Journal for Re Attach Therapy and Developmental Diversities
eISSN: 2589-7799
2023 August; 6 (1): 576-601

Matching Face with Voice Identity Using Static and Dynamic Stimuli

[1]Mustafa Rafid Abdul-Ally, [2]Prof. Balqis Issa Gatta Rashid (Ph.D.)
Dep. of English, College of Education for Human Sciences, University of Basrah

Received: 24- June -2023
Revised: 27- July -2023
Accepted: 21- August -2023

Abstract:

The current study tackles the ability of humans to match faces with voices depending on static and dynamic stimuli. The first aim of the current study is the possibility of matching a face with a voice using static stimuli and the second aim is the possibility of matching a face with a voice using dynamic stimuli and to find out which of the two gives more accurate matching, the static or dynamic stimuli. In this study, the Cross-Modal was adopted. The results show that the participants successfully selected the true speakers in most of the items that are shown to them. The dynamic stimuli gave more accurate results than the static ones.

Keywords: Face-voice matching, Static stimuli, Dynamic stimuli, Cross-modal.

1. Introduction

Both the dynamic and static features of our voices, like timbre, which are directly affected by many factors like age and gender, and the dynamic information, like patterns of pronunciation specific to a region (accent) or a person (such as unique laughs) are essential identifiers. Most listeners can correctly identify the speaker's gender (Lass et al., 1976, p.678) and age range (Hartman & Danahuer, 1976, p.715). It's more debatable whether a person can accurately estimate other physical qualities, such as height, weight, race, or even psychological attributes, like trustworthiness. Speaker identification is the pinnacle of identity-based technology; it enables us to recognize people by their voices with startling precision, even after extended intervals have passed (Papcun et al., 1989, pp.922-923).

Not only do people's faces convey emotion, but so do their voices. Essential insights into a person's emotional and motivational state can be gained by analyzing the acoustic parameters affected by involuntary impact and the patterns of muscular contractions associated with distinct affective states. Most research on how people interpret voices has been done in the field of linguistics (Ellis, 1989, p.211).

Emotional prosody is the speaker's employment of sonic qualities like amplitude, length, and pause that are directly related to affect to convey that state to the listener. Laughter, tears, shouts, and groans are all examples of non-speech interjections that are widely regarded as the audio counterpart of facial expressions (Scherer, 1995, p.246).

2. Objectives

The current study tries to achieve the following objectives:

1- Proving whether any information about a speaker’s face could be extracted from their voice.

2- exploring whether voice can be matched with dynamic face stimuli to determine the speaker’s face.

3- Investigating the static face stimuli and face-voice matching.

4- Investigating the accuracy of both dynamic and static stimuli and face-voice matching and which gives more accurate results.

*The present work is extracted from an M.A thesis written by the first author and supervised by the second one.

3. Hypotheses

The current study hypothesises the following:

1-Voice can provide information about a speaker’s face.

2- Dynamic and static face stimuli can be used to facilitate face-voice matching.

3- Dynamic stimuli give more accurate results than static stimuli.

4. Static and Dynamic Stimuli

Dobs, Bülthoff, & Schultz (2018, p.1) stated that facial features and their configuration, as well as the mobility of those features, provide a wealth of information about a face's structure when the face is in motion. Humans are continually decoding and incorporating these signs into their social interactions. Understanding human face perception requires research into the types of information conveyed by dynamic faces and how the human visual system processes and extracts the data. However, many face perception experiments still rely on static faces as stimuli, in part because of the difficulty of developing well-controlled dynamic face stimuli. Dynamic face stimuli studies have shown that the human visual system is very sensitive to natural facial motion, and they have consistently indicated the advantages of using dynamic face information when static face information is insufficient. These results lend credence to the theory that the human perceptual system combines sensory inputs to produce stable perception.

When we meet a friend, we make constant facial expressions like nodding, smiling, and talking; this is true of most people we encounter. People's emotional states (e.g., subtle or conversational facial expressions; Kaulard et al., 2012), where they are concentrating their attention (e.g., gaze motion; Nummenmaa and Calder, 2009), and the content of their speech can all be inferred from the information supplied by dynamic faces (e.g., lip movements; Ross et al., 2007). Despite the wealth of data given by moving faces, much of the data, such as sex, age, or basic emotions, is already present in the static version (Russell, 1994).

Understanding facial expressions is only one facet of the face that can be better perceived with the help of facial movements. For those with hearing loss, for instance, facial expressions might help them understand what others are saying (Bernstein et al., 2000; Rosenblum et al., 2002). A person's identity (Hill & Johnston, 2001; O'Toole et al., 2002; Girges et al., 2015) and gender can be inferred from their facial expressions. Facial movements can reveal a surprising lot about a person's identity, although the quantity varies with the type of movement. Recently, Dobs et al. (2016) captured numerous actors' facial expressions for a study; these expressions include emotional expressions (such as happiness), social-emotional expressions (such as laughing with a friend), and conversational expressions (e.g., introducing oneself).

5. The Data

The data in the current study are represented by 14 photos and videos of different people with their voices, then, there are 10 items in which there are either two photos of different persons with MP3 file or two videos with MP3 file. Those items are divided into 7 items which are represented as 14 videos, and the other 7 items which are represented as 14 photos, and each item contains one voice file. The videos are selected according to different criteria. And those criteria are stated below:

1- The speaker’s voice should be clear and audible.

2- Each speaker is chosen once; there are no items that present the same speaker at all.

3-The speakers do not have any kind of connection with English language teaching or linguistics at any level.

4- All the speakers are from the same region (according to their profiles).

5- The speakers of the same item should look alike to each other, there are some differences for example in the shape of the face, age, race…. etc.

6- Videos/ photos in the same item have the same environment.

7- The chosen videos and voices are understandable as well as easy to process, so the participants will not focus on the meaning rather than the faces and voices.

6. The Sample

The population of the current study is represented by the 3rd stage students at the Department of English College of Education for the humanities university of Basra of the academic year 2022-2023. The total number of the

population is 450, and the number of the students who participated in the test is 300. The number of the females is 250 and the males are 50, so it is safe to say the ratio is 1:5.

7. The procedures

After choosing the most suitable speakers for this study, the voices were separated from the videos by using Adobe Premiere Pro 2022, then after that, the videos were cut into parts each part lasts between 13-20 seconds. For each item, two different videos of different speakers in almost the same environment were prepared. On the other hand, some videos are chosen to take a photo of the speakers' faces only and separate their voices.

The videos were selected from YouTube by searching for speakers from the same region (USA), channels like TED Talks, TEDx Talks, and SysAid as well as UNSW which were great sources for different speakers with different faces.

After preparing the items, the participants were tested on different dates since their number was large and it was impossible to handle all of them in only one single session. The participants were tested by asking them to listen to the voice then they watched two muted videos or photos, and they must choose one of them and state the reason/s behind their choices.

The next stage was collecting and uploading the data to the Statistical Package for the Social Sciences program (SPSS) and the results are shown below. The model which is chosen in this study is the Cross Model which was developed by Lachs in 1999.

8. Static Stimuli

In the static stimuli group, there are 7 items, and each is discussed separately. In this group the participants listened to an MP3 voice file of a speaker, then, two different pictures of two different speakers are shown to the participants, one of the speakers is the true speaker. Each item will have its waveform figure in which one can indicate the type of voice of the item. Then, there is a spectrogram of the item, each spectrogram displays two lines, the first line (the blue one) represents the fundamental frequency (F_0), and the second line (the yellow line) represents the intensity of the voice. After that, two pictures are imported; those pictures are of the faces which are shown to the participants and one of the faces is that of the true speaker.

Item No. 1

In this item, there are two pictures of two different people, one of them is that of the true speaker. As can be seen in Pic.1 below, the two persons have almost the same pauses, and the dimensions of the pictures are the same, for example, one can see that only the upper parts of the two persons' bodies are shown to the participants and there are no hands shown to indicate any kind of movement or gestures.

From figure 1 below, one can tell that this is a model voice or can be described as a regular voice which means it is not harsh or creaky or any other type of voice quality.

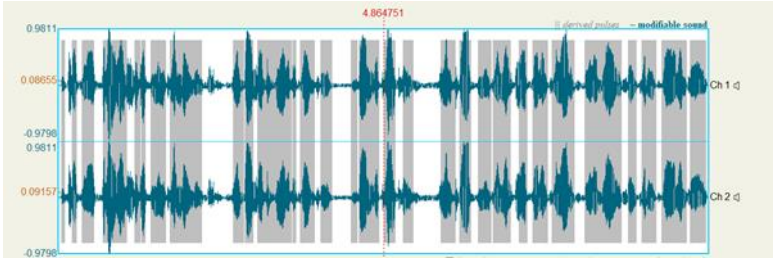
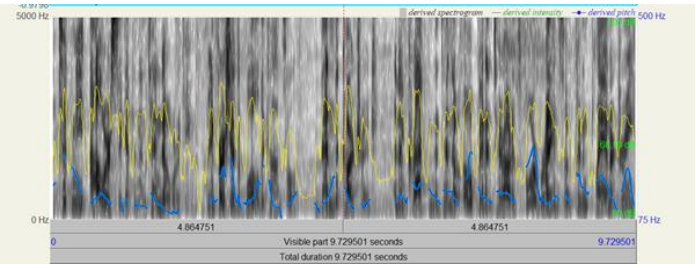


Fig. 1: Waveform of the voice of item no.1

Journal for Re Attach Therapy and Developmental Diversities
eISSN: 2589-7799
2023 August; 6 (1): 576-601



Spectrogram 1: The spectrogram of item no. 1

```
Pitch:
   Median pitch: 120.539 Hz
   Mean pitch: 128.970 Hz
   Standard deviation: 29.243 Hz
   Minimum pitch: 78.209 Hz
   Maximum pitch: 225.060 Hz
```

Pic. 1: the pitch of Item 1

From the spectrogram 1 and pic. 1 (which is taken from Praat) and knowing that the average range of pitch for male speakers is between 75-300 Hz, one can be sure that this is a modal voice since the pitch is between 78-225 Hz.

Putting in mind all the stated information, the question is whether the participants answered this item correctly or not.

Pic. 2: Item no.1 related faces

Item1					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		2	.7	.7	.7
	A	154	51.3	51.3	52.0
	B	144	48.0	48.0	100.0
	Total	300	100.0	100.0	

Table 1: Item no.1 results

Table no.1 displays the results of the first item in which the participants were asked to match a voice to a face, with the correct face being A's face. As the table shows, out of the 300 participants, 154 (51.3%) chose A's face and 144 (48%) chose B's face.

