Acoustic analysis of trill sounds

N. Dhananjaya, a) B. Yegnanarayana, and Peri Bhaskararao
International Institute of Information Technology, Hyderabad-500032, India

(Received 12 July 2011; revised 24 January 2012; accepted 26 January 2012)

In this paper, the acoustic–phonetic characteristics of steady apical trills—trill sounds produced by the periodic vibration of the apex of the tongue—are studied. Signal processing methods, namely, zero-frequency filtering and zero-time liftering of speech signals, are used to analyze the excitation source and the resonance characteristics of the vocal tract system, respectively. Although it is natural to expect the effect of trilling on the resonances of the vocal tract system, it is interesting to note that trilling influences the glottal source of excitation as well. The excitation characteristics derived using zero-frequency filtering of speech signals are glottal epochs, strength of impulses at the glottal epochs, and instantaneous fundamental frequency of the glottal vibration. Analysis based on zero-time liftering of speech signals is used to study the dynamic resonance characteristics of vocal tract system during the production of trill sounds. Qualitative analysis of trill sounds in different vowel contexts, and the acoustic cues that may help spotting trills in continuous speech are discussed.

© 2012 Acoustical Society of America. [http://dx.doi.org/10.1121/1.3688470]

PACS number(s): 43.72.Ar [SSN]

I. INTRODUCTION

Trills are a stricture type (Catford, 1977, p. 127), characterized primarily by ‘the vibration of one speech organ against another, driven by aerodynamic conditions’ (Ladefoged and Maddieson, 1996, p. 217). The most common trills involve the tip of the tongue vibrating against a contact point in the dental/alveolar region, and are called apical trills (McGowan, 1992; Ladefoged and Maddieson, 1996). Apical trills are the most common variety of trills among Indian languages. The objective of this paper is to derive the acoustic characteristics of apical trills from the speech signal. The effect of trilling on the glottal source of excitation and on the resonance characteristics of the vocal tract system is studied. The effect of different vowel context on the resonance characteristics of apical trills is also studied. The phonetic convention of the vowel context is indicated with a superscript to the base phoneme, such as [rə] to denote a voiced apical trill [r] adjacent to the vowel [a]. In this paper, characteristics of the voiced apical trill [r] are studied in the context of three different vowels [a], [i], and [u].

The phonological aspects of trills, such as their occurrences in various world languages, and their relationship with other phonemes, have been reported by Maddieson (1984), Ladefoged and Maddieson (1996), and Ruhlen (1987). The production of an apical trill involves satisfying several articulatory, as well as aerodynamic constraints. The articulatory constraints concern the lingual and vocal tract configurations. The aerodynamic constraints concern the maintenance of the right amount of tension at the apex (tongue tip) and the requisite volume velocity of air flow through the stricture, which are essential for the initiation and sustenance of the apical vibration. The articulatory mechanics of tongue-tip vibration have been described by Catford (1977), Ladefoged and Maddieson (1996), Recasens (1991), and Spajic et al. (1996), and modeled by McGowan (1992). The aerodynamic characteristics and the phonological patterns of trills across languages are studied in detail by Solé (2002). Estimates of the transglottal (subglottal and supraglottal) pressure values with respect to the atmospheric pressure, and the pressure gradient across the lingual constriction, essential for initiating and sustaining voicing and trilling, respectively, have been obtained based on oropharyngeal pressure and oral air flow measurements. Solé (2002) has also studied some of the phonological patterns, such as the absence of nasal trills, preference for voiced trills, alternation and co-occurrence of trilling and frication, and trill devoicing, from an aerodynamic point of view.

Ladefoged and Maddieson (1996) have reported that acoustic trills in linguistic use usually consist of two to five periods, whereas apical trills typically consist of two to three periods of vibration (geminate occurrences may be longer). Based on spectrographic measurements made for Finnish and Russian apical trills, Ladefoged and Maddieson (1996) report a typical trill period of 50 ms (open and closed phases each of 25 ms duration), and hence a trilling rate of about 20 cycles in a second. Lindau (1985) reports a mean trilling rate of 25 Hz (18–33 Hz) measured over 25 speakers from seven different languages. An estimate of the trilling frequency of the tongue tip based on mechanical lumped element modeling of trill aerodynamics is given by McGowan (1992). The trilling rate of the tongue tip can be estimated using the formula

$$ F_r = \frac{1}{2\pi\sqrt{MC}} $$

where $M$ is the mass of the tongue tip, estimated to be ~1 g (by assuming an approximate surface area of the tongue tip involved in vibration to be 1 cm²), and $C$ is the mechanical compliance (inverse of stiffness) per unit area of the tongue tip (approximated to be $3 \times 10^{-5}$ cm³/dyne) (McGowan, 1992; Stevens, 1999).

