Performance Analysis of Speech Features for Speaker Verification in Emotional Conditions

S K Das[1] & U Bhattacharjee[2]

[1, 2] Department of Computer Science & Engineering, Rajiv Gandhi University, Arunachal Pradesh, India

[1] satish.das@rgu.ac.in, [2] utpal.bhattacharjee@rgu.ac.in

Abstract

In the present study, we have analysed the following six spectral and source features: Mel Frequency Cepstral Coefficient (MFCC), Linear predictor cepstral coefficients (LPCC), Linear Predictor coefficient (LPC), Reflection coefficient (RC), Log area ratio (LAR) and Arc-sin reflection coefficients (ARC) for their relative speaker verification performance in emotional mismatched conditions. Emotions considered in our study are mainly Angry, Happy, Sad and Neutral. Further, it has been observed that prosodic features are extensively used in speaker verification task as well as emotion recognition. Since the prosodic features are highly sensitive to emotional conditions, we have not considered prosodic feature for further study. Our results show that, all the features are effected by the change in emotional condition of the speaker. However, ARC are found to be relatively robust to emotional changes. MFCC is the most robust feature in terms of deviation of the CDF statistics with emotional changes. LPCC and LPC features are found to be highly sensitive to emotional changes.

Key Words: *Speaker Verification, Emotion Recognition, MFCC, LPCC, LPC, RC, ARC*

1. Introduction

Feature is the compact representation of the acoustic properties manifested in the speech signal [1]. Choosing suitable features for developing any of the speech systems is a crucial design decision. A speech emotion recognition system should extract suitable features to characterise emotions of different kind efficiently [2]. Proper selection of features also related to proper selection of classifiers [3]. The features are to be chosen to represent the required information for the functioning of the proposed system.Different speech features represent different information of the speech signal in a highly overlapping manner.Therefore, for the development of a speech based system, the features are selected experimentally in most of the cases.In some of the cases, the features are also selected using mathematical approach like principal component analysis (PCA) [4].The speech features may be broadly classified into the following categories – (i) Excitation source features (ii) Spectral features and (iii) Prosodic features.

Speech features extracted from excitation source signal is called source features.Excitation source signal is obtained by discarding the vocal tract information from the speech signal. This is achieved by first predicting the vocal tract information using linear predictor filter coefficients extracted from the speech signal and then separating it by using inverse transformation. The resulting signal is called linear predictor residual signal [5]. The features extracted from LP residual signal is called excitation source features or source features.The state-of-the-art phone recognition systems are developed only with vocal tract information.

However, a sound unit is produced as a result of active involvement of excitation source and vocal tract. Just the shape of the vocal tract is not sufficient enough for the characterization of a sound unit.The bilabial plosive consonants b and p are produced by the same manner and place of articulation. The difference between these two sounds is coming as a result of difference in their excitation type. The consonant b is voiced and p is unvoiced. Similarly, for all the vowels, the excitation type is nearly similar. The difference between the vowel sounds are coming as a result of place and manner of articulation. Thus, we can conclude that each sound is produced as a result of unique combination of excitation source and vocal tract participation. Therefore, to characterize a sound unit, excitation source parameter as well as vocal tract parameter are necessary. The most commonly used source parameters for speaker verification are Reflection coefficient (RC), Log area ratio (LAR) and Arc-sin reflection coefficients (ARC) [6].

A sound unit is characterized by a sequence of shapes assumed by the vocal tract during production of the sound [7]. The vocal tract system can be considered as a cascade of cavities of varying cross sectional areas. During speech production, the vocal tractact as a resonator and emphasizes certain frequency components depending on the shape of the oral cavity. Formants are the resonances of the vocal tract at a given point of time characterized by bandwidth and amplitude [8]. These parameters are unique for a sound unit. The information about the sequence of shapes of vocal tract that produce the sound unit is captured by vocal tract features also called system or spectral features. The vocal tract features are clearly visible in the frequency domain. Frequency domain analysis of the speech signal is performed by segmenting the speech signal into frame of 20-30 ms, with the frame shift of 10 ms. Most commonly used spectral features are linear predictor cepstral coefficients (LPCC), mel frequency cepstral coefficients (MFCC), perceptual linear predictor coefficients (PLPC) and their derivations [9].

