Effect of Tongue Tip Trilling on the Glottal Excitation Source

V. K. Mittal[1]*, N. Dhananjaya*[2] *and B. Yegnanarayana*[3]

International Institute of Information Technlogy, Hyderabad, India
[1]vinay.mittal@iiit.ac.in, [2]dhanu@reserach.iiit.ac.in, [3]yegna@iiit.ac.in

Abstract

Recent studies have indicated changes in the glottal excitation source characteristics apart from vocal tract resonances due to tongue tip trilling. In this paper we study the significance of changing vocal tract system and the associated glottal excitation source characteristics due to trilling, from perception point of view. These studies are made by generating speech signal by either retaining the features of the vocal tract system or of the glottal excitation source of trill sounds. Experiments are conducted to understand the perceptual significance of the excitation source characteristics on production of different trill sounds. Speech sounds of sustained trill and approximant pair, and apical trills produced by four different places of articulation are considered. Features of the vocal tract system are extracted using linear prediction analysis, and those of the source by zero frequency filtering.

Index Terms: glottal excitation source, trill synthesis, zero frequency filtering, apical trills, tongue tip trilling

1. Introduction

Trilling is a phenomenon where the shape of the vocal tract changes rapidly with an approximate trilling rate of about 30 cycles per second. Analysis of trill sounds is limited to the study of production and the acoustic characterization in terms of spectral features. For example, the production mechanism of tongue tip trills were described and modeled in [1, 2, 3], from aerodynamic point of view. Trill cycle and trilling rate were studied in [1, 3]. Acoustic correlates of phonemic trill production are reported in [4]. In a recent study, the acoustic analysis of trill sounds was carried out using some new signal processing methods [5]. During production of trills, changes in the shape of the vocal tract system seem to affect the vibration of the vocal folds at the glottis due to pressure difference caused at the glottis. It was observed that tongue tip trilling produces changes in the period of the vibration of the vocal folds in each cycle [5].

In this paper we study the effect of changing vocal tract shape and the associated changing excitation characteristics on the perception of trill sounds. Features (such as epochs) of the glottal excitation source are extracted using zero-frequency filtering (ZFF) method [6]. The strength of excitation (SoE) of the impulse-like excitation is indicated by the slope of the ZFF signal at each epoch. The intervals between successive epochs represent the instantaneous pitch period (T_0). Both SoE and T_0 are used as source features. The shape of vocal tract system is captured and represented through linear prediction analysis [7]. Synthetic trill sounds are generated, retaining either the features of the vocal tract system or of the excitation source. The relative significance of these two components is evaluated using perceptual studies. Speech sounds of sustained trill and approximant pair are examined. Trill sounds produced at 4 different places of articulation are also examined.

The paper is organized as follows. In Section 2, the nature of apical trills is discussed along with trill production process, and methods for extracting features of the vocal tract system and of the glottal excitation source. Section 3 discusses the analysis and synthesis of apical trills, and also methods for generating trill sounds with desired features. The four different scenarios adopted for the study of perception of trills, are also discussed. In Section 4, experiments to study the effect of tongue tip trilling are described, along with perceptual evaluation. Section 5 presents a summary and the scope for further studies in this direction.

2. Nature of apical trills

Trills, a stricture type, involve vibration of an articulator (lower) against another articulator (upper) due to aerodynamic constraints. Trills involving the lower articulator as the tip of the tongue are called *apical trills* [1]. The tongue tip in apical trills vibrates against a contact point in the dental/alveolar region. Production of an apical trill involves several aerodynamic and articulatory constraints. Aerodynamic constraints are related to tension at the tongue tip and volume velocity of air flow through the stricture, both essential for the initiation and sustenance of the apical vibration. Articulatory constraints are related to lingual and vocal tract configuration aspects [1, 5].

Production of apical trills can be charactersized by three cyclic actions: (i) Repeated breaking of the apical stricture due to interaction between tongue tension and volume velocity of air flow. (ii) Partial falling of the tongue tip to partially release the positive pressure gradient in the oral cavity. (iii) Recoiling the tongue tip to meet upper articulator to form next event of stricture. One such closure-opening-closure cycle of the stricture, shown in Figure 1 (a), is called a *trill-cycle*. Typical rate of tongue tip trilling, as measured from acoustic waveform or the spectrogram, is about 20-30 Hz [2, 3]. Two to three cycles of apical trills are common in continuous speech, whereas more than three cycles may be produced in sustained production of the sound [5]. When the lower articulator (tongue tip) does not touch (or tap) the upper articulator completely, the production of trill is like in Figure 1(b). However, due to aerodynamic and articulatory constraints, the production of trill in this case, is mostly, as shown in Figure 1(c). This sound is called *approximant*. Apical trills are common among some languages like Telugu, Malyalam and Punjabi (Indian) languages, whereas approximants are more common in some languages, like English.

