Speaker Identification Based on Temporal Parameters in Colloquial Arabic

Eman M. Yousri[*1], Mervat Fashal[**2]

[*]*Post-graduate Student of Phonetics & Linguistics Dep., Faculty of Arts, University of Alexandria. Alexandria, Egypt.*

[1]emanyousri88@yahoo.com

[**]*Professor of Phonetic Science, Phonetics & Linguistics Dep., Faculty of Arts, University of Alexandria. Alexandria, Egypt.*

[2]mervat.fashal@alexu.edu.eg

Abstract: ***The subject of this study is to identify unknown speakers particularly from their speaking tempo represented in Speech Rate SR and Articulation Rate AR as temporal parameters. The fundamental goal of this study, on the acoustical level, is to prove acoustically that every speaker has a significant speech rate SR and articulation rate AR through which the unknown speaker can be discriminated and to investigate which of them (SR or AR) could be of more benefit for identifying unknown speakers and to what extent. Also, the present study is essentially concerned, on the perceptual level, with listeners' perceptual abilities in perceiving and differentiating different speaking tempo for identifying unknown speakers in order to utilize this exceptional ability in forensic speaker identification FSI; aiming to provide some useful acoustical and perceptual data to be used in forensic phonetic filed. The most important characteristic of the temporal aspects of speech, that they are not easily disguised or imitated by accent or fundamental frequency leveling; so they could be useful for identifying unknown speakers particularly in forensic phonetic field.***

The speech rate SR and articulation rate AR of ten unknown speakers / informants of colloquial Arabic are calculated. The speakers were recorded while talking spontaneously for a radio program. Only 30 seconds of speech are cut for each speaker from the entire episode. After that 60 naïve listeners are asked to listen carefully to the 10 unknown informants in order to mark the fastest speaker and the slowest speaker depending only on their ears.

Key words: ***Speaker Identification, Forensic Phonetics, Forensic Speaker Identification, Speech Rate, Articulation Rate, Speaking Tempo.***

1 INTRODUCTION

"*A voice is more than just a string of sounds. Voices are inherently complex. They signal a great deal of information in addition to the intended linguistic message.....*" (Rose 2002)

The human voice carries on the speech signal which is a multidimensional and a very complex acoustic wave. These signals convey the information about the words or message being spoken in order to convey *linguistic information (verbal)* as well as *non-linguistic information*. The non-linguistic information reveals information about the speaker's identity, personality and individuality (Extralinguistic component[1]). It also reveals information about the situation, the inner state and the current health of the speaker, as well as their attitudinal or emotional state (Paralinguistic component[2]). That is, virtually any utterance conveys information on several levels [11], [16], [25] & [28].

Speaker Identification is the task of deciding and determining a given sample of speech (uttered by unknown speaker), who among many candidate speakers said it. The unknown speaker is defined as the speaker whose model best matches the given utterance [7]. Nowadays voice identification analysis has matured into a sophisticated identification technique. The comparison of human voices now focuses on every aspect of the words spoken; the words themselves, the way the words flow together, and the pauses between them. There are two methods of speaker identification as evidence which are commonly applied: ***Naïve speaker identification*** (also called *aural identification*) which relays on human auditory perception and ***Technical speaker identification*** which uses acoustic analysis (through spectrograph) [21], [22] & [25].

[1]Extralinguistic information refers to aspects of the sound that are determined by the particular speaker's vocal tract anatomy and physiology such as their vocal tract length or the volume of their nasal cavity. This component is uncontrolled and involuntary via speakers, and it is also called ***informative language*** because it is telling information about the informant / speaker itself.

[2]Paralinguistic information refers to habitual muscular settings that an individual adopts when they speak, for example; a speaker may habitually speak with slightly rounded lips, nasalization, or a low pitch range. This setting component is under a speaker's control and voluntary. It is also called ***communicative language*** because it reflects the way speakers use language to communicate with others.

Naïve speaker identification involves the application of our natural abilities as human language users to the identification of a speaker (the term "naïve" means here the lack of specific training on the part of the person making the decision; for example: a witness to a crime may claim to identify a voice heard).*Technical speaker identification* is defined by the employment of any trained skill or any technologically-supported procedure in the decision-making process when there is an incriminating recording of a suspect. Both aural and spectrographic (visual) analyses are combined to form the conclusion about the identity of the speaker. Undoubtedly, the auditory method depends mainly on the naïve perceptual ability to recognize speakers' voices. Whereas; the acoustic analysis allows the phonetician expert to measure a number of parameters which reflect a person's vocal organs; including fundamental frequency (pitch), formant frequencies(resonant), and the dimensions of the vocal tract and its pattern of movement [1], [3], [5], [14], [17], [18], [25] & [26].

Forensic Speaker Identification FSI is considered as one of the most significant practical applications of speaker identification. FSI is defined as the most central aspect of forensic phonetics and acoustics which mainly concerned with solving problems related to identification of the unknown speaker in criminal investigation to identify suspects who were heard but not seen committing a crime including; murder, blackmail threats, ransom calls, kidnapping, political corruption, bomb threats, terrorist activities, etc. [5], [12], [17], [20], [22], [25] & [27]. *The fundamental theory of forensic speaker identification* relies primarily on that every voice is individually characteristic enough to distinguish itself through voice print analysis[3]. Many researchers support the theory that human voices are unique and could be used as a mean for identification. Besides, if everyone had the same voice, voices would not be used as discriminate evidence [2], [12], [17], [19], [21] & [25].

There will be always differences (which are always audible, measurable and quantifiable) between speech samples, even if they come from the same speaker. This is due to two kinds of variability: *1) organic vs. phonetic variability*, and *2) between speaker vs. within speaker variability*. Consequently, the main task of Forensic Speaker Identification FSI is to find all the sources of variability in order to make a clear distinction for the correct evaluation.

For speaker identification in forensic situation as evidence in the court, there are four main phonetic/acoustic parameters depending on the speaker through them he / she can be discriminated and identified:

1. The Fundamental Frequency F_0.
2. The formants frequencies of the vowels.
3. The resonance of the nasal consonants.
4. Tempo of speaking.

Tempo of speaking; the fourth parameter is our concern here; it is a multidimensional phenomenon and revealing the temporal aspects of the speech. It is also one of the prosodic cues which is considered as non-linguistic factor that signaling *paralinguistic* information (about the situation and the inner state of the speaker's attitudinal or emotional state) and also *extra-linguistic* information (about the speaker's identity, personality and individuality) [25] & [28]. ***Tempo of speaking*** can be exhibited by two methods, one is ***Speech Rate (SR)***, and another is ***Articulation Rate (AR)***. Both of SR and AR can be defined as "*the number of syllables per second*". The biggest difference between SR and AR is that the SR includes pause intervals but the AR does not [8] & [14].

