Speaker Diarization in Personal Video Recordings Based on LDA and User Feedback

Zeenat Afroze
Department of Electrical and Electronic Engineering, American International University Bangladesh, Dhaka, Bangladesh
Email: afroze@aiub.edu

***Abstract*—In this paper, we present the speaker diarization system which is used in personal video recordings. Speaker diarization begins by the extraction of relevant features from the input signal. Features are measurable characteristics which are important to the distinction between different classes. They should have low inter-class similarity and also low intra-class variability. So, LDA is used to cope intra-speaker variability. We demonstrate improvement of the performance over the baseline system based on LDA. This paper also reports the performance of the speaker diarization system. A new approach, called the user feedback has been developed here to improve the performance of the speaker diarization system.**

***Index Terms*—agglomerative clustering algorithm, bayesian information criterion, mel frequency cepstral coefficients, speaker diarization, user feedback**

I. Introduction

The goal of the speaker diarization system is to answer the question: "Who spoke when?" The task of the Diarization system consists in detecting homogenous audio segments. It means that each segment contains the information of only one speaker. Clusters are composed of many segments. Each cluster contains segments of only one speaker. The speaker diarization algorithm is a very useful part of the speech technology system. There are many applications of this system. The speaker segmentation algorithm is used to detect small segments from an audio file for the Automatic Speech Recognition to process [1]. It can be used as preprocessing module. It can be used before the speaker tracking, speaker identification, speaker verification etc. The speaker diarization system can be used in three application domain like Broadcast News (BN), meeting room data and telephone conversation [2]. Our target is to use it for the personal video recordings. The most of the speakers speak a very long time (e.g. more than 2 seconds) in the meeting and BN. But in a personal video recording, some speakers speak a few words like "YES" or "NO". It is very difficult to detect that speaker from the whole file. In a personal video, people sometimes tend to speak at the same time. So, recordings commonly include regions of overlapping speech. The success of the diarization depends on the choice of features space to represent speech [3]. These feature space directly affect model training process and Bayesian Information criterion (BIC) for the speaker change detection. The feature space should contain enough information to allow differentiating speakers. Mel Frequency Cepstrum Coefficients (MFCC) feature space is commonly used [4]. Models of different speakers cannot merge if the features inside segments are good to separate different speakers perfectly. In section 3, Linear Discriminant Analysis (LDA) has been used to discriminate among different classes [5]. LDA searches for a linear transformation. The feature clusters are the most distinct after that linear transformation. It can be achieved through scatter matrix analysis. The feedback from the user has been used to evaluate the performance of the system which is totally a new concept in the area of the speaker diarization system. It has been described in section 4. The user can select a segment from the output of the current system and give a feedback by saying "This is me". Our target is to improve the performance of the system by using the best user feedback. Section 2 contains the information of the overall speaker diarization system. Section 5 describes the experiment related to LDA and the user feedback. Section 6 shows the performance of the data.

II. Speaker Diarization System

The task of the speaker diarization system is to do segmentation of multiple speakers into speaker-homogenous parts. The system groups together all the segments that correspond to the same speaker. The number of speakers and speaker's characteristics are a priori unknown to the speaker diarization system. The speaker diarization system implemented is shown in Fig. 1 [6].

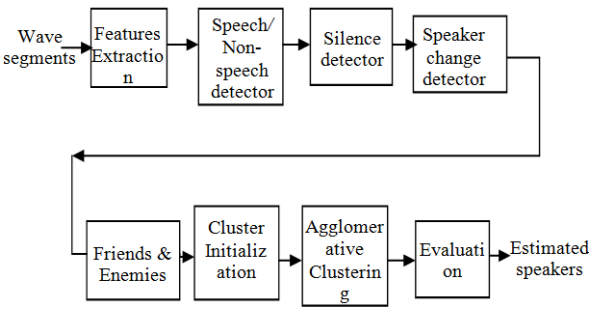

Figure 1. Speaker diarization system

Manuscript received August 12, 2013; revised December 20, 2013.

doi: 10.12720/ijsps.2.2.108-114

A. Features Extraction

The speaker diarization involves the parameterization of speech data into acoustic features such as MFCC [4]. It is computed with a 10ms frame rate. MFCC uses mel scale to address the sensitivity of the human ear. So, MFCCs approximate the human auditory system's response better than the normal cepstrum. The MFCC features have some advantages. The features dimension is approximately uncorrelated. Important and unique features which are extracted from each audio signal are taken at the speech and non-speech detector block.

