Noname manuscript No.
(will be inserted by the editor)

A tutorial on evaluation metrics for speaker diarization systems

Supratim Tribady · Shefali Waldekar · A Kishore Kumar · Goutam Saha · Md Sahidullah ·

Received: DD Month YEAR / Accepted: DD Month YEAR

Abstract In this article, we present a comprehensive review of the evaluation metrics for the SD systems. We demonstrate how they calculate the evaluation metrics from the ground truth and the system-generated output. Here, different errors are considered in each evaluation metrics, such as speaker error, false alarm, and missed speech with the help of ground truth and system output. We explain the importance of different error terms for computing these evaluation metrics for SD with the help of case studies. The limitations of different evaluation metrics are briefly explained in this article. Finally, we discuss the formulation of new or different evaluation metric for evaluation of SD systems.

Keywords Diarization error rate (DER) · Jaccard error rate (JER) · Rich transcription time marked (RTTM) · Speaker diarization (SD) · Speaker indexing · Un-partitioned evaluation map (UEM)

Supratim Tribady
Department of Electronics & Electrical Communication Engineering, IIT Kharagpur
E-mail: supratimtribedy96@gmail.com

Shefali Waldekar
Department of Electronics & Electrical Communication Engineering, IIT Kharagpur
E-mail: shefaliw@ece.iitkgp.ernet.in

A Kishore Kumar
Department of Electronics & Electrical Communication Engineering, IIT Kharagpur
E-mail: kishore@iitkgp.ac.in

Goutam Saha
Department of Electronics & Electrical Communication Engineering, IIT Kharagpur
E-mail: gsaha@ece.iitkgp.ac.in

Md Sahidullah
Université de Lorraine, CNRS, Inria, LORIA, F-54000, Nancy, France
E-mail: md.sahidullah@inria.fr

1 Introduction

Speaker diarization (SD) (also known as *speaker indexing* (Wilcox and Kimber, 1997)) aims to solve the problem “Who spoke When” for a given speech signal (Anguera et al., 2012). It mainly involves dividing a speech signal into segments followed by grouping of the homogeneous segments based on speaker similarity indexing. SD has many practical applications, such as automatic video captioning (Song et al., 2018), automatic transcript generation for spoken conversations (Bentley et al., 2018), smart speaker technology (Bentley et al., 2018), etc. In the present period with a growing number of broadcasting and online meeting, SD could play a key role in creating transcripts for *content summarization* and *sentiment analysis* in natural language processing application (Tiwary and Siddiqui, 2008). Most of the studies conducted in SD research focus on three kinds of audio-data: (i) *broadcast news audio* where speech data are usually collected from radio and TV programs containing commercial breaks and music (Wachob, 1992), (ii) *meeting audio* where multiple people are involved in a conversation (Mieczakowski et al.), and (iii) audio-data from *telephone conversation* (Elvins et al., 2003). However, studies on SD are also conducted with DIHARD corpora which consists of a wide variety of audio-data collected from a number of real-world conditions (Ryant et al., 2018).

The main challenge in SD system arises due to different practical problems, which mainly includes domain mismatch because of different acoustic environment (Himawan et al., 2018), incorrect detection of speakers in multi-speaker speech recognition from unsegmented recordings (Watanabe et al., 2020), and improper evaluation of the metrics during overlapping of speakers in a conversation (Vipperla et al., 2012). The system should be strong enough to deal with multiple speakers during overlapping.

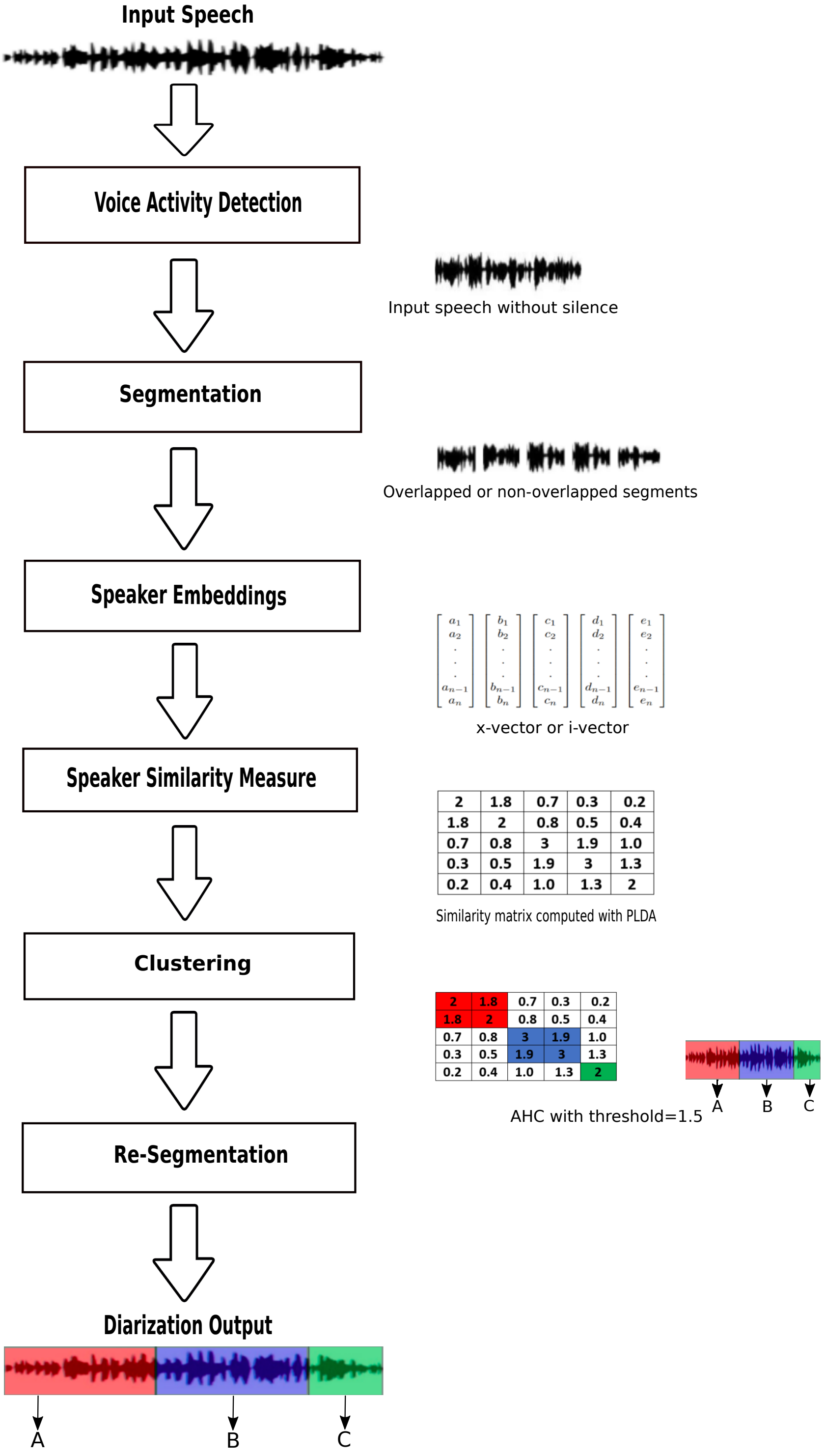
The main aim of this work is to review the evaluation metrics for SD system. Evaluation metrics play a very important role in determining the best system, based on the various shortcomings of the diarization process. The selection of an evaluation metric decides the system performance in different adversarial conditions. An important aspect of the evaluation metric is the capability to distinguish among various systems. Several metrics are used to check the performance of the speech processing systems, like for automatic speaker verification *Equal error rate (EER)* is used (Jyh-Min Cheng and Hsiao-Chuan Wang, 2004), for acoustic scene classification the standard is *accuracy* (Valenti et al., 2016), *F1 score* is used for sound event detection (Kong et al., 2019), metrics such as *Unweighted average recall (UAR)* is used for emotion recognition evaluation (Gamage et al., 2017), and *min t-DCF* is used for detecting spoofing countermeasures (Kinnunen et al., 2020),Word error rate (WER) is the primary evaluation metric for automatic speech recognition (Galibert, 2013), etc. These are some evaluation metrics used for checking the performance of the systems for the respective domain. Similarly, *Diarization error rate* and *Jaccard error rate* are the two widely used evaluation metric, used to check the performance of the SD system. Diarization error rate, remains the principal evaluation metric in this area which was introduced by *National Institute of Standards and Technology (NIST)* in the *Rich Transcriptions (RT)* evaluations[1] in the year 2000. Jaccard error rate, a metric introduced for *Second DIHARD Diarization Challenge, 2019*[2][3] that is based on the *Jaccard index* (Ryant et al., 2019).

