A NOVEL APPROACH TO NOISY SPEECH RECOGNITION USING DTW ALGORITHM WITH MEL-FREQUENCY CEPSTRAL COEFFICIENTS

Rishad Ahmed Shafik* **Fazli Qayyum Yousaf-Zai***

ABSTRACT

A new and effective approach to recognition of noisy speech is introduced. End-Point-Detection algorithm is used to measure the noise power and to automatically initiate recording of a spoken word. Unvoiced components of the recorded speec h, buried in noise, viz. ambient noise or hiss noise or telephone noise, were then optimally minimized by Yulewalk Finite Impulse Response (FIR) band pass Filter. The speech signal was then sampled and speech features were extracted using Mel-Frequency Cepstral Coefficients (MFCC), which were later dynamically time-warped to find the average minimal distance from Euclidean distance matrices to help facilitate the recognition of speech. For generalization, speech data from three speakers, of three different level of pitch, were collected and were compared to a mid-pitch speaker. This work will be extended to establish both speaker independent and speaker dependent efficacy and accuracy in future. Such a speech recognition system can be both fast and effective even in moderately hostile environments.

Keywords: End-Point-Detection algorithm (EPD), Yulewalk Finite Impulse Response (FIR) Filter, Mel-Frequency Cepstral Coefficients (MFCC), Dyanamic Time Warping (DTW).

1. INTRODUCTION

The history of speech recognition dates back to the 1870s. Extensive research has been carried out to establish speech recognition system to work in varying conditions. , , , Bendelac and Shallom[2] compared recognition rates for noiseless and simulated cellular car noise conditions using Dynamic Time Warping (DTW) technique, and reported accuracies above 95%. Awad et al[3] reported greater than 95% accuracy using an isolated 22-word vocabulary trained to a single speaker. Davis and Mermelstein[4] compared Cepstral and Linear Predictive Coding (LPC) distance measures using DTW methods for both the speaker dependent case as well as the independent cases on a vocabulary of 52 monosyllabic words from two different speakers, with 85% to as high as 98% accuracy. Rabiner et al[5] reported 5% to 35% error probabilities using DTW with LPC distance metric for 100 different speakers and isolated 10-word vocabulary. Sakoe and Chiba discussed various DTW local and global constraint options and reported error rates of less than 1% for their speaker dependent tests. Gray and Markel examined a number of distance measures, which are applicable to the isolated word recognition problem and reported that the Root Mean Square (RMS) logarithmic spectral distance measure using Cepstral coefficients performed the best. Kuitert and Boves[12Error!

* *EEE Department, Islamic University of Technology, Gazipur-1704, Bangladesh*

Bookmark not defined.] demonstrated speech recognition for Global System for Mobile (GSM) coded speech with hardware filter solutions. In the following sections, the recognition of speech buried in three different experimental noises will be investigated, viz. ambient noise, hiss noise and telephone noise. For this work, a 7-word vocabulary would be assumed as- 'STOP', 'FRONT', 'BACK', 'RIGHT', 'LEFT', 'FAST' and 'SLOW'. Finally, the results of average minimum distance calculation will be presented after applying DTW algorithm on MFCCs.

The speech recognition process that has been implemented can be best described by the following flow chart in figure-1. The main processes, viz. End Point Detection (EPD), Noise Filtering, vocabulary formation using MFCC and DTW will be illustrated in the following sections.

