Sensors & Transducers, Vol. 244, Issue 5, September 2020, pp. 7-12

www.sensorsportal.com

Sensors & Transducers

Published by IFSA Publishing, S. L., 2020
http://www.sensorsportal.com

Identification of a Speaker from Characteristics of a Voice

[1] Saritha KINKIRI, [2] Basel BAKARAT, [3] Simeon KEATES

[1] University of Greenwich, Chatham, ME4 4TB, United Kingdom.
[2] Edinburgh Napier University, Scotland, EH11 4DY, United Kingdom
[3] University of Chichester, Chichester, PO169 6PE, United Kingdom
E-mail: s.kinkiri@gre.ac.uk, B.Bakarat@napier.ac.uk, s.keates@chi.ac.uk

Received: 25 September 2020 /Accepted: 28 August 2020 /Published: 30 September 2020

Abstract: Speech is unique mode of communication among humans. Speech is a complex method of communication systems when compared with other methods. As humans, we also use non-speech, which is non-verbal communication to convey information. Nonverbal communication not only accentuates on the meaning of words, but also provides information such as, what kind of emotional state the person is in. non-verbal communication provided a higher level of information, which includes characteristics of a human voice and in this paper we will show how we can use these characteristics to identify person.

Keywords: Human speech, Verbal/non-verbal communication, Emotional state and Speaker identification.

1. Introduction

Human voice is extremely difficult for a computer to analyze and recognize [2]. There are two components in the human voices: verbal and non-verbal. Human life starts with non-verbal communication with other people. On average, children under the age of two, use the production of sounds instead of words to communicate. However, people who cannot speak use nonverbal communication too. Both children and non-speaking people can communicate efficiently to share information and emotions without using words.

Verbal communication is one of the most common methods used for interpersonal communication. It uses words to convey information to others and convey information about the speaker. Verbal communication often assists with the identification of the speaker too, but not all the time. Verbal speech includes a speaker's accent, speaking style, and pronunciation, etc. [1]. Typically, individuals can identify a familiar speaker with high accuracy, but we use a combination of parameters to identify a person such as a speaker accent, speaking style, and pronunciation, etc.

Table 1. Variations of Human Speech.

Variation in Speech	Modulation
Types of Speech	Reading a book in a normal/angry mode. Giving a lecture in a classroom.
Effects of Audience	With whom we are communicating with, for example: children/parents/friends/lectures.
Environments	Noisy place such as: traffic, noisy classroom
Life Span	Age children/adults/older
Emotional state	Happy/sad/angry/excited
Types of voices	Rough/loud/soft

2. Internal Mechanics of Human Voice Production

Identifying language independent features of a voice is key to investigating the unique characteristics of a speaker's voice. To be able to identify the language independent parameters, one should understand firstly how human speech works. A voice pattern can be considered as one of the biometrics that

is unique to an individual in the same way that fingerprints, iris pattern and DNA are [6].

2.1. Production of a Human Voice

The input for human voice is air, which passes through the lungs, then through the vocal folds to produce a sound, as shown in Fig. 1. This sound is a part of the means of communication, but it does not help us understand what speaker is trying to convey. Sound is carried through the vocal tract (combination of mouth, lips and tongue), which acts as a filter, making sound understandable when it leaves the lips. The average vocal tract length for males if 17 cm and 14 cm for females.

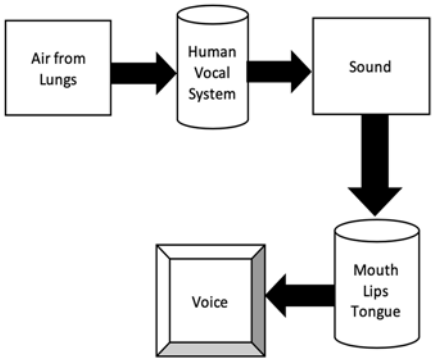

Fig. 1. Human Overall Voice Production.

2.2. Characteristics of a Human Voice

Humans can identify a speaker in a wide variety of situations. For example, imagine someone is sitting behind you. You can hear, but you cannot see, them and cannot understand what they are talking about since you do not know the language they are talking. However, you have enough data to build a picture of the speaker, which includes their gender, approximate age, and even their emotional state. The question is, what information is required to identify the speaker? To identify a speaker, one should be able to recognize the individual pattern of their voice.

