ANALYSIS AND SYNTHESIS OF DEVANAGARI FRICATIVES

Saurabh Kuhar
Tata Motors Limited, Engineering Research Centre, Lucknow, Uttar Pradesh, India
email: skuhar9@gmail.com

Nachiketa Tiwari
Indian Institute of Technology, Department of Mechanical Engineering, Kanpur, Uttar Pradesh, India

In this work a detailed analysis of Devanagari fricatives and simple recipes for their synthesis are presented. Pronunciations of these sounds vary with speaker’s age, gender, dialect etc. Yet listeners identify them with remarkable consistency. This work discusses the reason underlying such consistency, identifies key aural-differentiators for different fricatives, and develops recipes for these fricatives. A key differentiator of this work vis-a-vis past works is that it focuses on synthesis of naturally sounding fricatives through relatively simple recipes. Each of these recipes requires only two inputs. These are white noise and a simple fricative-specific filter. Our work also offers recipe for the uniquely Indian fricative /ʂ/. Listening tests showed that the auditory perception of our synthesized samples was very close to that of corresponding natural fricatives with recognition rates exceeding 88%. As these recipes use relatively simple inputs and produce close-to-natural sounds, they may accelerate data transmission, help speech therapy, and improve voice recognition technologies.

1. Introduction

A fricative is a class of consonants, which are generated as air passes through a constricted vocal tract. As there is significant pressure difference across the constriction, the airflow through the passage becomes turbulent thereby generating broadband noise. The initial portion of fricatives /s/ and /ʃ/ is made up of broadband noise as described in [1]. She combined initial parts of certain fricatives with vocalic parts of other fricatives. By listening to such recombinations, she found that identifying cues for /s/ and /ʃ/ reside in their initial parts. However, she did not characterize the nature of these signals. While our observations are consistent with her work, we also characterized the nature of these signals and used these characteristics for their synthesis. [2] investigated spectral properties of fricatives. They reported that differences between various classes of fricatives (labial, dental and palatal) are somewhat context invariant.

Most Indian fricatives are unvoiced. Unvoiced fricatives are associated with low amplitude vocal fold vibrations only when the fricative sound transits into a terminating vowel. This is true for all Devanagari fricatives discussed by [3]. They also identified specific acoustical characteristics that differentiate between voiced and unvoiced fricatives. They claimed one such characteristic to be the duration of period during which there are no vocal fold vibrations. [4] explained why fricatives tend to be generally unvoiced. [5] classified English fricatives according to the place of their articulation. They correlated variations in spectral peak location and spectral moment corresponding to changes in articulatory position for different fricatives. Similar approach was also used by [6] to identify the gender of speaker through analysis of /s/ and /ʃ/, and for characterization of fricatives by [7]. [8]
reported on laryngeal specifications of fricatives used in Sanskrit, Pali and several other languages. He found that voiceless fricatives are produced when glottis is spread wide open.

[9] concluded that the duration of signal for voiceless fricatives is not an effective metric in their identification and peak frequency of fricative-noise increases with outward movement of constriction in the oral cavity. Through such an observation, it was explained as to why the peak frequency for /ʃ/ is less than that for /s/. Similar observations were also reported in reference [2,5, and 10]. Perception of vowels and consonants was compared in different environments and these sounds were grouped in three categories based on their perception in noisy and quiet environments, and it was found by [11] that /s/ and /ʃ/ were relatively easy to recognize sounds in noisy environments.

Several efforts to understand fricatives’ production mechanism have been made in the past. Reference [12] synthesized fricatives by analysing spectra of recorded fricatives, and using electrical analogs of mechanical models. They reported that their synthesized signals for /s/ and /ʃ/ did not sound natural. In an exhaustive work, [13] mentioned that the presence of obstruction, length of front cavity, and air’s flow rate are significant parameters which influence production of fricatives. She concluded that the generation of /s/ is associated with a series pressure source, while other fricatives are associated with distributed pressure sources. Effects of the tongue’s position, the oral cavity shape, and other parameters on fricative production is explored by [14,15]. Noise source models to artificially reconstruct spectra of English voiced and unvoiced fricatives were developed by [16]. However, they did not report on the fidelity of such reconstructions. Other mechanical source models have also been discussed by [17,18]. The role of jaw position and tongue tip in production of German coronal consonants has been investigated by [19].

Earlier works paid less attention to Devanagari fricatives. We fill this gap by identifying those acoustical cues, which are fricative-specific but person-invariant, and which aid in fricative identification. We use this understanding for synthesis of Devanagari fricatives using simple recipes. We show that these recipes can generate fricatives with high recognition rates. In this work we also benefited from the understanding of ancient grammarian Panini whose work on Devanagari alphabet has been discussed in [20]. He terms fricatives as ushmaan, which in Sanskrit implies ‘carrying heat’ thereby suggesting the association of frictional energy with fricatives. He also relates to the nature of /s/, /ʃ/ and /ʂ/ through Sanskrit terms for widened glottis, relatively free outflow of air, lack of vibrations in vocal folds, and relatively large outflow of air.