In this article, we review different metrics used for the evaluation of SD system. We mainly analyse two widely used evaluation metric known as DER and JER, for synthetically prepared ground-truth and predicted output. We discuss the limitations of the currently used evaluation metrics and briefly discuss how to develop a new reliable metric for the evaluation of SD systems.

The rest of the paper is organized as follows. In Section 2, we present a brief overview of state-of-the-art SD system. In Section 3, we present the case studies with two evaluation metrics mainly DER and JER along with some other clustering metrics. In Section 4, we prepared synthetic data in the form of reference ground truth and system predicted labels and calculated the DER and JER. And lastly, in Section 5, and Section 6 we will discuss the limitations of DER and JER, and also give overview regarding development of new evaluation metrics.

2 An overview of the state-of-the-art speaker diarization system

SD is one of the different ways of processing done on audio signals (Anguera et al., 2012). A SD system usually consists of several components. The first important component is a *voice activity detector* (VAD) (Moattar and Homayounpour, 2009), which separates the speech segments from the non-speech segments in an audio-data. Then it applies a *speech based segmentation* technique to split the speech regions into different small segments (Tritschler and Gopinath, 1999). After segmentation *speaker embeddings*[4] are extracted. The state-of-the-art speaker diarization systems rely on speaker embeddings (Cyrta et al., 2017) for speaker similarity measure (Sell and Garcia-Romero, 2014). In the following step, it uses a *clustering* technique for clustering the segments into disjoint speakers. Finally, *re-segmentation* is used for further frame-level refinement of speaker diarization output (Sell and Garcia-Romero, 2015). Fig 1 illustrates the different components of the SD system.


Input Speech
Voice Activity Detection
Input speech without silence
Segmentation
Overlapped or non-overlapped segments
Speaker Embeddings
a_1 a_2 a_{n-1} a_n
b_1 b_2 b_{n-1} b_n
c_1 c_2 c_{n-1} c_n
d_1 d_2 d_{n-1} d_n
e_1 e_2 e_{n-1} e_n
x-vector or i-vector
Speaker Similarity Measure
Similarity matrix computed with PLDA
Clustering
AHC with threshold=1.5
A
B
C
Re-Segmentation
Diarization Output
A
B
C

2	1.8	0.7	0.3	0.2
1.8	2	0.8	0.5	0.4
0.7	0.8	3	1.9	1.0
0.3	0.5	1.9	3	1.3
0.2	0.4	1.0	1.3	2

2	1.8	0.7	0.3	0.2
1.8	2	0.8	0.5	0.4
0.7	0.8	3	1.9	1.0
0.3	0.5	1.9	3	1.3
0.2	0.4	1.0	1.3	2

Fig. 1 This figure tells about the standard SD structure with multiple modules. A raw audio recording of a conversation is given as an input, after that speech part of the signal is extracted or separated from the non-speech part of the audio signal. The speech part of the audio signal is segmented into small segments from which speaker embeddings (i-vector or x-vector) are extracted. The speaker similarity is measured and computed from the speaker embeddings and finally, based on the similarity measure the speaker embeddings are clustered using Agglomerative Hierarchical clustering with a threshold of 1.5 which assigns similar speaker segments to a global speaker ID. After clustering, again it is re-segmented and finally a timeline showing diarization output audio is found.

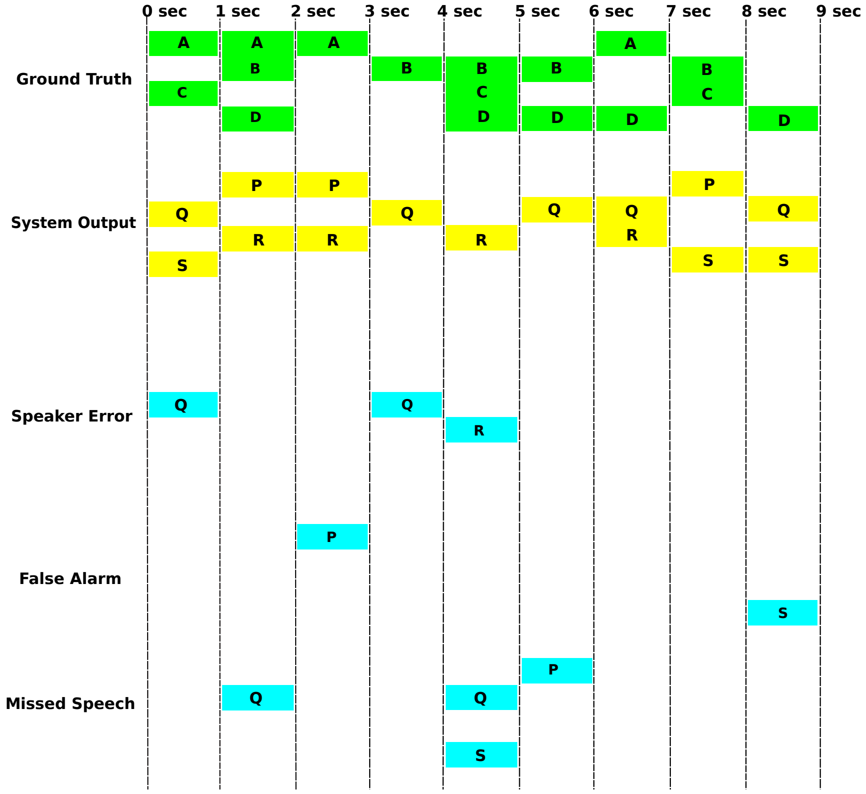


Fig. 2 Synthetic ground truth and system predicted labels to illustrate different types of error. Here green speaker labels indicate the reference ground truth , yellow color speaker labels indicates system predicted speaker labels and blue color speaker labels indicates speaker error, false alarm and missed speech