Several studies on phonemic trills in Spanish have been reported, such as categorization of the Spanish dialect continuum (Lipski, 1994), acoustic correlates to distinguish one...
phonemic trill from another (Colantoni, 2006), and acoustic characterization of trills (Henriksen and Willis, 2010). The acoustic correlates studied mostly concern the number of occlusions (or trill cycles) and the duration of the trill. Manual measurements of these parameters are made by observing the acoustic waveforms and the spectrograms. Based on these acoustic parameters, a detailed statistical analysis of trills in Spanish in terms of its sociolinguistic implications has been made by Diaz-Campos (2008) and Henriksen and Willis (2010). Studies made in deriving the acoustic—phonetic characteristics of trills from speech data are limited by the standard spectrographic tools derived from short-time spectral analysis. The dynamic nature of the vocal tract system during production of trills is likely to have an effect on the excitation source due to coupling of the excitation source and the vocal tract system.

Currently available signal processing techniques may not be adequate to study the dynamic source and system characteristics of the trill sounds. In this paper, some recently proposed signal processing techniques, together with conventional methods, are examined for the study of dynamic characteristics of the excitation source and the vocal tract system resonances. As discussed in the later sections, the new analysis techniques provide an interpretation of the results in terms of production characteristics of the trill sounds.

Murty and Yegnanarayana (2008) have proposed an approach based on the zero-frequency filtering (ZFF) of speech signals for analysis of impulse-like characteristics in the excitation source. The ZFF-based approach gives a simple but effective method for detection of the instants of glottal closure (GClS) or epochs in voiced sounds. The method also provides a measure of the strength of excitation at the epochs and the instantaneous fundamental frequency (Yegnanarayana and Murty, 2009; Murty et al., 2009). The regions around the GClS have high signal-to-noise ratio (SNR), and hence are useful as anchor points for analysis of the characteristics of the vocal tract system. Traditional short-time spectral analysis of speech involves processing the signal in blocks of 10–30 ms. Magnitude spectrum computed over block sizes less than 10 ms is not useful for analysis of the vocal tract system, due to issues caused by time-frequency resolution. Recently, a new technique called zero-time liftering (ZTL) of speech signals for analysis of resonance characteristics of the vocal tract system has been proposed (Dhananjaya, 2011). The ZTL technique provides high resolution of the spectral characteristics in temporal domain. Multiplication of the speech signal in time domain by an impulse-like window function provides the high temporal resolution. This is called the “liftering” operation analogous to the operation done in cepstrum analysis of speech (Bogert et al., 1963). Good resolution of the spectral characteristics in frequency domain is achieved using the group delay analysis (Yegnanarayana, 1978; Yegnanarayana and Murthy, 1992; Joseph et al., 2006), where group delay is defined as the negative derivative of the phase of the Fourier transform of the signal (Oppenheim and Schafer, 1975, p. 19).

The paper is organized as follows: Production characteristics of lingual trills, primarily the tongue-tip trills, are described in Sec. II. Section III describes the zero-frequency filtering-based analysis for extracting the features of the excitation source from the speech signal. Section IV describes the zero-time liftering technique for analyzing the spectral characteristics of trills. Analysis of trill sounds in terms of excitation source and vocal tract resonance characteristics is given in Sec. V. Characteristics of trills in different vowel contexts are examined. The acoustic features for spotting trills in continuous speech are discussed. Characteristics of the voiceless apical trill, as well as the voiced and voiceless labial trills are also examined in this section. Similarities and/or contrasts between apical and labial trills are discussed. A summary of the paper along with directions for further research is given in Sec. VI.

II. PRODUCTION CHARACTERISTICS OF APICAL TRILLS

In the production of an apical trill the apex is voluntarily positioned by the speaker to make a contact (Ladefoged and Maddieson, 1996, p. 218) with the corresponding upper articulator. Almost immediately, the pulmonic egress stream that is flowing into the oral cavity increases the pressure gradient across the stricture. Due to the fine interaction between “tongue-tension and volume-velocity of the air-flow” (Catford, 1977, p. 127), the apical stricture gets broken and the apex falls down to some extent releasing part of the positive pressure gradient in the oral cavity. Then due to the Bernoulli effect, the apex recoils to meet the upper articulator and forms the next event of stricture. Thus, the closure—opening cycle repeats itself a few times, and the total number of such cycles constitutes the complete trill. One such closure—opening cycle may be referred to as a “trill cycle,” and the different phases of an apical trill are depicted in Fig. 1. The typical rate of trilling of the tongue is ~20–30 Hz, and can be measured from the acoustic waveform or the spectrogram (McGowan, 1992; Stevens, 1999; Lindau, 1985; Ladefoged et al., 1977).