Tempo of speaking has significant importance in Forensic Speaker Identification FSI [4] because it is:

1. Carrying the individual-identifying information about the speaker.
2. Affected by the individuals variations in speaking.
3. Not affected by the frequency characteristics of the transmission systems and at the level at which the speaker talks.
4. Not easy to imitate or disguise.
5. Not controlled by the speaker.

Your speech says very much about you. It can reveal your age, your health, your level of education, your regional dialect, and many other factors; even the location in which a recording is made. Thus, for a forensic phonetician expert, there's a wealth of information hidden in voices and this data is collected, observed, documented, compared and processed for forensic speaker identification FSI.

2 METHODOLOGY

A. Data Collection

The experiment includes 10 unknown speakers (5 females and 5 males) of colloquial Arabic language, with no recorded speech disorders. Speaker's ages estimated between 19 to 40 years old. Natural spontaneous speaking style is elicited for

[3]The idea of the speaker identification is that every speaker has a unique voice pattern based on two factors: the first one is ***voice uniqueness***: including the uniqueness of the anatomical structure of every speaker. The second factor is ***the manner*** in which the articulators or muscles of speech are manipulated during speech: meaning the flexibility of the articulators which affected each other's.

30 seconds for each speaker trying to avoid the effect of any stress or the domination of any specific emotion. All the data are collected through a radio program called "the press in their eyes "الصحافة فى عيونهم" which is a daily program that announced every day at Alexandria Radio (Bakous Alex, frequency 101.1). The announcer of the program goes down to the street every day and asks one of the public. This one of the public could be a male or a female who was reading one of the daily newspaper and his or her identity is unknown for the announcer and for the listeners. The announcer asks a simple question which is: what's your comment about one of the news that you have been read at that daily journal? Then, the unknown speaker starts to talk spontaneously, without any recommended preparation, about any topic that he or she chooses. Accordingly, that unknown speaker is one of the Alexandrian populations who may get intermediate education (which enables that unknown speaker to read the daily journals) or may be well educated.

B. Recordings

The data are collected and elicited through the announcer who asks the unknown speaker about his/ her comments or opinions about any piece of news of the daily journal headlines. The whole duration of each episode is (about 5 minutes for every speaker) directly recorded from the radio channel using **Samsung mobile phone recorder as wav.files**; to avoid any transmission distortions. Then, all the episodes (10 episodes of 10 unknown speakers, each of which is 5 minutes) are transmitted into a laptop device for editing. Therefore, the researcher used cutter software for cutting only 30 seconds of continuous and spontaneous speech of each speaker from the whole speaking time (from the whole episode which is 5 minutes). This cutter software is called **"Easy audio ogg wma wav cutter software** (www.koyotesoft.com). At last all the edited data (only 30 seconds of spontaneous speech for 10 unknown speakers) are exposed to **Praat software** (www.praat.org) for the analysis (next step).

C. Analyses

All the data are analyzed manually with the aid of Praat software for all speakers. The analysis procedure is composed of three sequential steps which are:

The first step is the transcription process in which every 30 seconds of recording spontaneous speech for each unknown speaker are phonetically transcribed by using IPA symbols. The researcher transcribed all the data manually through the careful listening depending on the ears of the researcher with the aid of Praat software as a listening tool. Broad transcription type is used for this research because the main concern of that transcription process is counting the number of the pronounced syllables in a particular time (which is 30 seconds of spontaneous speech for each informant). So, no matter of how an informant is pronouncing a particular phoneme as long as does not affect the number of the pronounced syllables.

The second step is the segmentation process which means dividing the transcribed speech into syllables; this process is done manually by the researcher.

The third step is the calculation process in which speech rate SR and the articulation rate AR are calculated with their durations. Also the number of pauses and the duration of each pause are counted too.

D. Measurements

All the acoustical measurements illustrated with their mean of calculation for all the ten unknown speakers:

- *Fundamental frequency* f_0 is measured for all speakers using praat voice report.
- *Intensity* is measured for all the speakers with praat software through getting the mean intensity.
- *The number of the pronounced syllables* for each speaker, how many numbers of syllables the speaker has pronounced in only 30 seconds. The number of the pronounced syllables calculated manually by the researcher through counting all the produced syllables after segmentation process.
- *Speech rate SR* is measured according to the following definition “the number of syllables per second including the whole speaking time (with all pauses and hesitations)”; which is 30 seconds for each speaker.
- *Articulation rate AR* is measured according to the following definition “the number of syllables per second excluding the pause time and all the hesitation duration”. Note that the excluded pause time and hesitation duration will vary from one speaker to another.
- *All pauses durations* are measured by combining the duration of each pause in each speaker's utterance and the duration of each pause between utterances.
- *The number of pauses* for each speaker; is counted manually by the researcher, through counting the number of all pauses (filled and silent) occurred in the whole speaking utterance (occurred in 30 seconds for each speaker).
- *The duration of each pause* occurred in the whole speech sample (in 30 seconds) for each speaker with the aid of praat software, and also, determining the type of each pause.
- *Percentage of pause time* is measured manually by the researcher, through calculating the proportion of all the pauses time (the duration of all pauses) to the whole time of the speech sample (which is 30 seconds).
- *The degree of hesitancy* is measured manually by the researcher for each speaker through calculating the proportion of filled pauses to all pauses for the overall speech sample.

E. Perceptual Test

Sixty listeners of university students aged between 17 and 25 years old, with no recorded history of hearing impairments. Each listener was sitting directly in front of a laptop computer device with approximately three feet distance. The listeners were listening to the voice line-up (mp3 playlist, with 2 seconds interval between each informant and the following) through a loud speaker (attached to the laptop computer device) which was set up on medium volume.

The listeners were received some instructions from the researcher for doing the perceptual test perfectly:

1. Each listener received a "*listening sheet*" (See Figure 1) which contained the ten unknown speakers (5 females and 5 males listed one by one) titled as informant 1, informant 2, ……., informant 10.
2. The listeners are asked to listen carefully to the voice line-up of the ten unknown informants three times at most in order to enable them to select the fastest speaker and the slowest speaker.
3. Then, each listener selected the fastest speaker and the slowest one by marking (√) in front of his or her title at the listening sheet.

Observe that, the ten informants' voices intended to be listed one by one (male followed by female) in the voice-line up; in order to distract the listeners' attentions from the gender of the speaker. Because, almost all acoustic measurements and perceptual expert descriptions show experimentally that there are no significant differences in speech tempo between men and women. In other words, tempo of speech has no relation to the gender of the speaker.

Speakers المتكلمون	**The FASTEST** الأسرع	**The SLOWEST** الأبطأ
Informant 1		
Informant 2		
Informant 3		
Informant 4		
Informant 5		
Informant 6		
Informant 7		
Informant 8		
Informant 9		
Informant 10		

Figure 1: The Listening Sheet: where the involved listeners are marking (√) in front *The Fastest Informant* and *The Slowest Informant* as well.

