CONTRIBUTED PAPER

Extraction and Utilization of Excitation Information of Speech: A Review

This article reviews signal processing approaches used for extraction of excitation information from speech.

By Sudarsana Reddy Kadiri, *Member IEEE*, Paavo Alku, *Fellow IEEE*, and B. Yegnanarayana, *Life Fellow IEEE*

ABSTRACT | Speech production can be regarded as a process where a time-varying vocal tract system (filter) is excited by a time-varying excitation. In addition to its linguistic message, the speech signal also carries information about, for example, the gender and age of the speaker. Moreover, the speech signal includes acoustical cues about several speaker traits, such as the emotional state and the state of health of the speaker. In order to understand the production of these acoustical cues by the human speech production mechanism and utilize this information in speech technology, it is necessary to extract features describing both the excitation and the filter of the human speech production mechanism. While the methods to estimate and parameterize the vocal tract system are well established, the excitation appears less studied. This article provides a review of signal processing approaches used for the extraction of excitation information from speech. This article highlights the importance of excitation information in the analysis and classification of phonation type and vocal emotions, in the analysis of nonverbal laughter sounds, and in studying pathological voices. Furthermore, recent developments of deep learning techniques in the context of extraction and utilization of the excitation information are discussed.

KEYWORDS | Deep learning; emotional speech; glottal closure instant (GCI); glottal inverse filtering (GIF); glottal opening instant (GOI); nonverbal sounds; pathological voices; phonation type; speech analysis; speech excitation.

NOMENCLATURE

CCD	Complex cepstrum decomposition.
CP	Closed phase.
CPP	Cepstral peak prominence.
ClQ	Closing quotient.
CNN	Convolutional neural network.
dEGG	Derivative of the EGG signal.
DNN	Deep neural network.
DRF	Dominant resonance frequency.
DYPSA	Dynamic programming phase slope algorithm.
EGG	Electroglottography.
EoE	Energy of excitation.
FFNN	Feedforward neural network.
F_0	Fundamental frequency.
GCI	Glottal closure instant.
GOI	Glottal opening instant.
GIF	Glottal inverse filtering.
GMM	Gaussian mixture model.
GNE	Glottal-to-noise excitation.
H1-H2	Amplitude difference between the first and second harmonics.
HNR	Harmonic-to-noise ratio.
HRF	Harmonic richness factor.
HSV	High-speed video endoscopy.
IAIF	Iterative adaptive inverse filtering.
ILPR	Integrated linear prediction residual.
LFSD	Low-frequency spectral density.
LoMA	Lines of maximum amplitude.

Manuscript received August 8, 2021; accepted November 1, 2021. Date of publication November 30, 2021; date of current version December 8, 2021. This work was supported in part by the Academy of Finland under Project 330139. *(Corresponding author: Sudarsana Reddy Kadiri.)*

Sudarsana Reddy Kadiri and **Paavo Alku** are with the Department of Signal Processing and Acoustics, Aalto University, FI-00076 Espoo, Finland (e-mail: sudarsana.kadiri@aalto.fi; paavo.alku@aalto.fi).

B. Yegnanarayana is with the Speech Processing Laboratory, International Institute of Information Technology, Hyderabad (IIIT-H), Hyderabad 500032, India (e-mail: yegna@iiit.ac.in).

Digital Object Identifier 10.1109/JPROC.2021.3126493

LP	Linear prediction.
LPCCs	Linear prediction cepstral coefficients.
LSTM	Long short-term memory.
MBSC	Multiband summary correlogram.
MDQ	Maximum dispersion quotient.
MFCCs	Mel-frequency cepstral coefficients.
MLP	Multilayer perceptron.
NAQ	Normalized amplitude quotient.
OQ	Open quotient.
PLP	Perceptual linear prediction.
PS	Peak slope.
PSP	Parabolic spectral parameter.
QCP	Quasi-closed phase.
QOQ	Quasi-open quotient.
QPR	Quadratic programming.
RAPT	Robust algorithm for pitch tracking.
RNN	Recurrent neural network.
SEDREAMS	Speech event detection using the residual excitation and a mean-based signal.
SFF	Single frequency filtering.
SHRP	Subharmonics-to-harmonics ratio.
SIFT	Simplified inverse filter tracking.
SNR	Signal-to-noise ratio.
SoE	Strength of excitation.
SRH	Summation of residual harmonics.
SQ	Speed quotient.
STFT	Short-time Fourier transform.
SVM	Support vector machine.
SWIPE	Sawtooth waveform inspired pitch estimator.
YAGA	Yet another GCI algorithm.
YAAPT	Yet another algorithm for pitch tracking.
ZFF	Zero-frequency filtering.
ZTW	Zero-time windowing.
ZZT	Zeros of the z transform.

I. INTRODUCTION

Speech is the most sophisticated means of communication among people. The carrier of speech is the acoustic speech pressure signal. In the speech, a small number of basic elements, such as phones or syllables, are combined to form a large number of units, such as words and phrases. The complexity of speech is due to the many-to-one relationship between the speech sound and its perceived counterpart in the way that several phonetic contrasts can be produced by the same acoustic cue. Conversely, several acoustic cues may indicate the same phonetic contrast. In addition, the phonemic cues in conversational speech are enriched by characteristics, such as vocal emotions. Thus, the information conveyed through the speech signal is related not only to *what* is said but also *how* the spoken message is conveyed. While the former is useful in situations, such as information announcements, the latter is important in casual conversations.

Speech is produced by the physiological apparatus of the human speech production system. The function of this system can be divided into two main parts: *excitation*, the major component of it is generated at the larynx, and *filtering*, which refers to the effects of the dynamic articulators on the excitation during speech production. The characteristics of the excitation vary depending on the speech sound to be produced. For the most prevalent category of speech sounds in most languages, voiced sounds (such as the vowel [a] and the nasal [n]), the excitation is the air flow waveform generated by the vibration of the vocal folds. This excitation is called the glottal flow due to the air passing through the orifice between the two vibrating vocal folds at the glottis (see Fig. 1). The filtering process extends from the vocal folds to the lips and nostrils. It is influenced by the positioning of the tongue, the degree of opening of the mouth, and the movement of the lips. By varying the acoustical properties of the excitation, humans are capable of changing some essential cues of speech, such as pitch (e.g., generating low or high voices) and voice quality (e.g., coloring speech to sound breathy or pressed). By changing the articulators, humans can produce sounds (called phones) representing different phonemes (e.g., /a/ or /i/). Among the three (voiced, unvoiced, and plosive) categories of speech sounds, voiced sounds are of special interest in speech science [1], [2].

There are many situations where the *decomposition* of the speech signal into the excitation and filter components is needed. The source-filter decomposition helps to model the two components effectively in several speech technology applications, such as speech synthesis [3], [4], enhancement [5], [6], and coding [7]–[10]. Decomposition of speech signal helps to improve our understanding of the human speech production mechanism. Studies have shown that understanding the excitation component helps in generating acoustical cues of different voice qualities [11]–[13] and vocal emotions [14]–[20], as well as in the production of different paralinguistic and nonverbal sounds [21]–[23]. The excitation information is also useful for providing complementary information to the more widely used vocal tract spectral features to improve, for example, the detection of speech disorders [24]–[27]. The relatively less effort in the study of the excitation component is due to difficulty in the decomposition of the signal into the excitation component even though its importance is well established in many areas of speech science and technology [28]–[30]. The decomposition is difficult, for example, in expressive speech because of large variations in the characteristics of speech sounds. In addition, the nonstationarity of the speech production process compounds the difficulty. The speech production process also involves nonlinearities [31]–[33] that cannot be handled using linear source-filter models.

This article is a review of the methods to extract and utilize the excitation component of mainly voiced speech. An important and widely studied example of such a feature is the inverse of the glottal cycle duration, that is, the instantaneous *fundamental frequency* (F_0). In addition

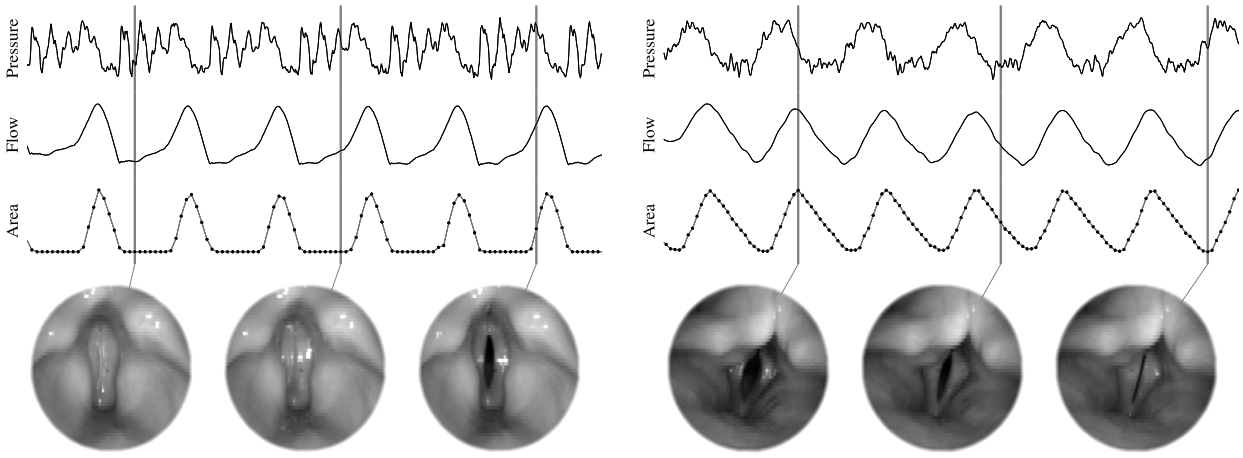


Fig. 1. *Demonstration of the production of voiced speech in two phonation types: normal (left) and breathy (right). The upper part of the figure shows three time-domain waveforms (the speech pressure signal, the glottal flow estimated by GIF, and the glottal area function) and the lower part shows images of the vocal folds. The gray vertical lines show the instants when the images of the vocal folds were taken by the transoral high-speed digital videoendoscopy system (adopted from [38]).*

to F_0, one of the most important features is the strong *impulse-like component* that is present in each cycle of the glottal flow waveform in the production of voiced speech. This impulse-like component is caused by the sudden deceleration of the air flow in the vicinity of the GCI due to adduction of the vocal folds. The characteristics of this impulse-like event in the excitation waveform, particularly its sharpness, are closely associated with several important speech attributes, such as voice quality [11], [34] and loudness [35], [36]. The excitation information also includes identifying the locations of GOIs, where secondary excitation of the vocal tract might take place. Furthermore, excitation information also involves estimating the *entire excitation waveform* from the microphone speech signal and then expressing this time-domain signal (or its spectrum) with a few parameters to quantify, for example, time ratios between the opening and closing phases of the glottal flow waveform.

The extraction of speech excitation information has been previously addressed in three review articles, which are more than five years old. In [30] and [37], the review focused on GIF and its applications. The review in [29] studied the GCI detection methods, F_0 extraction, and GIF methods, as well as applications for speech synthesis, speaker recognition, expressive speech processing, and biomedical applications. In this article, a more holistic review of the recent advances in extraction and utilization of excitation information is provided by studying different types of GIF methods, F_0 extraction methods, and GCI and GOI extraction methods. Furthermore, this review highlights the extraction and utilization of the excitation information based on recent developments in deep learning, an issue that is absent from all the previous reviews. In addition, the utilization of excitation information in speech-based biomarking of human health, an area that has become increasingly important in recent years, is discussed especially from the point of view of the detection of neurodegenerative diseases.

The organization of this article is given as follows. Section II describes the generation of speech signals by the physiological human speech production mechanism. Section III briefly describes the extraction of excitation information using nonacoustic techniques. In Section IV, the extraction of excitation information is studied by describing the estimation of glottal flow using GIF and describing the most important features of the excitation, namely, F_0, GCI, and GOI, and the issues underlying their extraction. Section V describes how excitation information has been used in four specific areas of speech research by addressing the study of phonation types, vocal emotions, laughter sounds, and pathological voices. Section VI describes recent trends in this area by discussing the use of deep learning in GIF and the extraction of F_0 and GCI, as well as the utilization of excitation information in the detection of neurological diseases. Finally, conclusions are given and future directions are discussed in Section VII.

The topics addressed in this article are thematically described in Table 1 (ranging from the speech production mechanism to the utilization of excitation information). The list of abbreviations used in this article is given in Nomenclature.

II. HUMAN SPEECH PRODUCTION MECHANISM

The speech production mechanism allows humans to produce a vast range of sounds ranging from verbal sounds (normal speech) to nonverbal sounds (laughter, cry, and so on) and sounds of different voice qualities and emotions. Understanding the physiological speech production mechanism helps in the analysis of speech signals. A simplified schematic presentation of the speech production system is shown in Fig. 2. The system consists of many organs, which

Table 1 Topics Addressed in This Article

- Speech excitation and vocal tract response components (Sections I and II)
- Speech excitation features
  - physiological movements of articulators and larynx (Sections II and III)
  - glottal flow waveform (Section IV(A))
  - impulse-like component, periodicity, subharmonicity, periodic and aperiodic components, harmonic-to-noise ratio (Section IV(A)-IV(D))
  - instantaneous fundamental frequency (F_0) (Section IV(B))
  - glottal closure instant (GCI), glottal opening instant (GOI) (Sections IV(C) and IV(D))
- Tools and methods to study the excitation component
  - high-speed video endoscopy (HSV) - glottal area measurements (Section I and III)
  - electroglottography (EGG) - vocal fold contact measurements (Section III)
  - glottal inverse filtering (GIF) - to estimate the (entire) glottal flow (Section IV(A))
  - direct extraction of the excitation component from speech (Sections IV(B)-IV(D))
  - DNN-based methods for the estimation of glottal source and excitation features (Section VI(A))
- Utilization of excitation information
  - study of phonation type (Section V(A))
  - study of vocal emotions (Section V(B))
  - study of laughter sounds (Section V(C))
  - study of pathological voices (Section V(D) and VI(B))
- Conclusions (Section VII)

can be categorized into three main groups: the lungs (the subglottal system), larynx, and vocal tract (the supraglottal system) [39], [40].

The lungs serve as the source of energy for the speech production process generating pressure in the larynx due to airflow. The (mean) lung pressure, also called subglottal pressure, is controlled by the speakers for producing sounds of different vocal intensity levels and phonation types. It has been found that lung pressure can rise to values as high as 6 kPa (60 cm of H_2O) in loud singing voices [41]. However, in the production of speech signals, lung pressure is typically much lower. For example, the measured lung pressure values in [42] were below 1 kPa (10 cm of H_2O) for most vowels of soft or normal loudness and lung pressure rose to larger values (around 4 kPa, that is, 40 cm of H_2O) only in loud and very loud speech signals.

During the production of voiced speech, the pressure from the lungs causes vibration of the vocal folds. The vocal folds (see Fig. 1), located in the larynx, are the key physiological organs in the production of voiced speech. The vocal folds have a layered structure consisting of five layers (for more details, see [43]). The vibration of the vocal folds forms the acoustical excitation signal for voiced speech, called glottal volume velocity waveform or simply glottal flow. Fig. 1 demonstrates the production of voiced speech in two phonation types: normal (left) and breathy (right). The figure shows the speech signal, the glottal flow estimated by GIF, the glottal area function, and the images of the vocal folds. The gray vertical lines show the instants when the images of the vocal folds were taken by the transoral high-speed digital videoendoscopy system.

There are two other types of excitation during speech production resulting in unvoiced and plosive sounds. The unvoiced sounds (e.g., [s] and [f]) are generated by forming a constriction at some point along the vocal tract and forcing air through this constriction to generate turbulence. Plosive sounds are generated by abruptly releasing the air pressure building behind closure along the vocal tract. These sounds are also called stops. Plosives can be both unvoiced (e.g., [k] and [t]) and voiced (e.g., [g] and [d]). The vocal tract system (consisting of the oral, nasal, and pharyngeal resonant cavities) shapes the excitation signal, and the resulting air flow signal is radiated at the lips to form the speech pressure signal.

The production of speech can be considered as exciting a filter (the vocal tract system) by an excitation. This is called the source-filter model of speech production. In the source-filter model, the source and filter are assumed to be independent. It is worth emphasizing that, even though the assumed independence of the source and filter enables using more straightforward technologies, for example, in speech analysis and synthesis, there is coupling between the source and tract in the production of speech as reported in many studies (e.g., [31]–[33]). In summary, according to the way the excitation signal of the human speech production system is generated, the produced speech signals can be roughly divided into the following three categories:

1) voiced sounds (excited by the quasi-periodic glottal flow);
2) unvoiced sounds (excited by aperiodic noise-type flow);
3) plosive sounds (excited by burst-type flow).

Production of these three broad categories of sounds is shown in the simplified diagram in Fig. 2. The current review focuses on the excitation information in the voiced sounds.

