International Journal of Speech Technology (2021) 24:1089–1100
https://doi.org/10.1007/s10772-021-09877-1

Mobile microphone robust acoustic feature identification using coefficient of variance

Nik Nur Wahidah Nik Hashim[1] · Mugahed Al-Ezzi Ahmed Ezzi[1] · Mitchell D. Wilkes[2]

Received: 23 December 2020 / Accepted: 27 July 2021 / Published online: 2 August 2021

Abstract
One of the most challenging techniques for speech analysis applications in mobile phones is acoustic feature extraction. The adverse environment noises, diversity of microphone specifications, and various recording software have a significant effect on the values of the extracted acoustic features. In this study, we investigate the robustness of different types of acoustic features related to time-based, frequency-based, and sustained vowel using 11 different mobile recording devices. 49 recordings of subjects reciting the Rainbow Passage and 25 recordings of sustained vowel /a/ were collected. By way of synchronous recording, we analyzed and compared the extracted 253-dimensional acoustic feature vectors in order to examine how consistent the data values between the different recording devices. The variability of data values was measured using the method of coefficient of variance. Data values with low variability were identified to be from features such as the transition parameters, amplitude modulation, contrast, Chroma, mean fundamental frequency and formants. These groups of features turn out to be more reliable than others in their dependency on the recording device specifications.

Keywords Acoustic features · Robust features · Microphones · Recording

Abbreviations

AFE	Acoustic feature extraction
SER	Speech emotion recognition
ASR	Automatic speech recognition
ADD	Automatic depression detection
ASC	Automatic scene classification
AGC	Automatic gain control
MEMS	Micro electromechanical system
TP	Transition parameter
ILpdf	Interval length probability density function
PSD	Power spectral density
MFCC	Mel-frequency cepstral coefficient
AM	Amplitude modulation
VUS	Voiced, unvoiced and silence
ch	Chroma
con	Contrast
f0	Fundamental frequency
HNR	Harmonics-to-noise ratio
COV	Coefficient of variance

✉ Nik Nur Wahidah Nik Hashim
niknurwahidah@iium.edu.my

[1] Department of Mechatronics, Faculty of Engineering, International Islamic University Malaysia (IIUM), Selangor, Malaysia

[2] Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN, USA

1 Introduction

1.1 Motivation and objective of the study

Smart technologies can be made a significant contributor to improve peoples' lives especially with the new information technologies such as big data, cloud computing, the internet of things and artificial intelligence. However, in the field of speech or acoustic technology such as speech emotion recognition (SER), automatic speech recognition (ASR), automatic depression detection (ADD) and automatic scene classification (ASC), capturing the embedded information can be quite challenging. For example, the human voice is encoded with a wealth of information regarding mood, stress condition, affective state and mental state of the subject. When most variabilities are removed and conditions are fulfilled, acoustic analysis of voice offers significant benefits in understanding the vocal output parameters depending on the objectives. These conditions include high-quality and consistent data acquisition, controlled environment with minimal noise background, powerful system and software

analysis, robust model algorithm and proper set up of speech task. Therefore, if either one of these variabilities is not eliminated, the information gathered can be challenging to compare and analyse.

Slight attention has been given to the speech technology in the mobile service environment. There are multiple applications that uses mobile application for sound analysis such as to assess noise exposure risk from the consumers' perspective (Sinha et al., 2016), for remote noise monitoring and data acquisition (Dickerson, 2016) and the accuracy study of using iPhone's app for routine collection of infants' noise exposures inside the *isolette* during air transport (Clark & Saunders, 2016). Mainly, the studies conducted in this field use a high-quality type of microphone with minor noise condition. A technical review by Svec and Granqvist (2010) suggested that microphone recommendation depends on phonation task and proximity, but not necessarily the price or quality. They also reported that spectral properties of sound are independent of proximity effect and dynamic range due to the production of voice is always at the comfortable levels, instead, it depends on the signal-to-noise ratio. However, for a more general and broader implementation such as mobile apps, the basis of these experiments tend to fail when speech signal query contains background noises that vary with the subjects' environments and obtained through variability of microphone types. Overall, due to the multiple microphones and environmental noises (Deliyski et al., 2006), a speech signal acquired by recording devices may cause significant performance reduction to the extracted speech features and parameters. However, combining findings by Svec and Granqvist and also by Faber (Faber, 2017), with a careful selection of device, app and microphone specification, accurate measurement can also be made with approximately high level of precision.

The essential problem in mobile phone technology for monitoring-based or detection-based application that uses voice as an input is that various built-in microphones in mobile devices are manufactured with customized automatic gain control (AGC), active noise cancellation, noise rejection strategy using directionality of two or more microphones and beamforming for speech enhancement. The built-in micro electromechanical system (MEMS) microphones cannot provide the same performance as compared to the professional microphone used in acoustic fields. For example, the presence of AGC is a unique circuit designed by the manufacturer to adjust the recording level when the input sound is too loud, or too soft. However, AGC is not able to distinguish between actual sound information or noise. The effect of AGC might significantly reduce the overall accuracy of sound level measurement but the effect on specific frequency components of the acoustic signal may or may not be influenced. In addition, the measurement performance is also influenced by the hardware components which possess variation in its frequency response, dynamic range and sensitivity. The study of acoustical energy in typical human speech conversation has typically been restricted to the frequency range of less than 8 kHz, where the term 'higher frequency' usually refers to frequencies within 2 to 8 kHz. Different smartphones have attenuation at different frequency ranges but minimizes the attenuation in the middle band of around 200 Hz to 2 kHz for human voice recording.

This work was proposed due to the limitation we encountered in the research of automatic depression detection using speech during the pandemic. Our movement was restricted and thus, we were not able to visit the hospitals for data collection. We then proceeded with gathering voice recordings through online platform such as using mobile phone devices. The research question we attempt to address is whether there are voice acoustic features that are not altered by different mobile recording devices. These unaltered features can be used to eliminate the bias in the multiple recording devices and apply the feature for further classification analysis. Although current trend may suggest the use of deep learning frameworks to learn robust features and see these variations in training, research in this field has one major challenges, which is limitation of database quantity. Even with a well-designed backend (classifier or regressor), it still requires sufficient amount of data in order to be robust to noise or disturbances in the features. Since this is the limitation, we had to focus on identifying the robust features in order to apply it to classifiers. These robust features can be collected and used for any application that requires voice as input. However, for this work, we demonstrated the application of robust features on depression detection.

1.2 Related work

The study on comparing vocal acoustic changes was initiated by Karnell et al. (1991) where the author began the investigation on particularly jitter and shimmer perturbation analysis across three voice laboratories. It was conducted in an attempt to find a standardize procedures in voice recording analysis and hardware in order to facilitate the interpretation of results from various laboratories. Jitter and magnitude of shimmer measurements differed significantly between the three laboratories due to multiple analysis techniques, different hardware digitization resolution and inadequate noise-free amplification.

