Matching Face with Voice Identity Using Static and Dynamic Stimuli in Accordance with Ethnicity*

Mustafa Rafid Abdul-Ally
Dept. of English, College of Education for Human Sciences, University of Basrah
Mus.ra.ab@gmail.com

Prof. Dr. Balqis I. G. Rashid
Dept. of English, College of Education for Human Sciences, University of Basrah
Balqis.Gatta@uobasrah.edu.iq

مطابقة الوجه مع الصوت باستعمال المحفزات الثابتة والديناميكية بحسب العرق

مصطفى رافد عبد العالي
قسم اللغة الإنكليزية ـ كلية التربية للعلوم الإنسانية ـ جامعة البصرة

الاستاذة الدكتورة بلقيس عيسى كاطع
قسم اللغة الإنكليزية ـ كلية التربية للعلوم الإنسانية ـ جامعة البصرة

Abstract:

The current study tackles the ability of humans to match faces with voices depending on static and dynamic stimuli with ethnicity. The first aim of the current study is the matching of face with voice using the static stimuli with the ethnicity and the second aim is the matching a face with a voice using dynamic stimuli with ethnicity and to find out which of the two gives a more accurate matching. 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 with the ethnicity gave more accurate results than the static ones.

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

الملخّص:

تتناول الدراسة الحالية قدرة البشر على مطابقة الوجوه مع الأصوات بناءً على المحفزات الثابتة والديناميكية بحسب العرق. الهدف الأول للدراسة الحالية هو اثبات إمكانية مطابقة الوجه مع الصوت باستخدام المحفزات الثابتة بحسب العرق، والهدف الثاني هو تقصي إمكانية مطابقة الوجه مع الصوت باستخدام المحفزات الديناميكية مع العرق ومعرفة أيًّا منهما يعطي نتائج مطابقة أكثر دقة. في هذه الدراسة، تم اعتماد النموذج المعاصرة. توصلت النتائج إلى **اختيار** المشاركين بالضبط المتحدثين الحقيقيين في معظم الفقرات التي عرضت لهم. أما المحفزات الديناميكية الخاصة بالعرق، فأعطت نتائج أكثر دقة مقارنة بالمحفزات الثابتة.

الكلمات المفتاحية: مطابقة الوجه والصوت، المحفزات الثابتة، المحفزات الديناميكية، المقاربة المتقاطعة، العرق.

1. Introduction

Both of the dynamic and static features of our voices, like timbre, are directly affected by many factors like age and gender, and 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 is 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 is directly related to the state of 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. Aims

The current study tries to achieve the following aims:

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

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

3- Exploring if by any means an individual’s ethnicity can be guessed or determined from his/her voice.

3. Hypotheses

The current study hypothesizes 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.

2- Voice can be used to define the ethnicity of a speaker.

4. Static and Dynamic Stimuli

Dobs, Bülthoff, & Schultz (2018, p.1) state that the 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 20 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 5 items which are represented as 10 videos, and the other 5 items which are represented as 10 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. Description of the experiment

The first part of the experiment is about "ethnicity and dynamic stimuli" and it consists of showing the participants two muted videos of two different persons, one of whom is an African American and the other is a white American, and asking them to use their knowledge of the persons' ethnicities to identify the true speaker. There are 5 items in this part. In the second part, the participants are shown two pictures of people of different ethnicities who are speaking a common language. There are 5 items in this part as well.

8.1 Ethnicity and Dynamic Stimuli

Item No. 1

In this item, A is an African American and B is a white American, depending on the voice of the true speaker the participants are asked whether they can select the true speaker or not. Here, the main variable is ethnicity; the participants will depend on it to determine the true speaker.

Figure 1indicates that the speaker has a modal voice, and from the spectrogram, we can tell that the F_0 range falls between 70 Hz – 275 Hz.

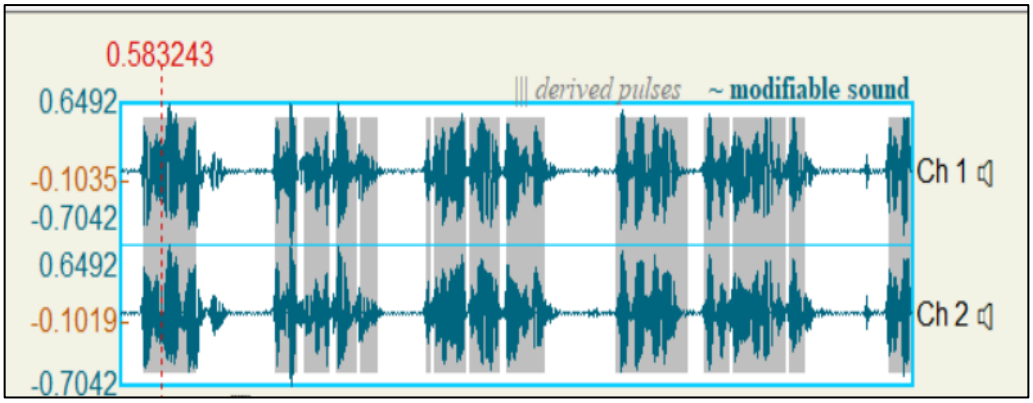
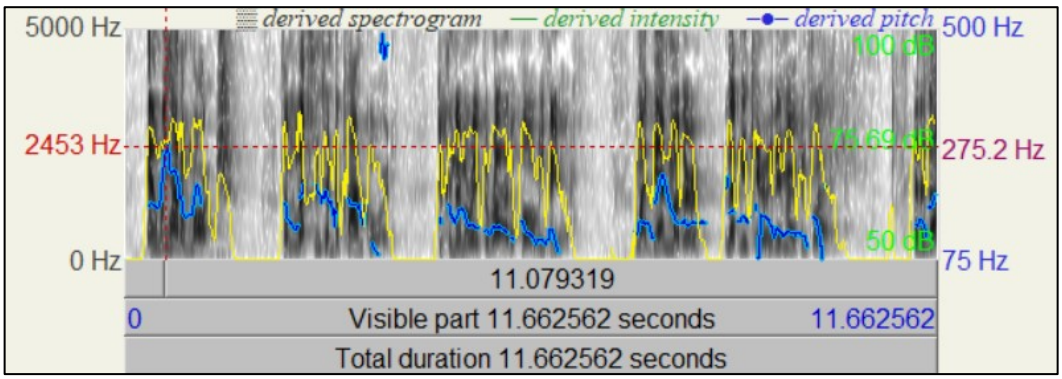


Fig. 1: Waveform of the voice of item no.1



Spectrogram 1: The spectrogram of item no. 1

Table 1: Item no.1 results

Item1					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		1	.3	.3	.3

	A	282	94.0	94.0	94.3
	B	17	5.7	5.7	100.0
	Total	300	100.0	100.0	

Out of the 300 participants, most of them (94.0%) selected A as the true speaker, while only a small proportion (5.7%) selected B.

Among the reasons given for selecting A or B, the most common reason (chosen by 38.7% of participants) was that A's voice sounds like that of an African American person, while the second most common reason (chosen by 20% of participants) was that A's voice is thicker than B's voice. R6 which states that "B's voice is thicker than A's" was selected by 5 participants (1.7) which is the least selected reason in this item.

It is also worth noting that in this item the main reason for selecting A as the true speaker was his ethnicity.