3 Evaluation metrics for speaker diarization

DER and JER, are used to measure the performance of a SD system. For speech regions, the diarization system specifies the locations of speaker labels to each homogeneous segment of speech. DER and JER provides a convenient way to compare different diarization approaches. The difference or error generated by a system in diarizing a speech, that is by comparing the error from

[1] http://www.xavieranguera.com/phdthesis/node147.html#NIS_rt_eval_plan_2006

[2] https://signalprocessingsociety.org/publications-resources/data-challenges/second-dihard-speech-diarization-challenge

[3] DIHARD II is the second in a series of diarization challenges focusing on ”hard” diarization; that is, SD for challenging recordings where there is an expectation that the current state-of-the-art will fare poorly https://coml.lscp.ens.fr/dihard/index.html

[4] Speaker embeddings are representation of speech segments created with deep neural network.

the ground truth (reference truth) Rich Transcription Time Marked (RTTM), and system predicted or system-generated RTTM using an Unpartioned Format Evaluation (UEM) file, which is used to specify the scoring within each recording. So, the motivation will be to decrease the DER and JER, in case of output RTTM which will help to improve the SD system and to match the relative speaker labels and location of speaker boundaries from reference RTTM compared to output predicted RTTM. The evaluation metrics are generated with the help of some files such as reference RTTM, system RTTM and UEM files. The DER of the system can be over 100 %, whereas the JER of the system cannot exceed over 100% (Anguera et al., 2005). The RTTM files are space-delimited text files containing one turn per line, each line containing ten fields whereas UEM files are used to specify the scoring regions within each audio recording. The UEM file contains a line with four space-delimited fields for each scoring region. For speech regions, the diarization system specifies the locations of speaker labels, to each homogeneous segment of speech. Computation of an error rate requires describing what are the errors present. The various types of error in the evaluation metrics of the SD system are:
Speaker error: Speaker error corresponds to the percentage of scored time that a reference speaker is assigned to a wrong speaker in the output reference speaker labels. Speaker error is mainly a diarization system error. Speaker error is assigned within a speech region, and it does not account for speaker errors in overlapping regions, or any other error coming from non-speech frames.
False alarm speech: False alarm speech corresponds to the percentage of a scored time, that a non-speech part is incorrectly labelled as a speech region in system-generated output.
Missed speech: Missed speech corresponds to the percentage of a scored time, that a speech part is incorrectly labelled as a non-speech part in system-generated output.

3.1 Diarization error rate

It is the most commonly used metric in the SD system. To compute DER, an optimal one-to-one mapping of reference speakers to system output speakers is determined. The DER is then the sum of the per speakers false alarm time, miss time and speaker error time that is not matched to the reference speaker divided by total speech time in an audio file. It is measured as the fraction of time that is not attributed correctly to a speaker or non-speech.

$$\text{DER} = \frac{\text{ERROR} + \text{FA} + \text{MISS}}{\text{TOTAL}} \tag{1}$$

Here TOTAL refers to the duration of the union of reference and system speaker segments and if the reference speaker was not paired with a system speaker, it is the duration of all reference speaker segments.

In Fig. 2, a synthetic speaker label has been generated to show the different errors generated in a SD system. In order to check the speaker mapping between the reference speaker and system speaker output, DER uses Hungarian algorithm and Weighted-Bipartite graph matching algorithm. Using Hungarian algorithm and Weighted-Bipartite graph matching we have found the reference speaker A is mapped with system speaker R, reference speaker B is mapped with system speaker P, reference speaker C is mapped with system speaker S and reference speaker D is mapped with system speaker Q. In Table 1, different types of error that are generated in the synthetic speaker labels are shown in Fig.2.

Time Frame	Reference Speaker	System Output	Error
0-1 second	A,C	Q,S	1 Speaker Error
1-2 second	A,B,D	P,R	1 Missed Speech
2-3 second	A	P,R	1 False Alarm
3-4 second	B	Q	1 Speaker Error
4-5 second	B,C,D	R	1 Speaker Error, 2 Missed Speech
5-6 second	B,D	Q	1 Missed Speech
6-7 second	A,D	Q,R	No Error
7-8 second	B,C	P,S	No Error
8-9 second	D	Q,S	1 False Alarm

Table 1 Demonstration of different types of error present in different time frames for the synthetically prepared data in Fig.2

3.2 Goodman-Kruskal tau (GKT)

GKT (Zarghami et al., 2009) is an unbalanced measure which was discovered by Goodman and Kruskal in 1954. For a reference speaker label 'ref' and a system speaker label 'sys', GKT(ref, sys) correlates to the fraction of change in sys that can be explained by ref. Therefore, GKT(sys,ref) is 1 when ref is exactly predictive compared to sys and is 0 when ref is not predictive compared to sys in system-output.

3.3 Conditional entropy

Another evaluation metric, which reports four information theoretic measures.

- H(X—Y) : conditional entropy in bits of the reference speaker label when system speaker label is present.
- H(Y—X) : conditional entropy in bits of the system speaker label when reference speaker label is present.
- MI : mutual information in bits between reference and system speaker labels.

- NMI : normalized mutual information between the reference and system speaker labels.

Here, X refers to the sequence of true frame-wise speaker labels whereas Y refers to the sequence of hypothesized speaker labels. NMI is basically derived from MI after being normalized in the interval between 0 to 1.

3.4 Purity, coverage and clustering metrics

Apart from DER and JER, purity (Cettolo, 2000) and coverage (Gauvain et al., 1998) also provide a convenient way to compare between systems of different diarization approaches. It is usually not sufficient to understand the type of error executed by the system. To understand the type of error performed by the system, purity and coverage play a key role to judge the behaviour of the system. A fourth approach or evaluation metrics to check the performance of the system uses both the reference and system output labels. Each recording is converted to a sequence of 10 msec out of which is a single speaker label is assigned to the following cases:

- frame containing no speech
- frame containing speech from a single speaker
- frame containing overlapping speech

B-cubed precision, recall, and F1 :The B-cubed precision for a single frame assigned speaker S in the reference diarization and C in the system diarization is the proportion of frames assigned C that are also assigned S. Similarly, the B-cubed recall for a frame is the proportion of all frames assigned S that are also assigned C. The overall precision and recall, then, are just the mean of the frame-level precision and recall measures and the overall F-1 their harmonic mean.

3.5 Speaker error rate

When speech or non-speech segments do not play an important role in the experiment, then the standard Speaker error rate (SER) comes into play, which does not include speech or non-speech errors. Speaker error rate (SER) corresponds to the amount of scored time when a reference speaker in the ground truth is mapped to a wrong speaker in the output speaker labels (Aronowitz, 2010). Speaker error rate (SER) is only assigned for speech regions, and it does not account for speaker errors in the non-speech part. For the evaluation of two-speaker segmentation task, Speaker error rate (SER) is computed according to the standard NIST protocol [5].

