Indian Journal of Forensic Medicine and Pathology
Volume 14 Number 2, April - June 2021 (Special Issue)
DOI: http://dx.doi.org/10.21088/ijfmp.0974.3383.14221.44

REVIEW ARTICLE

Study on Variation in Speaker Identification under Different Conditions

Sanchita Singh[1], Suneet Kumar[2], Akabar Ali[3]

Authors' Affiliations:
[1]M Sc Forensic Science Student , [2]Assistant professor, School of Basic and Applied Sciences, Galgotias University, Greater Noida, Uttar Pradesh 201310, India. [3]Senior Scientific Assistant, Forensic Science Laboratory, Patna 800001, Bihar, India.

Corresponding Author:
Sanchita Singh, M Sc Forensic Science Student, School of Basic and Applied Sciences, Galgotias University, Greater Noida, Uttar Pradesh 201310, India.

Email:
ssanchita672@gmail.com

ABSTRACT

Voice is a fundamental way to communicate with people in a natural atmosphere where we come across many distortions. Speaker identification is a new boon in forensic science which is essential to identify a specific speaker and that a voice cannot be changed and it will prove that it belongs to a single individual. Some voices are naturally or accidentally distorted whereas some are intentionally distorted to disguise the identity of the speaker. The disguised or distorted voices give different values than the authentic ones. The voices can be accidentally disguised by natural environment, by being in a hot or cold atmosphere or deliberately by changing the accent, by keeping hand on mouth, by pulling cheeks, by creating nasal voice etc. The analysis of these voice samples is done by examining using software like Gold Wave, Praat and SSL (speech sound lab). The software help us to examine the voice samples right from extracting clue words to their spectral analysis which are known as spectrograms. Calculating the hash values of the samples provide another authentication to the original samples. Hash value is an alpha-numeric value which gives unique identity to the samples. Hash value has different algorithms but MD5(Message digest) and SHA1(Secure hash algorithm) are more reliable and secured, SHA1 being even more secured than MD5. The differences are made between the samples by looking at the pitch and intensity of the voice of the speakers. The pitch of the two voice samples of the same speaker can also be different because of the natural variation present in the speaker's voice.

KEYWORDS | gold wave, praat, spectrograms, hash value, MD5, SHA1

INTRODUCTION

VOICE THEORY: HUMANS communicate with each other through speech. Speech is produced by the movement of the lips and tongue. The air is pushed out of the lungs and the sound is made in the mouth or the throat. There are three main organs of speech in humans:

Respiratory: When we talk, air from the lungs goes up from the trachea or windpipe, then to the larynx. It has to pass two muscular folds known as vocal chords. If the vocal chords are separated, then the air from the lungs has a free passage into the pharynx and the mouth. But when these folds are not apart from each other, the air from the lungs make the folds vibrate.[2]

Phonatory: Phonation is defined as the vibration of vocal folds or vocal cords. When they vibrate, sound is produced and when they don't, sound is not produced. The vocal folds vibrate by the action of sub glottal air pressure and by Bernoulli's effect.[2]

Articulatory: Articulation occurs by

How to cite this article
Garg Divesh. Study on Variation in Speaker Identification under Different Conditions. Indian J Forensic Med Pathol. 2021;14(2 Special):322-327.

the movement of the lips and tongue touching the roof of the mouth and the pharynx. The organs involved in speech production are known as articulators. There are two types of articulators:[2]

A. Active articulators: these are in motion helping in the process of articulation.

B. Passive articulators: these remain motionless and does not take part in articulation.

Sound waves: they are also known as speech sounds which may differ from each other on the basis of:

A. Pitch
B. Loudness
C. Quality

The vowels spoken may have the same pitch when they are said in the same note, they may also have the same loudness yet they differ when one vowel is said in a higher pitch or spoken more loudly than the other.[2]

When the sound waves reach the ear of the listener, they make the eardrums to vibrate.[2] This paper is based on voice recognition of different persons.