Another study by Titze and Winholtz (1993) explored the effects of microphones on the voice perturbation measures from sustained phonation. The study was designed based on five microphone characteristics and settings; (i) professional-grade and consumer-grade, (ii) microphone types which are dynamic and condenser, (iii) omnidirectional and cardioid pattern, (iv) distance and (v) angle between the source and

the microphone. Based on the analysis, the author recommended a professional-grade cardioid or omnidirectional condenser microphone which can be placed a few centimetres from the mouth at an angle of 45° to 90°. However, the sampled recordings were limited in number and analysis was only performed on sustained phonation. The results also demonstrate the variability of acoustic measures when multiple hardware characteristics were used.

The work done by Parsa et al. (2001) was an extend from Titze and Winholtz (1993) with an inclusion of glottal noise measure using three cardioid microphones and an omnidirectional microphone. The comparison between acoustical measurements reveals that the absolute jitter, which is calculated using the mean absolute difference of successive pitch periods, was not significantly different for all microphone types. This result contradicts the one reported by Karnell et al. However, other acoustic measures derived from the fundamental frequency were significantly different due to the alteration of temporal structure of the signals caused by the frequency responses of the microphones. Parsa's findings were then confirmed with the work done by Bottalico et al. (2018) where they stated that jitter and smooth cepstral peak prominence were the most independent to variations in microphones. Bottalico et al. also reported that the effect of room is higher than the incompatibility associated with the effect of microphones.

Kisenwether and Sataloff (2015) continued studying the effect of different microphone types to the acoustic measurements. This study compared the acoustical measurement from nine synthesized stimuli with the recorded stimuli sound through six types of microphones. They performed a mean subtraction of each acoustical measurement (f_0, f_0 standard deviation, absolute jitter shimmer, peak-to-peak amplitude, and Noise Harmonic Ratio) between synthesized and measured values and reported that the means were not statistically significant. However, they were cautious about generalizing the results due to the limited number of microphones tested. Similar to the previous studies, Krik et al. (2019) performed analysis on recordings from two different microphones using f_0, jitter, shimmer and Glottal Noise Excitation ratio on sustained vowel. They extracted the descriptive statistics and compared them between the two microphones. Results revealed that all three features were robust towards different microphones except for shimmer.

It would be an ideal method for an affective or mental health monitoring by voice if the features can be shown to be robust to be used on various mobile applications or devices. The issue for researchers is the distortions contained in the features of the query sound that are introduced by recording the sound through various mobile devices. If multiple microphones produce different acoustical measures, it will be difficult to identify which values are correct in representing the true measurement. Thus, comparability with the human voice and results can be questionable. This paper is a preliminary study which focuses on identifying robust acoustic features from voice recordings that are gathered through different mobile devices which also means that the devices have multiple microphone settings. The main question this paper addresses is what are the acoustic features from voice recordings that are not affected by the various mobile devices? The analysis will be performed on extracted multiple acoustic features of time-based and frequency-based using query speech captured by mobile phones.

The structure of the paper begins with presenting previous literatures that have particularly studied changes in speech acoustic features using multiple microphone types and specifications. Next section talks about the experiment including information on database and the extracted features Next, we presented our analysis on identifying which features are less affected by different microphones and discuss the results, specifically focusing on features that are robust. Finally, we presented our conclusion of the overall investigation.

2 Experiment

2.1 Data collection

Table 1 lists the brand, model and specifications for the mobile phones used in the recording, obtained from information on mobile specifications that are available online. A special test track was inputted into the device and played through the audio output and into the M-Audio Fast Track Pro external audio interface. The generated recordings were then analysed through the RightMark Audio Analyzer (RMAA) software which produces sound quality measurements such as frequency response, dynamic range, and harmonic distortion. Other specifications were obtained on the mobile's manual information.

For this work, two types of speech recordings were collected in order to determine the robustness of the acoustic measurements. All procedures performed in studies involving human participants were in accordance with the ethical standards and has been approved by the IIUM Research Ethical Committee (IREC 2019-006). Subject Informed Consent (SIC) was also obtained from all individual participants included in the study. Approximately 1.5–2 min of speech utterances reciting the rainbow passage were recorded using seven mobile phone devices, simultaneously. The rainbow passage is a short passage that contains alliterations and irregular consonant and vowel combinations that is commonly used by speech therapist to assess vocal abilities. The recordings were collected using the mobile phone's default recording application and then sent through the mobile application called

Table 1 Brands and specifications of mobile phones used for recording voice samples

ID	Brand	Model	Specifications					
			Frequency response	Dynamic range	Total harmonic distortion	Bit rate (Kb/s)	Channel	Sampling rate (KHz)
Speech utterance (rainbow passage)								
A1	Apple	iPhone 7	+0.06, − 0.10	92.3	0.0015	64	1	48
A2	Apple	iPhone XR	+0.03, − 0.04	93.5	0.0016	64	1	48
A3	Oppo	A71	+0.03, − 0.07	93.9	0.0012	320	2	48
A4	Samsung	Galaxy S10	+0.03, − 0.04	92.0	0.0015	128	1	44.1
A5	Huawei	Y9	+0.01, − 0.03	93.0	0.0013	148	2	48
A6	Apple	iPhone 6	+0.03, − 0.04	93.5	0.0016	64	1	44.1
A7	Oppo	Reno 2	+0.04, − 0.05	93.0	0.0015	320	2	48
Sustained vowel								
B1	Apple	iPhone SE	+0.01, − 0.06	93.0	0.0013	64	1	48
B2	Apple	iPhone 8 Plus	+0.07, − 0.01	93.2	0.0013	64	1	48
B3	Samsung	A30s	+0.03, − 0.05	93.0	0.0068	128	1	44.1
B4	Samsung	Note 9	+0.01, − 0.03	93.7	0.0017	256	1	48
B5	Apple	iPhone 6	+0.03, − 0.04	93.5	0.0016	64	1	44.1

Whatsapp voice note for collection. Each recording was prepared in a closed room from a single speaker. A total of 49 recordings were gathered from seven speakers that participated in the experiment.

We decided to gather additional acoustic measurements related to sustained vowel considering that the previous literature (Kisenwether & Sataloff, 2015; Krik et al., 2019; Parsa et al., 2001), studied these features in their experiments. In another session, approximately six seconds of sustained vowel /a/ were recorded following the previous methods on five mobile phone devices (device ID B1-B5) listed in Table 1 from five participants. A total of 25 recordings were collected from five speakers.