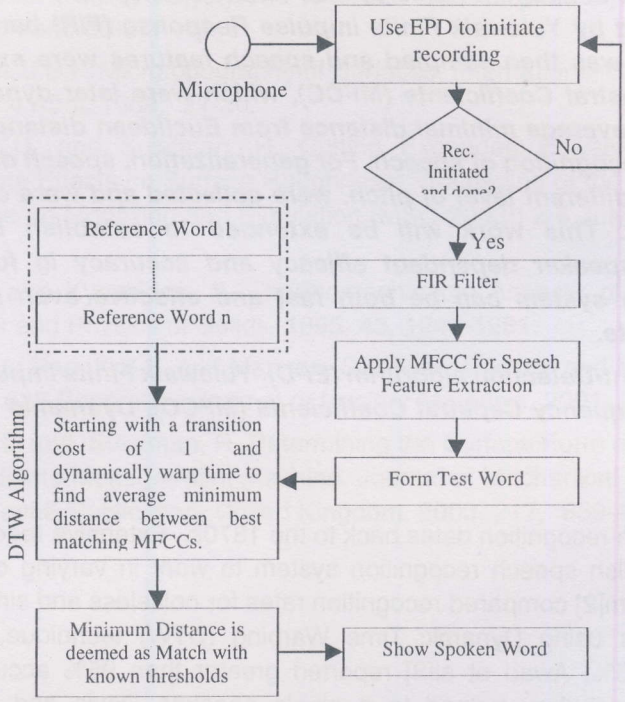

Figure 1: Flow Diagram of Optimal Speech Recognition

2. END POINT DETECTION AND RECORDING

Instead of manual parsing of recording of a speech, as was done in [5] and [6], the recording would be started after a continuous monitoring of average noise level by EPD algorithm. Several noises were simulated and examined to find the average minimum distance results, viz. ambient room noise, hiss noise, and telephone noise. A typical hiss noise for Signal-to-Noise Ratio (SNR) of 30dB had been simulated and used. Telephone line noise comes in many forms, such as, electrical interference from fluorescent fixtures, high frequency from the many amplifier stages in the voice path and sometimes from

cross talking. It usually ranges up to SNR of 45dB. End Point Detection (EPD) algorithm was used to calculate noise energy for frames up to 256 samples with frame overlap of 86 samples. Thus, for a frame i, the noise energy was found out using[9] -

$$p[i] = \sum_{k=1}^{j} (s[k]^2) \quad (1)$$

where k = 1, ...,m

Where s[k] is the speech data in each frame with j samples. Similarly, p[i] is calculated for all the frames over one second and an expectation for the final noise value, Enoise is found out as-

$$E_{noise} = \xi\{p[i]^2)\} \quad (2)$$

Where x operator stands for expectation over the total number of frames, p[i] is the noise power for frame i. Thus, when an utterance is heard, the recording is initiated only when the average energy level, Eavg, is higher than average noise level Enoise. Recorded ambient room noise, simulated hiss noise of SNR 30dB and recorded telephone noise of SNR 45dB measured at the same time interval is shown in figure-2. The energy profile of signal corresponding to the utterance, 'STOP' for each frame, showing where the average noise level is crossed, is illustrated in figure-3.

After the end point was detected, the program recorded the speech signal input for about

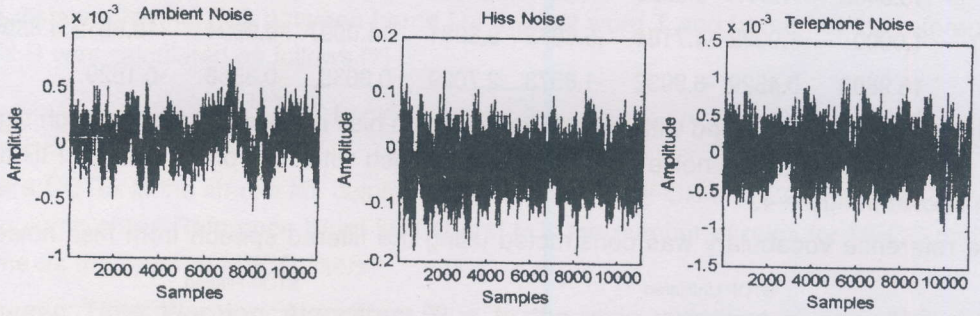

Figure 2: Noise Plot of Ambient Noise (Left), Hiss Noise (Centre), Telephone Noise (Right)

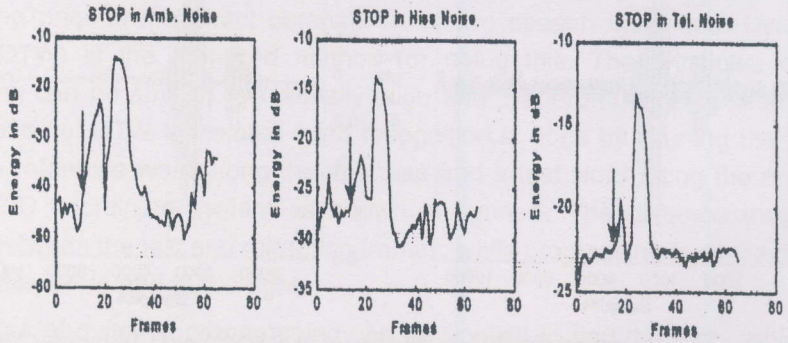

Figure 3: Average Energy Level (dB) and First End from EPD for Ambient Noise (Left), Hiss Noise (Centre) and Telephone Noise (Right)

1000 milliseconds in 8-bit format and sampled at a rate of 11025 Hz. This method of speech detection showed good synchronization of speech with the reference vocabulary as compared to manual parsing as in [5] and [6].