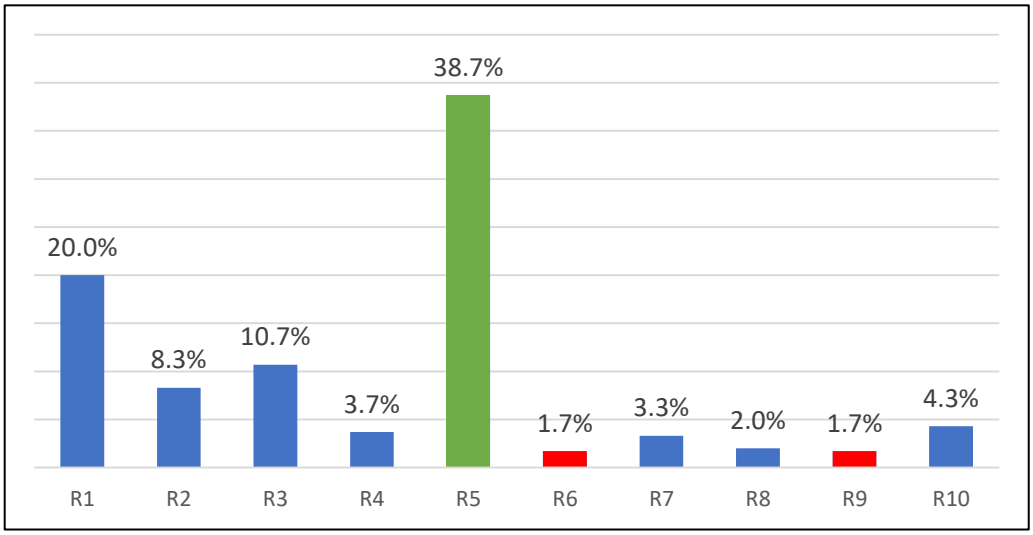


Chart 1: The selected reasons for answering item no.1

Item No. 2

In this item, there are two women from different ethnicities. A is a white American person and B is an African American one. In terms of the physical characteristics, both A and B are almost similar, they are even of almost the same age.

In figure 2, we can notice that the true speaker has a modal voice although there are some words or sounds that have high tones, and in the spectrogram, we can see that the blue line is so high compared to the previous item. The F_0 range falls between 120 Hz – 318 Hz, and this range is an average range for females.

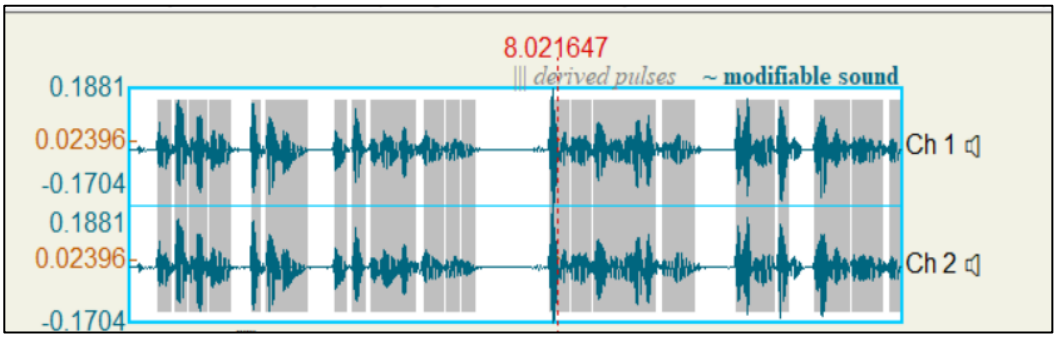
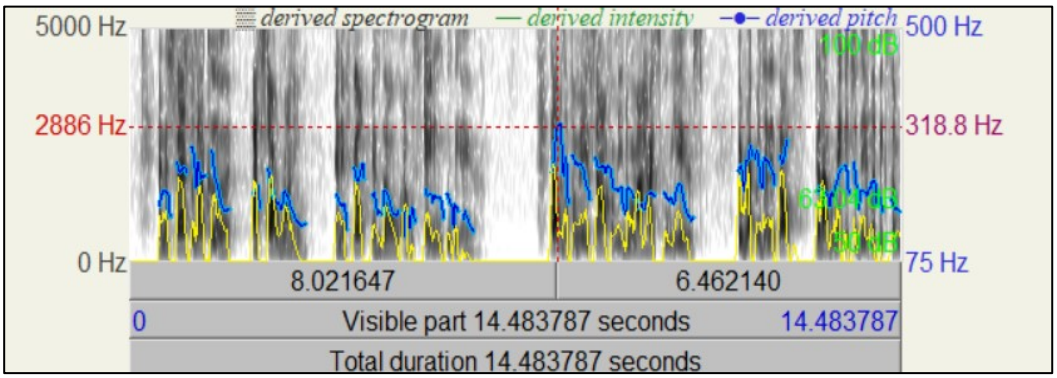


Fig. 2: Waveform of the voice of item no.2



Spectrogram 2: The spectrogram of item no.2

Based on the following table, 84.3% of participants selected B as the true speaker, while 15.3% of participants selected A as the true speaker. The true speaker in this item is B.

Table 2: Item no.2 results

Item2					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		1	.3	.3	.3
	A	46	15.3	15.3	15.7
	B	253	84.3	84.3	100.0

	Total	300	100.0	100.0	

As shown in the following chart, the most common reason for selecting B as the true speaker was that her voice sounds like that of an African American woman (46.0% of participants) which is consistent with the true speaker of this item. The second most common reason is that B's voice is thicker than A's voice (7.7% of participants).

Other reasons selected by the participants include A's voice being thicker than B's (7 responses or 2.3%), the vocal tract of A being larger than B's (15 responses or 5%), and B being older than A (5 responses or 1.7%).

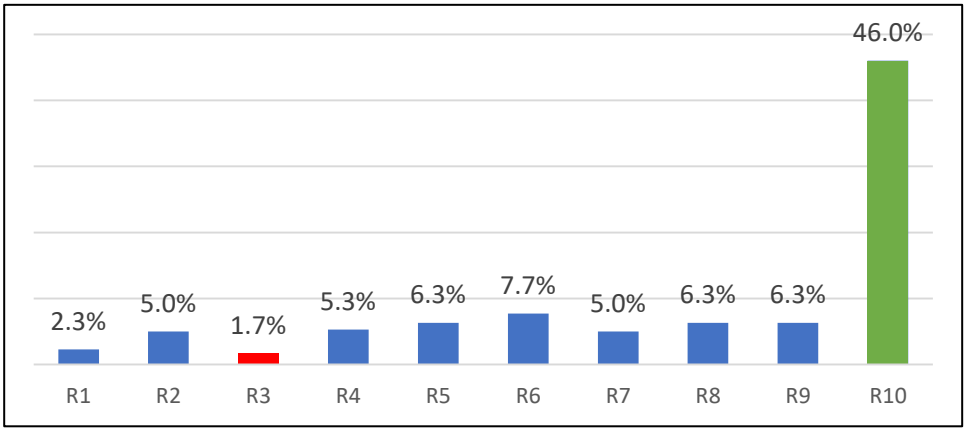


Chart 2: The selected reasons for answering item no.2

Item No. 3

In this item, the true speaker is a white American (A), and B is an African American. The age difference between them is not that obvious.

In Figure 3, one can see that the voice in this item is different from the last two items; the true speaker in this item has a creaky voice. In the spectrogram, it can be noticed that the blue line goes up and then suddenly drops down. The F_0 range falls between 73 Hz – 366 Hz.