Trill sounds usually have at least two trill cycles for them to be discriminated from another category of sounds, namely taps ([ɾ]), which have one single movement of the tongue from any arbitrary position to the roof of the oral cavity and back, analogous to a trill cycle. Lindau (1985, p. 166) observed that “from an acoustic point of view, a trill can be regarded as a series of taps.” On the other hand, as observed by Recasens (1991), and Recasens and Pallars (1999), an apical trill differs from an apical tap in the overall tongue body configuration. Based on electropalatographic
data, they observed that apical trills have a lower predorsum and a retracted postdorsum positions as compared to that of apical taps. Each of the constituent trill cycle is produced due to the Bernoulli effect, than due to voluntary movement of the apex of the tongue as in the case of a tap. But it should be noted here that the production of a trill still requires voluntarily maintaining the correct tongue body position, the right amount of tension (or stiffness) at the apex, and the requisite volume velocity of air flow across the stricture. In continuous speech two to three trill cycles are common (Lindau, 1985; Ladefoged and Maddieson, 1996; Henriksen and Willis, 2010), whereas in isolation they can be produced as a steady sustained sound with several (> 3) trill cycles.

III. ZERO FREQUENCY FILTERING FOR ANALYSIS OF EXCITATION CHARACTERISTICS

Recently a zero-frequency filtering method was proposed for extracting the impulse-like characteristics of the excitation source from the speech signal, such as the GCIs, instantaneous fundamental frequency ($F_0$), and strength of excitation (Murty and Yegnanarayana, 2008; Yegnanarayana and Murty, 2009). The idea behind filtering the speech signal at zero frequency is that the effect of an impulse-like excitation source is equally felt throughout the spectrum, including around the zero frequency, whereas the vocal tract information is predominantly concentrated around the formant peaks. The method involves filtering the speech signal through a cascade of zero-frequency resonators. A zero-frequency resonator is an all-pole system with two poles at $z = +1$ in the $z$-plane, which is equivalent to a sequence of two cumulative sum operations in time-domain. This leads to a polynomial-type growth/decay of the output signal. The polynomial-type growth/decay can be removed by a trend removal operation, which involves subtracting the local mean from the signal at each time instant (Murty and Yegnanarayana, 2008; Yegnanarayana and Murty, 2009). The resulting signal is referred to as the zero-frequency filtered signal. The positive zero crossings (negative to positive) of the filtered signal correspond to the instants of glottal closure, also referred to as epochs. The slope of the filtered signal around the epochs gives a measure of the strength of excitation.

Figure 2(a) shows an example of a steady or prolonged trill sound uttered in isolation as a CV (consonant–vowel) unit [r$\,\text{[a]}$] in the context of the vowel [a]. Note that in this paper, the repetition of the phone label [r$\,\text{[a]}$] is used to denote the prolonged utterance of the phone, and the superscript denotes the vowel context. The output of cascade of two zero-frequency resonators, and the ZFF signal obtained after trend removal are shown in Figs. 2(c) and 2(d), respectively. The epoch locations given by the positive zero crossings of the ZFF signal [Fig. 2(d)] are shown in Fig. 2(a) by downward-pointing arrows. The strength of excitation (measured at the epoch locations as the slope of the ZFF signal) and the instantaneous fundamental frequency (measured as the reciprocal of the time interval between adjacent epochs) are shown in Figs. 2(e) and 2(f), respectively. The epoch locations occur at regular instants in most of the voiced regions (0.1–0.8 s), governed by strong impulse-like excitations imparted at the instants of glottal closure. The measured strengths of excitation at these epoch locations are also large for the voiced regions. In the silence regions (0–0.1 s and 0.82–0.92 s) and voiceless regions (not shown), the epoch locations occur at irregular instants due to lack of any regular impulse-like excitations, and the excitation strengths measured at these epochs are significantly lower compared to those in the voiced regions (Murty and Yegnanarayana, 2008). A simple threshold on the excitation strength helps to isolate the regions of voiced excitation. The voiced/nonvoiced decision based on the excitation strength is shown in Fig. 2(d). The ZFF-based method for extraction of epoch locations and their strengths has been shown to be robust against additive noise (Murty and Yegnanarayana, 2008; Dhananjaya and Yegnanarayana, 2010).

Trilling of the tongue tip affects the measured strength of the glottal excitation, as can be seen in Fig. 2(e). The strength of excitation varies within a trill cycle, and it is less during the closed phase as compared to the open phase. This may be due to the loading of the vocal folds by the closing of the oral cavity. It is also seen from Fig. 2(f) that the instantaneous fundamental frequency varies due to the trilling of the tongue tip. In contrast, the contours of the excitation strength and the instantaneous $F_0$ are relatively smooth within the vowel region (0.65–0.8 s in Fig. 2). A portion of the trill region of the waveform in Fig. 2(a) is shown expanded in Fig. 3 to show the details of the excitation characteristics of the trill. The fundamental frequency $F_0$ seems to reach a minimum value in the closed phase just before the release of apical contact, and increases gradually as the apical contact is forced open. At the same time, the $F_0$ movement toward the point of apical contact is not smooth, which probably hints toward a faster recoil of the apex than the opening as is observed in the case of vocal folds.

