International Journal of Security and Its Applications
Vol. 10, No. 5 (2016) pp.57-62
http://dx.doi.org/10.14257/ijsia.2016.10.5.05

Use of Silence as an Altered Approach for Speaker Recognition

Rupali Pawar[1], R. M. Jalnekar[2] and J. S. Chitode[3]

[1]Research Scholar, Vishwakarma Institute of Technology, Pune
[2]Director & Professor, Vishwakarma Institute of Technology, Pune
[3]Professor, E &TC Department, Vishwakarma Institute of Technology, Pune
[1]rvspawar@rediffmail.com, [2]director@vit.edu, [3]j.chitode@gmail.coml

Abstract

Speaker recognition is an important application of Speech Signal Processing and has been used in public safety, authenticating users for important financial transactions, access control systems and many more. The conventional approach in speaker identification and speaker verification has been to remove the silence from the recorded speech signal and further extract the significant features from the residual signal for recognition. This paper presents an alternative approach and puts forth the experimental results of obtaining silence as a parameter to check if the pattern of pauses/silence for train and test files recorded for individual speaker match. The paper emphasizes the approach in which a paragraph is recorded for 8 speakers and used as train files. The duration of silence/pauses of the speaker in a paragraph are obtained. This silence obtained is compared with the silence obtained from test file the matching of pattern of the silences decides the identity of the speaker.

Keywords*: Framing, intra speaker variability, inter speaker variability, normalization, threshold*

1. Introduction

The Speaker recognition system has four phases Analysis, Feature Extraction, Recognition and Testing. In the speaker Recognition system the database used for processing can be standard or recorded. For a robust Speaker Recognition system the database should be free from inter speaker and intra speaker variability, as these are crucial parameters and their presence may affect the performance of the system. The intra speaker variation is due to variable speaking rate, changing emotions or other mental variables, and also due to environment noise, emotional state of the speaker at the time of recording *etc*. [1]. The inter speaker variability is variability in the speech signal due to individual variations in the speech of the speaker. This can be due to variations in excitation source, length of the vocal cord and movement of articulators. The conventional Speaker Recognition System extracts features like MFCC, pitch, duration as features for identification after the pre-emphasis stage during which the speech signal is made noise and silence free. This paper puts forth an approach where silence is obtained from the speech signal of a speaker and saved in the train folder which is further analyzed to check if the pattern of silence for the speaker remains the same any number of times it is recorded.

2. Implementation

To implement the above approach database of 8 speakers (7 Females, 1 male) is used. A paragraph of approximately 50 sec is recorded in a .wav file format and stored in the train folder. The same paragraph is again recorded for the above 8 speakers and stored in the test folder. The paragraph is divided into three parts as Initial paragraph, Intermediate

ISSN: 1738-9976 IJSIA

paragraph and End paragraph of around 16 sec each. The purpose or motive behind this is to capture the naturalness of the speaker. It is assumed that maximum spontaneity might be captured better in the intermediate part as a speaker might be conscious while recording the initial few lines and may be in a rush to complete the paragraph towards the end.

The recorded speech signals is framed in chunk of 20mS to 30mS and s analyzed to obtain silence in each frame. The overall speech is analyzed to demark silence, a threshold is fixed, the part of speech signal lying between this threshold values is marked as silence and the duration of this silence is captured in number of samples. Figure 1 depicts the speech signal of speaker 1, Intermediate paragraph and silence obtained from the same after applying threshold for test and train files respectively.