There are three principal characteristics of a human voice: frequency, timbre, and volume, as shown in Fig. 2. The frequency of a voice depends on the number of vibrations of the vocal cords per second. The vocal cords of men, who are perceived to have a lower number of vibrations per second, normally operate between 100-130 vibrations per second. On the other hand, the vocal cords of women, who are perceived to have a higher number of vibrations per second, normally operate between 180-220 vibrations per second [2-3]. The second characteristic, the timbre, distinguishes sounds that have the same frequency and loudness (volume). Timbre is also called as tone colour or tone quality. For example, each musical instrument has a different timbre, which is represented by comparing harmonics that are present besides the fundamental frequency [7]. Lastly, the volume or amplitude of a voice is the vibrations that affect loudness [8]. The higher the amplitude of the vibrations, the larger the amount of energy carried by the wave \& thus the louder it is. The units of volume are measured in decibels (dB), as shown in Fig. 3. Volume relates to how the waves, produced by the vocal cords, are amplified within the body based on factors such as the speakers' mood, with whom the person is conversing, the context of the conversation, how much physical effort the person is putting in to it and so on.

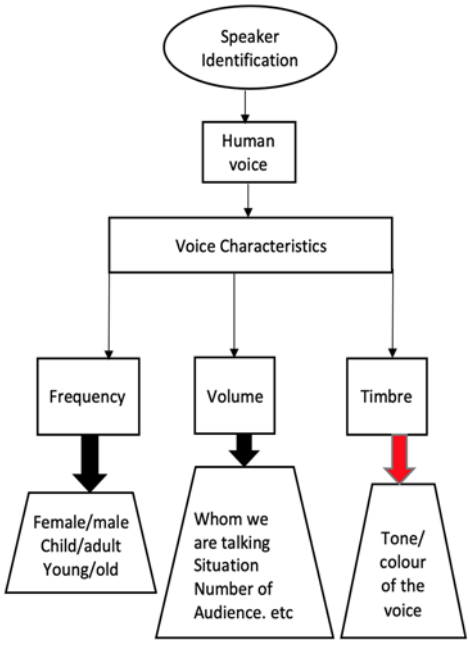

Fig. 2. Voice Characteristics.

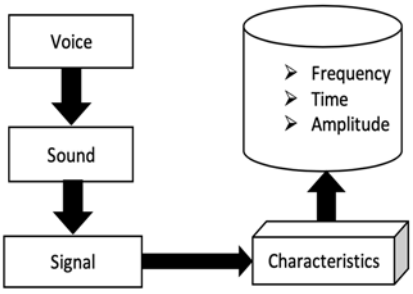

Fig. 3. Characteristics of Human Voice.

3. A Preliminary Study of Human Voice Characteristics

The experiment was conducted and 100 participants were involved; 35 female and 65 male, ages ranging from 20 to 40 years old. 30 participants are native English speakers and others are from different countries namely Egypt, India, Germany, France, Ethiopia, Saudi Arabia, Sri-Lanka etc. A script was developed for participants to read a list of sentences.

3.1. Initial Analysis

An ideal voice recognition system should aim to generate voice patterns that are independent of the language spoken. Only the participant's voice should be required to provide an input to the voice recognition system for testing and development purposes, i.e. no other constraints such as a specified language or content. A consent form was prepared for the participants, explaining the purpose of research and participants were asked to through the form before recoding was started. All participants were older than 20 years of age and understood the English language. Participants were asked to read out a pre-pared script, which consists of ten sentence that included all phonemes in the English language.

The script below shows a sample of what participants were asked to read, which was recorded for the study.

- The boys enjoyed playing dodge ball every Wednesday.
- Please give me a call in ten minutes.
- I love toast and orange juice for breakfast.
- There is heavy traffic on the highway.
- If you listen closely, you will hear the birds.
- My father is my inspiration for success.
- I will be in the office in ten minutes.
- I will go to India to meet my parents.
- Turn the music down in your headphones.
- It all happened suddenly.

3.2. Spectral Analysis

Spectral analysis transforms a sound wave into the frequency domain. The sound of a voice is created from vibrations produced by a person's vocal folds. But, voice from vocal folds needs to be filtered to be understandable. The filters in the voice production are nothing but vocal tract/resonators. The sound from the vocal folds are have to pass through by vocal tract, or else we cannot hear the sounds from the vocal colds on its own. The resonators are responsible for producing a unique voice for every individual. By applying Fast Fourier transform (FFT) to a participant's voice recording, the fundamental frequency has been observed for each participant and noted in the Table 2.

4. Potential Characteristics for Speaker Recognition

So far we have explored the possibility of recognizing a person from their fundamental frequency, but what if two participants have the same frequency range? What are the other parameters that one has to consider in order to identify a person?

Table 2. Analysis of fundamental frequency of people's voices.