FIG. 2. (Color online) Zero frequency analysis of a steady or sustained apical trill produced as an isolated CV (consonant–vowel) [r$\,\text{[a]}$]. (a) Speech waveform and the estimated epoch locations shown by downward arrows, (b) wideband (WB) spectrogram, (c) output of a cascade of two zero-frequency resonators, (d) ZFF signal after trend removal along with the V/NV decision, (e) excitation strength, and (f) instantaneous fundamental frequency ($F_0$).
evidence from other modalities such as eletroglottograph and/or magnetic resonance imaging (MRI) may be required to comment on the behavior of \( F_0 \) toward the point of apical contact. As per the aerodynamics of a pair of stretched membranes, analogous to the vocal folds, a minimum pressure gradient across the membrane is essential, depending on the mass and tension of the membrane, for the membrane to flutter or vibrate (Sole´, 2002; Herman, 2007). As the subglottal pressure builds up behind the closed glottis, the pressure gradient between the subglottal and supraglottal air pressures increases, forcing the vocal folds to open with a burst of air rushing across the glottis. This results in a temporary reduction of the subglottal pressure, and hence reduction in the pressure gradient, allowing the vocal folds to recoil back to their initial stretched position due to the inherent myoelastic tension in the membrane. This cycle repeats itself. The pressure gradient has a direct relationship with the rate of vibration of the vocal folds, meaning, higher the gradient, higher the rate of vibration (van den Berg, 1957; Fant, 1960, p. 266). When there is a supraglottal oral constriction, as in the case of closed phase of the trills, the supraglottal oral pressure increases, reducing the pressure gradient across the glottis. This in turn may lead to the reduction in the rate of vibration of the vocal folds temporarily, which increases again gradually as the oral constriction is released, causing an increase in the pressure gradient across the glottis. Such a phenomenon is also reported by van den Berg (1957) and Fant (1960, p. 266) during the production of voiced occlusions. Assuming a constant lung effort and a constant tension in the vocal fold membranes, vibration of the vocal folds is directly influenced by the pressure gradient across the glottis, which in turn is influenced by the trilling of the tongue tip. The pressure gradient decreases during the closed phase, thus reducing the rate of vibration of the vocal folds. The pressure gradient increases during the open phase, resulting in an increase in the rate of vibration of the vocal folds. Reduction in the excitation strength during the closed phase can be explained by the reduced air flow due to a reduction in the transglottal pressure gradient, as observed by Westbury (1983) in the case of voiced occlusions. It can be seen from Figs. 2(e) and 2(f) that the fluctuating pattern in the excitation strength and in the instantaneous \( F_0 \) repeats itself over a few glottal cycles.

IV. ZERO-TIME LIFTERING FOR ANALYSIS OF DYNAMIC FEATURES OF VOCAL TRACT SYSTEM

Recently, a new method for analysis, called zero-time liftering of speech signals was proposed (Dhananjaya, 2011). The method provides high temporal resolution, simultaneously maintaining a good spectral resolution. Liftering of speech signal in the time domain with a heavily decaying impulse-like window provides high temporal resolution, whereas the group delay analysis provides good resolution of the spectral characteristics. The use of a heavily decaying liftering function smoothes the spectrum severely resulting in a polynomial-type growth/decay, analogous to that in the zero-frequency filtering (Murty and Yegnanarayana, 2008). The masked or hidden spectral features can be highlighted by successive differencing of the numerator of the group-delay function. Phase inconsistencies of some weak higher formants are handled by computing the Hilbert envelope of the differenced numerator of the group delay function (Joseph et al., 2006). The resulting spectrum is referred to as HNGD function.

Figure 4 shows the HNGD plots computed at every sampled time instant for a prolonged utterance of the trill [\( r^3r^3 \)] in the context of following vowel [a]. The speech waveform and the instants of glottal closure (downward
arrows) are also shown along with the HNGD plots for reference. The HNGD plots are computed over segments of length 4 ms \((M = 40 \text{ at the sampling rate } F_s = 10,000)\), and using a discrete Fourier transform length of \(N = 2048\). The segment of speech shown in Fig. 4 has approximately five trill cycles over 200 ms of duration, which is equivalent to a trilling frequency of \(\sim 25\) Hz. The SNR of the speech signal varies continuously with time, and it can be seen that the HNGD plots around the high SNR regions (i.e., around the instants of glottal closure) are large compared to the HNGD plots in other portions of the signal. The time-varying nature of the spectral features can be seen better in Fig. 5, which shows one trill cycle of the speech signal and its HNGD plots computed at intervals of 1 ms. The GCIs are marked as downward pointing arrows. It can be seen from the waveform that at the beginning of the opening phase (\(\sim 118\) and \(\sim 150\) ms) there is a burst along with a bit of frication due to the sudden opening of the tongue tip. The effect of the burst and the frication can be seen in the spectrum as a large peak around 3 kHz. The burst is more prominent around the time instant \(150\) ms compared to that around \(118\) ms, depending on the synchronization between the instant of opening of the tapping and the instants at which the spectrum is computed. This shows that one may fail to capture the instantaneous changes in the spectral characteristics, if the signal is sampled only at the epochs or even at a finer sampling rate of every 1 ms. Also, the trill sound seems to have characteristics of a voice bar [predominant low frequency band around the fundamental frequency without any significant formant structure as in the case of voiced occlusions (Dhananjaya et al., 2008; Clark et al., 2007, p. 278)] during the closed phase, which is partly apparent from the signal (in the region \(145\)–\(150\) ms), but cannot be seen in the HNGD plots of Fig. 4. The large dynamic range between the HNGD plots in the open phase and the closed phase within a trill cycle, and between the closed and open phases of the glottal cycle (region around the GCI), makes it difficult to observe the dynamic spectral characteristics of all regions of a trill sound simultaneously. One way of observing the instantaneous dynamic nature of the trill sounds is by normalizing the HNGD plots computed at every time instant. Figures 6(a) and 6(b) show the HNGD plots for one trill cycle with and without normalization. In Fig. 6(b) the HNGD plots are normalized by dividing each plot by its maximum value, so that all the HNGD plots are now in the range of 0–1. The instantaneous or time-varying spectral characteristics of the trill sounds, such as the large spectral peaks around 3.5 kHz (\(\sim 118\) and \(\sim 150\) ms) due to bursts, and the voice-bar-like characteristics \(\sim 145\) ms, can be observed better in Fig. 6(b) compared to Fig. 6(a).

