Phonemes: An Explanatory Study Applied to Identify a Speaker

Saritha Kinkiri[1], Basel Barakat[2] and Simeon Keates[2]

[1] University of Greenwich, Chatham ME71PQ, UK
[2] Edinburgh Napier University, Edinburgh, EH11 4DY, UK
s.kinkiri@gre.ac.uk
B.N.Barakat@gre.ac.uk
s.keates@napier.ac.uk

Abstract. Speaker Identification (SI) is a process of identifying a speaker automatically via a machine using the speaker's voice. In SI, one speaker's voice is compared with *n*- number of speakers' templates within the reference database to find the best match among the potential speakers. Speakers are capable of changing their voice, though, such as their accent, which makes is more challenging to identify who is talking. In this paper, we extracted phonemes from a speaker's voice recording and investigated the associated frequencies and amplitudes to be assist in identifying the person who is speaking. This paper demonstrates the importance of phonemes in both speech and voice recognition systems. The results demonstrate that we can use phonemes to help the machine identify a particular speaker, however, phonemes get better accuracy in speech recognition than speaker identification.

Keywords: Accent, Human Speech, Phonemes and Speaker Identification.

1 Introduction

Speaker recognition is used to identify an individual person who is speaking, independent of what has been said. The production of speech involves the brain, vocal cords, lips, tongue, lungs, diaphragm, mouth and nasal/sinus cavities. The two steps in speaker recognition are perception and recognition. The brain receives a sound wave principally through the ears. The wave is transformed into electrical nerve impulses in the cochlea and those impulses are sent to the brain for processing and recognition.

Digital systems need to be given training on speech samples to identify a speaker. These speech samples are collected from each person speaking through a microphone and processed by a processor to recognise the voice/speech. Voice characteristics include both physical and behavioral components. The shape of the vocal tract is fundamental in the physiological component. The vocal tract is made up of the mouth, tongue, jaw, pharynx and larynx which articulate and control speech production by manipulating the airflow generated by the lungs and diaphragm. The behavioral component comprises emotion, accents, rate of speech and pronunciation. Some elements of speech, such as the ability to roll the letter 'r,' are controlled genetically.

Human speech conveys two levels of information [12]. At the primary level, speech signal conveys the words being spoken by a user, which helps us to recognise a user's pronunciation, accent, age and language. On secondary level, the signal conveys information that can identify a speaker on more fundamental characteristics rather than

what has been said. Humans are generally good at identifying a speaker in very limited time by listening to their speech/voice [1] especially if the speaker is familiar to them. However, even when the speaker is not known to the listener, it is still possible to leanr a lot about the person from how they speak. For example, if you are in a flight and some people are sitting behind you and they started talking with people who are sitting beside them; by listening to their speech, you would be able to identify the gender, predict their age, emotion and accent (if you are familiar with their accent), even though if you may have not seen them before.

Humans can even identify intent by listening to a sound that does not have any obvious semantic meaning. For example, parents of young children can often understand what the child or infant wants, irrespective of the fact that the child is not using proper words, or merely making sounds to indicate what they need or how they feel. Speech conveys different types of information such as message, language information, emotional and physiological characteristics [3] [10].

Machines can process audio signals in real-time such as speech recognizers e.g.: Siri and Alexa. However, it is difficult for a machine to distinguish sounds from different resources such as music, human voice, animal sound etc. as humans do. Thus, to make an algorithm that can identify the speaker, it is important to understand the components of the human voice. Current speaker identification systems extract short-term acoustic features from a human speech [2], as shown in Figure 1.

In this paper, we are investigating the differences in the frequencies of phonemes. Hence, we conducted an experiment, which includes collecting voice samples from ten participants and extracted phonemes. This paper is organized as follows, section 2 presents a brief overview of the background of the speaker identification system, followed by results, discussion and given the conclusion of using phonemes to identify a speaker.