B. Speech/Non-Speech

Features vector are analyzed in the speech and non-speech block [7]. Here, the frames which are sensed as speech and non-speech are denoted as speech label and non-speech label respectively. This block generates a theoretical threshold of the precision for the rest of the system. Because any error occurred in this block cannot be cured using further processing.

C. Silence Detector

In the silence detector block, a threshold of the frame energy is chosen to detect silence frames from speech frames which are taken from the speech/non-speech detector. A frame is detected as silence frame if the energy of a frame is less than a defined threshold value. Like speech/non-speech detector silence detector influences not only the quality of the clusters but also the number of the estimated speakers.

D. Speaker Change Detection

The speaker change detection algorithm identifies segments which contain probably only one event. If any error occurs in segmentation process, these errors degrade the performance of the overall system. The Bayesian Information Criterion (BIC) is used to identify possible speaker change points from the speech frames which are taken from the silence detector block [7]. ΔBIC is used to test whether two feature sets are represented by a single or two distinct models. Let us consider that we want to detect a possible change point at point t in Fig. 2.

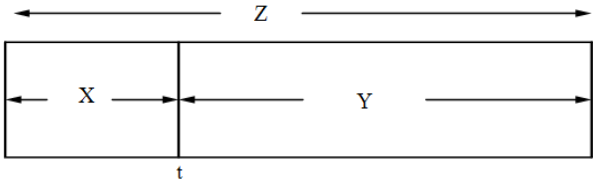

Figure 2. ΔBIC is used to detect whether two segments X and Y are from the same speaker or not

At first models are trained from each segment X and Y considering frames inside the segments X and Y. For acoustic modeling Hidden Markov Model (HMM) is used [3]. The BIC of each segment is computed using equation 1. The BIC of a model shows how good the data fit to its model. A combined model is trained using frames inside both segments.

$$BIC(\Theta_g) = \sum_i \log p(x_i, \Theta_g) - \lambda/2\, n \log N_g \qquad (1)$$

where,
x_i=features vector/ frame at instant i,
λ=a penalty parameter,
N_g=Number of features inside the model,
n=free parameters.

Now the ΔBIC is computed using equation 2.

$$\Delta BIC = BIC(\Theta_{xy}) - (BIC(\Theta_x) + BIC(\Theta_y)) \qquad (2)$$

where,
Θ_x represents the model which is trained from frames inside a test segment X,
Θ_y represents the model which is trained from frames inside another test segment Y,
Θ_{xy} represents the combined model which is trained from frames inside both segments (X and Y).

A possible speaker change point exists between these two segments if the ΔBIC is negative. The target of the speaker change detection block is to get pure segments which contain the information from the single speaker. Sometimes the algorithm ignores some genuine speaker change points or it causes some false alarm depending on the value of λ.

E. Friends and Enemies

The target of the Friends and Enemies block is to create clusters from segments which are taken from speaker change detection block [8]. The aim is that each cluster should contain the information of only one speaker. But the cluster can be mixed. One cluster can represent more than one speaker. More than one cluster can represent the same speaker. In this block, the most similar segments should be joined into one cluster to increase the cluster purity. According to Friends and Enemies algorithm, segments inside each cluster should be the most likely to each other and clusters are the most dissimilar to each other. The likeliness between the two segments is determined using equation 3.

$$\overline{xlkld}(S_1, S_2) = \frac{lkld(S_1/\Theta_{S2}) + lkld(S_2/\Theta_{S1})}{L_{S1} + L_{S2}} \qquad (3)$$

where,
$\overline{xlkld}$ =Normalized cross-likelihood between two segments, S_1 and S_2,
Θ_{s1}, Θ_{s2}=Models of segments S_1 and S_2 respectively,
L_{S1}, L_{S2}=Lengths of segments S_1 and S_2 respectively.

