta-Delta (2nd derivative) cepstral coefficients which represent the speech rate and speech acceleration respectively.

$$\Delta C_m(l) = 0.5\,[C_m(l+1) - C_m(l-1)] \qquad (5)$$

where $\boldsymbol{\Delta C_m(l)}$ is the delta or delta-delta cepstral coefficients at frame l.

B. I-Vector Algorithm

I-vector is a factor analysis front end approach that is more developed than joint factor analysis (JFA) approach, and has been recently used in speaker recognition/verification applications. Its main idea depends on representing the GMM-Super Vector by a low dimensional vector that contains both speaker and channel variability subspaces in a single total-variability space [12],[15].

In the proposed work, I-vector methodology have been used in the context of stuttering disfluencies classification through three main phases: training, enrollment (modeling) and testing, in the following subsections steps for each phase is illustrated:

1) *Building a UBM:* A universal background model (GMM-UBM) is built from the set of MFCC feature vectors that were extracted from the whole training speech segments to represent the general distribution of the training data features [14],[17].

The utterance GMM super vector M is defined as:

$$M = m + Tw \qquad (6)$$

where **m** is the UBM super vector, **w** is a random vector having a standard normal distribution, represents the total factors, and its mean called the identity vector or (I-vector), **T** is the total variability matrix.

2) *Sufficient Statistics:* Sufficient statistics represent Baum-Welch statistics that are calculated for each audio segment to be used next in extracting the corresponding I-vector [13]. Equations (7),(8) represent the extracted statistics for a segment with L frames, using the UBM Ω which consists of a C mixture components:

$$N_c = \sum_{t=1}^{L} Pr(c|y_t, \Omega) \qquad (7)$$

$$F_c = \sum_{t=1}^{L} Pr(c|y_t, \Omega)\, y_t \qquad (8)$$

where c = 1,2,...C and $Pr(c|y_t, \Omega)$is the posterior probability of mixture c generating the vector y_t.

3) *Total Variability Space:* The total variability space, represented by a total variability matrix T, is used as a feature extractor to extract the I-vector with a specific dimension/length from each speech segment. It contains the primary directions of variability from all training data and so it is trained using the training data sufficient statistics and the UBM through applying the Expectation Maximization (EM) algorithm, which calculates the maximum likelihood estimate of the total variability matrix [8], [13].

4) *I-Vector Extraction:* After obtaining the total variability matrix, it is used as a feature extractor to extract the I-vector for each segment which is the mean of a hidden variable w that is defined by its posterior distribution conditioned to the sufficient statistics of a segment [13]. The following equation (9) is used to extract the I-vector for each segment.

$$w = \left(I + T^t \sum^{-1} N(u)\, T\right)^{-1} T^t \sum^{-1} \tilde{F}(u) \qquad (9)$$

I: is the identity matrix, **T**: is the total variability matrix, ∑: is the diagonal covariance matrix which models the residual variabilities not covered by T, **N(u)**: is a diagonal matrix of diagonal elements N_cI, and $\widetilde{\boldsymbol{F}}(\boldsymbol{u})$: is the first order statistics for a segment u and is given by:

$$\widetilde{F_c} = \sum_{t=1}^{L} Pr(c|y_t, \Omega)\,(y_t - m_c) \qquad (10)$$

where m_c is the mean of mixture component c of the UBM.

5) *LDA:* In the proposed work, linear discriminant analysis (LDA) is used to maximize the between-class variations S_B while minimizing the within-class variations S_W through the I-vector dimensionality reduction, this is done after getting the eigenvector matrix(eigenvectors with the highest eigen values) and multiplying it by each I-vector to obtain the new comparable I-vectors [10]. This means maximizing the Fisher's criterion:

$$\max_v \left(\frac{V^t S_B V}{V^t S_W V}\right) \qquad (11)$$

and V is the eigenvector matrix obtained from solving the eigen value system: S_B V = D S_W V , where D is the eigen value.