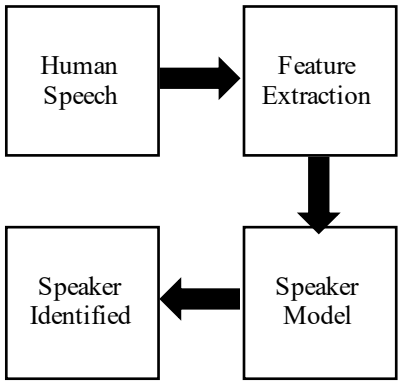

Fig. 1. Speaker Identification System

2 Background

The sounds of human speech are complex and have been studied for centuries and are still being researched [e.g., 13, 14]. Research suggests that phonetics has always been an important part of sound production. Phonetics is derived from a Greek word, *'phōnētikós'*; phone means a sound or voice. The small units of sounds are called phonemes, with each language having their own phonetic set. Phonetics have played the main role in learning and understanding a language rather than identifying a speaker. There are 20 letters that are considered to be "voiced," which, in English, include consonants B, D, G, J, L, M, N, NG, R, SZ, TH, V, W, Y, Z and vowels A, E, I, O and U. There are 8 "unvoiced" sounds: CH, F, K, P, S, SH, T and TH [4, 5].

There are three types of phonetics: acoustic, auditory and articulatory phonetics [5]. Acoustic phonetics is the physical property of the sounds of a language; that is the volume of sound, frequency of the sound waves, frequency of vibrations, etc. Auditory phonetics is focused on how speakers perceive the sounds of a language, with the help of the ears and the brain. Articulatory phonetics conveys how the vocal tract produces the sounds of a language that is, with the help of moving parts of our mouth and throat, also known as the articulators [5, 7]. Phonetics helps when learning and distinguishing within a language, or between multiple languages. By uttering a sequence of discrete sounds (or phonemes) with the help of our articulators, words are composed [8, 11]. A combination of coherent words leads to a sentence. Phonemes are discrete or different sounds within a particular language, but make up the building blocks of all speech. Thus, all words and sentences are ultimately collections of phonemes.

Feature extraction plays a crucial part in speech processing. Features should provide the necessary information to be able to identify a speaker. There are numerous feature extraction methods are available such as: Linear Predictive Codes (LPC), Perceptual Linear Prediction (PLP), Mel Frequency Cepstral Coefficients (MFCC), PLP-RASTA (PLP-Relative Spectra) etc. The most popular feature extraction method is MFCC, but extraction features would be difficult when speaker changes their voice such as: their emotional state, context, with whom they are talking etc. MFCC does not provide enough resolution in frequency regions and moreover, signal can not be reverted from frequency analysis by using MFCC [14].

Phonemes differ across languages; the frequency of the sounds varies in which they occur in words. Some phonemes may not be considered as phonemes in other languages. For example, the Chinese language is tonal, and sounds come from nasal cavities when compared with English [8]. The features will vary while patterns of sound also differ significantly in different languages. The fundamental frequency of "r" is the same for two British speakers. On the other hand, the way of pronouncing r can be used to distinguish between British and non-native speakers.

3 Methodology

In this paper, we extracted a phonemes from human speech. Each phoneme's amplitude and frequency values were measured and evaluated. The participant's task was to read

a script provided by the researcher and this took place in a silent/quiet room allocated especially for the research. The script was designed so that it could be read easily by all participants and prevented the use of foul language as well. The researcher will offer participants the option to do some trial recordings before the actual recording to allow the participant to become comfortable with the process. The equipment used in this research includes an audio recorder which in itself was not harmful, nevertheless, the researcher gave clear instructions on how to use the recording equipment (e.g. distance to the microphone) before the start of the recording.