3. NOISE FILTER

Yulewalk FIR band pass filters effectively reduce noise from signal buried in band-limited noise that may change slightly over time. The different noise assumed in this work can thus be effectively removed using Yulewalk FIR band pass filter. Since the desired signal would be voiced speech signal and the unvoiced noise components with minimum noise, it was extracted after filtering through a pass band of human voice frequency range of .03kHz to 3.4kHz. This reduces any high frequency noise present in the signal. However, this did not reduce the in-band telephone cross talking noise but worked satisfactorily with hiss noise. The polynomial form of transfer function for a Yulewalk FIR filter is given by-

$$H(z) = \frac{B(z)}{A(z)} \tag{3}$$

The coefficients of numerator B(z) and denominator A(z) of the 16th order filter are shown below with decreasing powers of polynomials -

B(z)	0.6379	1.0634	-2.9958	-5.9776	5.5451	14.8341	-4.2175	-21.0237	-0.9835
	18.3406	4.6417	-9.8326	-3.9104	2.9961	1.5217	-0.4001	-0.2389	
A(z)	1.0000	1.0032	-4.7184	-5.6645	9.5081	14.0097	-9.9924	-19.6912	4.8594
	16.9896	0.4529	-8.9932	-1.8573	2.7029	0.9035	-0.3556	-0.1529	

Major problems were faced with speech immersed in hiss noise. Unfiltered speech signal for 'STOP' buried in hiss noise and the filtered speech with 16th order Yulwalk FIR filter are shown in figure-4.

The reference vocabulary was constructed using the filtered speech from hiss noise in

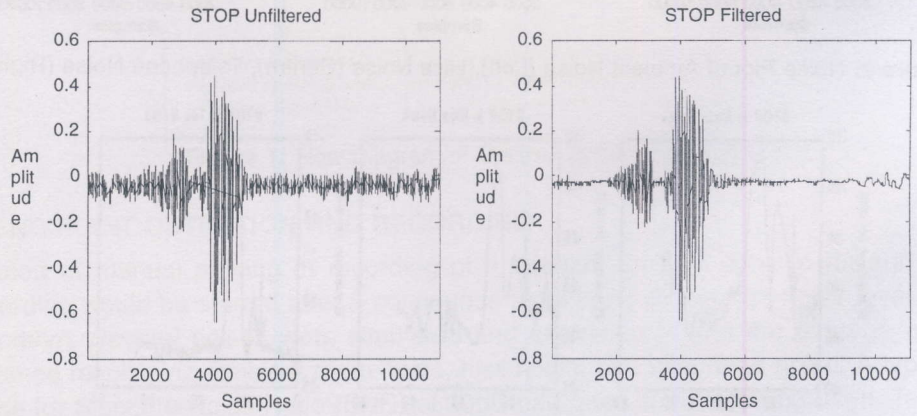

Figure 4: Unfiltered word 'STOP ' (Left) and filtered word 'STOP' using 16th order Yulewalk FIR filter (Right)

order to have the minimum distance output in noisy conditions.

4. SPEECH RECOGNITION ALGORITHM

Mel-Frequency Cepstral Coefficients: After the speech were recorded and noise reduction technique was applied, both the reference speech data and test speech data were divided in 23.2 milliseconds (n=256-point) frames, which were stepped by 11.6 milliseconds (m=128 points) between processing frames. A total of 10 overlapped 200-Hz triangle filters were spaced evenly between 0 and 1 kHz followed by 10 logarithmically increasing bandwidth filters spaced logarithmically from 1 kHz to the Nyquist frequency. The log-energy outputs Xk from the Mel-frequency filters were used along with the following equation to calculate the Mel-frequency Cepstral Coefficients[6]

$$MFCC_i = \sum_{k=0}^{20} X_k \cos\left(\frac{\pi i(k-0.5)}{20}\right) \quad i = 1,2,.....P \tag{4}$$