3 RESULTS AND DISCUSSION

This experiment is designed in order to be utilized in forensic case work to identify unknown speakers of Colloquial Arabic depending on their tempo of speaking (rate of speaking) on the basis of two approaches: 1) an acoustic approach; based on acoustical measurements and statistical analyses; and 2) a perceptual approach; based on the ability of the naïve listeners to recognize the differences in speaking tempo; specifically identifying *the fastest* speaking tempo and *the slowest* speaking tempo depending on their ears.

A. Perceptual Test Results

The following figure (Figure 2) showing the distribution of all the listeners' selections percentages for both the fastest speaker and the slowest speaker as well. Through glancing over Figure 2, it's noticed that, the percentages of listeners' selections are highly distributed across all the ten informants with varied degrees which reveal that there is no absolute agreement about a particular speaker whether the slowest or the fastest.

With respect to ***the fastest speaker identification***; 38 % of the listeners select informant 3 to be the fastest speaker (with the fastest speaking rate). And 23 % of the listeners select informant 7 to be the fastest speaker, followed by 21% of the listeners who select informant 2 to be the fastest speaker (See Figure 2).
On the other hand, the results of ***the slowest speaker identification***; showed that41% of the listeners expect informant 1 to be the slowest speaker. And 20 % of the listeners select informant 6 to be the slowest speaker. These results indicate that identifying the slowest speaker seems to be more problematic for the listeners (See Figure 2).

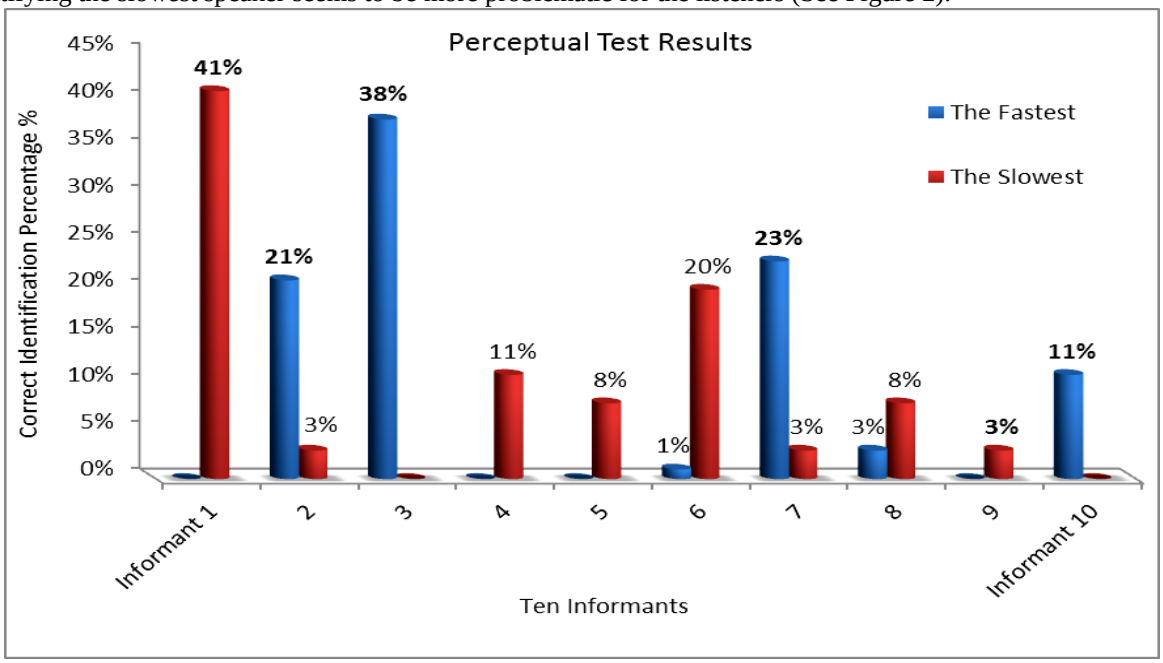

Figure 2: line- chart showing the distribution of all the listeners' selections (correct and false identifications) of both the fastest informant and the slowest informant depending only on their ears.

B. Acoustical Test Results

Table 1 showing a detailed description of all the acoustical results of the ten informants involved in this study. Accordingly we can deduce the following:
With respect to ***the speech rate SR and the articulation rate AR*** values (also see figure 3); ***the fastest speaker*** is informant 3 who pronounced the largest number of syllables in 30 seconds. The second fastest speaker is informant 2 and she also has the highest intensity across all speakers. The third fastest is Informant 7 who has the third place in speech rate SR but not in articulation rate AR; and he also has a relatively low F_0 across male speakers but not the least one.
Regarding ***the slowest speaker speech rate SR***; informant 9 is the slowest speaker followed by Informant 10 who has also a relatively high F_0 between female speakers; and a relatively high intensity across all speakers. Informant 10 also has a few numbers of pauses; moreover, their total duration is very short. According to ***the articulation rate AR of the slowest speaker***; it's noticed that, informant 10 has the slowest articulation rate AR, followed by informant 9; which indicates that the arrangement of informant 9 and informant 10 is reversed only in their articulation rates AR results. Informant 5 (male) has the third place at the slow speech rate and articulation rate too; he also has the lowest F_0 between male speakers and he has the second small intensity degree but not the least. Moreover, he has the highest degree of hesitancy.

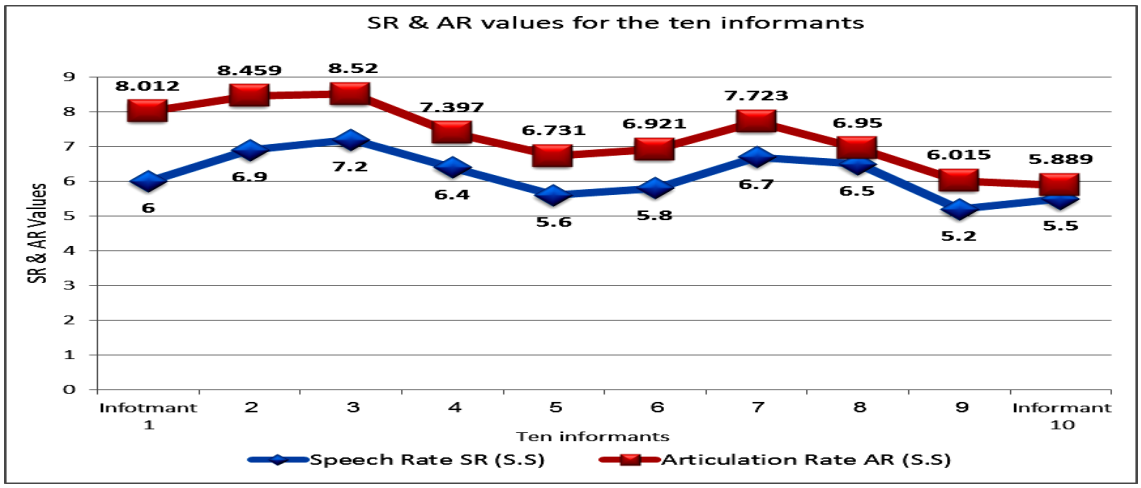

Figure 3: line- chart showing the values of the speech rate SR and the articulation rate AR for all the ten unknown informants.