In the case that the participant becomes anxious, the researcher would remind the participant that they are not obliged to take part in the study. Since one of the recording locations was enclosed (anechoic chamber), there was a chance that a participant did not want to record their voice in this location. The lead researcher was available in the anechoic chamber to calm the participants down if they were to appear to become anxious since it was an enclosed space. If the participant was still uncomfortable to do the recordings in that environment, then an alternative space could have been used. The alternative space would be outside of the anechoic chamber or any of the classrooms on campus. It was explained to the participants how his/her data would be used and handled in the project before the task started. The participants were given a choice to not take part if they decided to do so. The participants were given the option to leave the study at any point of the research. Assuming they gave their consent, their recorded voices were added anonymously to the database. Since the recorded voices could be used as a biometric identification means, there could be a consequent potential security risk. However, as all data was anonymised before storage and usage this risk was minimised as there was no personal ID linked with the recordings. All data was stored safely and will be deleted once the project is completed. The following were used for the recording as shown in Table 1.

Table 1. The Equipment used in this Research

Number of Participants	10
Recording Place	Anechoic Chamber, Nelson building
Recording Equipment	Scarlett 2i2 studio, MacBook
Software	Audacity
Programming Language	Python
Headphone/headset	Participant choice/option

4 Experiment

The speech was recorded from ten participants reading a script. Participants were asked to read the script, which comprised of ten sentences, which are shown below. They

aimed to cover the main phonemes used within the English language (there are 44 in total), these sentences are:

1. The boys enjoyed playing dodgeball every Wednesday.
2. Please give me a call in ten minutes.
3. I love toast and orange juice for breakfast.
4. There is heavy traffic on the highway.
5. If you listen closely, you will hear the birds.
6. My father is my inspiration for success.
7. I will be in the office in 10 minutes.
8. I will go to India to meet my parents.
9. Turn the music down in your headphones.
10. It all happened suddenly.

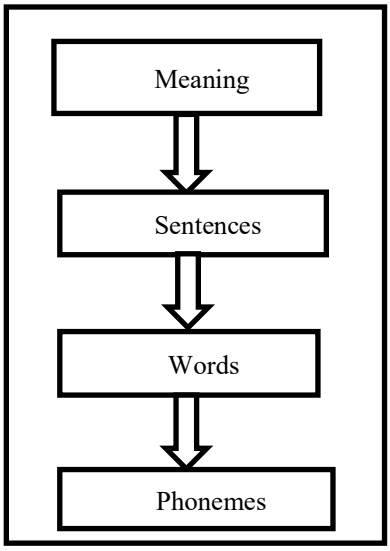

Fig. 2. Description of Phonemes

After recording the voices of all participants, the next step was to extract phonemes from a script. There is no software available to do extract phonemes from a speech, so it was extracted manually. To observe how the frequency and relative amplitude values changed for a specific phoneme a Fast Fourier Transform (FFT), was applied to the voice signal to observe the frequency spectrum. The FFT was applied to phonemes of all 10 participants.

5 Measuring Frequency Spectrum of Phonemes

Phonemes play an important role in human's speech. Phonemes help us to recognise the sound, as soon one's heard a sound. For example, when one speaks/say, "Hello". What is the first sound that comes first in the human's brain? The sound which comes as "/h". Identification of a phoneme helps to identify a common sound in different

words. For example, when some say: boys, breakfast, birds and ball. The first sound one can hear is "/b".

In this experiment, the frequency of phonemes in words has been observed at various points such as the position of a phoneme in several words. In this experiment, participants were asked to read a list of sentences. Participant read a bunch of sentences from a script which consists of several words, and then words have the same phoneme in different positions. For example, phoneme 'B' was read in different words by the same participant. However, their frequency and amplitudes values were changed drastically for a few participants which were noted in Table1. The hypothesis of this experiment was, phonemes would be individual to a speaker, then one can use phonemes to identify a speaker-independent of a language. But then, once experiment was conducted and results were observed, one can use phonemes to identify a speaker but with some boundaries, as explained in the conclusion.

6 Results

The voiced phonemes are extracted from participants and FFT is applied to observe how relative amplitude and frequency values of a phoneme vary for different words from the same participant. The highest peak of the frequency does not change. Each phoneme represents a different visual representation of the phonemes of a participant. Once the voiced phonemes of one participant are compared with another participant, it is observed that some phonemes are very similar to others and some of them are very distinctive. The frequency and relative amplitude values were derived and recorded, from each phoneme.