The segment should be the friend of any reference segment if it shows the highest $\overline{xlkld}$ with that reference segment considering other segments. If the segment shows the lowest $\overline{xlkld}$, then it is denoted as the enemy of the reference segment.

F. Cluster Initialization

In the cluster initialization block, a Viterbi decode is run to reassign the most probable frames from the file to each of the given clusters [6].

G. Agglomerative Clustering Block

The target of the agglomerative clustering block is to get final number of clusters which should be equal to the

real speakers [7]. All clusters belonging to the same speakers should be merged here. For the merging procedure and the cluster stopping criteria, BIC is used. The ΔBIC compares two possible models. The algorithm tries to find whether the two clusters belong to the same speaker or they belong to different speaker. There should be a stopping criterion to know when to stop the algorithm. Cluster pair is selected with the largest merge score (based on ΔBIC) that >0. If no such pair of clusters is found, the merging is stopped and the current clusters are used for further analysis. Here, we only consider blind clustering because there is no initial information at all.

H. Evaluation

Estimated clusters or speakers which are taken from the agglomerative clustering block are compared with the real speakers in the evaluation block. The Table I is used to define the performance of the system.

TABLE I. TABLE FOR EVALUATION PROCESS

	Estimated Speaker	Not Estimated Speaker	Non-Speech
Real Speaker	TP	FN	FNO_M
Not real Speaker	FP	-	-
Other	FPO_M	-	-

Recall, Precision and F-measure are used to evaluate the performance of the system using equation 4, 5 and 6 [9].

$$RECALL = \frac{TP}{TP + FN + FNO_M} \quad (4)$$

$$PRECISION = \frac{TP}{TP + FP + FPO_M} \quad (5)$$

$$F - MEASURE = \frac{2}{1/RECALL + 1/PRECISION} \quad (6)$$

III. FEATURES EXTRACTION USING LDA

Mel Frequency Cepstral Coefficients (MFCC) are the features used for the speaker diarization system. Speaker change points are detected by the ΔBIC computation of segments at the speaker change detection block. In the Friends and Enemies and the agglomerative clustering algorithm, models are trained based on frames or features inside each segment. Features play an important role to detect genuine speaker change points at the speaker change detection block and can affect the proper merging of the segment at the Friends and Enemies and agglomerative clustering block. If features inside segments are good to separate different speakers perfectly, then model of different speakers cannot merge which is expected to improve the performance. Linear Discriminant Analysis (LDA) searches those vectors in the underlying space which are used to discriminate among different classes. LDA yields the largest mean differences between the desired classes or speakers by creating a linear combination of independent features. The goal of LDA is to compute transformed features which are used to separate different speakers.

IV. USER FEEDBACK

The goal of the speaker diarization system (Fig. 3) is to identify the speaker from any video file. It also detects the number of estimated speakers. The number of estimated speakers can be more or less than the number of real speaker due to under-merging or over-merging of clusters. Our target is to use the feedback from the user to solve this merging problem. Positive acknowledgement by the user can be used as a user feedback. If any estimated labels which are found from the agglomerative clustering block are wrongly assigned by the system, the user can correct these labels by giving relevant feedbacks. The user can correct any part from the file which does not contain his information but the system wrongly assigned the user in that part.