C. Classification

The classification step in the proposed work was done with the minimum distance classifier using the Euclidean distance (12):

$$dist = \sqrt{\sum (model - test)^2} \qquad (12)$$

in which distance between each test I-vector is measured against all model i-vectors, and then the test segment is classified to belong to the model/class with the minimum distance.

5. EXPERIMENTAL RESULTS

In the proposed work, the dataset used was segmented manually into 1380 segments which are divided into the four categories: Normal, Repetitions, Prolongations, and Repetitions-Prolongations, with a part of these segments for each category divided as 852, 351, 85, and 92 for Normal, Repetitions, Prolongations, and Repetitions-Prolongations respectively.

Training phase has used around 80% of these segments from 14 speakers and the rest 20% from the 4 remaining speakers have been used for enrollment and testing phases with approximately half of the files used for enrollment and the other half for testing.

The proposed work has proved a fair accuracy in using the I-vector methodology in the area of stuttering disfluencies classification. The following subsections discuss the classification done by different approaches for the four classes of interest and for Repetition, Prolongation classes only.

The approaches used are two of the frequently features/classifiers sets that have been used in some research works before for solving the same problem of disfluencies classification specially repetition and prolongation disfluencies. These sets contain MFCC, LPCC as the features extraction techniques and KNN, LDA as the classifiers.

A. Four Classes Classification:

Tables 2 and 3 represent the resulting accuracies in the 4-classes classification with different combinations of (old I-vector length/new I-vector length after dimensionality reduction). **Table 2**: represents the accuracies when data segments classified into the four interested classes, while **Table 3** represents the accuracies when data is classified to be either normal or disfluent only, with the disfluent class contains the data of the (repetition, prolongation, rep-pro) classes.

Figure 2 represents the comparison done by the different approaches in classifying the pre-described segments into the four classes.

TABLE 2: THE 4-CLASSES CLASSIFICATION ACCURACIES FOR DIFFERENT I-VECTOR SIZES

I-Vector Size	Normal	Repetition	Prolongation	Rep-Pro
150/75	50.12%	64.24%	38%	17.5%
200/175	47.56%	69.56%	33.33%	50%
300/200	**52.43%**	**69.56%**	**40%**	**50%**