The recordings received in the voice note were in.OGG and.MP4 format. Speech files were converted to.WAV using the audio.online-convert.com at a sampling rate of 44.1 kHz and 32 bits per sample, with a mono channel. The recordings were normalized prior to the feature extraction.

Figure 1 shows the overall process for the analysis of robust features. Each subject will be recorded individually using all seven mobile devices simultaneously. After that, we perform feature extraction and measure the variability

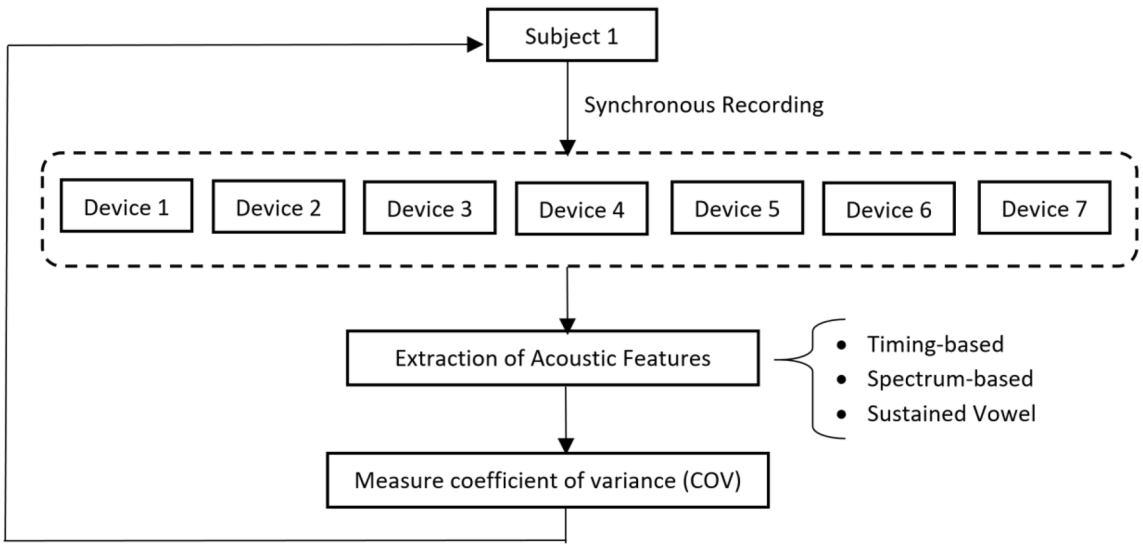


Fig. 1 Process flow for robust feature analysis

of each feature's data value using the method of coefficient of variance.

2.2 Feature extraction

Once the speech signals are obtained and pre-processed, the acoustic features were extracted. The same number of features are used for each voice sample. In other words, a 242-dimensional and 11-dimensional feature vectors are extracted from the utterance speech signal and sustained vowel, respectively. Table 2 lists the extracted speech features. Features ID 1 to 6 were extracted using MATLAB following the previous work done by Hashim et al. (2017), sustained vowel feature ID 7 was extracted using Praat software (Styler, 2013) and features ID 8 to 12 were extracted using Librosa library in Python. These features were chosen because they are commonly used in the studies related to voice or sound analysis.

These features can be divided into three categories which are the time-based features (ID 1, 2, 3 and 6 in Table 2), frequency-based features (ID 4, 5, and 8 to 12 in Table 2) and sustained vowel features (ID 7 in Table 2).

Figure 2 shows the overall process for acoustic feature extraction using MATLAB. The Transition Parameter (TP) feature captures the nine probabilities between each 40 ms frames labeled as voiced, unvoiced and silence (VUS) within one speech sample using the method of Markov Model. Parameter t11 represents transition probability from voiced-to-voiced frames in one whole voice sample and so on. For the interval length probability density function (ILpdf), the frequencies of consecutive interval ratios of 40 ms VUS frames were plotted in histogram and normalized on 4 bands of 40 ms to 0.8 s voiced segments and 5 bands of 40 ms to 2 s silence segments.

The method used to obtain the Amplitude Modulation (AM) is the 'square-law envelope detector' which squares the input signal and sends it through an averaging represented by a low-pass filter (gain = 1). The square root is then taken in order to reverse the scaling distortion from squaring the signal and to characterize a more accurate statistical measure.

Acoustic features based on pitch, loudness, and timber are psychoacoustic properties of auditory signals commonly used for speech and music analysis. These types of features can be categorized as spectrum-based features. Researchers in this field suggest aggregating acoustic features such as PSD, Chroma, Mel Spectrogram, MFCC, Spectral Contrast, and Tonnetz over performed single features in the automatic speech recognition systems (Ghosal & Kolekar, 2018; Su et al., 2020). These features are extracted by utilizing Librosa Python library (Mcfee, et al., 2015).

Power Spectral Density (PSD) describes the power present in the speech signal as a function of frequency and for this work, we obtained PSD using the method of Periodogram and normalized on the 4 bands of 0 to 2000 Hz.

Mel Frequency Cepstral Coefficients (MFCC) are commonly used in automatic speech and speaker recognition which are introduced by Davis and Mermelstein (Davis & Mermelstein, 1980). The MFCC feature extraction technique includes windowing the signal, applying the discrete fourier transform, taking the log of the magnitude, and then warping the frequencies on a Mel scale, followed by applying the

Table 2 List of extracted acoustic features

ID	Feature	Dimension	Parameter labels
1	Transition parameter (TP)	9	t11, t12, t13, t21, t22, t23, t31, t32, t33
2	Interval length probability density function (ILpdf)—silence	5	sil1 (0.4 s), sil2 (0.8 s), sil3 (1.2 s), sil4 (1.6 s), sil5 (2 s)
3	ILpdf—voiced	4	v1(0.2 s), v2(0.4 s), v3(0.8 s), v4(1.2 s)
4	Power spectral density (PSD)	4	PSD1 (0–50 Hz), PSD2 (501–1000 Hz), PSD3 (1001–1500 Hz), PSD4 (1501–2000 Hz)
5	Mel-frequency cepstral coefficient (MFCC)*	13	c1 to c13
6	Amplitude modulation (AM)	8	Minimum, maximum, range, variation, average, skewness, kurtosis, coefficient of variance
7	Sustained vowels	11	Mean fundamental frequency (f0), pitch std. dev., pitch sigma, jitter, shimmer, harmonics to noise ratio (HNR), formants (F1, F2, F3, F4)
8	MFCC**	40	mfcc1 to mfcc40
9	Chroma	12	ch1 to ch12
10	Mel-Spectrogram	128	m1 to m128
11	Contrast	7	con1 to con7
12	Tonnetz	6	tn1 to tn6