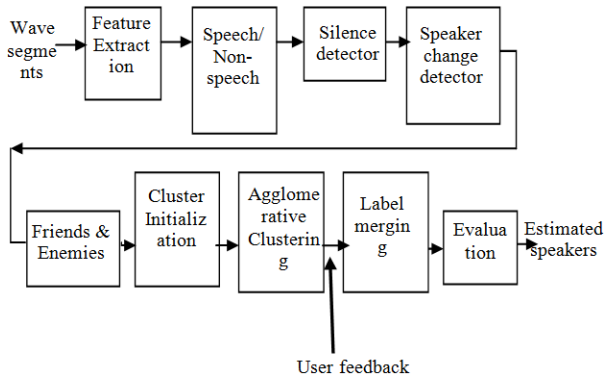

Figure 3. Speaker diarization system based on the positive acknowledgement

V. EXPERIMENTS

A. LDA

We can apply LDA in the speaker diarization system to transform our current features. At first different training classes should be created. Real labels of different speakers are taken. Then all segments of the same speaker are merged into one large segment. In this way, different segments or classes of different speakers are created. To get transformed features using LDA, the following steps have been done: [5]

Within-class scatter matrix, S_w is calculated using equation 7.

$$S_w = \sum_{j=1}^{C} \sum_{i=1}^{N_j} (X_i^j - \mu_j)(X_i^j - \mu_j)^T \quad (7)$$

where,

X_i^j : i^{th} feature of a class j,

μ_j: Mean of a class j,

C: Number of classes,

N_j: Number of features in j.

Within-class scatter matrix represents the characteristic of features within the same class. So, it is necessary to minimize the within-class scatter matrix.

The difference of features among separate classes is depicted by the between-class scatter matrix. So, between class scatter matrix should be maximized. The between-class scatter matrix, S_b is calculated using equation 8.

$$S_b = \sum_{j=1}^{C} (\mu_j - \mu)(\mu_j - \mu)^T \quad (8)$$

where μ: Mean of all classes.

The eigenvectors of the projection matrix is calculated using equation 9.

$$W = eig(S_w^{-1} S_b) \quad (9)$$

where W is the Transform matrix.

As our goal is to maximize the between-class scatter matrix and minimize the within-class scatter matrix, so rows of the transform matrix are obtained by choosing the M largest eigenvectors of $(S_w^{-1} S_b)$. The transformed features are obtained using equation 10.

$$Y = WX \quad (10)$$

where,
Y: Transformed features,
W: Transform matrix,
X: Original features.

Now, these transformed features are used in the speaker diarization system and the performance is expected to be increased.

B. *Positive Acknowledgement by the User*

Estimated speakers are found after using the agglomerative clustering algorithm. The goal of the agglomerative clustering algorithm is to obtain the same number of estimated clusters and estimated models as real speakers. Sometimes more than one estimated models represent one real speaker due to lack of merging of clusters which decrease the performance. Now, our target is to use the user feedback which can solve this merging problem. The user can assign a model as his model from estimated models. Then user can select another model from the rest of the estimated models as his model which is wrongly assigned by the system. Let us consider Fig. 4. The real speakers are denoted as SPK1, SPK2 and the estimated speakers are denoted as 1, 2, 3, 4. At first the user assigns a model as his model which is 1. Now the user can choose another estimated model which model number is not same as the assigned model but the user claims that it is his model. Here, this feedback model is 2. So, the user can give a feedback by pointing two different models and say "This is me and that is also me." It means that the user selects two different models which should be the same model but the system makes them separated due to under-merging. The speaker diarization system using the user feedback is shown in Fig. 3. These feedback models are taken at the label merging block (Fig. 3).

Here, the labels of these two models are merged into the first user assigned model's label. Then the performance of the whole system is evaluated using this user feedback.

We have taken feedback automatically from the estimated label considering the real label to evaluate the performance of the system. The necessary steps that we have followed to take feedbacks are described below.