In this paper, voice samples of female speakers have been taken to analyze the difference between the various features of voice like their pitch, intensity, formant frequencies. This experiment is done with the help of Praat, SSL and GoldWave software in which different kinds of analysis have been done. The sampling is done in such a way that both original and disguised voice samples have been recorded in two ways: direct and telephonic. This research has been done to understand the difference between original and disguised voice recordings and how speaker identification in forensic science is important in solving different types of crimes happening now a days. As voice is also a biometric identifier because, like fingerprints, retina scan, palm prints etc, each voice is unique to a single individual. Therefore, an individual's voice known as voiceprint is always different comparing to the other person. And even if one tries to conceal it in some way, it is not possible for them to withhold the real qualities and characters of their voice.

Literature review

- Annu Choudhary *et al.,* (2013) proposed a programmed discourse acknowledgment framework for disengaged and associated expressions of Hindi language by utilizing Hidden Markov Model Toolkit (HTK). Hindi words are utilized for dataset separated by MFCC and the acknowledgment framework accomplished 95% exactness in secluded words and 90% in associated words.[7]
- Kersta L.G said that experimental data encourages me to believe in the fact that voiceprint could be used to make a unique identification of a particular individual. It is my opinion that perceptible uniqueness exists in each voice and that it cannot be changed by distorting or disguising and masking it will not escape identification if the speech is comprehensible.[4]
- Hirano *et al.,* did the acoustic analysis with respect to the perturbation of pitch, amplitude perturbation and energy related to noise. The evaluation efforts on the pitch, amplitude and noise energy. It emphasizes on the measurement of hoarse voices, vocal cord vibration, unsettled noise pathologies.[6]
- Bhuta *et al.,* (2004) determined the parameters related to noise of the Multi-Dimensional Voice Program (MDVP) in relation to the perceptual rating system. This algorithm or system is used to reduce the noisy background or the voice quality of the distressed voice signal. The resulting output produces the reliable, standard, consistent and valid measure against the voice pathology. The voice turbulence and soft phonation index and noise harmonic ratio for coarseness exposure and reduction of breath in the voice sample detection improvement tools are used.[5]

METHODS

The study includes samples collected from 6 female speakers and 6 voice samples from each speaker, comprising a total 36 voice recordings analyzed separately. The literature selected for sample collection is in Hindi language. The reason behind the collection of female voices is because the females have a very high pitch and it can go up to 3000-4000 Hz also. The samples were collected in two ways:

1. Direct recording
2. Telephonic recording

The samples collected are recorded under different conditions like:

- Normal voice
- Keeping cloth on phone
- Keeping hand on mouth
- Pulling cheeks
- Pinching nose making it a nasal voice

There are three types of voice samples collected for this study:

- Normal
- Disguise 1
- Disguise 2

The tools used in the study are:

1. Mobile recorder
2. High quality head phones
3. Hash calculator
4. Gold Wave software
5. Praat software
6. Speech Sound Lab

High quality headphones are used to hear the voices clearly with minimum background noise interference.

Gold Wave software has been used to extract the cue words from the voice recordings shown as spectrums. The original voice recordings were in .mp3 format which was then converted to .wav format. It was resampled from 48000Hz to 11025Hz and the bandpass was adjusted from 200-4000db. This is done because the normal hearing range of humans is 20-20,000 Hz.