Table 1: showing a comprehensive exhibition of all the measurements for all the ten informants for 30 seconds for each informant (Speech Rate Time). The ten informants (5 males and 5 females) arranged one by one.

Informants / *Measurements*	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Gender	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female
F_0 *(Hz)*	174	279	187	212	114	245	120	245	174	247
Intensity *(dB)*	60	82	75	77	69	71	73	71	77	79
Number of all pronounced syllables *(syll.)*	181	207	216	192	167	173	201	194	155	166
Speech Rate SR *(Syll. /Sec.)*	6.0	6.9	7.2	6.4	5.6	5.8	6.7	6.5	5.2	5.5
Articulation Rate AR *(Syll. /Sec.)*	8.012	8.459	8.520	7.397	6.731	6.921	7.723	6.95	6.015	5.889
Articulation Time *(Sec.)*	21.59	21.87	23.24	23.93	22.88	23.55	24.99	26.90	24.77	26.49
The number of pauses *(frequencies of occurrences)*	25	22	17	22	18	15	11	10	16	14
Total pauses time *(Sec.)*	8.41	8.13	6.46	6.07	7.12	6.45	5.01	3.1	5.23	3.51
Percentage of all pauses and hesitations to the overall speech *%*	28.03	27.1	21.5	20.23	23.73	21.5	16.7	10.33	17.43	11.7
The degree of hesitancy *%*	28	54.5	23.5	50	66.7	26.7	27.3	50	31.3	35.7

- **Correlation between speech rate SR and articulation rate AR**

Acoustically, there are many more acoustic cues which modify and determine the perceived speaking tempo. According to the results of the present experiment, the fastest speaker in speech rate SR has the fastest articulation rate AR as well (See Figure 3). But the slowest speaker in speech rate SR is not having the slowest articulation rate AR (See Figure 3) which indicates that the articulation rate AR has a relatively significant effect in identifying the fastest or the slowest speaker. Those results imply that speech rate SR is more prominent in identifying speakers with the fastest speaking tempo as well as slowest speaking tempo. Moreover, the number of the pronounced syllables which depend on the velocity of the speech organs is the main factor that directly influence the speech rate SR specifically.

Romito, Lio & Galata (2005) states that the articulation rate AR depends on the intrinsic duration of the various phones, the rate of the articulation movements and on the rules of coarticulation, but speech rate SR depends on speaker-specific features and on the communicative situation. Koreman (2006) investigates the role of the articulation rate AR in distinguishing fast and slow rates. Koreman finds out that the intended and realized phone rates alone (Articulation Rate AR) can't explain the perceived rates of the speakers, which indicates that there are other factors than the articulation rate

AR, that determine the perceived speaking tempo such as pauses and hesitations (which are excluded from his data). According to his results, we can't assume a direct relationship between articulation rate AR and the perceived rate. Moreover, Laver (1994), reports that Goldman - Eisler (1968) shows, experimentally, that the speech rate SR and the articulation rate AR have no significant correlation. But, this doesn't exclude the possibility that articulation rate AR is a concomitant of speech rate SR.

On the contrary, Gold (2012), investigates the role of articulation rate AR as a discriminant in forensic speaker comparisons. Gold shows experimentally that the articulation rate AR performs much better as a discriminated parameter *within* the same speaker comparisons (indicating different styles from the same speaker) than the different speaker comparisons. He manifested that articulation rate AR is still considered in forensic speaker comparisons in conjunction with other speech parameters. This is based on Rose (2006) who points out that "not all speakers differ from each other in the same." Therefore, there will be individuals where articulation rate AR is a highly discriminated parameter for them. Also Jessen (2008) assures that when articulation rate is taken into account, considering only the fluent speech and all pauses ignored, the *between* speaker variation is relatively greater, and the *within* speaker variation seems to be smaller in articulation rate than syllable rate (Goldman Eisler, 1968; and Künzel, 1997), which makes articulation rate AR the more promising measure for forensic purposes.

According figure 4 that showing us the ***Mean Intensity*** of all the ten unknown speakers, regarding *high intensity degrees*, informant 2 recorded the highest degree of intensity; followed by informant 10. Observe that, informant 2 who has the highest intensity degree, is the second fast speaker. And informant 10 who has the second high degree of intensity is the second slow speaker. Regarding *low intensity degrees*, informant 1 (male) recorded the lowest intensity degree; followed by informant 5 (male). Perceptually, 41% of the listeners expected informant 1 to be the slowest speaker (see the perceptual test results).

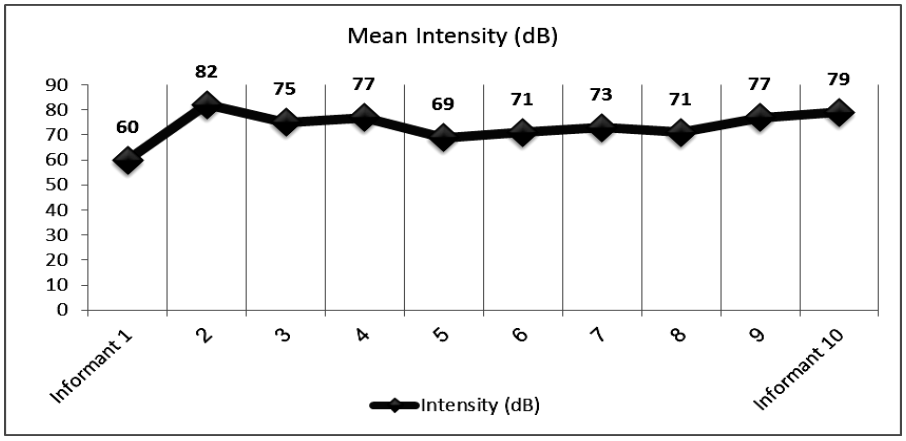

Figure 4 : Line chart used to visualize the mean intensity of the total speech duration for the ten informants.

Primarily ***the percentage of all pauses to the overall speech sample*** is depending on both; *the number of pauses* (*pauses' frequencies of occurrences*) as well as *their durations*; for more details see table 1.