[5] http://www.itl.nist.gov/iad/mig/tests/sre/2002/SpkrSegEval-v07.pl

3.6 Jaccard error rate

In addition to the principal metric, the JER is based on the Jaccard Index, which is a similarity measure used to evaluate the output of speaker segmentation. JER was newly introduced in the second DIHARD Diarization Challenge as another evaluation metric along with DER (Ryant et al., 2019). The JER calculates the missed speech and false alarm speech for each individual speaker. An optimal mapping between reference speakers and system output speakers is determined. The Jaccard index is computed, for each such speaker pairs. The JER is defined as 1 minus the average of these speaker pair scores.

More specifically, “N“ reference speakers and “M“ system speakers are assumed from the ground truth and system predicted output. An optimal mapping between speakers is determined using the Hungarian algorithm 7 (Jonker and Volgenant, 1986) so that each reference speaker is paired with at most one system speaker and each system speaker with at most one reference speaker (Bell and Dee, 2016). Then, for each reference speaker “ref“ the speaker-specific Jaccard error rate is “(FA + MISS)/TOTAL“, where “TOTAL“ denotes the duration of the union of reference and system speaker segments; if the reference speaker was not paired with a system speaker, it is the duration of all reference speaker segments - “FA“ is the total system speaker time not attributed to the reference speaker; if the reference speaker was not paired with a system speaker, it is 0 - “MISS“ is the total reference speaker time not attributed to the system speaker; if the reference speaker was not paired with a system speaker, it is equal to “TOTAL“.The Jaccard error rate is the average of the speaker-specific Jaccard error rate.

JER and DER are highly correlated with the JER typically being higher, especially in recordings where one or more speakers is particularly dominant. When a i^{th} speaker from reference output corresponds to the j^{th} speaker in

$$\mathrm{JER_i} = \frac{\mathrm{FA_i} + \mathrm{Miss_i}}{\mathrm{Union\ of\ ref_i} + \mathrm{system_j}} . \tag{2}$$

$$\mathrm{Overall_{JER}} = \frac{1}{\mathrm{N}} \sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{JER_i} \tag{3}$$

Here N refers to number of speakers present in the conversation.

4 Examples demonstrating the computation of evaluation metrics

In this section, we demonstrate with examples how evaluation metrics are computed from the ground-truth and system predicted output. For better understanding of the computation process, we show each intermediate steps. We considered five different examples as summarized in Table. All the speech recordings are nine seconds in length.

	$\{\#\mathrm{ref}_{\mathrm{spk}}, \#\mathrm{sys}_{\mathrm{spk}}\}$
Example 1	{1, 1}
Example 2	{2, 2}
Example 3	{4, 4}
Example 4	{4, 3}
Example 5	{3, 4}

Table 2 Summary of the five examples for the computation of DER and JER. Here $\mathrm{ref}_{\mathrm{spk}}$ denotes the number speakers in reference (or ground-truth) and $\mathrm{sys}_{\mathrm{spk}}$ denotes the number of speakers in system output.

4.1 Example 1

In this example (as shown in Fig. 3), we show how the DER and JER are computed for single speaker in both ground-truth and system predicted output.

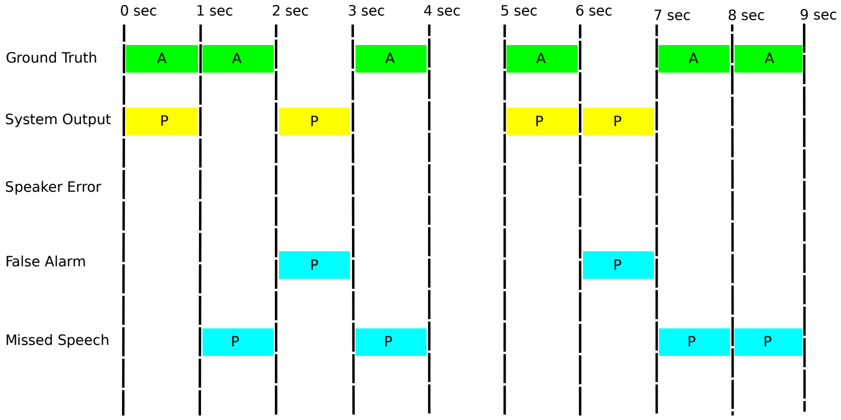


Fig. 3 Synthetic ground truth and system predicted labels and illustration different types of error for **Example 1**. The green boxes indicate the reference ground truth, yellow boxes indicate system predicted speaker labels and cyan boxes indicates speaker error, false alarm and missed speech.

Example 1

In order to compute the DER, we first need to compute the three basic errors: speaker error, false alarm and missed speech as shown in Eq. 1. In this case, we have no speaker error as the single speaker in ground-truth (*i.e.*, Speaker A) is paired with the single speaker in predicted output (*i.e.*, Speaker P). We observe two seconds of false alarm due and four seconds missed speech as shown in Fig. 3. In this case, the total amount of speech for ground-truth speaker is six seconds. Therefore, the DER for Example 1 will be,

$$\mathrm{DER}_{\mathrm{ex1}} = \frac{0+2+4}{6} \times 100\% = 100\%.$$

In JER computation, first speaker correspondence between each of the reference speakers and system output is computed with Hungarian algorithm. Then we compute individual JERs for each reference speakers as shown in Eq. 3. Finally, overall JER is computed by taking average of the individual JERs. In this example, we have single speaker in both reference and system output. Therefore, the overall JER is computed as,

$$\mathrm{JER}_{\mathrm{ex1}} = \mathrm{JER}_{\mathrm{A}} = \frac{\mathrm{FA}_{\mathrm{A}} + \mathrm{Miss}_{\mathrm{A}}}{\cup(\mathrm{A},\mathrm{P})} = \frac{2+4}{8} \times 100\% = 75\%$$

4.2 Example 2

In Fig. 4, we demonstrate the evaluation metric computation for two speakers in both ground-truth and system predicted output in a recording of 9 seconds.

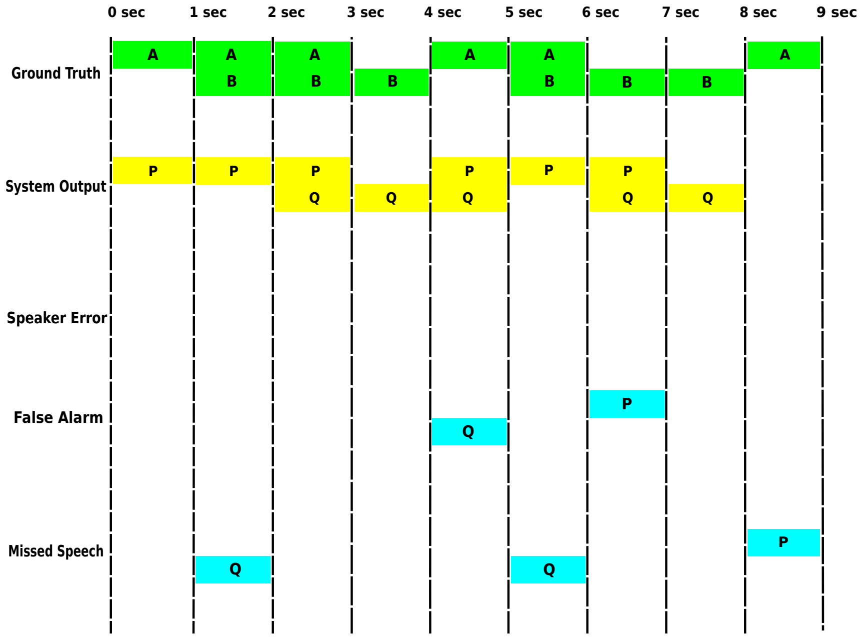


Fig. 4 Synthetic ground truth and system predicted labels and illustration different types of error for **Example 2**. The green boxes indicate the reference ground truth, yellow boxes indicate system predicted speaker labels and cyan boxes indicates speaker error, false alarm and missed speech.