The contact point of the upper articulator, against which the tongue tip vibrates, can be in different regions, like bilabial, dental, alveolar and post-alveolar. These are called in this paper as 'trill sounds produced at different places of articulation'.

Features of the glottal excitation source are extracted using *zero-frequency filtering* (ZFF) method [6]. The impulse-like characteristics of the excitation source are extracted by filtering

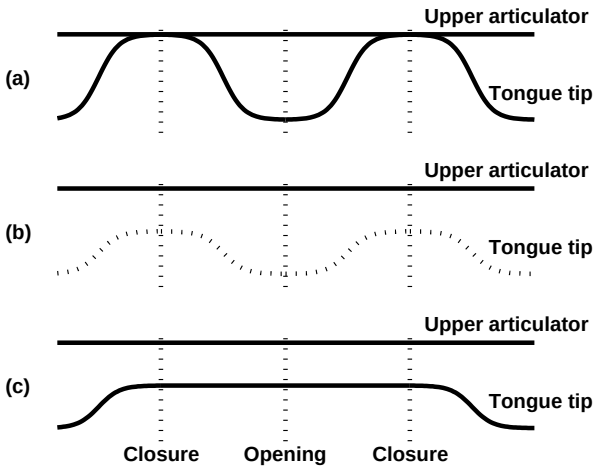

Figure 1: *Illustration of stricture for (a) an apical trill, (b) theoretical approximant and (c) an approximant in reality. The relative closure/opening positions of the tongue tip (lower articulator) with respect to upper articulator are shown.*

it through a cascade of two zero-frequency resonators. A zero-frequency resonator (ZFR) is an all-pole system with two poles at $z = +1$ in the z-plane. It is equivalent to a sequence of two cumulative sum operations in time-domain, which leads to polynomial-like growth/decay of the ZFR output signal. The fluctutations in the ZFR output signal can be emphasized using a trend removal operation, involving subtraction of the local mean from the signal at each time instant. The resultant zero-frequency filtered (ZFF) signal is illustrated in Figure 2(b). The positive zero-crossings (negative to positive going) correspond to the instants of glottal closure, also referred to as *epochs* [6]. The slope of the filtered signal around the epoch gives a measure of the *strength of* the impluse-like *excitation (SoE)* [6]. The interval between successive epochs corresponds to fundamental period (T_0), inverse of which gives the instantaneous fundamental frequency (F_0).

It was established [1] that the strength of excitation *(SoE)* varies within a trill cycle. The *SoE* is less during the closed phase as compared to the open phase of a trill cycle. Also, the instantaneous fundamental frequency (F_0) varies due to the trilling of the tongue tip. Apparently, F_0 is at minimum value in the closed phase just before the release of apical contact, and increases gradually with the opening of apical contact stricture. Contours of F_0 and *SoE* variation for apical trills are shown in Figures 2(c) and 2(d), respectively.

Features of the vocal tract system are usually extracted by LP analysis [7], using a frame size of 20 msec and a frame shift of 5 msec. The LPCs capture the change of the vocal tract shape information, if the analysis frame size is less than the period of a trill cycle (i.e., 33 msec for a trill cycle of 30 Hz).

Synthetic speech is generated by exciting the time-varying vocal tract system represented by LPCs with an impulse sequence with intervals corresponding to the pitch period, and with amplitudes corresponding to the strength of excitation. LPC based synthesis model is used, as it provides flexibility of modifying the excitation parameters such as pitch period and gain, and at the same time maintains naturalness by using the residual signal information, if necessary [8].

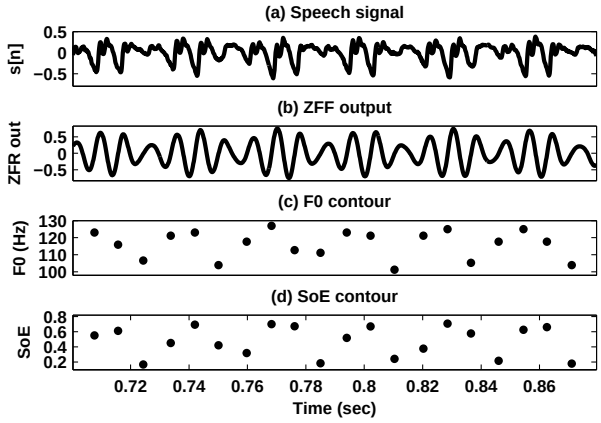

Figure 2: *(a) Signal waveform, (b) ZFF output, (c) F_0 contour and (d) SoE contour for an apical trill region.*