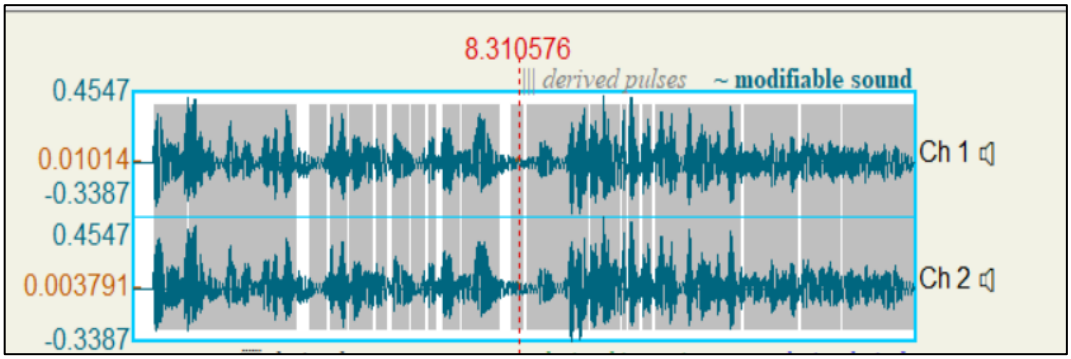
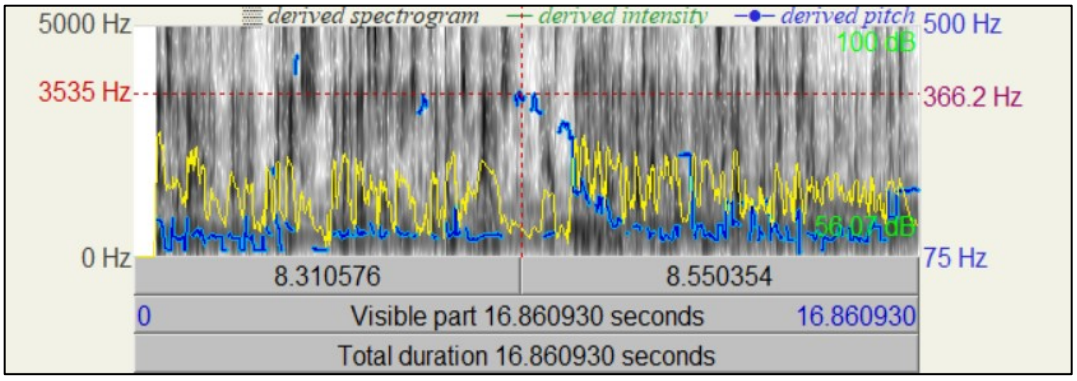


Fig. 3: Waveform of the voice of item no.3



Spectrogram 3: The spectrogram of item no.3

Table 3: Item no.3 results

Item3					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A	211	70.3	70.3	70.3
	B	89	29.7	29.7	100.0
	Total	300	100.0	100.0	

Table 3 confirms that out of the 300 participants, 211 (70.3%) selected A as the true speaker, while 89 (29.7%) selected B. This shows that many

of the participants were able to correctly select A as the true speaker based on the reasons in the following chart.

The most common reason selected was R10 (B's voice sounds like an African American), with 55 participants (18.3%) selecting this reason. R4 (A's voice is slower than B's voice) and R8 (B looks older than A) were also selected by a number of the participants, with 40 (13.3%) and 38 (12.7%) participants selecting these reasons respectively. R1, R2, R3, R5, R6, R7, and R9 were selected by fewer participants, with percentages ranging from 13% to 27%.

The results show that the participants relied on a variety of factors when making their decision, with some factors being more common than others. However, most of the participants were able to correctly identify A as the true speaker depending on the ethnicity of the speaker.

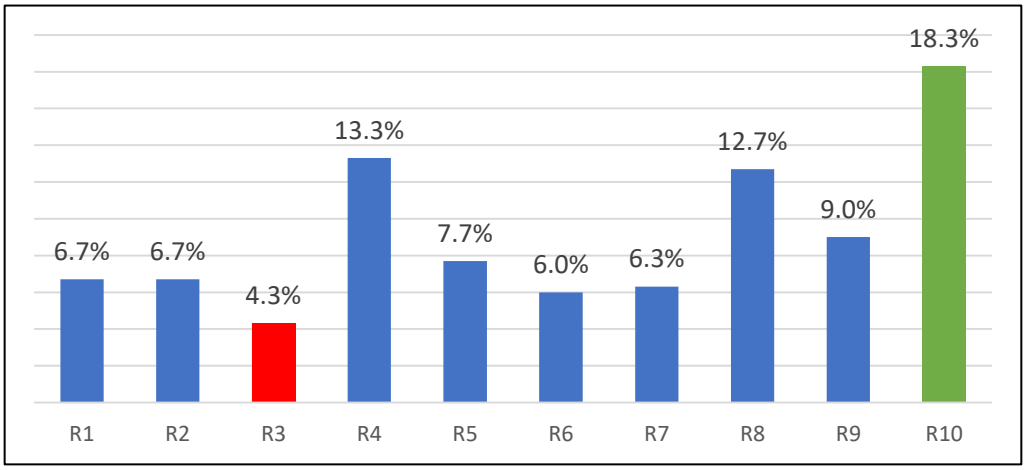


Chart 3: The selected reasons for answering item no.3

Item No. 4

In item no. 34, the first video is about an African American young woman in her 20s while B is about a white American in her 30s. Both tend to use a lot of facial gestures and hands movement.

From the following figure, we can tell that the true speaker has a creaky voice, and in the spectrogram, we can notice that the blue line keeps

getting up and then drops down suddenly, and this is a sign of creakiness. The F_0 range falls between 74 Hz - 342 Hz.

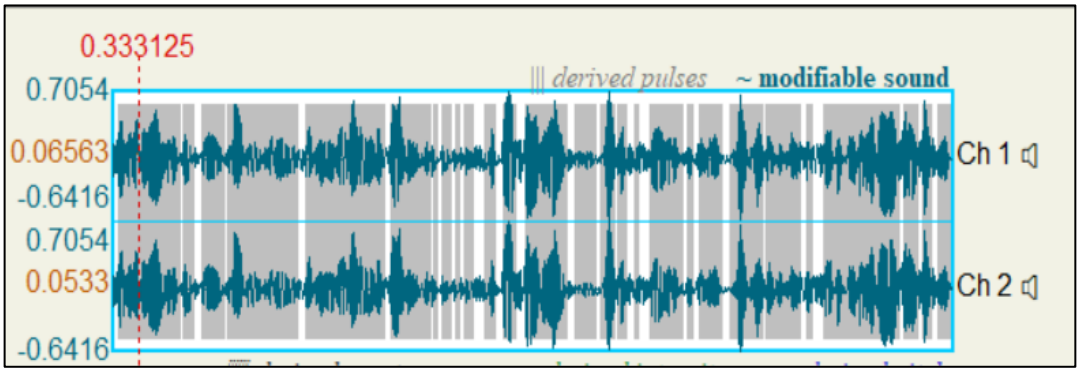
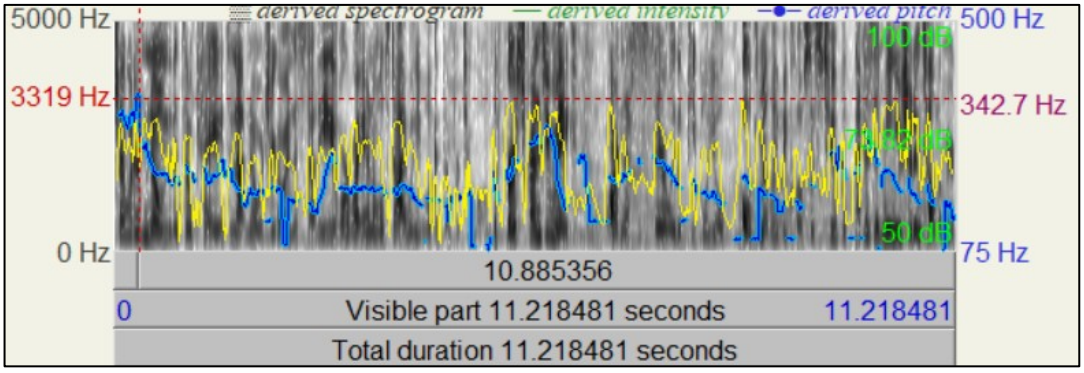