Prosody represents the suprasegmental aspects of speech production. Prosody is concern with those aspects of speech signal that modulate and enhance its meaning [10]. It makes the human speech natural. It is associated with longer unit of speech such as syllable, words, phrases and sentences. Prosody is acoustically represented by duration, intonation (F0 contour) and energy [11]. Mary and Yegnnarayana [12] analyzed the effectiveness of prosodic features for speaker verification. They observed that shape of the F0 contour reflects certain speaking habits of a person. In order to represent the shape of the F0 contour, tilt parameters have been used [13]. A 7-dimensinal feature vector was proposed, which includes mean value of pitch (F_{0_μ}), peak fundamental frequency (F_{0_p}), change of F0 (ΔF0), distance of F0 peak with respect to vowel onset point (VOP) (D_p), amplitude tilt (A_t), Duration tilt (D_t) and change of long energy (ΔE). Each region between two consecutive VOPs is represented using the above mentioned parameter set. They have conducted a study on NIST SRE 2003 extended database and concluded that there is a potential for these prosodic features for speaker verification. Further, it was suggested that due to the complementary nature of the prosodic and spectral features, the overall speaker verification performance can be improved while combing the evidences. Carey et al [14] used prosodic feature for speaker identification. They used mean, variance, skew and kurtosis of the pitch and energy and their first two derivatives as feature vector. They combine these features with the cepstral features. The feature vector has been used with a Hidden Markov Model (HMM) based speaker identification system. Six NIST 1995 evaluation test were conducted and a 30% performance gain was achieved. They have observed that prosodic features are more robust in handset mismatched conditions. Many other researches also

reported enhanced speaker verification performance when prosodic features are combined with spectral features [15, 16]. Prosodic features are also extensively used in emotion recognition. Many attempts have been made to detect the emotional content of a speech utterance by using prosodic features such as pitch and energy [17, 18, 19, 20].

In the present study, we have analysis the following six spectral and source features: Mel Frequency Cepstral Coefficient (MFCC), Linear predictor cepstral coefficients (LPCC), Linear Predictor coefficient (LPC), Reflection coefficient (RC), Log area ratio (LAR) and Arc-sin reflection coefficients (ARC) for their relative speaker verification performance in emotional mismatched conditions. Further, it has been observed that prosodic features are extensively used in speaker verification task as well as emotion recognition. Since the prosodic features are highly sensitive to emotional conditions, we have not considered prosodic feature for further study.

2. Analysis of features

Speech signal has been segmented into 20 msec frame with 10 msec overlapping. Hamming window has been applied to each frame. An energy based voiced activity detection algorithm has been applied to eliminate the silent portion. For each windowed speech segment, all the six features have been extracted. Performance of the features have been evaluated for:

Inter-speaker variability

Intra-speaker variability due to emotion

To analyse the relative performance of different speech features for speaker recognition at emotional condition, three statistical evaluation methods have been used. They are, probability density function (PDF) characteristics, Analysis of variance (F-ratio) and Kolmogorov-Smirnov (K-S) test.

Data distribution of a class close to the normal distribution leads to better classification [21]. Probability density function (PDF) of the coefficients have been computed for the same speaker at different emotional conditions. The results of the experiment are given below:

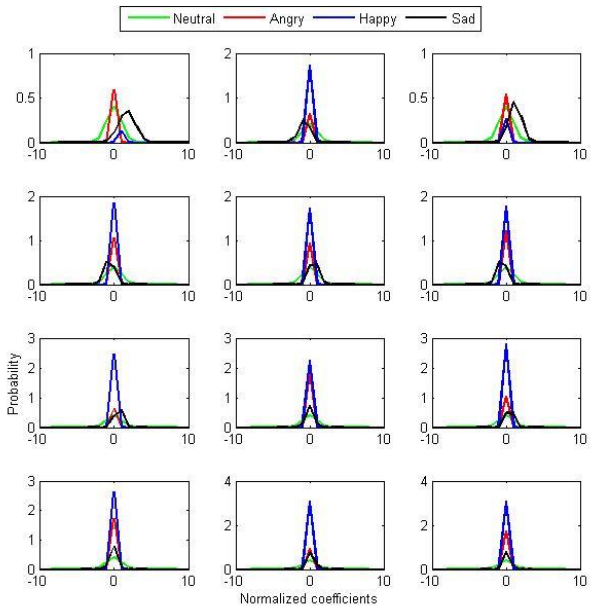

Figure 1. PDF for first 12 coefficients of MFCC at Neutral and Emulated Emotional condition