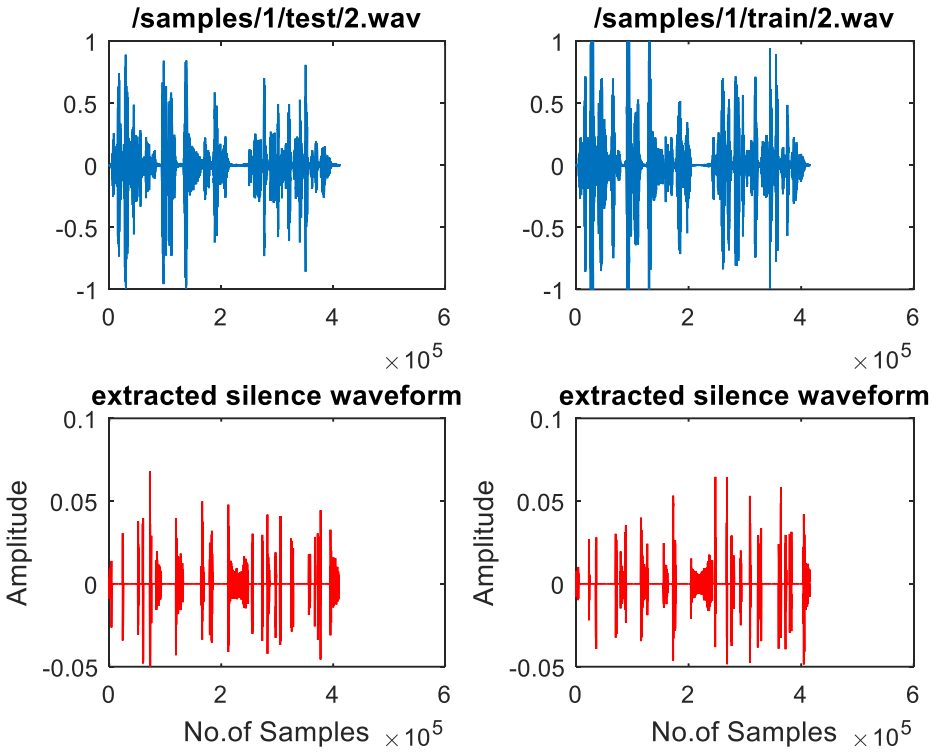

Figure 1. Speech Signal (blue) and Silence Extracted (red) for Test and Train Files of Intermediate Paragraph of Speaker 1

2.1. Algorithm

1. Record the data at different instances and save as train and test file
2. Frame the speech signal into chunks of 10-30mS
3. To find the duration of pauses/silence occurring at various instances in the sentence, compare the amplitude of speech signal with positive and negative threshold value. (chosen by permutation combination for the recorded speech)
4. Normalize the silence duration noted.
5. Correlate the normalized data to find the match in the pattern of silence for each Speaker

3. Experimentation

As shown in Figure 1 the speech signal is recorded and it is observed that the sentence in each paragraph has certain pauses. Instead of removing this silence, the

speech signal is framed, and the silence duration which is in between the threshold values is obtained and noted. The duration of silence obtained for each instance is recorded for both the test and train file of every speaker for each initial, intermediate and end paragraph as shown in the Table 1 below. The duration of silence/pauses obtained has large variation in the value and hence the data is normalized. The normalized values of train and test data are compared to find the similarity in the two signals thereby allowing us to compare whether the pauses are of similar pattern for a speaker. Certain observations are put forth after plotting graph to compare the normalized train and test data for respective paragraph recorded for all the speakers, as depicted in Figure 2a & 2b

1. The pattern may shift to certain extent during different recording sessions.
2. The pattern of silence for intermediate paragraph show better similarity compared to Initial and End paragraph.

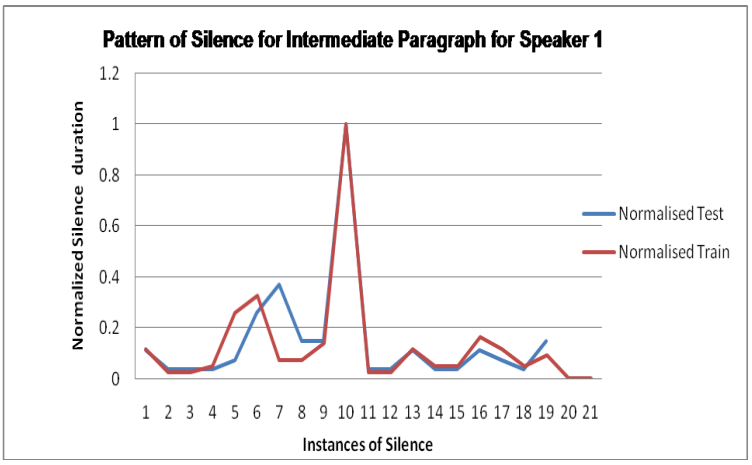

Figure 2a: Pattern of Silence for Intermediate Paragraph for Speaker 1 Normalised Values of Silence Duration (y-axis) Vs Instances of Silence (x-Axis)