Fig. 4: Waveform of the voice of item no.4



Spectrogram 4: The spectrogram of item no.4

Out of the 300 participants, 181 (60.3%) correctly selected speaker A as the true speaker, while 119 (39.7%) incorrectly selected speaker B as the true speaker.

Table 4: Item no.4 results

Item4					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A	181	60.3	60.3	60.3
	B	119	39.7	39.7	100.0
	Total	300	100.0	100.0	

In chart 4, we can see that the reasons selected by the participants for making their choices of the true speaker varied. R5 (A's voice sounds like an African American) was the most selected reason, with 82 (27.4%) participants selecting it. R5 (B's voice sounds like an than of an African American) was the second most selected reason with 46 (15.3%) participants selecting it. The least selected reason was R3 (A looks older than B), with only 9 (3.0%) participants selecting it.

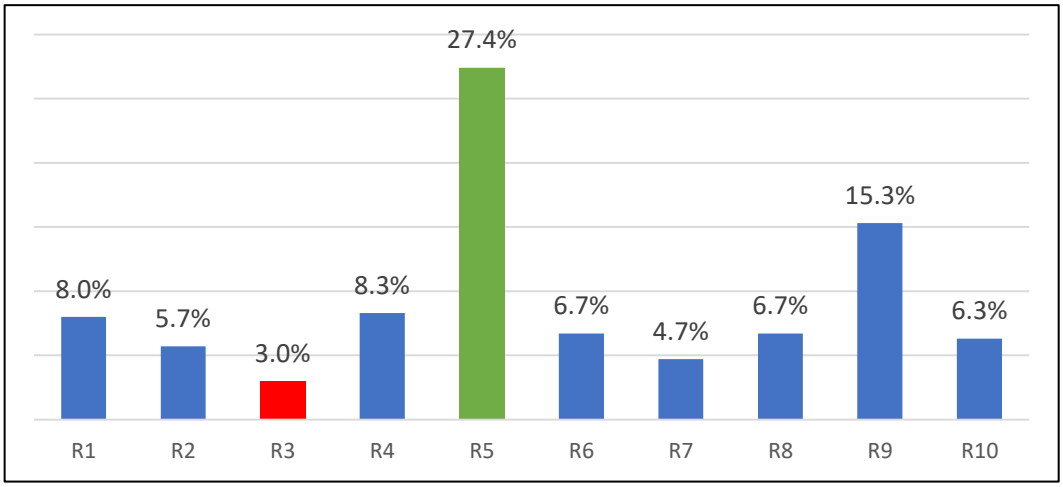


Chart 4: The selected reasons for answering item no.4

Item No. 5

In this item A is an African American woman and B is a white American woman. Both are giving a speech on a stage, and both used a lot of facial gestures and hands movement.

In figure 5, we can see that the true speaker has a modal voice but with some high tones, and in the spectrogram, we can notice that the blue line is higher than the previous item. The F_0 range falls between 78 Hz – 354 Hz.

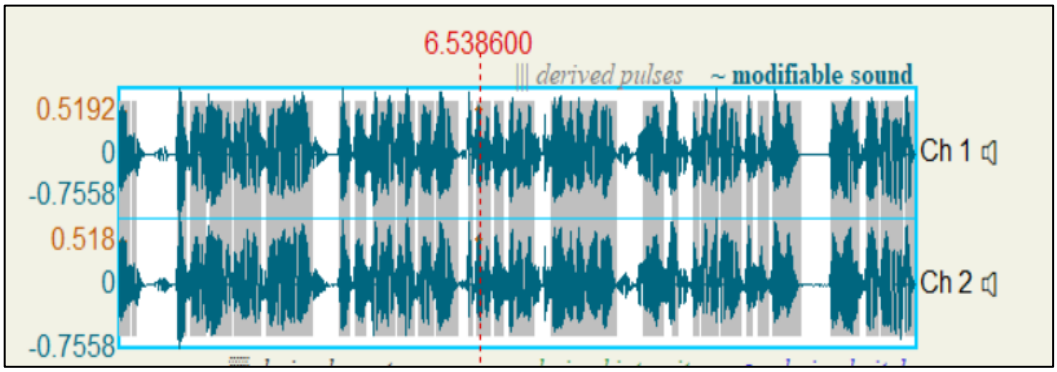
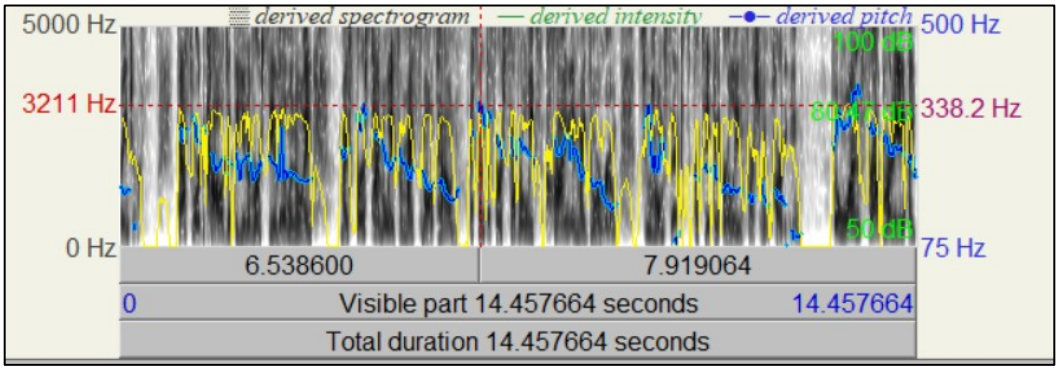


Fig. 5: Waveform of the voice of item no.5



Spectrogram 5: The spectrogram of item no.5

In the following table, we can tell that out of the 300 participants, 69% (207) correctly identified speaker B as the true speaker, 30.3% (91) incorrectly chose speaker A, while only 0.7% (2 participants) did not answer this item.

Table 5: Item no.5 results

Item5					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		2	.7	.7	.7
	A	91	30.3	30.3	31.0

	B	207	69.0	69.0	100.0
	Total	300	100.0	100.0	

From chart 5, we can see that the most selected reason is R5 which states that "A's voice sounds like an African American "; it was selected by 68 participants (22.7%). The participants here are certain that the true speaker is a white woman and A is that of an African American woman that is why they excluded the chance that A is the true speaker by selecting that her voice should sound like an African American. This is consistent with the fact that the true speaker was B, who is a white American woman. The second most selected reason was R6 ("B's voice is thicker than A's voice") which was selected by 35 participants (11.7%). Other reasons included R6 ("B's voice is thicker than A's voice") and R9 ("B's voice is slower than A's voice"), which were selected by 35 (11.7%) and 29 (9.7%) participants, respectively.