TABLE 3: THE CLASSIFICATION ACCURACIES FOR NORMAL AND DISFLUENT CLASSES

I-Vector Size	Normal	Disfluent
150/75	64.75%	71.63%
200/175	61.49%	71.42%

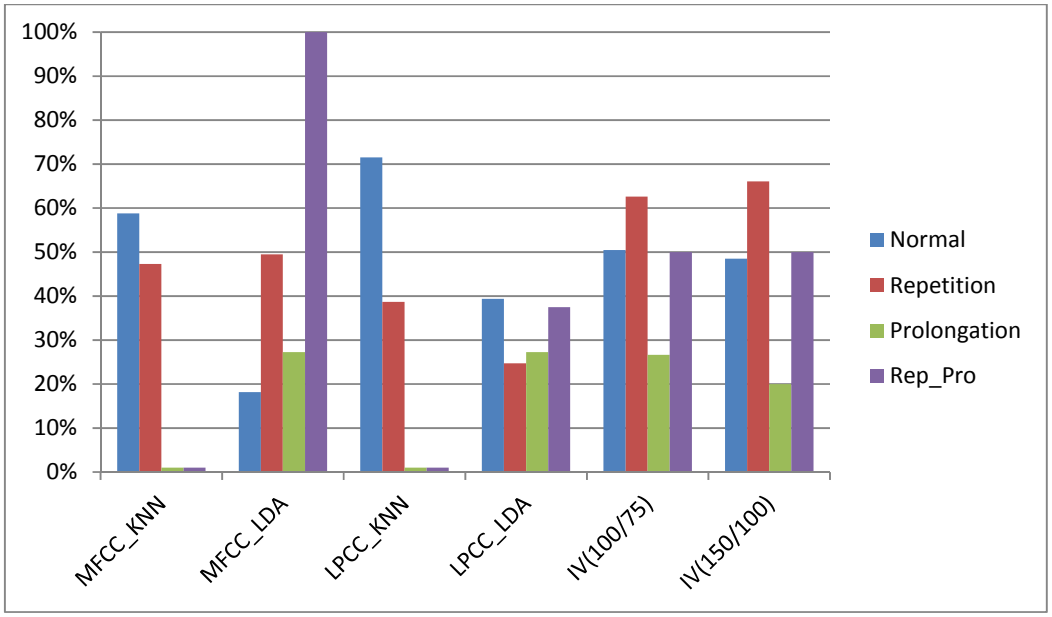

Figure 2: Different Approaches with 4-Classes Classification

From the presented tables, it is noticed that I-vector results vary between classes according to number of training, enrollment and testing segments relative to the total number of segments used in the training process of the UBM and total variability matrix.

Regarding the results in figure 2, when calculating the average accuracy of all classes for each approach, the I-vector will produce 47.43%, 46.15% for IV(100/75) and IV(150/100) respectively where LDA average accuracies are 48.72% (MFCC/LDA), and 32.22% (LPCC/LDA) and the KNN with 26.52% (MFCC/KNN) and 27.55% (LPCC/KNN). This indicates that KNN as expected will not perform well for classes with variable length, but LDA and I-vector can produce a better but not optimal performance.

B. Repetition and Prolongation Classification:

Almost all research works done in stuttering classification area concentrate on classification of repetition and prolongation classes due to their frequently occurrences in stuttering speech. So that, a comparison done with the two

different approaches regarding these two classes only. **Tables 4 and 5** present these comparisons in terms of different frame lengths, different combinations of feature/classifier set and two different old/new I-vector lengths.

Table 4: shows the result of this comparison with a total number of samples of 351,85 for repetition and prolongation respectively with a 80% / 20% of samples are chosen for training and testing while **Table 5:** shows the same comparison presented in table 4 but with an equal number of training and testing samples in both classes (69 samples for training and 16 samples for testing).

TABLE 4: THE CLASSIFICATION ACCURACIES FOR REPETITION AND PROLONGATION CLASSES

		MFCC/KNN	MFCC/LDA	LPCC/KNN	LPCC/LDA	IV(100/75)	IV(150/100)
10ms	**Rep.**	**98.57%**	58.57%	90%	51.42%	65.55%	66.66%
	Pro.	12.5%	43.75%	18.75%	62.5%	**92.5%**	90%
20ms	**Rep.**	**95.71%**	58.57%	91.43%	51.42%	65.55%	65.55%
	Pro.	18.75%	50%	18.75%	62.5%	**92.5%**	72.5%
30ms	**Rep.**	**98.57%**	57.14%	91.43%	54.28%	62.22%	63.88%
	Pro.	12.5%	43.75%	18.75%	62.5%	95%	**97.5%**

TABLE 5: CLASSIFICATION ACCURACIES FOR EQUAL REPETITION AND PROLONGATION CLASSES

		MFCC/KNN	MFCC/LDA	LPCC/KNN	LPCC/LDA	IV(100/75)	IV(150/100)
10ms	**Rep.**	50%	50%	50%	43.75%	60%	**65%**
	Pro.	68.75%	68.75%	43.75%	75%	**77.5%**	42.5%
20ms	**Rep.**	56.25%	43.75%	50%	37.5%	**62. 5%**	57.5%
	Pro.	50%	56.25%	50%	75%	**82.5%**	65%
30ms	**Rep.**	50%	43.75%	50%	37.5%	62.5%	**77.5%**
	Pro.	50%	56.25%	56.25%	68.75%	**82.5%**	67.5%

It can be noticed from these tables that in case of different classes size, the KNN produced the best accuracy in repetition classification due to the large amount of data in this class (351 samples) with relative to the amount of data in prolongation class (86 samples) which dominate the neighbors of test samples and lead to the high misclassification rate of prolongation samples, while I-vector approach produced the best accuracy in prolongation classification due to the small number of test segments which equal to half of the test samples number as the other half is used for enrollment step and production of the each class reference I-vector which lead to a high accuracy with a small number of truly classified samples.