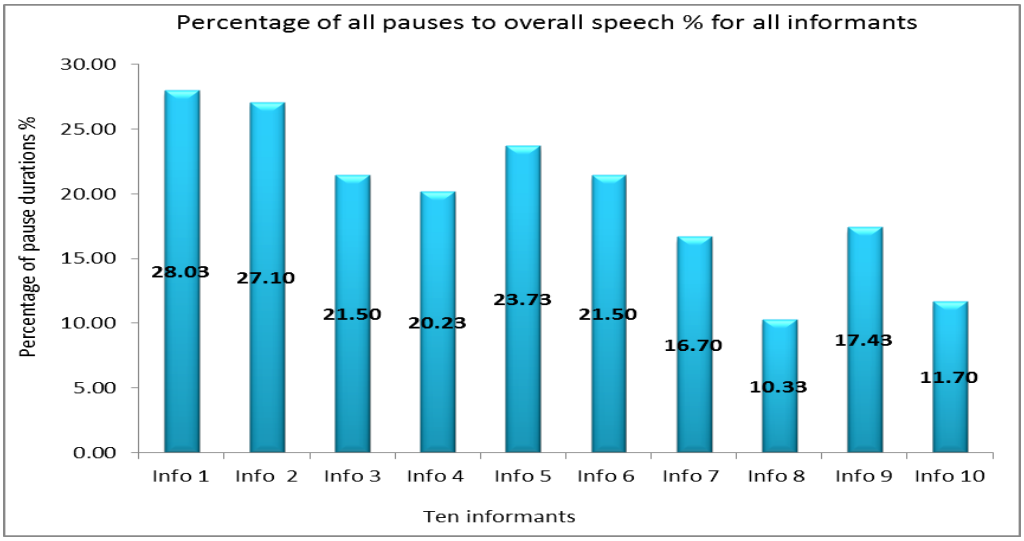

Figure 5: Column - chart showing the percentage of all pauses and hesitations to the overall speech sample for each speaker.

According to the preceding figure (See Figure 5), respecting ***the highest percentage of all pauses to overall speech sample***, informant 1 has the highest percentage of pauses (28 % of his speech sample consists of pauses); followed by informant 2; followed by informant 5. Regarding ***the lowest percentage of all pauses to the overall speech sample***, informant 8 has the lowest percentage of pauses (only 10.33 % of her speech sample consists of pauses); followed by informant 10; followed by informant 7. Hence, informant 8 and informant 10 should be expected to be the fastest speakers according to their percentage of all pauses to the overall speech sample because they have the lowest percentages; and reversely, informant 1 and informant 2 should be expected to be the slowest speakers because they have the highest percentages of pauses to overall speech. But this is not true acoustically, which implies that the percentage of all pauses to the overall speech sample has no direct influence on the speech rate and they cannot determine or modify the speech rate. The results of the present experiment indicate that the percentage of all pauses to overall speech sample plays a double-edged role:

The first role of the percentage of all pauses is on the acoustical level, they don't have any obvious effectiveness on the speech rate SR; i.e., there is no remarkable correlation between the speech rate SR and pause (the fastest speaking rate SR must not have the minimum percentage of pauses and vice versa). However, Laver (1994), reports that Goldman - Eisler (1968) shows, experimentally, that speech rate SR may be positively correlated with the percentage of pauses and durations in the speech sample. Moreover, Vaane (1982) states that the most languages' variations in speech rate SR are mainly due to variations in the durations of pauses. However, in light of the results of the present experiment, it does not appear to be a verity. The results of the present experiment also contradict with Fashal (1991) which, experimentally, studies the tempo of reading texts of news-bulletins in Egyptian broad casting. The results conclude that the number and the duration of pauses is an effective factor in changing the rate of speech i.e., decreasing the number and the duration of the pauses increases the speech tempo.

The second role of the percentage of all pauses is on the perceptual level, large percentage of pauses durations is considered one of the most important factors that influence the listeners' perceptions; in particular, the perception of slow speaking tempo. Conversely, small percentage of pauses durations does not mean fast speaking tempo for listeners' perceptions. That's confirmed by Koreman (2006) who states that; it is possible that; pausing or dis-fluencies determine the perceived speaking tempo.

The degree of hesitancy for each informant shows the proportion of filled pauses to all pauses for the overall speech sample to indicate large differences between speakers (*intra speaker variation*) and relatively small differences within speaker (*inter speaker variation*). Figure 6 indicates that informant 5 (who is arranged as the third slow speaker according his speaking rate and he has the lowest F_0 between male speakers) has *the highest degree of hesitancy* (66.7 %), which may indicate that the high degree of hesitancy may negatively affect the perceived speaking rate. In other words; high degree of hesitancy may be considered as a sign of slow speaking rate. To confirm this, we need more experimental research. Figure 6 also indicates that informant 3 (who is the fastest speaker according to his speaking rate and he has the highest F_0 between male speakers) has *the lowest degree of hesitancy* (23.5 %). Regarding the results of the present experiment; the degree of hesitancy seems to have an inverse relation with speaking tempo particularly at fast speaking tempo. In other words; the fastest speaker (according to speech rate SR and articulation rate AR) has the least degree of hesitancy. And this relation is compatible with only the fast speaking rate.

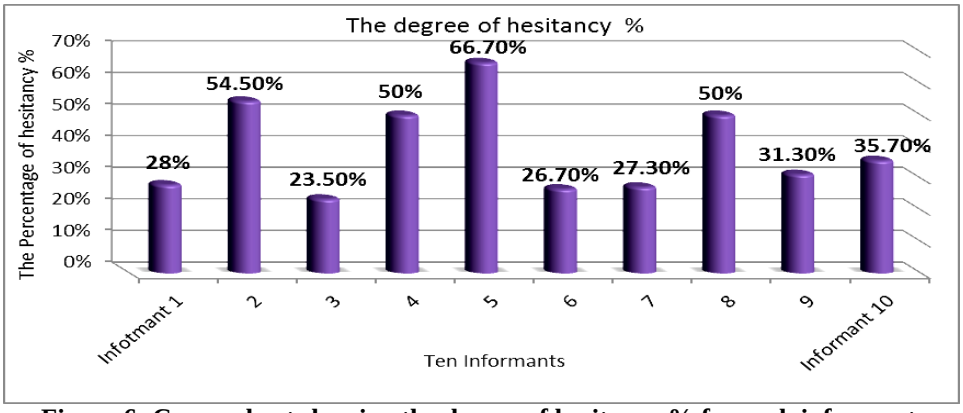

Figure 6: Cone - chart showing the degree of hesitancy % for each informant.

- **<u>Correlation between the acoustical results and the perceptual ones</u>**

Hayward (2000) said that, "*experiments in speech perception have commonly focused on a single acoustic cue and a single impressionistic-phonetic dimension. Nevertheless, it would be highly exceptional for a single cue to be solely responsible for signaling a simple phonetic contrast*". For more illustration; if listeners are asked to classify stimuli which vary in two or more cues, there is typically a *trading relation* between the cues. Hayward (P. 118) adopted this trading relation hypothesis; explaining that this trading relation is used in the sense of *trade-off* which is used in everyday

situations. In other words, for the listeners' perception while listening to a speech sample, some acoustic cues are more dominant or more prominent above some other cues.