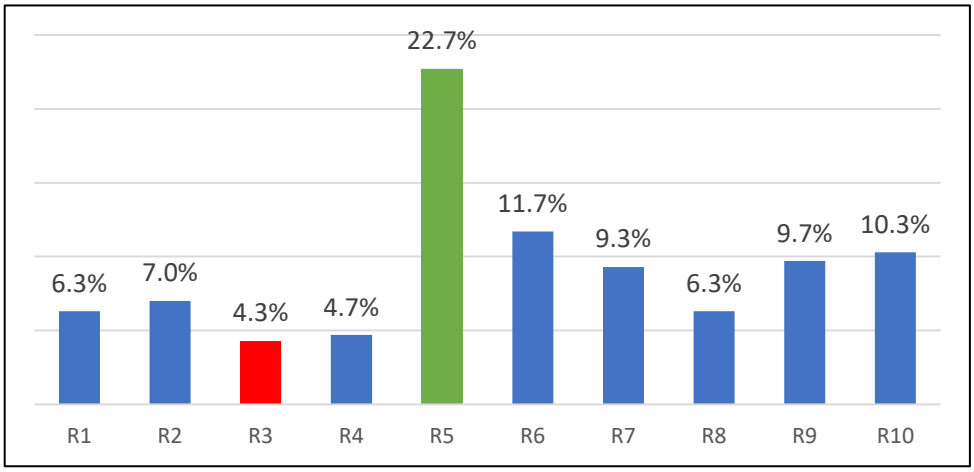


Chart 5: The selected reasons for answering item no.5

8.2 Ethnicity and Static Stimuli

Item No. 6

In item no.6, there are two pictures of two persons (A and B). A is a white American person and B is an African American one. The differences between them are obvious. The true speaker in this item is B.

In the following figure, one can notice that the true speaker has a hoarse (thick) voice, and in the spectrogram, we can see that the blue line is so low and the F_0 range falls between 75 Hz – 226 Hz.

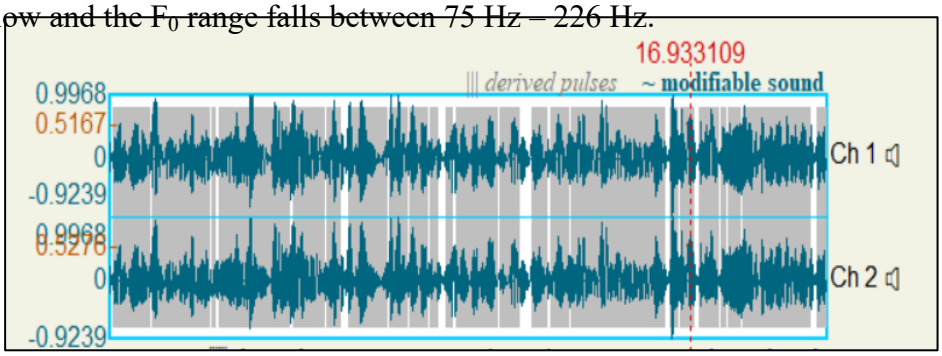
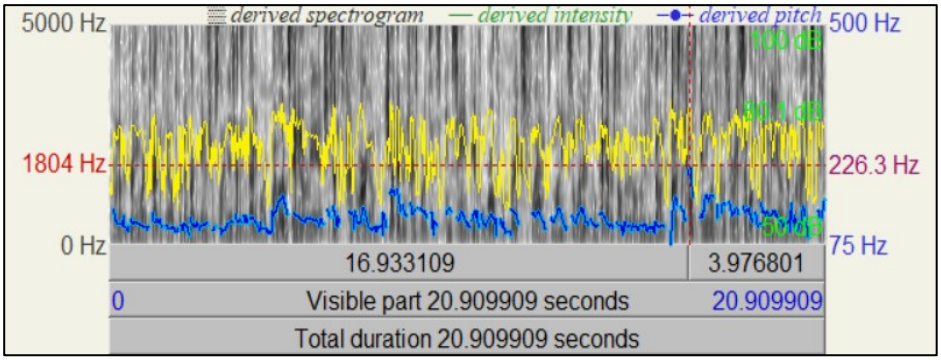


Fig. 6: Waveform of the voice of item no.6



Spectrogram 6: The spectrogram of item no.6

Pic. 1: Item no.6 related faces

Picture (1) above displays that the differences between the two persons are so obvious. Both of them show different kinds of facial gestures, but no one shows any hand movement. Both are paused at a certain point in time.

Table 6: Item no.6 results

Item6					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A	213	71.0	71.0	71.0
	B	87	29.0	29.0	100.0
	Total	300	100.0	100.0	

Table 6 shows the results which are as follows: out of the 300 participants, 213 (71%) incorrectly selected the true speaker as A (the white American), while 87 participants (29%) correctly selected B (the African American) as the true speaker.

Chart 6 indicates that the most selected reason by the participants was that “B’s voice sounds like an African American one” which was selected by 49 participants (16.3%). The second most selected reason is R6 which states that "B's voice is thicker than A's voice," which was chosen by 39 participants (13%). Other reasons included "A's voice is slower than B's voice" (19 participants, 6.3%) and "B's voice sounds like an African American " (26 participants, 8.7%). The results show that the participants did not correctly select the true speaker.

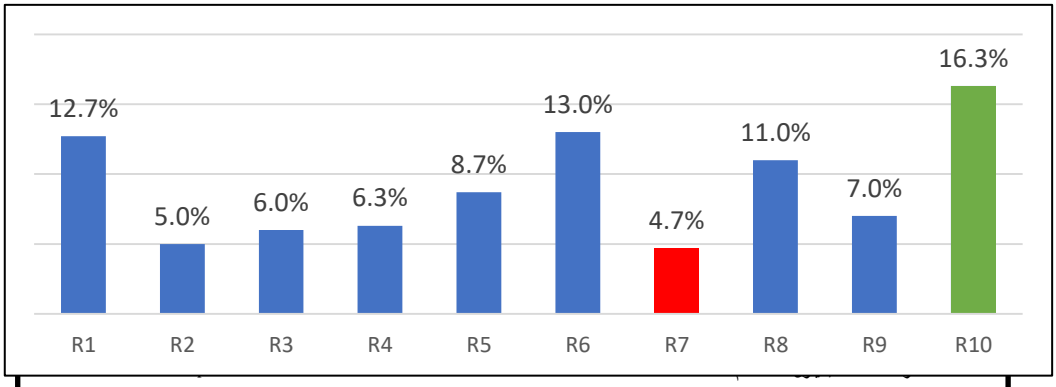


Chart 6: The selected reasons for answering item no.6

Item No. 7

In this item, A is a young African American person and B is a young white American. They are almost of the same age. Both do not show any kind of hand movement.

From the following figure, one can tell that this is a modal voice, and in the spectrogram, we could notice that the blue line is normal; it is not so high or low. The F_0 range falls between 75 Hz – 164 Hz.