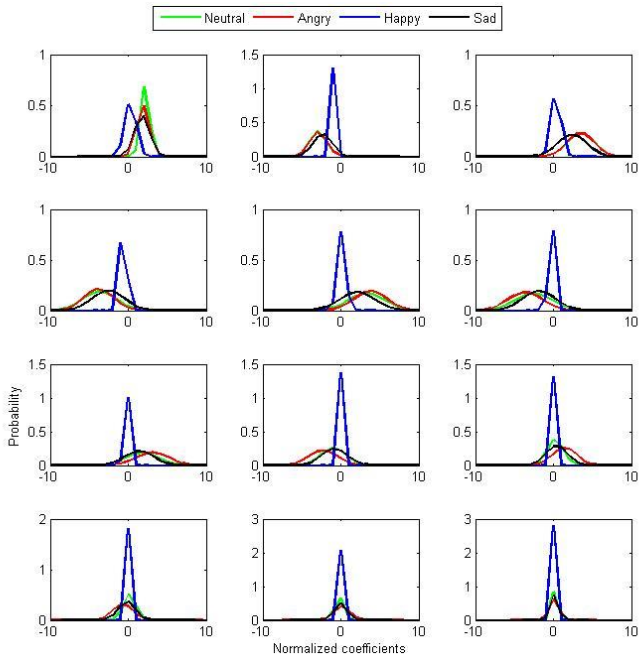

Figure 2. PDF for first 12 coefficients of LPCC at Neutral and Emulated Emotional condition

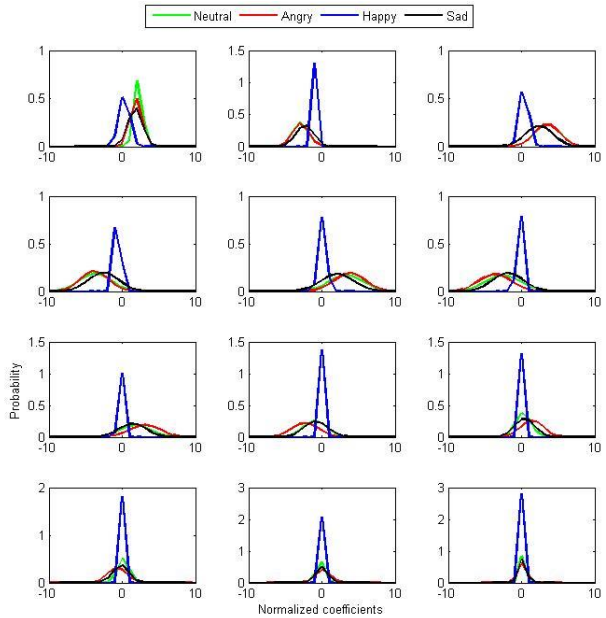

Figure 3. PDF for first 12 coefficients of LPC at Neutral and Emulated Emotional condition

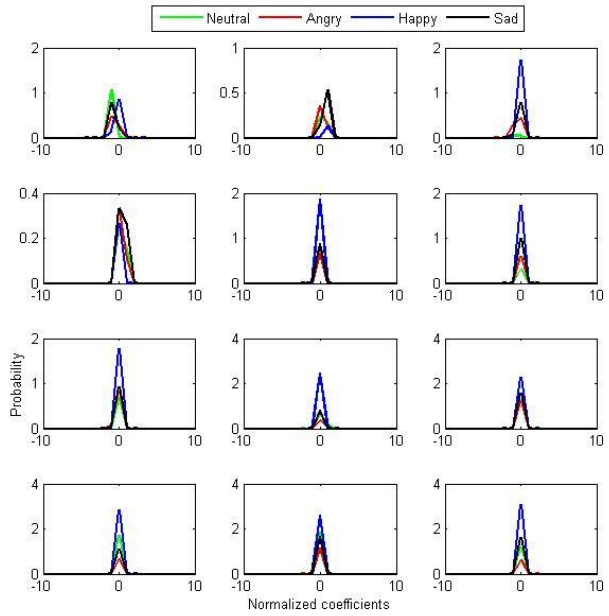

Figure 4. PDF for first 12 coefficients of RC at Neutral and Emulated Emotional condition