It is used to show a difference between the normal and disguised voice and to find out if they are from the same speaker or not. Gold Wave has many features including:[3]

- Real time graphic visuals like bar, waveforms, spectrograms
- Basic and advanced filters like noise reduction, volume enhancer, effects like resampling, bandpass etc.
- It can support different file formats like .mp3, .wav, Ogg, FLAC, AIFF etc
- Supports large file editing

Praat is used for the spectrographic analysis of the voice samples. Praat provides different features including:

- Pitch
- Intensity
- Resonating frequency
- Formants
- Pulses etc

It has been used to study the differences between the pitch, intensity and the formants

SL. NO.	CUE WORDS	WAVE FILE: P1 S1(A).wav			WAVE FILE: P1 S1(C).wav			WAVE FILE: P1 S1(D).wav		
		FROM (SEC:MSEC)	TO (SEC:MSEC)	DURATION	FROM (SEC:MSEC)	TO (SEC:MSEC)	DURATION	FROM (SEC:MSEC)	TO (SEC:MSEC)	DURATION
1.	nahi	0.537	0.779	0.241	1.789	2.030	0.241	0.949	1.189	0.241
2.	haar	0.848	1.020	0.172	2.089	2.261	0.172	1.275	1.447	0.172
3.	rahi	1.054	1.248	0.194	2.296	2.491	0.194	1.486	1.680	0.194
4.	nirantar	2.279	2.697	0.418	4.098	4.517	0.418	3.100	3.519	0.418
5.	jaal	5.381	5.621	0.240	7.250	7.490	0.240	6.235	6.475	0.240
6	pura	5.684	5.945	0.261	6.965	7.226	0.261	6.569	6.829	0.261
7	liya	6.324	6.578	0.254	8.298	8.552	0.254	7.149	7.402	0.254
8	kar	7.397	7.544	0.147	9.618	9.765	0.147	8.181	8.329	0.147
9	raja	7.539	7.787	0.248	9.751	9.999	0.248	8.237	8.575	0.248
10	baar	9.836	10.057	0.221	11.857	12.077	0.221	10.257	10.487	0.221
11	yaad	14.905	15.115	0.209	16.362	16.572	0.209	14.957	15.166	0.209
12	Jo	15.896	16.077	0.181	17.384	17.565	0.181	15.794	15.976	0.181
13	usse	16.101	16.331	0.230	17.550	17.780	0.230	15.954	16.185	0.230
14	bada	17.085	17.278	0.193	18.410	18.603	0.193	17.036	17.229	0.193
15	gayi	18.043	18.224	0.181	19.242	19.424	0.181	17.950	18.162	0.181
16	jab	19.015	19.184	0.169	20.181	20.349	0.169	19.006	19.175	0.169
17	nah	20.607	20.706	0.099	21.296	21.395	0.099	20.197	20.296	0.099
18	tak	21.182	21.309	0.127	22.099	22.226	0.127	21.039	21.166	0.127
19	bina	22.094	22.295	0.201	22.906	23.107	0.201	21.848	22.048	0.201
20	lagataar	22.407	27.883	0.476	23.168	23.644	0.476	22.076	22.552	0.476

