Statistical analysis for the pitch of mask-wearing Arabic speech

Hasan M. Kadhim, Alaa H. Ahmed, Saif A. Abdulhussien
Electrical Engineering Department, Engineering College, Mustansiriyah University, Baghdad, Iraq

Article Info

Article history:

Received Nov 02, 2021
Revised May 29, 2022
Accepted Jun 08, 2022

Keywords:

Arabic language
F0
Masked-face speech
PDA
Pitch

ABSTRACT

The study is a comparison between the statistical properties of pitch (F0) for mask-wearing speech and unmasked. The speakers are Arab, of different ages and genders. A robust algorithm for pitch tracking (RAPT) is used for estimating F0. The subjective tests denote that masked speech is attenuated, and noisy-background speech has fewer F0 candidates. Using objective tests, 60% of female and male F0s do not change when wearing masks. The remaining 40% of speech F0s change (the percentage gross error), by an approximately 20% increase and 20% decrease. The percentage classification error is about 10%. The F0 changes in females younger than 12 years old are fewer compared with similarly-aged males. The F0 changes of females older than 12 years old were approximately equal compared with similarly-aged males. An average of F0 (M) is used for each speech to divide its F0 band (50-500) Hz into two bands, lower-band (LB) (50-M) Hz and upper-band (UB) (M-500) Hz. The attributes of the two bands have been statistically analyzed. The F0 classification error (CE) for females is higher than for males, but the gross error (GE) for males is higher than for females. The F0 change values are directly proportional to the probability of F0 change.

Corresponding Author:

Hasan M. Kadhim
Electrical Engineering Department, Engineering College, Mustansiriyah University,
10047, Baghdad, Iraq
Email: hasanalmgotir@uomustansiriyah.edu.iq

1. INTRODUCTION

During the COVID-19 pandemic, face masks became a necessary part of a person’s attire. Despite their preventative advantages, they have some negative impacts on speaking and hearing. Physically, all face masks have some effects on the components of the human speaking systems. Wearing a face mask reduces the movability of the lips, the oral cavity, the jaws, and the tongue. Face masks act as a physical obstacle against the airflow of the mouth and the nose. Face mask-wearing has subjectively changed the audio features of human speech.

According to Fourier analysis, any periodic function can be analyzed as an infinite series of trigonometric functions (sets of sines and cosines). The frequencies of these functions are discrete and multiple. The first frequency (harmonic) is called the fundamental and has the greatest magnitude. The waveform of speech is not completely periodic, but it has many periodicities in the specific short term. Scientists and researchers in audio, language and speech signal processing have been efficiently exploiting this merit. For speech, this first frequency is called pitch and is abbreviated as F0. F0 has the greatest energy among the other frequencies during multiples of 10 ms periods, due to the short-time discrete Fourier transform (STDFT). The F0 detection process is related to the category of the speech in that period, voiced or unvoiced speech. Generally, the pitch description is the correlation perceptual of an audio fundamental frequency F0. Perceiving the F0 is important for general attribution to compare audio with various timbres due to the possibility of errors in hearing. Alternatively, because F0 is the compact measurement for a group of frequency harmonics, F0 efficiently provides an abstract representation of audio for human memory storage [1].

To estimate F0, pitch detection algorithms (PDA) are utilized. The range of an estimated F0 is between 40 Hz and 600 Hz. Usually, the F0 of females is higher than that of males [2]. Many algorithms have been proposed to detect the pitch F0 of audio and speech signals. Typically, one of the PDA algorithms detects and measures the duration of the quasiperiodic speech and audio signals, and then reciprocates the calculated value to extract the pitch frequency F0. Some PDA algorithms (e.g., ones based on autocorrelation) need two or more F0 periods (about 50 ms) to estimate the pitch. There are three main approaches for PDA: time-domain e.g., YIN "the yin and yang is an oriental philosophy" [3], frequency-domain e.g., Tolonen and Karjalainen algorithm (TK) [4], and spectral/ temporal approach e.g., the yet another algorithm for pitch tracking (YAAPT) [5].

The efficient algorithms and applications of PDA are: The PRAAT "the imperative form of to speak in Dutch" application was presented by Boersma [6], to detect speech signal periodicity robustly and directly by lag autocorrelation. After tests, the PRAAT has obvious immunity against jitter and additive noise to periodic signals. For speech signal analysis, the PRAAT has more accuracy for magnitude orders than the commonly used methods. The PRAAT measures the harmonics-to-noise ratio (HNR) for the lag domain. The measurements are reliable and accurate when compared with traditional frequency-domain approaches. An online open-source PRAAT application is available [7]. Sun [8] proposed the sub-harmonic/harmonic (SHRP) algorithm to find F0, depending on the ratio of SHR. By following alternate cycles of speech, pitch via spectrum shifting is estimated. The scale of frequency is logarithmic, and the SHR is calculated. The algorithm is evaluated via two databases. Performance of SHRP exceeded other PDAs. An online open-source SHRP Matlab toolbox is available. The maximum likelihood was adapted by Noll [9] with a cepstral analysis of frequency components product to detect F0. The harmonics have been matched to pre-defined spectrum schemes. Gruber [10], the possibility of spectrum polyphonic detection was investigated. To detect the harmonics spectrum, a periodogram to transform the time-domain waveform was used. Brown and Puckette [11], an improvement of F0 detection by the discrete cosine transform (DCT) spectrum was derived using the phase. The short-time Fourier transform (STFT) bins can be utilized to increase the accuracy of harmonics re-assignment using the phase. In addition to phase, magnitude is used to increase the accuracy. By Zahorian *et al.* [5], the YAAPT uses a time-domain tracking with an auto-correlation to normalize cross-correlation. In the frequency domain, the researchers used spectral attributes to find the pitch precisely. Tolonen and Karjalainen [4] suggested the TK algorithm to find multi-F0 by analyzing the periodicity of the speech signal. They partitioned the signal into two bands, lower and higher than 1 kHz. They invoked the summary autocorrelation function (SACF), and enhancement auto-correlation function (ESACF) between two signals to detect the speech signal periodicity. Medan *et al.* [12] super-resolution pitch-detector (SRPD) was derived. The procedure is based on the similarity of speech excitation techniques. The procedure has an infinite resolution, greater accuracy for F0, robustness against noise, more reliability, and less computational complexity. The procedure is applicable for speech processing that needs analysis of synchronous spectral F0. The speech transformation and representation using adaptive interpolation of weighted spectrum (STRAIGHT) paradigm was introduced by Kawahara *et al*. [13] Analysis of time frequency is used with group delay and instantaneous frequency. To extract signal attributes, the paradigm measures the aperiodicity of the frequency domain and the energy concentration in the time domain. A modified method is executed by minimizing perceptual disturbances, according to errors in the extracted attributes. The sawtooth waveform inspired pitch estimator (SWIPE) was developed by Camacho and Harris [14] to process music and speech to detect the F0 as the first harmonic (fundamental) of the sawtooth signal. The sawtooth has the best spectrum that matches the input speech signal spectrum. The kernel of decay cosine yields an extension for the previous frequency-based, sieve-type detection algorithm by giving smooth peaks for decaying amplitudes with the harmonics of the signal correlation.