Regarding ***the fastest speaker identification;*** the acoustical results are corresponded with the perceptual results indicating that informant 3 is the fastest speaker. According to the perception of the listeners, 38 % of the listeners select informant 3 to be the fastest speaker (see figure 2 and table 1). According to the acoustical measurements, as well, informant 3 pronounced the largest number of syllables in 30 seconds; so, he has the highest SR and AR; moreover, he has the minimum degree of hesitancy and the highest F_0 between male speakers. *Perceptually*, 23 % of the listeners expected that informant 7 to be the fastest speaker. This proportion can't be easily ignored; informant 7 is arranged acoustically as the third fastest speaker according to the number of the pronounced syllables and consequently in speech rate SR value. But informant 7 has a relatively small articulation rate AR and a low F_0 across male speakers. He also has a small number of pauses. By gathering all the preceding acoustical cues together; perceptually, informant7 is considered to be a fast speaker but not the fastest. 21 % of the listeners expected that informant 2 to be the fastest speaker. Informant 2 acoustically is organized as the second fastest speaker; moreover, she has 22 pauses in her total speech sample with relatively high degree of hesitancy (see table1). Informant 2 also has the highest F_0 between female speakers; and she also has the highest mean intensity across all speakers. All the preceding cues together gave an impression of fast speaking tempo for the listeners. *Perceptually*, listeners' perceptions seem to be highly affected by her remarkable highest degrees in F_0 and intensity across all speakers more than the number of pauses and the degree of hesitancy.

11 % of the listeners also expected that informant 10 to be the fastest speaker. Whereas, she is arranged acoustically as the second slowest speaker according to her speech rate SR and she has the slowest articulation rate AR across all the involved speakers. However, no one of the listeners selects her or expects her to be the slowest speaker (See Figure 2 and table 1).On the other hand, Informant 10 has a relatively high F_0 between female speakers; and also, she has the second high intensity degree across all speakers. Informant 10 also has a few numbers of pauses; moreover, their total duration is very short. These preceding results indicated that, high F_0 and high intensity; in addition to her few number of pauses and their short duration; all of these cues together may be responsible for perceiving fast speaking tempo more than the other acoustic cues.

With respect to ***the slowest speaker identification***, the perceptual results did not match with the acoustical results. *Perceptually*, 41 % of the listeners expected that Informant 1 is the slowest speaker. Acoustically, informant 1 has recorded the largest percentage of all pauses to the overall speech sample (See table 1 and Figure 5). Informant 1 also has the largest number of pauses' frequency and duration (see table 1). Moreover; he has the minimum intensity degree across all speakers. The preceding acoustical cues imply that there is a correlation between the number and the duration of pauses with the perception of the listeners' particularly in determining the slow speaking rates. In other words; increasing the number and the duration of the pauses gives a perceptual impression of slow speaking rate. In addition to, the mean intensity seems to have a remarkable effect on the listeners' perception i.e., the lower the loudness, the slower the perceived speaking rate and vice versa.

On the other hand; *perceptually*, only 3 % of the listeners select informant 9 to be the slowest speaker. According to the acoustical results informant 9 is the slowest speaker, because of his smallest speaking rate SR. He also has a relatively high F_0 between male speakers, in addition to a relatively high mean intensity. These preceding results indicated that; because of his relatively high F_0 and relatively high mean intensity only 3 % of the listeners select him to be the slowest speaker; and no one selects him to be the fastest speaker as well.

To summarize, from the preceding exhibition of the correlation between the acoustical results and the perceptual results, we can deduce the following: F0, intensity and pauses' numbers and durations have significance effect on the perceived speaking tempo. And further, those acoustical cues are used as perceptual parameters that almost all of the listeners relay on, to identify the slowest and the fastest speaking rates. There are so many parameters that influence the speaking tempo and its' perception. Naïve listeners may focus on some of these cues more than others; which give the possibility of false identification to occur. Particularly, in identifying the slowest speaking tempo; most of naïve listeners clearly relate the slow speaking tempo directly with the pauses frequencies and durations which confuse most of them in identifying the slowest speaking tempo. Whereas the phonetician expert must find all the sources and types of these cues and combine them together for identifying the identity of the unknown speaker in the forensic case work.

4 CONCLUSIONS

It has become clear over the years that speaker identification should not be confined to a single method, but a variety of different methods should be used. To find out the most effective acoustical cues which affect the rate of speech and its perception, an acoustic – perceptual line must be drawn. Clearly, the acoustical results should explain and clarify the perceptual ones. According to the acoustical results and the perceptual results of the present study, we deduced the following conclusions:

1. With respect to **speech rate SR** and **Articulation rate AR:** Speech rate SR is more prominent in identifying the unknown speaker. However, this does not mean to exclude the articulation rate AR.

On the perceptual level; in identifying the rate of speech, listeners are listening to the speech rate SR (with all pauses i.e. filled, silent and hesitations). They are not listening to the articulation rate AR only; therefore, articulation rate AR should be combined with other acoustic cues in order to reflect perceptually a fast or a slow speaking tempo.
On the acoustical level; articulation rate AR is distinctive for some speakers because it referred directly to the pure pronounced syllables which is related primarily to the velocity of the speech organs of the speaker (whoever male or female). Substantially, velocity of the speech organs is an effective factor in modifying the tempo of speaking. As a result of that; speech rate SR should be combined with the articulation rate AR in order to help in identifying the identity of the unknown speaker under the forensic circumstances.

2. **The degree of hesitancy:** acoustically; it is considered as a remarkable factor for *the fastest speaking tempo*. In other words, according to the results of the present experiment; the fastest speaker in speaking tempo has the least degree of hesitancy. But the highest degree of hesitancy does not take place at the slowest speaker's speaking tempo.
3. **F_0:** it is an important acoustic cue in identifying the speaker's speaking rate acoustically and perceptually as well. The importance of F_0, according to our results, appeared in identifying *the fastest speaking tempo*; but not in identifying the slowest speaking tempo. High F0 (for male or female speaker) indicated fast speaking tempo perceptually as well as acoustically.
4. **Mean intensity:** *on the perceptual level*, mean intensity is a remarkable cue for listeners' perception in identifying the rate of speaking of the speaker (whether the slowest or the fastest). In other words, high intensity indicated fast speaking tempo; and low intensity indicated slow speaking tempo. Whereas, *on the acoustical level*, and according to our results, mean intensity results are inconsistent with speaking tempo results.
5. **The percentage of all pauses**, **pauses' numbers** and **pauses' durations** have a double aged role.
   On the perceptual level; naïve listeners have a great ability in detecting different speaking tempi. However, particularly in identifying the slowest speaking tempo, there is some inconsistency between the acoustical results and the listeners' selections. Seemingly, listeners received the high percentage of all pauses as a sign of low speaking tempo that may confuse them in identifying the slowest speaking tempo, unlike identifying the fastest speaking tempo.
   More specifically, informant 1 has the lowest mean intensity and the highest percentage of all pauses. Therefore;
   41 % of the naïve listeners expected him to be the speaker with the slowest speaking tempo. Sequentially as a result, on the perceptual level and according to the results of the present experiment, low loudness and high percentage of all pauses, when occurred together in that way, seem to have a great influence on the listeners' perception in identifying the slowest speaking tempo particularly.
   On the acoustical level; according to the results of this study, the percentage of all pauses, pauses' numbers and pauses' durations did not show significance role in modifying the rate of speech. For more illustration, the fastest speaker with the fastest speaking rate does not have the least degrees of anyone of these cues.