Example 2

For Example 2, we observe two seconds of false alarm, two seconds of missed speech, and no speaker error as shown in Fig. 4. We also compute the total amount of speech spoken by two speakers in reference is 12 seconds. Therefore, we can compute DER as,

$$\mathrm{DER}_{\mathrm{ex2}} = \frac{0+2+3}{12} \times 100\% = 41.66\%.$$

Now to compute the JER, we first need to find the speaker correspondence. Using Hungarian algorithm, we have found that reference speaker A pairs with system predicted Speaker P and reference speaker B pairs with system predicted Speaker Q. Then, we can compute the JERs of individual referene speakers as,

$$\mathrm{JER}_{\mathrm{A}} = \frac{\mathrm{FA}_{\mathrm{A}} + \mathrm{Miss}_{\mathrm{A}}}{\cup(\mathrm{A}, \mathrm{P})} = \frac{1+1}{7} \times 100\% = 28.57\%.$$

$$\mathrm{JER}_{\mathrm{B}} = \frac{\mathrm{FA}_{\mathrm{B}} + \mathrm{Miss}_{\mathrm{B}}}{\cup(\mathrm{B}, \mathrm{Q})} = \frac{2+1}{7} \times 100\% = 42.86\%.$$

Therefore, the overall JER will be,

$$\mathrm{JER}_{\mathrm{ex2}} = \frac{1}{2}\Big[28.57 + 42.86\Big] \times 100\% = 35.71\%.$$

4.3 Example 3

In Fig. 5, we demonstrate the evaluation metric computation for four speakers in both ground-truth and system predicted output in a recording of 9 seconds.

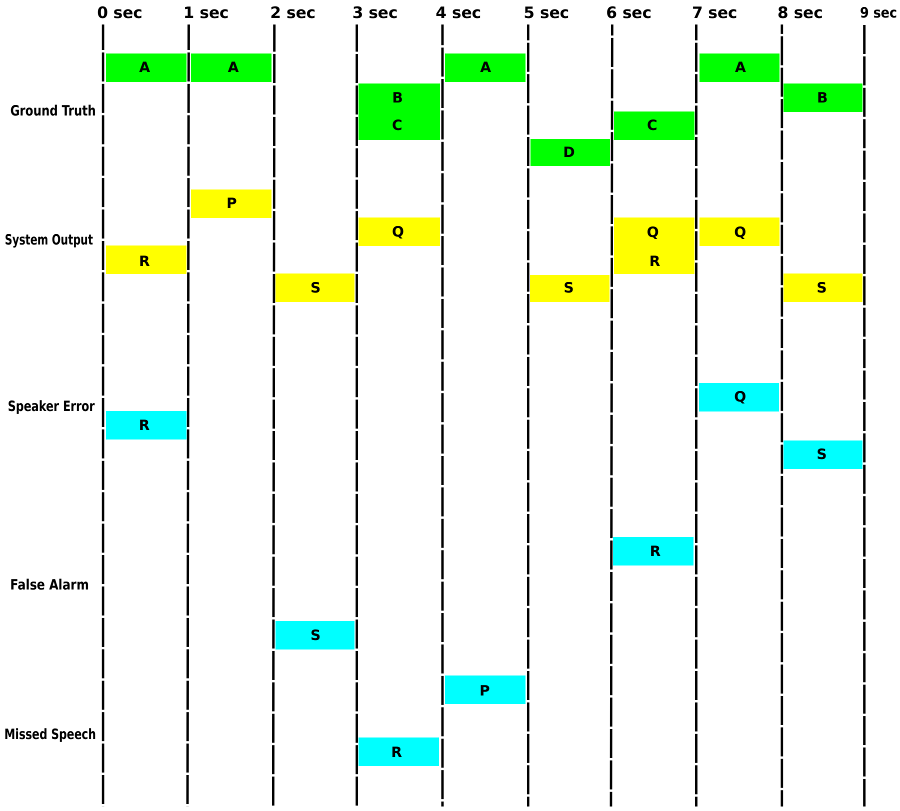


Fig. 5 Synthetic ground truth and system predicted labels and illustration different types of error for **Example 3**. The green boxes indicate the reference ground truth, yellow boxes indicate system predicted speaker labels and cyan boxes indicates speaker error, false alarm and missed speech.

Example 3

For Example 3, we observe two seconds of false alarm, two seconds of missed speech, and three seconds of speaker error as shown in Fig. 5. We also compute the total amount of speech spoken by four speakers in reference is nine seconds. Therefore, we can compute DER as,

$$\mathrm{DER_{ex3}} = \frac{3+2+2}{9} \times 100\% = 77.77\%.$$

Now to compute the JER, we first need to find the speaker correspondence. Using Hungarian algorithm, we have found that reference speaker A pairs with system predicted Speaker P, reference speaker B pairs with system predicted Speaker R, reference speaker C pairs with system predicted Speaker Q and reference speaker D pairs with system predicted Speaker S. Then, we can compute the JERs of individual reference speakers as,

$$\mathrm{JER_A} = \frac{\mathrm{FA_A} + \mathrm{Miss_A}}{\cup(\mathrm{A},\mathrm{P})} = \frac{3+0}{4} = \frac{3}{4} \times 100\% = 75.00\%.$$

$$\mathrm{JER_B} = \frac{\mathrm{FA_B} + \mathrm{Miss_B}}{\cup(\mathrm{B},\mathrm{Q})} = \frac{2+2}{4} = \frac{4}{4} \times 100\% = 100.00\%.$$

$$\mathrm{JER_C} = \frac{\mathrm{FA_C} + \mathrm{Miss_C}}{\cup(\mathrm{C},\mathrm{R})} = \frac{0+1}{3} = \frac{1}{3} \times 100\% = 33.33\%.$$

$$\mathrm{JER_D} = \frac{\mathrm{FA_D} + \mathrm{Miss_D}}{\cup(\mathrm{D},\mathrm{S})} = \frac{0+2}{3} = \frac{2}{3} \times 100\% = 66.66\%.$$

Therefore, the overall JER will be,

$$\mathrm{JER_{ex3}} = \frac{1}{\mathrm{N}}[\mathrm{JER_A} + \mathrm{JER_B} + \mathrm{JER_C} + \mathrm{JER_D}]. \tag{4}$$