2. Method

We recorded voice samples of 10 speakers, aged between 13 and 48, in wav format. None of the speakers had had any known speaking or hearing problems. The sampling rate was set at 51.2 kHz. Each recorded sample was analysed in time and frequency domains. We used several custom designed digital filters to assess the importance of specific frequency bands in signature cues corresponding to each fricative. In this way, we identified characteristic cues for each fricative. We used this information to synthesize fricatives. We did this by applying fricative specific filters to a base signal made up of white noise. These filters work as transfer functions and modify the source sound in the same way as shape and motion of articulators modulate noise generated due to passing of air through the constricted oral passage. Next, we presented these synthesized fricatives in a random order to listeners for assessing the goodness of our recipe.

3. Results

Figure 1 shows how normalised sound pressure level varies in time during the production of different fricative sound. We note from this figure that this is a three-stage process.

To understand this, we divided the time signal for each fricative into three regions (see Figure 1), corresponding to three stages of sound production. Here, Region 1 represents sound produced due to air passing through the constricted passage created by articulators.
Next, there is Region 2, a transitory zone representing sound produced when the articulators start moving back to their relaxed position. During this period, the vocal folds begin to vibrate and start producing the terminating vowel. Finally, there is Region 3, corresponding to sound produced when articulators are relaxed and only vocal folds are vibrating and producing the terminating vowel. For Devanagari fricatives, this terminating sound is schwa (ə). Such a production process has been described in detail by [13-18]. Such a decomposition of time signals helps us differentiate these against unvoiced unaspirated stops. Table 1 provides an overview of these contrasts. We note that the duration of Region 1 for fricatives is significantly larger than that for unvoiced unaspirated stops. We also note that unlike fricatives, which have broadband spectral content, specific frequencies in narrow bands tend to dominate for unvoiced unaspirated stops. This was also shown by [21, 22].

We also found that Region 2 does not materially influence our auditory perception of a fricative. To show this, we removed Region 2 from each of the fricative samples. Next, listeners evaluated such modified samples in a random order. Listeners recognized such modified samples correctly in 97% of cases, and could not distinguish them vis-à-vis original signals. We attribute this to the fact that the duration of Region 2, as well as energy contained in it is small vis-à-vis those for Regions 1 and 3. We also note that Region 3 has little bearing on auditory perception of the fricatives studied in this work, as this region corresponds to the same terminating vowel, i.e. schwa. Hence, in this work we conducted a detailed analysis of only Region 1 as this region has all appropriate acoustical signatures specific to a fricative. These observations are applicable to all four voiceless fricatives. A logical extension of this work in future could include understanding the extent of influence of context on the nature of these fricative sounds. It has been shown in [23, 24] that for some classes of stop consonants (e.g. velars and retroflex), context plays an important role in influencing spectral characteristics of the burst signal, while for other stops (e.g. labials), this role is not significant. Next, we discuss important fricative-specific observations and their recipes.

3.1 The Structure of, and the Recipe for Post Alveolar Fricative /ʃ/

Ten samples of Region 1 corresponding to /ʃ/ were analyzed. The mean duration of these signals was 0.17 s. Further, we note in Figure 1 that the amplitude of the signal increases with time but its peak value remains significantly less than that of Region 3. Figure 2 shows spectral content of two such signals. We note that the bulk of the signal’s energy is concentrated in 1-10 kHz band, and its spectral amplitude rolls off rapidly post 8 kHz. We designed a filter shown in Figure 3 to
accomplish such a modulation and also overlaid it on Figure 2 for comparison. We used this information to synthesize /ʃ/. First, we generated a 0.17 s long white noise signal and increased its amplitude parabolically over time such that the amplitude did not exceed 0.05. Next, we applied a digital filter shown in Figure 3, and concatenated this signal with the terminating and person-specific natural schwa. In this way, we synthesized ten different samples of /ʃ/. Figure 4 is the time-series representation of one such synthesized /ʃ/. Listening tests showed that in 88% cases (70 out of 80) synthesized /ʃ/ sounded same as natural /ʃ/. We also noted that the parabolic increase in the amplitude of the base signal is acoustically important. We found that sans such amplification, synthesized /ʃ/ has an element of a hiss sound. This hiss is imperceptible when the base signal is amplified over time parabolically.

![Figure 2: Frequency spectrum for Region 1 of fricative /ʃ/.

![Figure 3: Filter for synthesizing /ʃ/.

![Figure 4: Synthesized /ʃ/ sound signal.