Participant	Mean Freq. (Hz)	Median Freq. (Hz)	Min Freq. (Hz)	Max Freq. (Hz)
1	223.16	231	192	239
2	580.83	587	520	604
3	533	533	515	558
4	441	441	434	448
5	128.83	128.43	121	142
6	118.16	120	109	126
7	136.5	136.5	133	139
8	130.33	129	123	139
9	571.66	575	534	616
10	213.66	213.66	179	235
11	162.66	163	156	167
12	214	220	184	225
13	119.83	119.83	101	141
14	120.16	120.16	110	130
15	138.33	140	110	155
16	452	452	403	479
17	221.83	223	200	237
18	265.33	259	243	293
19	225.33	225.33	203	251
20	224.16	224.16	199	240
21	227.16	227.16	191	249
22	261.16	259	252	275
23	177.16	177.16	143	223
24	144.5	144	140	156
25	240.33	240.33	228	252
26	258.66	249	225	339
27	258.66	249	245	286
28	111	113	90	1119
29	141.33	142	119	160
30	126.83	126.83	110	142
31	335.5	335.5	311	369
32	331.66	335.16	314	378
33	376.83	376.83	330	402
34	241.5	241.5	220	261
35	251.16	251.16	227	285
36	226.83	226.83	201	256
37	224.83	224.83	191	268
38	149.5	137	113	260
39	431.33	408	403	486
40	129.66	127	139	123
41	180.66	170	142	223
42	142.66	145	109	164
43	163.66	163.66	131	198
44	247.83	247	217	288
45	166	160	149	196
46	430.66	430.66	420	440
47	518.83	520	472	552
48	545	545	504	591
49	255.33	255	247	266
50	421	453.5	421	488

4.1. Fundamental Frequency

Frequency range values have been observed from Spectral analysis. Each person has a specific frequency range for their Fundamental frequency, by looking at the frequency range, we can eliminate people whose Fundamental frequency falls outside of the any observed readings.

Minimum and maximum fundamental frequencies for all participants. For example, let say a participant frequency is 100 Hz, we can eliminate the people who does not fall under 100 Hz frequency range, with this we can eliminate on average 40 to 50 % of the population from a database as shown in Fig. 4.

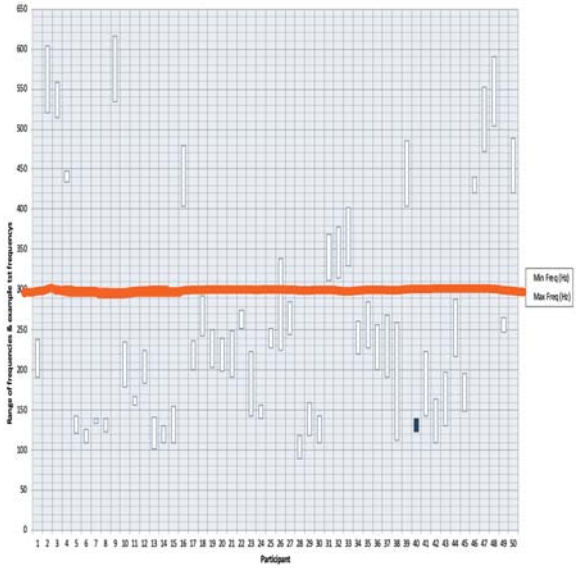

Fig. 4. Elimination of possible list of people from a database.

4.2. Speech Rate

Speech rate is another factor to be considered in order to identify a speaker. People communicate with each other at different speech rates [3]. An experiment was conducted where participants (speaking in the English language) were recorded 6 times, in a noiseless room. 100 participants were requested to read a script as mentioned in the. Their speech rate was calculated was calculated as number words per minute, as shown in Table 3.

Speech rate involves both physical and psychological characteristics of a person, such as their: gender, age, emotional state and movement of lips, and tongue etc. [9]. Speakers can change their speaking rate, if they would like to do so. However, changes of speech rate can happen without a speaker's knowledge, because speakers cannot always control the way they are speaking. The following factors impact speech rate of a speaker and perception of a listener as shown in Fig. 5.

4.2.1. Natural (Relaxed) Speaking Rate

This is the rate of speech that people use to communicate with their family, close friends and people whom they spend more time with. Culture plays an important role and it is where a person's natural speaking rate develops. Even geographical locations can have a major impact on speaking rate.

Table 3. Participants speech rate.