Robust algorithm for pitch tracking (RAPT)

The well-known PDA detection is the RAPT. The algorithm was proven by Talkin and Kleijn [15] and is based on cross-correlation. The main improvement of this estimator is the reduction of computational complexity. The sequential outline of the RAPT algorithm is:

1) Providing speech samples with their sampling rate and with a reduced sampling rate.
2) Periodically, computing normalized cross-correlation function (NCCF) of the reduced sampling rate speech signal with lags in the F0 range.
3) Indicating the locations of maximum at the 1st pass of NCCF.
4) For the vicinity of the peaks in that 1st pass, calculate the NCCF for the original sampling rate.
5) Again, finding the maximum in that NCCF. Obtaining the location and amplitude of the modified peak.
6) For each peak obtained from the NCCF (high resolution), estimate the F0 of the processed frame.
7) The hypothesis of the frame for unvoiced/voiced is advanced for each frame.

8) Finding the group of the NCCF peaks via optimization process for the unvoiced/voiced hypotheses for all the frames which have the best match with the above characteristics.
9) Using the well-known speech pitch tracking algorithm (PTA), RAPT has the following differences:
   - PTA computes the NCCF in the linear prediction coding (LPC). RAPT computes the NCCF in the original speech signal.
   - Two stages of NCCF are used to reduce the overall computational load. There is a similarity between RAPT and the simplified inverse filtering technique (SIFT) in NCCF double-stage computing. To increase accuracy, RAPT uses maximum interpolation for the samples with a high sampling rate.

Kaldi pitch tracker is an improved RAPT [16]. Performance of tonal languages has been improved, when the Kaldi F0 tracker with the estimator of “probability-of-voicing” is used in automatic speech recognition (ASR). The original RAPT makes the hard decisions either voiced or unvoiced for each frame. The F0 tracker assigns the pitch for the unvoiced frame while constraining each pitch trajectory is continuous [17].

2. UNMASKED AND MASKED FACE SPEECH

For Arabic speech, there is a lack of research and literature focusing on the acoustic effects of face masks. The reasons for that, are the short period of the pandemic, the small number of mask-wearing people before the pandemic, and the difficulty with the unfeasibility of research on that subject. The following references and literature supporting this research are:

Atcherson *et al.* [18] found a clear difference between spectral analyzes of the speech stimuli with masks and without. The root mean square (RMS) value of that difference is about 2 dB. Consistently, the national health service (NHS) listeners performed the tests across different conditions. Transparent masks provide a benefit of visual input for listener groups with hearing impairment. Speech perception with noise was greatest on the improved magnitude scale for the group with severe-to-profound hearing loss. Corey *et al.* [19] denote the muffled speech due to face masks with more difficulty of communication for people with hearing loss. Acoustic attenuation caused by different face masks is examined using different types of masks. A human talker and head-shaped speaker are used. The resulting speech for all masks is attenuated above 1 kHz frequencies. The greatest attenuation occurs in front of the speaker. Between cloth masks, there is substantial variation due to weaves and material types. Compared to cloth and medical masks, transparent masks have bad acoustic performance. Lapel microphones have a negligible effect against most masks. The researchers suggested that assistive listening systems and existing sound reinforcement are useful for masked verbal communication. Deshpande and Schuller [20] summarized the researcher’s community efforts toward helping society and individuals, against the pandemic by speech and audio digital signal processing. Deep techniques are summarized to contribute short-term solutions. The article is an overview of the contributions from modalities of non-speech. These modalities serve or complement as inspiration for speech and/or audio analysis. The researchers discussed the observations with challenges, feasible solutions, and the achievements of significant technologies. Ribeiro *et al.* [21] discussed the following difficulties due to face masks: intelligibility of speech, vocal effort perception; auditory feedback, and speech coordination. The researchers concluded that for necessary and professional activities, face mask-wearing had a higher perception for symptoms of discomfort and vocal fatigue, increased vocal effort, difficult intelligibility of speech, and speech coordination. Bottalico *et al.* [22] studied the effects of face mask-wearing on communication in a classroom. Evaluation of speech intelligibility variations due to traditional face masks (N95, medical/surgical, and fabric). Auralized classroom students have presented that speech intelligibility. Realistically, under 0.4 s and 3.1 s reverberation times, classroom conditions are simulated. With a 3 dB signal-to-noise ratio (SNR), speech-shaped noise presents speech stimuli. A greater drop in speech intelligibility was yielded due to fabric masks in comparison to N95 and medical/surgical masks. For teaching environments, they recommend N95 and/or medical/surgical masks. Das and Li [23] studied audio features using linear filter-banks. An instantaneous phase with long-term attributes captures classified artifacts of the speech signal with a face mask and without. The extracted features were used alongside the following toolkits: deepspectrum, bag-of-audio-words, audeep [24], and computational paralinguistic evaluation functional (ComParE). They revealed the capability of audio features, and the score fusion level using the baselines of ComParE2020 produces 73.5% test sets of average recall. With a noisy background, Thibodeau *et al.* [25] investigated the auditory-visual recognition of 154 talkers with and without opaque and transparent masks. The researchers continued their smaller study with 29 talkers. Observed differences between opaque and transparent masks have been attributed to acoustic differences and visual gestures. In a quiet room, online sessions, listeners heard 40 minutes via listening devices. The devices are assistive hearing aids and earbuds. The talkers had normal hearing, suspected or confirmed hearing loss, and were using listening assistance devices and without the devices. Nguyen *et al.* [26] compared speech measurement via a record of 16 adults with a KN95 or a medical mask and without. Average of spectral levels for the 2 bands, below 1 kHz and between 1 kHz to 8 kHz, the researchers analyzed the first band to the second energy ratio, HNR for the 2 bands, vocal intensity, and smooth cepstral peak prominence (CPPS) of the 2 bands. There is an obvious average spectral level attenuation at the 1-8 kHz band; meanwhile, the attenuation is negligible at less than 1 kHz band. For face masked speech, vowel average spectral levels had little change. HNR is greater for the face masked speech than the unmasked speech. Mask-wearing did not affect the vocal intensity and CPPS much. Cohn *et al*. [27] tested the influence on comprehension of a fabric face masked speech. Three styles of speech (clear, casual, positive-emotional) with and without masks were compared. Subjectively, listeners had tested the speeches. In word identification, the tests were denoted as highly accurate for babbling clear conversation, and for casual conversation, they were denoted as not very accurate. The accuracy for emotional speech was moderate. For clear style, face masked speech had greater intelligibility than the unmasked. For emotional style, the face masked speech had lower intelligibility than unmasked. No significant difference was observed for the casual style. This may imply that emotional/ casual styles had less of an intent to be understood clearly by the listeners. Toscano and Toscano [28] studied N95 respirator, surgical, and cloth masks' effects on speech recognition with multi-talker-babble. On quieter backgrounds, masks had insignificant effects, less than a 5.5% reduction in accuracy relative to unmasked conditions. In background with higher noise levels, average accuracy reduction ranged from 2.8% to 18.2% relative to unmasked conditions, except for surgical masks. The study demonstrated that most mask types yielded similar accuracy for low noise level backgrounds; however, differences among the masks were more pronounced in high noise level backgrounds.