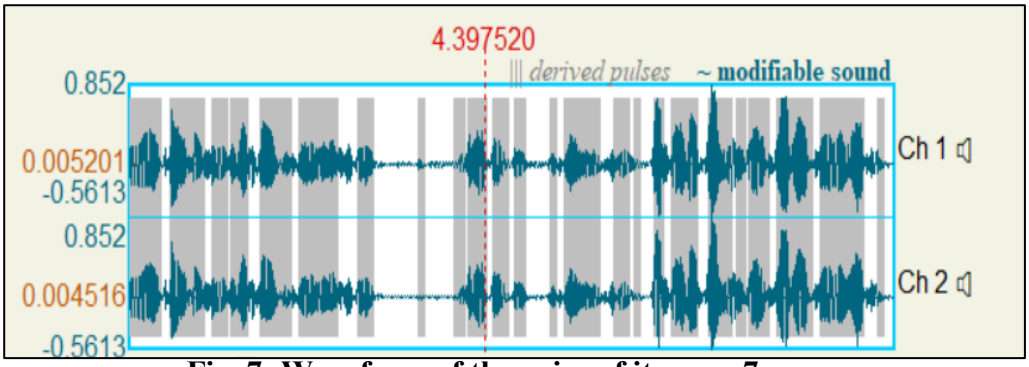
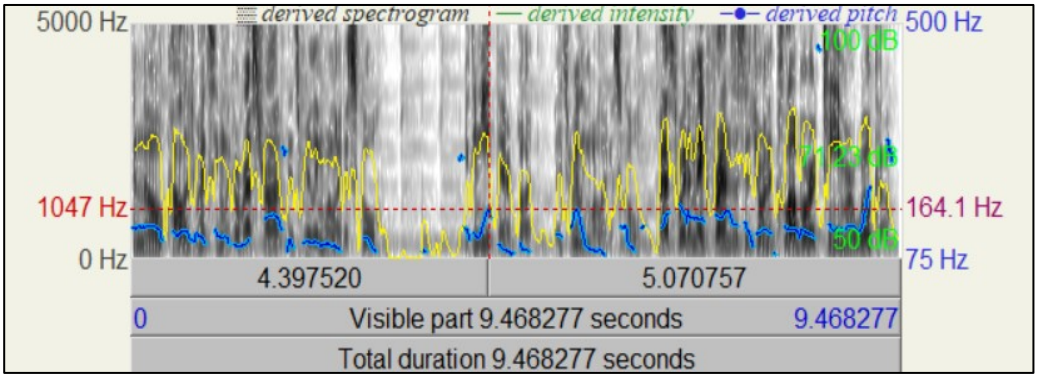


Fig. 7: Waveform of the voice of item no.7



Spectrogram 7: The spectrogram of item no. 7

Pic. 2: Item no.7 related faces

The following table shows the frequency and percentage of the participants who correctly identified A as the true speaker and those who selected B. Out of the 300 participants, 181 (60.3%) correctly identified A while 118 (39.3%) selected B.

Table 7: Item no.7 results

Item7					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		1	.3	.3	.3
	A	181	60.3	60.3	60.7
	B	118	39.3	39.3	100.0
	Total	300	100.0	100.0	

Chart 7 shows that the most common reason that was selected was R5 ("A's voice sounds like an African American ") with 139 participants (46.3%) selecting it. The second most common reason was R8 ("B looks older than A") with 27 participants (9%) selecting it. The least common

reason was R3 ("A looks older than B") with only 4 participants (1.3%) selecting it.

Table 7 and chart 7 suggest that the participants may have relied on different factors when deciding on which individual was the true speaker. The most common reason was that A's voice sounds like an African American. These findings may have implications for how people perceive and interpret voices based on ethnicity and other factors.

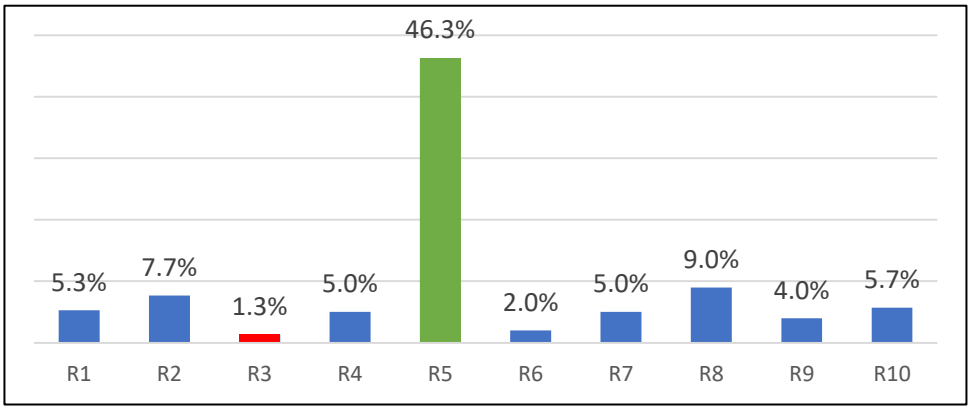


Chart 7: The selected reasons for answering item no.7

Item No. 8

In this item, A is a white American and B is an African American. They are almost of the same age, and neither A nor B shows any kind of facial gestures or hand movements.

The following figure shows that the true speaker has a modal voice, and the spectrogram displays that the blue line goes so high and that is because the true speaker is a young woman. The F_0 range falls between 77 Hz – 404 Hz.

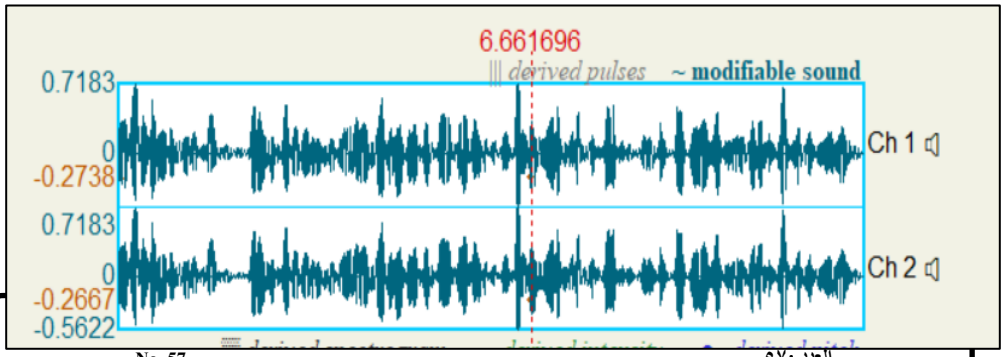
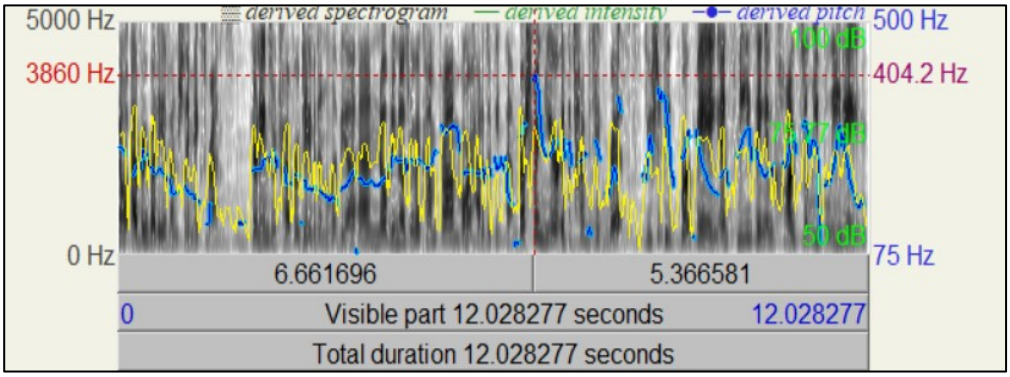


Fig. 8: Waveform of the voice of item no.8



Spectrogram 8: The spectrogram of item no.8

Picture 3 shows that both speakers paused at a certain point while they were speaking. It seems that both have almost the same vocal tract size.