3. Analysis and synthesis of apical trills

In order to study the relative significance of the dynamic vocal tract system and the excitation source in the perception of trill sounds, synthetic speech signals are generated by controlling the system and source characteristics of trill sounds separately. For this the natural trill sounds are analyzed to extract the source characteristics in terms of epochs and the strength of impulses at epochs. The dynamic characteristics of the vocal tract shape are captured using LP analysis on a frame size of 20 msec with a frame shift of 5 msec.

Four scenarios are considered for synthesis of trills: (i) Retaining the characteristics of both source and system. (ii) Only source. (iii) Only system. (iv) Neither source nor system. Perceptual evaluation of the synthesized speech in each of these four scenarios of selective retention is carried out.

Changes in the glottal excitation source characteristics are made by first disturbing the fundamental frequency (F_0) information and then the amplitude information. For each epoch, the next epoch is located at an interval corresponding to the average of several pitch periods around this epoch. The new impulse sequence reflects the averaged pitch period information. The amplitude of each impulse corresponds to the average of the *SoE* around that epoch. This new impulse sequence is referred to in this paper as 'changed excitation source' information. The impulse sequence with changed source information is used as excitation for generating speech for scenarios (iii) and (iv) of selective retention. The effect of this averaging is shown in Figure 3 for a trill and for an approximant. Figure 3 shows the changed source charcteristics, i.e., the F_0 and *SoE* contours of excitation sequence, which are more evident for the trill (first) sound as compared to the approximant (second) sound. This can be contrasted with the corresponding contours of the F_0 and *SoE* of excitation sequence, for the original trill and approximant sounds, shown in Figure 4.

Since the trill cycle is of 20-30 Hz, the LPCs computed using a frame size of 100 msec can be considered as the 'changed characteristics of vocal tract system'. The changed characteristics of the system is used for generating speech for scenarios (ii) and (iv) of selective retention.

To establish the significance of coupling between the system and source characteristics (in scenario (i)), the scenario (iv), where both the source and system information are changed, is used for comparison.

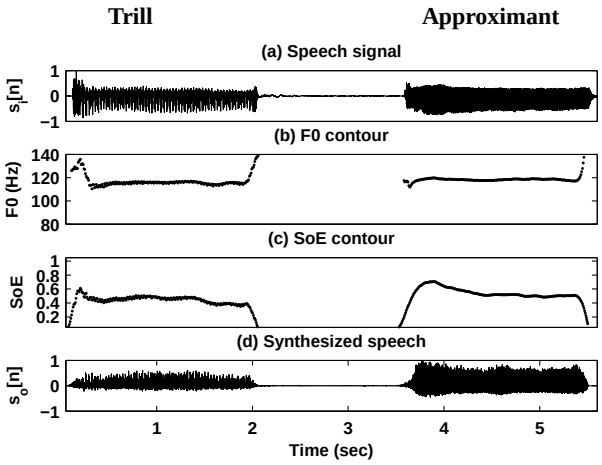

Figure 3: *(a) Signal waveform, (b) F_0 contour and (c) SoE contour of excitation sequence, and (d) Synthesized speech waveform (x_{13}), for a sustained apical trill-approximant pair. Source information is changed (system only retained) in synthesized speech.*

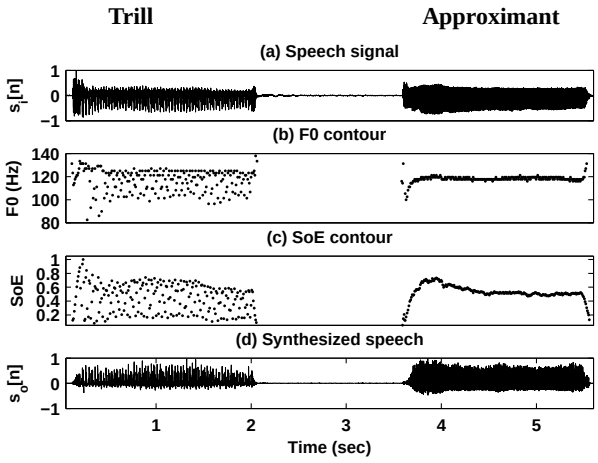

Figure 4: *(a) Signal waveform, (b) F_0 contour and (c) SoE contour of excitation sequence, and (d) Synthesized speech waveform (x_{11}), for the sustained apical trill-approximant pair. System and source both information (that of original speech) are retained in synthesized speech.*

Perceptual evaluation of the 4 synthesized speech files (one for each of the 4 scenarios of selective retention), with reference to original speech file is carried out. Similarity score criterion as given in Table 1, is used.