Many of the participants chose the wrong face (B's Face) and to know why they did so we asked them to state the reason/s behind their choices. Ten predicted reasons with one open reason are provided.

The bar chart 1 illustrates that the most common reasons provided were related to the thickness of the voice and age of the speaker. The participants thought that A's voice should be thicker than B's and A should be older, that's why they chose A as the right answer which is not.

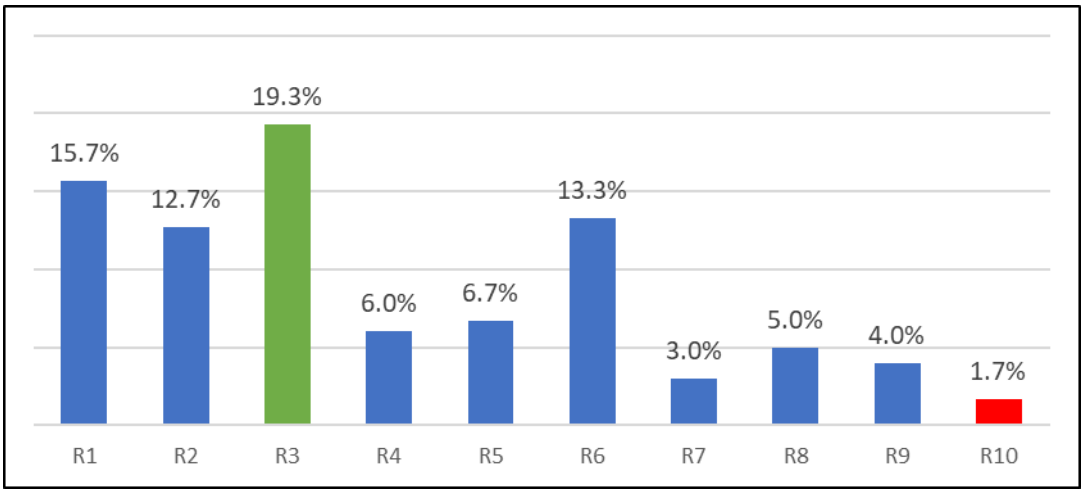


Chart 1: The selected reasons for answering item no.1

The most common reason provided was reason number 3 (henceforth R + the reason number), fifty-eight participants, or 19.3%, stated that A looked older than B, and 15 participants, or 5%, stated the opposite). The speed rate of the voice and perceived ethnicity are not relevant to the current item, but some participants have chosen both as reasons for their answers. They chose those two irrelevant reasons either they did not understand the task very well or they wanted to finish the task without paying much attention to the real reasons.

The results show that the participants used a variety of factors to match the voice to the true face, with the thickness of the voice and the size of the vocal tract being the most cited reasons.

Item No.2

In this item, as can be seen in Pic.3, both speakers seem old, but one can indicate that A is somehow older than B even though they are almost of the same age.

Fig 2 shows that this is a hoarse voice. By looking at the waveform of the voice, one cannot be sure whether this is a hoarse voice. The best way to know what type of voice we have is either by listening to the voice or having a detailed look at the spectrogram.

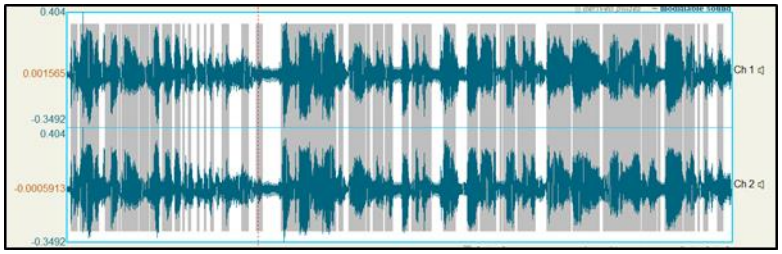
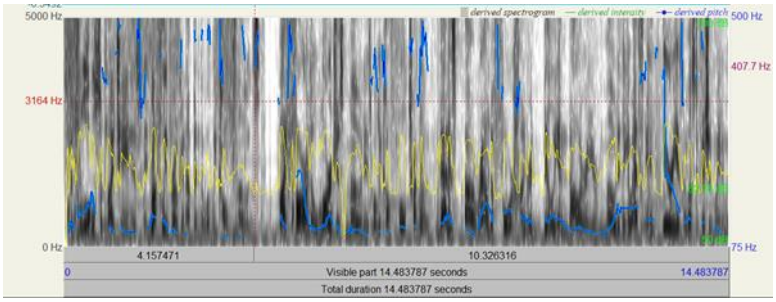
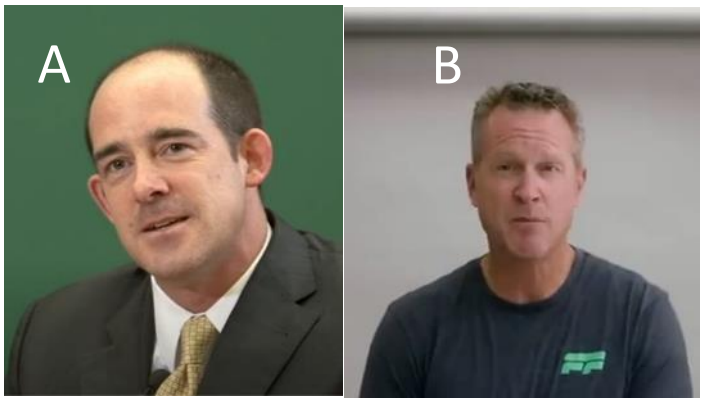


Fig. 2: Waveform of the voice of item no.2



Spectrogram 2: The spectrogram of item no. 2

One can see that in the above spectrogram, there are darker areas compared to the previous spectrogram. It is safe to say that this is a hoarse voice. Looking at picture 3, one can assume that B should have a thicker voice than A. Even B's facial expressions and the physical characteristics seem tougher than A's.



Pic. 3: Item no.2 related faces

Table 2: Item no.2 results

Item2					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		2	.7	.7	.7
	A	109	36.3	36.3	37.0
	B	189	63.0	63.0	100.0
	Total	300	100.0	100.0	

In Table 2, we can see that out of the 300 participants, 189 (63%) correctly chose option B as the correct face corresponding to the voice file. This suggests that most of the participants were able to match the voice to the correct face. Considering the following chart, we can see that the majority has chosen B because he has a thicker voice than A.

The most cited reason for choosing B was R3 (A looks older than B), followed by R8 (B looks older than A), and R6 (B's Voice is thicker than A's Voice). These results indicate that the participants relied on multiple factors when matching the voice to the face. Overall, these results indicate that the participants were generally successful in matching the voice to the correct face. They also show that multiple factors, such as voice pitch and age, were considered when making this decision.

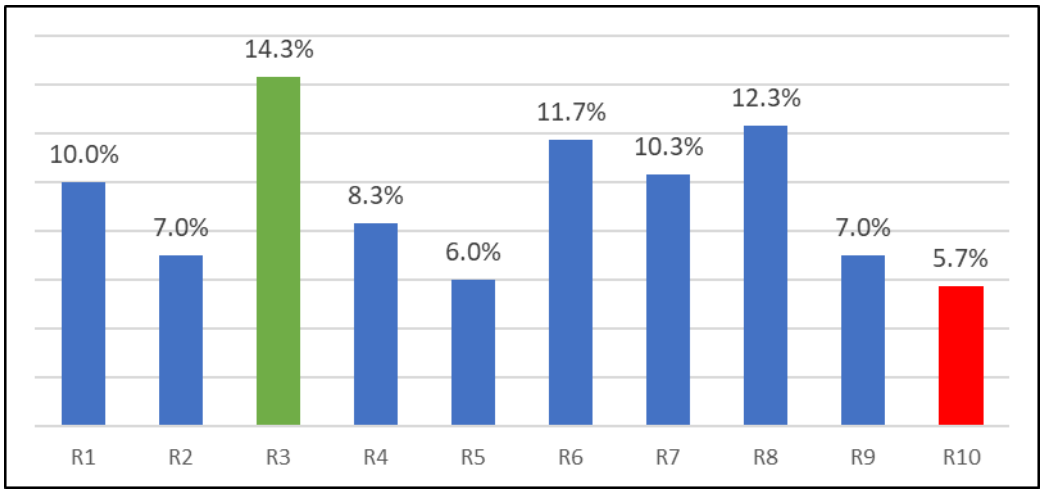


Chart 2: The selected reasons for answering item no.2

Item No.3

Item no. 3 is different from the two previous items. In this item, the sample is different. The participants listened to a woman's voice then they were shown two pictures of two women. The two women have the same pauses, and both seem that they were talking. In both pictures, it is noticeable that the hands of the women are shown. The participants could not use the women's gestures and body movements to indicate the true speaker.

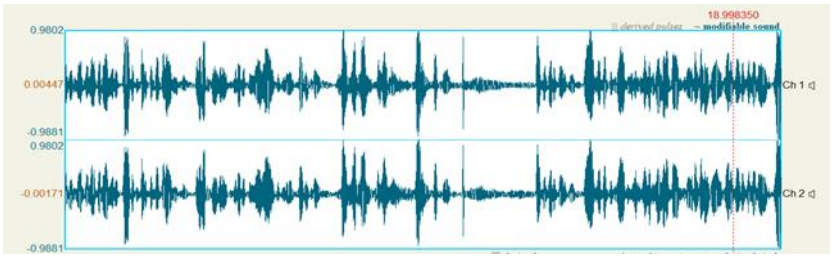
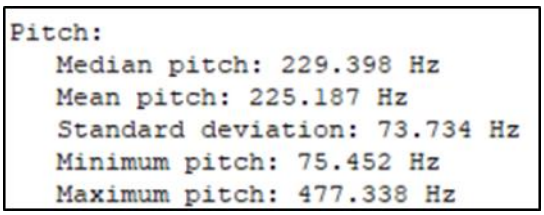


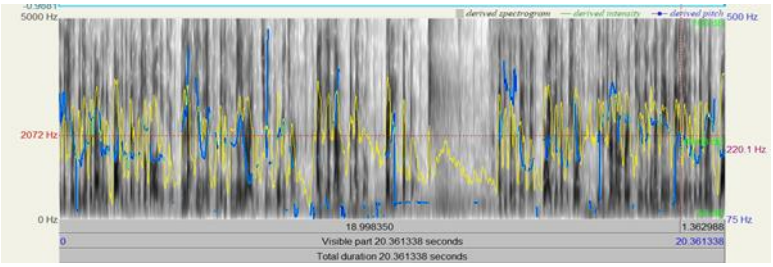
Fig. 3: Waveform of the voice of item no.3

The figure above is completely different from the previous ones. In the first and second figures, the amplitude is most of the time low and the waves are compressed to each other. But in this figure, the amplitude is higher and at the same time, one can see that the waves are not compressed. The reason for such differences is the gender of the speaker, but from the waveform and the spectrogram only, it is not clear if this voice is for a male or a female. Using Praat to show the pitch of this voice could solve this issue.