Pic. 3: Item no.8 related faces

Table 8: Item no.8 results

Item8					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A	207	69.0	69.0	69.0
	B	93	31.0	31.0	100.0
	Total	300	100.0	100.0	

Table 8 above illustrates that out of the 300 participants, most of the participants, 207 (69.0%), selected A as the true speaker, while 93 (31.0%) selected B. This indicates that the participants successfully selected A as the true speaker.

From the following chart, one could tell that the most common reason chosen by the participants was R10 which states that “B’s voice sounds like an African American voice", it was selected by 79 (26.3%) participants.

The next most common reasons were R4 which states that “A's voice is slower than B's voice” (30 participants or 10%) and R7 which states that “The vocal tract of B is larger than A’s” (29 participants or 9.7%).

The reasons chosen by the participants show that their decision was influenced by the physical characteristics of the speakers, such as voice pitch, vocal tract size, and the perceived ethnicity.

The following chart provides insight into the participants' decision-making process when identifying the true speaker based on voice recordings and visual cues. The results indicate that the physical characteristics, such as voice pitch and perceived ethnicity, play a significant role in the participants' decisions.

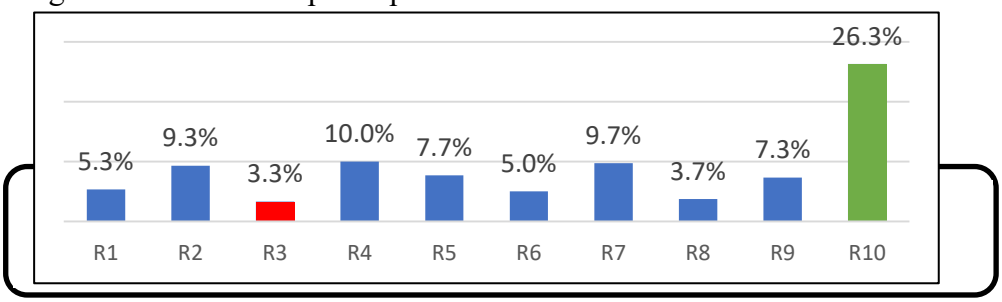


Chart 8: The selected reasons for answering item no.8

Item No. 9

In this item, A is an African American man and B is a white American one; the ethnic difference is so obvious. B shows some hand movements while B does not show any.

Figure 9 shows that the speaker has a modal voice, and in the following spectrogram, it is noticeable that the blue line is neither so low nor so high. The F_0 range falls between 75 Hz – 212 Hz.

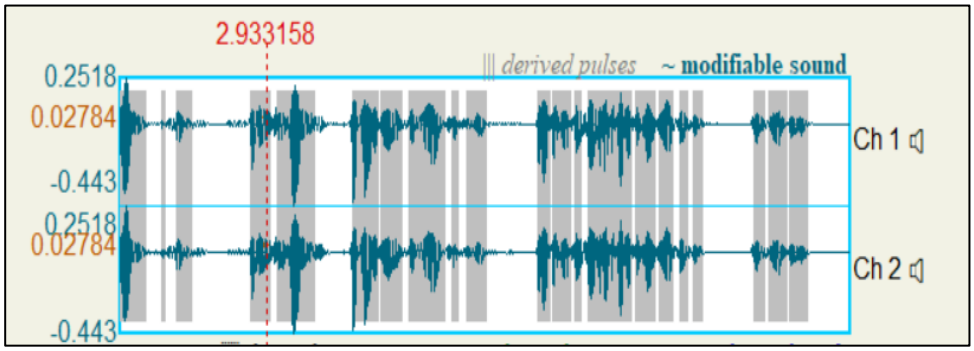
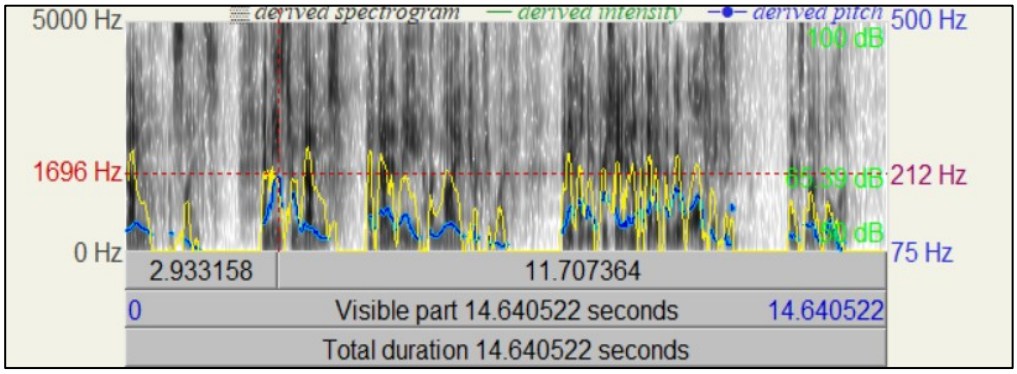


Fig. 9: Waveform of the voice of item no.9



Spectrogram 9: The spectrogram of item no. 9

Pic. 4: Item no.9 related faces

The ethnic difference between the two men is so obvious and we can see that B shows some hand movements and some facial gestures.

Table 9: Item no.9 results

Item9					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A	61	20.3	20.3	20.3
	B	239	79.7	79.7	100.0
	Total	300	100.0	100.0	

In table 9, we can see that out of the 300 participants, 239 (79.7%) selected B as the true speaker while only 61 (20.3%) selected A. This suggests that most of the participants correctly identified B as the true speaker.

In the following chart, we can see the frequency and percentage of reasons selected by the participants for making up their decision of selecting speaker A or B. The most common reason selected was R9 which states that "B's voice is slower than A's" with a frequency of 58 (19.3%). Other frequently selected reasons were R8 which states that "B looks older than A" with a frequency of 51 (17.0%) and R5 which states that "A's voice sounds like an African American " with a frequency of 45 (15.0%). The participants seemed to rely on a combination of vocal characteristics and perceived age and ethnicity to make their decision. This shows that they relied heavily on the ethnicity of the speaker as well as the speech rate of the voice to make up their decision.

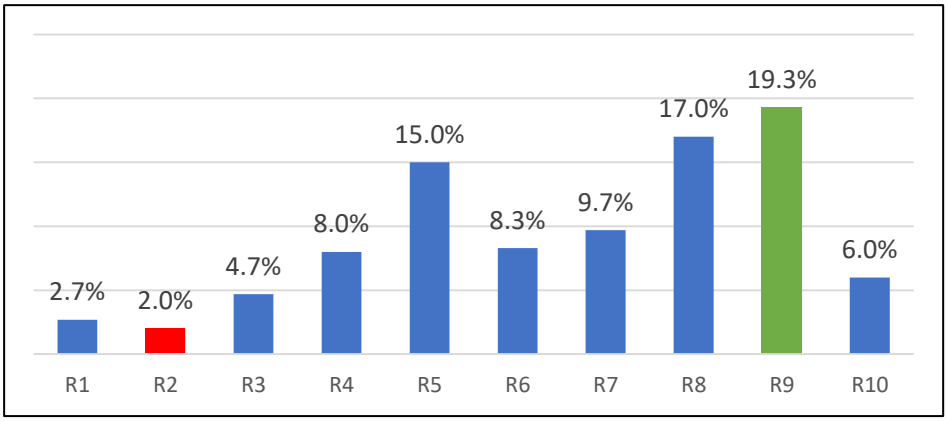


Chart 9: The selected reasons for answering item no.9

Item No. 10

In this item, A is an African American woman in her 30s and B is a white American in her 30s too. The difference is so obvious between the two women.