DIRECT

CUE WORD	FORMANT	Sample File (Normal)	Sample File (Disguise-1)	Sample File (Disguise-2)
nahi	F1	541	548	565
	F2	**1486**	**1529**	**1162**
	F3	2055	2281	2457
	F4	2866	3060	3048
haar	F1	882	883	741
	F2	**1496**	**1486**	**1482**
	F3	2091	2402	2090
	F4	2913	3059	2998
rahi	F1	454	540	451
	F2	**1893**	**1958**	**1678**
	F3	2232	2253	2188
	F4	2741	2805	2764
nirantar	F1	483	515	490
	F2	**1646**	**1848**	**1721**
	F3	2055	2281	2457
	F4	2866	3060	3048
jaal	F1	767	713	720
	F2	1431	1477	1771
	F3	2229	2844	1790
	F4	2969	3038	3061
poora	F1	591	545	593
	F2	1141	1211	1421
	F3	2253	2249	2593
	F4	2969	3038	3209
liya	F1	586	474	443
	F2	1573	1923	1429
	F3	2325	2306	2550
	F4	3005	3012	2887
kar	F1	585	668	692
	F2	1621	1827	1766
	F3	2488	1885	2310
	F4	2845	2848	2754
raja	F1	781	668	692
	F2	1490	1513	1626
	F3	2401	2810	1785
	F4	2942	3239	2906
baar	F1	808	794	775
	F2	1512	1426	1229
	F3	2718	2779	2422
	F4	3219	3490	2680
yaad	F1	859	729	850
	F2	1585	1585	1710
	F3	2199	2148	2357
	F4	2822	2892	2619
joh	F1	543	583	525
	F2	1747	1772	1718
	F3	2317	2848	2591
	F4	2726	2952	2978
usse	F1	541	617	559
	F2	1457	1419	1235
	F3	2517	2313	2275
	F4	3014	3032	3118
bada	F1	638	677	564
	F2	1714	1677	1546
	F3	2133	2295	2285
	F4	2958	3051	2718
gayi	F1	555	449	452
	F2	1856	1988	1640
	F3	2432	2355	2632
	F4	2806	3069	3117
jab	F1	700	660	586
	F2	1877	1918	1750
	F3	2690	2686	2059
	F4	2762	3172	2886
nah	F1	821	700	762
	F2	1752	1567	1258
	F3	1911	2415	1984
	F4	2926	3122	3075
tak	F1	601	810	765
	F2	1784	1651	1614
	F3	2549	2825	2301
	F4	2887	3042	2960
bina	F1	512	667	497
	F2	1952	1430	1425
	F3	2220	2223	2386
	F4	2395	3004	2892
lagatar	F1	677	671	571
	F2	1421	1494	1104
	F3	2645	2703	2387
	F4	3019	3397	3048

TELEPHONIC

CUE WORD	FORMANT	Sample File (Normal)	Sample File (Disguise-1)	Sample File (Disguise-2)
nahi	F1	509	601	500
	F2	**1661**	**1540**	**1611**
	F3	2756	1970	2431
	F4	3126	3220	3475
haar	F1	807	716	733
	F2	**1661**	**1540**	**1611**
	F3	2756	1970	2431
	F4	3126	3220	3475
rahi	F1	446	539	504
	F2	**1845**	**1762**	**1731**
	F3	2529	2408	2271
	F4	2928	2863	3227
nirantar	F1	605	553	680
	F2	**1750**	**1674**	**1453**
	F3	2485	1809	1936
	F4	3195	3237	2861
jaal	F1	788	836	768
	F2	1686	1606	1367
	F3	2085	2186	1723
	F4	3216	3130	3003
poora	F1	424	483	495
	F2	1503	1244	1130
	F3	2364	1752	1781
	F4	3136	3105	2936
liya	F1	425	463	611
	F2	1676	1198	1221
	F3	2558	1712	1912
	F4	3085	2887	3137
kar	F1	606	640	600
	F2	1583	1637	1415
	F3	2233	2505	1846
	F4	3031	2867	2639
raja	F1	779	805	788
	F2	1704	1539	1469
	F3	3213	2229	2228
	F4	3711	2885	2912
baar	F1	980	769	836
	F2	1726	1737	1640
	F3	3009	2218	2601
	F4	3766	2534	3133
yaad	F1	936	870	775
	F2	1825	1701	1769
	F3	2406	2679	2476
	F4	3240	2831	3953
joh	F1	515	559	541
	F2	1067	1423	1228
	F3	2963	2538	2310
	F4	3007	3546	3934
usse	F1	410	492	504
	F2	1543	1567	1538
	F3	2478	3117	2623
	F4	3334	3686	3322
bada	F1	713	625	650
	F2	1564	1480	1432
	F3	2488	2314	2305
	F4	2901	3218	2926
gayi	F1	534	459	522
	F2	1834	1884	1085
	F3	2591	2355	2632
	F4	2991	2970	2783
jab	F1	552	486	545
	F2	1785	1637	835
	F3	2529	2545	1634
	F4	3140	3085	2955
nah	F1	586	702	727
	F2	1711	1549	1409
	F3	2881	2886	2783
	F4	3140	3849	3576
tak	F1	556	672	570
	F2	1683	1701	1614
	F3	3007	2951	2672
	F4	3694	3848	3960
bina	F1	290	344	722
	F2	1761	1632	1804
	F3	2826	2782	3441
	F4	3113	3555	3617
lagatar	F1	443	700	529
	F2	1506	1446	1430
	F3	3102	2599	2669
	F4	4133	3988	3847

varying even in the voice samples of the same person speaking in a normal and disguised way. This study contains four different values of the formants (F1, F2, F3, F4). The value of F2 has been considered standard.