Pic 4 shows that the pitch range is between 75-477 Hz, and taking into consideration that the average range of pitch for a male speaker is 75-300 Hz, and for a female speaker is up to 600 Hz, one can be sure that this is a modal voice of a female speaker. As shown in the spectrogram 3 given below, the blue line is higher in this item than in the previous items.



Pic 4: the pitch of the voice of item 3



Spectrogram 3: The spectrogram of item no. 3

Pic. 5: Item no.3 related faces

The two women in the picture above are of almost the same age but it seems that B is a little bit older than A. Another thing is that B seems to be thinner; however, both have the same pauses.

From the following table, we can tell that most of the participants chose option A. Out of the 300 participants, 68.6% (206) chose face A while 31.3% (94) chose face B. Notably, the correct face is A's face, which suggests that a good number of the participants were able to accurately match the voice to the corresponding face.

Table 3: Item no.3 results

Item3					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A	206	68.6	68.6	68.6
	B	94	31.3	31.3	100.0
	Total	300	100.0	100.0	

The following chart provides insight into the reasons behind the participants' choices. The most common reasons for choosing A's face were R5, R8, and R7. R7 which states "The vocal tract of B is larger than A's" was chosen by 10% (30) of the participants, on the other hand, reason R8 which states that "B looks older than A" represents 12% of the participants. R5 which states that "A's voice sounds like African American" is irrelevant to this item, so it is excluded.

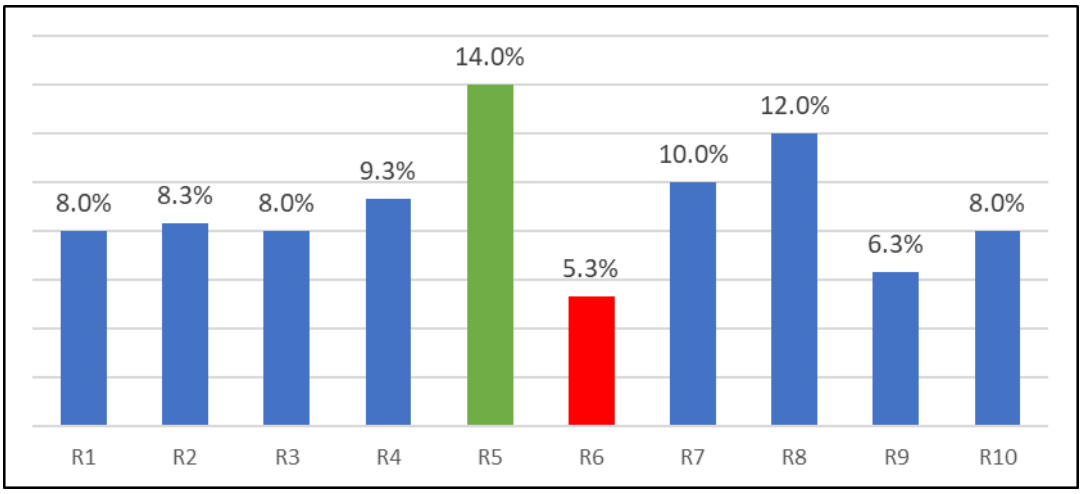


Chart 3: The selected reasons for answering item no.3

R6 which states that "B's voice is thicker than A's voice" is the least selected reason, which indicates that the participants did not think that B should have a thick voice, and this is because B is thinner. Out of the 300 participants, 10.7% did not choose anything; they left this item empty.

These findings show that the participants relied on multiple cues when matching a voice to a face, including vocal characteristics such as voice thickness and vocal tract size, as well as ethnicity. The percentage of accuracy in matching the voice to the correct face could also imply that the participants were able to use a combination of cues to make up their decisions.

Item No.4

In item no. 4, there are two males, the first one (A) seems a little bit older than B. They differ in the size of the vocal track and its shape. The following figure shows that the speaker has a modal voice quality.

From the spectrogram, we can say that the average of the F_0 falls between 75Hz-200Hz, though there are some areas where it gets so high.

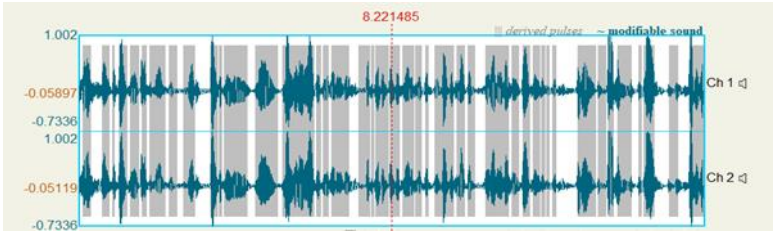
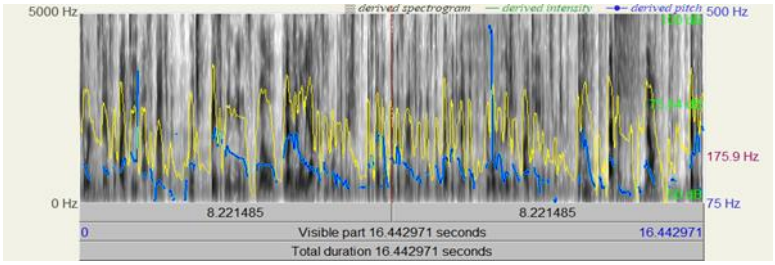


Fig. 4: Waveform of the voice of item no.4



Spectrogram 4: The spectrogram of item no. 4

From Pic 6, one can assume that A should have a thick attractive voice while B seems to have a normal voice. Now looking at table 4, it shows that out of the 300 participants, 198 (66%) correctly identified Face B as the

matching face for the MP3 voice file presented. This suggests that Face B is more likely to be associated with the voice characteristics heard in the MP3 file than Face A.

Pic. 6: Item no.4 related faces

Table 4: Item no.4 results

Item4					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A	102	30	30	30
	B	198	66.0	66.0	100.0
	Total	300	100.0	100.0	

Chart 4 illustrates the reasons behind the participants' choice of face A or face B. The most frequent reason given for choosing face B is R8 which states that "B looks older than A" It was selected by 13.7% of the participants. Other reasons given for choosing face B were R4 and R7. R4 which states that "A's Voice is Slower than B's" was selected by 12% of the participants. Moreover, R7 which states that "The vocal tract of B is larger than A's" was selected by 10.7% of the participants.

The table and chart indicate that the participants used a combination of vocal and physical characteristics to match the voice to the correct face. They relied heavily on the thickness of the voice, as well as the size of the vocal tract and the age of the person.

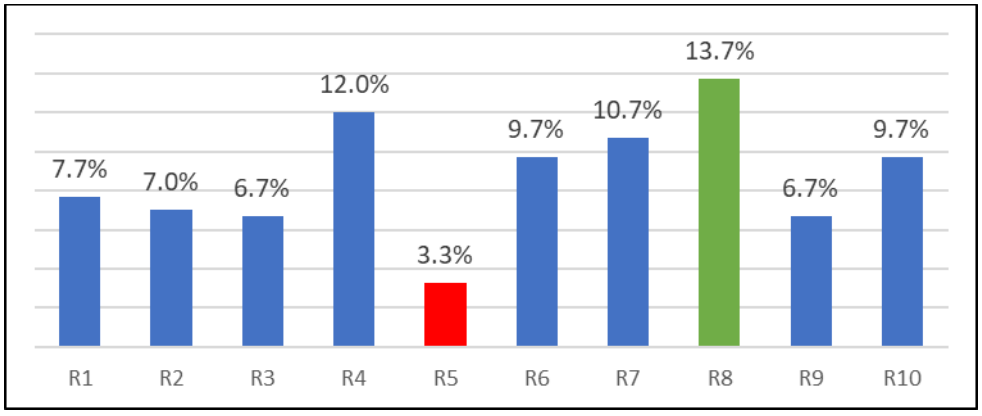


Chart 4: The selected reasons for answering item no.4

Item No. 5

In item no.5, there are two old men, but B looks older than A. Taking a glance at the two persons, the first assumption that comes to mind is that B should have a thicker voice than A and that A should have a normal voice (modal voice). The physical features of their faces are different, this might help the participants to identify the speaker.

From the waveform figure, we could tell that this is not a thick voice but a modal one. In spectrogram No.5 we could notice that the blue line falls between 75Hz – 149 Hz which suggests that it is a modal voice.

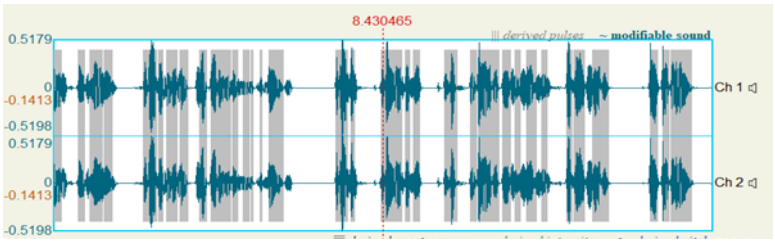
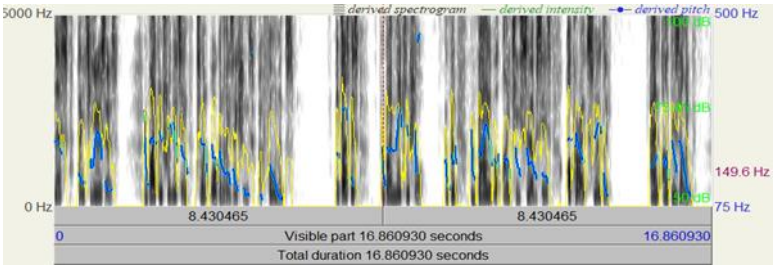


Fig. 5: Waveform of the voice of item no.5



Spectrogram 5: The spectrogram of item no. 5

Pic. 7: Item no.5 related faces

In table 5, we can see that most of the participants chose option A, that is out of 300 participants, 62% (186) chose face A while 38% (114) chose face B. The correct face is A's face, which suggests that a good number of the participants were able to accurately associate the voice with the corresponding face. According to the reasons given by the participants, A looks older than B and that is why his voice should be slower and thicker. The voice of this item was sharp and modal, using these parameters they excluded A's face from their answers. This does not apply to all the participants since 114 chose B's face as the correct answer.