Where, 20 is the number of Mel-frequency filters and P=10 is the order. The coefficients for each reference word were saved in a P×m matrix to form vocabulary. Later, with the reference vocabulary and test word, the distance metric was formed by using a Euclidean distance for the Cepstral coefficients over all frames after dynamic time warping was applied to align the frames optimally. All paths were given a transition cost of 1 initially. The distance metric, Di,j, between frame i of the test word T and frame j of the reference word R was calculated as follows.[7]

$$D_{i,j} = \left(\frac{1}{P}\right)\sqrt{\sum_{k=1}^{P}\left(CCT_{i,k} - CCR_{j,k}\right)^2} \quad i = 1,2,.....P \tag{5}$$

Where CCTi,k is the ith row kth column of the Test Word MFCC and CCRj,k is the jth row kth column of the Reference Word MFCC and P is the number of rows for MFCC matrix (same as the order of MFCC filters).

Dynamic Time Warping Algorithm: Due to the wide variations in speech among speakers and among different instances, it is necessary to apply some kind of non-linear time warping prior to the direct comparison of two speech instances. Dynamic Time Warping, (DTW) is the preferred method for doing this. The principles of dynamic programming can be applied to optimally align and synchronize the speech signals.[6] The application of DTW in isolated word recognition is done by aligning the processing frames of a reference word along the abscissa and a test word along the ordinate of a Cartesian 2-D coordinate system as shown in figure-5. The distance metric is then computed between the test and reference frames, while progressing from the origin at the left bottom corner up and to the right.

The principles of dynamic programming can be applied to find the path, which has the

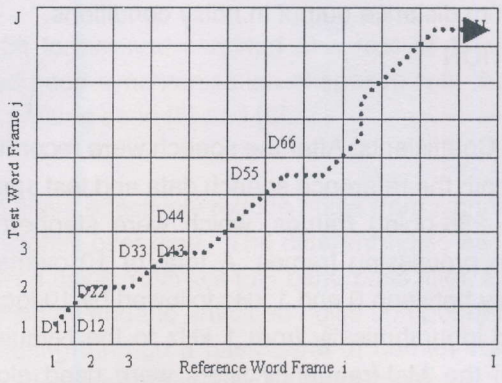

Figure 5: Dynamic Time Warping

minimum accumulated distance metric. After performing this test, using all of the reference vocabulary words for each test word, the reference word with the minimum accumulated distance metric is deemed to be a match. For a speech signal, there are a number of constraints on the search path, which can be applied to help decrease the complexity of the search. The primary constraint is that the search should be monotonic. This can be forced by the application of global and local constraints.[9] Global constraints are simply overall constraints on the valid overall search path. Boundaries are placed by a maximum and minimum slope (Smax , Smin) as well as by allowing the search to begin and/or end within a given frame tolerance (Eps) of the initial and final frames. Local constraints determine the valid search path on a local basis. Sakoe and Chiba[6] presented a number of variations of local constraints. For this work, the only locally imposed constraint was monotonic. This local constraint simply requires that the only valid paths to a given point must pass through a point from the left and/or below. Global and local constraints used in DTW algorithm are shown in figure-6. The average minimum distances found for all the points for coefficient matrices are then averaged for each sample.

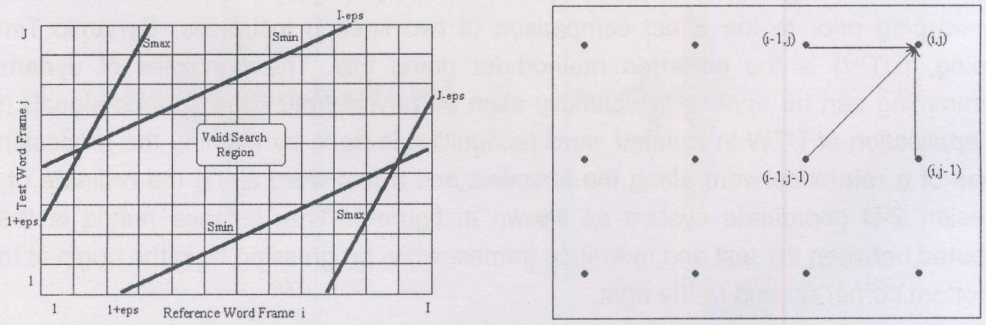

Figure 6: Global Constraints of DTW (left) and Local Constraints on DTW (right)