The source-filter model of speech production [shown schematically in Fig. 3(a)] can be expressed mathematically in the time domain as follows:

$$s[n] = e[n] * v[n] \tag{1}$$

where $s[n]$ is the speech signal, $e[n]$ is the excitation (i.e., the derivative of the glottal flow waveform), $v[n]$ is the impulse response of the vocal tract (filter), and $*$ denotes convolution operation. In the z-domain, the corresponding

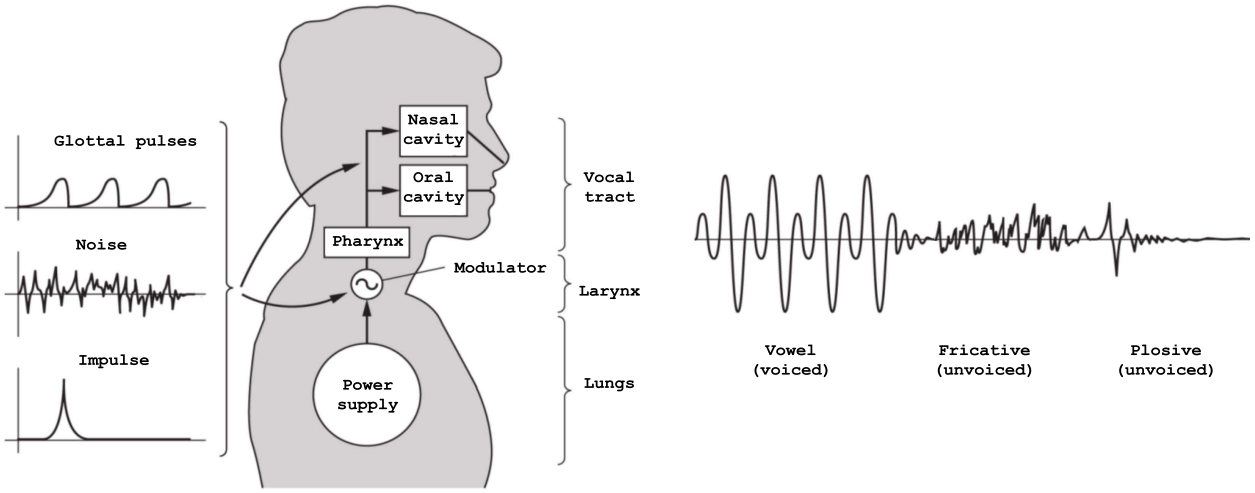


Fig. 2. *Schematic presentation of the human speech production mechanism (adopted from [39]). Left: three excitation waveforms. Right: corresponding speech waveform.*

equation is given as follows:

$$S(z) = E(z)V(z) \tag{2}$$

where $S(z), E(z)$, and $V(z)$ correspond to the z transforms of the speech signal, source, and filter, respectively.

Given the speech signal ($s[n]$), the excitation can be obtained as follows:

$$E(z) = \frac{1}{V(z)} S(z). \tag{3}$$

Equation (3) shows that the excitation can be computed by canceling the effect of the vocal tract filter ($1/V(z)$) from the speech signal [see Fig. 3(b)]. This forms the basis for the GIF method for the extraction of excitation information (discussed in Section IV-A1). The objective of this review is to discuss signal processing approaches to extract information in the excitation signal $e[n]$, given the speech signal $s[n]$.

III. EXTRACTION OF EXCITATION INFORMATION USING NONACOUSTICAL TECHNIQUES

In this section, three nonacoustical techniques to extract excitation information (EGG, HSV, and videokymography) are briefly discussed. These techniques have been used to obtain the *ground truth* for the excitation features, such as F_0, GCI, and GOI. The ground truth is useful for the evaluation of the methods developed for extracting these features from speech.

EGG is an electrical method to study voice production by feeding high-frequency-modulated current through two electrodes placed on either side of the glottis [45]. The electrical impedance between the electrodes decreases as the vocal folds adduct, and the impedance increases when the vocal folds abduct. Hence, the EGG signal provides information about the area of contact between the vocal folds during the production of voiced speech. As a method to compute the ground truth, EGG benefits from being a low-cost approach and can be applied not only for isolated sounds but also for continuous speech.

Fig. 4 shows the EGG signal in one glottal cycle. The EGG signal consists of four distinct phases [46]: *the closing phase, the CP, the opening phase*, and *the open phase*. In the closing phase (between t_1 and t_3), the vocal folds first start contacting at the lower margins (between t_1 and t_2) and then moving the contact to the upper margins (between t_2 and t_3). Generally, the closing of the vocal folds is faster than the opening, and the instant of the maximum slope occurs at t_2, which can be seen as a prominent negative peak in the dEGG shown in Fig. 4(b). The vocal folds are in full contact during the CP (between t_3 and t_4), blocking the passage of air through the glottis. In the opening phase (between t_4 and t_6), the lower margins of the vocal folds begin to separate slowly from each other (between t_4 and t_5), followed by separation along the upper margins of the vocal folds (between t_5 and t_6). The instant of the

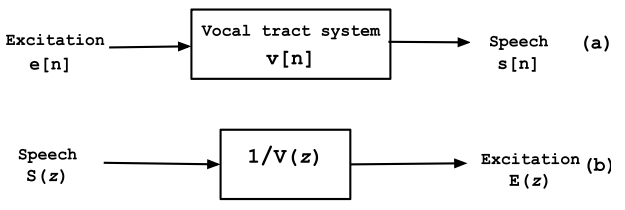


Fig. 3. *(a) Source-filter model of speech production. (b) Extraction of excitation using inverse filtering.*

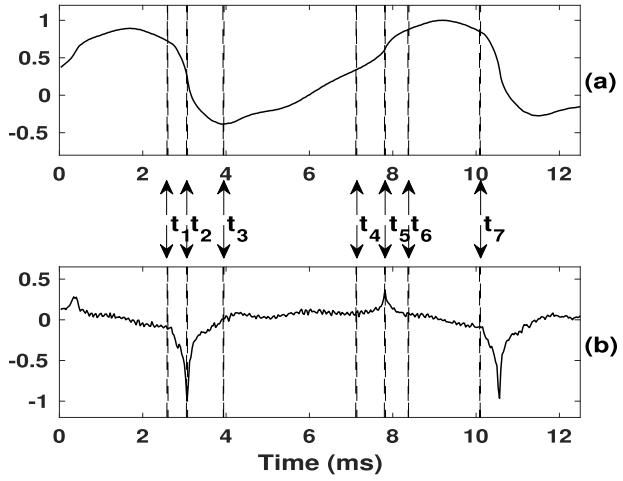


Fig. 4. *Segment of (a) EGG signal and (b) corresponding dEGG signal. Four parts of the glottal cycle are defined as follows: the closing phase (from t_1 and t_3), the CP (from t_3 and t_4), the opening phase (from t_4 to t_6), the open phase (from t_6 and t_7), and the pitch period (from t_1 to t_7) [44].*

maximum slope occurs at t_5, which can be seen as the positive peak in the dEGG signal [see Fig. 4(b)]. The vocal folds are apart during the open phase (between t_6 and t_7).

The locations of the peaks in the dEGG signal, i.e., the negative peak at t_2 and the positive peak at t_5, are considered to be GCI and GOI, respectively. F_0 is estimated as the inverse of the time difference between two consecutive GCIs. The values of F_0, GCI, and GOI extracted from dEGG are used as the ground truth in evaluating the corresponding features extracted from the acoustic speech signal. In general, the glottal opening is a relatively slow phenomenon compared to the glottal closing. Therefore, the glottal opening may not appear in the dEGG as a clear impulse. Note that the EGG signal does not carry any information about the variations in the acoustic pressure signal [47]. A recent review of EGG for applications, including basic voice science, clinical practice, and singing, is given in [48].

In addition to EGG, laryngeal imaging methods, such as HSV and videokymography, have been used to compute the ground truth for the evaluation of various methods to extract excitation information from speech [49]. HSV is a technology to extract 2-D images from the motion of the vibrating vocal folds, and it is widely used in voice clinics. Videokymography is a simplified version of HSV based on high-speed imaging of the vocal folds at a specifically selected location along a horizontal line. For more details on HSV, the reader is referred to the review article published in [50], and for more details on videokymography, the reader is referred to the reviews published in [51] and [52]. Compared to EGG, the use of HSV and videokymography is more challenging in the computation of the ground truth because both of these methods require expensive equipment. Also, the obtained imaging data might be of low temporal and spatial resolution, and the methods do not enable a noninvasive analysis of voice production. Laryngeal imaging has been used jointly with acoustical analysis of speech excitation information in studying, for example, glottic cancer [53], diplophonia [54], and phonation onsets [55]. For visualization, simultaneously recorded HSV and EGG signals are shown in Fig. 5 for the closing and opening phases for a nonpathological vowel production by a male speaker.

IV. EXTRACTION OF EXCITATION INFORMATION FROM SPEECH SIGNALS

In this section, the extraction of excitation information from speech signals is described by first discussing the estimation of the glottal flow waveform using GIF and the parameterization methods developed to express excitation information from the glottal flow waveforms. Next, the most important excitation information features, which can be extracted directly from speech signals, such as F_0, GCI, and GOI, are discussed.

A. Extraction of Excitation Information Using GIF

GIF refers to the approach to estimate the glottal source from speech signals. In this section, we will first give an overview of the GIF methods and then describe the parameters derived from the estimated glottal source waveforms.

1) GIF Methods: The estimation of the glottal source waveform by GIF is based on estimating the vocal tract filter. The effect of the vocal tract resonances is reduced by filtering the speech signal through the inverse of the estimated vocal tract transfer function. The idea of GIF was proposed in the 1950s [56] using analog antiresonance circuits. Since the 1970s, GIF methods are using digital signal processing tools. These methods differ mainly in the way the vocal tract transfer function is estimated. Most methods are based on LP analysis, which assumes that the vocal tract transfer function can be approximated by an all-pole filter [57]. A widely used LP-based GIF method, i.e., the CP analysis, was proposed in [58]. The CP analysis is based on computing the vocal tract transfer function with LP using the covariance criterion that is computed from speech samples in the CP of the glottal cycle (i.e., this method calls for the extraction of GCI and GOI). Another popular GIF method is the IAIF [57]. In this method, the average effect of the glottal source on the speech spectrum during the open phase and CP of the glottal cycle is first estimated with a low-order all-pole filter. By removing this estimated average effect of the glottal source, a vocal tract model is computed without using the knowledge of GCI or GOI.

More recent GIF methods are based on the QCP analysis [59] and QPR [60]. In the former, the CP analysis is replaced by a temporally weighted LP analysis, called weighted LP. QPR together with physically motivated

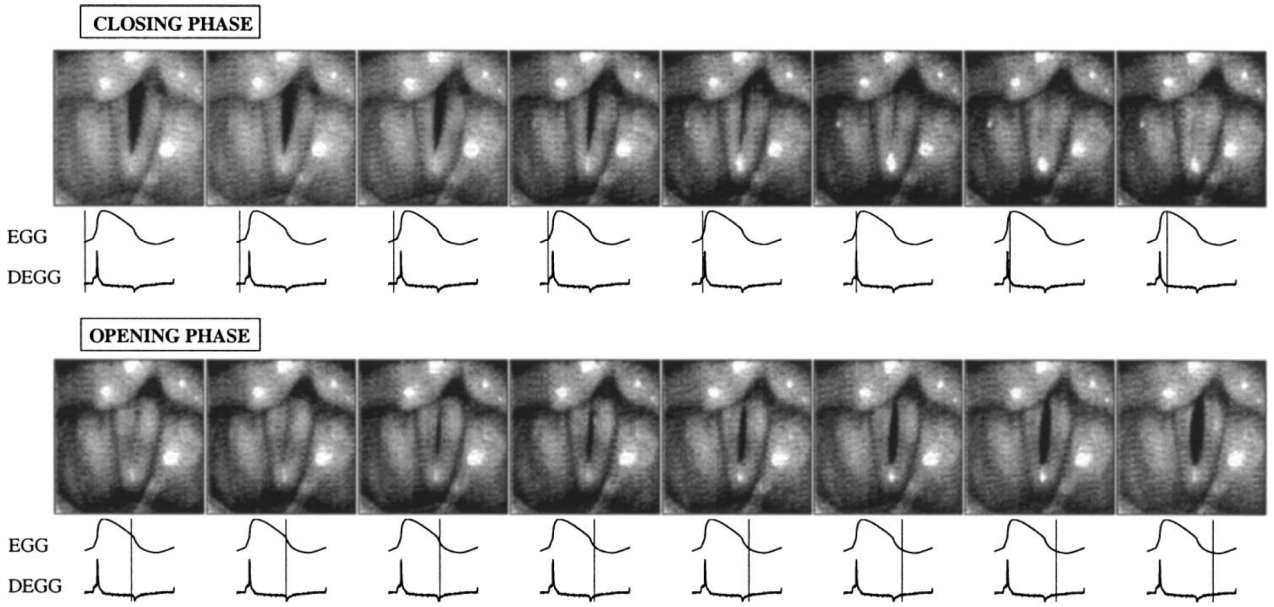


Fig. 5. *Visualization of the closing and opening phases of the glottal cycle by simultaneous electroglottographic and high-speed recordings. Vertical bars to the EGG and dEGG signals indicate the moment in time at which the visual image occurs. The EGG sampling frequency is 44 444 Hz, and the high-speed camera sampling frequency is 3704 frames/s (reproduced from [46] with permission of the publisher, the Acoustical Society of America).*

optimization (e.g., the flatness of the CP) is used to model jointly the vocal tract and the lip radiation.

GIF methods have also been developed based on the joint optimization of the source and filter [61]–[64]. In these methods, glottal source models, such as the Liljencrants–Fant (LF) [65] model and the Rosenberg–Klatt (RK) model [30], [66], are used to represent the glottal flow pulse or its derivative in a parametric form. Due to the use of predefined mathematical functions for the glottal source, these GIF methods are limited in their ability to capture the behavior of the glottal source in natural speech, particularly for phonation types. Moreover, the use of multiparameter source models usually prohibits the use of classical optimization methods due to the nonconvex nature of the error surface, thus increasing the computational complexity [62]. The joint optimization of the source and filter has also been applied in GIF using acoustical tube models of the vocal tract [67]. The GIF proposed in [67] uses state-space modeling based on a concatenated tube model of the vocal tract and the LF model of the source. By optimizing the model using extended Kalman filtering, estimates of the glottal source and intermediate pressure values within the vocal tract are obtained.

GIF methods have also been developed using a combination of causal (minimum phase) and anticausal (maximum phase) components of the speech signal. The ZZT method [68], [69] and the CCD method [70] are two methods in this category. In these methods, the response of the vocal tract and the return phase of the glottal flow are considered as causal signals, and the open phase of the glottal flow is considered as an anticausal signal. These signals are separated by the mixed-phase decomposition using analysis synchronized with the GCIs. The performances of the ZZT and CCD methods are limited due to the use of short speech segments and also due to computational cost [69], [70]. Moreover, the assumption that speech can be expressed as a combination of causal and anticausal components may not hold when the speech data are degraded due to noise.

In all GIF methods, the ultimate goal is to try to estimate the ground truth, that is, the true glottal volume velocity waveform produced by the vocal folds, with maximum accuracy. Unfortunately, noninvasive recording of the true glottal flow is not possible in the natural production of speech. This absence of the ground truth is an inherent obstacle in the assessment of all GIF methods. The problem has been circumvented in some studies by synthetic test vowels generated using artificial glottal flow waveforms, such as the LF model [61], [67]. In addition, some studies have used physical modeling of the human voice production [59], [71]–[73]. In this approach, the test data are generated by simulating physical laws in sound production and transmission, instead of using preselected artificial source waveforms, which are linearly filtered with digital vocal tract models. A few recent studies [74], [75] proposed using a physical apparatus, where synthetic speech signals are produced by using known voice source waveforms as inputs. The physical vocal tract replica is made of stacked plexiglass disks or 3-D-printed in plastic using MRI images of the true vocal tract. In the above studies, the glottal flow estimated by GIF was compared with the information of the glottal area. There

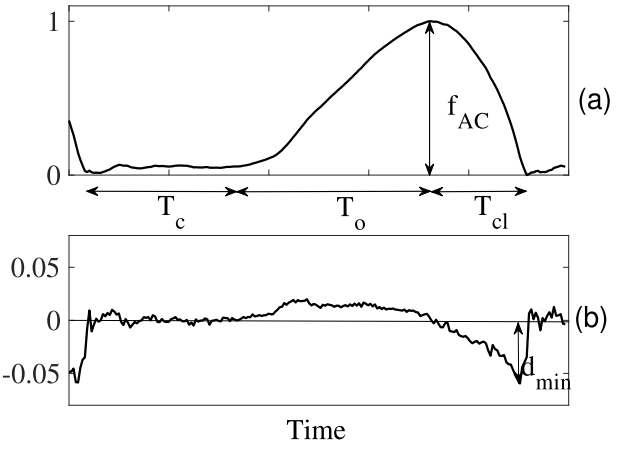


Fig. 6. *Computation of time-based and amplitude-based parameters from (a) glottal pulse and (b) its first time-derivative. The ac-flow (f_{ac}), minimum flow (f_{min}), and the minimum of the derivative (d_{min}).*

are many recent investigations where GIF methods have been studied jointly with glottal area information extracted using HSV or with physical models of voice production. These investigations have addressed issues, such as the relationship between the glottal flow and glottal area in the presence of source-filter interaction [76], [77] and in phonation onsets [55], the computation of parameter values for physical models [78], [79], and the estimation of subglottal pressure, laryngeal muscle activation, and vocal fold contact pressure [80]. We argue that the strategy used in these investigations to study excitation information of speech (i.e., using GIF jointly with HSV and with physical modeling approaches) will become increasingly important and also increasingly feasible in the future due to the progress in HSV [50], [81], physical modeling [82], [83], and GIF [59], [67], [71].

2) Parameterization of the GIF-Based Glottal Flow Estimates: Glottal flows estimated by GIF are parameterized by expressing some important features in a compressed numerical form. Methods for parameterization of the glottal flow estimates can be grouped into time- and frequency-domain methods.

a) Time-domain parameterization methods: The traditional way to parameterize the glottal flow waveform in the time domain is to compute time-based quotients. This involves measuring ratios of time durations of different phases of glottal flow waveform in one cycle. These time-based measures require the identification of GCI and GOI in the estimated glottal waveforms. For illustration, one cycle of the estimated glottal flow and its derivative are shown in Fig. 6(a) and (b), respectively. In the figure, the glottal pulse is divided into three parts: the CP (T_c), the opening phase (T_o), and the closing phase (T_{cl}). The most widely used time-domain parameters are the OQ, SQ, and ClQ, which are defined as follows:

$$\mathrm{OQ} = \frac{T_o + T_{\mathrm{cl}}}{T} \tag{4}$$

$$\mathrm{SQ} = \frac{T_o}{T_{\mathrm{cl}}} \tag{5}$$

$$\mathrm{ClQ} = \frac{T_{\mathrm{cl}}}{T} \tag{6}$$

where $T = T_c + T_o + T_{\mathrm{cl}}$ is the period of the glottal cycle.