Next, voiceless phonemes of all participants are extracted to find out if there is any consistency, sufficient enough to identify a speaker. Surprisingly, voiceless phonemes of some participants are very distinctive to recognize a person. FFT graphs are prepared for both phonemes of all participants and voiced versus voiceless phonemes is compared to draw a conclusion. Moving forward, voiced and voiceless phonemes of all participants are compared. Participants voices are used as an initial data set and only their phonemes are extracted. Participant 1 and participant 3 have the same similarity, when they pronounce the letter "P" as shown in fig 3 and 4; on the other hand, participant 6 and participant 9 have 100% of similarity of producing phonemes "r". Participant 1 is similar to participant 4 when pronouncing the phoneme "th". Lastly, participant 5 is the only one with a distinctive pronunciation of the phoneme "SZ".

There are several factors, which make a phoneme sound different and represent different relative amplitude and frequency values. Another phenomenon that will be questioned is how easy it is for a participant to pronounce a phoneme?

Phonemes are extracted for ten participants. For each participant, a boundary is set up for dominant frequency, independent of phonemes, meaning he/she can say any phoneme but, the dominant frequency should lie between the range. For several phonemes, like Sz, Y, P, W and TH, the dominant frequency lies between 50 to 200 Hz. For the phoneme T and V, the dominant frequency lies between 200 to 285 Hz. It was observed that the dominant frequency of phoneme "V" of participant 1 and 2 are the same. The

frequency of “SZ” of participant 1 and 2 are same but differs in amplitude values. The highest peak of participant 1 of “Sz” is same as “w” of participant 2.

In our daily conversation, as a listener we recognize/concentrate words to understand the meaning, which helps us, communicate with each other. Hence, it is not practical to use frequencies to differentiate the phonemes. For example, if someone is continuously saying b, b, b, … several times and say p 20 times in between and then continue saying b, we don’t recognize the ‘p’ and perceive as if the participant said b only.