3. DATABASE FOR FACE MASKED ARABIC SPEECH

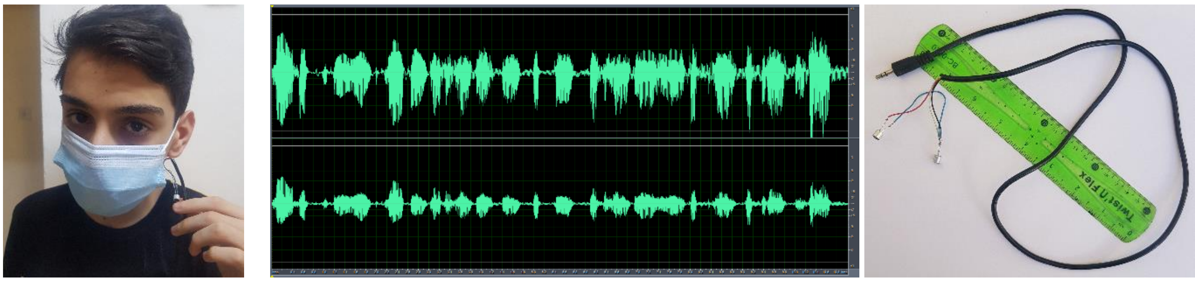
The main issue facing the research was the unavailability of a face-masked Arabic speech database. Standard speech databases and Arabic academics did not provide the required face masked speech which is recorded under restricted conditions (the restriction is very important to achieve a fair comparison). This article's researchers made efforts and consumed a lot of time to produce such a database. The difficulty is due to middle-eastern culture, privacy, and security matters. The female speech recording was more difficult than the male.

The researchers have built the required face masked/unmasked Arabic speech database with the two genders and wide-range age of speakers. The female/male speakers were divided into five age groups: under 12, 12–18, 18–40, 40–55, and older than 55 years old Figure 1(a). For each previous age range, 6–10 persons have recorded their speech. About 50% of them are female, and the others are male. All of them, have recorded the same Arabic (Iraqi accent) counting sentence: "*wahid* واحد*, thnain* ثنين*; thelatha* ثلاثه*; arba'a* أربعه*, khemsa* خمسه*, cita* سته*, seba'a* سبعه*, thamanya* ثمانيه*; tissa'a* تسعه*; eshra* عشره*,* and *hde'ash* هدعش". They mean counting from one to eleven; Figure 1(b). To get more reliable results and conclusions, this article's Arabic database has been compared subjectively and objectively with the standard acoustic databases such as "TIMIT acoustic-phonetic continuous speech corpus". Our database can be evaluated as a small Arabic reliable database for face masked/unmasked speech for different ages and genders.

Typically, each person masked his mouth, nose, and jaw with a medical mask. PC sound-card with double channel (stereo) and 3.5 mm audio jack were used for the recording. On the other terminal of the audio cable, two microphones are connected. The microphones are fully identical in the installation and brand. The first microphone is connected to the left channel, and the second is connected to the right. The first microphone is located under the mask (records the face-unmasked speech) and the second is located outside the mask (records the face masked speech). Different types of microphones were tested, and then the least noisy microphones were chosen and installed with similar lengths and types of terminals; Figure 1(c).

Each person said the counting sentence one time only, so the sound-card records simultaneously two speeches (the face-unmasked and the face masked). The recorded speech is saved as a ".wav" double channel (stereo) audio file. The sampling rate of the recorded speech is 16 kHz with 16 bit/sample resolution. Depending on the talker, the period of the counting sentence is 10 to15 seconds (s). Before processing, long durations of silence are removed manually. Recorded speech that's suffered from clipping is deleted and then re-recorded carefully. Subjectively, there's a clear similarity for the envelope of normalized speech, but there's a difference between their amplitude (sample-by-sample); Figure 1(b). The differences are due to the physical effects of the face mask. The Adobe-Audition® and the Audacity® audio signal processing applications were used for recording, playing, editing, spectrogram displaying, frequency-domain analyzing, subjective testing, and visual plotting.

Before the main process to detect F0, the recorded ".wav" files were tested in the time and frequency domains by using the Audacity® and the Adobe-Audition® integrated development environment (IDE). Those domains reflect the main views of the recorded audio speech. The other secondary parameters have been tested subjectively by those IDEs such as noise, silence, and tones. The analysis features have been used to check the spectrum, the contrast, and the clipping. For more details, the "tools" tag was used for the cross-checking between the two channels of each speaker, between the speech of different genders of each channel, and then among the same gender speakers for each audio channel.