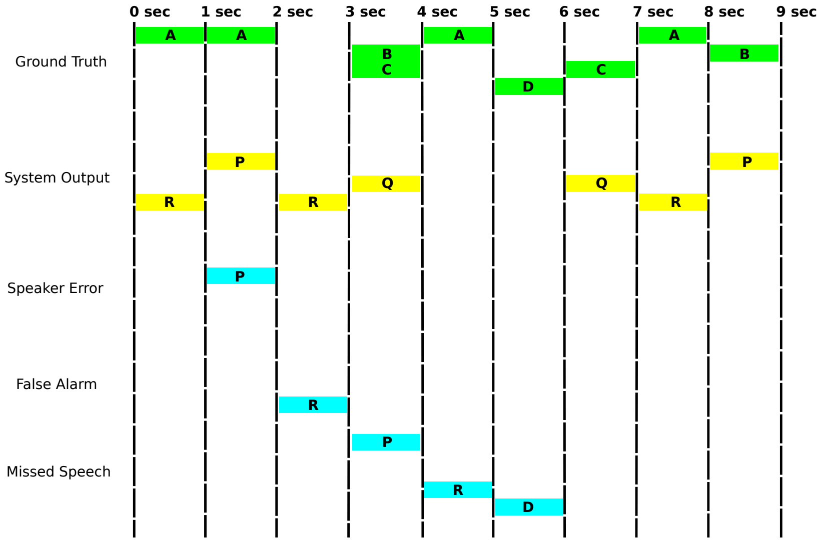


Fig. 6 Synthetic ground truth and system predicted labels and illustration different types of error for **Example 4**. The green boxes indicate the reference ground truth, yellow boxes indicate system predicted speaker labels and cyan boxes indicates speaker error, false alarm and missed speech

4.4 Example 4:

In Fig. 6, we demonstrate the evaluation metric computation for four speakers in both ground-truth and system predicted output in a recording of 9 seconds.

Example 4

In Example 4, we observe one second of false alarm, three seconds of missed speech and one second of speaker error as shown in Fig. 6. We also compute the total amount of speech spoken by four speakers in reference is 9 seconds. Therefore, we can compute the DER as,

$$\mathrm{DER_{ex4}} = \frac{3+1+1}{9} \times 100\% = 55.56\%.$$

Now to compute the JERs, we first need to find the speaker correspondence. Using Hungarian algorithm, we have found that reference speaker A pairs with system predicted speaker R, reference speaker B pairs with system predicted speaker P, reference speaker C pairs with system predicted speaker Q. Then, we can compute the JERs of individual reference speakers as,

$$\mathrm{JER_A} = \frac{\mathrm{FA_A} + \mathrm{Miss_A}}{\cup(\mathrm{A},\mathrm{Q})} = \frac{2+1}{5} = \frac{3}{5} \times 100\% = 60.00\%.$$

$$\mathrm{JER_B} = \frac{\mathrm{FA_B} + \mathrm{Miss_B}}{\cup(\mathrm{B},\mathrm{R})} = \frac{1+1}{3} = \frac{2}{3} \times 100\% = 66.667\%.$$

$$\mathrm{JER_C} = \frac{\mathrm{FA_C} + \mathrm{Miss_C}}{\cup(\mathrm{C},\mathrm{P})} = \frac{0+0}{5} = \frac{0}{5} \times 100\% = 0.00\%.$$

$$\mathrm{JER_D} = \frac{\mathrm{FA_D} + \mathrm{Miss_D}}{\cup(\mathrm{D},\mathrm{D})} = \frac{0+1}{1} = \frac{1}{1} \times 100\% = 100.00\%.$$

So, the overall JER will be,

$$\mathrm{JER_{ex4}} = \frac{1}{\mathrm{N}}[\mathrm{JER_A} + \mathrm{JER_B} + \mathrm{JER_C} + \mathrm{JER_D}]. \tag{5}$$

$$\mathrm{JER_{ex4}} = \frac{1}{4}[60.00 + 66.667 + 0.00 + 100.00] \times 100\% = 56.67\%..$$

4.5 Example 5:

In Fig. 7, we demonstrate the evaluation metric computation for four speakers in both ground-truth and system predicted output in a recording of 8 seconds.

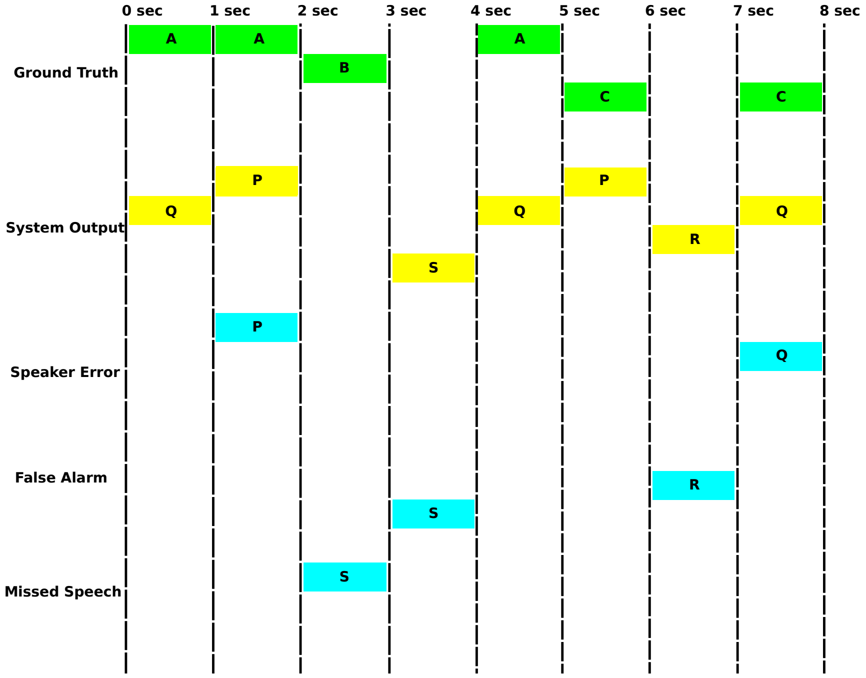


Fig. 7 Synthetic ground truth and system predicted labels and illustration different types of error for **Example 5**. The green boxes indicate the reference ground truth, yellow boxes indicate system predicted speaker labels and cyan boxes indicates speaker error, false alarm and missed speech

Example 5

In Example 5, we observe one second of false alarm, two seconds of missed speech and two seconds of speaker error as shown in Fig. 6. We also compute the total amount of speech spoken by four speakers in reference is 6 seconds. Therefore, we can compute the DER as,

$$\mathrm{DER_{ex5}} = \frac{2+2+1}{6} \times 100\% = 83.33\%.$$

Now to compute the JERs, we first need to find the speaker correspondence. Using Hungarian algorithm, we have found that reference speaker A pairs with system predicted speaker Q, reference speaker B pairs with system predicted speaker S and reference speaker C pairs with system predicted speaker P. Then, we can compute the JERs of individual reference speakers as,

$$\mathrm{JER_A} = \frac{\mathrm{FA_A} + \mathrm{Miss_A}}{\cup(\mathrm{A},\mathrm{R})} = \frac{1+1}{4} = \frac{2}{4} \times 100\% = 50.00\%.$$

$$\mathrm{JER_B} = \frac{\mathrm{FA_B} + \mathrm{Miss_B}}{\cup(\mathrm{B},\mathrm{Q})} = \frac{1+3}{4} = \frac{4}{4} \times 100\% = 100.00\%.$$

$$\mathrm{JER_C} = \frac{\mathrm{FA_C} + \mathrm{Miss_C}}{\cup(\mathrm{C},\mathrm{P})} = \frac{1+1}{3} = \frac{2}{3} \times 100\% = 67.66\%.$$