Table 1: *Criterion for similarity score for perceptual evaluation of two trill sounds (synthesized and original speech)*

Perceptual similarity	Similarity score
both sound very much similar	5
both sound quite similar	4
both sound somewhat similar	3
both sound quite different	2
both sound very much different	1

4. Perceptual evaluation

Two experiments were conducted in this study, each with 4 scenarios of selective retention of source/system information. Since most databases of continuous speech have very limited trill data suitable for this study, the required speech sounds of sustained trill-approximant pair and the trills with 4 different places of articulation, are recorded with the help of an expert male phonetician.

Experiment 1 was conducted with a trill-approximant pair speech file (*x10.wav*). From this reference file (x_{10}), the features of the glottal excitation and the vocal tract system are extracted. Using these source and system features, the four synthesized speech files (x_{11}, x_{12}, x_{13} and x_{14}) are generated, one for each of the 4 scenarios of selective retention of source/system information. Figure 4 shows the sustained apical trill-approximant pair with F_0 and *SoE* contours of excitation sequence, and synthesized speech for scenario (i) (i.e., retaining both source and system information). Figure 3 shows the changed source characteristics as in scenario (iii) (i.e., retaining only the system information) of the trill region. The effect of 'changed source information' can be observed in F_0 and *SoE* contours of excitation sequence in Figure 3, as compared to those in Figure 4. Perceptual evaluation is carried out by comparing each of the synthesized speech file (x_{11} to x_{14}) with reference original speech (x_{10}). A total of 20 subjects, all speech researchers from Speech and Vision Lab at IIIT-Hyderabad, participated in this evaluation. The subjects were asked to give the similarity scores for each of the 4 synthesized trill-approximant pairs. The averaged scores of perceptual evaluation for Experiment 1 are given in Table 2.

Table 2: *Experiment 1: Results of perceptual evaluation. Average similarity scores between synthesized speech files (x_{11}, x_{12}, x_{13} and x_{14}) and original speech file (x_{10}) are displayed.*

x_{11} vs x_{10}	x_{12} vs x_{10}	x_{13} vs x_{10}	x_{14} vs x_{10}
(Source, System retained)	(Source only retained)	(System only retained)	(Source, system changed)
3.95	3.48	2.82	1.75

Another experiment (Experiment 2) was conducted with speech file (*x20.wav*) consisting of trill sounds corresponding to the 4 different places of articulation, namely, bilabial, dental, alveolar and post-alveolar. From this reference speech file (x_{20}), the features of the glottal excitation and vocal tract system are extracted. Four synthesized speech files (x_{21}, x_{22}, x_{23} and x_{24}) are generated for each of the 4 scenarios of selective retention of source/system information. Perceptual evaluation was carried out by comparing each of the trill sound in a synthesized speech file (x_{21} to x_{24}), with the corresponding original trill utterance (x_{20}). Similarity score for each of the 4 different places of articulation with respect to corresponding original speech was obtained. All the 20 subjects gave similarity scores for each of the 4 synthesized speech files (x_{21}, x_{22}, x_{23} and x_{24}), for each place of articulation. The results of perceptual evaluation for Experiment 2 are given in Table 3.

In Table 2 the high average score in column 1 is due to the fact that source and system information both are retained in the synthesized speech, which is perceptually close to the

Table 3: *Experiment 2: Results of perceptual evaluation. Average similarity scores between each place of articulation in synthesized speech files (x_{21}, x_{22}, x_{23} and x_{24}), and corresponding sound in original speech file (x_{20}) are displayed.*

File name: synthesized vs reference speech	Bilabial trill	Dental trill	Alveolar trill	Post-alveolar trill	Average score (for all 4 trills)
x_{21} vs x_{20}	3.15	3.55	3.58	3.13	3.35
x_{22} vs x_{20}	2.55	2.90	2.85	2.85	2.79
x_{23} vs x_{20}	2.25	2.45	2.30	2.30	2.33
x_{24} vs x_{20}	1.20	1.30	1.40	1.40	1.33

original speech. The lower average score in column 4 confirms the fact that when both the source and system are changed, the resulting sound is different from the original trill utterance. The trill sound is perceptually close to an approximant, in this case. The lower average score in column 4 in contrast to high average score in column 1, is indicative of the fact that vocal tract system and glottal excitation source information both jointly contribute to the production and perception of trill sounds.