(a) (b) (c)

Figure 1. Recording database: (a) 14 years old male speaker; (b) a typical sample of stereo unmasked and masked speech of Arabic for counting from 1 to 11; and (c) installation of 3.5 mm 2-channel identical microphones; the first is for unmasked speech and the second is for masked

4. EXPERIMENTS AND SUBJECTIVE TESTS

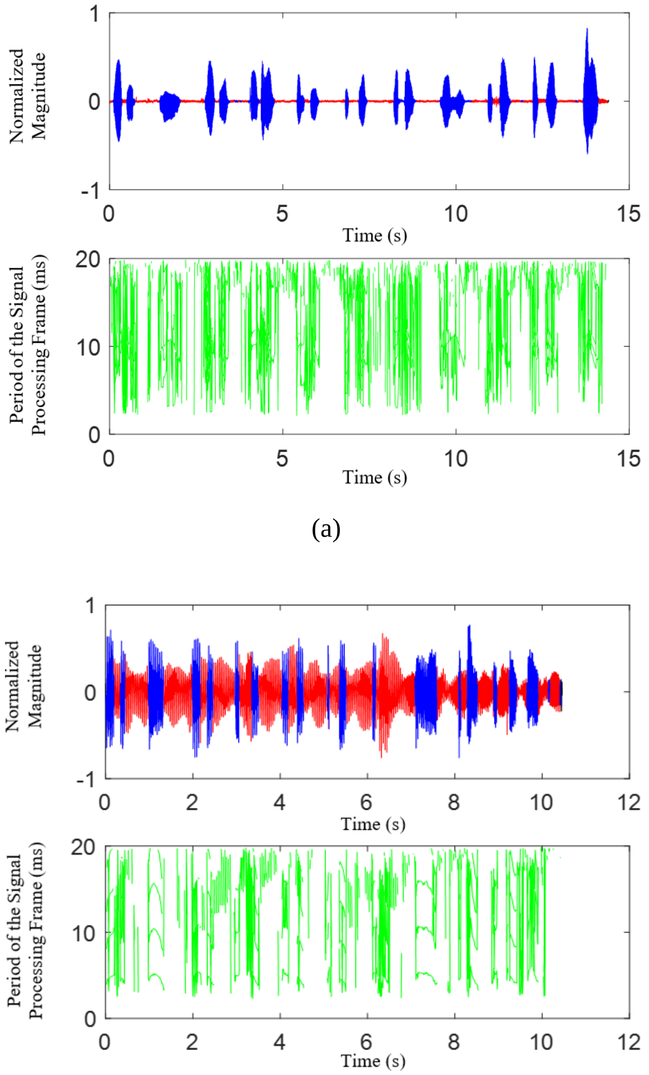
By running the RAPT algorithm Matlab® programs, the F0 is estimated for the voiced speech. The unvoiced speech (without the F0) is also detected. The instants of the voiced speech (with its F0) and the unvoiced speech are precisely extracted. Besides the Matlab integrated development environment (IDE), the Notepad++® edits the required speech parameters (e.g., sampling rate) in the source file, and the Audacity® plays and displays the 2-channel speech signals.

The RAPT default values of Talkin and Kleijn [15], Gonzalez and Brooks [29] are: Hanning window, the sampling rate of 16 kHz, minimum F0 of 50 Hz, maximum F0 of 500 Hz, frame time of 10 ms, low pass (LP) filter window size of 5 ms, the correlation window size of 7.5 ms, minimum peak in normalized cross-correlation function of 0.3, taper factor of 0.3 (linear lag), F0 change of 0.02 cost factor, transition cost of 0.005 (voice state fixing), transition cost of 0.5 for delta modulation, the bias for encouraging voice of 0 (hypotheses), doubling/halving of 0.35 cost for exact values, the noise level of 0 for absolute RMS, a level relative of noise (RMS noise) floor of 2, SNR of 0.001 (peak S to floor R), window length of 30 ms (RMS measurement), window spacing of 20 ms (RMS measurement), maximum hypothesis for each frame of 20, position in s-plane of -7000 (pre-emphasis 0.0), and the number of full lags to try of 7. The above default parameters have been changed to adapt to the masked/unmasked speech signal of the research. The sampling rate is still at the standard 16 kHz with 2 bytes (16-bit resolution for each sample of the two observation signals). The range of F0 was increased to cover the 40 Hz to 600 Hz band. The LP filter window was 5 ms with an additive of 2.5 ms for the correlation-window size (i.e., 7.5 ms). Most of the other parameters have been changed by ± 25% of the standard default values. The length of the main window (speech frame) was 25 ms to 40 ms with 15 ms to 30 ms spacing.

The time-domain plots by graph-mode implementation illustrate the lag candidates and F0 larynx frequency of the voiced speech frame-by-frame. For unvoiced speech and/or silence, "nan" is returned. For the candidate F0, start and/or end samples for the frame, a flag is returned at the beginning of each speech spurt Figure 2(a). Suggestions and bugs include backward dynamic programming (DP) for the pass with true-cost output for any F0 candidates; discrimination between the silent and/or the voiceless state, the best DP for the long-period penalties such as twice or half frequency F0. After the necessary implementations, the resulting data are collected. The subjective tests of the data denote that:

- More than 3 dB attenuation in the signal energy of the face masked speech.
- The attenuation and the noisy-condition effects are similar for both genders.
- Age groups did not affect the noisy background and the attenuation.

The above subjective tests gave us rough indications of the time-domain formulation of the signals for the Arabic masked and the unmasked speech of the two genders and the five ranges of age groups. The tests did not have exact numerical values for the attenuation and how much SNR for the two signals with the comparisons between these ratios, Figure 2(a) illustrates the low-noise background and Figure 2(b) on the high-noise background. Since the above subjective tests did not have enough evaluations to decide the effects of the medical mask on the F0 of Arabic speech, the researchers used two approaches for that task. The first approach is the objective standard and non-standard tests. The second approach is the mathematical model which supports the objective tests. For that, the researchers used statical analysis, because they have different useful parameters. The analysis was used to support the standard objective tests. The next title has more details about the analysis and standard objective tests.



(a)

(b)

Figure 2. Typical graphs for face masked speech: (a) on the low-noise background and (b) on high-noise background

5. OBJECTIVE TESTS

The subject tests of the experiments' results did not indicate exactly the numerical values of F0 increase/decrease shifting due to the mask-wearing. The SNR ratio also cannot be calculated by subjective tests. Using the following statistical process, objective tests are invoked to investigate the effects of mask-wearing on the F0 for the above 5 age groups of female/ male speakers. The standard and non-standard following objective-tests criteria are measured to compare the F0 of the unmasked and face masked speech [30]: i) minimum (F0) of the F0 minima (MnF0) of the conversations of each age group; ii) maximum (F0) of the F0 maxima (MxF0) of the conversations of each age group; and iii) mean (M) is the average value of the averages of the conversations of each age group.