*Using Malcolm Slaney algorithm (Slaney, 1993)

** Using librosa library (Mcfee et al., 2015)

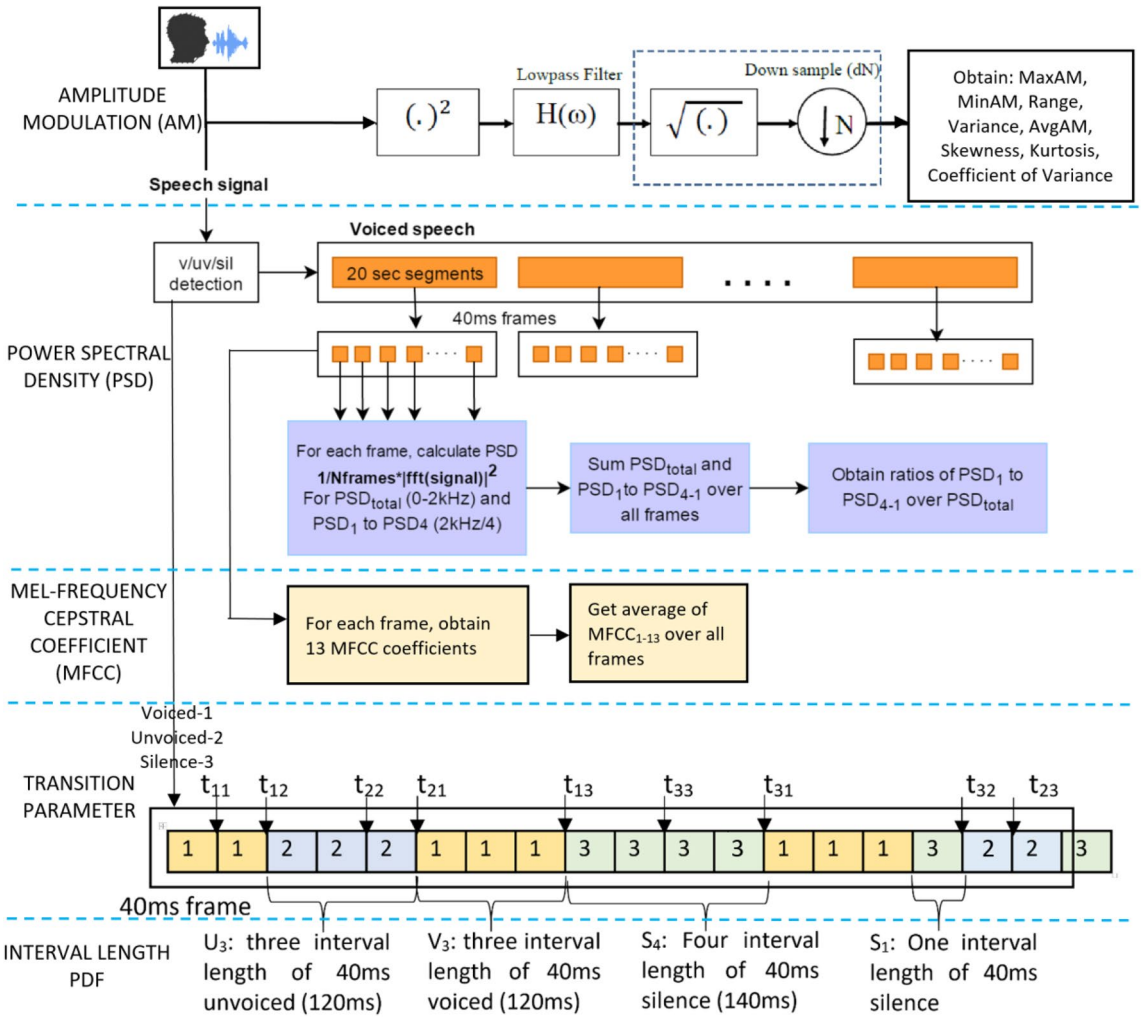


Fig. 2 Overall process of the feature extraction in MATLAB

inverse discrete cosine transform. Two methods of obtaining MFCC are used here. The first method uses Malcolm Slaney algorithm (Slaney, 1993) and the other uses the Python Librosa library (Mcfee, et al., 2015).

Chromagram or Chroma (ch) feature is recognized for its robustness to the changes in timbre and closely correlate to the musical aspect of harmony. It is also known as pitch class profiles. For feature extraction, the audio file is translated into a series of chroma features, and each sequence explains how the short-time energy of the signal is spread over the twelve chroma band (Ellis, 2007).

According to Cohn (1998), tonal centroid features or known as the Harmonic Network (Tonnetz), signifies pitch. The tonal centroid vector t_n of time frame n is the result of multiplication of the chroma vector c_n and a transformation matrix T. Then, the t_n divided by the L_1 norm of chroma vector to prevent numerical instability and ensure that the tonal centroid vector dimension is always six. The tonal centroid vector is given as:

$$t_n(d) = \frac{1}{\|c_n\|_1} \sum_{i=0}^{11} T(d,l) c_n(l) \quad \begin{matrix} 0 \le d \le 5 \\ 0 \le l \le 11 \end{matrix}$$

where d is the index of which of six dimensions is being evaluated, and l is the chroma vector pitch class index.

In the Spectral Contrast feature, each frame of a spectrogram is divided into sub-bands. For each sub-band, the energy contrast is estimated by difference between the mean energy in the top quantile (peak energy) and that of the bottom quantile (valley energy). Clear, narrow-band signals mostly have high contrast values, while broad-band noise have low contrast values (Jiang et al., 2002).

A Mel-Spectrogram is a spectrogram where the frequencies are converted to the Mel scale. The spectrogram can be found by computing the FFT on overlapping windowed segments of the signal.

Sustained vowels are commonly used in voice analysis for perturbation measures and vocal characteristics. These

features were extracted using Praat software. Features that are obtained from sustained vowels are fundamental frequency (f0), jitter, shimmer, pitch, harmonics-to-noise ratio (HNR) and formants. f0 can be used to represent the number of cycles of opening/closing of the glottis; however, the measurements vary with sex and age. Pitch measure the quality of sound and has a non-linear correlation with the f0. Jitter is a parameter that measures the frequency variation or instability from cycle to cycle whereas, shimmer relates to the amplitude variation or instability of a sound wave. The HNR represents the ratio between periodic and non-periodic components in a voiced speech. Finally, formant represents the resonant frequencies of the vocal tract and are especially prominent in vowels. In this work, four formant peak frequencies were extracted for analysis.