Time-domain parameters are affected by distortions, such as ripple, caused by incomplete canceling of formants. To counter the effects of the ripple, time-domain parameters are sometimes computed by replacing the true closure and opening instants with the time instants when the glottal flow crosses a level, which is set to a value between the minimum and maximum amplitudes of the glottal pulse [84].

The time-domain parameterization of the glottal flow can also be computed using amplitude-based measures. The most widely used amplitude-based time-domain parameterization methods take advantage of two prominent amplitude values of the glottal flow and its derivative: the ac amplitude of the glottal flow pulse and the amplitude of the negative peak of the flow derivative [65], [85]–[87]. An amplitude-based parameter called the NAQ proposed in [86] is given by

$$\mathrm{NAQ} = \frac{f_{\mathrm{ac}}}{d_{\mathrm{min}} \cdot T}. \tag{7}$$

b) Frequency-domain parameterization methods: Frequency-domain parameters of the glottal flow are obtained from the Fourier transform of the estimated glottal flow. In practice, only the power spectrum is used to derive the frequency-domain parameters. A widely used frequency-domain parameter is the alpha ratio, which measures spectral tilt by computing the ratio between the spectral energies below and above a certain frequency (typically $\leq$1 kHz) [88]. Another frequency-domain glottal flow parameter is the HRF [89]. The HRF measures the tilt of the glottal flow spectrum as the ratio between the sum of the amplitudes of harmonics above F_0 and the amplitude of F_0. Another measure for the spectral tilt of the glottal flow is the dB difference between the amplitude of the fundamental and the second harmonic, i.e., H1-H2 [90]. It is also possible to quantify the glottal flow using the ratio between the harmonic and nonharmonic components of the glottal flow spectrum, which is referred to as the HNR [91], [92].

B. Extraction of F_0

F_0 of the vocal fold vibration is one of the important components of excitation information in voiced speech. The value of F_0 varies from about 60 Hz in low-pitched male voices to about 1500 Hz in sopranos' singing voices [93]. The temporal variation of F_0 corresponds to

intonation, which contributes to vocal emotions [94]. The factors affecting the performance of F_0 estimation methods are the effects of vocal tract resonances, the rapid variation of F_0 (e.g., in emotional speech and children's speech), and signal degradation due to noise and reverberation.

In the production of some speech sounds, the glottal excitation is inherently aperiodic containing more noise (such as in breathy phonation) or diplophony (such as in vocal fry) [95], which needs further investigation. As F_0 extraction is covered in several tutorials/books, this topic is not handled in detail in this review article, but we, instead, discuss the general aspects of F_0 extraction briefly here and focus more on recent deep learning-based progress of the topic in Section VI-A. For more details on F_0 extraction, please see [96]–[104], where various methods are reviewed by for the study of clean and noisy speech, as well as singing voices.

The F_0 extraction methods can be grouped into three broad categories: 1) time-domain; 2) frequency-domain; and 3) time–frequency-domain methods. Time-domain methods take advantage of the periodicity of the speech signal or the LP residual. In this category, autocorrelation-based methods are popular due to their simplicity. The autocorrelation function measures the degree of similarity between a signal and its delayed version [105]. An estimate of the pitch period, i.e., the inverse of F_0, is obtained by using the location of the peak in the autocorrelation function computed from a segment of speech or LP residual. This approach is used in many F_0 extraction methods, such as SIFT [97], [106], RAPT [107], YAAPT [108], and PRAAT [109]. Several modifications to the autocorrelation-based methods were proposed in the YIN method [93].

The spectra of periodic time-domain signals consist of high-energy amplitude components, located at F_0 and its harmonics. This property forms the basis for frequency-domain methods. Examples of methods belonging to this category are the SHRP [110], the SRH [111], the summation of impulse-sequence harmonics [104], the method of dominant harmonics [112], and the SWIPE [113].

In the time–frequency-domain methods, the speech signal is first decomposed into several frequency bands, and then, the time-domain methods are applied to each subband signal. The auditory-model correlogram-based algorithm [114] is a popular method, in which speech is decomposed using an auditory filter bank, and an autocorrelation function is computed for each subband signal. MBSC-based F_0 estimation [115] uses four wideband FIR filters to capture multiple harmonics in every subband. Different weighting schemes are used to obtain the peak of the enhanced summary correlogram for robust F_0 estimation.

C. Extraction of GCI

The derivative of the glottal flow waveform estimated from natural speech typically shows a prominent negative peak during the closing phase [28], [86]. This negative peak serves as the main excitation of the vocal tract system in each glottal cycle. The time instant of the negative peak is called GCI. The GCI is used in different areas of speech research, such as study of glottal activity [116], estimation of pitch [104], [117]–[119] and formants [120], [121], and the analysis of loudness [36] and nonverbal sounds (such as laughter [23] and shouting [122]). GCIs are also used in the time delay estimation [123]–[125], in determining the number of speakers from mixed signals [126], speech enhancement [5], [6], multispeaker separation [127], prosody modification [128], and speech synthesis [3], [4].

The widely used GCI detection methods are grouped into three categories [129]. The first category is based on processing the excitation signal, the second category involves processing the speech signal, and the third category uses both the speech signal and the excitation signal.

1) Methods Based on Processing the Excitation Signal: The methods in this category use the excitation signal derived from the speech signal after removing the contribution of the vocal tract. This is usually carried out by using the LP analysis. The location of the large error value in the LP residual within a glottal cycle corresponds to the GCI. Identification of GCI locations from the LP residual is sometimes difficult due to the polarity of the residual values around the GCI. To overcome this difficulty, the use of the Hilbert envelope of the LP residual was proposed in [130]. In [131], the Gabor filtering of the Hilbert envelope of the LP residual was used to detect GCIs. Some methods use the group delay function of the LP residual to locate the GCIs [131], [132]. It was found in [133] that the group delay-based methods gave high false alarms. Dynamic programming-based techniques were proposed to reduce false alarms. Methods in [134] and [135] use the glottal flow waveform instead of the LP residual to detect the GCIs. The ILPR was used to detect the GCIs by searching for transients in the ILPR using the dynamic plosion index [136].

2) Methods Based on Processing the Speech Signal: Earlier methods for GCI detection were based on short-time energy of the speech signal in the time–frequency representation [137]. For the energy computation and the time–frequency representation, block processing of the speech signal is required, which may affect the accuracy of the GCI detection. In [138], GCIs were detected by searching for the maximum of the determinant of the autocovariance matrix of the speech signal.

Some methods exploit the properties of the impulse-like excitation present in the speech signal due to GCI. ZFF is one such method that takes advantage of the nature of the impulse-like excitation. In ZFF, the speech signal is filtered around 0 Hz using a cascade of two digital resonators [139]. The negative-to-positive zero crossings of the ZFF signal correspond to GCIs for a signal with positive polarity [140]. Another technique in this category is the LoMA method, which uses the time-scale representation to locate GCIs [141]. The idea of the LoMA method is

that discontinuities in the speech signal at GCIs and GOIs are reflected as amplitude maxima at each scale of the wavelet transform. Within a pitch period, an optimal LoMA is computed using dynamic programming to detect the GCIs. In [142], singularity/discontinuity behavior present in the speech signal was exploited using a nonlinear technique, called the microcanonical multiscale formalism, for GCI detection. The method was shown to be robust in conditions of low SNR. Recently, the magnitude spectral properties of the time-domain impulses were exploited to detect the GCIs using the SFF method [143]–[145]. The method was shown to be robust in detecting GCIs in emotional speech and telephone quality speech.

3) Methods Based on Processing Both Speech Signal and Excitation Signal: In this category, the methods use the speech signal to first identify possible GCI locations within a certain interval. After this, discontinuities in the excitation signal are used to locate the GCIs. SEDREAMS is one such method [146]. SEDREAMS uses the mean-based signal to find the possible GCI locations in an interval, after which the peak of the LP residual in the interval is used to detect the GCI. The mean-based signal oscillates around the local pitch period, thus guaranteeing good performance in terms of reliability, i.e., reduction in the number of false alarms and misses. In [147], SEDREAMS was modified to handle speech of different voice qualities. This method uses postprocessing techniques and dynamic programming, in addition to SEDREAMS. Other methods, such as DYPSA [133] and YAGA [134], use the excitation signal (LP residual in DYPSA and glottal flow waveform in YAGA), wavelet transform, group delay, and dynamic programming by minimizing various cost functions. The cost function consists of various elements, such as the interpulse similarity, normalized energy values, pitch deviation, costs derived from the projected phase slope, and deviations from an ideal phase slope function. More details on the GCI detection methods and the GCI-based analysis of speech processing can be found in [28], [29], [129], [146], and [148].

D. Extraction of GOI

In comparison to the detection of GCIs, the detection of GOIs is generally more difficult from speech signals because the abduction of the vocal folds is typically a more gradual phenomenon compared to the abduction of the vocal folds [28]. Methods for the detection of GOIs are mainly based on first detecting the GCIs, after which a suitable duration is assumed for the open phase, either by fixing a value or by using a ratio with respect to the pitch period. The detection of GOIs is needed for the CP analysis and characterizing speech production using the OQ [149], [150].

It is to be noted that there is no unique definition for GOI [134]. Three main definitions of GOI are reported in the literature [134]. Each one of these definitions is limited to a specific application of interest. In the first definition, the GOI occurs at the end of the CP, where an increase in the LP residual error occurs [58], [134]. This definition is used in the estimation of the glottal flow with the CP analysis. The second definition is based on the dEGG signal, where the GOI is identified as the location of the maximum value of the dEGG signal, corresponding to the maximum rate of change of the glottal impedance/conductance [46], [151]. This definition has been used to compute the OQ to describe pathological voices [149], [150]. The third definition of GOI is based on the EGG signal, by defining GOI as the time instant where the amplitude of the EGG signal is equal to a given percentage of the maximum value of the EGG signal within the glottal cycle [152]. Since the glottal opening is typically more gradual compared to glottal closing, it is appropriate to define the GOI as an *interval* within a glottal cycle rather than a time instant.

In [153], the Hilbert envelope of the LP residual was used for the detection of GOIs, after first detecting GCIs. In [134], [154], and [155], the multiscale product of the decomposed wavelet signals was shown to be effective for the GCI/GOI detection from speech and EGG signals. In [134], the YAGA method was proposed for the detection of GCIs/GOIs using wavelet transform, group delay, glottal flow waveform, and dynamic programming. SEDREAMS uses the LP residual and mean-based signal to detect GCIs/GOIs [146].

From the speech production's point of view, when the vocal folds are completely open in a glottal cycle, the subglottal system is maximally coupled to the supraglottal system, and the resultant vocal tract is longer compared to the tract during the CP. The effect of opening on the response of the vocal tract system is different during different stages of the open phase. When the glottis starts to open, the bandwidth of the first formant of the supraglottal vocal tract begins to increase. On the other hand, at the end of the opening phase, the effective vocal tract length will be larger due to coupling, and therefore, the center frequency of the lowest resonance will decrease and its bandwidth will increase. This results in the increased spectral flatness of the response of the vocal tract system. Motivated by this phenomenon, the lower DRF is used for deriving the open phase using the ZTW method [156], [157]. The glottal open phase is determined using a threshold value of 0.5 over the normalized DRF contour. The interval below this threshold is identified as the open phase and the remaining part of the glottal cycle as the CP. It was shown in [145] that the spectral flatness computed at each instant of the ZTW spectrum highlights glottal opening, as the effective vocal tract length is longer in the glottal open phase, which increases the bandwidths of the resonances, making the spectrum flatter, compared to the CP. In [145], the open phase is identified as the interval between the peak in the spectral flatness plot within a glottal cycle to the following GCI.

Research has also been conducted to extract impulse-like sequences and their relative strengths in each glottal cycle directly from the speech signal [23], [158].

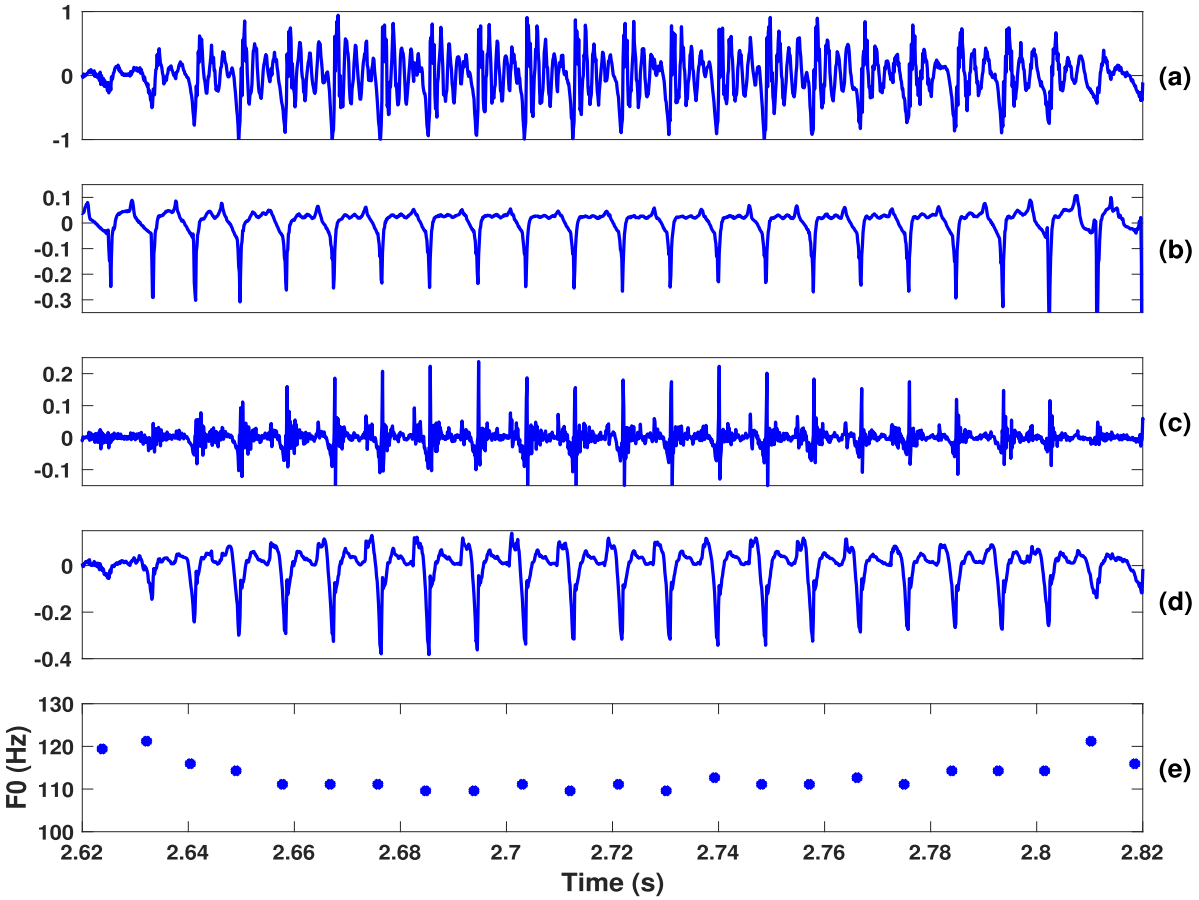


Fig. 7. *Illustration of some excitation features. (a) Speech signal. (b) dEGG signal. (c) LP residual. (d) Glottal flow derivative. (e) Instantaneous* F_0.

In [9], [159], [160], the excitation component was represented as a multiple-pulse sequence for the purpose of speech synthesis. For this, the LP analysis and synthesis methods were used to determine the locations and strengths of the impulses by considering one pulse at a time or by jointly optimizing the strengths of several pulses (such as the regular pulse excitation and the random pulse excitation). In a more recent study [158], a method was proposed to extract a sequence of impulses from the signal by modifying the ZFF method using various levels of trend removals. This approach is justified by the pitch perception of expressive voices [104], [161]–[163]. However, there is a need for signal processing techniques that can exploit the impulse-like sequences derived directly from the input signal without using block processing and vocal tract system characteristics.

In addition to the issues described above, some studies have been proposed for extracting the excitation information using features, such as the strength of the impulse-like excitation at glottal closure (as in NAQ) and the sharpness or the abruptness of glottal closure [36], [164]. Fig. 7 illustrates some of the excitation features extracted from the speech signal. In Fig. 7, (a) shows a segment of voiced speech, (b) shows the dEGG signal, (c) shows the LP residual, (d) shows the glottal flow derivative, and (e) shows the instantaneous F_0.

V. UTILIZATION OF EXCITATION INFORMATION IN DIFFERENT AREAS OF SPEECH RESEARCH

In this section, we discuss how the excitation information is utilized in different areas of speech research. The section is divided into four research areas, where extraction of excitation information plays a significant role:

1) study of phonation types;
2) study of vocal emotions;
3) study of laughter sounds;
4) study of pathological voices.

A. Study of Phonation Types

Humans are capable of coloring their speech by changing phonation type, i.e., the vibration mode of the vocal folds. The analysis and classification of different phonation types are needed in applications, such as in speech synthesis and modification systems [89], [165], [166], and tagging expressive speech corpora [167]. Furthermore, the identification of phonation type is useful in the assessment of the cognitive load of the speaker, speaker recognition, emotion recognition, and speech recognition [14], [17], [29], [168]–[173].