ACKNOWLEDGMENT

I want to give the great thanks to my supervisor Prof. Dr. Mervat Fashal, the professor of phonetic science at the department of phonetics and linguistics, for her guidance, supporting and for giving me this great opportunity. I want also to express my sincere appreciation to Dr. Sameh Al-Ansary, the headmaster of the phonetics and linguistics department for his remarkable effort for our department.

REFERENCES

[1] Bonastre, J. F., Bimbot, F., Boë, L. J., Campbell, J. P., Reynolds, D. A., and Magrin, C. I. (2003) Person Authentication by Voice: A Need for Caution. In Eurospeech 2003, Interspeech 2003. *Proceedings of the 8th European conference on speech communication and technology*. pp. 33-6. Geneva, Switzerland.http://www.afcpparole.org/doc/AFCP_SpLC_HotTopicsEurospeech03_final.pdf

[2] Cain, S., Smrkovski, L. and Wilson, M. (1990) Voiceprint Identification - Money Laundering and Narcotics Update, Department of Justice, 1988, and the Legal Investigator. NDAA Bulletin December 1993.

[3] Dellwo, V. (2003) Forensic Phonetics. Phonetics and linguistics, *Speech science*.

[4] Demenko G. (2000) Analysis of supra-segmental features for speaker verification. Institute of linguistics, Adam Mickiewicz University, Poznan, Poland.

[5] Eriksson, A. (2005) Tutorial on forensic speech science. Part I: Forensic phonetics. In Interspeech, Eurospeech 2005. Proceedings of the 9th *European Conference on Speech Communication and Technology*. Department of Linguistics, Gothenburg University, Gothenburg, Sweden 2005.
Website:http://www.york.ac.uk/media/languageandlinguistics/documents/currentstudents/Eriksson_tutorial_paper.pdf

[6] Fashal, M. (1991) *Acoustic Analysis of Tempo in News-Bulletins*. PH. D. thesis. Alexandria University.

[7] Furui, S. (2008) Speaker Recognition. Tokyo Institute of Technology. Scholarpedia, 3(4):3715. doi:10.4249/scholarpedia.3715. Website: http://www.scholarpedia.org/article/Speaker_recognition

[8] Gold, E. (2012) Articulation Rate as a Descriminant. In Forensic Speaker Comparisons. *UNSW Forensic Speech Science Conference*, Sydney, Australia. Website: http://sydney2012.forensic-voice-comparison.net/
[9] Goldman-Eisler, F. (1968) Psycholinguistics: experiments in spontaneous speech. London: New York, Academic Press.
[10] Hayward, K. (2000) Experimental Phonetics: An Introduction. Harlow: Longman.
[11] Hicks J.W. (1979) *An Acoustical/Temporal Analysis Of Emotional Stress In Speech* .PhD dissertation, University of Florida
[12] Jessen M. and Bundeskriminalamt BKA (2008) Forensic Phonetics. *Language and Linguistics* Compass 2/4 (2008): 671–711.
[13] Koenig, B. J. (1986) Spectrographic Voice Identification: A Forensic Survey. Letter to the editor of JASA 79/6: 2088–90.
[14] Koreman, J. (2006), The role of articulation rate in distinguishing fast and slow speaker. Institute of Phonetics Saarland University, Germany. jkoreman@coli.uni-saarland.de.
[15] Künzel, H. J. (1997), Some general phonetic and forensic aspects of speaking tempo. Forensic Linguistics, 4.48–83.
[16] Laver, J. M. D. (1994) *Principles of Phonetics.* Cambridge: Cambridge University Press.
[17] Lindh, J. (2004) Handling the "Voiceprint" Issue. In Proceedings, FONETIK 2004, Dept. of Linguistics, Stockholm University.
[18] Lindh, J. (2009) *Robustness of recognition of voices, speech and speakers through a forensic perspective - tools and methods.* A PhD Thesis EMBRYO, University of Gothenburg, Sweden, ISBN 978-91-977196-7-4, 2.
[19] Mcpeek, T. (2013) *American Voice Types: Towards A Vocal Typology for American English*. PHD dissertation, University of Florida 2013.
[20] Nolan F., McDougall K., Jong D. G. & Hudson T. (2006) A Forensic Phonetic Study of 'Dynamic' Sources of Variability in Speech: The DyViS Project. Department of Linguistics, University of Cambridge, United Kingdom, fjn1@cam.ac.uk , kem37@cam.ac.uk , gd288@cam.ac.uk , toh22@cam.ac.uk.
[21] Nolan, F. (1997) Speaker recognition and forensic phonetics. In W. J. Hardcastle & J. Laver (Eds.), the *handbook of phonetic sciences*, 744–767. Oxford: Blackwell.
[22] Nolan, F. (2001) Speaker identification evidence: its forms, limitations and roles, in Proceedings of *the Conference Law and Language*: Prospect and Retrospect. 12-15 Levi, Finnish Lapland. nolan2001proclanglaw.pdf
[23] Nolan, F. and Catalin G. (2005) A case for formant analysis in forensic speaker identification. *International Journal of Speech, Language and the Law*, 12 (2): 143-173.
[24] Romito, L., Lio, R. & Galata, V. (2005) Fluency articulation and speech rate as new parameters in the speaker recognition. Article presented at *the 3rd conference on experimental phonetics (CEP),* Santiago de Compostela.
[25] Rose, P. (2002) *Forensic Speaker Identification*. London, UK: Taylor and Francis.
[26] Saeed, K. A. (2006) Note on Biometrics and Voice Print: Voice-Signal Feature Selection and Extraction – A Burg-Toeplitz Approach. Faculty of Computer Science, Bialystok Technical University, Wiejska 45A, 15-351 Bialystok, Poland. aida@ii.pb.bialystok.pl. http://aragorn.pb.bialystok.pl/~zspinfo/
[27] Singh, N., Khan, R. A. and Shree, R. (2012) Applications of Speaker Recognition. *International conference of modeling, optimization and computing (ICMOC 2012)*, Procedia Engineering Conference 2012.
[28] Trouvain, J. (2003) *Tempo variation in speech production, implications for speech synthesis*. PH.D dissertation.
[29] Vaane, E. (1982) Subjective Estimation of Speech Rate. Phonetica 39, pp. 136- 149.