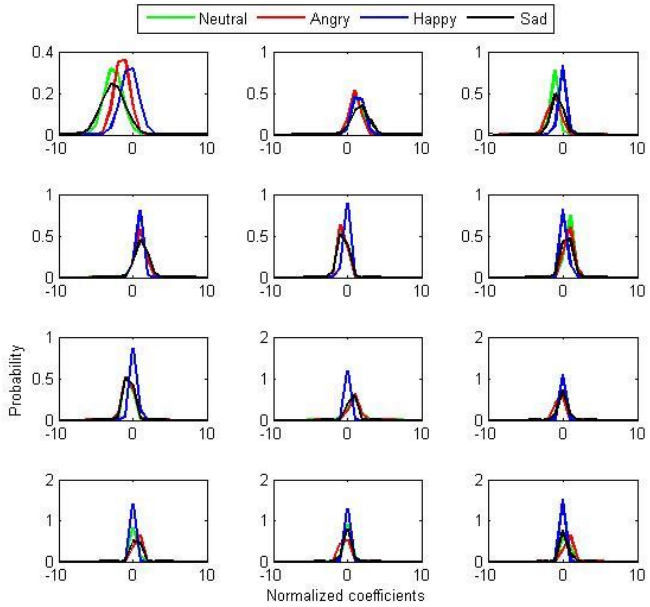

Figure 5. PDF for first 12 coefficients of LAR at Neutral and Emulated Emotional condition

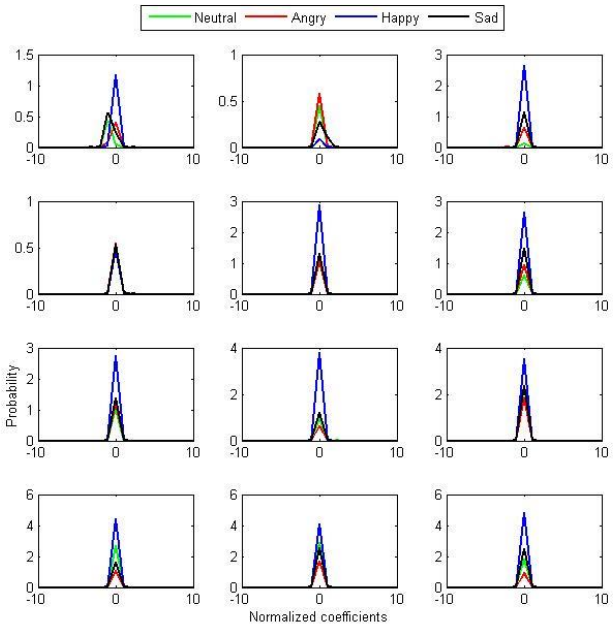

Figure 6. PDF for first 12 coefficients of ARC at Neutral and Emulated Emotional condition

The average mean of the PDF and their relative displacement due to change is emotional condition of the speaker has been calculated and summarised in the table given below :

Table 1. Average mean of the PDF at different emotional conditions for each feature type

Feature Name	Neutral	Angry	Happy	Sad	Average Difference between two pdf means
MFCC	1.4507	0.1203	0.3969	0.1422	0.6318
LPCC	2.1484	0.1551	0.0386	0.0419	1.0797
LPC	2.7871	0.1809	0.0394	0.0665	1.3950
RC	0.4849	0.1809	0.0397	0.0665	0.2438
ARC	0.3410	0.1203	0.0255	0.0437	0.1717
LAR	1.2211	0.3964	0.0809	0.1422	0.6157

From the above experiments, it has been observed that all the features are effected by the change in emotional condition of the speaker. However, ARC are found to be relatively robust to emotional changes. LPCC and LPC features are found to be very much sensitive to emotional changes.

For quantitative evaluation of the impact of emotion on speech data, F-ratio value has been computed. F-ratio is the ratio of the variability between two groups by variability within the group

$$F = \frac{between-group\ variability}{within-group\ variability} \quad (1)$$

In the present study, we have considered two group of features, extracted at two different emotional conditionsand their F-ratio has been computed. If the F-ratio is higher, it indicates that the feature is sensitive to the change in emotional condition of the speaker. Lower value

of F-ratio indicates more robustness towards emotional changes.Table given below shows the average F-ratio for each feature.