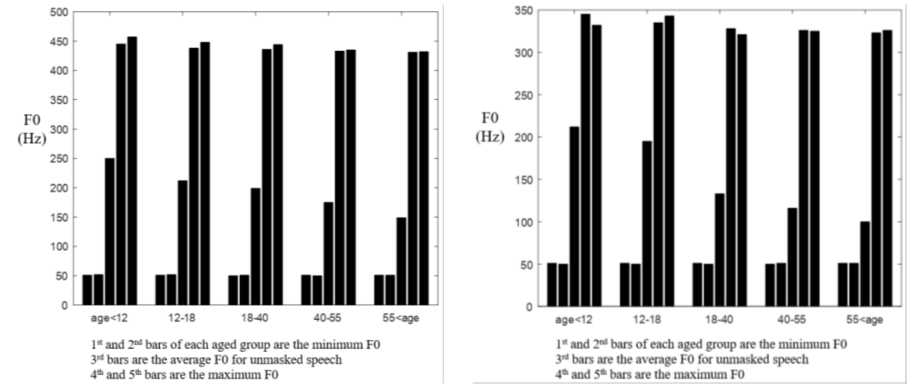
The listed data in Table 1 (for females) and Table 2 (for males) confirm the fact that F0 for the female is higher than F0 for the male for the masked speech and the unmasked speech. The tables provide good indications of the minimum, average, and maximum F0s, but do not clarify the overall details about other F0s: Figure 3(a) is for females, and Figure 3(b) is for males. For that, statistical analysis is exploited.

Table 1. The minimum of F0 minima (MnF0), an average of F0 averages (M), and the maximum of F0 maxima (MxF0) for unmasked (U) and masked (MSK) speech of 5 female age groups of speakers F0 (Hz)

	Age < 12 years old		12–18 years old		18–40 years old		40–55 years old		55 years old < age	
	U	MSK	U	MSK	U	MSK	U	MSK	U	MSK
MnF0	51	52	51	52	50	51	51	50	51	51
M	250		212		199		175		149	
MxF0	445	457	438	448	436	444	433	435	431	432

Table 2. The minimum of F0 minima (MnF0), the average of F0 averages (M), and the maximum of F0 maxima (MxF0) for unmasked (U) and masked (MSK) speech of 5 male age groups of speakers F0 (Hz)

	Age < 12 years old		12–18 years old		18–40 years old		40–55 years old		55 years old < age	
	U	MSK	U	MSK	U	MSK	U	MSK	U	MSK
MnF0	51	50	51	50	51	50	50	51	51	51
M	212		195		133		116		100	
MxF0	345	332	335	343	328	321	326	325	323	326

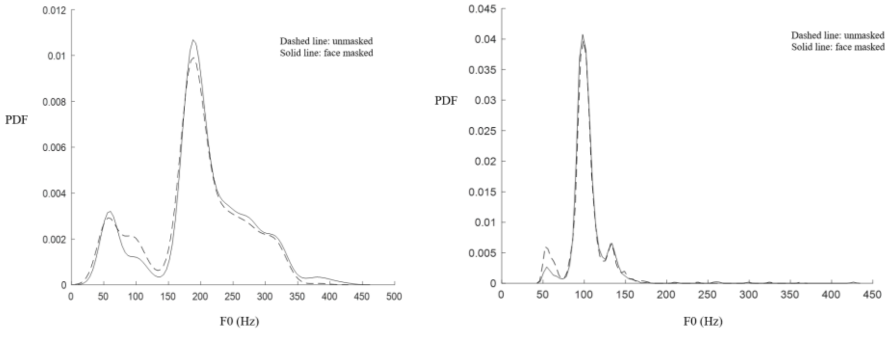


(a) (b)

Figure 3. For unmasked and masked speech: (a) for female speakers and (b) for male speakers

a) Probability density function (PDF) of F0, by using the histogram calculations, the F0 for the unmasked and face masked speech of male and female speakers are illustrated in Figure 4(a) for female and Figure 4(b) for males. The F0 range is from 50 to 500 Hz. The subjective test denotes that the F0 distribution for the male has more variance than the female F0 distribution. Male F0 is concentrated on the higher frequencies, while the female is concentrated on the lower.

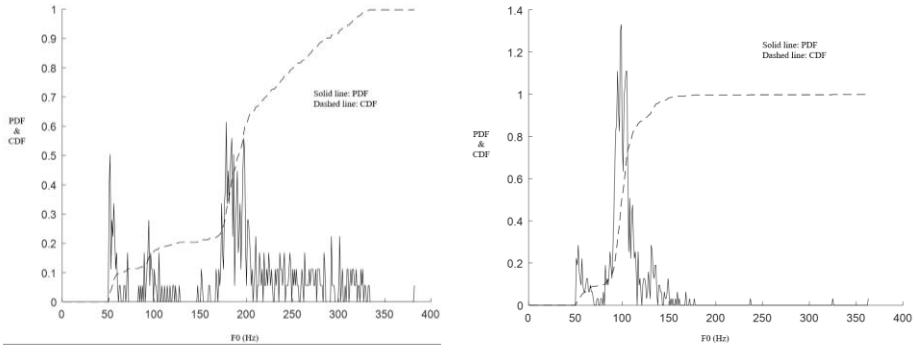
Subjectively, most conversations in the recorded database have two major lobes in the PDF distributions and many minor lobes in those distributions. The first major lobe, at the lower frequencies, has less energy than the second major lobe at the higher frequencies. The F0 distribution concentrates on those two major lobes, i.e., most of the F0 energy is located inside those two major lobes.



(a) (b)

Figure 4. Probability density function (PDF) of F0 for the unmasked and face masked speech: (a) female and (b) male

b) For the original unmasked speech, the cumulative distribution functions (CDF) are calculated from their corresponding PDFs and illustrated in Figure 5(a) for females and Figure 5(b) for males. According to statistical axioms, CDF = 0.5 at the average value of PDF, i.e., at the average F0 of the speech. The F0 frequencies less than average F0 are considered as a lower band of the speech from 50 Hz to M. The F0 frequencies higher than average F0 are considered as an upper band of the speech from M to 500 Hz. Since the PDF distributions of most conversations in the recorded database have two major lobes in the F0 distribution, the CDF of those distributions has two main risings in the configuration through frequency domain distribution of them. The second rise is greater than the first because the energy content of the second is higher than the first across the PDF distribution.



(a) (b)

Figure 5. Normalized PDF and it's CDF of F0 for the unmasked speech: (a) for female face unmasked speaker and (b) for male face unmasked speaker

c) The classification error (CE) is the percentage ratio of the unvoiced-speech frames (for the unmasked speech), and those classified as voiced-speech frames (for the face masked speech). It's the ratio of the voiced-speech frames (for the unmasked speech), and those classified as unvoiced-speech frames (for the face masked speech). The CE is a standard objective test for the PDA algorithm. The CE percentage (%) tests are tabulated in Table 3 and Figure 6(a).