3 Results

3.1 Identification of robust features

The extracted feature values in this dataset have different order of magnitude and we consider these feature sets to be different from each other. Therefore, in order to capture the variability between data values, we decided to use the method of coefficient of variance (COV). The COV measures how consistent the values of each set from their respective mean of the data set. The smaller the percentage COV value, the higher is the uniformity within the values present in the data set. Even if the set of data has a low standard deviation, it does not mean the data has less variability. There are sets of data with low standard deviation but high COV. The COV can be calculated as the ratio of the standard deviation to the mean.

$$\%CV = \frac{\sigma}{\mu} \cdot 100\%$$

Figures 3 and 4 plots the COV for all features listed in Table 2, for each subject. The graphs were shown for COV up to 100 percent. However, the COV percentage can go beyond 100% if the standard deviation exceeds the mean value. This usually happens when the data set has majority values that are too small or close to zero.

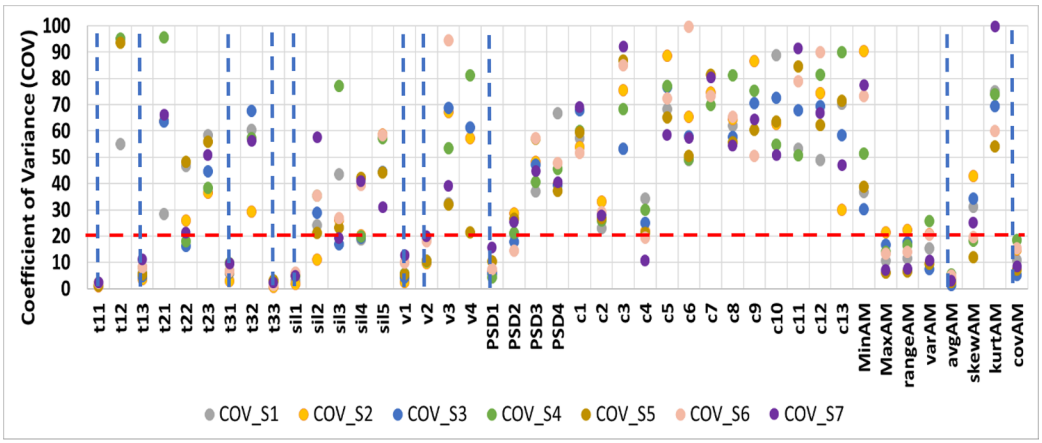
There are no standard threshold values for the percentage of COV and it commonly depends on the field of study. For this study, we consider the percent of COV of less than 10% to be significantly low variability, less than 20% to be marginally low variability and less than 30% to be acceptable. Table 3 list all the features that are within these range of COV.

3.2 Validation of robust features on depression detection

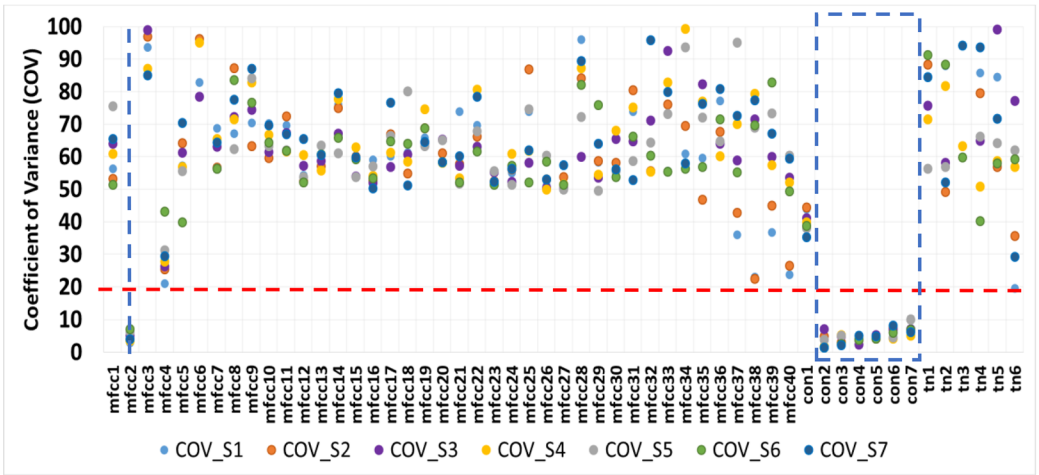
We demonstrated the use of robust features in one experiment related to depression detection. For this experiment, we gathered speech recordings of depressed and healthy subjects via an online application called WhatsApp Voice Note due to the limitation of visiting the hospitals and risk of having a face-to-face meeting with subjects on site during the pandemic. All procedures performed in studies involving human participants were in accordance with the ethical standards and has been approved by the IIUM Research Ethical Committee (IREC 2019-006). The database was divided based on gender and diagnostic groups of depressed and healthy. Subjects consisted of 43 depressed and 47 healthy were required to sign an informed consent. Subjects were then asked to read a standardized Bahasa Malaysia passage called Cerita Datuk that is commonly used by speech therapists and to fill in the Malay Beck Depression Inventory-II (Malay BDI-II) and Patient Health Questionnaire-9 (PHQ-9) for ground truth reference.

The recordings received in the Voice Note were in.OGG and.MP4 format. Speech files were converted to.WAV using the audio.online-convert.com at a sampling rate of 44.1 kHz and 32 bits per sample, with a mono channel. The recordings were normalized prior to the acoustic parameter extraction. Each audio signal was then divided into 20 s segments and acoustic features were extracted for each segment. The total number of 20 s segments for female-controlled speech and male-controlled speech are 198 and 73, respectively. We proceed with the classification analysis on the robust features obtained using MATLAB as shown in Table 3, which are the time-based features and the Power Spectral features from the spectrum-based category. We performed Exhaustive Feature Selection (EFS) on four selected classifiers which are the Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Random Forest (RF) and Extreme Gradient Boosting (XGBoost). We present the classifier and feature set with the best performance for female and male speech depression detection in Table 4.