From figure 10, we can tell that the voice of the true speaker is a modal voice, it has a lot of high areas. From the spectrogram, we can tell that the blue line is so high and the F_0 range falls between 100 Hz – 425 Hz.

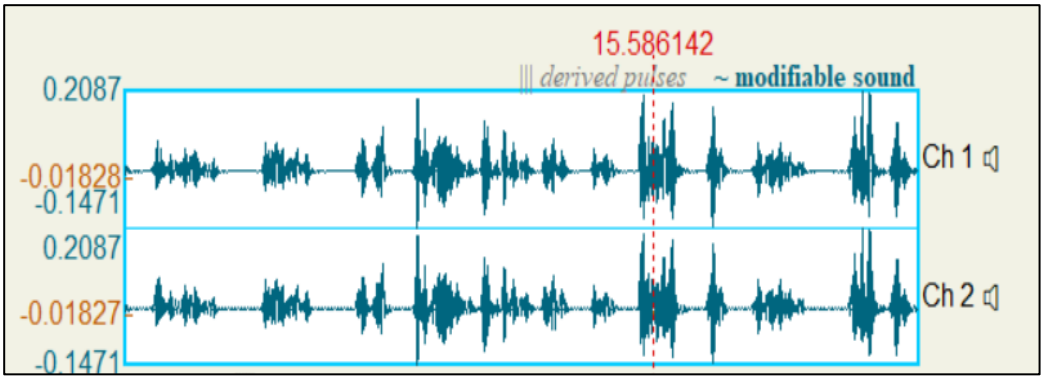
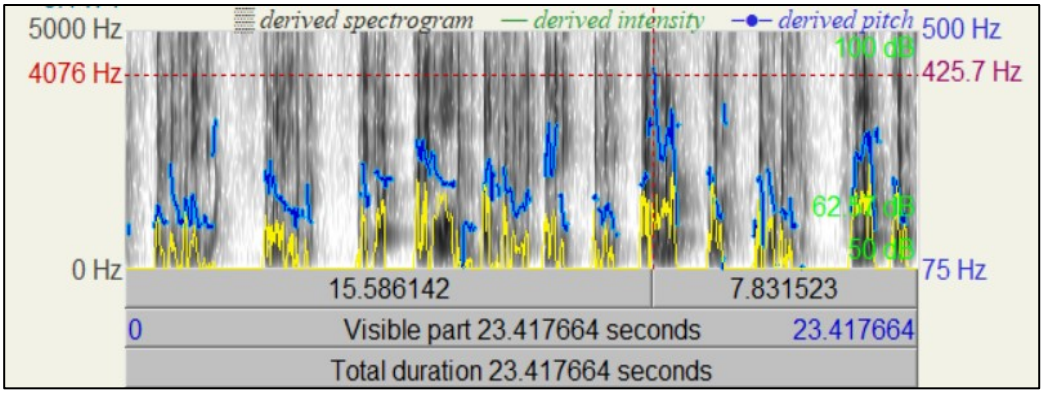


Fig. 10: Waveform of the voice of item no.10



Spectrogram 10: The spectrogram of item no. 10

Pic. 5: Item no.10 related faces

Picture 5 shows that A does not show any kind of hand movements and facial gestures, unlike B.

Table 10: Item no.10 results

Item10					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid		2	.7	.7	.7
	A	207	69.0	69.0	69.7
	B	91	30.3	30.3	100.0
	Total	300	100.0	100.0	

In table 10, it appears that a good number of the participants (69%) correctly selected speaker A, while the remaining 30.3% incorrectly selected B. Two of the participants (0.7%) did not choose A or B.

Looking at the reasons given for their choices in chart 10, the most selected reason was that A's voice sounds like an African American one (41% of participants). Other reasons such as R3 “A looks older than B” was selected by 17 participants (5.7%), R4 which states that “A's voice is slower than B's voice” was selected by 17 participants (5.7%).

These results imply that a good number of the participants were able to accurately identify the true speaker based on their voice and the images provided and that the most selected factor in their decision-making was the perceived ethnic identity of the speaker.

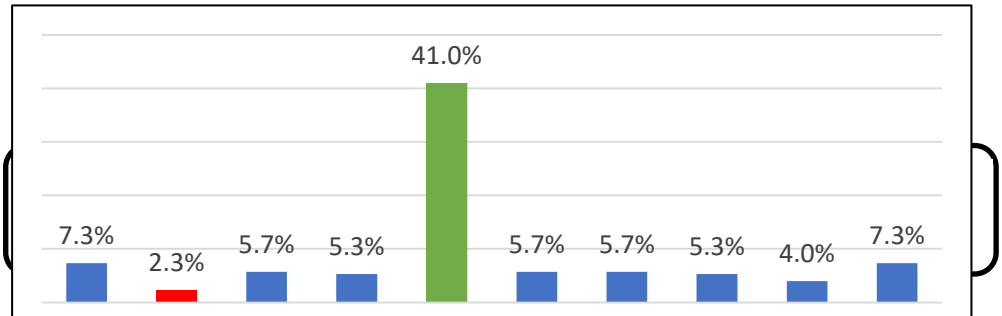


Chart 10: The selected reasons for answering item no.10

Conclusion

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

According to the tables and charts above, it was that most of the participants answered the above items correctly, they were able to correctly select the true speakers; in the first variable (dynamic stimuli), all the items were answered correctly and the same with the static stimuli. In the ethnicity and dynamic stimuli part, the total percentage of the participants who successfully selected the true speaker was 75.78% and 24.22% failed to select the true speaker. On the other hand, in the ethnicity and static stimuli part, it was found that only one item was answered incorrectly out of five. The total percentage of the participants who selected the true speaker was 61.50% and 38.50% failed to select the true speaker.

One can notice that the dynamic stimuli with ethnicity gave more accurate results compared to the static stimuli with ethnicity.

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” and “Voice can be used to define the ethnicity of the speaker” were answered and verified respectively and hence they are accepted.

References

Bernstein, L. E., Tucker, P. E., & Demorest, M. E. (2000). Speech perception without hearing. Percept. Psychophys. 62, 233–252. doi: 10.3758/BF03205546

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

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

Girges, C., Spencer, J., & O'Brien, J. (2015). Categorizing identity from facial motion. Q. J. Exp. Psychol. 68, 1832–1843. doi: 10.1080/17470218.2014.993664 Gold, J. M., Barker, J. D., Barr, S., Bittner, J. L., Bromfield, W. D., Chu, N., et al.

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

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

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

Lass, N.J. et al. (1976) Speaker sex identification from voiced, whispered, and filtered isolated vowels. J. Acoust. Soc. Am. 59, 675–678

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

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

Papcun, G. et al. (1989) Long-term memory for unfamiliar voices. J. Acoust. Soc. Am. 85, 913–925

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

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

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

Scherer, K.R. (1995) Expression of emotion in voice and music. J. Voice 9, 235–248

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