Here, Speaker D is only present in the system generated speaker labels. So, it will not be considered for calculation of JER in case of speaker D. So, the overall JER will be,

$$\mathrm{JER_{ex5}} = \frac{1}{\mathrm{N}}[\mathrm{JER_A} + \mathrm{JER_B} + \mathrm{JER_C}]. \tag{6}$$

$$\mathrm{JER_{ex5}} = \frac{1}{4}[50.00 + 100.00 + 67.66] \times 100\% = 72.22\%..$$

5 Limitations of the existing evaluation metrics

DER and JER one of the current existing evaluation metrics in the field of SD. There are various drawbacks of the existing evaluation metrics. According to the formulae of the DER, the denominator part ”TOTAL” is the duration of all reference speaker segments, but it does not include the system-speaker segments for calculation of the error rate. If there is a data imbalance of two or more speakers in terms of duration of the active speaker in a conversation in reference speaker level, then irrespective of the system-generated speaker levels it will produce a good result which in turn will give less DER, which is not correct will respect to the ground scenario. DER has no upper limit, as it can exceed 100%. After that JER was introduced in the DIHARD II challenge, 2019 which also has some drawbacks. The drawback for upperlimit in DER is solved in the JER, as it cannot exceed more than 100%. The JER is used to calculate the error rate from the weighted average of each individual speaker present in a conversation. It performs speaker correspondence using the Hungarian algorithm. But from our experiments, we see that these speaker correspondence does not reflect actual speaker mapping for the calculation of JER during the overlapping of more than two active speakers. DER and JER gives the overall error rate of an audio file, but it does not provide the error rate of each segment-wise speaker boundaries of each speaker, which might be helpful to analyse and reduce the overall DER and JER of the entire audio file. Systems which do not consider overlapping will always acknowledge the considerable amount of error. Ignoring the overlappings decreases overall Jaccard error rate, but it does not portray the actual scenario of the number of speakers present in the conversation, and the actual identity of the speaker error, false alarm and missed speech. In case of synthetically prepared data from Fig. 5, there is an overlapping of speaker B and speaker C and in the system output only speaker Q is present, but the system considered it as speaker R instead of speaker Q. So, to get the minimum JER, the system is considering speaker R instead of speaker Q. The main challenge of implementing the metric is establishing the mapping between reference speaker ID and system-generated speaker ID. These are the major limitations of the DER and JER.

6 Proposal for new evaluation metrics

One of the new findings from our experiments is the importance of using a correct evaluation metric for SD system. The existing evaluation metrics or the evaluation tools are reaching the limits under certain conditions, so there is a need to build and generalize new metric for evaluation and rebuild it based on their application to make them usable under the new challenges and conditions along with making it comparable with the previous results. Though, it is very difficult to define a precise point in time boundaries about when a speaker starts or stop, especially when overlapping speech is present, it is better to build a new state-of-the-art evaluation metric which will detect a proper error during the speaker overlappings. A new evaluation metrics should be developed to give segment-wise errors between speaker labels of reference ground truth and system-generated speaker labels thereby properly detecting speaker errors. The evaluation metric should also detect errors during speaker overlappings, such as speaker error, missed speech and false alarm and denoting it for the respective speakers. The new evaluation metrics should calculate the DER and JER with respect to the system generated speaker labels irrespective of calculating it concerning to generate optimum DER and JER.

7 Conclusions

Following on from the previous study, we draw different conclusions on the evaluation metrics for SD system. Due to the increased used of online meetings, and smart speakers, SD has become very important. The DER and JER is still a relevant evaluation metric used to measure the standard of a diarization system, with much more composite setup including:

- Cross-show diarization, where re-occurring speakers in multiple shows have to be acknowledged.
- Speaker overlappings, where multiple speakers speak concurrently.

More eminently, we described the implementation method along with the algorithms to ensure a better understanding of the evaluation metric. These evaluation metrics serve as an important parameter for checking the overall performance of the system.

Appendix A: Speaker correspondence

Hungarian Algorithm: The Hungarian method is a combinational optimization algorithm that solves the speaker correspondence assignment issue. In this section, we explore the working of the Hungarian algorithm which is used to compute the best matches between ground-truth speaker sequence and system output sequence of speakers. Here, we will demonstrate the calculation of JER for speaker correspondence with the help of Hungarian algorithm for **Example 3** in the Fig. 5. The JER is calculated for all possible cases or conditions. In our case, we calculate for all possible combinations thereby mapping from one speaker in reference speaker sequence to another speaker in system-speaker sequence. After calculating the JER, we put the values in the matrix corresponding to the speaker levels and then we go for the calculation of the optimal value of JER.

In **Example 3**, there are four speakers in the ground-truth (*i.e.*, Speaker A, Speaker B, Speaker C and Speaker D) and four system speaker output (*i.e.* Speaker P, Speaker Q, Speaker R, and Speaker S). So, using the Hungarian algorithmwe will explain the speaker correspondence in case of JER calculation. The matrix below shows the cost of assigning a speaker from reference

Reference speaker sequence
Speaker A
Speaker B
Speaker C
Speaker D

System speaker sequence
Speaker P
Speaker Q
Speaker R
Speaker S

speaker level to a speaker in the system speaker level. The main objective is to minimize the total JER in the system.

Speaker	P	Q	R	S
A	0.750	0.833	0.800	1.000
B	1.000	0.750	1.000	0.750
C	1.000	0.330	0.660	1.000
D	1.000	1.000	1.000	0.660

Step 1: Substraction of row minima from each row.

Speaker	P	Q	R	S
A	0.000	0.083	0.050	0.250
B	0.250	0.000	0.250	0.000
C	0.670	0.000	0.330	0.670
D	0.340	0.340	0.340	0.000

Step 2: Substraction of column minima from each column

Speaker	P	Q	R	S
A	0.000	0.083	0.000	0.250
B	0.250	0.000	0.2000	0.000
C	0.670	0.000	0.280	0.670
D	0.340	0.340	0.290	0.000

Step 3: Covering all the zero rows and columns with minimum number of lines.

Speaker	P	Q	R	S
A	0.000	0.083	0.000	0.250
B	0.250	0.000	0.200	0.000
C	0.670	0.000	0.280	0.670
D	0.340	0.340	0.290	0.000

Step 4: Creating additional zeros in the matrix. For example, we find the smallest number from all uncovered rows and columns and substract it from all uncovered elements and add it to all elements that are covered by boxes twice.

Speaker	P	Q	R	S
A	0.000	0.283	0.000	0.450
B	0.050	0.000	0.000	0.000
C	0.470	0.000	0.080	0.670
D	0.140	0.340	0.090	0.000

Now in order to cover all the minimum number of zero rows and columns , we return to step 3.