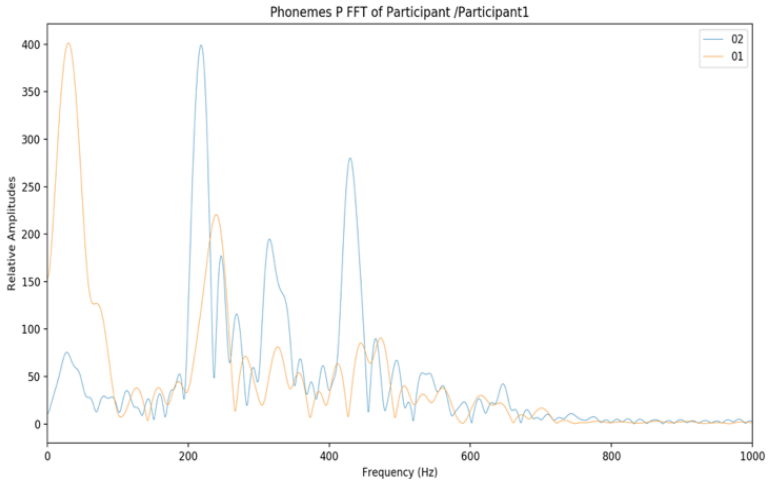

Fig. 3. Spectrograph of Phoneme ‘p’ of a participant1

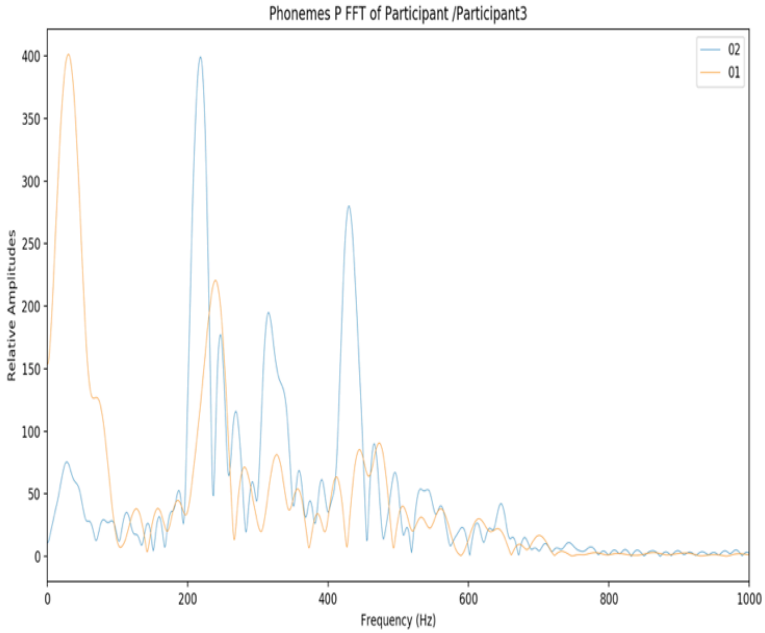

Fig.4. Spectrograph of phoneme ‘p’ of a participant3

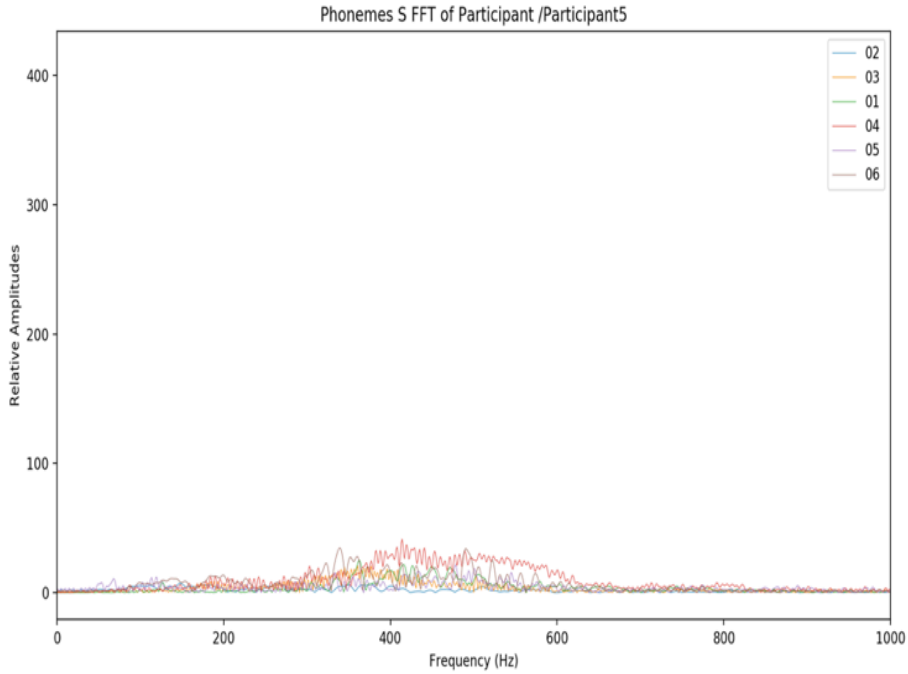

Fig. 5. Spectrograph of phoneme ‘S’ of a participant5

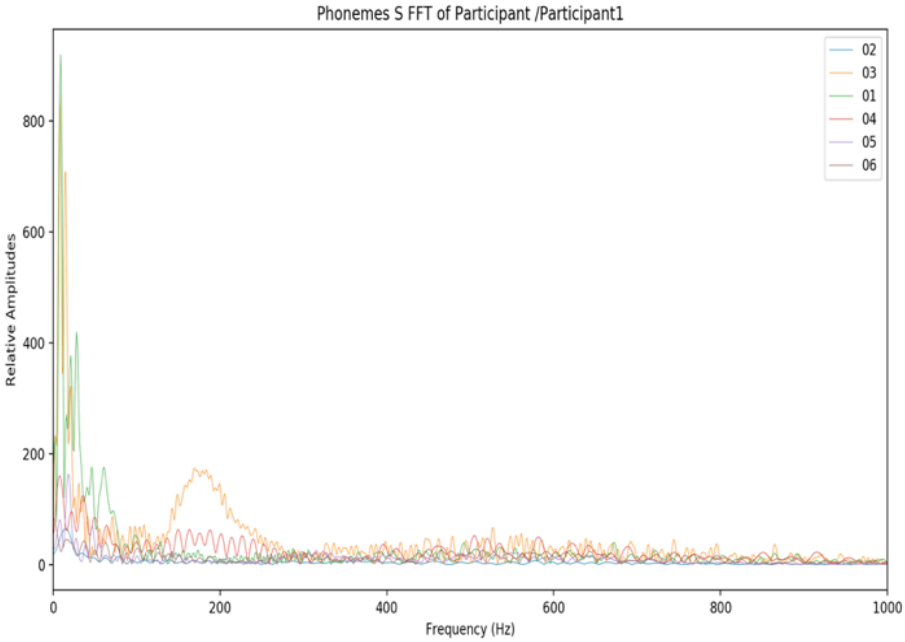

Fig.6. Spectrograph of phoneme ‘S’ of a participant1

7 Discussion

Several questions arise about the recording of the above data. These questions include: why phonemes are changing even though the same person is speaking? Why some phonemes are distinctive to a participant and a few of them are very similar between different participants?

There are several limitations of using phonemes as a fundamental factor that affects voice recognition: Phonemes produce different sounds because of the exaltation of air from our mouth. It is hard to keep track of how these different sounds are produced, as it is dependent on many factors such as how much air is exhaled whilst speaking, how big the vocal cord is open, the shape of lips, placement of tongue etc.

Some of the other factors included are the actual placement of the phoneme in a word; emotions can alter the phonetic emphasis on a word and context of the word (paint and pain/ sell and cell). Phonemes are good enough to identify their origin, but not consistent to identify a person. Even in the linguistics, the aim of the listener is not to concentrate on individual phoneme, but to understand the meaning of the words/sentences. It is difficult to extract phonemes from a voice signal manually.

The same sound may be represented by different letters or combination of letters. One should be knowledge of phonemes completely or can use of IPA to find out the phonemes in a word. The same letter produces a different sound. Different combination of letters represents a single sound. Some letters do not even produce a sound. There is no letter but still represents a sound. Phonemes change their frequency based on their place, that