V. ANALYSIS OF TRILLS IN CONTINUOUS SPEECH

Like other continuants, trills are influenced by the adjacent vowel(s). In this section we examine the characteristics of a trill sound in the context of vowels [a], [i], and [u], which form the vertices of the vowel triangle in the \(F_1–F_2\) formant space. Trills can also undergo transitions from one

FIG. 5. (Color online) Waveform for one trill cycle of an apical trill [triː], and the corresponding HNGD plots computed for every 1 ms shift.

FIG. 6. (Color online) Waveform for one trill cycle of an apical trill [rɪɾ] and the corresponding HNGD plots. (a) Unnormalized HNGD plots and (b) normalized HNGD plots.
vowel context to another, when they occur in between two vowels. Other categories of trills, namely, voiceless apical trills and bilabial trills, are also examined in comparison with the voiced apical trills.

A. Effect of vowel context on trills

A comparative study between the tap [ɾ] and the apical trill [ɾ] by Recasens and Pallars (1999) shows that trills are less affected by adjacent vowels as compared to taps, which is mainly due to the more constrained lingual position required for the production of an apical trill as against a tap. Recasens and Pallars (1999) studied the coarticulation effects of apical trills with adjacent vowels using electropalatographic data, as well as formant frequency data, to show that a trill cannot be considered as a geminate correlate of a tap. Nevertheless, apical trills can be produced with varying vocal tract configurations irrespective of the small degree of freedom for variability. In this section, apical trills produced by a trained phonetician (male) in the context of three vowels [a], [i], and [u] are used to study the effect of vowel context on the trill. The three vowels provide three distinct vocal tract configurations.

Figure 7 shows the HNGD plots for trills uttered in three different vowel contexts, [aɾaɾa], [iɾi], and [uɾu]. The trill sounds have been uttered as isolated VCVs (vowel—consonant—vowel), where the consonant C is the trill [ɾ] and the vowel V is one of [a], [i], and [u]. The HNGD plots clearly show that the spectral peaks in the region of trill sounds are different in each of the three vowel contexts. This shows that the production of trill sounds need not have a unique vocal tract shape, although it has a highly constrained lingual configuration. Another observation that can be made from the HNGD plots is that the resonances of the trill sounds are more aligned with that of the vowels [a] and [u], as compared to that of [i]. This may be because the front-high tongue dorsum position for [i] needs to be retracted back considerably for the production of trill, which can be seen in terms of a decrease in second formant from [i] to [ɾ].

The amount of reconfiguration required by the tongue body in the transition from a vowel to trill can be observed in terms of the changes in the vocal tract resonances. Figure 8 shows the locations of the context-dependent trills in the vowel triangle formed by the vowels [a], [i], and [u], and the relative positions of the corresponding trills [ɾa], [ɾi], and [ɾu], respectively. The $F_1$ and $F_2$ values shown here are the mean values, obtained by averaging the $F_1$ and $F_2$ values estimated at every 1 ms shift from a single utterance of the VCVs [aɾaɾa], [iɾi], and [uɾu] by a trained male phonetician.

---

**FIG. 7.** (Color online) Waveform and HNGD plots of the apical trill [ɾ] in the context of three different vowels [a], [i], and [u]. (a) [aɾaɾa], (b) [iɾi], and (c) [uɾu].