Participant	Words per minute			
	Min SR	Max SR	Mean SR	Median SR
1	98	110	104	103
2	110	120	113.83	113
3	106	118	113	113
4	118	134	126.83	127.5
5	110	135	121.33	120
6	92	100	96.66	97
7	110	135	121.11	120
8	126	135	129.83	128.5
9	100	120	112.5	112.5
10	140	142	140.83	140.5
11	110	135	111.83	111
12	108	120	113.83	112.5
13	106	118	111.66	111
14	125	134	129.83	129
15	110	135	129.5	128.5
16	126	134	129.5	128.5
17	145	150	148.83	150
18	135	140	137.16	136.5
19	125	126	125.16	125
20	115	130	121.33	120
21	90	95	92	91
22	135	138	136.66	136.5
23	126	150	144.83	149
24	128	132	129.66	130
25	140	140	140	140
26	90	98	93.33	93.5
27	115	120	117.66	116.5
28	128	132	130	130
29	100	106	101.83	100
30	110	115	111.83	111.5
31	140	145	141.5	141.5
32	124	135	129.83	129.5
33	120	140	128.83	127.5
34	90	100	94.66	95
35	110	140	121.16	117.5
36	100	105	101.5	101
37	145	150	148.16	149.5
38	110	118	114.16	115
39	130	140	136.66	137.5
40	125	135	129.33	128.5
41	120	133	127.66	129
42	100	140	110	100
43	124	129	125.5	125
44	130	138	132.83	133
45	125	130	127.83	128.5
46	130	137	132.5	131.5
47	140	143	141.5	141.5
48	90	96	93.16	93.5
49	121	130	125.33	125
50	110	120	115.33	115

4.2.2. Impact of Behavior

The most common impact of behavior on speech rate is when strangers communicate to each other. Individuals present emotions such as nervousness, and reluctance when they converse with unfamiliar people. For example, presenting in front of an audience for the

first time is always nerve wrecking, causing speech rate to be faster or slower rate than usual.

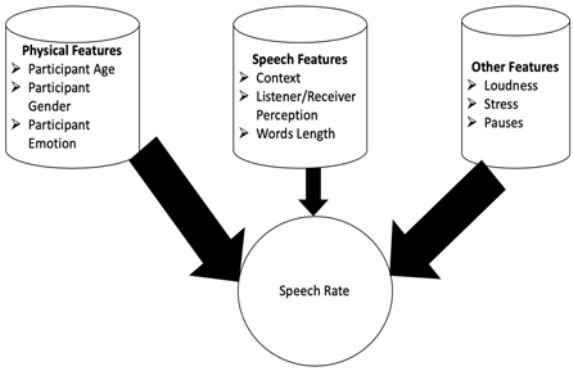

Fig. 5. Causes of variations of speech rate.

4.2.3. At a Work Place

Work places usually involve working at a fast pace to produce quicker results, which undoubtedly causes stress. If a person is unable to work under pressure, they might be mentally processing information at a slower rate, which can cause them to talk slowly too, thus, reducing their speaking rate. However, they may equally talk faster as a result of increased alternative levels such as talk faster in order to keep up with the fast work pace.

4.2.4. Speeches

During a speech in front of an audience, the speaker would normally take more pauses than usual in order to gain maximum attention from the listeners, also to allow them time to pick and choose their words carefully. This is usually beneficial to convey a message or gain support from the audience. Such practices are most commonly seen by leaders, politicians, and professional speakers.

4.2.5. Emergency Situation

People will talk faster when they are in an emergency so that they can convey their problems to the listener as soon as possible. Normally, this is observed in situations where help is required such as an emergency call for an ambulance, or the police.

4.2.6. Contexts of a Speech

Sometimes people either talk slowly or quickly, depending upon their knowledge of what they are talking about. If their understanding of the subject is clear and thorough, they might talk comparatively faster than normal. Equally, if they are unsure, they may talk more slowly.

4.2.7. Vocabulary

Generally, if sentences have longer words and are difficult to pronounce, speakers take a longer time than normal to finish their sentences.

4.3. Articulation Rate

Articulation Rate (AR) is defined as, the number of speech units delivered per second. The speech units can be syllables or words. AR is similar to SR, but the main difference between them is that, SR includes pauses, whereas, AR does not. The speed of speaking can either be defined as SR or AR as "speech rate". AR and SR depend on the continuity of the speech. For example, both SR & AR can be fast speaking rates when speech is fluent.

In summary, AR & SR are the same if the speaker has no pauses at all. SR is the mean of the words or syllables per minute including pauses. AR is the mean of words or syllables per minute recording between pauses. Thus AR will always greater than SR.

4.4. Accent

Accent is one of the keys to human speech to identify their locality. An accent provides various details about a speaker, such as an ethnicity, social status, and their first language.