is in the starting/middle/end of the sentences. The spectral analysis showed that, participant information is non-uniformly distributed. Some of the frequency domains clearly showing the differences to be able to identify a speaker. However, the problem is, how one can decide the frequency bands for individual when other participants also have the same differences, for example phoneme ‘p’ is exactly same as shown in Figure 3 and 4.

8 Conclusion

Nowadays, most people tend to go abroad to pursue their higher studies or for their dream job. One tends to learn or adopt a foreign language in terms of accent and pronunciation. However, some individuals pronounce certain words in a unique style, which helps identify their origin. For instance, the emphasis on a certain letter of a word is different in different accents like ‘water’ in British English, has the ‘t’ silent when pronounced, whereas, in an Indian accent that “ter” in ‘water’ is pronounced as turr, with an emphasis on the “r”. Production of sounds in the vocal tract during speech describes and characterizes the sounds. There are two types of sounds, voiced and unvoiced/voiceless. A voiced sound will produce vibrations in the vocal cord as compared to unvoiced sounds. Unvoiced sounds produce no vibrations in the vocal cord but still generate sounds through the mouth and lips.

Participants can adjust the boundaries of a phonemes' frequency based on the context. For example, the participant will learn how to say words in different ways. There are numerous papers focused on how phonemes are used to identify a person, but it is only available for a few languages [4] [5] [11]. This is mainly because they have their own systems because of know which phonemes are used very often. Phonemes mainly arise from a language perceptive. As humans, we don't listen to phonemes on its own, however, we do listen to complete phonemes to understand the language, but not to identify a speaker. Language carries information from a human speech, by using words.