**FIG. 8.** (Color online) Vowel triangle formed by the vowels [a], [i], and [u], and the relative positions of the corresponding trills [ɾa], [ɾi], and [ɾu], respectively. The $F_1$ and $F_2$ values shown here are the mean values, obtained by averaging the $F_1$ and $F_2$ values estimated at every 1 ms shift from a single utterance of the VCVs [aɾaɾa], [iɾi], and [uɾu] by a trained male phonetician.
mostly in the tongue dorsum position for the production of trills, whereas the largest $F_2$ movement (~500 Hz) is observed in the context of [i] as the frontoral tongue dorsum position for [i] is retracted for the production of the trill. Comparison of trills produced in the three different vowel contexts (marked as circles in Fig. 8) shows that $F_1$ has a narrow spread of ~150 Hz (from approximately 350 to 500 Hz), whereas $F_2$ has a broader spread of ~900 Hz (from approximately 800 to 1700 Hz). The small variation in $F_1$ shows that the length of the vocal tract does not vary much, and is the longest (lowest $F_1$) in the context of vowel [u]. The large variations in $F_2$ may be attributed to the flexibility in the tongue dorsum position for the production of trills, with a highly retracted or back position in the context of [u] producing the lowest value of $F_2$. This observation based on the acoustic data may probably be verified by an analysis of data from other modalities, such as electropalatography, x-ray, and/or MRI. Although apical trills tend to take on the spectral characteristics of the adjacent vowel to a certain extent, they also tend to move toward a common space, due to the inherent articulatory constraints in their production.

Figures 7 and 8 show the characteristics of trills produced with a fixed vocal tract configuration (one of the three vowels [a], [i], and [u]) on either side. Trills can also be produced with a continuously changing vocal tract when the vowels on either side of the trill are not the same. Figure 9 shows the spectral characteristics of trills uttered as isolated $V_1C V_2$ units, where $C$ denotes the apical trill [r], and $V_1$ and $V_2$ ($V_1 \neq V_2$) are one of the three vowels [a], [i], and [u]. The continuous transition of the spectral peaks from one vowel context to another can be clearly seen. In the case of trills transiting between vowels [a] and [i] ([ar] and [ir]) as in Figs. 9(a) and 9(b), the key feature is the exaggerated movement of the second formant at the boundary between [r] and [i], whereas the transition of formants between [r] and [a] is more gradual. This is mainly due to significant reconfiguration required in the tongue body position between [i] and [r], as seen from Figs. 7 and 8. In the case of trills transiting between [u] and [a] ([ur] and [ur]) as in Figs. 9(c) and 9(d), the key feature is the absence of any significant movement in the formants, as opposed to the case between [a] and [i], or [i] and [u]. Again trills transiting between [i] and [u] ([ir] and [ur]) as in Figs. 9(e) and 9(f) indicate a clear but gradual movement of the tongue body, as can be construed from the gradual movement of $F_2$.

**B. Features for spotting steady trills in continuous speech**

Acoustic cues for spotting steady trills in continuous speech are explored in this section. Unlike isolated utterances of trills, the majority of the trills in continuous or spontaneous speech tend to have fewer (less than three) trill cycles. They may have either one or two trill cycles, or the trilling may be totally absent at times, with the resulting sound being an approximant. Analysis of trills in the previous section shows the dynamic nature of the spectral characteristics of the trills, which vary with the vowel context. Hence, representation of the spectral characteristics of trills for spotting in continuous speech is a difficult issue. For spotting steady trills in continuous speech, an approach based on acoustic–phonetic knowledge using the excitation source characteristics seems to be more appropriate than a statistical approach using the spectral features. The excitation features that are useful in spotting the trills are the excitation strength and the instantaneous fundamental frequency. It can be recalled, Fig. 2 and Sec. III, that the excitation strength and the instantaneous $F_0$ vary in the trill region, whereas these parameters are almost steady and smooth for other voiced sounds. Figure 10(a) shows the waveform of a short utterance in Telugu, an Indian language, containing an apical trill. It can be seen that there are only two trill cycles in the utterance around the time instance 0.4 s (from approximately 0.35 to 0.45 s). Figures 10(d) and 10(e) show that the fluctuations in the excitation strength and the instantaneous $F_0$ can be observed for the trill sounds in continuous speech, which may be useful for distinguishing these regions from other voiced regions.

The speech waveform for a trill sound reflects two kinds of periodicity—a longer periodicity originating from the trill cycles and a shorter periodicity reflecting the glottal cycles. The presence of these two periodicities can be used as an additional cue for spotting trills in continuous speech. Normalized cross-correlation (NCC) can be used to measure the periodicity in a given signal. The NCC values for a given sequence $x[n]$, starting at an arbitrary time instant $n$, are computed as

$$\rho[k] = \frac{\sum_{n=0}^{M-1} x[n] x[n+k]}{\sqrt{\left(\sum_{n=0}^{M-1} x[n]^2\right) \left(\sum_{n=0}^{M-1} x[n+k]^2\right)}} ,$$