In case of equal classes size, the I-vector approach has outperformed the other approaches and achieved a sufficient accuracy in both classes as they have the same number of training, enrollment, and test samples.

6. Conclusion

A new applied approach to the classification of stuttering disfluencies was presented in this paper. This approach depends on using I-vector methodology, and extracting one model I-vector for each class of the four classes we are interested in during this study (normal, repetition, prolongation and rep-pro) with also extracting an I-vector for each test speech segment, then comparing each test I-vector with the four models I-vectors using the minimum distance classifier to decide the class this test segment relates to. This approach has achieved a classification accuracy of 52.43%, 69.56%, 40%, 50% for the normal, repetition, prolongation and rep-pro classes respectively and a classification accuracy of 64.75%, 71.63% for the main normal and disfluent classes. It is concluded from testing different combinations of I-vector lengths before and after dimensionality reduction that as the number of classes increases, the accuracies will be increased also while exceeding the I-vector length.

The results of I-vector are also compared with MFCC/LPCC, KNN/LDA features/classifiers sets to classify the four classes of interest and to classify repetition and prolongation classes only. The results indicate that for equally sized classes with a small amount of data used for training and testing, the I-vector resulted in the best accuracy with relative to other approaches used and so as the size of classes increase equally, it is expected to improve the accuracy of the classification process. While in different classes size, the I-vector may be not the optimal approach although it achieved a fair accuracy according to the amount of data for each class.

The presented results introduce using the I-vector approach in speech recognition tasks in addition to its use in speaker recognition applications which can help researchers to try improving its performance in speech recognition tasks.

Future work will focus on increasing the amount of training and testing segments for all classes, combine I-vector with other features and test different classifiers than the minimum distance one. Comparing the results with different approaches used in the same problem, testing the I-vector approach on different stuttering events like(blocks, hesitations and so on), and testing the approach on Arabic speech disfluencies segments.

REFERENCES

[1] I. Esmaili et al., "Automatic Classification of Speech Disfluencies in Continuous Speech Based on Similarity Measures and Morphological Image Processing Tools", *Journal of Biomedical Signal Processing and Control*, vol.23, pp.104-114, (2016).

[2] Shashidhar G. Koolagudi, Pravin B. Ramteke and Fathima Afroz, "Repetition Detection in Stuttered Speech", *Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics*, Vol.43, India, (2016).

[3] Pravin B. Ramteke, P.S. Savin and Shashidhar G. Koolagudi, "Recognition of Repetition and Prolongation in Stuttered Speech Using ANN", *Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics*, Vol.43, India, (2016).

[4] P. Mahesha and D.S. Vinod, "Combining Cepstral and Prosodic Features for Classification of Dysuencies in Stuttered Speech", *Intelligent Computing, Communication and Devices(ICCD)*, Vol.308, pp.623-633, India, (2015).

[5] P. Mahesha and D.S. Vinod, "Gaussian Mixture Model Based Classification of Stuttering Dysuencies", *Journal of Intelligent Systems,* vol.25, pp.387-399, (2015).

[6] P. Mahesha and D.S. Vinod, "Support Vector Machine-Based Stuttering Dysfluency Classification Using GMM Supervectors", *International Journal of Grid and Utility Computing*, vol.6, (2015).