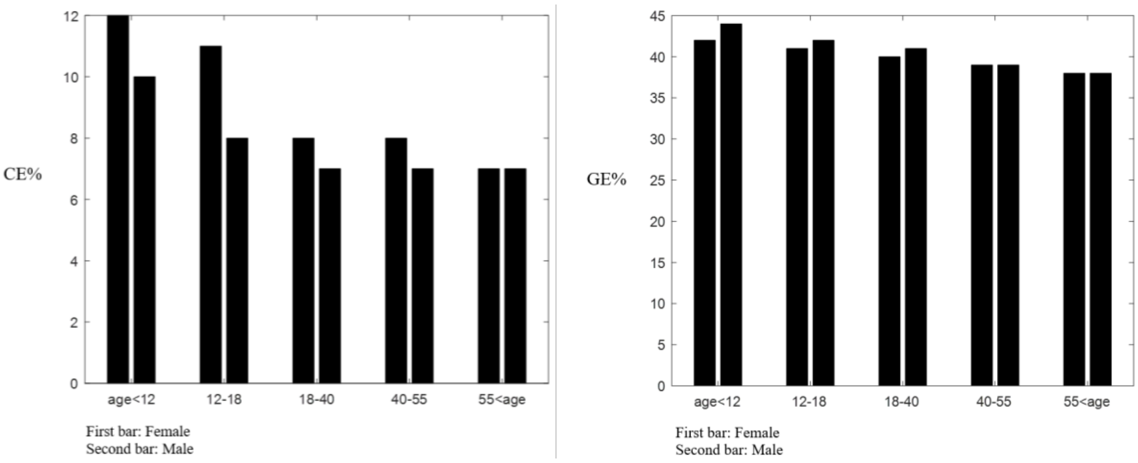
d) The gross error (GE) is the percentage ratio of the estimated pitch value of the voiced-speech frames (for the face masked speech), which deviates about 20% of the reference pitch value (for the unmasked speech). The gross error is a standard objective test for the PDA algorithm. The percentage gross error GE% tests are tabulated in Table 3 and Figure 6(b). More details will be illustrated in the next paragraphs.

e) The researchers proposed the following procedure to manipulate the details of F0 on the frequency domain. The F0 bandwidth range 50 Hz to 500 Hz is divided into two bands, LB from 50 Hz to average F0 (M), and the UB from average F0 (M) to 500 Hz. For the five female and male age groups of speakers, and Table 5 contain the comparison measurements of these groups concerning gender and sub-band. Figure 7 (for females) and Figure 8 (for males) illustrate these data (Figure 7(a) and Figure 8(a) for the lower band, and Figure 7(b) and Figure 8(b) for the upper band). According to these experimental measurements:

- Most of the lower-band F0 (55% to 70%) remained unchanged for different ages and genders. The average of them is about 40%.
- Most of the upper-band F0 (about 70%) remained unchanged for males older than 12 years old. For males less than 12 years old, there is a similarity with female upper-band F0 changing.
- For different female ages, the unchanged upper-band F0s fluctuate from 60% to 73%.
- For the lower band, the range of the number of F0 increases is (15% to 25%) of the total F0 for different ages and genders. Its average is about 20%.
- The average increase of F0 value is about 22% of the original F0 unmasked speech.
- For the lower band, the range of the number of F0 decreases is (15% to 22%) of the total F0 for different ages and genders.
- The decrease of the F0 value is less than 18% of the original F0 unmasked speech.
- For the upper band, the number of F0 increases and decreases is less than 16% of the total F0 for different ages and genders.

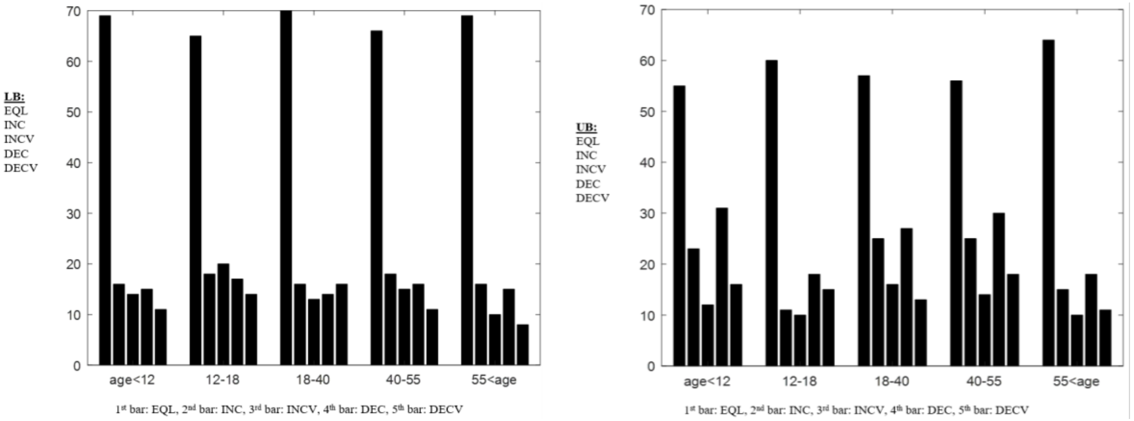
Table 3. The percentage CE, the percentage GE, and the mean value (M) of F0 for face masked speech of the 5 female/male age groups of speakers. Their original unmasked speech is the reference

	Age < 12 years old		12–18 years old		18–40 years old		40–55 years old		55 years old < age	
	Female (F)	Male (M)	F	M	F	M	F	M	F	M
CE%	12%	10%	11%	8%	8%	7%	8%	7%	7%	7%
GE%	42%	44%	41%	42%	40%	41%	39%	39%	38%	38%
M (Hz)	250	212	212	195	199	133	175	116	149	100



(a) (b)

Figure 6. The objective tests of F0 for face masked speech of the 5 age groups of speakers: (a) The CE% and (b) The GE%. Their original unmasked speech is the reference

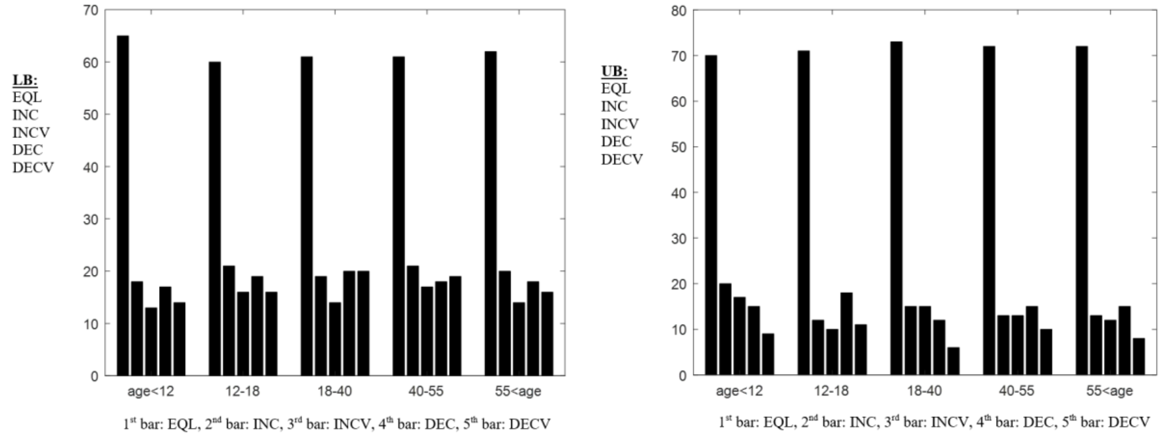


(a) (b)