5. RESULTS

The present speech recognition system was tested with three different speakers A, B, and C with a reference vocabulary set using speech signals from speaker A, chosen with medium pitch. Speaker B and C are chosen with high and low pitch respectively. The reference vocabulary was set up using filtered speech from ambient noise and the distances shown in the result tables show average minimum distances for speech words immersed in three noises discussed before. Results were observed for maximum and minimum slope of Smax=3.0 and Smin=0.5 and frame tolerance of Eps=3 with order of P=10. The minimum values (shown bold in the result/tables 1, 2 and 3 in Appendix A) of the average minimum accumulated distances set, generated from comparison of the coefficients of test word with that of seven reference words using DTW algorithm, was always picked up as the spoken word. However, since hiss noise consists of critical pass-band noise, the performances were worse than the other two cases. Accuracies were measured using average correct recognitions out of total attempts and variances were minimized for values of Smax and Smin and Eps. Thus, as good as 100% test accuracy was obtained for speaker dependent tests for speech words immersed in ambient noise and telephone noise but for that in hiss noise, the test accuracy was measured to be 71.43%. In speaker independent tests, high pitched speaker B showed test accuracy of 42.86% in hiss noise and 85.71% accuracy was observed in both ambient noise and telephone noise. Low pitched speaker C showed test accuracy of 71.43% in hiss noise and 85.71% accuracy in ambient noise and 100% accuracy in telephone noise was observed. It was also seen previously that for perfect matching of two distances matrices, the minimum distance would be 1.000, which was set to be the initial path weight. Due to the bulk of filtration and monotonic DTW algorithm, the time required for each speech recording to recognition was estimated to be about 7.83 seconds on an average with an 833MHz microprocessor, which is quite acceptable and fast with efficacy concerned. Speaker A showed the best results for its pitch and feature similarity. The minimum distances results varied slightly for different speakers in speaker independent tests for dissimilarities of pitch and length of speech. As seen from the tabular set of data, most of the inaccuracies took place in the case of either 'FAST' or 'SLOW', because of speech similarities with 'FRONT' and 'STOP'. But this can be made up with introduction of automatic learning system using database approach.

6. CONCLUSION

In this work, detailed results have been found for speech recognition of a 7-word vocabulary, with sample speeches taken in three different noisy environments. It has been found that, for ambient noise the recognition accuracy is better than for telephone noise and hiss noise. Telephone noise can be random and performances may vary for different instances. Hiss noise carries pass-band spectrum in human voice frequency range and as such, to improve performance of speech in such noise, adaptive noise cancellation technique can be used and is being considered for work for future. Research will also be carried out in future on the optimization of speech recognition impaired by Additive White Gaussian Noise (AWGN) and power noise signals. Such recognition systems will be greatly useful in space and military applications.

6. Bibliography

[1] http://florin.stanford.edu/~t361/Fall2000/TWeston/history.html [Last Accessed 02/10/2004].

[2] Bendelac G., Shallom I., "Eyes-Free Dialing for Cellular Telephones" Proceedings, IEEE Vehicular Technology Conference, 1991, vol18#1, pp.512-515.

[3] Awad S., Wagner D., Flaherty, M."A Voice Controlled Telephone Dialer" IEEE Trans on Instrumentation and Measurement, 2/89, vol.38#1, pp.38-39.

[4] Davis S., Mermelstein P., "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences" IEEE Trans on Acoustics, Speech, and Signal Processing. 8/80, vol.ASSP-28#4, pp.142-146.

[5] Rabiner L., Rosenberg A., Levinson S., "Considerations in Dynamic Time Warping Algorithms for Discrete Word Recognition" IEEE Trans on Acoustics, Speech, and Signal Processing vol.ASSP-26 #6 12/78, pp.129-134.

[6] Sakoe H., Chiba S., "Dynamic Programming Algorithm Optimization for Spoken Word Recognition" IEEE Trans on Acoustics, Speech, and Signal Processing vol.ASSP-26 #1 2/78, pp.1008.

[7] Gray A., Markel J., "Distance Measures for Speech Processing" IEEE Trans on Acoustics, Speech, and Signal Processing vol.ASSP-24 #5 10/76, pp.1721-1724.

[8] http://www.epanorama.net/links/telephone.html [Last Accessed 23/09/2004].