Generally, three broad phonation types are considered. They are breathy, modal (or normal), and pressed (or

tense). When phonation type changes from breathy to modal and pressed, the characteristics of the glottal flow pulse change considerably. The glottal flow pulse changes from a smooth symmetric waveform in breathy phonation to an asymmetric waveform with sharp edges in pressed phonation [11], [174]. This variation in the time domain is reflected as the decrease in the decay of the spectral envelope of the glottal pulse in the frequency domain [175], [176].

Glottal source parameters were explored for discriminating breathy, modal, and tense voices in [11] and [147]. Frequency-domain parameters, such as H1-H2 [176], HRF [89] and the PSP [177], were used for the discrimination task. In addition, time-domain parameters, such as ClQ, QOQ, OQ, and SQ, and amplitude-based parameters, such as NAQ, were also used [11], [30], [86]. Some studies measured the amount of aspiration noise present in the signal for detection of breathy voice based on the observation that breathy voices are noisier compared to modal voices [176], [178]. In [179] and [180], parameters were derived for various voice qualities by fitting the estimated glottal source waveform with the LF model.

In [164] and [175], it was observed that H1-H2 and NAQ were the best parameters for discrimination of different phonation types. However, it was observed that the accuracy of the estimated glottal source parameters reduces for high-pitched voices and expressive voices [29], [30]. To overcome this, attempts have been made recently to extract the excitation information directly from the speech signal. In [164], a parameter called the MDQ was proposed to capture the sharp changes in the glottal closure characteristics from the LP residual. In [175], using the spectral parameter LFSD, it was observed that pressed voices show smaller OQ, and breathy voices show higher OQ. The effect of the subglottal system on the spectrum is stronger for breathy voices due to larger OQ compared to the pressed voices. Larger OQ results in the increase in LFSD for breathy voices, typically around the region of the glottal formant (which is lower in frequency than the first formant). In [175], it was observed that LFSD and MDQ are close to NAQ, and HNR seems to provide poor discrimination for the three phonation types. However, HNR was shown to provide good discrimination of breathy and modal voices compared to pressed and modal voices. It was observed that H1-H2 performs poorly for female speakers, and it is as good as NAQ for male speakers. This may be due to the overlap of the second harmonic with the first formant for female voices. In general, it was observed that no single parameter performed consistently well for all the speakers in the discrimination of phonation type.

Kadiri *et al.* [34], Kadiri and Yegnanarayana [182], [183], and Kadiri and Alku [184] explored the features derived from the ZFF, ZTW, and SFF methods for discriminating phonation types. In these studies, cepstral coefficients were obtained from the spectra estimated by the three methods, and the cepstral coefficients were used in addition to excitation information scalar features (such as spectral statistics). Recently, in [184], the MFCCs computed from the glottal source waveforms estimated by the QCP method and the ZFF method were shown to be effective for the classification of different phonation types from speech signals.

Table 2 Trend in Spectral Features of Emotional Utterances With Respect to Neutral State Utterance (Increase: ↑ and Decrease: ↓) [185], [186]

	Mean F_1	Mean F_1 bandwidth	Mean F_2	Spectral tilt
Anger	↑	–	↑/↓	↓
Happiness	↓	↑	–	↓
Sadness	↑	↓	↓	↑

B. Study of Vocal Emotions

The features used for emotion recognition can be broadly characterized as spectral and prosodic features. The general trend of four spectral features, i.e., changes in the lowest two formant frequencies, the bandwidth of the first formant (F_1), and spectral tilt, is indicated in Table 2 for anger, happiness, and sadness [185], [186]. The trend is indicated as an increase or decrease in the parameter value relative to the neutral state. Similarly, the trend of prosodic features, i.e., F_0, energy, and speaking rate, is indicated in Table 3 [185], [186].

The basic technological principles of emotion recognition systems are similar to those used in speech and speaker recognition, as well as in language identification [187]–[189]. In most emotion recognition studies, short segments of speech are represented in terms of spectral features, such as MFCCs or LPCCs, prosody features, and their statistics [185], [187], [190]–[194]. These features are available in open toolkits, such as openSMILE [192], [195]–[197]. The features extracted from the emotional speech are used to develop nondiscriminative/discriminative models, such as GMMs, FFNNs, and DNNs [187], [198]–[200]. Binary classification techniques, such as SVMs and Bayesian logistic regression, have been used for the multiclass problem by adopting them in hierarchical binary decision tree framework [188], [196], [201].

Emotion recognition systems generally use the features representing the vocal tract system characteristics. There are fewer studies of emotional speech involving the use of excitation information [14]–[17], [199], [202], [203]. Most of these studies use the voice source features computed from a specific category of speech sounds, such as vowels [14], [15], [17], [202], [204]. In [15] and [16], the role of the voice source in the perception

Table 3 Trend in Prosody Features of Emotional Utterances With Respect to Neutral State Utterance (Increase: ↑ and Decrease: ↓) [185], [186]

	Mean F_0	Variance of F_0	Energy	Speaking rate
Anger	↑	↑	↑	↑
Happiness	↑	↑	↑	↑
Sadness	↓	↑	↓	↓

of emotional arousal (active and passive) and valence (positive and negative) attributes was studied from short vowels (with a duration of 150 ms). The results showed that NAQ correlates better with arousal than valence for both genders. Similarly, in [14], emotions in short vowel segments of [a :] in the continuous speech were analyzed. Significant differences were found in NAQ between most emotions. Even though NAQ correlates with emotions and voice quality changes, it was found that NAQ by itself is not sufficient for discriminating between emotions accurately [14]. The interdependencies among glottal source features were studied in [17] between five emotions using six glottal source parameters extracted from the glottal flows estimated by GIF [17]. In [202] and [204], the robustness of the glottal source features was examined across databases for four emotions (anger, happiness, neutral state, and sadness).

Studies (e.g., [18]–[20], [206], and [207]) have investigated excitation features derived directly from the speech signal. Gangamohan *et al.* [18] examined four excitation features (F_0, the SoE, the EoE, and loudness) for discriminating different emotions. These features were used to build emotion detection [206] and emotion recognition systems [19], [20], [205]. The effectiveness of excitation during the production of emotional speech was examined in [207]–[209] using prosody modification to convert neutral speech to emotional speech.

C. Study of Laughter Sounds

Nonverbal sounds, such as laughter, convey nonlinguistic information. Production of these sounds is typically involuntary and spontaneous. Nonverbal sounds do not have any clear description of articulation. In laughter, changes occur in the excitation due to involuntary bursts of activity. Laughter conveys a variety of functions, such as indication of affection, aggressive behavior (laugh in someone's face), bonding behavior (such as in early infancy), or appeasement behavior (such as in situations of dominance) [210]. Detection of laughter can help in understanding the emotional state of a speaker [211]. The analysis of laughter also helps in spotting regions of laughter in continuous speech. Characterization of laughter helps in laughter synthesis.

Laughter sounds have been classified in different ways in different studies. In [212], laughter was classified into three classes: 1) spontaneous laughter; 2) voluntary laughter; and 3) speaking or singing laughter. In spontaneous laughter, there is an urge to laugh without restraining its expression. Voluntary laughter is a kind of fake laughter to produce a sound pattern that is similar to that in natural laughter. The laughter in speaking/singing is not based on forced breathing but on well-dosed air supply, which results in breathiness and aspiration. The continuum from speech to laughter was divided into three categories [213], [214]: speech, speech–laughter, and laughter. The duration of vocalization was observed to increase in speech–laughter. This is likely due to changes in one/more features of vowel elongation, pitch, breathiness, and syllabic pulsation [213]. Voiced laughter was shown to induce a significantly more positive emotional response in listeners compared to unvoiced laughter [215].

In [216], laughter analysis was carried out using features, such as F_0, time duration, root mean square amplitude, and formant frequencies. It was observed that laughter has significantly longer unvoiced regions compared to voiced regions. The mean F_0 of laughter sounds was reported to be 472 Hz for (Italian and German) female speakers, and the F_0 values ranged between 246 and 1007 Hz [210], [217]. The average F_0 of normal speech sounds was reported to be 214 and 124 Hz for female and male speakers, respectively. A group of acoustic features, including F_0, the number of calls per bout, formant clusters (F1 versus F2), and spectrograms, were investigated in [218] to analyze temporal features of laughter, their production modes, and source-filter effects. Their study proposed a subclassification of F_0 contours in each laughter call into rising, falling, flat, sinusoidal, and arched. The acoustic features of laughter–speech continuum, such as the pitch range, voice quality, and formant space, were studied in [214]. Two specific acoustic features (the rhythm and the change in F_0) of the laughter series were investigated in [219]. In [211], combinations of several features (pitch, energy, voicing features, modulation spectrum, and PLP features) were used to model laughter and speech. The voice source characteristics were investigated using the OQ along with spectral tilt in [214]. Voice source features, including the instantaneous pitch period, the SoE at glottal closure, and their slopes and ratio, were used for the analysis of laughter in [218] and [221].

D. Study of Pathological Voices

Excitation information of speech is also used in studying pathological voices. Voice pathologies are disorders in which the phonation process in the larynx is disturbed due to, for example, dysphonia, polyps, and vocal nodules [221], [222]. Voice disorders are complex, and they often do not have a single etiology [221]. Voice pathologies arise due to infections, psychogenic, and physiological causes, and due to vocal misuse, which is prevalent in professions, such as teaching, singing, and client service representatives [223], [224]. Change in voice from normal to pathological may indicate early neurodegenerative disease, such as Parkinson's disease [225]. The utilization of excitation information of speech has attracted increasing interest in the area of speech-based detection of neurodegenerative diseases (discussed in Section VI-B). Automatic detection of voice pathology is important because it enables early intervention for the diagnosis.

The features used in investigating pathological voices can be generally classified into the following three categories [226], [227]: 1) perturbation measures; 2) spectral and cepstral measures; and 3) complexity measures. The perturbation measures aim to capture the presence of

aperiodicity and aspiration noise in the voice signals that occur due to irregular movements of the vocal folds and incomplete glottal closure. The most widely used parameters in this category are jitter, shimmer, HNR, normalized noise entropy, and GNE ratio [228]–[240]. The popular features in the category of spectrum/cepstrum measures are MFCCs [227], [241]–[243]. In addition, LPCCs [229], [244], [245] and PLP [227], [246] have been used in voice pathology detection. The complexity measures have been proposed to capture nonlinearity and nonstationarity of voice signals using estimators based on nonlinear dynamic analysis [231], [247]–[252]. The popular features in this category are computed using the fractal and correlation dimension [246], [247], [253]–[255]. More details on the features used for pathology detection can be found in recent review articles [227], [256].

Since voice pathologies may affect different parts of the speech production mechanism, both the vocal tract system and the glottal source need to be parameterized for the analysis and detection. Existing studies have captured the characteristics of the vocal tract effectively by utilizing spectral and cepstral features (such as MFCCs and PLPs). However, there is less research in the analysis and detection of voice pathologies using glottal source features. Recently, a systematic analysis of glottal source features in normal and pathological voices was carried out in [24]. In that study, the glottal source features were derived from the ZFF signal and the glottal flow waveform estimated using the QCP method [59]. The features derived from the ZFF signal consisted of the SoE, EoE, loudness measure, and ZFF signal energy [34], [181], [182]. The glottal flow signals estimated using QCP were parameterized in terms of time- and frequency-domain glottal features [30], [257]. In addition to these, features derived directly from speech signals that capture the specific property of the glottal source were also studied. These features were the CPP [176], PS [258], MDQ [164], and Rd shape parameter [259], [260]. Furthermore, MFCCs derived from the glottal source waveforms were shown to be effective for voice pathology detection. In [26], [262], and [263], it was shown that glottal source features were useful in the automatic detection of dysarthria and also in the assessment of intelligibility in speakers with dysarthria. In [263], glottal parameters computed by GIF were used to identify pathophysiological phonatory mechanisms for phonotraumatic and nonphonotraumatic vocal hyperfunction. In [264], detection of pathological voices caused by vocal nodules was investigated using several glottal parameters and a classifier based on a genetic algorithm. In [265], automatic detection of voice pathology was studied by using a random forest classifier and including several voice disorders, both functional and organic pathologies. The study compared glottal flow features with the widely used openSMILE feature set [266]. The results indicated that the best detection accuracy was obtained by combining glottal features with the openSMILE features. Similar results have been obtained in recent investigations on automatic speech-based detection of diseases, such as heart disease [267] and specific language impairment [268].

VI. RECENT TRENDS IN EXTRACTION AND UTILIZATION OF EXCITATION INFORMATION

This section describes recent developments in the extraction and utilization of excitation information. The section addresses the issue in two parts by first describing the use of deep learning for GIF and extraction of F_0 and GCI. In the second part, the utilization of the excitation information in a popular health topic, the automatic detection of neurodegenerative diseases from speech signals, is discussed.

A. Deep Learning for GIF and for Extraction of F_0 and GCI

Inspired by the success of deep learning in many areas of speech technology, the extraction of excitation information has been recently studied using approaches based on deep learning both in GIF and the detection F_0 and GCI. It is known that signal processing-based GIF methods are affected by distortions in the speech signal due to ambient noise, the poor audio quality of the recording equipment, and compression and bandwidth limitation caused by speech transmission [30], [269]. To address this issue, a few recent studies [269]–[271] have proposed using DNN-based methods for estimation of the glottal source waveform. In [269], coded telephone quality speech was studied using a DNN-based GIF method by using both clean and coded speech in training. DNN was used to map the speech features (line spectral frequencies) extracted from the coded speech to the time-domain glottal flow waveforms estimated from the corresponding clean speech. The glottal flow estimated from clean speech (using an existing signal processing-based GIF method and the QCP method) was used to train the DNN. It was observed that the DNN-based GIF method showed good performance in the estimation of glottal flows under the coded condition for both high- and low-pitched vowels.

As described in Section IV-B, the existing F_0 extraction methods are based on handcrafted signal processing frameworks working in the time-domain and/or frequency-domain. These signal processing approaches are known to be prone to pitch doubling/halving errors. In [102] and [273]–[275], machine learning models for F_0 extraction were proposed. The method proposed in [272] first extracts spectral domain features (the normalized log-frequency power spectrogram) and then adopts a neural network to compute the F_0 estimate. To capture the variation of F_0, RNNs were explored. Specifically, the authors investigated both DNN- and RNN-based methods to produce reasonably accurate probabilistic outputs for pitch. From the pitch probability in each frame, a Viterbi decoding algorithm was used to derive continuous

pitch contour. By removing feature extraction and Viterbi decoding modules, mapping the raw waveform directly to the F_0-corresponded states was proposed in [275]. In [276], the CREPE model, which is an end-to-end CNN that uses the raw waveform, was proposed. The network is trained in a supervised fashion by minimizing the cross-entropy loss between the output of the model and the ground-truth pitch. In [102], a voicing detection was proposed as a classification problem and pitch estimation as a regression problem. For both tasks, various acoustic features and traditional machine learning methods were used. In [277], vocoder-based modifications for speech data augmentation for neural network estimation (such as CREPE) of F_0 were explored.

As described in Section IV-C, existing robust GCI detection methods use a two-stage approach. The initial stage involves the transformation of speech into a representative excitation signal (such as an LP residual), where GCIs can be localized better. The later stage involves the detection of locations of the GCIs. The initial stage uses signal processing approaches based on, for example, the source-filter model of speech production, and the later stage adopts algorithms, such as peak picking and dynamic programming. Recent developments in the area of data-driven representation learning have shown that it is possible to operate directly on the raw speech signal, and let the learning algorithm learn the abstract representations of the underlying task. As an example of this kind of approach, CNNs were utilized in [278] for the GCI detection by operating on low-pass filtered speech and regarding the negative peaks of the filtered signal as the correct GCIs. In [279], the GCI detection was posed as a temporal event detection problem, relaxing the constraints used in [278]. In [279] and [280], the GCI detection was formulated using a representation learning perspective, where an appropriate representation is implicitly learned from the raw signal. In [281] and [282], a deep CNN-based GCI detection method was proposed by fusing raw speech and LP residual features. In [283] and [284], classification-based data-driven algorithms were studied for the GCI detection, using conventional machine learning methods, such as SVMs, extremely randomized trees, k-nearest neighbors, and MLP with handcrafted features extracted from speech. In these studies, the problem was viewed as a two-class classification problem, where a peak in the speech signal could either correspond or not correspond to GCI. The handcrafted features are peak-based features comprising the amplitudes of the negative peak and the neighboring negative peaks, the time difference between the negative peak and each of the neighboring negative peaks, the amplitudes of the neighboring positive peaks, the width of the negative peak and each of the neighboring negative peaks, and the correlation of the signal around each of the neighboring negative peaks. In [285], features, such as voiced/unvoiced, harmonic/noise, and spectral features, were added to the handcrafted features for improving the performance of GCI detection.

B. Utilization of Excitation Information for Detection of Neurodegenerative Diseases

Neurodegenerative diseases, particularly Parkinson's disease and Alzheimer's disease, are becoming increasingly prevalent globally due to the aging of the population. The early detection of these diseases is essential, and speech provides an effective means of biomarking these diseases at an early stage of the disease's progress. Speech-based detection of neurodegenerative diseases has attracted increasing interest as an automatic, low-cost, and easy-to-administer method [231], [286]. The detection methods proposed can be divided into traditional pipeline systems and modern end-to-end systems. In the former, selected handcrafted features are computed from speech to train classifiers (such as SVMs) to predict one of the two labels (disordered versus healthy). Many different speech features have been used in these studies. In the detection of PD, speech has been parameterized with handcrafted features based on articulation, phonation, and prosody [287]–[289]. In the end-to-end systems, the use of handcrafted features is replaced by training deep learning networks that directly map the raw speech signal waveform (or its spectrogram) to the output labels (disordered versus healthy). Deep learning models, such as CNNs, MLPs, and LSTM [288], [290]–[292], for example, have been used for this purpose.