Changes in the position of a phoneme create a lot of difference that would reflect a different pattern of a human speech. Then it would become more difficult to identify a speaker. Moreover, English is not a phonetical language. In the English language, one phoneme can be represented by using other letters. For example, \k\ in Cat, kite, KitKat. \k\ is represented by using 'c'.

Participants have used knowledge of phonemes from their original language that helps us identify which country they belong to. It is difficult to extract a phoneme, if you don't know observe/listen what has been said. For example, /p/ in cap and /b/ in cab. If system is trained based on phonemes only, without context/situation system cannot figure out which phoneme is pronounced. The dominant frequency of phoneme 'p' did not change for all ten participants as shown in Figure 6.

Phoneme "B" of participant 1 and 2 were extracted from word: boys, ball, breakfast and birds. The frequency values were varied, even though same person is speaking the same phoneme shown in figure 6 and values were noted in the Table 2.

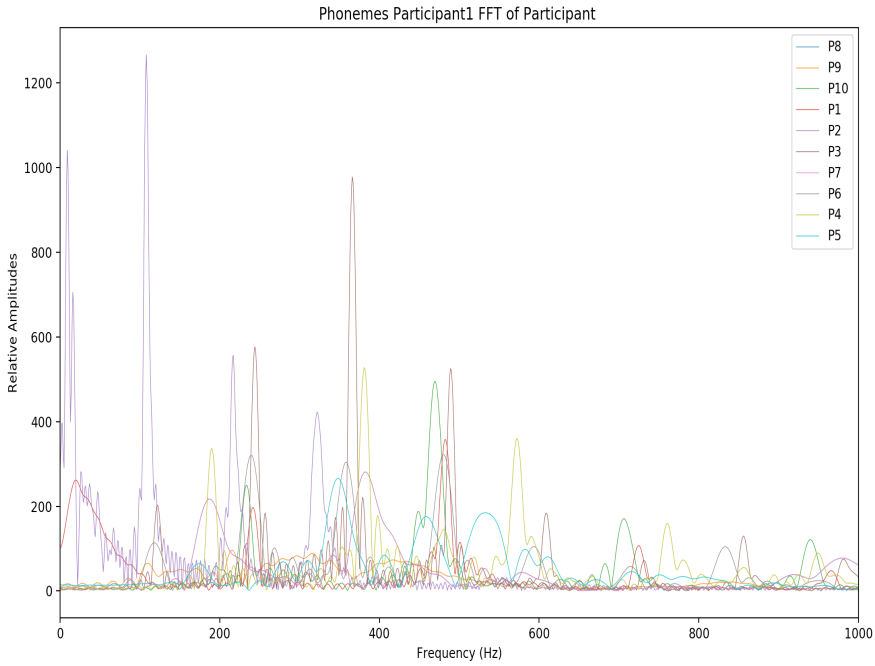

Fig.7: Spectrograph of phoneme 'B' of 10 participants

Table 2. Frequency of Phoneme "B".

Phoneme "B"	Boys	Ball	Breakfast	Birds	Mean	Median
Partici-pant 1	248	528	482	242	375	365
Partici-pant 2	100	32	108	32	68	66
Partici-pant 3	409	343	366	551	417.25	387.5
Partici-pant 4	335	345	381	472	383.25	363
Partici-pant 5	336	342	480	285	360.75	339
Partici-pant 6	449	370	348	336	375.75	359
Partici-pant 7	222	236	280	240	246.5	242
Partici-pant 8	280	450	320	440	372.5	380
Partici-pant 9	270	150	320	440	247.5	260
Partici-pant 10	320	387	420	450	394.25	403.5

When we consider the Chinese language, it is a tonal language. The way of expressing phonemes would be different to convey the message/information. After observing the data, it is concluded that phonemes will not help us identify a speaker, but instead help us find out their nativity. Phonemes can play an important role in the linguistic theory of speech. One of the main problems with phonemes is that participants had an influence from their native language on the other familiar language (English). Participants pronounced differently or mispronounced, phonemes in words. They tend to use

their native language phonology skill on other languages that helped us recognise their nativity. If would be helpful to understand the language, so we can use it in speech recognition and language identification. we can presume which language a person is speaking and/or what is their origin (for example, Indians, British and so.).