Figure 7. Comparison between the LB and the UB of the 5 female age groups: (a) for the LB and (b) for the UB

Table 4. Comparison between the LB and the UB of the 5 female age groups

		age<12years old	12-18 years old	18-40 years old	40-55years old	55years old< age
EQL	LB	69%	65%	70%	66%	69%
	UB	55%	60%	57%	56%	64%
INC	LB	16%	18%	16%	18%	16%
	UB	23%	11%	25%	25%	15%
INCV	LB	0.14	0.20	0.13	0.15	0.10
(PU)	UB	0.12	0.10	0.16	0.14	0.10
DEC	LB	15%	17%	14%	16%	15%
	UB	31%	18%	27%	30%	18%
DECV	LB	0.11	0.14	0.16	0.11	0.08
(PU)	UB	0.16	0.15	0.13	0.18	0.11



(a) (b)

Figure 8. Comparison between the LB and the UB of the 5 male age groups: (a) for the LB and (b) for the UB

EQL is of the unchanged F0. INC and DEC are the F0 increase and decrease changes respectively. INCV and DECV are the per-unit (PU) values of the F0 increase and decrease changes respectively. The calculations are per-unit to the sub-band number of F0.

Table 5. Comparison between the LB and the UB of the 5 male age groups

		age<12years old	12-18 years old	18-40 years old	40-55years old	55years old< age
EQL	LB	65%	60%	61%	61%	62%
	UB	70%	71%	73%	72%	72%
INC	LB	18%	21%	19%	21%	20%
	UB	20%	12%	15%	13%	13%
INCV	LB	0.13	0.16	0.14	0.17	0.14
(PU)	UB	0.17	0.10	0.15	0.13	0.12
DEC	LB	17%	19%	20%	18%	18%
	UB	15%	18%	12%	15%	15%
DECV	LB	0.14	0.16	0.20	0.19	0.16
(PU)	UB	0.09	0.11	0.06	0.10	0.08

EQL is of the unchanged F0. INC and DEC are the F0 increase and decrease changes respectively. INCV and DECV are the values of the F0 increase and decrease changes respectively. The calculations are per-unit (PU) to the sub-band number of F0.

6. CONCLUSION

For the effect of mask-wearing against the Arabic speech F0, from the above data, tables, and graphs, the F0 changed less for males than females. The change is less for the older than the younger males and females. The change is less significant for the low-frequency F0 than the high-frequency F0 of the F0 bandwidth from 50 Hz to 500 Hz. The F0 changes in females younger than 12 years old are fewer compared with similarly-aged males. The F0 changes of females older than 12 years old were approximately equal compared with similarly-aged males. The probability density function and cumulative distribution function of F0 for different ages and genders have little shifting due to mask-wearing.

For future research, the study could be expanded by using several algorithms of F0 detection, such as YAAPT, PRAAT, YIN, and/or STRAIGHT (for these algorithms/ applications, more details in the introduction section of this article). In this research, the F0 band has been divided into two sub-bands according to the average value of F0 for each conversation. The F0 band can be divided into several sub-bands by using the standard filter-bank scheme or by using the wavelet configuration of the frequency domain for the F0 band. For the mathematical model which can be modified in this research, other models could be used instead of the statistical model (e.g., the stochastic model). Other statistical parameters could sustain the results of the research, such as the fourth-order (the kurtosis) of the central moment of the analyzed data.

ACKNOWLEDGEMENTS

We thank our Electrical Engineering Department; Engineering College; Mustansiriyah University “www.uomustansiriyah.edu.iq” / Iraqi MoHE and SR for supporting this research.

REFERENCES

[1] V. J. Williamson, A. D. Baddeley, and G. J. Hitch, "Musicians' and nonmusicians' short-term memory for verbal and musical sequences: Comparing phonological similarity and pitch proximity," *Memory & Cognition*, vol. 38, pp.163-175, 2010, doi: 10.3758/MC.38.2.163.

[2] P. C. Bagshaw, S. M. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching," 1993. [Online]. Available: https://www.cstr.ed.ac.uk/downloads/publications/1993/Bagshaw_1993_a.pdf

[3] A. D. Cheveigné, and H. Kawahara, "Comparative evaluation of F0 estimation algorithms," In *Seventh European Conference on Speech Communication and Technology*, 2001. [Online]. Available: https://www.ee.columbia.edu/~dpwe/papers/deChevK01-ptrack.pdf

[4] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE transactions on speech and audio processing,* vol. 8, no. 6, pp. 708-716, 2000, doi: 10.1109/89.876309.

[5] S. A. Zahorian, P. Dikshit, and H. Hu, "A spectral-temporal method for pitch tracking" In *Ninth International Conference on Spoken Language Processing,* 2006. [Online]. Available: http://www.ws.binghamton.edu/zahorian/yaapt/Zahorian2006Spectral.pdf

[6] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the institute of phonetic sciences*, 1993, vol. 17, no. 1193, pp. 97-110. [Online]. Available: https://www.fon.hum.uva.nl/paul/papers/Proceedings_1993.pdf

[7] PRAAT v6.1.49**,** P. Boersma, June 2021. [Online]. Available: "https://github.com/praat/praat/releases/tag/v6.1.49".

[8] X. Sun, "Pitch determination and voice quality analysis using Subharmonic-to-Harmonic Ratio," *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, pp. I-333-I-336, doi: 10.1109/ICASSP.2002.5743722.

[9] A. M. Noll, "Cepstrum pitch determination," *The journal of the acoustical society of America*, vol. 41, no. 2, pp. 293-309, 1967, doi: 10.1121/1.1910339.

[10] M. H. J. Gruber, "Statistical digital signal processing and modeling," *Technometrics*, vol. 39. no. 3, pp. 335-336, 1997, doi: 10.1080/00401706.1997.10485128.