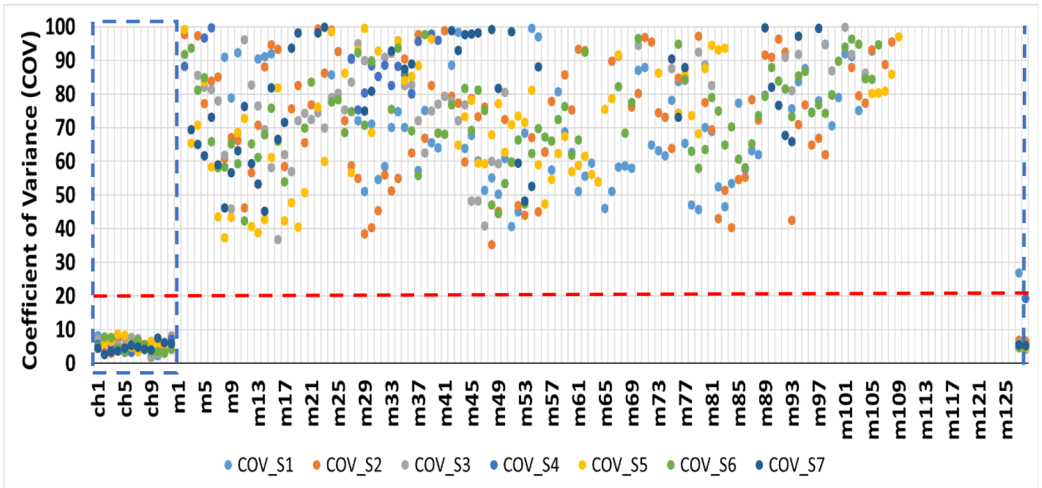
For the male speech, the diagnostic accuracy and Area Under the Curve (AUC) score of the XGBoost classifier shows a good performance of 86% and 79%, respectively. The precision and recall values are also in the high range (above 80%) which shows a balance high true positive and high true negative rate. For the female speech, KNN classifier was able to achieve an acceptable performance of AUC score and accuracy with a percentage of an approximately in the low range of 70%. KNN produced a consistent value of precision and recall, captured by the f1 score of 70%. F1 score conveys the balance between the value of precision and recall.


Coefficient of Variance (COV)
t11 t12 t13 t21 t22 t23 t31 t32 t33 sil1 sil2 sil3 sil4 sil5 v1 v2 v3 v4 PSD1 PSD2 PSD3 PSD4 c1 c2 c3 c4 c5 c6 c7 c8 c9 c10 c11 c12 c13 MinAM MaxAM rangeAM varAM avgAM skewAM kurtAM covAM
COV_S1 COV_S2 COV_S3 COV_S4 COV_S5 COV_S6 COV_S7

(a)


Coefficient of Variance (COV)
mfcc1 mfcc2 mfcc3 mfcc4 mfcc5 mfcc6 mfcc7 mfcc8 mfcc9 mfcc10 mfcc11 mfcc12 mfcc13 mfcc14 mfcc15 mfcc16 mfcc17 mfcc18 mfcc19 mfcc20 mfcc21 mfcc22 mfcc23 mfcc24 mfcc25 mfcc26 mfcc27 mfcc28 mfcc29 mfcc30 mfcc31 mfcc32 mfcc33 mfcc34 mfcc35 mfcc36 mfcc37 mfcc38 mfcc39 mfcc40 con1 con2 con3 con4 con5 con6 con7 tn1 tn2 tn3 tn4 tn5 tn6
COV_S1 COV_S2 COV_S3 COV_S4 COV_S5 COV_S6 COV_S7

(b)


Coefficient of Variance (COV)
ch1 ch5 ch9 m1 m5 m9 m13 m17 m21 m25 m29 m33 m37 m41 m45 m49 m53 m57 m61 m65 m69 m73 m77 m81 m85 m89 m93 m97 m101 m105 m109 m113 m117 m121 m125
COV_S1 COV_S2 COV_S3 COV_S4 COV_S5 COV_S6 COV_S7

(c)

◂**Fig. 3** Plot of Coefficient of Variances (COV) for seven mobile phone recordings for each seven subjects on the utterance speech. The features are, **a** transition matrix, silence, voiced, power spectral density, 13-MFCC coefficients and amplitude modulation statistics. **b** 40-Mel-cepstral coefficients, contrast and tonnetz. **c** Chroma and mel-spectrogram

4 Discussion

For the TP feature, transition probabilities with unvoiced frames have significantly higher COV due to the fact that the utterance of rainbow passage is mostly made up of voiced and silence segments. Thus, the transition probability that has unvoiced frames, are mostly zeros. This includes t12, t21, t22, t23 and t32. Referring to Fig. 5a–g, the variability in the time-domain signals of the waveforms is greatly visible in the form of noise interference. However, the horizontal axis of the time-domain signals and the envelope of the signals are not significantly different from each other, especially after filtering the high frequency noise interference. Lower order silence (sil1) and voiced (v1 and v2) segments are more certain due to the fact that their consecutive 40 ms frames are wider and are considerably detectible. Higher order silence and voiced segments are too small and prone to be mislabelled within the respective bands.

In musical analysis, chroma feature has been used in the application of music synchronization and is known to be robust to variation in instrumentation, timbre and dynamics (Müller et al., 2009). The steps of obtaining chroma are similar to MFCC. Müller et al. reported that the lower MFCC coefficients are related to the variation of timbre. However, after applying the discrete cosine transform (DCT), the variation to timbre was removed by discarding all the lower Mel-cepstral coefficients and keeping only the upper coefficients. The resulting vector was then transformed using inverse-DCT and projected onto the 12 chroma bins.

Another spectrum-based feature that is robust towards multiple mobile recording devices is the spectral contrast. Although this feature is considered to be in the spectrum-based, the feature values are obtained by calculating the vertical difference between peaks and valleys that are measured in octave-scale filter sub-bands. Thus, the amplitude difference of the spectral might be robust towards multiple microphone frequency responses.

The components in MFCC are derived from the DCT coefficients that represents the uneven spectral shape. Table 3 lists the second Mel-cepstral coefficient (mfcc2) as one of the robust features. This feature estimates the broad shape of the spectrum and is commonly associated with the spectral centroid. The higher order of MFCC are used to represent the shape of the spectrum and capture the pitch and tone information. Although, MFCC is a widely used feature in the field of music, sound, and speech analysis, this feature is not robust towards different recording devices and might also be more prone to get affected by the noise environment. A study by Pan and Waibel (2000) demonstrated the influence of background noise in MFCC with the signal-to-noise ratio affecting the frequency bands differently. This is due to the noise spectrum produced by the microphone proximity where the noise mean spectrum of distant microphone is much higher than the close one.

The power spectral density (PSD) also demonstrates the characteristic of robustness, although not in the most significant category of COV. In this work, four equal PSD bands of 500 Hz were extracted. Commonly in voice signals, more than 90% of voice energy is in the first two bands (PSD1 and PSD2) with the ratio of PSD1 higher than PSD2. Therefore, we assume that less variability will occur due to the bulk of energy in these bands.

On the sustained vowel category, mean fundamental frequency (f0) and formants have shown significant precision in the feature values with COV percentage of less than 10. The sustained vowel was used instead of speech utterance in order to avoid performing the voiced and unvoiced frames characterization and minimizing the range variability of f0 values. We extracted the features using PRAAT software. For the f0 estimation algorithm, PRAAT uses the time-domain approach that relies on autocorrelation (Boersma, 1993). For formants, the speaker's vocal tract resonance frequencies are estimated using the method of Linear Prediction. In PRAAT, formant tracks appear as red lines overlaid the spectrogram of the selected signal. Meanwhile, the HNR feature or also known as the degree of periodicity, was calculated based on the relative heights of the maximum normalized autocorrelation (Boersma, 1993). We can conclude that due to the timing aspects of these algorithms, these features were shown to be more robust towards multiple mobile recording devices.