Table 2. F-ratio for MFCC feature under different emotional condition

	Sad	Angry	Happy
Neutral	1.0017	0.1914	0.3626
Happy	1.2149	2.3461	
Angry	0.5455		

Table 3. F-ratio for LPCC feature under different emotional condition

	Sad	Angry	Happy
Neutral	1.2863	2.9497	2.0721
Happy	4.8383	3.3470	
Angry	4.3807		

Table 4. F-ratio for LPC feature under different emotional condition

	Sad	Angry	Happy
Neutral	2.0136	3.0731	2.5569
Happy	4.4580	3.7763	
Angry	4.2870		

Table 5. F-ratio for RC feature under different emotional condition

	Sad	Angry	Happy
Neutral	4.7179	6.6942	6.4411
Happy	4.7086	3.6133	
Angry	5.9334		

Table 6. F-ratio for ARC feature under different emotional condition

	Sad	Angry	Happy
Neutral	4.1796	5.6142	5.6644
Happy	4.3190	3.5750	
Angry	5.3351		

Table 7. F-ratio for LAR feature under different emotional condition

	Sad	Angry	Happy
Neutral	2.5463	4.5504	4.8143
Happy	3.8222	3.5263	
Angry	4.6514		

Table 8. Average F-ratio for all the speech features

Feature	F-ratio value
MFCC	0.9437
LPCC	3.1457
LPC	3.3608
RC	5.3574
ARC	4.7813
LAR	3.9852

From the above results, it has been observed that MFCC is most robust to the change in the emotional condition whereas RC is found to be the most sensitive to the change in emotional changes.

Cumulative Distribution function (CDF) of a real valued random variable X evaluated at *x* is the probability that X will take a value less than or equal to *x*. In this work KS test is used to determine the maximum difference between the cumulative distribution function of the feature vector. The maximum distance between the distributions serve as the test statistics.

KS test has been conducted between each pair of feature vectors extracted for the same speaker at different emotional condition. The figure given below shows the CDF plot for a speaker at different emotional conditions

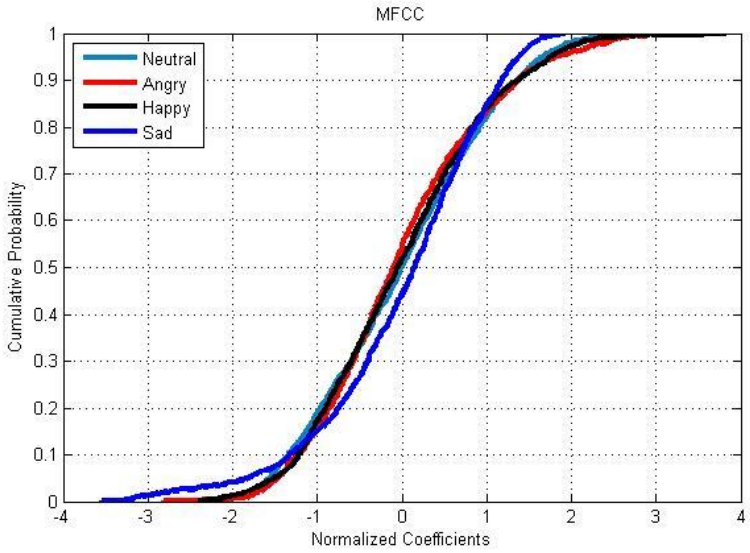

Figure 7. Maximum KS test distance for same speaker speech at neutral and different simulated emotional conditions for 1st coefficient of MFCC feature

The table given below shows the maximum distance between speech features extracted for from the same speaker at different emotional condition. It is the average of emulated as well as simulated emotions.

Table 9. Maximum distance between the MFCC features for the same speaker at different emotional conditions

Coeff.No	Neutral – Angry	Neutral-Happy	Neutral-sad	Happy-Sad	Happy-Angry	Sad-Angry	Average maximum distance between the CDFs
1	0.0292	0.0284	0.0405	0.0552	0.0337	0.0455	0.0388
2	0.0397	0.0439	0.0337	0.0527	0.0449	0.0345	0.0416
3	0.0692	0.0564	0.0564	0.0207	0.0316	0.0239	0.0430
4	0.0629	0.0619	0.0698	0.0771	0.0442	0.0775	0.0656
5	0.0518	0.0333	0.0392	0.0432	0.0575	0.0657	0.0485
6	0.0311	0.0305	0.0553	0.0466	0.0260	0.0383	0.0380
7	0.0411	0.0776	0.0340	0.0930	0.0442	0.0643	0.0590
8	0.0336	0.0508	0.0525	0.0301	0.0505	0.0473	0.0441
9	0.0426	0.0319	0.0459	0.0402	0.0470	0.0586	0.0444
10	0.0260	0.0285	0.0373	0.0447	0.0393	0.0297	0.0343
11	0.0364	0.0339	0.0483	0.0317	0.0407	0.0538	0.0408
12	0.0397	0.0358	0.0381	0.0282	0.0239	0.0290	0.0325
13	0.0513	0.0204	0.0298	0.0417	0.0386	0.0305	0.0354
	0.0427	0.0410	0.0447	0.0465	0.0402	0.0460	0.0427