where $M$ is the window size over which NCC is computed and $k$ is the shift or time-lag. The NCC values computed for a segment of trill are plotted in Fig. 11(b). The window size $M$ used is 30 ms (240 samples at a sampling rate of 8 kHz). It can be seen that the NCC plot has multiple peaks. The highest peak (marked as TC—trill cycle), excluding the peak at 0 time-lag, corresponds to the periodicity due to apical vibration, and the first peak (marked as GC—glottal cycle) corresponds to the periodicity due to glottal vibration. Note that for a typical voiced segment, the highest peak in the NCC plot [Fig. 11(d)] is the peak due to glottal periodicity. The highest peak value $\rho_{\text{max}}$ in the NCC plot is given by

$$\rho_{\text{max}} = \max_k \{\rho[k]\}, \quad k = [N_1 : N_2],$$

where $N_1$ and $N_2$ are the lower and upper limits (in number of samples) for finding the highest peak, and may correspond typically to 2 and 50 ms, respectively. The time-lag $N_{\text{max}}$ (in number of samples) of the maximum NCC value $\rho_{\text{max}}$ and the corresponding frequency of periodicity $F_{\text{max}}$ in Hz are given by
\[ N_{\text{max}} = \arg \max_k \{ \rho[k] \}, \quad k = [N_1 : N_2] \]
and
\[ F_{\text{max}} = \frac{F_s}{N_{\text{max}}}, \]
(3) where \( F_s \) is the sampling frequency.

Figure 10(f) shows the \( F_{\text{max}} \) values computed for segments of speech starting at time instants \( n \) corresponding to the GCIs, using Eqs. (1), (3), and (4). It can be seen that the \( F_{\text{max}} \) values are low (~30 Hz) in the region of trill, whereas they are close to the instantaneous \( F_0 \) values in Fig. 10(e).
obtained using the ZFF signal. Pitch halving or doubling is a common problem when correlation measures are used to estimate the periodicity in speech signals, especially in steady voiced regions. It can be seen in Fig. 10(f) that, at around the time instance 0.1 s, the $F_{\text{max}}$ value is half that of the instantaneous $F_0$. Such spurious estimates can be identified using the knowledge that the contours of the excitation strength and the instantaneous $F_0$ are smoother in the steady voiced regions.

Figure 10(g) shows the maximum NCC values $\rho_{\text{max}}$ (marked as diamonds). The peak NCC values $\rho_0$ corresponding to the glottal periodicity are marked as dots, and they are obtained by locating the highest peak in $p[k]$ in the range of 2–20 ms. Note that the upper limit of the range is less than the period of the trill cycle. It can be noticed that $\rho_0$ and $\rho_{\text{max}}$ values are exactly the same in most voiced regions, as the peak of NCC values around the glottal period also happens to be the highest peak. In the region of trill, the $\rho_0$ values are smaller than the $\rho_{\text{max}}$ values, which correspond to the periodicities due to glottal and trill cycles, respectively.

### C. Voiceless trills

Trills produced with the absence of glottal excitation (or voicing) are referred to as voiceless trills. Analogous to their voiced counterparts, voiceless trills can be produced in different vowel contexts. The spectral characteristics of the voiceless trill [ɾ] in the context of [a] are shown in Fig. 12(b). It should be noted that the first formant (typically around 600 Hz) for the vowels [a] (between 30 and 100 ms and 370 and 450 ms) is not clearly visible. In the first [a] (between 30 and 100 ms), the first formant is faintly visible but appears to merge with the second formant at ~1500 Hz. In the second [a] (between 370 and 450 ms), the first formant is completely invisible, whereas the second and third
formants seem to be merging. This may be due to the presence of a strong turbulent noise-like excitation, which typically has less energy in the low frequency region. Similar characteristics are observed in the case of other vowel contexts as well. Figure 13 shows the acoustic cues discussed in the previous sections for spotting voiced trills in continuous speech. The figure shows a voiceless trill \( \text{[B]} \) followed by two bilabial trills, which will be discussed in the next section. It can be seen that the energy of the zero-frequency filtered signal [Fig. 13(c)] is significantly high in the region of voiceless trills, compared to the energy of the voiceless vowel [a] on either side. This is primarily because of the presence of impulse-like excitations generated by the trilling of the tongue tip. Hence, it is seen that the measured excitation strengths are large [Fig. 13(d)], and are detected as voiced region [Fig. 13(e)]. The epoch locations estimated within the voiceless trill are random, due to the noise-like characteristic of the excitation, and are irregularly spaced compared to a typical voiced region (vowel [a] between 1 to 1.2 s in Fig. 13). A similar trend can be observed for the instantaneous \( F_0 \) in Fig. 13(e). The signal periodicity \( F_{\text{max}} \) measured using the normalized cross-correlation is low (around 30 Hz) in the trill region as compared to other voiced regions (around 100 Hz), as can be seen in Fig. 13(f). It is to be noted that there is doubling of the estimated periodicity (or halving of \( F_{\text{max}} \)) at some instants toward the beginning of the trill. This can happen due to the presence of three or more steady trill cycles. This by itself may not be an issue for spotting voiceless trills, but needs to be addressed if an accurate estimate of the trill frequency is required.