References

1. M. Bazyar and R. Sudirman, "A new speaker change detection method in a speaker identification system for two-speakers segmentation," *2014 IEEE Symposium on Computer Applications and Industrial Electronics (ISCAIE)*, Penang, 2014, pp. 141-145 .
2. F.R Chowdhury, S. Selouani and D. O'Shaughnessy, "Distributed automatic text-independent speaker identification using GMM-UBM speaker models," 2009 Canadian conference on Electrical and Computer Engineering, St. John's, NL, 2009, pp. 372-375.
3. B. G. Nagaraja and H.S. Jayanna, "Efficient window for monolingual and cross lingual speaker identification using MFCC," 2013 International Conference on Advanced computing and communication systems, Coimbatore, 2013, pp. 1-4.
4. Al-Hattami, Abdulghani. (2010). A Phonetic and Phonological Study of the Consonants of English and Arabic. Language in India. 10. Pp. 242-365.
5. S. Bacha, R. Ghozi, M. Jaidane and N. Gouider-Khoujia, "Arabic Adaption of Phonology and Memory test using entropy-based analysis of word complexity." 2012 11th International Conference on Information Science, Signal Processing and their Applications, (ISSPA), Montreal, QC, 2012, pp. 672-677.
6. G. H Ngo, M. Nguyen and N. F. Chen, "Phonology-Augmented Statistical Framework for Machine Transliteration Using Limited Linguistic Resources," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27. no. 1, pp. 192-211, Jan. 2019.
7. Shih, Stephanie S, Inkelas and Sharon, "Auto segmental Aims in Surface-Optimizing Phonology," 2018 Linguistic Journal, pp. 137 -196.
8. F. R. Chowdary, S. Selouani and D. O'Shaughnessy, "Distributed automatic text independent speaker identification using GMM-UBM speaker models," 2009 Canadian Conference on Electrical and Computer Engineering, St. John's, NL, 2009, pp. 372-375.
9. N. Uma Maheswari, A. P. Kabilan and R. Venkatesh, "Speaker independent speech Recognition system based on phoneme identification," 2008 International Conference on Computing, Communication and Networking, St. Thomas, VI, 2008, pp. 1-6.
10. R. A. Rashid, N. H. Mahalin, M. A. Sarijari and A. A. Abdul Aziz, " Security system using biometric technology: Design and implementation of Voice Recognition System (VRS)," 2008 International conference on Computer and communication Engineering, Kuala Lumpur, 2008, pp. 898-902.
11. Akhila K S and R. Kumaraswamy, "Comparative analysis of Kannada phoneme recognition using different classifies," 2015 International Conference on Trends in automation, communications and computing Technology (I-TACT-15), Banglore, 2015, pp. 1-6.
12. S. P. Panda, "Automated speech recognition system in advancement of human-computer interaction," 2017 International Conference on computing Methodologies and Communication (ICCMC), Erode, 2017, pp. 302-306.
13. Xue, Ming & Zhu, Changjun, "A Study and Application on Machine Learning of Artificial Intelligence," 2009, pp. 272-274.
14. Chen Zhao, Hongcui Wang, Songgun Hyon, Jianguo Wei and Jianwu Dang, "Efficient Feature Extraction of Speaker Identification Using Phoneme Mean F-Ration for Chinese", 2013 8th International Symposium on Chinese Spoken Language Processing, pp. 345-348.

15. Avan, Nadine and Burton, A. Mike and Scott, Sophie K. and McGettigan, Carolyn, “Flexible voices: Identity perception from variable vocal signals,” 2019 Psychonomic Bulletin \& Review Journal, pp. 90-102.
16. Saritha Kinkiri and Simeon Keates, “Identification of a Speaker from Familiar and Unfamiliar Voices,” 2019 5th International Conference on Robotics and Artificial, pp. 94-97.