Table.10. Maximum distance between the LPCC features for the same speaker at different emotional conditions

Coeff.No	Neutral – Angry	Neutral-Happy	Neutral-sad	Happy-Sad	Happy-Angry	Sad-Angry	Average maximum distance between the CDFs
1	0.0587	0.0355	0.2668	0.2535	0.0589	0.2958	0.1615
2	0.4250	0.6021	0.4957	0.8754	0.2611	0.7930	0.5754
3	0.3516	0.4516	0.5788	0.8727	0.1495	0.8270	0.5385
4	0.3167	0.4128	0.5504	0.8355	0.1768	0.7877	0.5133
5	0.3244	0.3849	0.5280	0.8095	0.1432	0.7735	0.4939
6	0.3000	0.3859	0.4599	0.7856	0.1649	0.7364	0.4721
7	0.2947	0.3410	0.4404	0.6956	0.1067	0.6759	0.4257
8	0.2599	0.3137	0.4443	0.6569	0.0835	0.6189	0.3962
9	0.1658	0.1275	0.3455	0.4526	0.0498	0.4910	0.2720
10	0.0529	0.1446	0.0441	0.1693	0.1186	0.0672	0.0995
11	0.2184	0.2613	0.5287	0.7300	0.1284	0.7244	0.4319
12	0.1660	0.1919	0.5316	0.6771	0.0982	0.6855	0.3917
13	0.0985	0.1772	0.3587	0.2872	0.0940	0.3536	0.2282
	0.2333	0.2946	0.4287	0.6231	0.1257	0.6023	0.3846

Table 11. Maximum distance between the LPC features for the same speaker at different emotional conditions

Coeff.No	Neutral – Angry	Neutral-Happy	Neutral-sad	Happy-Sad	Happy-Angry	Sad-Angry	Average maximum distance between the CDFs
1	0.1468	0.2100	0.0864	0.2584	0.0872	0.2190	0.1680
2	0.1139	0.0484	0.2528	0.2282	0.1404	0.3512	0.1892
3	0.1883	0.1037	0.4071	0.4866	0.1031	0.5409	0.3050
4	0.2193	0.1414	0.5053	0.6051	0.0922	0.6548	0.3697
5	0.2555	0.1895	0.5303	0.6640	0.0789	0.7087	0.4045
6	0.3018	0.2572	0.5430	0.7065	0.0517	0.7481	0.4347
7	0.3232	0.3340	0.4225	0.6787	0.0464	0.6924	0.4162
8	0.3761	0.4276	0.2032	0.6094	0.0619	0.5548	0.3722
9	0.3882	0.4485	0.2116	0.5243	0.0688	0.4594	0.3501
10	0.3594	0.3463	0.2858	0.3674	0.0362	0.3792	0.2957
11	0.2951	0.2660	0.3265	0.2534	0.0513	0.2735	0.2443
12	0.2078	0.1800	0.2800	0.1980	0.0635	0.1759	0.1842
13	0.1476	0.0376	0.2291	0.2129	0.1245	0.1464	0.1497
	0.2443	0.2139	0.3229	0.4257	0.0707	0.4373	0.2858

Table 12. Maximum distance between the RC features for the same speaker at different emotional conditions

Coeff.No	Neutral – Angry	Neutral-Happy	Neutral-sad	Happy-Sad	Happy-Angry	Sad-Angry	Average maximum distance between the CDFs
1	0.4284	0.6224	0.2660	0.6913	0.2941	0.5466	0.4748
2	0.0809	0.0739	0.1122	0.1641	0.0795	0.1686	0.1132
3	0.1932	0.3160	0.2407	0.4828	0.1645	0.3862	0.2972
4	0.1331	0.1004	0.2766	0.3430	0.0536	0.3629	0.2116
5	0.1824	0.0878	0.2770	0.3561	0.1124	0.4399	0.2426
6	0.3676	0.4347	0.2991	0.6193	0.0920	0.6012	0.4023
7	0.2838	0.2783	0.3022	0.5606	0.0649	0.5522	0.3403
8	0.1114	0.2210	0.0972	0.2954	0.1408	0.1981	0.1773
9	0.4470	0.6243	0.2634	0.7729	0.2609	0.6645	0.5055
10	0.1284	0.0517	0.2916	0.2921	0.0981	0.3824	0.2074
11	0.1979	0.2856	0.4069	0.1504	0.1052	0.2401	0.2310
12	0.3876	0.4672	0.3852	0.7446	0.0830	0.6878	0.4592
13	0.0914	0.1905	0.4205	0.5604	0.1214	0.4796	0.3106
	0.2333	0.2888	0.2799	0.4641	0.1285	0.4392	0.3056