The formant frequencies of the speaker are mentioned in the above table. There is variation in the formant frequencies because of intraspeaker variation. The formants—F1, F2 and F3—are the intensification of the frequencies in the spectrum and indicates the resonance of the vocal tract. The first two formants are ample to label the said vowel.

The spectrograms taken using spectrographic tools have been examined and the values of pitch, intensity and all the formant frequencies have been noted down. The blue line represents pitch, yellow lines indicating intensity and the red dots are the formants. The black patches indicate the vowels. Each formant frequency is set to a similar value and then the desired formant value is remarked down. This procedure is reiterated for all the four formants. This spectral analysis adds another validation to the fact that each of the 6 voice samples taken from every speaker is same for each specific speaker. The values of F3 and F4 have been found to be very high. This has been seen mostly in the vowels I and E. The value for vowel A is comparatively lower than I and E. The formant frequency of the vowel U and O is low too.

DISCUSSION

The spectral analysis was done for all the speakers and their respective voice recordings. But only the result and data of a single speaker has been mentioned in this paper. The rest of the work was done in the same way.

The concept of spectrographs depends on the Fourier theorem. Putting into practice of the technique depends on the refined use of the electronic filtering or on the approach of complex computational algorithms. The Fourier theorem affirms that any periodic waveform can be analyzed into a series of sine waves with a number of frequencies, amplitudes and phase relationships.

The most conjoint method for spectrographic filtering of the speech signal are the bandpass filters that conduct frequencies within lower and higher range of frequencies passing. The lower and higher limits of the bandpass are demarcated in those frequencies where reduction is compared to the center of the band. These are known as the cut off frequencies of the filter. Filters with narrow bandpass are lethargic or inactive in their response whereas wide band filters respond in a very swift manner. Their time resolution is quite good except for their frequency resolution, which is very poor.[3]

The spectrum of an acoustic wave is basically the result of a Fourier analysis of the waves under examination, i.e., it is a proclamation of what frequencies are present and what their amplitudes are. Each frequency component (harmonic) of the wave is represented by a line sited approximately positioned on the frequency axis. The height of each harmonic line shows its amplitude in dB.[4]

The graph is not continuous and there are no points between the harmonics. The square waves are composed of discrete frequency components. The top of the harmonic lines cannot be joined together to form a continuous and a smooth curve. The blank spaces or the blank lines imply the absence of frequencies and not the absence of any data. This type of spectrum is known as line spectrum. [11,12]

The formant frequencies of the speaker are mentioned in the above table. There is variation in the formant frequencies because of intraspeaker variation.

These are the values of pitch and intensity of all the female speakers mentioned above in the table. The pitch of the female speakers is naturally very high. The average pitch and intensities have been noted down for each voice sample of each speaker individually to show that they vary each time even in the voice samples of the same speaker due to natural variation among them.

Cue words are similar-sounding words which have been selected from all the voice recordings (normal, disguise 1, disguise 2) for one speaker and this process is repeated for the rest of the speakers. This has been done for both ways of recordings (direct and telephonic). The clue words have been selected on the basis of CV (consonant-vowel), CVC (consonant-vowel-consonant), CVVC (consonant-vowel-vowel-consonant) format. The analysis is based on picking up the same words

available in the voice sample of a speaker. This is done for every possible word found from the samples. This authenticates that the particular voice sample belongs to a particular speaker.