Since neurodegenerative diseases affect phonation, obtaining parameters based on speech excitation information is a justified approach to build traditional pipeline systems for the detection of neurodegenerative diseases from speech signals. A few recent studies [27], [293] have investigated the use of speech excitation information in the detection of PD with the traditional pipeline approach by estimating the glottal flow using the IAIF method (as described in Section IV-A1) and by training SVM classifiers using the computed parameters. These studies indicated that glottal parameters carry useful information to improve detection accuracy. In [294], excitation information was studied in PD by first estimating the glottal flow from speech using GIF, after which parameters of a biomedical two-mass model were determined by fitting the glottal flow spectrum to the model. The study showed that the biomedical model can be used to measure the instability of phonation, and the features are good biomarkers of PD.

Some recent investigations have studied the use of time-domain excitation information to build end-to-end systems for the detection task. In this approach, voice excitation information is represented by the estimated glottal flow waveform, which is then used as input to a deep learning-based end-to-end system. There are two justifications for studying this kind of end-to-end system for the detection of neurodegenerative diseases. First, the glottal flow captures the phonation information, which is known to be affected by neurodegenerative diseases [287]–[289]. Second, compared to the speech signal, which is the default input in most of the end-to-end detection systems, the glottal

flow is a more basic signal due to the absence of vocal tract resonances. Using such time-domain signals, deep learning systems can be trained with smaller amounts of training data, as indicated in [295]. This is particularly useful because long voice recordings cannot be obtained from patients easily. The end-to-end systems were recently studied in the detection of voice pathologies [25]. The data for this study included voice pathologies caused by different diseases, including neurodegenerative ALS disease. The study indicated improvements in the detection accuracy when the glottal flow was used as input to deep learning-based classifiers, instead of the speech signal. Similar results were recently reported in [296] for the detection of PD.

VII. CONCLUSION AND DIRECTIONS FOR FUTURE RESEARCH

In this article, a review was provided on the extraction and utilization of the excitation information of speech signals. First, the motivation of the topic was explained. Second, the functioning of the human speech production mechanism was briefly described. Third, the extraction of the main components of excitation information was presented by describing the GIF-based estimation of the glottal flow, the underlying excitation information parameters, and the extraction of F_0, GCI, and GOI. Fourth, the utilization of excitation information in various speech processing tasks was discussed by including analysis and classification of phonation type, the study of emotional speech, the study of nonverbal laughter sounds, and the study of pathological voices. Finally, recent trends of the review topic were discussed by addressing two issues, the utilization of deep learning in GIF, the extraction of F_0 and GCI, and the utilization of excitation information in studying neurodegenerative diseases.

Even though the fundamental theory underlying the review topic, that is, the linear source-filter theory of speech production [297], [298], has been known for more than five decades, the technologies discussed in the review are still topical, and the utilization of speech excitation information has attracted increasing interest in a few areas in recent years. One such area is speech-based biomarking of the state of health, especially the automatic detection and classification of neurodegenerative diseases. This research topic has gained momentum due to the aging of the population, a recognized global grand challenge. In the area of speech-based classification of neurodegenerative diseases, the traditional model-driven systems consisting of separate feature and classification stages are currently increasingly replaced with data-driven end-to-end systems based on deep learning. The end-to-end approach is attractive because it enables building health monitoring systems that do not need any domain expertise in the system training phase. It can, however, be argued that, when the traditional approach is used together with effective speech excitation parameters (e.g., those discussed in Section IV), the analysis benefits from its better capability to demonstrate which particular functions of the speech production mechanism have been affected by the underlying disease. This demonstration capability of traditional speech excitation features can be easily taken advantage of by clinicians and speech-language pathologists. Even though the end-to-end approach has shown better accuracy compared to the traditional, feature-based approach in a few studies [288], [290], [292], the end-to-end technology can be criticized for providing a black box-type of solution with poor interpretability to the detection task [299]. Moreover, the end-to-end approach requires larger amounts of training data than the traditional feature-based pipeline approach. Collecting large amounts of speech data from patient populations is not as easy as it is from healthy speakers.

In addition to the health-related research area described above, we argue that the methods to extract excitation information from acoustic speech signals discussed in this review can be used to improve our knowledge of human speech production, particularly when these methods are used jointly with the latest imaging technologies and physical modeling approaches of voice production. In this area, we emphasize, particularly, the recent progress in HSV (e.g., [81]) and GIF (e.g., [59] and [67]), which, in principle, enables obtaining glottal area and glottal flow signals with good spatial and time resolutions from natural voice production, not only for isolated vowel sounds but also for continuous speech. Information extracted jointly by HSV and GIF can be used both to acquire new fundamental research knowledge about the human speech production process and compute parameter values for physical models of voice production.

The review shows that, despite the fact that many methods have been developed over the past few decades to extract excitation information from speech, the development of new methods is still continuing, and new research is needed in order to tackle known limitations in current methods. One such limitation is related to the extraction of GCI where the performance of the state-of-art methods is good, but the performance is limited by the need for issues, such as the computation of the average pitch period and the use of block processing. The limitation of the performance of the GCI extraction due to these issues is severe, particularly in the analysis of expressive voices due to rapid variations in F_0 and source-filter coupling. In addition, improved robustness is needed in GCI extraction methods to enable their utilization in realistic environments with noise and reverberation. The second topic that calls for new research is the extraction of GOI. The performance of existing GOI extraction methods is poor because the glottal opening is a relatively slow phenomenon (compared to glottal closing), and therefore, it manifests itself weakly in the amplitude characteristics of the speech signal. Hence, a more fine-grained detection of excitation components within a glottal cycle (including instants of secondary excitation near the glottal opening) is needed because they contribute, in addition to the major excitation at the

instant of glottal closure, to the production and perception characteristics of speech signals. Moreover, improved robustness of GIF analysis to noise and other nonideal recording conditions is still needed, despite it having been shown recently in [270] and [296] that conducting inverse filtering with DNNs helps to improve the robustness of GIF. To improve robustness further, deep learning architectures other than DNNs, such as CNNs and LSTMs, could be studied as computational inverse networks of the vocal tract. Furthermore, features that better reflect the physical functioning of the vocal folds in the production of pathological speech or different vocal emotions, for example, need to be developed further to enhance speech analysis and classification, as well as the general understanding of human speech production. ■

Acknowledgment

B. Yegnanarayana would like to acknowledge the Indian National Science Academy for the support.

REFERENCES

[1] J. Laver, *The Phonetic Description of Voice Quality*. Cambridge, U.K.: Cambridge Univ. Press, 1980.

[2] K. N. Stevens, *Acoustic Phonetics*. Cambridge, MA, USA: MIT Press, 1998, pp. 55–126.

[3] T. Raitio *et al.*, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 1, pp. 153–165, Jan. 2011.

[4] T. Drugman and T. Dutoit, "The deterministic plus stochastic model of the residual signal and its applications," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 3, pp. 968–981, Mar. 2012.

[5] S. R. M. Prasanna, "Event based analysis of speech," Ph.D. dissertation, Dept. Comput. Sci. Eng., Indian Inst. Technol. Madras, Chennai, India, Mar. 2004.

[6] K. T. Deepak and S. R. M. Prasanna, "Foreground speech segmentation and enhancement using glottal closure instants and mel cepstral coefficients," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 7, pp. 1205–1219, Jul. 2016.

[7] T. Bäckström, *Speech Coding With Code-Excited Linear Prediction*. Cham, Switzerland: Springer, 2017.

[8] N. S. Jayant, "Digital coding of speech waveforms: PCM, DPCM, and DM quantizers," *Proc. IEEE*, vol. 62, no. 5, pp. 611–632, May 1974.

[9] P. Kroon, E. Deprettere, and R. Sluyter, "Regular-pulse excitation—A novel approach to effective and efficient multipulse coding of speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 5, pp. 1054–1063, Oct. 1986.

[10] B. Atal and B. Caspers, "Beyond multipulse and CELP towards high quality speech at 4 Kb/s," in *Advances in Speech Coding* (The Springer International Series in Engineering and Computer Science), vol. 114, B. Atal, V. Cuperman, and A. Gersho, Eds. Boston, MA, USA: Springer, 1991, pp. 191–201.

[11] M. Airas and P. Alku, "Comparison of multiple voice source parameters in different phonation types," in *Proc. INTERSPEECH*, Aug. 2007, pp. 1410–1413.

[12] D. Liu, E. Kankare, A.-M. Laukkanen, and P. Alku, "Comparison of parametrization methods of electroglottographic and inverse filtered acoustic speech pressure signals in distinguishing between phonation types," *Biomed. Signal Process. Control*, vol. 36, pp. 183–193, Jul. 2017.

[13] C. Gobl and A. N. Chasaide, "Acoustic characteristics of voice quality," *Speech Commun.*, vol. 11, nos. 4–5, pp. 481–490, Oct. 1992.

[14] M. Airas and P. Alku, "Emotions in vowel segments of continuous speech: Analysis of the glottal flow using the normalised amplitude quotient," *Phonetica*, vol. 63, no. 1, pp. 26–46, Mar. 2006.

[15] T. Waaramaa-Mäki-Kulmala, *Emotions in Voice: Acoustic and Perceptual Analysis of Voice Quality in the Vocal Expression of Emotions* (Acta Univesitatis Tamperensis). Tampere, Finland: Tampere Univ. Press, 2009.

[16] T. Waaramaa, A.-M. Laukkanen, M. Airas, and P. Alku, "Perception of emotional valences and activity levels from vowel segments of continuous speech," *J. Voice*, vol. 24, no. 1, pp. 30–38, 2010.

[17] J. Sundberg, S. Patel, E. Björkner, and K. R. Scherer, "Interdependencies among voice source parameters in emotional speech," *IEEE Trans. Affect. Comput.*, vol. 2, no. 3, pp. 162–174, Jul. 2011.

[18] P. Gangamohan, S. R. Kadiri, and B. Yegnanarayana, "Analysis of emotional speech at subsegmental level," in *Proc. INTERSPEECH*, Aug. 2013, pp. 1916–1920.

[19] P. Gangamohan, S. R. Kadiri, S. V. Gangashetty, and B. Yegnanarayana, "Excitation source features for discrimination of anger and happy emotions," in *Proc. INTERSPEECH*, Sep. 2014, pp. 1253–1257.

[20] S. R. Kadiri, P. Gangamohan, S. V. Gangashetty, and B. Yegnanarayana, "Analysis of excitation source features of speech for emotion recognition," in *Proc. INTERSPEECH*, Sep. 2015, pp. 1324–1328.

[21] J. Pohjalainen, T. Raitio, S. Yrttiaho, and P. Alku, "Detection of shouted speech in noise: Human and machine," *J. Acoust. Soc. Amer.*, vol. 133, no. 4, pp. 2377–2389, Apr. 2013.

[22] S. A. Thati, S. K. Kumar, and B. Yegnanarayana, "Synthesis of laughter by modifying excitation characteristics," *J. Acoust. Soc. Amer.*, vol. 133, no. 5, pp. 3072–3082, May 2013.

[23] V. K. Mittal and B. Yegnanarayana, "Analysis of production characteristics of laughter," *Comput. Speech Lang.*, vol. 30, no. 1, pp. 99–115, Mar. 2015.

[24] S. R. Kadiri and P. Alku, "Analysis and detection of pathological voice using glottal source features," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 2, pp. 367–379, Feb. 2020.

[25] N. P. Narendra and P. Alku, "Glottal source information for pathological voice detection," *IEEE Access*, vol. 8, pp. 67745–67755, 2020.

[26] N. P. Narendra and P. Alku, "Automatic assessment of intelligibility in speakers with dysarthria from coded telephone speech using glottal features," *Comput. Speech Lang.*, vol. 65, Jan. 2021, Art. no. 101117.

[27] M. Novotný, P. Dušek, I. Daly, E. Ržička, and J. Rusz, "Glottal source analysis of voice deficits in newly diagnosed drug-naïve patients with Parkinson's disease: Correlation between acoustic speech characteristics and non-speech motor performance," *Biomed. Signal Process. Control*, vol. 57, Mar. 2020, Art. no. 101818.

[28] B. Yegnanarayana and S. V. Gangashetty, "Epoch-based analysis of speech signals," *Sadhana*, vol. 36, no. 5, pp. 651–697, Oct. 2011.

[29] T. Drugman, P. Alku, A. Alwan, and B. Yegnanarayana, "Glottal source processing: From analysis to applications," *Comput. Speech Lang.*, vol. 28, no. 5, pp. 1117–1138, Sep. 2014.

[30] P. Alku, "Glottal inverse filtering analysis of human voice production—A review of estimation and parameterization methods of the glottal excitation and their applications," *Sadhana*, vol. 36, no. 5, pp. 623–650, Oct. 2011.

[31] I. R. Titze, "Nonlinear source–filter coupling in phonation: Theory," *J. Acoust. Soc. Amer.*, vol. 123, no. 5, pp. 2733–2749, May 2008.

[32] I. Titze, T. Riede, and P. Popolo, "Nonlinear source–filter coupling in phonation: Vocal exercises," *J. Acoust. Soc. Amer.*, vol. 123, no. 4, pp. 1902–1915, Apr. 2008.

[33] M. Rothenberg, *Acoustic Interaction Between the Glottal Source and the Vocal Tract, Vocal Fold Physiology*. Tokyo, Japan: Univ. Tokyo Press, 1981, pp. 305–323.

[34] S. R. Kadiri, P. Alku, and B. Yegnanarayana, "Analysis and classification of phonation types in speech and singing voice," *Speech Commun.*, vol. 118, pp. 33–47, Apr. 2020.

[35] S. Guruprasad, "Significance of processing regions of high signal-to-noise ratio in speech signals," Ph.D. dissertation, Dept. Comput. Sci. Eng., Indian Inst. Technol. Madras, Chennai, India, Apr. 2011.

[36] G. Seshadri and B. Yegnanarayana, "Perceived loudness of speech based on the characteristics of glottal excitation source," *J. Acoust. Soc. Amer.*, vol. 126, no. 4, pp. 2061–2071, Oct. 2009.

[37] J. Walker and P. Murphy, "A review of glottal waveform analysis," in *Progress in Nonlinear Speech Processing* (Lecture Notes in Computer Science), vol. 4391. Berlin, Germany: Springer, 2007, pp. 1–21.

[38] H. Pulakka, "Analysis of human voice production using inverse filtering, high-speed imaging, and electroglottography," M.S. thesis, Dept. Comput. Sci. Eng., Helsinki Univ. Technol., Finland, 2005. [Online]. Available: http://urn.fi/urn:nbn:fi:tkk-007925

[39] T. F. Quatieri, *Discrete-Time Speech Signal Processing*. Singapore: Pearson, 2004.

[40] K. N. Stevens, *Acoustic Phonetics*. Cambridge, MA, USA: MIT Press, 1998.

[41] A. Bouhuys, J. Mead, D. F. Proctor, and K. N. Stevens, "Pressure-flow events during singing," *Ann. New York Acad. Sci.*, vol. 155, no. 1, pp. 165–176, 1968.

[42] P. Alku, M. Airas, E. Björkner, and J. Sundberg, "An amplitude quotient based method to analyze changes in the shape of the glottal pulse in the regulation of vocal intensity," *J. Acoust. Soc. Amer.*, vol. 120, no. 2, pp. 1052–1062, Aug. 2006.

[43] M. Hirano, "Clinical examination of voice," *Disorders Hum. Commun.*, vol. 5, pp. 1–99, 1981.

[44] S. R. Kadiri, R. Prasad, and B. Yegnanarayana, "Detection of glottal closure instant and glottal open region from speech signals using spectral flatness measure," *Speech Commun.*, vol. 116, pp. 30–43, Jan. 2020.

[45] P. Fabre, "Etude comparée des glottogrammes et des phonogrammes de la voix humaine," *Annuaire Oto-Rhino Laryngologie*, vol. 75, pp. 767–775, 1958.

[46] N. Henrich, C. d'Alessandro, B. Doval, and M. Castellengo, "On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation," *J. Acoust. Soc. Amer.*, vol. 115, no. 3, pp. 1321–1332, Mar. 2004.

[47] K. N. Stevens, "Physics of laryngeal behavior and larynx modes," *Phonetica*, vol. 34, no. 4, pp. 264–279, Jul. 1977.

[48] C. T. Herbst, "Electroglottography—An update," *J. Voice*, vol. 34, no. 4, pp. 503–526, Jul. 2020.

[49] C. T. Herbst, J. Lohscheller, J. G. Švec, N. Henrich, G. Weissengruber, and W. T. Fitch, "Glottal opening and closing events investigated by electroglottography and super-high-speed video recordings," *J. Exp. Biol.*, vol. 217, no. 6, pp. 955–963, Mar. 2014.