Step 3: Again, covering all the rows and columns with minimum number of zeros.

Speaker	P	Q	R	S
A	0.000	0.283	0.000	0.480
B	0.050	0.000	0.000	0.000
C	0.470	0.000	0.080	0.670
D	0.140	0.340	0.090	0.000

Now in order to cover all the minimum number of zero rows and columns , we return to step 3.

Step 5: Therefore, the zeros in each row shows optimal assignment.

Speaker	P	Q	R	S
A	0.000	0.283	0.000	0.450
B	0.050	0.000	0.000	0.000
C	0.470	0.000	0.080	0.670
D	0.140	0.340	0.090	0.000

Step 6: Now, corresponding to the optimal matrix for cost function.

Speaker	P	Q	R	S
A	0.750	0.833	0.800	1.000
B	1.000	0.750	1.000	0.750
C	1.000	0.330	0.660	1.000
D	1.000	1.000	1.000	0.660

Hence, the optimal Jaccard error rate value will be:

$$\mathrm{JER}_{\mathrm{min}} = \frac{1}{4}[0.750 + 1.000 + 0.330 + 0.660] \times 100\% = 68.73\%.$$

(7)

Hence, using Hungarian algorithm we found the speaker correspondence between reference speaker sequence and system-speaker sequence.

Speaker A →Speaker P
Speaker B →Speaker R
Speaker C →Speaker Q
Speaker D →Speaker S

References

X. Anguera, C. Woofers, J. Hernando, Speaker diarization for multi-party meetings using acoustic fusion, in *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, IEEE, 2005, pp. 426–431. IEEE

X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, O. Vinyals, Speaker diarization: A review of recent research. IEEE Transactions on Audio, Speech, and Language Processing **20**(2), 356–370 (2012)

H. Aronowitz, Unsupervised Compensation of Intra-Session Intra-Speaker Variability for Speaker Diarization., in *Odyssey*, 2010, p. 25

J. Bell, H.M. Dee, The subset-matched jaccard index for evaluation of segmentation for plant images. arXiv preprint arXiv:1611.06880 (2016)

F. Bentley, C. Luvogt, M. Silverman, R. Wirasinghe, B. White, D. Lottridge, Understanding the long-term use of smart speaker assistants. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **2**(3), 1–24 (2018)

M. Cettolo, Segmentation, classification and clustering of an Italian broadcast news corpus, in *Proc. of RIAO*, Citeseer, 2000. Citeseer

P. Cyrta, T. Trzciński, W. Stokowiec, Speaker diarization using deep recurrent convolutional neural networks for speaker embeddings, in *International Conference on Information Systems Architecture and Technology*, Springer, 2017, pp. 107–117. Springer

T.T. Elvins, R.T. Fassett, P. Shinn, *System and method for gathering, personalized rendering, and secure telephonic transmission of audio data* (Google Patents, 2003). US Patent 6,529,586

O. Galibert, Methodologies for the evaluation of speaker diarization and automatic speech recognition in the presence of overlapping speech., in *INTERSPEECH*, 2013, pp. 1131–1134

K.W. Gamage, V. Sethu, E. Ambikairajah, Salience based lexical features for emotion recognition, in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2017, pp. 5830–5834. IEEE

J.-L. Gauvain, L.F. Lamel, G. Adda, Partitioning and transcription of broadcast news data, in *Fifth International Conference on Spoken Language Processing*, 1998

I. Himawan, M.H. Rahman, S. Sridharan, C. Fookes, A. Kanagasundaram, Investigating deep neural networks for speaker diarization in the dihard challenge, in *2018 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2018, pp. 1029–1035. IEEE

R. Jonker, T. Volgenant, Improving the hungarian assignment algorithm. Operations Research Letters **5**(4), 171–175 (1986)

Jyh-Min Cheng, Hsiao-Chuan Wang, A method of estimating the equal error rate for automatic speaker verification, in *2004 International Symposium on Chinese Spoken Language Processing*, 2004, pp. 285–288

T. Kinnunen, H. Delgado, N. Evans, K.A. Lee, V. Vestman, A. Nautsch, M. Todisco, X. Wang, M. Sahidullah, J. Yamagishi, et al., Tandem assessment of spoofing countermeasures and automatic speaker verification: Fundamentals. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2020)

Q. Kong, Y. Xu, I. Sobieraj, W. Wang, M.D. Plumbley, Sound event detection and time–frequency segmentation from weakly labelled data. IEEE/ACM Transactions on Audio, Speech, and Language Processing **27**(4), 777–787 (2019)

A. Mieczakowski, J. Goodman-Deane, J. Patmore, J. Clarkson, Conversations, conferencing and collaboration

M.H. Moattar, M.M. Homayounpour, A simple but efficient real-time voice activity detection algorithm, in *2009 17th European Signal Processing Conference*, IEEE, 2009, pp. 2549–2553. IEEE

N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, M. Liberman, First dihard challenge evaluation plan. 2018, tech. Rep. (2018)

N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, M. Liberman, The second dihard diarization challenge: Dataset, task, and baselines. arXiv preprint arXiv:1906.07839 (2019)

G. Sell, D. Garcia-Romero, Speaker diarization with PLDA i-vector scoring and unsupervised calibration, in *2014 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, 2014, pp. 413–417. IEEE

G. Sell, D. Garcia-Romero, Diarization resegmentation in the factor analysis subspace, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2015, pp. 4794–4798. IEEE

J. Song, Y. Guo, L. Gao, X. Li, A. Hanjalic, H.T. Shen, From deterministic to generative: Multimodal stochastic rnns for video captioning. IEEE transactions on neural networks and learning systems **30**(10), 3047–3058 (2018)

U. Tiwary, T. Siddiqui, *Natural language processing and information retrieval* (Oxford University Press, Inc., ???, 2008)

A. Tritschler, R.A. Gopinath, Improved speaker segmentation and segments clustering using the bayesian information criterion, in *Sixth European Conference on Speech Communication and Technology*, 1999

M. Valenti, A. Diment, G. Parascandolo, S. Squartini, T. Virtanen, DCASE 2016 acoustic scene classification using convolutional neural networks, in *Proc. Workshop Detection Classif. Acoust. Scenes Events*, 2016, pp. 95–99

R. Vipperla, J.T. Geiger, S. Bozonnet, D. Wang, N. Evans, B. Schuller, G. Rigoll, Speech overlap detection and attribution using convolutive non-negative sparse coding, in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2012, pp. 4181–4184. IEEE

D.E. Wachob, *Method and apparatus for providing demographically targeted television commercials* (Google Patents, 1992). US Patent 5,155,591

S. Watanabe, M. Mandel, J. Barker, E. Vincent, Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings. arXiv preprint arXiv:2004.09249 (2020)

L.D. Wilcox, D.G. Kimber, *Unsupervised speaker clustering for automatic speaker indexing of recorded audio data* (Google Patents, 1997). US Patent 5,659,662

A. Zarghami, S. Fazeli, N. Dokoohaki, M. Matskin, Social trust-aware recommendation system: A t-index approach, in *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, vol. 3, IEEE, 2009, pp. 85–90. IEEE