Table 5: Item no.5 results

Item5					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A	186	62.0	62.0	62.0
	B	114	38.0	38.0	100.0
	Total	300	100.0	100.0	

Chart 5 shows that the most frequent reason given for choosing face A was R3, 65 (21.7%) of the participants chose this reason which is an unexpected answer since B looks older than A. Other reasons given for choosing face A were R4 and R9, both reasons contradict each other but R9 is much more relevant to this item than R The participants relied on many cues to match the voice to the correct face. However, they succeeded in correctly selecting the right face for this item.

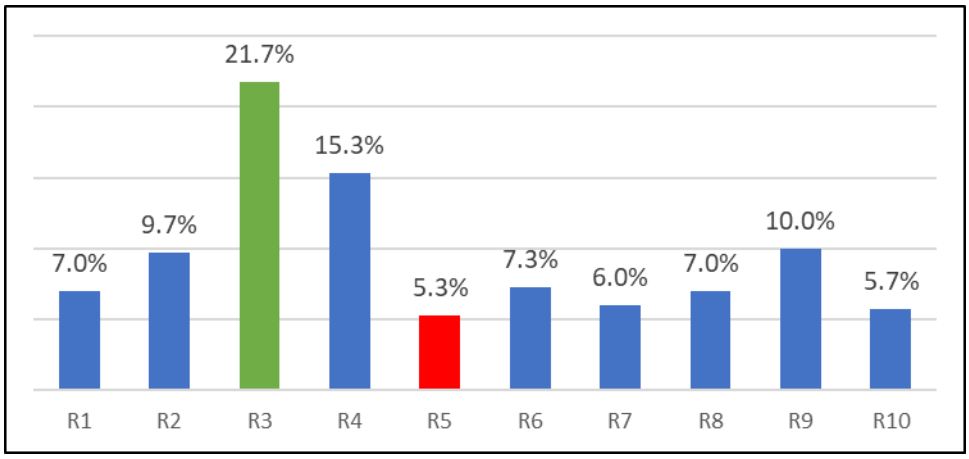


Chart 5: The selected reasons for answering item no.5

Item No. 6

Item no. 6 contains two different speakers, both are females. A seems thinner and has a larger vocal tract. On the other hand, B seems to be younger and has a smaller vocal tract. Both have the same pauses and there are no unique movements or gestures.

Figure 6 shows the waveform of the voice of item no.6 and as can be seen this voice is a modal voice since the sound waves are clear and not compressed together.

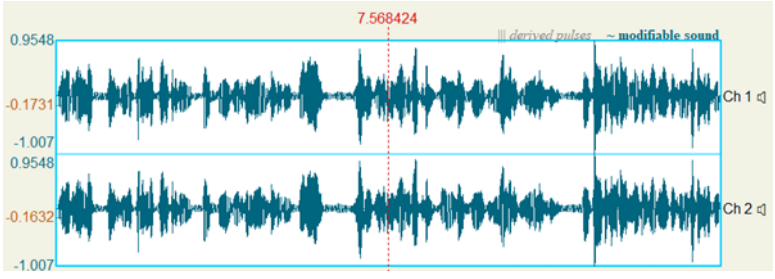
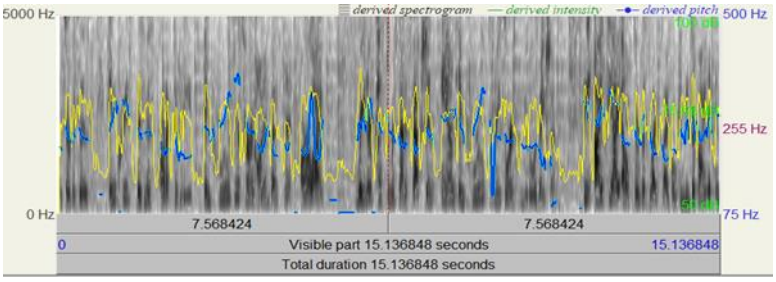


Fig. 6: Waveform of the voice of item no.6



Spectrogram 6: The spectrogram of item no. 6

In spectrogram 6, it is noticeable that the blue line which represents the F_0 (or the pitch since the relationship between F_0 and pitch is positive, thus a high F_0 means a high pitch) is higher than the previous item, this is because females have a range of F_0 which falls between 100 Hz – 600 Hz.

Pic. 8: Item no.6 related faces

From table 6, one can see that out of the 300 participants, 203 (67.7%) correctly chose A's face, while 97 (32.3%) chose B's face. This indicates that the participants were not completely successful in matching the voice to the correct face.

Table 6: Item no.6 results

Item6					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A	203	67.7	67.7	67.7
	B	97	32.3	32.3	100.0
	Total	300	100.0	100.0	

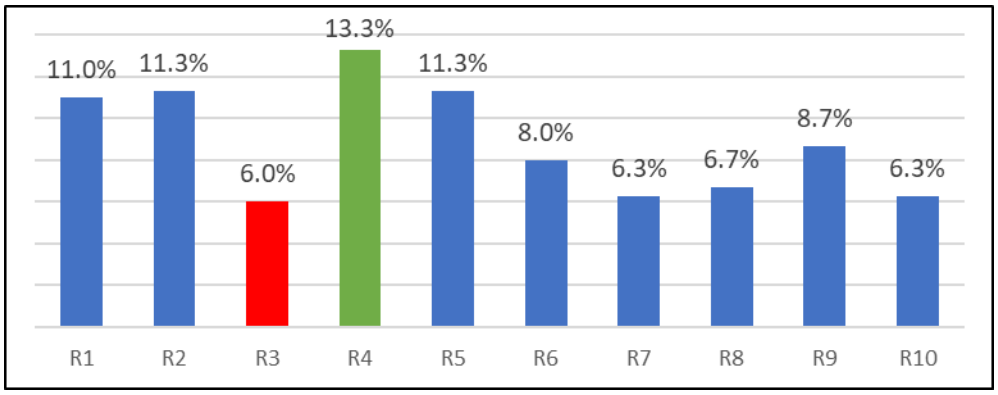


Chart 6: The selected reasons for answering item no.6

According to chart 6, R4 which states that "A's voice is slower than B's" was selected by 40 (13.3%) participants. Then R2 and R5 are equal in percentage. R5 is not relevant to the current item, so it is excluded. R2 states that "the vocal tract of A is larger than B's" which is correct. The participants used this cue to help them to choose the true speaker. Another reason was R1 (A's voice is thicker than B's voice) which was selected by 33 participants (12.4%) as the reason why they thought the voice was for A's face.

These results show that the participants relied on multiple cues to match the voice to the correct face. However, the success rate varied depending on the specific item given by the participants for their choices.

Item No. 7

This item is slightly different from the others. Here the participants are tested to see if they can choose the true speaker depending on the voice only, although one of the selected speakers shows hand gestures while the other one does not.

From fig. 7, one can note that this is a modal voice, and in the spectrogram we can see that the range of the F_0 falls between 75 Hz – 195 Hz which indicates that this is a modal voice.

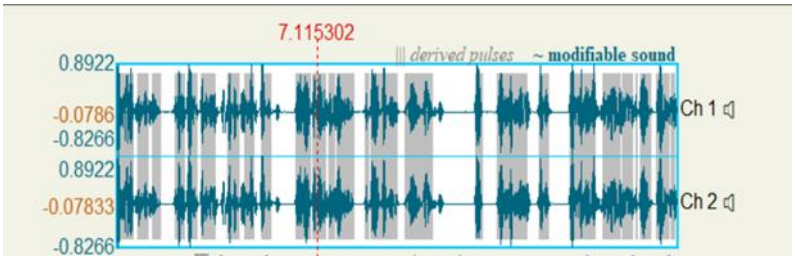
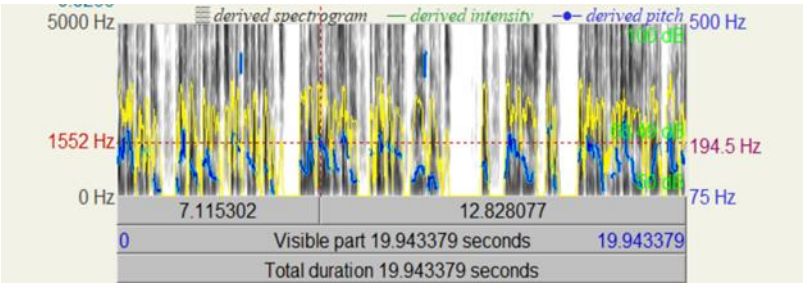


Fig. 7: Waveform of the voice of item no.7



Spectrogram 7: The spectrogram of item no. 7

From the picture of the two persons, one can tell that A is using his hand to indicate that he is explaining or talking about something; on the other hand, B does not show any kind of gestures, he holds steady. The question is whether the participants depended on gestures only to identify the true speaker or not.

Table 7 shows that 100 participants (33.3%) incorrectly selected face A, while 200 participants (66.7%) correctly identified face B. This indicates that the participants did not rely on gestures to choose the true speaker. They mostly identified the true speaker based on the voice quality.

Pic. 9: Item no.7 related faces

Table 7: Item no.7 results

Item7					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A	100	33.3	33.3	33.3
	B	200	66.7	66.7	100.0
	Total	300	100.0	100.0	

As shown in chart 7, the most selected answer is R8 which states that "B looks older than A", in other words, the participants depended on the age factor in the first place to identify the true speaker, not the gesture. Then, the second most selected reason is R7 which states that "The vocal tract of B is larger than A's". The least selected reason was R4 which states that "A's voice is slower than B's".

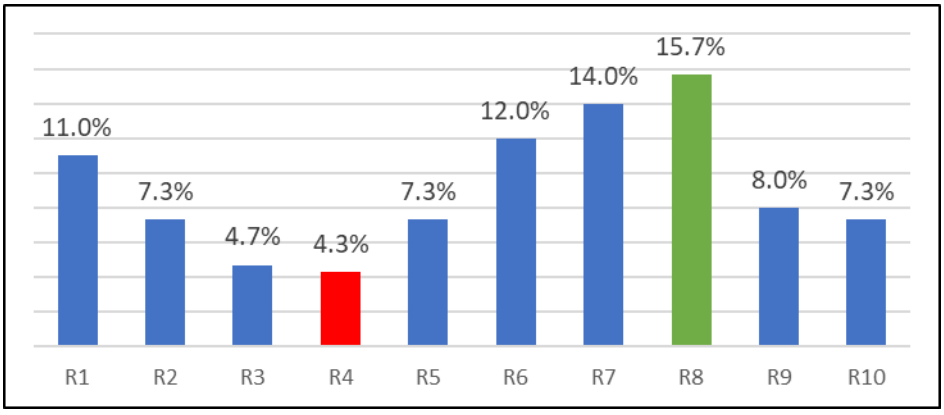


Chart 7: The selected reasons for answering item no.7

9. Dynamic Stimuli

In this group (the dynamic stimuli group) there are 8 items. Each item is discussed separately as was the case with the first group. The participants in this group listened to an MP3 voice file of a speaker before being shown two different muted videos of two different speakers, one of which is that of the true speaker. The same steps as in the first group are followed here.