[50] D. D. Deliyski, P. P. Petrushev, H. S. Bonilha, T. T. Gerlach, B. Martin-Harris, and R. E. Hillman,

"Clinical implementation of laryngeal high-speed videoendoscopy: Challenges and evolution," *Folia Phoniatrica Logopaedica*, vol. 60, no. 1, pp. 33–44, 2008.
[51] J. G. Švec and H. K. Schutte, "Kymographic imaging of laryngeal vibrations," *Current Opinion Otolaryngol. Head Neck Surg.*, vol. 20, no. 6, pp. 458–465, Dec. 2012.
[52] A. M. Kist, S. Dürr, A. Schützenberger, and M. Döllinger, "OpenHSV: An open platform for laryngeal high-speed videoendoscopy," *Sci. Rep.*, vol. 11, no. 1, pp. 1–12, Dec. 2021.
[53] D. D. Mehta, D. D. Deliyski, S. M. Zeitels, T. F. Quatieri, and R. E. Hillman, "Voice production mechanisms following phonosurgical treatment of early glottic cancer," *Ann. Otol., Rhinol. Laryngol.*, vol. 119, no. 1, pp. 1–9, Jan. 2010.
[54] P. Aichinger *et al.*, "Comparison of an audio-based and a video-based approach for detecting diplophonia," *Biomed. Signal Process. Control*, vol. 31, pp. 576–585, Jan. 2017.
[55] T. Murtola, J. Malinen, A. Geneid, and P. Alku, "Analysis of phonation onsets in vowel production, using information from glottal area and flow estimate," *Speech Commun.*, vol. 109, pp. 55–65, May 2019.
[56] R. L. Miller, "Nature of the vocal cord wave," *J. Acoust. Soc. Amer.*, vol. 31, no. 6, pp. 667–677, Jun. 1959.
[57] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Commun.*, vol. 11, nos. 2–3, pp. 109–118, Jun. 1992.
[58] D. Wong, J. Markel, and A. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 4, pp. 350–355, Aug. 1979.
[59] M. Airaksinen, T. Raitio, B. Story, and P. Alku, "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 3, pp. 596–607, Mar. 2014.
[60] M. Airaksinen, T. Bäckström, and P. Alku, "Quadratic programming approach to glottal inverse filtering by joint norm-1 and norm-2 optimization," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 5, pp. 929–939, May 2017.
[61] Q. Fu and P. Murphy, "Robust glottal source estimation based on joint source–filter model optimization," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 2, pp. 492–501, Mar. 2006.
[62] O. Schleusing, T. Kinnunen, B. Story, and J.-M. Vesin, "Joint source–filter optimization for accurate vocal tract estimation using differential evolution," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 8, pp. 1560–1572, Aug. 2013.
[63] H. Auvinen, T. Raitio, M. Airaksinen, S. Siltanen, B. H. Story, and P. Alku, "Automatic glottal inverse filtering with the Markov chain Monte Carlo method," *Comput. Speech Lang.*, vol. 28, no. 5, pp. 1139–1155, Sep. 2014.
[64] G. A. Alzamendi and G. Schlotthauer, "Modeling and joint estimation of glottal source and vocal tract filter by state-space methods," *Biomed. Signal Process. Control*, vol. 37, pp. 5–15, Aug. 2017.
[65] G. Fant, "The LF-model revisited. Transformations and frequency domain analysis," *Speech Transmiss. Lab. Quart. Prog. Status Rep.*, vol. 36, nos. 2–3, pp. 119–156, 1995.
[66] R. Veldhuis, "A computationally efficient alternative for the Liljencrants–Fant model and its perceptual evaluation," *J. Acoust. Soc. Amer.*, vol. 103, no. 1, pp. 566–571, 1998.
[67] S. Sahoo and A. Routray, "A novel method of glottal inverse filtering," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 7, pp. 1230–1241, Jul. 2016.
[68] B. Bozkurt, B. Doval, C. d'Alessandro, and T. Dutoit, "Zeros of Z-transform representation with application to source–filter separation in speech," *IEEE Signal Process. Lett.*, vol. 12, no. 4, pp. 344–347, Apr. 2005.
[69] T. Drugman, B. Bozkurt, and T. Dutoit, "A comparative study of glottal source estimation techniques," *Comput. Speech Lang.*, vol. 26, no. 1, pp. 20–34, Jan. 2012.
[70] T. Drugman, B. Bozkurt, and T. Dutoit, "Causal–anticausal decomposition of speech using complex cepstrum for glottal source estimation," *Speech Commun.*, vol. 53, no. 6, pp. 855–866, Jul. 2011.
[71] A. Sasou, "Glottal inverse filtering by combining a constrained LP and an HMM-based generative model of glottal flow derivative," *Speech Commun.*, vol. 104, pp. 113–128, Nov. 2018.
[72] Y.-R. Chien, D. D. Mehta, J. Guðnason, M. Zañartu, and T. F. Quatieri, "Evaluation of glottal inverse filtering algorithms using a physiologically based articulatory speech synthesizer," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 8, pp. 1718–1730, Aug. 2017.
[73] P. Mokhtari, B. Story, P. Alku, and H. Ando, "Estimation of the glottal flow from speech pressure signals: Evaluation of three variants of iterative adaptive inverse filtering using computational physical modelling of voice production," *Speech Commun.*, vol. 104, pp. 24–38, Nov. 2018.
[74] D. T. W. Chu, K. Li, J. Epps, J. Smith, and J. Wolfe, "Experimental evaluation of inverse filtering using physical systems with known glottal flow and tract characteristics," *J. Acoust. Soc. Amer.*, vol. 133, no. 5, pp. EL358–EL362, May 2013.
[75] P. Alku *et al.*, "OPENGLOT—An open environment for the evaluation of glottal inverse filtering," *Speech Commun.*, vol. 107, pp. 38–47, Feb. 2019.
[76] P. Alku, T. Murtola, J. Malinen, A. Geneid, and E. Vilkman, "Skewing of the glottal flow with respect to the glottal area measured in natural production of vowels," *J. Acoust. Soc. Amer.*, vol. 146, no. 4, pp. 2501–2509, Oct. 2019.
[77] A. Palaparthi and I. R. Titze, "Analysis of glottal inverse filtering in the presence of source–filter interaction," *Speech Commun.*, vol. 123, pp. 98–108, Oct. 2020.
[78] T. Murtola, P. Alku, J. Malinen, and A. Geneid, "Parameterization of a computational physical model for glottal flow using inverse filtering and high-speed videoendoscopy," *Speech Commun.*, vol. 96, pp. 67–80, Feb. 2018.
[79] T. Murtola and P. Alku, "Indicators of anterior–posterior phase difference in glottal opening measured from natural production of vowels," *J. Acoust. Soc. Amer.*, vol. 148, no. 2, pp. EL141–EL146, Aug. 2020.
[80] G. A. Alzamendi *et al.*, "Bayesian estimation of vocal function measures using laryngeal high-speed videoendoscopy and glottal airflow estimates: An *in vivo* case study," *J. Acoust. Soc. Amer.*, vol. 147, no. 5, pp. EL434–EL439, May 2020.
[81] A. M. Yousef, D. D. Deliyski, S. R. C. Zacharias, A. de Alarcon, R. F. Orlikoff, and M. Naghibolhosseini, "Spatial segmentation for laryngeal high-speed videoendoscopy in connected speech," *J. Voice*, Nov. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0892199720304082, doi: 10.1016/j.jvoice.2020.10.017.
[82] B. H. Story and K. Bunton, "A model of speech production based on the acoustic relativity of the vocal tract," *J. Acoust. Soc. Amer.*, vol. 146, no. 4, pp. 2522–2528, Oct. 2019.
[83] I. R. Titze, A. Palaparthi, K. Cox, A. Stark, L. Maxfield, and B. Manternach, "Vocalization with semi-occluded airways is favorable for optimizing sound production," *PLOS Comput. Biol.*, vol. 17, no. 3, Mar. 2021, Art. no. e1008744.
[84] C. Dromey, E. T. Stathopoulos, and C. M. Sapienza, "Glottal airflow and electroglottographic measures of vocal function at multiple intensities," *J. Voice*, vol. 6, no. 1, pp. 44–54, Jan. 1992.
[85] P. Alku and E. Vilkman, "A comparison of glottal voice source quantification parameters in breathy, normal and pressed phonation of female and male speakers," *Folia Phoniatrica Logopaedica*, vol. 48, no. 5, pp. 240–254, 1996.
[86] P. Alku, T. Backstrom, and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow," *J. Acoust. Soc. Amer.*, vol. 112, pp. 701–710, Aug. 2002.
[87] C. Gobl and A. N. Chasaide, "Amplitude-based source parameters for measuring voice quality," in *Proc. VOQUAL*, 2003, pp. 151–156.
[88] B. Frokjaer-Jensen and S. Prytz, "Registration of voice quality," *Bruel Kjaer Tech. Rev.*, vol. 3, pp. 3–17, 1973.
[89] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *J. Acoust. Soc. Amer.*, vol. 90, no. 5, pp. 2394–2410, 1991.
[90] I. R. Titze and J. Sundberg, "Vocal intensity in speakers and singers," *J. Acoust. Soc. Amer.*, vol. 91, no. 5, pp. 2936–2946, 1992.
[91] P. J. Murphy, "Perturbation-free measurement of the harmonics-to-noise ratio in voice signals using pitch synchronous harmonic analysis," *J. Acoust. Soc. Amer.*, vol. 105, no. 5, pp. 2866–2881, May 1999.
[92] B. Yegnanarayana, C. d'Alessandro, and V. Darsinos, "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 1, pp. 1–11, Jan. 1998.
[93] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.
[94] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *Proc. Int. Conf. Spoken Lang. Process.*, Denver, CO, USA, Sep. 2002, pp. 2037–2040.
[95] C. Manfredi, M. D'Aniello, P. Bruscaglioni, and A. Ismaelli, "A comparative analysis of fundamental frequency estimation methods with application to pathological voices," *Med. Eng. Phys.*, vol. 22, no. 2, pp. 135–147, 2000.
[96] B. Gold, N. Morgan, and D. Ellis, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. Hoboken, NJ, USA: Wiley, 2011.
[97] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 5, pp. 399–418, Oct. 1976.
[98] O. Babacan, T. Drugman, N. d'Alessandro, N. Henrich, and T. Dutoit, "A comparative study of pitch extraction algorithms on a large variety of singing sounds," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7815–7819.
[99] S. Strömbergsson, "Today's most frequently used F0 estimation methods, and their accuracy in estimating male and female pitch in clean speech," in *Proc. INTERSPEECH*, Sep. 2016, pp. 525–529.
[100] D. Jouvet and Y. Laprie, "Performance analysis of several pitch detection algorithms on simulated and real noisy speech data," in *Proc. 25th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2017, pp. 1614–1618.
[101] V. Pannala, G. Aneeja, S. R. Kadiri, and B. Yegnanarayana, "Robust estimation of fundamental frequency using single frequency filtering approach," in *Proc. INTERSPEECH*, Sep. 2016, pp. 2155–2159.
[102] T. Drugman, G. Huybrechts, V. Klimkov, and A. Moinet, "Traditional machine learning for pitch detection," *IEEE Signal Process. Lett.*, vol. 25, no. 11, pp. 1745–1749, Nov. 2018.
[103] R. Bittner, "Data-driven fundamental frequency estimation," Ph.D. dissertation, Dept. Music Performing Arts Professions, New York Univ., New York, NY, USA, 2018.
[104] S. R. Kadiri and B. Yegnanarayana, "Estimation of fundamental frequency from singing voice using harmonics of impulse-like excitation source," in *Proc. INTERSPEECH*, 2018, pp. 2319–2323.
[105] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, 1999.
[106] J. Markel, "The SIFT algorithm for fundamental

frequency estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, no. 5, pp. 367–377, Dec. 1972.
[107] D. Talkin, "Robust algorithm for pitch tracking," *Speech Coding Synth.*, vol. 495, pp. 497–518, Nov. 1995.
[108] K. Kasi and S. Zahorian, "Yet another algorithm for pitch tracking," in *Proc. ICASSP*, vol. 1, 2002, pp. 361–364.
[109] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot Int.*, vol. 5, nos. 9–10, pp. 341–345, 2002.
[110] D. J. Hermes, "Measurement of pitch by subharmonic summation," *J. Acoust. Soc. Amer.*, vol. 83, no. 1, pp. 257–264, Jan. 1988.
[111] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proc. INTERSPEECH*, Aug. 2011, pp. 1973–1976.
[112] T. Nakatani and T. Irino, "Robust and accurate fundamental frequency estimation based on dominant harmonic components," *J. Acoust. Soc. Amer.*, vol. 116, no. 6, pp. 3690–3700, Dec. 2004.
[113] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 124, pp. 1638–1652, Jun. 2008.
[114] A. de Cheveigne, "Speech F0 extraction based on Licklider's pitch perception model," in *Proc. ICPhS*, 1991, pp. 218–221.
[115] L. N. Tan and A. Alwan, "Multi-band summary correlogram-based pitch detection for noisy speech," *Speech Commun.*, vol. 55, nos. 7–8, pp. 841–856, 2013.
[116] K. S. R. Murty, B. Yegnanarayana, and M. A. Joseph, "Characterization of glottal activity from speech signals," *IEEE Signal Process. Lett.*, vol. 16, no. 6, pp. 469–472, Jun. 2009.
[117] B. Yegnanarayana and K. S. R. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 4, pp. 614–624, May 2009.
[118] K. S. R. Murty, "Significance of excitation source information for speech analysis," Ph.D. dissertation, Dept. Comput. Sci. Eng., Indian Inst. Technol. Madras, Chennai, India, Mar. 2009.
[119] B. Yegnanarayana and S. R. M. Prasanna, "Analysis of instantaneous F0 contours from two speakers mixed signal using zero frequency filtering," in *Proc. ICASSP*, 2010, pp. 5074–5077.
[120] J. M. Anand, S. Guruprasad, and B. Yegnanarayana, "Extracting formants from short segments of speech using group delay functions," in *Proc. INTERSPEECH*, Sep. 2006, pp. 1009–1012.
[121] D. Gowda, S. R. Kadiri, B. Story, and P. Alku, "Time-varying quasi-closed-phase analysis for accurate formant tracking in speech signals," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1901–1914, 2020.
[122] V. K. Mittal and B. Yegnanarayana, "Effect of glottal dynamics in the production of shouted speech," *J. Acoust. Soc. Amer.*, vol. 133, no. 5, pp. 3050–3061, May 2013.
[123] B. Yegnanarayana, S. Prasanna, R. Duraiswami, and D. Zotkin, "Processing of reverberant speech for time-delay estimation," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 6, pp. 1110–1118, Nov. 2005.
[124] B. H. V. S. N. Murthy, B. Yegnanarayana, and S. R. Kadiri, "Time delay estimation from mixed multispeaker speech signals using single frequency filtering," *Circuits, Syst., Signal Process.*, vol. 39, no. 4, pp. 1988–2005, Apr. 2020.
[125] S. Thakallapalli, S. R. Kadiri, and S. V. Gangashetty, "Spectral features derived from single frequency filter for multispeaker localization," in *Proc. Nat. Conf. Commun. (NCC)*, Feb. 2020, pp. 1–6.
[126] R. K. Swamy, K. S. R. Murty, and B. Yegnanarayana, "Determining number of speakers from multispeaker speech signals using excitation source information," *IEEE Signal Process. Lett.*, vol. 14, no. 7, pp. 481–484, Jul. 2007.
[127] B. Yegnanarayana, R. K. Swamy, and S. R. M. Prasanna, "Separation of multispeaker speech using excitation information," in *Proc. NOLISP*, 2005, pp. 11–18.
[128] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 3, pp. 972–980, May 2006.
[129] S. R. Kadiri, "A quantitative comparison of epoch extraction algorithms for telephone speech," in *Proc. ICASSP*, May 2019, pp. 6500–6504.
[130] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 4, pp. 309–319, Aug. 1979.
[131] K. S. Rao, S. R. M. Prasanna, and B. Yegnanarayana, "Determination of instants of significant excitation in speech using Hilbert envelope and group delay function," *IEEE Signal Process. Lett.*, vol. 14, no. 10, pp. 762–765, Oct. 2007.
[132] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 325–333, Sep. 1995.
[133] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, "Estimation of glottal closure instants in voiced speech using the DYPSA algorithm," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 1, pp. 34–43, Jan. 2007.
[134] M. R. P. Thomas, J. Gudnason, and P. A. Naylor, "Estimation of glottal closing and opening instants in voiced speech using the YAGA algorithm," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 82–91, Jan. 2012.
[135] A. I. Koutrouvelis, G. P. Kafentzis, N. D. Gaubitch, and R. Heusdens, "A fast method for high-resolution voiced/unvoiced detection and glottal closure/opening instant estimation of speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 2, pp. 316–328, Feb. 2016.
[136] A. P. Prathosh, T. V. Ananthapadmanabha, and A. G. Ramakrishnan, "Epoch extraction based on integrated linear prediction residual using plosion index," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 12, pp. 2471–2480, Dec. 2013.
[137] C. Ma, Y. Kamp, and L. Willems, "A Frobenius norm approach to glottal closure detection from the speech signal," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 258–265, Apr. 1994.
[138] H. W. Strube, "Determination of the instant of glottal closure from the speech wave," *J. Acoust. Soc. Amer.*, vol. 56, no. 5, pp. 1625–1629, Nov. 1974.
[139] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 8, pp. 1602–1613, Nov. 2008.
[140] S. R. Kadiri and B. Yegnanarayana, "Speech polarity detection using strength of impulse-like excitation extracted from speech epochs," in *Proc. ICASSP*, Mar. 2017, pp. 5610–5614.
[141] C. d'Alessandro and N. Sturmel, "Glottal closure instant and voice source analysis using time-scale lines of maximum amplitude," *Sadhana*, vol. 36, no. 5, pp. 601–622, Oct. 2011.
[142] V. Khanagha, K. Daoudi, and H. M. Yahia, "Detection of glottal closure instants based on the microcanonical multiscale formalism," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1941–1950, Dec. 2014.
[143] S. R. Kadiri and B. Yegnanarayana, "Epoch extraction from emotional speech using single frequency filtering approach," *Speech Commun.*, vol. 86, pp. 52–63, Feb. 2017.
[144] G. Aneeja, S. R. Kadiri, and B. Yegnanarayana, "Detection of glottal closure instants in degraded speech using single frequency filtering analysis," in *Proc. INTERSPEECH*, Sep. 2018, pp. 2300–2304.
[145] S. R. Kadiri and B. Yegnanarayana, "Determination of glottal closure instants from clean and telephone quality speech signals using single frequency filtering," *Comput. Speech Lang.*, vol. 64, Nov. 2020, Art. no. 101097.
[146] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit, "Detection of glottal closure instants from speech signals: A quantitative review," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 3, pp. 994–1006, Nov. 2011.
[147] J. Kane and C. Gobl, "Evaluation of glottal closure instant detection in a range of voice qualities," *Speech Commun.*, vol. 55, no. 2, pp. 295–314, Feb. 2013.
[148] S. R. Kadiri, P. Alku, and B. Yegnanarayana, "Comparison of glottal closure instants detection algorithms for emotional speech," in *Proc. ICASSP*, May 2020, pp. 7379–7383.
[149] P. Lieberman, "Some acoustic measures of the fundamental periodicity of normal and pathologic larynges," *J. Acoust. Soc. Amer.*, vol. 35, no. 3, pp. 344–353, Mar. 1963.
[150] D. G. Silva, L. C. Oliveira, and M. Andrea, "Jitter estimation algorithms for detection of pathological voices," *EURASIP J. Adv. Signal Process.*, vol. 2009, no. 1, Dec. 2009, Art. no. 567875.
[151] E. R. M. Abberton, D. M. Howard, and A. J. Fourcin, "Laryngographic assessment of normal voice: A tutorial," *Clin. Linguistics Phonetics*, vol. 3, no. 3, pp. 263–296, 1989.
[152] M. Rothenberg and J. J. Mahshie, "Monitoring vocal fold abduction through vocal fold contact area," *J. Speech, Lang., Hearing Res.*, vol. 31, no. 3, pp. 338–351, Sep. 1988.
[153] K. Ramesh, S. R. M. Prasanna, and D. Govind, "Detection of glottal opening instants using Hilbert envelope," in *Proc. INTERSPEECH*, Aug. 2013, pp. 44–48.
[154] A. Bouzid and N. Ellouze, "Voice source parameter measurement based on multi-scale analysis of electroglottographic signal," *Speech Commun.*, vol. 51, no. 9, pp. 782–792, Sep. 2009.
[155] M. R. P. Thomas and P. A. Naylor, "The SIGMA algorithm: A glottal activity detector for electroglottographic signals," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 8, pp. 1557–1566, Nov. 2009.
[156] R. S. Prasad and B. Yegnanarayana, "Determination of glottal open regions by exploiting changes in the vocal tract system characteristics," *J. Acoust. Soc. Amer.*, vol. 140, no. 1, pp. 666–677, Jul. 2016.
[157] Y. Bayya and D. N. Gowda, "Spectro-temporal analysis of speech signals using zero-time windowing and group delay function," *Speech Commun.*, vol. 55, no. 6, pp. 782–795, Jul. 2013.
[158] V. K. Mittal and B. Yegnanarayana, "Significance of aperiodicity in the pitch perception of expressive voices," in *Proc. INTERSPEECH*, Sep. 2014, pp. 504–508.
[159] B. S. Atal and B. E. Caspers, "Periodic repetition of multi-pulse excitation," *J. Acoust. Soc. Amer.*, vol. 74, no. S1, p. S51, Nov. 1983.
[160] B. Caspers and B. Atal, "Role of multi-pulse excitation in synthesis of natural-sounding voiced speech," in *Proc. ICASSP*, 1987, pp. 2388–2391.
[161] S. R. Kadiri and B. Yegnanarayana, "Analysis of singing voice for epoch extraction using zero frequency filtering method," in *Proc. ICASSP*, Apr. 2015, pp. 4260–4264.
[162] S. R. Kadiri, "Analysis of excitation information in expressive speech," Ph.D. dissertation, Speech Process. Lab., IIIT Hyderabad, Hyderabad, Telangana, Dec. 2018.
[163] S. R. Kadiri and B. Yegnanarayana, "Analysis of aperiodicity in artistic Noh singing voice using an impulse sequence representation of excitation source," *J. Acoust. Soc. Amer.*, vol. 146, no. 6, pp. 4446–4457, Dec. 2019.
[164] J. Kane and C. Gobl, "Wavelet maxima dispersion for breathy to tense voice discrimination," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 6, pp. 1170–1179, Jun. 2013.
[165] A. Roebel, S. Huber, X. Rodet, and G. Degottex, "Analysis and modification of excitation source characteristics for singing voice synthesis," in *Proc. ICASSP*, Mar. 2012, pp. 5381–5384.