3.2 The Structure of and the Recipe for Retroflex Fricative /ʂ/

Our analysis of ten samples for Region 1 of /ʂ/ sound shows that mean duration for these signals is 0.17 s. Figure 1 depicts one such sample. We noted that the signal for Region 1 of /ʂ/ increases in amplitude with time but its peak value remains significantly lower than that for Region 3. Figure 5 depicts the spectral content of two such signals. We observe that most of their energy is concentrated in 500 - 2500 Hz band. We also noted that signal’s spectral amplitude rapidly ramps up and rolls off near 500 Hz, and 2500 Hz in all samples. To mimic such spectral modulation, we designed a custom filter shown in Figure 6. This filter is overlaid on Figure 5 for comparison purposes.
The 23rd International Congress on Sound and Vibration

ICSV23, Athens (Greece), 10-14 July 2016

Next, we synthesized /ʂ/. For this, we first generated a 0.17 s long white noise signal with parabolically increasing amplitude but limited to a maximum of 0.05. The rationale for such amplification has been explained earlier. Next, we applied the custom designed digital filter as depicted in Figure 6 to this signal. Finally, we concatenated this signal with the schwa of /ʂ/ sound corresponding to different speakers and generated ten different samples of /ʂ/ sound. Figure 7 represents one such synthesized /ʂ/ signal. Listening tests showed that in 89% (71 out of 80) cases the synthesized /ʂ/ sounded same as natural /ʂ/.

3.3 The Structure of and the Recipe for Alveolar Fricative /s/

Our analysis for Region 1 of /s/ sound shows that the mean duration of these signals is 0.17 s. Figure 1 depicts one such sample. The amplitude of this signal increases with time but its peak value remains significantly lesser than that for Region 3. Figure 8 depicts spectra for two of these signals. It shows that bulk of signal’s energy is contained in 4-8 kHz band, and the spectral amplitude rolls off rapidly post 8 kHz. Next, we synthesized /s/. For this, we first generated a 0.17 s long white noise signal with parabolically increasing amplitude but not exceeding 0.05. Subsequently, a digital filter as shown in Figure 9 was applied to it. Finally, we concatenated this signal with the schwa of /s/ sound corresponding to different speakers. Figure 10 is the time-series representation of one such synthesized /s/. Listening tests showed that in 96% (77 out of 80) cases the synthesized /s/ sounded same as natural /s/.
3.4 The Structure of and the Recipe for Glottal Fricative /h/

Our analysis for Region 1 of /h/ sound shows that the mean duration of these signals is 0.11 s. Figure 1 depicts one such sample. The amplitude of this signal increases with time but its peak value remains significantly lesser than that for Region 3. Figure 11 depicts spectra for two such signals.

We see that bulk of signal’s energy is contained in 500–2000 Hz band, and the spectral amplitude rolls off rapidly post 2 kHz. For its synthesis, we generated a white noise signal with parabolically increasing amplitude over time but limited its maximum value to 0.05. We noted that sans such a
parabolic amplification, synthesized /h/ was confused with other aspirated consonants including /kh/, and /f/. Next, we applied the digital filter shown in Figure 12. Finally, we concatenated this signal with the schwa of /h/ sound corresponding to different speakers. Listening tests showed that in 100% cases the synthesized /h/ sounded same as natural /h/.

Figure 12: Filter for synthesizing /h/.

4. Conclusions

In this work, we have analyzed and synthesized all Devanagari fricatives. For this, we delineated their signals into three regions corresponding to the physiology of their articulation. Noting that the transitory Region 2 does not materially influence the perception of these fricatives, we focussed our analysis on Region 1 only, as it carries necessary acoustic cues required to identify these fricative sounds. We show that the spectral content of these fricatives is concentrated in specific bands. These insights helped us develop recipes for Devanagari fricatives as shown in Table 2.

<table>
<thead>
<tr>
<th>Fricative</th>
<th>Articulation</th>
<th>Recipe</th>
<th>Signal’s duration</th>
<th>Recognition rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>/ʃ/</td>
<td>Post Alveolar</td>
<td>Application of filter shown in Figure 3 on a parabolically increasing white noise with a peak amplitude of 0.05</td>
<td>0.17 s</td>
<td>88%</td>
</tr>
<tr>
<td>/ʂ/</td>
<td>Retroflex</td>
<td>Application of filter shown in Figure 6 on a parabolically increasing white noise with a peak amplitude of 0.05</td>
<td>0.17 s</td>
<td>89%</td>
</tr>
<tr>
<td>/s/</td>
<td>Alveolar</td>
<td>Application of filter shown in Figure 9 on a parabolically increasing white noise with a peak amplitude of 0.05</td>
<td>0.17 s</td>
<td>96%</td>
</tr>
<tr>
<td>/h/</td>
<td>Glottal</td>
<td>Application of filter shown in Figure 12 on a parabolically increasing white noise with a peak amplitude of 0.05</td>
<td>0.15 s</td>
<td>100%</td>
</tr>
</tbody>
</table>

Our listening tests showed that recognition rates for synthesized fricatives are very high and they sound fairly close to their respective natural counterparts. This establishes the validity of these recipes. As these recipes use relatively simple inputs and produce close-to-natural sounds, they may accelerate data transmission, enhance speech therapy, and improve voice recognition.

REFERENCES