Item No. 8

This item contains two different muted videos, the first muted video (A) is about a blond woman speaking in a conference. She is thinner than B. B is in the same environment, she is also a speaker in some conference, and she seems fatter than A. One could propose that B should have a thicker voice than A.

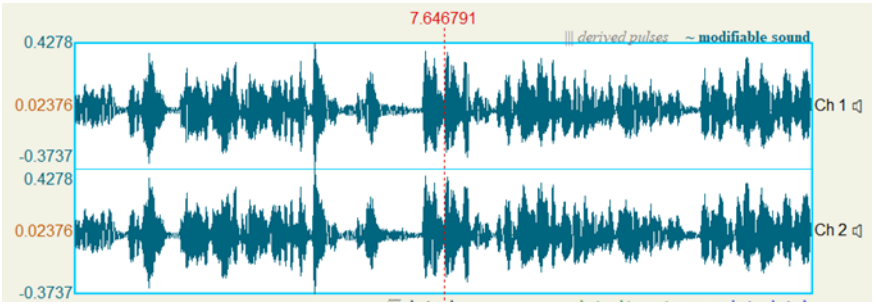
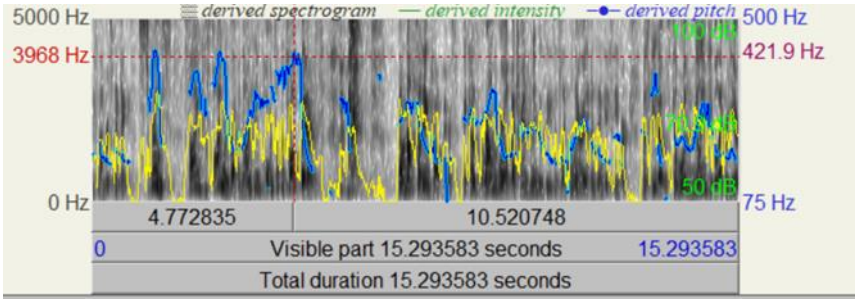


Fig. 8: Waveform of the voice of item no.8

In the above figure, we can notice that this is a modal voice. And following the spectrogram shows that the F_0 range falls between 75 Hz – 421 Hz which is above male's range. There are some areas where it is harsh, this mix between the two types of voice quality could help the participants to identify the correct face. The type of voice and the facial gestures of the speakers could be used as a hint to do so.



Spectrogram 8: The spectrogram of item no. 8

Table 8 shows that 230 (76.7%) participants have chosen the right answer (B), and 70 (23.3%) participants chose the wrong face (A). These results are completely different from the first variable (static stimuli). We can notice that the dynamic stimuli have given us a higher success rate than the static stimuli.

Item 8					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A	70	23.3	23.3	23.3
	B	230	76.7	76.7	100.0
	Total	300	100.0	100.0	

Table 8: Item no.8 results

The participants stated some reasons for their answers. As shown in the chart below, the most common reasons for their answers are R4 and R7. Fifty-seven (19%) of the participants stated that A seems to speak at a slower rate than B (R4) and using this and the MP3 voice they matched the right face, while 37 (12.3%) of the participants stated that the vocal tract of B is larger than A's (R7).

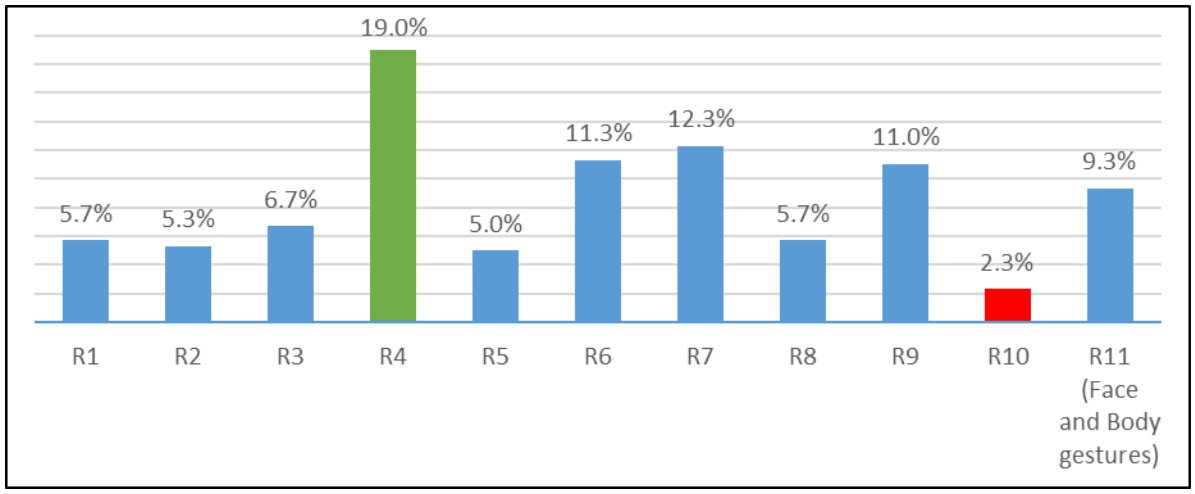


Chart 8: The selected reasons for answering item no.8

Item No. 9

Two muted videos are shown to the participants in item 12. Video A is about a man in his 30s and video B is the same, but A seems so serious and there is not any kind of emotions in his speech, while B has shown a lot of facial expressions indicating that he is affected. The participants used this cue to help them to match the voice to the face.

From figure 9 and spectrogram 9, we can tell that the true speaker has a modal voice. The F_0 range falls between 75 Hz – 235 Hz which is the normal range of a male speaker.

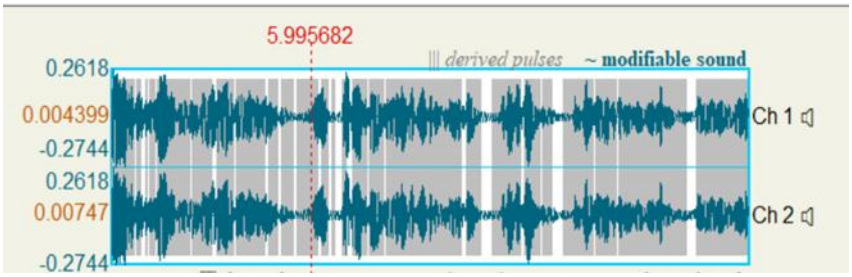
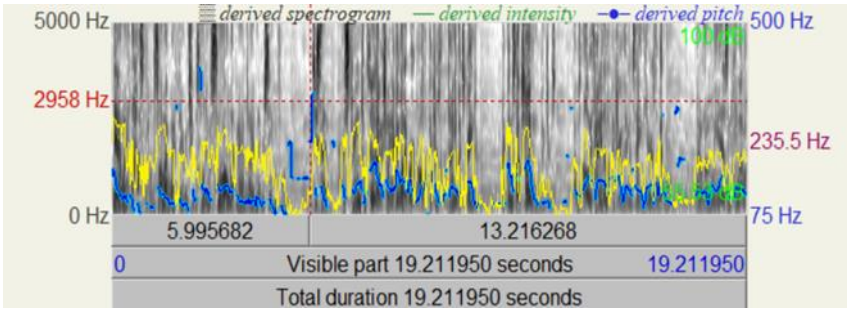


Fig. 9: Waveform of the voice of item no.9



Spectrogram 9: The spectrogram of item no. 9

In the following table we can recognize that many of the participants selected the correct face which is B. Out of 299 (one of the participants did not answer this item), 196 of the participants selected B which is equal to 65.3% of the total number. One hundred and three (33%) of the participants selected the wrong face.

Journal for Re Attach Therapy and Developmental Diversities
eISSN: 2589-7799
2023 August; 6 (1): 576-601

Table 9: Item no.9 results

Item 9					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		1	.3	.3	.3
	A	103	33	33	37
	B	196	65.3	65.3	100.0
	Total	300	100.0	100.0	

After having the results of the reasons behind this percentage, it could be noticed that R9 is the most selected reason; by watching the videos participants assumed that B's voice should be slower than A's (R9). The least selected reasons were R1 and R10, for sure the two speakers are not African Americans so we can exclude those results. On the other hand, most of the participants did not think that A's voice should be thicker than B's since they both are in their 30s.

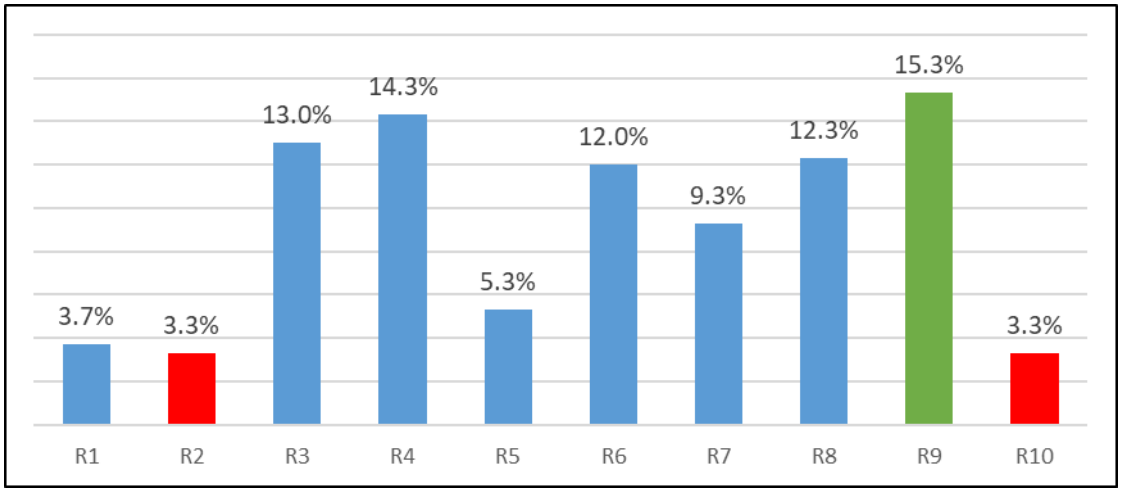


Chart 9: The selected reasons for answering item no.9

Item No. 10

The two videos in item 10 were almost similar except for one thing, video A is about a skinny woman who seems a little bit tired and talk normally, whereas in B the woman looks fatter.