[166] J. Lorenzo-Trueba *et al.*, "Towards glottal source controllability in expressive speech synthesis," in *Proc. INTERSPEECH*, Sep. 2012, pp. 1618–1621.

[167] E. Székely, J. Kane, S. Scherer, C. Gobl, and J. Carson-Berndsen, "Detecting a targeted voice style in an audiobook using voice quality features," in *Proc. ICASSP*, Mar. 2012, pp. 4593–4596.

[168] M. Tahon, G. Degottex, and L. Devillers, "Usual voice quality features and glottal features for emotional valence detection," in *Proc. Speech Prosody*, Shanghai, China, 2012, pp. 693–696.

[169] M. Lugger and B. Yang, "Cascaded emotion classification via psychological emotion dimensions using a large set of voice quality parameters," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2008, pp. 4945–4948.

[170] T. F. Yap, J. Epps, E. Ambikairajah, and E. H. C. Choi, "Voice source features for cognitive load classification," in *Proc. ICASSP*, May 2011, pp. 5700–5703.

[171] K. W. Godin, T. Hasan, and J. H. L. Hansen, "Glottal waveform analysis of physical task stress speech," in *Proc. INTERSPEECH*, Sep. 2012, pp. 1648–1651.

[172] E. Shriberg *et al.*, "Effects of vocal effort and speaking style on text-independent speaker verification," in *Proc. INTERSPEECH*, Sep. 2008, pp. 609–612.

[173] M. S. P. Zelinka and J. Schimmel, "Impact of vocal effort variability on automatic speech recognition," *Speech Commun.*, vol. 54, no. 6, pp. 732–742, 2012.

[174] P. Alku, J. Vintturi, and E. Vilkman, "Measuring the effect of fundamental frequency raising as a strategy for increasing vocal intensity in soft, normal and loud phonation," *Speech Commun.*, vol. 38, nos. 3–4, pp. 321–334, Nov. 2002.

[175] D. Gowda and M. Kurimo, "Analysis of breathy, modal and pressed phonation based on low frequency spectral density," in *Proc. INTERSPEECH*, Aug. 2013, pp. 3206–3210.

[176] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, "Acoustic correlates of breathy vocal quality," *J. Speech, Lang., Hearing Res.*, vol. 37, no. 4, pp. 769–778, Aug. 1994.

[177] P. Alku, H. Strik, and E. Vilkman, "Parabolic spectral parameter—A new method for quantification of the glottal flow," *Speech Commun.*, vol. 22, no. 1, pp. 67–79, Jul. 1997.

[178] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Amer.*, vol. 87, no. 2, pp. 820–857, Feb. 1990.

[179] C. Gobl and A. N. Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Commun.*, vol. 40, nos. 1–2, pp. 189–212, Apr. 2003.

[180] M. Swerts and R. Veldhuis, "The effect of speech melody on voice quality," *Speech Commun.*, vol. 33, no. 4, pp. 297–303, Mar. 2001.

[181] S. R. Kadiri and B. Yegnanarayana, "Breathy to tense voice discrimination using zero-time windowing cepstral coefficients (ZTWCCs)," in *Proc. INTERSPEECH*, Sep. 2018, pp. 232–236.

[182] S. R. Kadiri and B. Yegnanarayana, "Analysis and detection of phonation modes in singing voice using excitation source features and single frequency filtering cepstral coefficients (SFFCC)," in *Proc. INTERSPEECH*, Sep. 2018, pp. 441–445.

[183] S. R. Kadiri and P. Alku, "Mel-frequency cepstral coefficients derived using the zero-time windowing spectrum for classification of phonation types in singing," *J. Acoust. Soc. Amer.*, vol. 146, no. 5, pp. EL418–EL423, Nov. 2019.

[184] S. R. Kadiri and P. Alku, "Mel-frequency cepstral coefficients of voice source waveforms for classification of phonation types in speech," in *Proc. INTERSPEECH*, Sep. 2019, pp. 2508–2512.

[185] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Commun.*, vol. 40, nos. 1–2, pp. 227–256, 2003.

[186] D. Morrison, R. Wang, and L. C. De Silva, "Ensemble methods for spoken emotion recognition in call-centres," *Speech Commun.*, vol. 49, no. 2, pp. 98–112, Feb. 2007.

[187] M. Kockmann, L. Burget, and J. H. Černocký, "Application of speaker- and language identification state-of-the-art techniques for emotion recognition," *Speech Commun.*, vol. 53, nos. 9–10, pp. 1172–1185, Nov. 2011.

[188] A. Hassan, R. Damper, and M. Niranjan, "On acoustic emotion recognition: Compensating for covariate shift," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 7, pp. 1458–1468, Jul. 2013.

[189] A. Milton and S. T. Selvi, "Class-specific multiple classifiers scheme to recognize emotions from speech signals," *Comput. Speech Lang.*, vol. 28, no. 3, pp. 727–742, May 2014.

[190] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Audio Speech Language Process.*, vol. 13, no. 2, pp. 293–303, Mar. 2005.

[191] I. Luengo, E. Navas, and I. Hernáez, "Feature analysis and evaluation for automatic emotion identification in speech," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 490–501, Oct. 2010.

[192] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Commun.*, vol. 53, nos. 9–10, pp. 1062–1087, Nov./Dec. 2011.

[193] M. J. Gangeh, P. Fewzee, A. Ghodsi, M. S. Kamel, and F. Karray, "Multiview supervised dictionary learning in speech emotion recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 6, pp. 1056–1068, Jun. 2014.

[194] P. Gangamohan, S. R. Kadiri, and B. Yegnanarayana, "Analysis of emotional speech—A review," in *Toward Robotic Socially Believable Behaving Systems—Modeling Emotions*, vol. 1. Cham, Switzerland: Springer, 2016, pp. 205–238.

[195] B. Schuller *et al.*, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Trans. Affect. Comput.*, vol. 1, no. 2, pp. 119–131, Jul. 2010.

[196] A. Hassan and R. I. Damper, "Classification of emotional speech using 3DEC hierarchical classifier," *Speech Commun.*, vol. 54, no. 7, pp. 903–916, Sep. 2012.

[197] W. Zheng, M. Xin, X. Wang, and B. Wang, "A novel speech emotion recognition method via incomplete sparse least square regression," *IEEE Signal Process. Lett.*, vol. 21, no. 5, pp. 569–572, May 2014.

[198] A. Stuhlsatz, C. Meyer, F. Eyben, T. ZieIke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *Proc. ICASSP*, May 2011, pp. 5688–5691.

[199] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: A review," *Int. J. Speech Technol.*, vol. 15, no. 2, pp. 99–117, Jun. 2012.

[200] L. Li *et al.*, "Hybrid deep neural network-hidden Markov model (DNN-HMM) based speech emotion recognition," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, Sep. 2013, pp. 312–317.

[201] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Commun.*, vol. 53, nos. 9–10, pp. 1162–1171, Nov. 2011.

[202] R. Sun, E. Moore, and J. F. Torres, "Investigating glottal parameters for differentiating emotional categories with similar prosodics," in *Proc. ICASSP*, Apr. 2009, pp. 4509–4512.

[203] S. R. M. Prasanna and D. Govind, "Analysis of excitation source information in emotional speech," in *Proc. INTERSPEECH*, Sep. 2010, pp. 781–784.

[204] R. Sun and E. Moore, II, "A preliminary study on cross-databases emotion recognition using the glottal features in speech," in *Proc. INTERSPEECH*, Sep. 2012, pp. 1628–1631.

[205] S. R. Kadiri, P. Gangamohan, S. V. Gangashetty, P. Alku, and B. Yegnanarayana, "Excitation features of speech for emotion recognition using neutral speech as reference," *Circuits, Syst., Signal Process.*, vol. 39, no. 9, pp. 4459–4481, Sep. 2020.

[206] S. R. Kadiri and P. Alku, "Excitation features of speech for speaker-specific emotion detection," *IEEE Access*, vol. 8, pp. 60382–60391, 2020.

[207] D. Govind, S. R. M. Prasanna, and B. Yegnanarayana, "Neutral to target emotion conversion using source and suprasegmental information," in *Proc. INTERSPEECH*, Aug. 2011, pp. 2969–2972.

[208] A. K. Vuppala and S. R. Kadiri, "Neutral to anger speech conversion using non-uniform duration modification," in *Proc. ICIIS*, 2014, pp. 1–4.

[209] H. Vydana, S. Kadiri, and A. Vuppala, "Vowel-based non-uniform prosody modification for emotion conversion," *Circuits, Syst., Signal Process.*, vol. 35, no. 5, pp. 1643–1663, 2016.

[210] H. Rothgänger, G. Hauser, A. C. Cappellini, and A. Guidotti, "Analysis of laughter and speech sounds in Italian and German students," *Sci. Nature*, vol. 85, no. 8, pp. 394–402, Aug. 1998.

[211] K. P. Truong and D. A. V. Leeuwen, "Automatic detection of laughter," in *Proc. INTERSPEECH*, 2005, pp. 485–488.

[212] W. Ruch and P. Ekman, "The expressive pattern of laughter," in *Emotion, Qualia, and Consciousness*, A. W. Kaszniak, Ed. Tokyo, Japan: Word Scientific, 2001, pp. 426–443.

[213] E. E. Nwokah, H.-C. Hsu, P. Davies, and A. Fogel, "The integration of laughter and speech in vocal communication: A dynamic systems perspective," *J. Speech, Lang., Hearing Res.*, vol. 42, no. 4, pp. 880–894, Aug. 1999.

[214] C. Menezes and Y. Igarashi, "The speech laugh spectrum," in *Proc. 6th Int. Seminar Speech Prod. (ISSP)*, Dec. 2006, pp. 157–524.

[215] J.-A. Bachorowski and M. J. Owren, "Not all laughs are alike: Voiced but not unvoiced laughter readily elicits positive affect," *Psychol. Sci.*, vol. 12, no. 3, pp. 252–257, May 2001.

[216] C. A. Bickley and S. Hunnicutt, "Acoustic analysis of laughter," in *Proc. ICSLP*, 1992, pp. 927–930.

[217] V. K. Mittal, "Analysis of nonverbal speech sounds," Ph.D. dissertation, Dept. Electron. Commun. Eng., Int. Inst. Inf. Technol.-Hyderabad, Hyderabad, India, 2014.

[218] J.-A. Bachorowski, M. J. Smoski, and M. J. Owren, "The acoustic features of human laughter," *J. Acoust. Soc. Amer.*, vol. 110, no. 3, pp. 1581–1597, Sep. 2001.

[219] S. Kipper and D. Todt, "The role of rhythm and pitch in the evaluation of human laughter," *J. Nonverbal Behav.*, vol. 27, no. 4, pp. 255–272, 2003.

[220] K. S. Kumar, M. S. H. Reddy, K. S. R. Murty, and B. Yegnanarayana, "Analysis of laugh signals for detecting in continuous speech," in *Proc. INTERSPEECH*, Sep. 2009, pp. 1591–1594.

[221] A. E. Aronson and D. Bless, *Clinical Voice Disorders*. New York, NY, USA: Thieme, 2009, pp. 1–336.

[222] B. Schuller *et al.*, "A survey on perceived speaker traits: Personality, likability, pathology, and the first challenge," *Comput. Speech Lang.*, vol. 29, no. 1, pp. 100–131, Jan. 2015.

[223] A. Aronson, *Clinical Voice Disorders: An Interdisciplinary Approach*. New York, NY, USA: Thieme, 1985.

[224] N. R. Williams, "Occupational groups at risk of voice disorders: A review of the literature," *Occupational Med.*, vol. 53, no. 7, pp. 456–460, Oct. 2003.

[225] P. Carding, "Voice pathology in the United Kingdom," *Brit. Med. J.*, vol. 327, no. 7414, pp. 514–515, Sep. 2003.

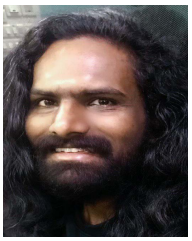
[226] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez, "Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients," *IEEE J. Trans. Bio-Med. Eng.*, vol. 58, no. 2, pp. 370–379, Feb. 2011.