Table 13. Maximum distance between the ARC features for the same speaker at different emotional conditions

Coeff.No	Neutral – Angry	Neutral-Happy	Neutral-sad	Happy-Sad	Happy-Angry	Sad-Angry	Average maximum distance between the CDFs
1	0.4729	0.6807	0.3078	0.7615	0.7615	0.6508	0.6059
2	0.1004	0.1014	0.1307	0.2257	0.2257	0.2090	0.1655
3	0.1729	0.3150	0.2369	0.4824	0.4824	0.3740	0.3439
4	0.1152	0.0818	0.2752	0.3284	0.3284	0.3482	0.2462
5	0.1831	0.0945	0.2762	0.3500	0.3500	0.4372	0.2818
6	0.3661	0.4313	0.2969	0.6160	0.6160	0.5969	0.4872
7	0.2807	0.2670	0.2990	0.5597	0.5597	0.5532	0.4199
8	0.1081	0.2223	0.0910	0.2881	0.2881	0.1895	0.1979
9	0.4442	0.6276	0.2638	0.7734	0.7734	0.6594	0.5903
10	0.1256	0.0522	0.3017	0.2968	0.2968	0.3867	0.2433
11	0.1992	0.2847	0.4067	0.1520	0.1520	0.2407	0.2392
12	0.3869	0.4614	0.3839	0.7403	0.7403	0.6857	0.5664
13	0.0915	0.1905	0.4185	0.5595	0.5595	0.4782	0.3830
	0.2344	0.2931	0.2837	0.4718	0.4718	0.4469	0.3670

Table 14. Maximum distance between the LAR features for the same speaker at different emotional conditions

Coeff.No	Neutral – Angry	Neutral-Happy	Neutral-sad	Happy-Sad	Happy-Angry	Sad-Angry	Average maximum distance between the CDFs
1	0.5180	0.7056	0.3899	0.8392	0.2988	0.7522	0.5840
2	0.1296	0.1387	0.1520	0.2812	0.0801	0.2517	0.1722
3	0.1569	0.3090	0.2170	0.4691	0.1608	0.3474	0.2767
4	0.0940	0.0568	0.2706	0.3042	0.0751	0.3368	0.1896
5	0.1806	0.0965	0.2681	0.3510	0.0967	0.4305	0.2372
6	0.3661	0.4331	0.2875	0.6073	0.0854	0.5923	0.3953
7	0.2791	0.2597	0.3074	0.5596	0.0704	0.5546	0.3385
8	0.1034	0.2168	0.0898	0.2826	0.1494	0.1708	0.1688
9	0.4403	0.6287	0.2592	0.7753	0.2616	0.6570	0.5037
10	0.1106	0.0488	0.3016	0.3013	0.0928	0.3857	0.2068
11	0.1987	0.2778	0.4050	0.1533	0.1095	0.2393	0.2306
12	0.3876	0.4625	0.3820	0.7389	0.0844	0.6841	0.4566
13	0.0890	0.1925	0.4190	0.5602	0.1252	0.4814	0.3112
	0.2349	0.2943	0.2884	0.4787	0.1300	0.4526	0.3132

The average distance for each feature type may be summarized as follows:

Table 15. Average maximum distance for each feature type

Feature	Average maximum distance between the CDFs
MFCC	0.0427
LPCC	0.3846
LPC	0.2858
RC	0.3056
ARC	0.3670
LAR	0.3132

From the above table it has been observed that MFCC is found to be the most robust feature in terms of deviation of the CDF statistics with emotional changes. LPCC is found to be the most sensitive to the change in emotional condition with change in emotion. it has been observed that MFCC is most robust to the change in the emotional condition whereas RC is found to be the most sensitive to the change in emotional changes.

Conclusion

In the present study, we have analysed the following six spectral and source features: Mel Frequency Cepstral Coefficient (MFCC), Linear predictor cepstral coefficients (LPCC), Linear Predictor coefficient (LPC), Reflection coefficient (RC), Log area ratio (LAR) and Arc-sin reflection coefficients (ARC) for their relative speaker verification performance in emotional mismatched conditions.