(i) At first the user assigns a model from all estimated speaker's model. The user can assign a model considering a large segment (Fig. 4) or more than one small segment (Fig. 5). If the user assigns that model which has enough information of him, then the assignment will be almost correct and is expected to improve the performance of the system. Each real label is considered to do the assignment of the model automatically. The ratio of the number of frames of each estimated model that belongs to each real label with respect to the number of frames inside that real label is calculated. If the ratio is more than 40%, then this model is assigned by that real label as the main model. So, the model 1 in Fig. 4 and models 1 and 4 in Fig. 5 can be assigned according to the above 40% rule. If the real label assigns more than one model, then the label of these models are merged into one label such that all of them represent one main model as shown in Fig. 5.

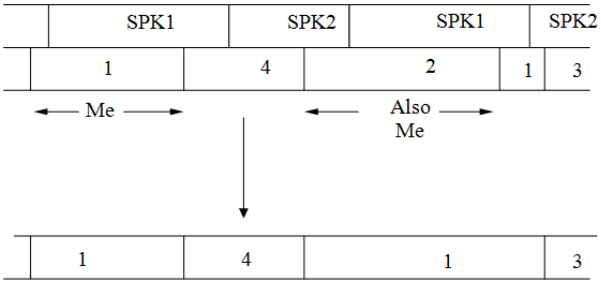

Figure 4. User feedback based on the positive acknowledgement

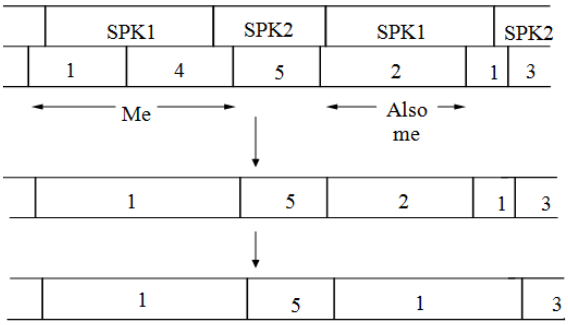

Figure 5. Assignment of small segments based on the positive acknowledgement by the user

(ii) The real label which assigns a model in the previous step is searched from the whole file. If the real label is found, then the ratio of the number of frames of each estimated model except the main model that belongs to that real label with respect to the number of frames inside that real label is calculated. If any model covers 90% of that real label, then it can take as a feedback model. Then the label of this feedback model is merged into the main model and the whole performance is evaluated using this information. Analysis has been done by taking 50% or 0% coverage of the feedback model with respect to the real label. We are interested to know the performance of the system, if the feedback model does not contain enough information of the user.

So, the assignment of a model and searching feedback models are performed for each real label. F-measure, precision and recall are computed for each feedback model. Then the average performance is calculated considering all feedbacks for each file.

VI. Results

A. LDA

Linear Discriminant Analysis (LDA) discriminates between different classes by creating independent feature vectors for each class. Here, LDA is applied to MFCC and we are interested to know whether it improves the performance or not. The performance of the speaker diarization system (baseline) is shown in Table II.

TABLE II. Performance of Speaker Diarization System

File name	F-measure (%)	Precision (%)	Recall (%)
Video01$_1$	22.57	13.98	58.47
Video01$_2$	34.71	24.63	58.76
Video01$_3$	42.38	44.91	40.12
Video02$_1$	34.52	41.65	29.48
Video02$_3$	26.71	20.62	37.90
Video03$_1$	44.18	33.52	64.81
Video03$_2$	56.19	46.92	70.03
Video03$_3$	25.32	24.4	26.32
Video 04$_1$	25.14	31.43	20.94
Video04$_3$	63.05	64.18	61.95

TABLE III. Performance Based on Transformed Features (Full Transform Matrix)

File name	F-measure (%)	Precision (%)	Recall (%)
Video01$_1$	19.44	12.00	51.04
Video01$_2$	40.46	42.23	38.83
Video01$_3$	48.30	36.08	73.04
Video02$_1$	46.47	46.02	46.93
Video02$_3$	27.25	22.66	34.16
Video03$_1$	45.58	35.99	62.15
Video03$_2$	47.99	46.89	49.15
Video03$_3$	28.78	27.88	29.75
Video 04$_1$	26.66	30.32	23.80
Video04$_3$	70.05	68.23	71.98