[9] http://www.cnel.ufl.edu/~kkale/6825Project.html [Last Accessed 23/09/2004].

[10] Haigh J.A., Mason J.S., "A Voice Activity Detector based on Cepstral Analysis", Journal of Signals and Systems, April, 1993, vol.11, pp.74-79

[11] Kassianides C., Otung I.E., "A Dynamic model of Tropospheric Scintillation", First International Workshop on Radiowave Propagation Modelling for SatCom Services at Ku-band and above, 28-29 October 1998.

[12] Kuitert M., Boves L., "Speaker Verification with GSM coded Telephone Speech" EUROSPEECH-97 Proceedings, Sept 22-25 1997, pp.11-14.

[13] Openheim and Schaefer, Discrete-Time Signal Processing, 3rd Edition, Prentice-Hall, Englewood Cliffs, 1989.

APPENDIX A: Test Results

Table 1: Average Minimum Distance, D_{min} for Speaker A for different noises **[Speaker Dependent Test]**

Test→	'STOP'			'FRONT'			'BACK'			'RIGHT'			'LEFT'			'FAST'			'SLOW'		
Ref.↓	Amb	Hiss	Tel.	Amb	Hiss	Tel.	Amb	Hiss	Tel.	Amb	Hiss	Tel.	Amb	Hiss	Tel.	Amb	Hiss	Tel.	Amb	Hiss	Tel.
'STOP'	**2.4269**	**2.6897**	**2.2365**	2.8239	3.2381	2.8638	3.2134	3.3222	3.0933	3.5443	3.2466	3.1841	3.2995	3.4743	3.1983	3.1059	3.3067	3.1877	2.6854	**2.9674**	2.4457
'FRONT'	2.7214	3.2682	2.8044	**2.3446**	3.0887	**2.2986**	3.0046	3.4781	2.9914	2.7723	3.2738	2.6822	2.9440	3.7168	2.8914	2.8497	3.4938	2.7593	2.9416	3.4371	2.7564
'BACK'	3.0125	3.7969	3.0480	2.9487	3.5154	2.8716	**2.3117**	**2.8498**	**2.2524**	2.7536	3.5629	2.9141	2.7155	3.4465	2.7149	2.7681	3.4665	2.8249	3.0778	3.7725	3.1867
'RIGHT'	3.1535	3.7016	2.8250	2.6504	3.1463	2.5050	2.7076	3.2536	2.6353	**2.3901**	**2.9246**	**2.2798**	2.8970	3.4115	2.7868	2.6535	3.3915	2.9086	3.3013	3.6666	3.1211
'LEFT'	3.0759	3.2923	2.8291	3.0946	3.1397	2.7822	2.6802	3.0888	2.6820	2.7957	3.0160	2.5876	**2.4152**	**2.8242**	**2.3757**	2.8111	3.2648	2.5540	3.1244	3.3439	3.0811
'FAST'	3.0726	3.1506	2.6744	2.6728	**2.8976**	2.6365	2.8493	3.2305	2.7669	2.8155	2.9467	2.6773	2.7415	3.1542	2.7046	**2.2329**	**2.7086**	**2.3659**	3.2266	3.1635	3.1123
'SLOW'	2.5248	3.1054	2.5119	2.9172	3.5331	2.9700	3.3575	3.5513	3.3138	3.6990	3.5938	3.3549	3.3240	3.6405	3.2782	3.2718	3.6902	3.4702	**2.5361**	3.1210	**2.2442**

Table 2: Average Minimum Distance, D_{min} for Speaker B for different noises **[Speaker Independent Test]**