BIOGRAPHY

Eman M. Yousri: was graduated from the Department of Phonetics and Linguistics in 2009. Her Graduation Project was about Voice Print Analysis for Forensic Speaker Identification in courts; her grade was Excellent for this graduation project. She obtained her Master Degree in Phonetic Science in June 2015. The thesis was also about Forensic Speaker Identification Depending on Temporal Parameters.

Prof. Mervat Fashal: is a professor of phonetic science, and in particular in psych-acoustics field, Dept. of Phonetics and Linguistics, Faculty of Arts, Alexandria University. She had studied Ph.D. in

Germany. She has articles and experiments on various areas of phonetics. The significant contributions in the field are mainly in the following topics:

- Speech production and perception
- Prosodic and discourse analysis
- Acoustic analysis of normal and abnormal speech
- Speech recognition and speaker identification in the field of forensic phonetics

Mervat Mohamed Ahmed Fashal is a Full Professor since 2008, a Head of Phonetics Department from 2003 – 2012

TRANSLATED ABSTRACT

التعرف على المتكلم اعتماداً على معايير السرعة الزمنية فى العامية المصرية

هدف هذه الدراسة هو التعرف على هوية المتكلمين غير المعروفين من سرعة كلامهم، وقد تم فى هذا البحث -**على المستوى الإدراكي** - عمل تقييم للقدرة الإدراكية للمستمعين غير المدربين فى التعرف المتكلم اعتمادا على سرعة كلامه، وإدراك ما إذا كان الأسرع أم الأبطأ بين المتكلمين العشرة الذين تم اختيارهم للتجربة. أما **على المستوى الأكوستيكي**، فقد تم رصد المعايير الفيزيائية الأساسية للتعرف على صوت المتكلم وهى كالآتى:

1. التردد الأساسى F_0
2. الترددات المكونة للصوائت(F1, F2, F3.)
3. الرنين الأنفي للصوامت (الغُنة).
4. معدل سرعة الكلام (SR) ومعدل سرعةالمنطوقات (AR)

وقد اختير العنصر الأخير وهو سرعة الكلام (Speech Tempo) كموضوع لهذه الدراسة. وقد تم عمل التحليل الفيزيائى لكلام المتحدثين وقياس معدل سرعة الكلام ومعدل سرعة المنطوقات والوقفات فى كلام كل متحدث (أطوالهم وأعدادهم). هذا فضلا على قياس التردد الأساسى لكل متكلم F_0 وشدة الصوت (I).
هناك العديد من الأسباب الأساسية التي توضح مدى أهمية المعايير الزمنية ومعدل سرعة الكلام فى التعرف على المتكلم للأغراض القضائية وهى كلآتى:

1. لا يمكن محاكاة سمات السرعة الزمنية للكلام
2. لا يمكن للمتكلم السيطرة على السرعة الزمنية لكلامه بشكل واع.
3. الفروق الفردية بين المتكلمين تُعد من أهم مصادر التغير التي تُؤثر على معدل سرعة الكلام.

تشمل هذة التجربة عشرة أشخاص (خمس نساء وخمسة رجال) غير معروفين الهوية ومتحدثين أصليين للهجة العامية العربية وتقدر أعمارهم بين 19 و 40 عام. تتكون المادة من كلام تلقائي لمدة نصف دقيقة (30 ثانية) لكل متكلم مع تجنب تأثير أو سيطرة أي نوع من أنواع المشاعر السلبية للمتكلمين.
تم تسجيل المادة من خلال برنامج "الصحافة فى عيونهم" الذي يذاع يومياً على راديو إذاعة الإسكندرية. وتم تحليل المادة المسجلة لكل متكلم يدويا وكتابتها بالرموز الصوتية Transcription وذلك عن طريق الإستماع الجيد لهذه المادة المسجلة مراراً وتكراراً بواسطة Praat Software. ثم تمت عملية فصل المقاطع Segmentation Process وذلك لحساب SR & AR. كما تم أيضا قياس التردد الأساسي و شدة الصوت لكل متكلم ودرجة التلعثم و عدد الوقفات وزمن كل وقفة ونوعها ونسبة كل الوقفات إلى مدة الكلام الكاملة .

ستون مستمع من طلبة الجامعات ومتحدثين أصليين أيضاً للعامية العربية المصرية وتتراوح أعمارهم بين 17 و25 عام ، جميعهم تطوعوا للإشتراك فى هذا الإختبار. المهمة الأساسية للمستمعين هى الإستماع بحرص شديد إلى المتكلمين العشرة وتحديد المتكلم الأسرع وأيضا المتكلم الأبطأ من حيث سرعة الكلام عن طريق وضع علامة (√) أمام الرمز الدال عليه.

تشير النتائج إلى أن:

1. أكوستيكياً و إدراكياً: سرعة الكلام موضحة فى معدل سرعة الكلام (SR)هى المعيار الأقوى فى التعرف على المتكلمين غير المعروفين; بينما معدل سرعة نطق الأصوات (الصوامت والصوائت) (AR) كان أقل تأثيرا على تحديد سرعة المتكلم.
2. النسبة المئوية للوقفات تلعب دوراً مهما جداً على المستويين الإدراكي والأكوستيكي؛ على المستوى الإدراكي فإن زيادة النسبة المئوية للوقفات تُعدّ من أهم العناصر التي تؤثر على إدراك المستمعين للسرعة الزمنية للكلام، حيث تشير إلى سرعة الكلام البطيئة. أما على المستوى الأكوستيكي: فليس لها أي تأثير واضح على زيادة أو نقصان سرعة الكلام للمتكلم.
3. درجة التلعثم فى الكلام (الوقفات المملوءة filled pauses)، تُعد من العناصر المُميزة فى التعرف على المتكلم الأسرع من حيث سرعة الكلام للمتكلم. ومع ذلك فليس لها أي دور فعّال فى التعرف على المتكلم الأبطأ .
4. التردد الأساسي للمتكلم يُعد من العناصر الأكوستيكية المُميزة لتحديد سرعة الكلام للمتكلم ،بحيث زيادة التردد الأساسي للمتكلم تشير إلى زيادة معدل سرعة كلامه إدراكيا وأكوستيكيا.
5. متوسط شدة الصوت لدى المتكلم يُعد من الناحية الإدراكية من العناصر المُميزة بالنسبة إلى آذان المستمعين، بحيث زيادة شدة الصوت تشير إلى زيادة معدل سرعة الكلام للمتكلم ,وأيضا نقصان شدة الصوت تدل على نقصان معدل سرعة الكلام للمتكلم. ولكن هذه النتائج لا تنطبق على المستوى الأكوستيكي.