In figure 10, one can notice that the true speaker has a modal voice but since in this item the true speaker is a woman, we can notice the difference between this figure and that of the previous item. The F_0 range falls between 80 Hz – 463 Hz.

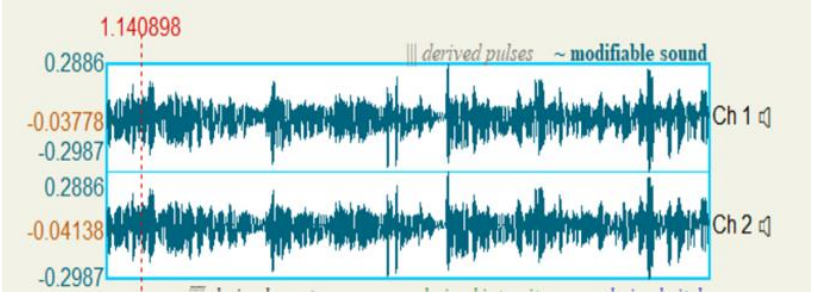
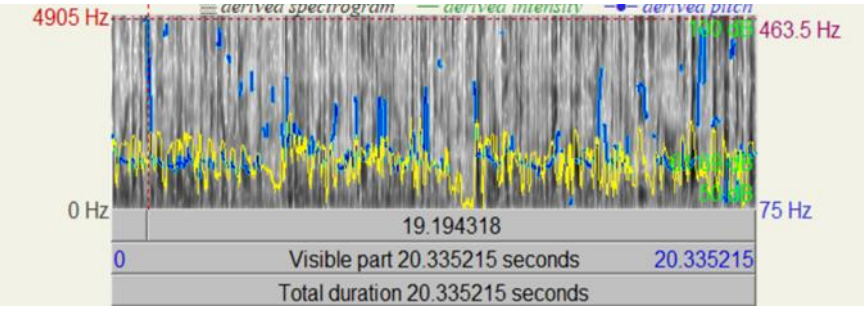


Fig. 10: Waveform of the voice of item no.10



Spectrogram 10: The spectrogram of item no. 10

Table 10 provides information about the frequency and percentage of the participants who correctly identified the speaker (A) from the two videos. According to the first table, out of the 300 participants, 215 (71.7%) correctly identified the speaker from the two videos, while 85 (28.3%) did not. This shows that the task was much easier than the previous items, as a larger number of the participants correctly identified the true speaker.

Table 10: Item no.10 results

Item10					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A	215	71.7	71.7	71.7
	B	85	28.3	28.3	100.0
	Total	300	100.0	100.0	

Chart 13 provides information about the reasons that the participants gave behind making their choice. The most frequently cited reason for identification was "A's voice is slower than B's" (R4, 32.4%), followed by "The vocal tract of A is larger than B's" (R3, 15.8%). Other reasons cited by the participants include age differences (13%), voice thickness (18.4%), and facial gestures (4%).

It is worth noting that some participants did not provide a reason for their identification (9.3%), which may suggest that they were unable to identify any distinctive characteristics between the two speakers. Overall, these results suggest that the participants relied on a combination of vocal characteristics and non-verbal cues to identify the speaker, with the most salient characteristic being the speech rate of the speaker's voice.

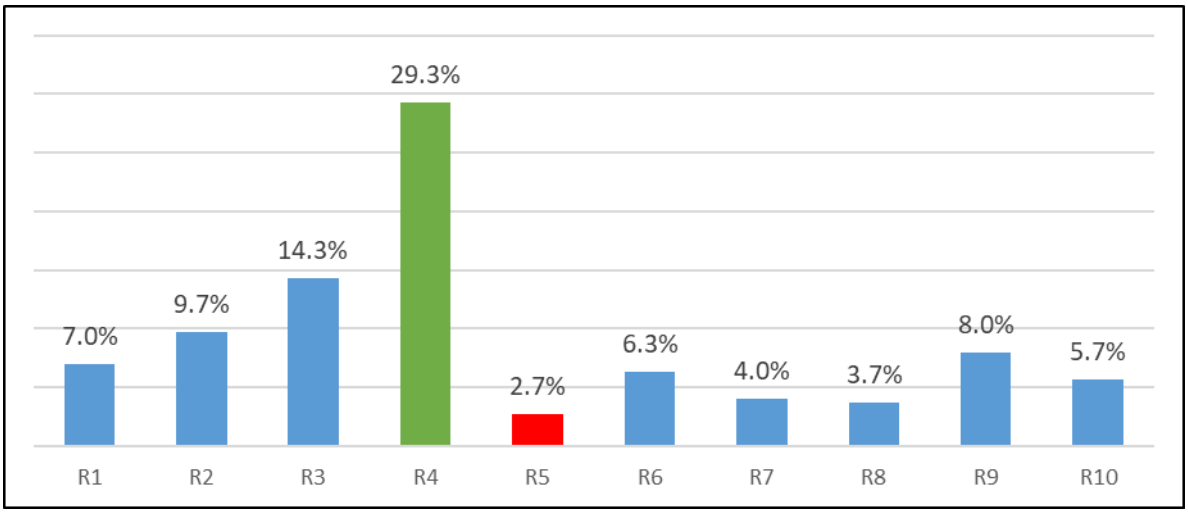


Chart 10: The selected reasons for answering item no.10

Item No. 11

Item no. 11 is about two young males, A looks a little bit younger than B. B in this item was presenting a certain topic and he used his hand a lot. The question is whether the participants used those gestures to select the true speaker.

Figure 11 and spectrogram 11 show that the true speaker has a modal voice. The F_0 range is between 75 Hz – 180 Hz. But there are some areas where F_0 goes down and this is a sign of creakiness.

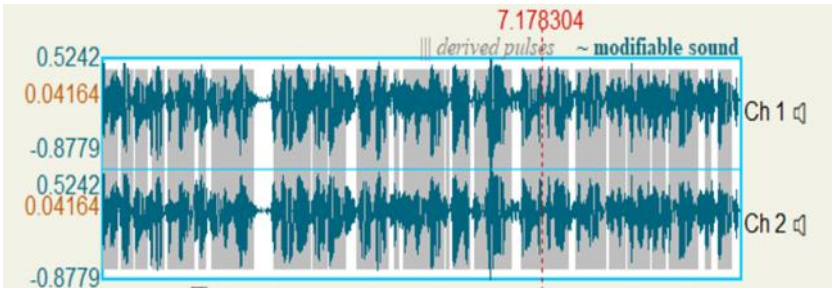
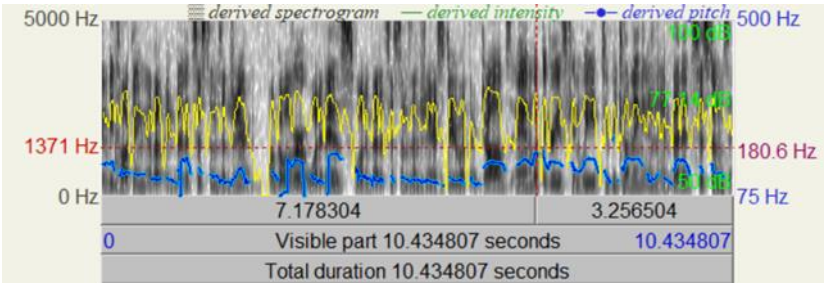


Fig. 11: Waveform of the voice of item no.11



Spectrogram 11: The spectrogram of item no. 11

Table 11 shows the results of the participants' responses in identifying the correct face from a list of two videos (A and B), where the correct video is A. Out of the 300 participants, 146 (48.7%) identified video A as the correct one, while 153 (51%) wrongly selected video B.

Table 11: Item no.11 results

Item11					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		1	.3	.3	.3
	A	146	48.7	48.7	49.0
	B	153	51.0	51.0	100.0
	Total	300	100.0	100.0	

Chart 11 shows the frequency of the distribution of the reasons why participants believed video A was the correct one. The most frequently cited reason was R6 ("B's voice is thicker than A's voice"), which was mentioned by 51 participants (17% of the total sample). The least cited reason was reason 9 ("B's voice rate is slower than A's"), which was mentioned by only 9 participants (3% of the total sample).

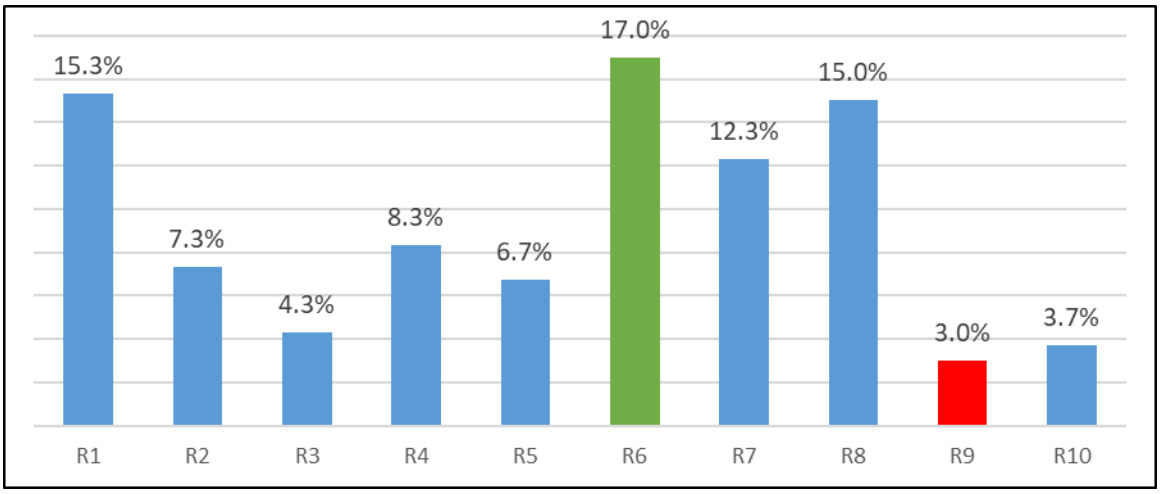


Chart 14: The selected reasons for answering item no.11

The results indicate that the participants did not correctly identify the true speaker; it seems that B's gestures and hand movements affected their decision.