TABLE IV. Performance Based on Transformed Features (24×15 Transform Matrix)

File name	F-measure (%)	Precision (%)	Recall (%)
Video01$_1$	18.29	10.85	58.15
Video01$_2$	36.37	24.93	67.21
Video01$_3$	57.65	49.29	69.42
Video02$_1$	33.95	22.74	66.94
Video02$_3$	21.82	16.03	34.14
Video03$_1$	46.85	33.05	80.46
Video03$_2$	41.00	39.53	42.59
Video03$_3$	19.51	16.25	24.40
Video 04$_1$	25.85	18.11	45.13
Video04$_3$	72.80	66.63	80.23

TABLE V. Performance Based on Transformed Features (24×10 Transform Matrix)

File name	F-measure (%)	Precision (%)	Recall (%)
Video01$_1$	15.18	8.89	52.02
Video01$_2$	39.69	29.02	52.73
Video01$_3$	54.89	46.88	66.20
Video02$_1$	36.75	24.91	69.99
Video02$_3$	18.10	10.38	70.87
Video03$_1$	44.52	30.55	82.08
Video03$_2$	41.10	32.73	55.22
Video03$_3$	22.84	14.65	51.75
Video 04$_1$	28.53	22.26	39.72
Video04$_3$	56.56	57.64	55.52

If the system uses the full transform matrix which is discussed in section 5, then recall, precision and F-measure increase. Because now transformed features can distinguish between different speaker's segment and models trained from these segments are separated in the best way. It increases the performance because it prevents over-merging or under-merging of the segment. The performance of the speaker diarization system using full transform matrix is shown in Table III. Different dimensions of the transform matrix like 24×15 and 24×10 are used to get transformed features. Here, we select the highest eigenvector. The dimensionality reduction can reduce the irrelevant noise and can increase computation performance. The performance of the speaker diarization system using different dimension is shown in Table IV and Table V.

If the dimension is decreased, precision decreases and recall increases which is shown in Table VI. If the system uses few dimensions of feature vectors, few data are available for model training process which makes models too general. As models become too general, they may perform over-merging which decrease the performance like precision.

TABLE VI. Average Performance Based on Transformed Features

Concept	F-measure (%)	Precision (%)	Recall (%)
Base line	37.48	34.62	46.88
Full transform matrix	40.1	36.83	48.09
24×15 transform matrix	37.41	29.74	56.87
24×10 transform matrix	35.82	27.79	60.61

TABLE VII. Performance Based on the Positive Acknowledgement by the User (Model Cover 50% of the Real Label)

File name	F-measure (%)	Precision (%)	Recall (%)	Feedback #
Video01$_1$	22.57	13.98	58.47	0
Video01$_2$	33.51	21.39	77.28	10
Video01$_3$	47.16	41.85	54.23	259
Video02$_1$	36.67	37.20	36.41	185
Video02$_3$	26.08	19.03	41.85	13
Video03$_1$	40.61	27.91	74.49	6
Video03$_2$	37.92	25.06	77.94	6
Video03$_3$	21.92	16.68	31.95	2
Video 04$_1$	26.71	29.54	24.57	36
Video04$_3$	65.80	59.33	73.97	1218

TABLE VIII. Performance Based on the Positive Acknowledgement by the User (Model Cover 90% of the Real Label)

File name	F-measure (%)	Precision (%)	Recall (%)	Feedback #
Video01$_1$	22.57	13.98	58.47	0
Video01$_2$	33.52	21.40	77.28	3
Video01$_3$	47.25	42.00	54.29	128
Video02$_1$	36.51	37.18	36.04	70
Video02$_3$	25.72	18.44	42.83	6
Video03$_1$	40.6	27.91	74.49	1
Video03$_2$	37.92	25.05	77.94	3
Video03$_3$	21.91	16.68	31.95	1
Video 04$_1$	26.54	29.08	24.73	31
Video04$_3$	65.22	58.97	73.04	331