People often mimic the other person's accent subconsciously when they are conversing. Everyone has an accent in their speech community, and some words are more pronounced than others. Hence, we may be able to use an accent to identify a speaker, however, it will be difficult if all the speakers are from the same locality.

4.5. Pause

During speech, there are two types of pauses: intentional (conscious) pauses and Natural (unconscious) pauses. Intentional pauses may occur when:

- While giving a presentation, we make sure to give a pause in between our words to ensure that the, audience is listening & make our speech clearer.
- While discussing a project topic with our supervisor/project leader.
- Trying to choose words more carefully.

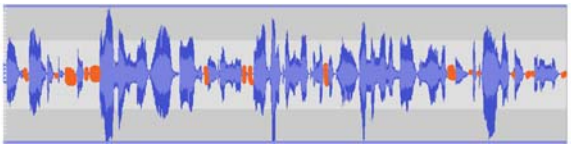

Fig. 6. Pause by a user.

Natural pauses will occur without our knowledge or consciousness, in situations such as:

- While talking in our first language.
- Relaxed conversations with parents/family/friends.
- When giving a speech on something you are very well-versed in.

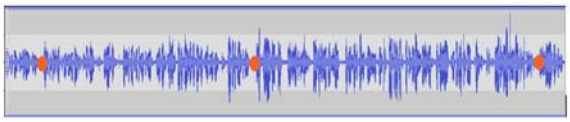

Fig. 7. Less pauses by a user.

4.6. Speech Variation

Human voices change over time, from birth through puberty & into old age. For example, children/infants sound different as compared with adults. Voices sometimes change during day and night. Recordings of people's voices in the morning and evening time show their relative amplitude and frequency values changed based on various reasons.

- Participants are more active in the morning and they became tired by the night because of work during the daytime.
- Some participants were active in the evening since they were about to go home.
- Some participants sounded the same during morning and evening hours.

5. Results

A speaker's voice varies based on several factors and situations. However, there is a list of parameters that can be used to identify a speaker. Frequency of the highest peak is one of the parameters used to identify a speaker. Female and male participants typically have different frequency range. There are two ways of identifying a speaker based on frequency values. Firstly, one has to decide whether a speaker is female or male. Secondly, comparing the frequency value of a person with all the participants. And, finally, eliminating the ones which do not match.

6. Conclusions

In the candidate list, initially frequency parameters can be used to delete people who cannot be the speaker, then the next parameter we could use would be speech rate that is participants can be eliminated based on the number of words spoken per minute and both minimum and maximum words per minute can be compared. Those who do not meet these parameters will be eliminated and the list of people remaining to be compared will be narrower. Next, the accent, pronunciation and repetitively used words (such as some people have a tendency to use some words very often) will be compared.

Acknowledgements

I would like say thanks to my friend Arwa for her support and feedback.

References

[1]. Z. Win Aung, A Robust Speaker Identification System, *International Journal of Trend in Scientific Research and Development*, Vol. 2, Issue 5, August 2018, pp. 2057-2064.

[2]. B. Monson Brian, J. Hunter Eric, J. Lotto Andrew, H. Story Brad, The perceptual significance of high-frequency energy in the human voice, *Journal of Frontiers in Psychology*, Vol. 4, 2014, pp. 587-597.

[3]. B. Shuren, Extraction of Instantaneous Frequency Characteristic Using Time-frequency Ridges, *Chinese Journal of Mechanical Engineering*, Vol. 10, 2008.

[4]. T. H. Kinnunen, Optimizing spectral feature based text-independent speaker recognition, *University of Joensuu*, 2005.

[5]. D. George, J. Hawkins, A hierarchical Bayesian model of invariant pattern recognition in the visual cortex, in *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)*, 2005.

[6]. P. Saini, P. Rao, Multimodal Biometrics Security: A Review, *International Journal of Innovative Research in Engineering and Multidisciplinary Physical Sciences*, Vol. 6, Issue 1, January-February 2018.

[7]. Acoustical Terminology, American Standard Association, NY, 1960.

[8]. R. Plomp, The Intelligent Ear: On the Nature of Sound Perception, 1st ed. Book, *New York: Psychology Press*, 2001.

[9]. D. R. Beukelman, P. Mirenda, Augmentative and alternative communication, *Paul H. Brookes*, Baltimore, 1998.

[10]. M. S. Hawley, S. P. Cunningham, P. D. Green, P. Enderby, R. Palmer, S. Sehgal, P. O'Neill, A Voice-Input Voice-Output Communication Aid for People with Severe Speech Impairment, *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 21, Issue 1, 2013.