It seems that the participants relied more on the visual cues (facial gestures and hand movements) rather than on the auditory cues (voice thickness and speech rate) to identify the true speaker. It is worth noting that this is the only incorrectly identified item in this group (dynamic Stimuli).

Item No. 12

Two muted videos of two different very young males are shown to the participants. A shows a lot of gestures and hand movements, while B sits on a chair without any kind of hand movements but only some face gestures.

Figure 12 shows that the true speaker has a creaky voice but from the waveforms only it is not clear if it is creaky or modal. In the spectrogram 12, it is noticeable that the F_0 goes up and then suddenly down, this is a sign of creakiness.

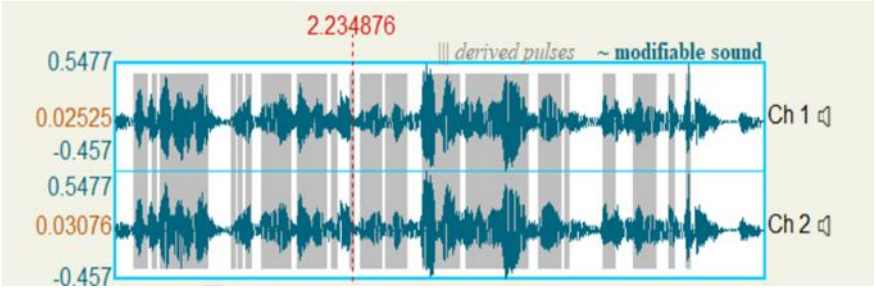
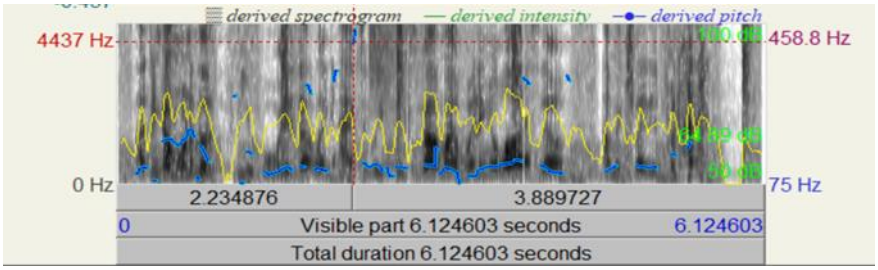


Fig. 12: Waveform of the voice of item no.12



Spectrogram 12: The spectrogram of item no. 12

Table 12 shows the frequency and percentage of the participants who chose either A or B as the correct video. Regarding this item, the correct video is B, and the table shows that 76% of the participants correctly identified B as the true speaker, while 23.7% incorrectly identified A.

Journal for Re Attach Therapy and Developmental Diversities
eISSN: 2589-7799
2023 August; 6 (1): 576-601

Table 12: Item no.12 results

Item12					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		1	.3	.3	.3
	A	71	23.7	23.7	20
	B	228	76.0	76.0	100.0
	Total	300	100.0	100.0	

Chart 15 shows the frequency and percentage of the participants who selected each reason as the basis for believing that either A or B was the correct speaker. The valid percentages are calculated based on the total number of the responses for that reason, and the cumulative percentages show the total percentage of the responses up to that point.

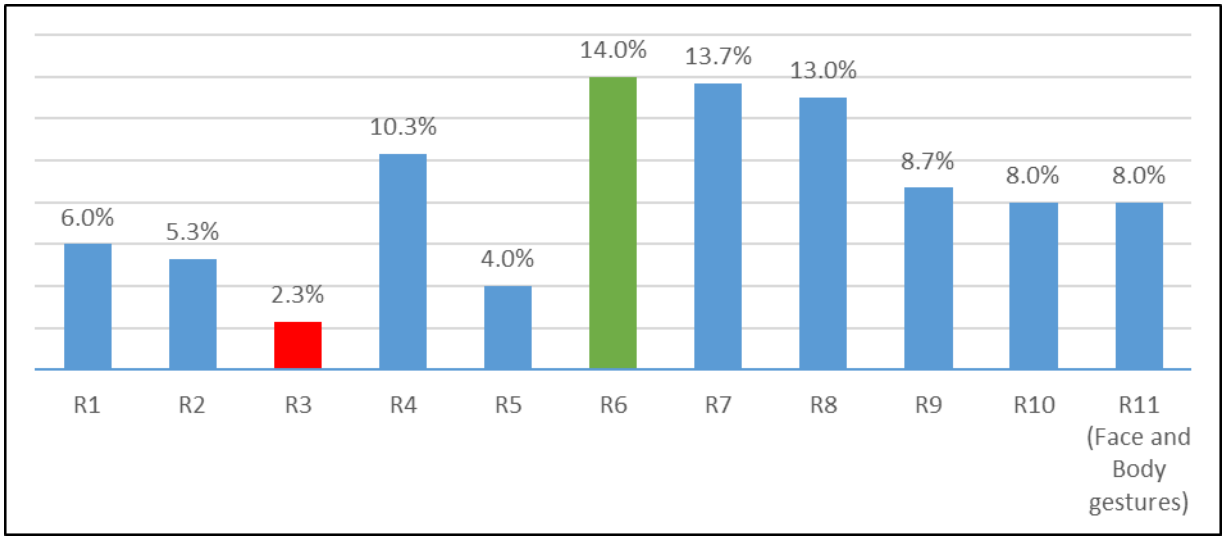


Chart 15: The selected reasons for answering item no.15

It seems that the most common reasons for the participants to correctly identify the true speaker were R6 and R7. R6, which states that "B's voice should be thicker than A's" was selected by 42 (14%) participants, and R7 which states that "The vocal tract of B is larger than A's" was selected by 41 (13.7%) participants. The participants selected other reasons such as R2 which states that "the vocal tract of B is larger than A", R3 (B looks older than A), and R4 (A's voice is slower than B's).

Moreover, some participants chose reasons that are not related to voice or speech but were used by the participants to match the voice to the face, such as R11 (face and body gestures).

Item No. 13

In this item, there are two muted videos: the first one (A) is of a fat man with a large vocal tract, and the second video (B) is of a very handsome man with a well-shaped vocal tract.

From figure 13 and the following spectrogram, one could tell that this is not a modal voice, there are a lot of compressed areas, and we can notice that the amplitude is higher in this item than in the previous one. The range of the F_0 falls between 72 Hz – 185 Hz, but we can notice that there are high areas of the F_0, and those areas go up to 499 Hz.

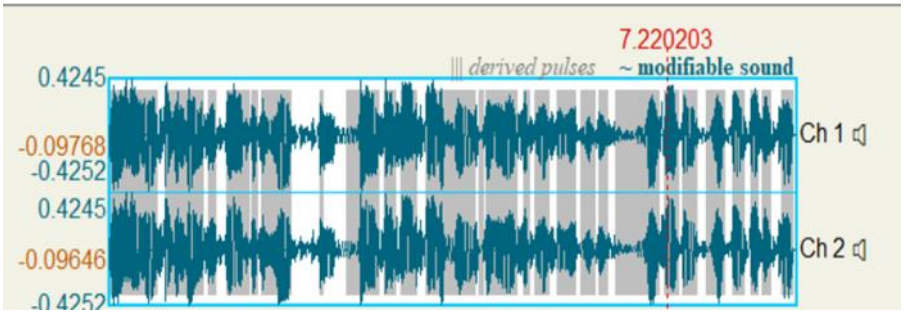
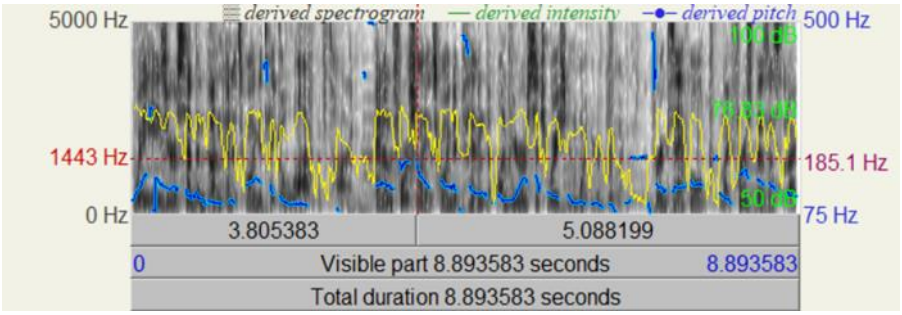


Fig. 13: Waveform of the voice of item no.13



Spectrogram 13: The spectrogram of item no. 13

Table 13: Item no.13 results

Item13					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A	75	25.0	25.0	25.0
	B	225	75.0	75.0	100.0
	Total	300	100.0	100.0	

Based on the table above we can see that 75% of the participants correctly identified video B as the true speaker, while 25% incorrectly chose video A. The reasons which the participants gave for making their choices as the following: the most common reason was that they thought there should be a difference in the thickness of the voice between the two speakers (R6 and R1), hence 19.3% of the participants selected R6.

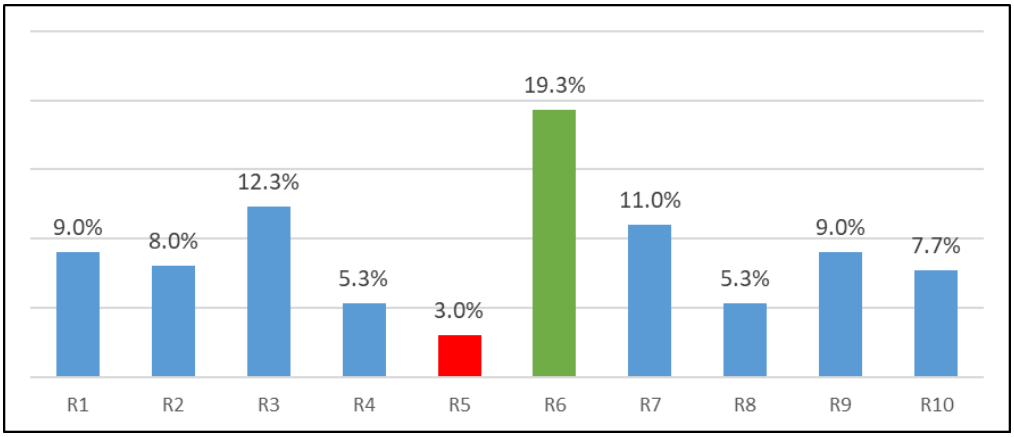


Chart 13: The selected reasons for answering item no.13

Other reasons given included differences in vocal tract size (R2 and R7), perceived age differences (R3 and R8), and differences in speech rates (R4 and R9). Only a small percentage of the participants (3.3%) selected R5, which states that one of the speakers has an African American sounding voice. Interestingly, face and body gestures (R11) were only selected by 12.2% of the participants, which means that these cues may have been not as important as the vocal cues in identifying the correct speaker. The vocal cues were the most important factor in making the participants' decision to correctly identify the true speaker.