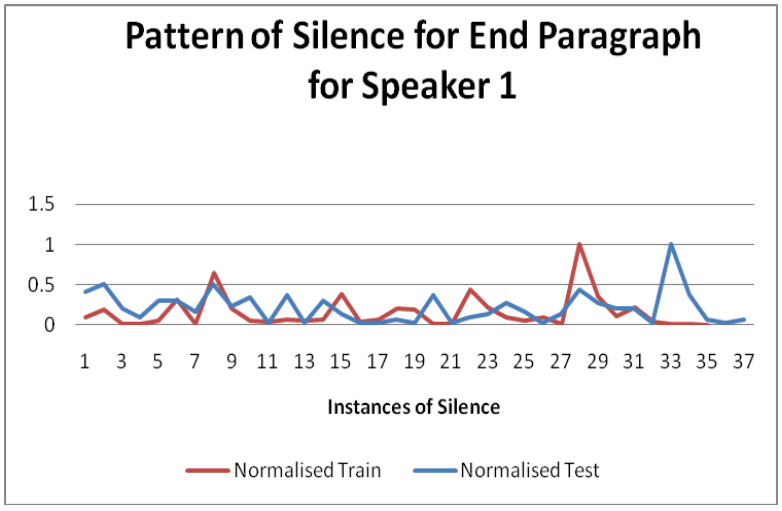

Figure 2b: Pattern of Silence for End Paragraph for Speaker 1 Normalised values of silence duration (y-axis) Vs Instances of Silence (x-axis)

Table 1. Test and Train Data Recorded and Normalized for Initial, Intermediate and End Paragraph for Speaker 1

Speaker 1											
Start Paragraph				Intermediate Paragraph				End Paragraph			
Test	Normalized Test	Train	Normalized Train	Test	Normalized Test	Train	Normalized Train	Test	Normalized Test	Train	Normalized Train
58080	1	89760	1	3960	0.111	4400	0.116	15840	0.414	4400	0.094
1320	0.023	880	0.01	1320	0.037	880	0.023	19800	0.517	8800	0.189
1320	0.023	3520	0.039	1320	0.037	880	0.023	7920	0.207	880	0.019
2640	0.045	880	0.01	1320	0.037	1760	0.047	3960	0.103	880	0.019
1320	0.023	13200	0.147	2640	0.074	9680	0.256	11880	0.31	2640	0.057
10560	0.182	880	0.01	9240	0.259	12320	0.326	11880	0.31	14960	0.321
3960	0.068	2640	0.029	13200	0.37	2640	0.07	6600	0.172	880	0.019
1320	0.023	2640	0.029	5280	0.148	2640	0.07	19800	0.517	29920	0.642
3960	0.068	2640	0.029	5280	0.148	5280	0.14	9240	0.241	9680	0.208
1320	0.023	2640	0.029	35640	1	37840	1	13200	0.345	2640	0.057
1320	0.023	31680	0.353	1320	0.037	880	0.023	1320	0.034	1760	0.038
2640	0.045	880	0.01	1320	0.037	880	0.023	14520	0.379	3520	0.075
3960	0.068	2640	0.029	3960	0.111	4400	0.116	1320	0.034	2640	0.057
1320	0.023	9680	0.108	1320	0.037	1760	0.047	11880	0.31	3520	0.075
31680	0.545	1760	0.02	1320	0.037	1760	0.047	5280	0.138	17600	0.377
13200	0.227	3520	0.039	3960	0.111	6160	0.163	1320	0.034	1760	0.038
3960	0.068	2640	0.029	2640	0.074	4400	0.116	1320	0.034	3520	0.075
1320	0.023	11440	0.127	1320	0.037	1760	0.047	2640	0.069	9680	0.208
1320	0.023	1760	0.02	5280	0.148	3520	0.093	1320	0.034	8800	0.189
10560	0.182	3520	0.039	0	0	0	0	14520	0.379	880	0.019
1320	0.023	7040	0.078	0	0	0	0	1320	0.034	880	0.019
3960	0.068	1760	0.02	0	0	0	0	3960	0.103	20240	0.434
9240	0.159	34320	0.382	0	0	0	0	5280	0.138	10560	0.226
1320	0.023	2640	0.029	0	0	0	0	10560	0.276	4400	0.094
3960	0.068	13200	0.147	0	0	0	0	6600	0.172	2640	0.057
1320	0.023	3520	0.039	0	0	0	0	1320	0.034	4400	0.094
34320	0.591	2640	0.029	0	0	0	0	5280	0.138	880	0.019
2640	0.045	2640	0.029	0	0	0	0	17160	0.448	46640	1
9240	0.159	0	0	0	0	0	0	10560	0.276	16720	0.358
2640	0.045	0	0	0	0	0	0	7920	0.207	5280	0.113
2640	0.045	0	0	0	0	0	0	7920	0.207	10560	0.226
0	0	0	0	0	0	0	0	1320	0.034	1760	0.038
0	0	0	0	0	0	0	0	38280	1	880	0.019
0	0	0	0	0	0	0	0	14520	0.379	880	0.019
0	0	0	0	0	0	0	0	2640	0.069	0	0
0	0	0	0	0	0	0	0	1320	0.034	0	0
0	0	0	0	0	0	0	0	2640	0.069	0	0