TABLE IX. PERFORMANCE BASED ON THE POSITIVE ACKNOWLEDGEMENT BY THE USER (TAKING ALL MODELS AS FEEDBACK)

File name	F-measure (%)	Precision (%)	Recall (%)	Feedback #
$Video01_1$	18.71	10.80	70.06	5
$Video01_2$	33.51	21.39	77.28	14
$Video01_3$	45.68	40.78	52.11	447
$Video02_1$	35.86	37.36	34.75	418
$Video02_3$	25.80	18.92	41.23	54
$Video03_1$	42.90	30.10	74.94	15
$Video03_2$	37.92	25.06	77.94	6
$Video03_3$	21.92	16.68	31.95	2
$Video04_1$	25.65	28.09	23.77	336
$Video04_3$	65.72	59.34	73.76	2139

B. *Positive Acknowledgement by the User*

In this feedback method, the user assigns a model as his model and selects a feedback model which is wrongly assigned as another user by the system. Then the label of these two models is merged and the performance of the system is evaluated. The result is shown in Table VII, Table VIII and Table IX.

If the feedback is used, recall increases and precision decreases in all files. It is expected because labels of two different models which were separated by the system are merged into one label. The precision decreases due to over-merging of the label. The overall performance or F-measure increases in some files. We find that F-measure increases in files $video01_3$, $video02_1$, $video04_1$ and $video04_3$. These video files have more feedbacks than the other files.

If the user gives that type of feedbacks where the model contain very less information about him as shown in Fig. 6, then this type of feedback decreases the performance of the system. Here the user assigns model 1 as his model. If the user chooses model 2 as the feedback model which contains more information about another user, then it degrades the performance.

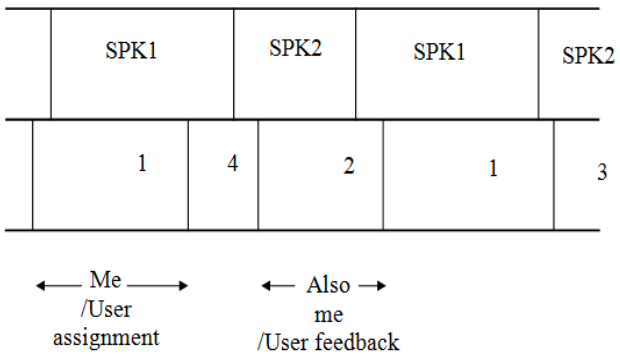

Figure 6. Feedback based on the positive acknowledgement by the user where feedback models contain little information by the user.

Now the average performance is compared between the baseline and the user feedback considering those four files which have enough feedbacks to analyze the performance. And this result is shown in Table X. Analysis has been done considering three different cases. They are given below:

- Case A: The Feedback model covers more than 90% of the real label
- Case B: The Feedback model covers more than 50% of the real label
- Case C: The Feedback model covers more than 0% of the real label or taking all models as the feedback.

The average recall and average F-measure increase in all cases (A, B, C) than the baseline due to proper merging of clusters. As the feedback model contains enough information of the user in case A and case B, it prevents wrong merging of labels and increases the performance compared to case C.

TABLE X. AVERAGE PERFORMANCE BASED ON THE POSITIVE ACKNOWLEDGEMENT BY THE USER

Concept	F-measure (%)	Precision (%)	Recall (%)
Base line	41.27	45.54	38.12
Case A	43.88	41.81	47.03
Case B	44.09	41.98	47.30
Case C	43.23	41.39	46.1

VII. CONCLUSION

Features help to emphasize the characteristics of the speech. It is important for discriminating between relevant speech sounds. It also suppresses unwanted variations and noise. If features vectors extracted from each frame contain unique characteristics to describe the individual speaker, then it is expected to improve the performance. In this paper, LDA has been used to create the transformed features which are able to maximize the distance between different classes. It improves the performance of the speaker diarization system. In this paper, user feedback has been used to improve the performance of the current speaker diarization system. This is totally a new idea which has been developed here. The user is expected to give a feedback by pointing a segment and claims the segment as his segment. The label of the estimated speaker is merged according to the positive acknowledgement by the user and the performance of the speaker diarization system has been improved.