Test→	'STOP'			'FRONT'			'BACK'			'RIGHT'			'LEFT'			'FAST'			'SLOW'		
Ref.↓	Amb	Hiss	Tel.	Amb	Hiss	Tel.	Amb	Hiss	Tel.	Amb	Hiss	Tel.	Amb	Hiss	Tel.	Amb	Hiss	Tel.	Amb	Hiss	Tel.
'STOP'	**2.6838**	3.1978	**2.4323**	3.1109	3.2414	2.9232	3.1635	3.2223	3.2557	3.4535	3.5032	3.1350	3.3074	3.3543	3.1075	3.4408	3.4615	3.3263	**2.8519**	**3.0816**	2.5993
'FRONT'	3.0509	3.3617	2.6648	**2.7103**	3.2016	**2.4929**	2.9444	3.4497	2.8924	2.9584	3.3825	2.6007	3.1511	3.4887	2.8554	3.0130	3.3251	2.8083	3.0004	3.3670	2.9027
'BACK'	3.1801	3.6793	3.0667	3.2979	3.7033	2.7996	**2.7438**	**2.9805**	**2.4686**	3.0274	3.5962	2.7204	3.0818	3.5737	**2.5140**	3.0085	3.4611	3.0798	3.3482	3.6665	3.3253
'RIGHT'	3.1656	3.4866	3.1663	2.9537	3.3899	2.8171	2.8946	3.2774	2.7474	**2.7655**	3.2583	**2.3847**	2.8365	3.2874	2.6664	2.8462	3.1566	2.7682	3.3747	3.4780	3.3893
'LEFT'	3.2895	3.3603	3.1654	3.2889	3.3815	3.0562	3.0275	3.1259	2.8499	3.0732	3.3593	2.9394	**2.5774**	**2.9095**	2.5435	3.0174	3.2330	3.0911	3.4674	3.4727	3.5320
'FAST'	3.1539	**3.1011**	3.0105	3.0683	**3.0935**	2.8798	3.0349	3.3112	2.8147	2.9835	**3.1014**	2.7732	2.8618	2.9840	2.8682	**2.6637**	**2.7741**	**2.4438**	3.3812	3.2701	3.2144
'SLOW'	3.0018	3.5981	2.6818	3.3488	3.5800	2.9037	3.3212	3.4645	3.3775	3.7131	3.7408	3.2832	3.5188	3.5136	3.0615	3.5682	3.7183	3.4156	2.8797	3.3161	**2.5344**

Table 3: Average Minimum Distance, D_{min} for Speaker C for different noises **[Speaker Independent Test]**

Test→	'STOP'			'FRONT'			'BACK'			'RIGHT'			'LEFT'			'FAST'			'SLOW'		
Ref.↓	Amb	Hiss	Tel.	Amb	Hiss	Tel.	Amb	Hiss	Tel.	Amb	Hiss	Tel.	Amb	Hiss	Tel.	Amb	Hiss	Tel.	Amb	Hiss	Tel.
'STOP'	**2.7290**	**3.0478**	**2.6182**	3.1276	**3.1068**	2.8198	3.3556	3.4041	3.3097	3.4627	3.3619	3.3593	3.4828	3.4474	3.4563	3.4829	3.5733	3.1083	2.9441	3.0579	2.6041
'FRONT'	2.9031	3.4224	2.9061	**2.5124**	3.2379	**2.2762**	3.2038	3.5905	3.1664	2.8747	3.3346	2.6384	3.3062	3.4960	3.0965	3.3275	3.7864	3.0834	3.1466	3.4048	2.9805
'BACK'	3.0601	3.6361	3.0216	3.3888	3.7121	2.9833	**2.6642**	3.2704	**2.4436**	3.0608	3.6133	2.7477	3.1098	3.6211	2.7988	3.1695	3.7508	2.6799	3.6703	4.0449	3.1863
'RIGHT'	3.1130	3.5743	3.0717	2.7272	3.1332	2.6153	2.9516	3.4696	3.1013	**2.5276**	**3.0782**	**2.4489**	2.9901	3.4615	2.8046	3.4578	3.9530	2.7715	3.9207	4.1627	3.2181
'LEFT'	3.0564	3.3393	3.0026	3.4599	3.5103	2.8536	3.1040	3.3109	2.6926	3.1564	3.2610	2.8129	**2.6103**	**3.3151**	**2.3689**	**3.1473**	3.3708	2.6227	3.4452	3.6576	3.2066
'FAST'	3.1047	3.0775	3.0237	3.0075	3.0865	2.6847	3.0993	**3.2074**	2.9877	2.8146	3.0213	2.7481	3.1342	3.2031	2.6935	3.1937	**3.3479**	**2.4356**	3.2241	3.8483	3.1710
'SLOW'	3.0446	3.4085	2.8372	3.1199	3.4621	3.0480	3.4231	3.6319	3.4767	3.5778	3.6207	3.3594	3.6203	3.6884	3.4377	3.6248	3.8477	3.3622	**2.7953**	**3.0125**	**2.5544**