Item No. 14

For item 14, it is in the same situation as the previous item. One of the speakers is an overweight person and the other is a thin person. But in this item, the true speaker is not the thin person but the opposite. A uses a lot of hand movement and face gestures while B uses face gestures a lot and rarely used his hand. This might help the participants to decide on the true speaker.

From the figure and the spectrogram below we could notice that the voice is not thick as in the previous item, and the F_0 range falls between 75 Hz – 179 Hz. As can be seen, the blue line in the spectrogram is stable and there are no high areas. And even for the intensity, we could notice that the waves are stable, and their movements are regular.

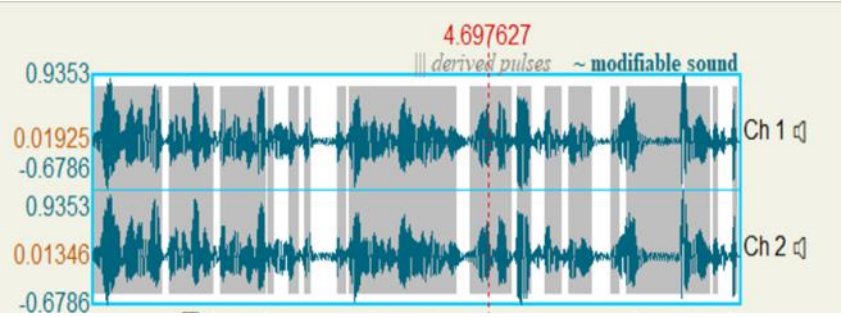
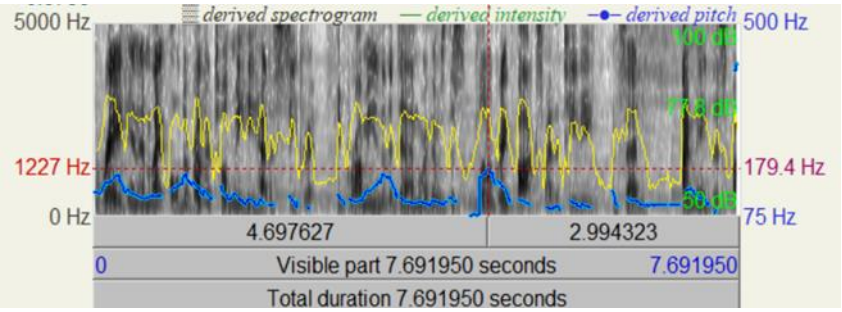


Fig. 14: Waveform of the voice of item no.14



Spectrogram 14: The spectrogram of item no. 14

Out of the 300 participants, 154 (51.3%) correctly identified video B as being the true speaker, while 145 (48.3%) incorrectly selected video A.

As for the reasons given (see chart 14), the most selected reason was R4 (A's voice is slower than B's) with 47 (15.7%) participants selecting it, followed by R7 (the vocal tract of B is larger than A's) with 37 (12.3%) participants selecting it. The least selected reasons were R1 (A's voice is thicker than B's) and R10 (B's voice sounds like an African American voice).

Table 14: Item no.14 results

Item14				
	Frequency	Percent	Valid Percent	Cumulative Percent

Valid		1	.3	.3	.3
	A	145	48.3	48.3	48.7
	B	154	51.3	51.3	100.0
	Total	300	100.0	100.0	

Notably, R11 (Face and body gestures) was selected by 14 (7%) participants, even though the videos were muted. This means that the participants are still able to identify the true speaker depending on non-verbal cues.

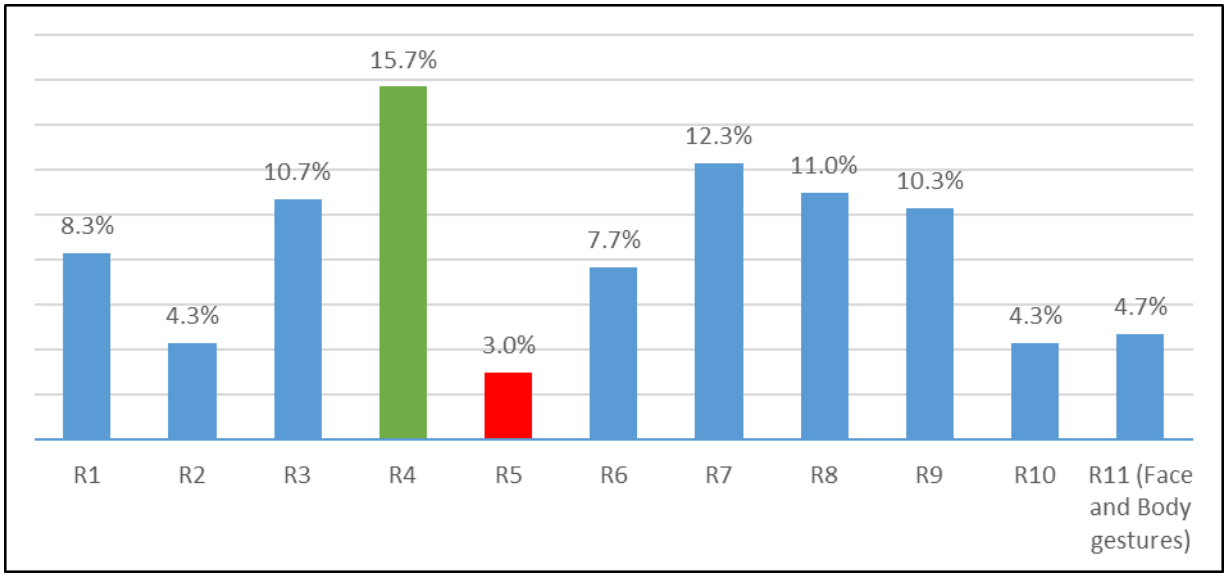


Chart 14: The selected reasons for answering item no.14

9. Conclusions

Face-voice matching refers to the process of identifying a person by both their voice and their facial features.

The process of face-voice matching involves analyzing both the audio and visual characteristics of a person's identity. Voice analysis involves features such as pitch, accent, tone, and speech patterns, while facial analysis may involve features such as facial structure, vocal tract size, and gestures.

In this study, face-voice matching is tested according to two different variables. The variables are the static stimuli as well as the dynamic stimuli. They were tested separately. In the static stimuli group, it was found that out of the 10 items, the participants failed to successfully select the true speaker in 3 items only. After gathering all the results of the static stimuli group, it was found that 60.3% of the participants successfully selected the true speaker and 39.7% of the participants failed to select the true speaker.

With the dynamic stimuli, there was a lot of information conveyed by the speakers and that information and cues helped the participants to select the true speaker and sometimes the opposite. After gathering the findings of this group, it turns out that the participants failed to correctly identify the true speaker for one item only out of ten. The total percentage of the participants who successfully selected the true speaker is 67.24% and 32.76% failed to select the true speaker.

One can notice that the dynamic stimuli gave more accurate results than the static stimuli.

The hypotheses of the current study which state that “Voice can provide information about a speaker’s face.”, “Dynamic and static face stimuli can be used to facilitate face-voice matching.” “Dynamic stimuli give more accurate results than static stimuli” were answered and verified respectively and hence they are accepted.

References

1. Bernstein, L. E., Tucker, P. E., & Demorest, M. E. (2000). Speech perception without hearing. Percept. Psychophys. 62, 233–252. doi: 10.3758/BF03205546
2. Dobs, K., Bülthoff, I., & Schultz, J. (2016). Identity information content depends on the type of facial movement. Sci. Rep. 6, 1–9. doi: 10.1038/srep34301
3. Ellis, A.W. (1989) Neuro-cognitive processing of faces and voices. In Handbook of Research on Face Processing (Young, A.W. and Ellis, H.D., eds), pp. 207–215, Elsevier
4. Girges, C., Spencer, J., & O'Brien, J. (2015). Categorizing identity from facial motion. Q. J. Exp. Psychol. 68, 1832–1843. doi: 10.1080/17470218.201993664 Gold, J. M., Barker, J. D., Barr, S., Bittner, J. L., Bromfield, W. D., Chu, N., et al.
5. Hartman, D.E. & Danahuer, J.L. (1976) Perceptual features of speech for males in four perceived age decades. J. Acoust. Soc. Am. 59, 713–715
6. Hill, H., & Johnston, A. (2001). Categorizing sex and identity from the biological motion of faces. Curr. Biol. 11, 880–885. doi: 10.1016/S0960-9822(01)00243-3
7. Kaulard, K., Cunningham, D. W., Bülthoff, H. H., & Wallraven, C. (2012). The MPI facial expression database — a validated database of emotional and conversational facial expressions. PLoS ONE 7:e32321. doi: 10.1371/journal.pone.0032321.s002
8. Lass, N.J. et al. (1976) Speaker sex identification from voiced, whispered, and filtered isolated vowels. J. Acoust. Soc. Am. 59, 675–678
9. Nummenmaa, L., & Calder, A. J. (2009). Neural mechanisms of social attention. Trends Cogn. Sci. 13, 135–143. doi: 10.1016/j.tics.2008.12.006
10. O'Toole, A. J., Roark, D. A., & Abdi, H. (2002). Recognizing moving faces: a psychological and neural synthesis. Trends Cogn. Sci. 6, 261–266. doi: 10.1016/S1364-6613(02)01908-3
11. Papcun, G. et al. (1989) Long-term memory for unfamiliar voices. J. Acoust. Soc. Am. 85, 913–925
12. Rosenblum, L. D., Yakel, D. A., Baseer, N., Panchal, A., Nodarse, B. C., & Niehus, R. P. (2002). Visual speech information for face recognition. Percept. Psychophys. 64, 220–229. doi: 10.3758/BF03195788
13. Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what i am saying? exploring visual enhancement of speech comprehension in noisy environments. Cereb. Cortex 17, 1147–1153. doi: 10.1093/cercor/bhl024
14. Russell, J. A. (1994). Is there universal recognition of emotion from facial expressions? A review of the cross-cultural studies. Psychol. Bull. 115, 102–141. doi: 10.1037/0033-2909.115.1.102
15. Scherer, K.R. (1995) Expression of emotion in voice and music. J. Voice 9, 235–248