[11] J. C. Brown and M. S. Puckette, "A high resolution fundamental frequency determination based on phase changes of the Fourier transform," *The Journal of the Acoustical Society of America,* vol. 94, no. 2, pp. 662-667, 1993, doi: 10.1121/1.406883.

[12] Y. Medan, E. Yair and D. Chazan, "Super resolution pitch determination of speech signals," in *IEEE Transactions on Signal Processing*, vol. 39, no. 1, pp. 40-48, Jan. 1991, doi: 10.1109/78.80763.

[13] H. Kawahara, A. de Cheveigné, H. Banno, T. Takahashi, and T. Irino, "Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT," in *Ninth European Conference on Speech Communication and Technology*, 2005, doi: 10.21437/Interspeech.2005-335.

[14] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *The Journal of the Acoustical Society of America,* vol. 124, no. 3, pp. 1638-1652, 2008. doi: 10.1121/1.2951592.

[15] D. Talkin and W. B. Kleijn, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis,* vol. 495, p. 518, 1995. [Online]. Available: https://www.ee.columbia.edu/~dpwe/papers/Talkin95-rapt.pdf

[16] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 2494-2498, doi: 10.1109/ICASSP.2014.6854049.

[17] Kaldi-ASR, J. H. University, 2020 . [Online]. Available"https://github.com/kaldi-asr/kaldi."

[18] S. R. Atcherson *et al.*, "The effect of conventional and transparent surgical masks on speech understanding in individuals with and without hearing loss," *Journal of the American Academy of Audiology,* vol. 28, no. 1, pp. 58-67, 2017, doi: 10.3766/jaaa.15151.

[19] R. M. Corey, U. Jones, and A. C. Singer, "Acoustic effects of medical, cloth, and transparent face masks on speech signals," *The Journal of the Acoustical Society of America,* vol. 148, no. 4, pp. 2371-2375, 2020, doi: 10.1121/10.0002279.

[20] G. Deshpande and B. Schuller, "An overview on audio, signal, speech, & language processing for covid-19," *arXiv*, 2020, doi: 10.48550/arXiv.2005.08579.

[21] V. V. Ribeiro, A. P. D. -Leite, E. C. Pereira, A. D. N. Santos, P. Martins, and R. de A. Irineu, "Effect of wearing a face mask on vocal self-perception during a pandemic," *Journal of Voice*, 2020, doi: 10.1016/j.jvoice.2020.09.006.

[22] P. Bottalico, S. Murgia, G. E. Puglisi, A. Astolfi, and K. I. Kirk, "Effect of masks on speech intelligibility in auralized classrooms," *The Journal of the Acoustical Society of America*, vol. 148, no. 5, pp. 2878-2884, 2020, doi: 10.1121/10.0002450.

[23] R. K. Das and H. Li, "Classification of Speech with and without Face Mask using Acoustic Features," *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2020, pp. 747-752. [Online]. Available: https://arxiv.org/pdf/2010.03907

[24] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, "Audeep: Unsupervised learning of representations from audio with deep recurrent neural networks," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6340-6344, 2017, doi: 10.48550/arXiv.1712.04382.

[25] L. M. Thibodeau, R. B. T. -Nielsen, C. M. Q. Tran, and R. T. D. S. Jacob, "Communicating during COVID-19: The effect of transparent masks for speech recognition in noise," *Ear and Hearing,* vol. 42, no. 4, pp. 772-781, 2021, doi: 10.1097/AUD.0000000000001065.

[26] D. D. Nguyen *et al.*, "Acoustic voice characteristics with and without wearing a face mask," *Scientific Reports,* vol. 11, pp. 1-11, 2021, doi: 10.1038/s41598-021-85130-8.

[27] M. Cohn, A. Pycha, and G. Zellou, "Intelligibility of face-masked speech depends on speaking style: Comparing casual, clear, and emotional speech," *Cognition*, vol. 210, p. 104570, 2021, doi: 10.1016/j.cognition.2020.104570.

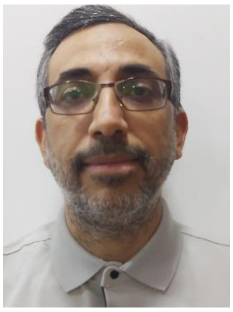
[28] J. C. Toscano and C. M. Toscano, "Effects of face masks on speech recognition in multi-talker babble noise," *PloS one,* vol. 16, no. 2, p. e0246842, 2021, doi: 10.1371/journal.pone.0246842.

[29] S. Gonzalez and M. Brookes, "A Pitch Estimation Filter robust to high levels of noise (PEFAC)," *2011 19th European Signal Processing Conference*, 2011, pp. 451-455.

[30] I. Luengo, I. Saratxaga, E. Navas, I. Hernaez, J. Sanchez and I. Sainz, "Evaluation of Pitch Detection Algorithms Under Real Conditions," *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, 2007, pp. IV-1057-IV-1060, doi: 10.1109/ICASSP.2007.367255.

BIOGRAPHIES OF THE AUTHORS

Hasan M. Kadhim is a lecturer in the Electrical Engineering Department, College of Engineering, Mustansiriyah University; Baghdad, Iraq. He received his B.Eng., M.Sc., and Ph.D. degrees in Electrical Engineering from Baghdad University, Mustansiriyah University, and Newcastle University in 1985, 2000, and 2017, respectively. His research field is Speech Separation and Speaker Diarization. He can be contacted at: hasanalmgotir@uomustansiriyah.edu.iq, https://uomustansiriyah.edu.iq/e-learn/profile.php?id=5466, https://github.com/HasanAlmgotir, https://uomustansiriyah.academia.edu/HasanMKadhim, https://www.kaggle.com/hasanmakadhim, https://www.researchgate.net/profile/Hasan-Ma-Kadhim.



Alaa H. Ahmed is a lecturer in the Electrical Engineering Department, College of Engineering, Mustansiriyah University; Baghdad, Iraq. He received his B.Eng., M.Sc., and Ph.D. degrees in Electrical Engineering from Baghdad University and Newcastle University in 1997, 2000, and 2017, respectively. His research field is Mobile Communication. He can be contacted at: alaa75hs@uomustansiriyah.edu.iq, https://uomustansiriyah.edu.iq/e-learn/profile.php?id=4566.

Saif A. Abdulhussien is a lecturer in the Electrical Engineering Department, College of Engineering, Mustansiriyah University; Baghdad, Iraq. He received his B.Eng., and M.Sc. degrees in Electrical Engineering from Mustansiriyah University in 2006 and 2012 respectively. His research field is Electronics and Communication Engineering. He can be contacted at: saifabc0001@uomustansiriyah.edu.iq, https://uomustansiriyah.edu.iq/e-learn/profile.php?id=1907.