[7] C.C. Yang, P.H. Yeh, S.L. Yang and M.D. Shieh, "Automatic Recognition of Repetitions in Stuttered Speech: Using End-point Detection and Dynamic Time Warping", *10th Oxford Dysfluency Conference(ODC), Procedia-Social and Behavioral Sciences*, Vol.193, p.356, Oxford, United Kingdom, (2015).

[8] Man-Wai Mak, Wei Rao and Kong-Aik Lee, "Normalization of Total Variability Matrix for i-vector /PLDA Speaker Verification", *IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP)*, (2015).

[9] Monica Mundada et al., "Recognition and Classification of Speech and Its Related Fluency Disorders", *International Journal of Computer Science and Information Technologies (IJCSIT)*, vol.5, no.5, pp.6764-6767, (2014).

[10] P. Xanthopoulos et al., *Robust Data Mining, Springer Briefs in Optimization, Chap.4*, Springer, (2013).

[11] O. Chia Ai et al., "Classification of Speech Dysfluencies with MFCC and LPCC Features", *International Journal of Expert Systems with Applications*, vol.39, pp.2157-2165, (2012).

[12] Ahilan Kanagasundaram et al., "I-vector Based Speaker Recognition on Short Utterances", *Proceedings of the 12th Annual Conference of the International Speech Communication Association* (ISCA), (2011).

[13] Dehak et al., "Front-end Factor Analysis for Speaker Verification", *IEEE Transactions on Audio, Speech, and Language Processing*, vol.19, no.4, pp.788-798, (2011).

[14] Taufiq Hasan and John H. L. Hansen, "A Study on Universal Background Model Training in Speaker Verification", *IEEE Transactions on Audio, Speech, and Language Processing*, vol.19, no.7, pp.1890-1899, (2011).

[15] Jean-Francois Bonastre, Pierre-Michel, Bousquet and Driss Matrouf, "Intersession Compensation and Scoring Methods in the i-vectors Space for Speaker Recognition", *INTERSPEECH*, (2011).

[16] J. Bartrip, P. Howell and S. Davis, "The University College London Archive of Stuttered Speech (UCLASS)", *Journal of Speech, Language, and Hearing Research JSLHR*, vol.52, no.2, pp.556-569, (2009).

[17] Daniel Povey et al., "Universal Background Model Based Speech Recognition", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (2008).

[18] Fang Zheng, Guoliang Zhang, and Zhanjiang Song, "Comparison of Different Implementations of MFCC", *Journal of Computer Science& Technology*, vol.16, no.6, pp.582-589, (2001).

[19] Lawrence Rabiner and Biing-Hwang Juang, *Fundamentals of speech recognition*, 1st ed, Prentice-Hall, (1993).

[20] William H. Perkins, "What Is Stuttering?", *Journal of Speech and Hearing Disorders*, vol.55, pp.370-382, (1990).

BIOGRAPHY

Samah A. Ghonem, received the B.Sc. degree from Faculty of Computers and Information, Cairo University in 2010, in Information Technology and Information Systems as the major and minor specializations respectively. From 2011 to date, she is working as a Teaching Assistant at Faculty of Computers and Information, Cairo University. Her research interests are Speech Processing, Signal Processing, and Machine learning.

Sherif Mahdy Abdou received his B.Sc. and M.Sc. degrees in computer science and automatic control from University of Alexandria, Egypt in 1993 and 1997, respectively. He received a Ph.D degree in Electrical and Computer Engineering from University of Miami, USA in 2003. In 2003 Dr. Abdou joined BBN Technologies as a senior staff scientist in the Arabic language team of the Ears project to provide affordable reusable speech-to-text decoding for the Defense Advanced Research Projects Agency, DARPA. In 2005 Dr. Abdou was appointed as the research

and development manager of the Research and Development International (RDI) company where he is leading a team to develop several products for natural language processing, computer aided language learning, speech recognition, speech synthesis, optical character recognition, handwriting recognition with special focus on the technologies of the Arabic language. In 2005 Dr. Abdou joined Cairo University as an Assistant Professor at the Information Technology department at the Faculty of Computers and Information. Dr. Abdou is one of the holders of the patent "Systems and Methods for Quran Recitations Rules: HAFSS". Dr. Abdou is a member of the review committee in several conferences and journals in the HLT fields and is the Principal Investigator and Co- Principal Investigator of several research projects in the areas of Language learning, Virtual tutors, Web monitoring and Intelligent Contact Centers.