Table 1 depicts silence samples from Test and Train data recorded and normalized for Initial, Intermediate and End paragraph for speaker 1. The graph of the normalized train and test data for all the speakers for initial , intermediate and end part of the paragraph after experimentation show that the silence pattern in intermediate and end paragraph have better matching compared to the initial paragraph.

Figure 2 a & b represent the graph of normalized silence sample values of Initial paragraph and end paragraph for speaker 1 Vs the instances of silence is plotted. The comparison shows that silence in intermediate paragraph have better matching results compared to the end paragraph.

It is observed that Intermediate paragraph shows better correlation values for normalized Train and Test samples compared to end paragraph.

4. Testing & Results

The correlation of the normalized silence values for each speaker in the train and Test folders are depicted in the Table 2

Table 2: Correlation between Normalized Silence Sample Values of Test and Train Dta for Initial, Intermediate and Final Paragraph of 8 speakers

Correlation								
Paragraph	SP1	SP2	SP3	SP4	SP5	SP6	SP7	SP8
Initial	0.666	0.0058	0.144393	0.222037	-0.13835	0.005843	-0.04049	0.064
Intermediate	0.93614	0.193856	0.049007	0.239726	0.129241	0.362003	0.035346	0.005
End	0.282068	-0.05442	0.119878	-0.04535	-0.08544	0.044086	0.225506	0.763
Max-Value	0.93614	0.193856	0.144393	0.239726	0.129241	0.362003	0.225506	0.76307

For 8 speakers database used the value of correlation is high for 5 instances out of 8 for intermediate paragraph, twice for End Paragraph and once for initial paragraph. Thus the percentage of matching pattern of silences is 87.5% for intermediate paragraph and end paragraph considered together.

5. Conclusion

The analysis and correlation of the data experimented for above 8 speakers' shows that the similarity of the pauses is on the higher side in the intermediate paragraph and end paragraph than in the initial paragraph. Also the pattern of taking pause of each speaker is unique and can be used to identify the speaker. This observation allows us to conclude that the natural manner of a speaker may be reflected after speaking a few sentences in the beginning. This experimentation would lead the researchers to explore this alternative approach of obtaining silence as a unique feature of a speaker for identifying a speaker.

1) The performance of the system is dependent on the speakers fluency, command over the language, the time when both recordings are done, the environment of recording and many more factors.

2) The matching percent can improve by Multi-session recording instead of single session as this might help reduce the intra-speaker variability.

3) Experimentation can be done by varying the duration of recordings, a larger paragraph could be selected which would give a database of more than the current one of 50 sec .

4) Proper mechanism to set a threshold for identifying the silence in the speech can be devised instead of the trial based or permutation combination approach.

References

[1] M. A. Anusuya and S.K. Katti, “Speech Recognition by Machine: A Review”, International journal of computer science and Information Security, **(2009)**.