Several issues remain to be analyzed or investigated to improve the overall efficiency of the system. There are several possibilities to take feedback from the user except our proposed technique. The user can detect some segments from the output of the speaker change detection block as his or her segments. All the user detected segments can be joined together to make a large segments which is good for the model training process as it contains more data to represent single speaker. Now this large model can be enough fit to select similar segments using the Friends and Enemies algorithm. The user can select more than one segment from the output of the Friends and Enemies block as his or her segments. Then segments should be merged at the Agglomerative clustering block to make a large cluster. This large cluster is good to represent single speaker's model and can cause perfect merging with other clusters using the agglomerative clustering algorithm which is expected to improve the performance of the speaker diarization system.

ACKNOWLEDGMENT

The author wishes to thank Dr. Marko Lugger and Dr. Wilhelm Hagg for helping with new ideas and suggestions.

REFERENCES

[1] J. Pylkkönen, "Towards efficient and robust automatic speech recognition: Decoding techniques and discriminative training," Doctoral diss., School of Science, Department of Information and Computer Science, Aalto University, Mar. 2013.

[2] X. A. Miro, S. Bozonnet, *et al*., "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356-364, Feb. 2012.

[3] M. Zelenák, "Detection and handling of overlapping speech for speaker diarization," Doctoral diss., TALP Research Center, Department of Signal Theory and Communications, Universitat Politécnica de Catalunya, Barcelona, Spain, Oct. 2011.

[4] Z. Xiongy, R. Radhakrishnan, A. Divakaran, and T. S. Huang, "Comparing MFCC and MPEG-7 audio features for feature extraction, maximum likelihood HMM and entropic prior HMM for sports audio classification," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Dec. 2003.

[5] Önsen Toygar and Adnan Acan, "Face recognition using PCA, LDA and ICA approaches on colored images," *Istanbul University-Journal of Electrical and Electronics Engineering*, vol. 3, no. 1, pp. 735-738, 2003.

[6] X. Anguera, C. Wooters, and J. M. Pardo, "Robust speaker diarization for meetings: ICSI RT06S meetings evaluation system," in *Machine Learning for Multimodal Interaction*, Springer Berlin Heidelberg, 2006, pp. 346-358.

[7] Xavier Anguera Miró, "Robust speaker diarization for meetings," PhD thesis, Speech Processing Group, Department of Signal Theory and Communications, Universitat Politécnica de Catalunya, Barcelona, Oct. 2006.

[8] X. Anguera, C. Wooters, and J. Hernando, "Friends and enemies: A novel initialization for speaker diarization," in *Proc. Interspeech - ICSLP*, 2006.

[9] J. Ajmera, I. McCowan, and H. Bourlard, "Robast speaker change detection," *IEEE Signal Processing Letters*, vol. 11, no. 8, pp. 651, Aug. 2004.

Zeenat Afroze was born in Dhaka, Bangladesh on 1981. She completed B.Sc on Electrical and Electronic Engineering from Bangladesh University of Engineering and Technology (BUET), Dhaka, Bangladesh in 2004. She also completed M.Sc on Communication Engineering and Media Technology from Universität Stuttgart, Stuttgart, Germany in 2008.

She worked as a lecturer in the department of Electrical and Electronic Engineering at American International University Bangladesh, Dhaka, Bangladesh during September 2004-February 2006. Currently, she is working as an assistant professor in American International University Bangladesh. Her research interest includes Communication and Digital Signal Processing (Digital Communication and Digital System).