In this work, the selected robust features were then used in a classification analysis. The classifiers were able to identify 86% of male depressed speech and 70% of female depressed speech using combination of time-based and spectrum-based features. We have shown that although the speech recordings were collected using multiple mobile devices, we were able to proceed with performing the classification and the results are validated due to the removal of features variability.

5 Conclusion

For applications that require input signals from multiple devices or microphones especially in a mobile application, it is important to recognize whether different microphones alter the acoustic measurements of voice signals. Researchers also have to be aware that the instrumentations used for collecting acoustic signals and algorithms

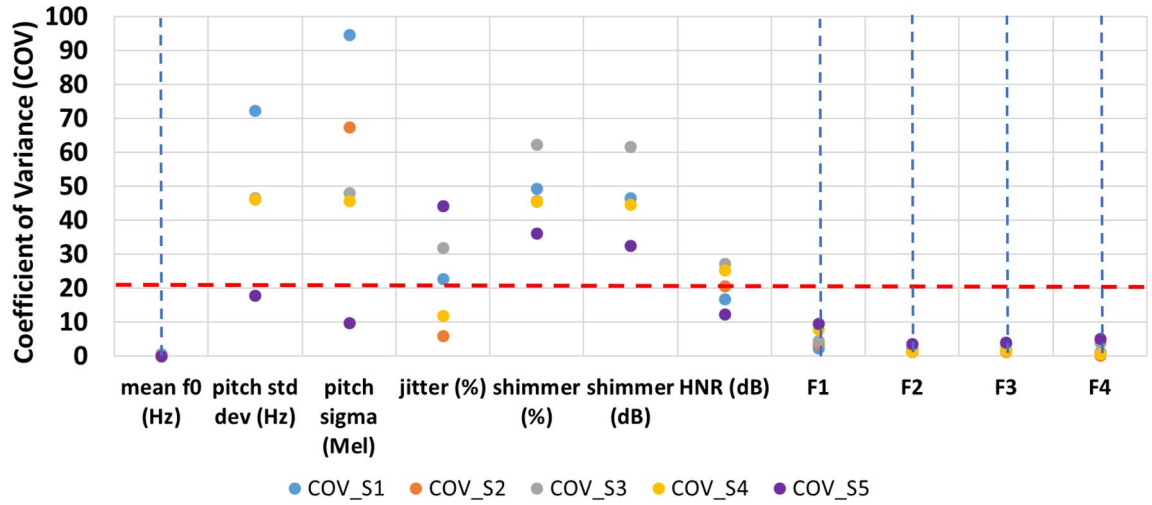


Fig. 4 Plot of coefficient of variances (COV) for seven mobile phone recordings for each of the seven subjects on the sustained vowel /a/. List of features are mean fundamental frequency, pitch standard deviation, pitch sigma, jitter, shimmer and harmonics-to-noise ratio (HNR), Formants (F1–F4)

Table 3 List of features that are less than 10% COV, 20% COV (*) and 30% COV (**)

Feature type	List of features
Time-based	t11, t31, t33, sil1, avgAM, covAM, t13*, v1*, v2* varAM**, rangeAM**, maxAM**
Spectrum-based	mfcc2, con2, con3, con4, con5, con6, con7, ch1, ch2, ch3, ch4, ch5, ch6, ch7, ch8, ch9, ch10, ch11, ch12, m128 PSD1* PSD2**
Sustained vowels	Mean f0, F1, F2, F3, F4 HNR**

*less than 20% COV

**less than 30% COV

Table 4 Best feature set and classifier model for female and male speech

Speech signal type	Classifier	Features	Mean AUC score (cv = 10)	Accuracy score	Classification report			
					Class/avg	Precision	Recall	f1
Male speech	XGBoost	t33, sil1, psd1	0.7903	0.8636	Healthy	0.66	0.75	0.70
					Depressed	0.75	0.66	0.70
					Macro avg	0.70	0.70	0.70
					Weighted avg	0.71	0.70	0.70
Female speech	KNN	13, t33	0.7548	0.7000	Healthy	0.91	0.83	0.87
					Depressed	0.82	0.90	0.86
					Macro avg	0.86	0.87	0.86
					Weighted avg	0.87	0.86	0.86

used for the analysis, are not standardized. Especially when dealing with clinical research, value objective measures that are generated from the experiments are mostly of concern. In this study, we identified acoustic features that are robust towards multiple mobile recording devices. Acoustic features extracted using the time-domain information are less prone to suffer from feature variability due to different microphone specifications. However, we would consider this as a preliminary study and requires further numerical justification on a larger dataset in order