From the experimental results it has been observed that all the features are effected by the change in emotional condition of the speaker. However, ARC is found to be relatively robust to emotional changes. MFCC is the most robust feature in terms of deviation of the CDF statistics with emotional changes. LPCC and LPC features are found to be highly sensitive to emotional changes.

References

[1] L Rabiner and B Juang, "*Fundamentals of speech recognition",* New Jersey: Prentice Hall Inc (1993)

[2] S K Das , U Bhattacharjee and A K Mandal, "Review of Performance Factors of Emotional Speaker Recognition System: Features, Feature Extraction Approaches and Databases", *Reliability: Theory & Applications*, vol.16(3), 2021, pp. 132-148.

[3] A K Mandal, S K Das and U Bhattacharjee, "Speech recognition classifiers: a literature review", *Solid State Technology, vol* 64(2), 2021, pp. 5215-5230.

[4] K S Rao and S G Koolagudi, "*Emotion recognition using speech features",* New York: Springer Science & Business Media, (2012)

[5] J Makhoul, "Linear prediction: A tutorial review" *Proc. IEEE* 63(4) pp 561-580. 1975

[6] G S Raja and S Dandapat, "Speaker recognition under stressed condition", *Int. J. of Speech Technology, vol.* 13(3), (2010), pp. 141-161.

[7] B S Atal, J J Chang, M V Mathews and J W Tukey, "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting

technique", *The J. of the Acoustical Society of America,* vol. 63(5) (1978), pp. 1535-1555.

[8] G Fant, "*Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations",* vol. 2, (1971), ed R Jackson and C H Van Schooneveld (Paris: Mouton).

[9] D Ververidis and C Kotropoulos, "Emotional speech recognition: Resources, features, and methods", Speech *Communication, vol.* 48, (2006), pp. 1162–1181.

[10] Shriberg L D, Paul R, McSweeny J L, Klin A, Cohen D J and Volkmar F R 2001 Speech and prosody characteristics of adolescents and adults with high-functioning autism and Asperger syndrome *J. of Speech, Language, and Hearing Research* 44(5) pp 1097-1115.

[11] K S Rao, S G Koolagudi and R R Vempada, "Emotion recognition from speech using global and local prosodic features", *Int. J. of Speech Technology, vol.* 16, (2013), pp. 143-160.

[12] L Mary, "Prosodic Features for Speaker Recognition", *Forensic Speaker Recognition* ed Neustein A and Patil H (New York: Springer), (2012)

[13] P Taylor, "Analysis and synthesis of intonation using the tilt model", *J. Acoust. Soc. Am.* 107(3), (2000), pp. 1697–1714.

[14] M J Carey, E S Parris, H Lloyd-Thomas and S Bennett, "Robust prosodic features for speaker identification", *Proc. Fourth Int. Conf. on Spojen Language,* 3, (1996), pp. 1800-1803.

[15] N Dehak, P Dumouchel and P Kenny, "Modeling prosodic features with joint factor analysis for speaker verification", *IEEE Trans. on Audio, Speech, and Language Processing, vol.* 15(7), (2007), pp. 2095-2103.

[16] K Sönmez, E Shriberg, L Heck and M Weintraub, "Modeling dynamic prosodic variation for speaker verification", *Proc. 5th Int. Conf. on Spoken Language Processing*, paper 0920. (1998)

[17] R Cowie, E Douglas-Cowie, N Tsapatsoulis, S Kollias, W Fellenz and J Taylor, "Emotion recognition in human–computer interaction" , IEEE Signal Processing Magazine, vol. 18(1), (2001), pp 32–80.

[18] C Busso, S Lee and S Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection", *IEEE Trans. on Audio, Speech and Language Processing, vol.* 17(4), (2009) , pp. 582–596.

[19] L Bosch, "Emotions, speech and the ASR framework", *Speech Communication,* vol.40 (1-2), (2003), pp. 213–225.

[20] K Koike, H Suzuki and H Saito, "Prosodic parameters in emotional speech", *5th Int. Conf. on Spoken Language Processing,* pp. 679-682, (1998)

[21] D C Ambrus, "Collecting and recording of an emotional speech Database", *Technical Report, Faculty of Electrical Engineering and Computer Science, Institute of Electronics, University of Maribor*. (2000)