The relatively high average score in column 2 (for x_{12}), where only source information is retained (system information is changed), and relatively low average score in column 3 (for x_{13}), where only system information is retained (source information is changed) are interesting results. These scores indicate the relatively higher significance of the glottal excitation source information in the perception of apical trills.

In Table 3, the last column gives the average similarity scores across all the 4 different trill sounds. These results are in line with the results of Experiment 1 (in Table 2).

The average scores in row 3 (for x_{22}) in Table 3, where only source information is retained, are relatively higher in comparison to the average scores in row 4 (for x_{23}), where only system information is retained. This pattern is consistent for each of the 4 places of articulation. It reconfirms the inference drawn from the results of Experiment 1 (in Table 2) that source information contributes relatively more as compared to the system information, in perception of tongue tip trills.

The relatively higher average scores in 3^{rd} and 4^{th} columns (in Table 3, for dental and alveolar trills, respectively), in row 2 especially (for x_{21}) where both source and system are retained, also indicate relatively better perceptual closeness of dental and aleveolar synthesized trill sounds to the corresponding natural trill sounds.

The least average scores in last row (for x_{24}) in Table 3, when both source and system are changed, and high average scores in row 2 (for x_{21}), when both source and system are retained, are similar to the results of Experiment 1 (in Table 2). These average scores highlight the fact that there is some amount of system-source coupling in the production and perception of the tongue tip trilling. It also indicates that production of tongue tip (apical) trilling does affect the glottal excitation source due to coupling with the vocal tract system.

5. Conclusions

The effect of tongue tip (apical) trilling on glottal excitation source is indicated by the fact that system alone or source alone information is not sufficient for production and perception of apical trills. Both the source and system are involved in some coupled way, in the production/perception of apical trills, due to interaction between aerodynamic and articulatory components. Glottal excitation source appears to contribute relatively more, for perception of apical trills, as indicated by the perceptual evaluation results of experiments 1 and 2. Also, the synthesized apical dental/alevolar trills are perceptually closer to the corresponding natural trill sounds.

This study can be useful further in automatic spotting of trills, synthesis/modification of trill sounds and trill-based discrimination of different languages and dialects. The study can also be helpful in distinguishing the trill sounds and approximants from signal processing point of view, and in understanding the production/perception of different apical trill sounds at different places of articulation.

The data-files used and synthesized speech files can be downloaded from the link below:

http://speech.iiit.ac.in/index.php/demos/trill-is2012.html

Acknowledgement

Authors are thankful to Prof Peri Bhaskararao for bringing attention to subtle phonetic differences between sounds of apical trills/approximants and different trills as per place of articulation, and also in recording the required date files.

6. References

[1] P. Ladefoged and I. Maddieson, "Sounds of World's Languages", *Blackwell publishing*, Oxford, UK, chapter 7, pp. 217-236, 1996.

[2] P. Ladefoged, A. Cochran and S. F. Disner,"Laterals and trills", *J. Intl. Phon. Ass.*, vol. 2, pp. 46-54, 1977.

[3] R. S. McGowan,"Tongue tip trills and vocal-tract wall compliance", *J. Acoust. Soc. Am.*, vol. 91, no. 5, pp. 2903-2910, May 1992.

[4] N. C. Henriksen and E. W. Willis, "Acoustic characterization of phonemic trill production in Jerezano Andalusian Spanish", *Selected Proceedings of the 4^{th} Conference on Labortatory Approaches to Spanish Phonology*, edited by M. Ortega-Llebaria, pp. 115-127 (Cascadilla Proceedings Project, Somerville, MA), 2010.

[5] N. Dhananjaya, B. Yegnanarayana and Peri Bhaskararao, "Acoustic analysis of trill sounds", *J. Acoust. Soc. Am.*, vol. 131, no. 4, pp. 3141-3152, Apr. 2012.

[6] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals", *IEEE Trans. Acoust., Speech, Lang. Process.*, vol. 16, no. 8, pp. 1602-1614, Nov. 2008.

[7] J. E. Markel and A. H. Gray, "Linear Prediction of Speech", *Springer-Verlag New York, Inc.*, Secaucus, NJ, USA, 1982.

[8] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation", *IEEE Trans. Acoust., Speech, Lang. Process.*, vol. 14, pp. 972-980, May 2006.