D. Labial trills

Production mechanism of bilabial trills (voiced—[B] and voiceless—[B]) is similar to that of apical trills. They are produced at the bilabial place of articulation. They are reported to occur at the phonemic level in a few languages such as Piraha and Wari, in South America (Ladefoged and Maddieson, 1996). Figures 12(c) and 12(d) show the spectral characteristics, respectively, of voiced and voiceless bilabial trills. Although it is found that the bilabial trills have highly nonstationary spectral characteristics, a more careful analysis is essential to study the effect of vowel context, and to discriminate them from apical trills. The presence of a voice bar just before the start of trilling can be clearly seen from Fig. 12(c). Figure 13 shows the acoustic cues that can be used for spotting bilabial trills in continuous speech. The utterance contains a voiceless apical trill (discussed in the previous section), followed by a voiced bilabial trill (0.7–1.4 s) and a voiceless bilabial trill (1.4–1.8 s). It can be seen that the zero-frequency analysis brings out prominently the feeble voicing present during the closure region (~0.8 s) in the production of [\( \text{B} \)]. The excitation strength and the instantaneous \( F_0 \) values have much larger fluctuation, compared to apical trills. It is seen from Fig. 13(f) that bilabial trills have a trilling
frequency similar to that of apical trills (~30 Hz). The normalized cross-correlation values have a trend similar to that discussed for apical trills.

VI. SUMMARY AND CONCLUSIONS

Trill sounds, especially the apical trills, are produced with a rapidly changing vocal tract geometry due to the trilling of the apex of the tongue. Although the time-varying vocal tract dynamics result in continuously changing spectral characteristics, it also seems to influence the source of excitation. In this paper the characteristics of voiced apical trills were studied. The excitation source characteristics were studied by examining the features of excitation derived using the zero-frequency analysis of speech signals. The zero-frequency analysis provides information on epoch locations, strength of impulses at epochs and the instantaneous fundamental frequency ($F_0$). The instantaneous $F_0$ values vary within a trill cycle, with lower $F_0$ values in the closed phase of the trill cycle compared to the open phase. This fluctuation could be due to an increase in the pressure gradient across the glottis, when the tapping of the oral cavity by the tongue tip is released. The increase or reduction in the pressure gradient also affects the strengths of the impulses at epochs.

The dynamic spectral characteristics of trills were studied using the zero-time liftering method for analysis of speech. Through this method the instantaneous response of the vocal tract system could be obtained. This new method of analysis enabled us to examine the details of the spectral features during each trill cycle, and also the effect of vowel context on the spectral features of the trills. The spectral characteristics of the trills were examined using the HNGD plots derived using zero-time liftering analysis and group-delay analysis. Together these analysis methods provide good temporal resolution without affecting the spectral features significantly.

The acoustic cues for spotting trills in continuous speech were also explored. The fluctuations of the values of the excitation strength at epochs and the instantaneous $F_0$ help in discriminating trills from other voiced sounds. Normalized cross-correlation provides an additional cue for spotting trills. These acoustic cues for spotting will be useful if there are steady trill sounds in continuous speech with at least two complete trill cycles. But steady trill sounds with two or more trill cycles are less frequent in continuous speech. The acoustic characteristics of voiceless trills and bilabial trills...
were also examined briefly, mainly from the point of view of spotting them in continuous speech.

In summary, this paper demonstrates that trilling in the vocal tract affects the source of excitation. The reduced values of $F_o$ and the strengths of excitation observed during the closed phase, as compared to the open phase of a trill cycle, are in agreement with the observations made on voiced occlusions (stops). The influence of vowel context on the resonant characteristics of trill was examined using the high temporal resolution of the spectral features provided by the zero-time liftering analysis of speech. Although trills have been reported to have a highly constrained lingual configuration so as to meet the articulatory and aerodynamic requirements, it was shown in this paper that trills do have a significant amount of flexibility in terms of tongue body position, as seen from a large variation in the second formant.

The main contribution of this paper is in our ability to extract the dynamic characteristics of both excitation and vocal tract system resonances, using the new signal processing tools like zero-frequency filtering and zero-time liftering.

In this paper only a qualitative description of acoustic features for spotting trills in continuous speech is given. It is important to note that as the vocal tract shape is changing rapidly, and the system is excited mainly due to impulse-like excitation at epochs, the dynamic spectral characteristics sampled near the epoch locations may appear somewhat random during each trill cycle. As the spectral characteristics are changing continuously during the production of trills, and also due to vowel context, it is a challenge to represent them for pattern matching. Further study is needed to develop methods for discriminating different categories of trills such as apical and labial trills. The present study may provide some direction toward such an investigation. It would be interesting to study the contrast between trills and creaky voices, as both produce dynamic characteristics that need to be resolved both in temporal and frequency domain. Although trills are produced with a time-varying vocal tract system excited by a reasonably steady glottal source, creaky voice is produced by exciting a reasonably steady vocal tract by a time-varying glottal source.