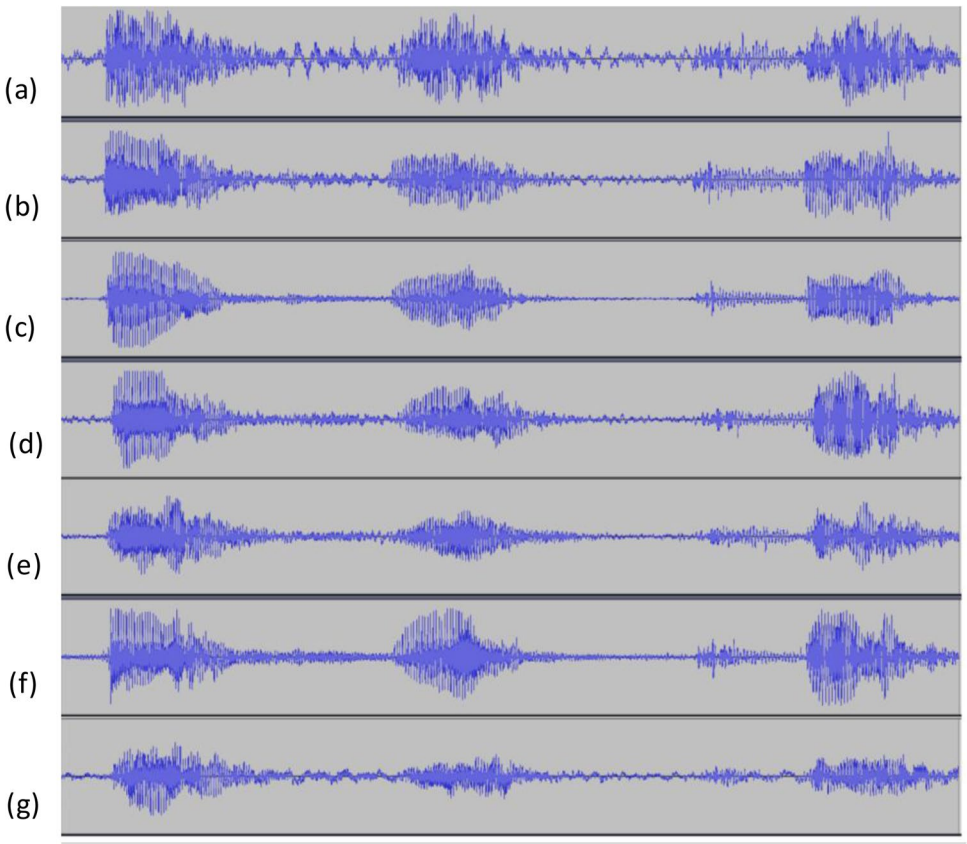


Fig. 5 Seven time-domain signal waveforms for the utterance of /path/, /high/, /above/ in the rainbow passage recorded from one subject, synchronously using seven mobile devices, **a** device A1, **b** device A2, **c** device A3, **d** device A4, **e** device A5, **f** device A6 and **g** device A7

to conclude the accuracy and repeatability of the robust features.

Acknowledgements This work was supported by funding from the Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme (FRGS19-051-0659).

Author contribution NNWNH and MDW conceived of the presented idea during research discussion. NNWNH and MAEAE carried out the experiments. NNWNH wrote the manuscript with support from MDW and MAEAE. MDW also verified the analytical method. All authors discussed the results and contributed to the final manuscript.

Funding This work was supported by funding from the Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme (FRGS19-051-0659). The funding is used to pay the graduate research assistance's stipend.

Data availability The datasets generated during and/or analysed during the current study are available in the Mendeley Data repository, http://dx.doi.org/10.17632/8kw826c2x7.1 (Nik Hashim, 2020).

Declarations

Competing interest The authors declare that they have no competing interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards and has been approved by the IIUM Research Ethical Committee (IREC 2019-006). Subject Informed Consent (SIC) was also obtained from all individual participants included in the study.

References

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound.

Bottalico, P., et al. (2018). Reproducibility of voice parameters: The effect of room acoustics and microphones. *Journal of Voice*. https://doi.org/10.1016/j.jvoice.2018.10.016

Clark, W. W., & Saunders, S. (2016). Assessment of noise exposures for pre-term infants during air transport to neonatal intensive care units using iPhone sound meter apps. *Journal of the Acoustical Society of America*. https://doi.org/10.1121/1.4950019

Cohn, R. (1998). Introduction to Neo-Riemannian to theory: A survey and a historical perspective. *Journal of Music Theory, 42*(2), 167–180.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 28*(4), 357–366.

Deliyski, D. D., Shaw, H. S., Evans, M. K., & Vesselinov, R. (2006). Regression tree approach to studying factors influencing acoustic voice analysis. *Folia Phoniatrica Et Logopedica*. https://doi.org/10.1159/000093184

Dickerson, M. (2016). Investigating the feasibility of using mobile devices for remote noise monitoring and data acquisition. *Journal of the Acoustical Society of America*. https://doi.org/10.1121/1.4950018

Ellis, D. P. W. (2007). Classifying music audio with timbral and chroma features.

Faber, B. M. (2017). Acoustical measurements with smartphones : Possibilities and limitations. *Acoutics Today*.

Ghosal, D., & Kolekar, M. H. (2018). Music genre recognition using deep neural networks and transfer learning. In *Proc. annu. conf. int. speech commun. assoc. INTERSPEECH*, vol. 2018-Septe, no. September, pp. 2087–2091. https://doi.org/10.21437/Interspeech.2018-2045.

Hashim, N. W., Wilkes, M., Salomon, R., Meggs, J., & France, D. J. (2017). Evaluation of voice acoustics as predictors of clinical depression scores. *Journal of Voice*. https://doi.org/10.1016/j.jvoice.2016.06.006

Jiang, D.-N., Lu, L., Zhang, H.-J., Tao, J.-H., & Cai, L.-H. (2002). Music type classification by spectral contrast feature. In *IEEE Int. Conf. Multimed. Expo,* Vol. 1, 113–116.

Karnell, M. P., Scherer, R. S., & Fischer, L. B. (1991). Comparison of acoustic voice perturbation measures among three independent voice laboratories. *Journal of Speech and Hearing Research*. https://doi.org/10.1044/jshr.3404.781

Kisenwether, J. S., & Sataloff, R. T. (2015). The effect of microphone type on acoustical measures of synthesized vowels. *Journal of Voice*. https://doi.org/10.1016/j.jvoice.2014.11.006

Krik, V. M., Ribeiro, V. V., Siqueira, L. T. D., Rosa, M. D. O., & Leite, A. P. D. (2019). Análise acústica da voz: comparação entre dois tipos de microfones. *Audiology Communication Research*. https://doi.org/10.1590/2317-6431-2018-2113

Mcfee, B., *et al*. (2015). Librosa—audio processing Python library. In *Proc. 14th python sci. conf.*

Müller, M., Ewert, S., & Kreuzer, S. (2009). Making chroma features more robust to timbre changes. https://doi.org/10.1109/ICASSP.2009.4959974.

Pan, Y., & Waibel, A. (2000). The effects of room acoustics on MFCC speech parameter.

Parsa, V., Jamieson, D. G., & Pretty, B. R. (2001). Effects of microphone type on acoustic measures of voice. *Journal of Voice*. https://doi.org/10.1016/S0892-1997(01)00035-2

Sinha, S., et al. (2016). Real-time sound measurements of exercise classes with mobile app demonstrate excessive noise exposure. *Journal of the Acoustical Society of America*. https://doi.org/10.1121/1.4950021

Slaney, M. (1993). Auditory toolbox. *Apple Comput. Co. Apple Tech. Rep*.

Styler, W. (2013). Using Praat for linguistic research. *Savevowels*.

Su, Y., Zhang, K., Wang, J., Zhou, D., & Madani, K. (2020). Performance analysis of multiple aggregated acoustic features for environment sound classification. *Applied Acoustics*. https://doi.org/10.1016/j.apacoust.2019.107050

Švec, J. G., & Granqvist, S. (2010). Guidelines for selecting microphones for human voice production research. *American Journal of Speech-Language Pathology*. https://doi.org/10.1044/1058-0360(2010/09-0091)

Titze, R., & Winholtz, W. S. (1993). Effect of microphone type and placement on voice perturbation measurements. *Journal of Speech and Hearing Research*. https://doi.org/10.1044/jshr.3606.1177