Mahmoud A. Ismail, Ph.D., *ACM Member***:** Mahmoud Shoman is a Full Professor of Information Technology at the Faculty of Computers and Information, Cairo University. He is currently the Vice Dean of Postgraduate Studies and Research. He also serves as the Chief Information Officer (CIO) of Cairo University where he plans and monitors the implementation of all strategic IT projects in the university. He has supervised more than 30 Ph.D. and M.Sc. students and published more than 45 research papers in international journals and conferences. His research interests include Pattern Recognition, Digital Signal Processing, Evolutionary Computation, Speech Processing, Information Security and Data Hiding. Dr. Mahmoud received his B.Sc. (Distinction with Honors) and M.Sc. in Electronics and Communication Engineering in 1990 and 1994, respectively, from Cairo University. He received the Ph.D. in Computers and Systems Engineering in 1998 from Ain Shams University.

Nivin Ghamry, received the B.Sc., M.Sc., and Ph.D. degrees from Cairo University of Engineering, Egypt in 1995, 1998, and 2003, respectively, all in electrical engineering, Electronics and Communication department. From 1995 to 2006, she was a researcher in the Electronics Research Institute, Cairo, VLSI departments. From 2006 to 2010, she was a lecturer at the Faculty of Engineering, Fayoum University, Egypt, Electronics and Communication department. From 2008 to 2010 she was a visiting researcher at the Technical University of Berlin, German, control and Diagnosis department as well as the Institute of Informatics of the Humboldt University of Berlin. From 2010 to 2012 she is working as an assistant Professor (Lecturer and Researcher) and from 2012 to date as an Associate Professor at Faculty of Computers and Information, Cairo University. Her research interests include the areas of Design of embedded systems, digital signal processing, Linear and Nonlinear System Identification and VLSI applications.

ARABIC ABSTRACT

تصنيف عوامل التمتمة باستخدام I-Vector

سماح أ. غنيم[1]، شريف عبده [2] ، محمود إ. شومان [3] ، نيفين غمري [4]

قسم تكنولوجيا المعلومات،كلية الحاسبات و المعلومات، جامعة القاهرة

ملخص

التمتمة تمثل المشكلة الرئيسية في التخاطب من خلال أهم عاملين و هما التكرار و الإطالة. و من المهم تصنيف هذه العوامل أوتوماتيكيا بدلا من التصنيف اليدوى الغير موضوعي، المستهلك للوقت و المعتمد علي خبرة أطباء التخاطب. في العمل المقترح يتم تقديم طريقة جديدة في التصنيف الأوتوماتيكي و المعتمدة علي استخدام منهجية ال I-vector المعتاد استخدامها في تطبيقات التعرف و التأكد من المتحدث. و ارتباطا بكمية البيانات المستخدمة تم الحصول علي النتائج 52.43%، 69.56%، 40% و 50% للتصنيفات (طبيعي، تكرار، إطالة، تكرار و إطالة) و 64.75% و 71.63% للتصنيف إلي طبيعي و متلعثم علي الترتيب. أفضل نتائج لتصنيف التكرار و الإطالة عند استخدام عدد متساوي من العينات لكل فئة تم الحصول عليها من استخدام ال i-vector بدقة 77.5% و 82.5% للتكرار و الإطالة و ذلك بالمقارنة بمنهجيات ال MFCC/LPCC – KNN/LDA المختبرة علي نفس مجموعة البيانات.