Forensic speaker identification is given more importance nowadays than it was earlier. It is now argued that voice print identification is as valuable as fingerprint identification. By looking at different features of speech and by analyzing it on different software, it is concluded that voice print identification is a unique and helpful research tool in cases such as ransom calls, tapped phone conversations, etc. This experiment is based on mining of cue words of the speaker of all the 6 voice samples recorded directly and telephonically for each speaker. The analysis is based on selection of the similar sounding words available in the voice samples of the speaker. To substantiate the results found from gold wave, it then has been analyzed on SSL and Praat, to get a spectrographic analysis of the recorded samples. The spectrograms show pitch, intensity and the formants clearly as different colored lines. The varying values of the pitch, intensity and the formants have been determined and it shows that they vary even for the same specific speaker, too. This article concludes that all the 6 voice recordings of a specific person has been matched and they belong to the same person. This applies for the rest of the samples recorded by different speakers. The physical parameters mentioned above in the objectives responsible for distorted voice samples are:

- Bad throat condition
- Due to cold and cough
- Stammering in speech which may be original or fear-induced
- Some people tend to talk with a nasal tone

But these are the accidental distortions caused in a voice sample, as some people try to change their voice so as not to reveal their true identity by disguising it, but however hard one may try to hide one's real voice, one cannot succeed. With the help of above-mentioned tools, the individual features and the difference between original and disguised voice is acquired. IJFMP

REFERENCES

1. ***Magdin, M., Sulka, T., Tomanová,et al.*** *(1970, January 01). Voice Analysis Using PRAAT Software and Classification of User Emotional State. Retrieved from https://www.ijimai.org/journal/bibcite/reference/2713.*
2. ***Ladefoged, P. N., & Johnson, K.*** *(2011). A course in phonetics. Boston etc.: Wadsworth.*
3. ***Dunn, H. K.*** *(1961). Methods of Measuring Vowel Formant Bandwidths. The Journal of the Acoustical Society of America, 33(12), 1737-1746. doi:10.1121/1.1908558*
4. ***Kersta, L. G.*** *(1948). Amplitude Cross-Section Representation with the Sound Spectrograph. The Journal of the Acoustical Society of America, 20(6), 796-801. doi:10.1121/1.1906439*
5. ***Bhuta, T., Patrick, L., & Garnett, J. D.*** *(2004). Perceptual evaluation of voice quality and its correlation with acoustic measurements. Journal of Voice, 18(3), 299-304. doi: 10.1016/j.jvoice.2003.12.004*
6. ***Hirano, M., Hibi, S., Yoshida, et al.*** *(1988). Acoustic Analysis of Pathological Voice: Some Results of Clinical Application. ActaOto-Laryngologica, 105(5-6), 432-438. doi:10.3109/00016488809119497*
7. ***No authors listed.****Automatic Speech Recognition System for Isolated ... (n.d.). Retrieved from https://www.academia.edu/36830074/Automatic_Speech_Recognition_System_for_Isolated_and_Connected_Words_of_Hindi_Language_By_Using_Hidden_Markov_Model_Toolkit_HTK*
8. ***Selvakumari, N.S. (n.d.)*** *Acoustic Analysis for Human Voice Disorder Classification Using Optimization And Machine Learning Technique*
9. **Li, R. J.** *(2009). Introduction. Handbook of Fourier Analysis& Its Applications. doi:10.1093/oso/9780195335927.003.0006*
10. ***Introduction.*** *(1988). Fourier Analysis, 221-225. doi:10.1017/cbo9781107049949.048*
11. ***No authors listed.*** *Line Spectrum Analysis. (2017). Statistical Signal Processing in Engineering, 347-368. doi:10.1002/9781119294016.ch16*
12. ***Tahilramani, N., & Bhatt, N.*** *(2017). Information hiding in line spectrum pair feature of non-voice part of speech signal. 2017 International Conference On Smart Technologies For Smart Nation (Smart Tech Con). doi:10.1109/smarttechcon.2017.8358353*