[227] J. A. Gómez-García, L. Moro-Velázquez, and J. I. Godino-Llorente, "On the design of automatic voice condition analysis systems. Part I: Review of

concepts and an insight to the state of the art," *Biomed. Signal Process. Control*, vol. 51, pp. 181–199, May 2019.
[228] J. I. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, P. Gómez-Vilda, M. Blanco-Velasco, and F. Cruz-Roldán, "The effectiveness of the glottal to noise excitation ratio for the screening of voice disorders," *J. Voice*, vol. 24, no. 1, pp. 47–56, Jan. 2010.
[229] V. Parsa and D. G. Jamieson, "Identification of pathological voices using glottal noise measures," *J. Speech, Lang., Hearing Res.*, vol. 43, no. 2, pp. 469–485, 2000.
[230] D. Michaelis, T. Gramss, and H. W. Strube, "Glottal-to-noise excitation ratio—A new measure for describing pathological voices," *Acta Acustica United With Acustica*, vol. 83, no. 4, pp. 700–706, 1997.
[231] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 4, pp. 1015–1022, Apr. 2009.
[232] D. G. Silva, L. C. Oliveira, and M. Andrea, "Jitter esti mation algorithms for detection of pathological voices," *EURASIP J. Adv. Signal Process.*, vol. 2009, no. 1, pp. 9:1–9:9, Dec. 2009.
[233] M. Vasilakis and Y. Stylianou, "Voice pathology detection based eon short-term jitter estimations in running speech," *Folia Phoniatrica Logopaedica*, vol. 61, no. 3, pp. 153–170, 2009.
[234] Y. Zhang, J. J. Jiang, L. Biazzo, and M. Jorgensen, "Perturbation and nonlinear dynamic analyses of voices from patients with unilateral laryngeal paralysis," *J. Voice*, vol. 19, no. 4, pp. 519–528, Dec. 2005.
[235] V. Parsa and D. G. Jamieson, "Acoustic discrimination of pathological voice," *J. Speech, Lang., Hearing Res.*, vol. 44, no. 2, pp. 327–339, 2001.
[236] J. R. Orozco-Arroyave *et al.*, "Characterization methods for the detection of multiple voice disorders: Neurological, functional, and laryngeal diseases," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 6, pp. 1820–1828, Nov. 2015.
[237] J. Mekyska *et al.*, "Robust and complex approach of pathological speech signal analysis," *Neurocomputing*, vol. 167, no. 1, pp. 94–111, 2015.
[238] Y. Qi and R. E. Hillman, "Temporal and spectral estimations of harmonics-to-noise ratio in human voice signals," *J. Acoust. Soc. Amer.*, vol. 102, no. 1, pp. 537–543, Jul. 1997.
[239] J.-W. Lee, S. Kim, and H.-G. Kang, "Detecting pathological speech using contour modeling of harmonic-to-noise ratio," in *Proc. ICASSP*, May 2014, pp. 5969–5973.
[240] H. Kasuya, S. Ogawa, K. Mashima, and S. Ebihara, "Normalized noise energy as an acoustic measure to evaluate pathologic voice," *J. Acoust. Soc. Amer.*, vol. 80, no. 5, pp. 1329–1334, Nov. 1986.
[241] C. R. Watts and S. N. Awan, "Use of spectral/cepstral analyses for differentiating normal from hypofunctional voices in sustained vowel and continuous speech contexts," *J. Speech, Lang., Hearing Res.*, vol. 54, no. 6, pp. 1525–1537, Dec. 2011.
[242] J. I. Godino-Llorente and P. Gomez-Vilda, "Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 2, pp. 380–384, Feb. 2004.
[243] S. R. Kadiri, R. Kethireddy, and P. Alku, "Parkinson's disease detection from speech using single frequency filtering cepstral coefficients," in *Proc. INTERSPEECH*, 2020, pp. 4971–4975.
[244] J. I. Godino-Llorente, S. Aguilera-Navarro, and P. G. Vilda, "LPC, LPCC and MFCC parameterisation applied to the detection of voice impairments," in *Proc. INTERSPEECH*, 2000, pp. 965–968.
[245] J. C. Saldanha, T. Ananthakrishna, and R. Pinto, "Vocal fold pathology assessment using mel-frequency cepstral coefficients and linear predictive cepstral coefficients features," *J. Med. Imag. Health Informat.*, vol. 4, no. 2, pp. 168–173, 2014.
[246] M. A. Little, D. A. E. Costello, and M. L. Harries, "Objective dysphonia quantification in vocal fold paralysis: Comparing nonlinear with classical measures," *J. Voice*, vol. 25, no. 1, pp. 21–31, 2011.
[247] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. E. Costello, and I. M. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *BioMed. Eng. OnLine*, vol. 6, no. 1, p. 23, 2007.
[248] J. Orozco, J. F. Vargas-Bonilla, and E. Delgado-Trejos, "Acoustic analysis and non linear dynamics applied to voice pathology detection: A review," *Recent Patents Signal Process.*, vol. 2, no. 2, pp. 96–107, 2012.
[249] P. Henriquez, J. B. Alonso, M. A. Ferrer, C. M. Travieso, J. I. Godino-Llorente, and F. Diaz-de-Maria, "Characterization of healthy and pathological voice through measures based on nonlinear dynamics," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 6, pp. 1186–1195, Aug. 2009.
[250] V. Péan, M. Ouayoun, C. Fugain, B. Meyer, and C. Chouard, "A fractal approach to normal and pathological voices," *Acta Otolaryngol.*, vol. 120, no. 2, pp. 222–224, 2000.
[251] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity," *J. Roy. Soc. Interface*, vol. 8, no. 59, pp. 842–855, Jun. 2011.
[252] G. Vaziri, F. Almasganj, and R. Behroozmand, "Pathological assessment of patients' speech signals using nonlinear dynamical analysis," *Comput. Biol. Med.*, vol. 40, no. 1, pp. 54–63, Jan. 2010.
[253] C. M. Travieso, J. B. Alonso, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, E. Nöth, and A. G. Ravelo-García, "Detection of different voice diseases based on the nonlinear characterization of speech signals," *Expert Syst. Appl.*, vol. 82, pp. 184–195, Oct. 2017.
[254] A. Giovanni, M. Ouaknine, and J.-M. Triglia, "Determination of largest Lyapunov exponents of vocal signal: Application to unilateral laryngeal paralysis," *J. Voice*, vol. 13, no. 3, pp. 341–354, Sep. 1999.
[255] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 4, pp. 884–893, Apr. 2010.
[256] J. A. Gómez-García, L. Moro-Velázquez, and J. I. Godino-Llorente, "On the design of automatic voice condition analysis systems. Part II: Review of speaker recognition techniques and study on the effects of different variability factors," *Biomed. Signal Process. Control*, vol. 48, pp. 128–143, Feb. 2019.
[257] M. Airas, "TKK Aparat: An environment for voice inverse filtering and parameterization," *Logopedics Phoniatrics Vocol.*, vol. 33, no. 1, pp. 49–64, 2008.
[258] J. Kane and C. Gobl, "Identifying regions of non-modal phonation using features of the wavelet transform," in *Proc. INTERSPEECH*, Aug. 2011, pp. 177–180.
[259] G. Degottex, A. Roebel, and X. Rodet, "Phase minimization for glottal model estimation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 5, pp. 1080–1090, Jul. 2011.
[260] S. Huber, A. Roebel, and G. Degottex, "Glottal source shape parameter estimation using phase minimization variants," in *Proc. INTERSPEECH*, Portland, OR, USA, Sep. 2012, pp. 1644–1647.
[261] N. P. Narendra and P. Alku, "Automatic intelligibility assessment of dysarthric speech using glottal parameters," *Speech Commun.*, vol. 123, pp. 1–9, Oct. 2020.
[262] N. P. Narendra and P. Alku, "Dysarthric speech classification from coded telephone speech using glottal features," *Speech Commun.*, vol. 110, pp. 47–55, Jul. 2019.
[263] V. M. Espinoza, M. Zañartu, J. H. Van Stan, D. D. Mehta, and R. E. Hillman, "Glottal aerodynamic measures in women with phonotraumatic and nonphonotraumatic vocal hyperfunction," *J. Speech Lang., Hearing Res.*, vol. 60, no. 8, pp. 2159–2169, 2017.
[264] K. Szklanny and P. Wrzeciono, "The application of a genetic algorithm in the noninvasive assessment of vocal nodules in children," *IEEE Access*, vol. 7, pp. 44966–44976, 2019.
[265] Y. Wu, C. Zhou, Z. Fan, D. Wu, X. Zhang, and Z. Tao, "Investigation and evaluation of glottal flow waveform for voice pathology detection," *IEEE Access*, vol. 9, pp. 30–44, 2021.
[266] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The Munich versatile and fast open-source audio feature extractor," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 1459–1462.
[267] M. K. Reddy *et al.*, "The automatic detection of heart failure using speech signals," *Comput. Speech Lang.*, vol. 69, Sep. 2021, Art. no. 101205.
[268] M. K. Reddy, P. Alku, and K. S. Rao, "Detection of specific language impairment in children using glottal source features," *IEEE Access*, vol. 8, pp. 15273–15279, 2020.
[269] N. P. Narendra, M. Airaksinen, B. Story, and P. Alku, "Estimation of the glottal source from coded telephone speech using deep neural networks," *Speech Commun.*, vol. 106, pp. 95–104, Jan. 2019.
[270] M. Airaksinen, T. Raitio, and P. Alku, "Noise robust estimation of the voice source using a deep neural network," in *Proc. ICASSP*, Apr. 2015, pp. 5137–5141.
[271] M. Airaksinen and P. Alku, "Effects of training data variety in generating glottal pulses from acoustic features with DNNs," in *Proc. INTERSPEECH*, 2017, pp. 3946–3950.
[272] K. Han and D. Wang, "Neural network based pitch tracking in very noisy speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 2158–2168, Dec. 2014.
[273] S. Xu and H. Shimodaira, "Direct F0 estimation with neural-network-based regression," in *Proc. INTERSPEECH*, Sep. 2019, pp. 1995–1999.
[274] B. Gfeller, C. Frank, D. Roblek, M. Sharifi, M. Tagliasacchi, and M. Velimirovic, "Pitch estimation via self-supervision," in *Proc. ICASSP*, May 2020, pp. 3527–3531.
[275] P. Verma and R. W. Schafer, "Frequency estimation from waveforms using multi-layered neural networks," in *Proc. INTERSPEECH*, 2016, pp. 2165–2169.
[276] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *Proc. ICASSP*, 2018, pp. 161–165.
[277] M. Airaksinen, L. Juvela, P. Alku, and O. Räsänen, "Data augmentation strategies for neural network F0 estimation," in *Proc. ICASSP*, 2019, pp. 6485–6489.
[278] S. Yang, Z. Wu, B. Shen, and H. Meng, "Detection of glottal closure instants from speech signals: A convolutional neural network based method," in *Proc. INTERSPEECH*, 2018, pp. 317–321.
[279] M. Goyal, V. Srivastava, and A. P. Srivastava, "Detection of glottal closure instants from raw speech using convolutional neural networks," in *Proc. INTERSPEECH*, 2019, pp. 1591–1595.
[280] L. Ardaillon and A. Roebel, "GCI detection from raw speech using a fully-convolutional network," in *Proc. ICASSP*, 2020, pp. 6739–6743.
[281] M. Reddy, T. Mandal, and K. S. Rao, "Glottal closure instants detection from pathological acoustic speech signal using deep learning," in *Proc. Mach. Learn. Health Workshop*, 2018, pp. 1–6.
[282] M. G. Reddy, K. S. Rao, and P. P. Das, "Glottal closure instants detection from speech signal by deep features extracted from raw speech and linear prediction residual," in *Proc. INTERSPEECH*, Sep. 2019, pp. 156–160.
[283] J. Matoušek and D. Tihelka, "Classification-based detection of glottal closure instants from speech

signals," in *Proc. INTERSPEECH*, 2017, pp. 3053–3057.
[284] J. Matousek and D. Tihelka, "Glottal closure instant detection from speech signal using voting classifier and recursive feature elimination," in *Proc. INTERSPEECH*, 2018, pp. 2112–2116.
[285] J. Matoušek and D. Tihelka, "Using extreme gradient boosting to detect glottal closure instants in speech signal," in *Proc. ICASSP*, 2019, pp. 6515–6519.
[286] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1264–1271, May 2012.
[287] J. R. Orozco-Arroyave *et al.*, "NeuroSpeech: An open-source software for Parkinson's speech analysis," *Digit. Signal Process.*, vol. 77, pp. 207–221, Jun. 2018.
[288] T. Arias-Vergara, J. C. Vasquez-Correa, J. R. Orozco-Arroyave, P. Klumpp, and E. Nöth, "Unobtrusive monitoring of speech impairments of Parkinson's disease patients through mobile devices," in *Proc. ICASSP*, 2018, pp. 6004–6008.
[289] E. Vaiciukynas, A. Verikas, A. Gelzinis, and M. Bacauskiene, "Detecting Parkinson's disease from sustained phonation and speech signals," *PLoS ONE*, vol. 12, no. 10, Oct. 2017, Art. no. e0185613.
[290] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, and E. Nöth, "A multitask learning approach to assess the dysarthria severity in patients with Parkinson's disease," in *Proc. INTERSPEECH*, 2018, pp. 456–460.
[291] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, B. Eskofier, J. Klucken, and E. Nöth, "Multimodal assessment of Parkinson's disease: A deep learning approach," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 4, pp. 1618–1630, Jul. 2019.
[292] J. C. Vásquez-Correa, J. R. Orozco-Arroyave, and E. Nöth, "Convolutional neural network to model articulation impairments in patients with Parkinson's disease," in *Proc. INTERSPEECH*, 2017, pp. 314–318.
[293] E. A. Belalcázar-Bolaños, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, T. Haderlein, and E. Nöth, "Glottal flow patterns analyses for Parkinson's disease detection: Acoustic and nonlinear approaches," in *Text, Speech, and Dialogue*. Cham, Switzerland: Springer, 2016, pp. 400–407.
[294] P. Gómez-Vilda, D. Palacios-Alonso, V. Rodellar-Biarge, A. Álvarez-Marquina, V. Nieto-Lluis, and R. Martínez-Olalla, "Parkinson's disease monitoring by biomechanical instability of phonation," *Neurocomputing*, vol. 255, pp. 3–16, Sep. 2017.
[295] L. Juvela, B. Bollepalli, V. Tsiaras, and P. Alku, "GlotNet—A raw waveform model for the glottal excitation in statistical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 27, no. 6, pp. 1019–1030, Jun. 2019.
[296] N. P. Narendra, B. Schuller, and P. Alku, "The detection of Parkinson's disease from speech using voice source information," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 1925–1936, 2021.
[297] G. Fant, *Acoustic Theory of Speech Production*. The Hague, The Netherlands: Mouton, 1960.
[298] J. L. Flanagan, *Speech Analysis Synthesis and Perception*, 2nd ed. New York, NY, USA: Springer-Verlag, 1972.
[299] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019.

ABOUT THE AUTHORS



Sudarsana Reddy Kadiri (Member, IEEE) received the B.Tech. degree from Jawaharlal Nehru Technological University (JNTU), Hyderabad, India, in 2011, with a specialization in electronics and communication engineering (ECE), and the Ph.D. degree from the Department of ECE, International Institute of Information Technology, Hyderabad (IIIT-H), Hyderabad, in 2018.

He was a Teaching Assistant for several courses at IIIT-H from 2012 to 2018. Since 2019, he has been involved in teaching and mentoring activities at Aalto University, Espoo, Finland, where he is currently a Postdoctoral Researcher with the Department of Signal Processing and Acoustics. His research interests include signal processing, speech analysis, speech synthesis, paralinguistics, affective computing, voice pathologies, machine learning, and auditory neuroscience. He has published over 50 research papers in peer-reviewed journals and conferences in these areas.

Dr. Kadiri was awarded the Tata Consultancy Services (TCS) Fellowship for his Ph.D. degree.

Paavo Alku (Fellow, IEEE) received the M.Sc., Lic.Tech., and Dr.Sc.(Tech.) degrees from the Helsinki University of Technology, Espoo, Finland, in 1986, 1988, and 1992, respectively.

He was an Assistant Professor with the Asian Institute of Technology, Bangkok, Thailand, in 1993, and an Assistant Professor and a Professor with the University of Turku, Turku, Finland, from 1994 to 1999. He is currently a Professor of speech communication technology with Aalto University, Espoo. He has published more than 200 peer-reviewed journal articles and more than 200 peer-reviewed conference papers. His research interests include the analysis and parameterization of speech production, statistical parametric speech synthesis, spectral modeling of speech, speech-based biomarking of human health, and cerebral processing of speech.

Dr. Alku is a Fellow of the International Speech Communications Association (ISCA). He has served as an Academy Professor assigned by the Academy of Finland from 2015 to 2019. He is also an Associate Editor of *Journal of the Acoustical Society of America*.

B. Yegnanarayana (Life Fellow, IEEE) received the B.Sc. degree from Andhra University, Visakhapatnam, India, in 1961, the B.E., M.E., and Ph.D. degrees from the Indian Institute of Science (IISc), Bengaluru, India, in 1964, 1966, and 1974, respectively, and the D.Sc. degree (Honoris Causa) from Jawaharlal Nehru Technological University, Anantapur, India, in February 2019.

He was an Institute Professor from 2012 to 2016 and a Professor and the Microsoft Chair from 2006 to 2012 with the International Institute of Information Technology, Hyderabad (IIIT-H), Hyderabad, India. He was a Professor from 1980 to 2006 and the Head of the Computer Science and Engineering (CSE) Department from 1985 to 1989 with Indian Institute of Technology (IIT) Madras (IITM), Chennai, India, a Visiting Associate Professor at Carnegie Mellon University (CMU), Pittsburgh, PA, USA, from 1977 to 1980, and a member of the Faculty of IISc from 1966 to 1978. He was a Professor Emeritus with Birla Institute of Technology & Science Pilani (BITS-Pilani), Hyderabad, in 2016. He was a Visiting Professor with IIT Dharwad, Dharwad, India, and Carnegie Mellon University Africa (CMU Africa), Kigali, Rwanda, in 2019. He is currently an Adjunct Faculty with IIT Tirupati, Tirupati, India, a Distinguished Professor with IIT Hyderabad, Hyderabad, and a Distinguished Adjunct Professor with the International Institute of Information Technology (IIIT), Naya Raipur, India. He is also the INSA Senior Scientist with IIIT-H. He is the author of the book *Artificial Neural Networks* (Prentice-Hall of India, 1999). He has supervised 36 Ph.D. and 42 M.S. theses at IISc, IITM, and IIIT-H. His research interests are in signal processing, speech, image processing, and neural networks. He has published over 400 papers in these areas.

Dr. Yegnanarayana is a Fellow of the Indian National Academy of Engineering (INAE), the Indian National Science Academy (INSA), the Indian Academy of Sciences (IASc), and the International Speech Communications Association (ISCA). He was a recipient of the Third IETE Prof. S. V. C. Aiya Memorial Award in 1996. He received the Prof. S. N. Mitra Memorial Award for the year 2006 from INAE. He was awarded the 2013 Distinguished Alumnus Award from IISc Bangalore. He was awarded "The Sayed Husain Zaheer Medal (2014)" of INSA in 2014. He received the Prof. Rais Ahmed Memorial Lecture Award from the Acoustical Society of India in 2016. He was the General Chair of Interspeech2018 held in Hyderabad in September 2018. He was an Associate Editor of IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING from 2003 to 2006. He is also an Associate Editor of *Journal of the Acoustical Society of America*.