QUALITY ANALYSIS OF NON-NASAL AND NON-ASPIRATED URDU STOPS, SYNTHESIZED USING HLSYN (A HIGH LEVEL SYNTHESIZER)

MUHAMMAD IRFAN RAFIQ

1. ABSTRACT

HLSyn is a quasi-articulatory synthesizer (independent of any particular language) with a small set of parameters, which model the complex acoustic consequences of vocal tract. The primary motivation of using HLSyn is to combine the ‘simplicity of control’ of articulatory synthesizers with ‘accuracy and efficiency’ of formant synthesizers. The aim of this paper is to analyze the quality of non-nasal and non-aspirated Urdu Stops, synthesized using HLSyn.

2. INTRODUCTION

The work described here is the initial step towards the development of Urdu text-to-speech Engine. The first major task of text-to-speech conversion is the phonemic transcription including syllabification and lexical stress assignment, for all commonly used words in the language. Normally large lexicons are used for this purpose. Novel words (which are absent in the lexicon) are transcribed by implicit analogy with existing lexicon entries. Implicit analogy refers to technique of automatic learning of context-dependent grapheme-to-phoneme rules from existing lexicon entries (Bagshaw, 1998, p.119-142).

The second major task is to synthesize the speech, given the transcription and other information about syllabification and lexical stress. Speech synthesizers used for this purpose, can fall into any of the three broad categories;(1) Articulatory Synthesizers that attempt to model the vocal tract directly instead of modeling the acoustic output from it; (2) Formant Synthesizers which derive the approximation of the speech waveform, with the help of a set of rules formulated in the acoustic domain;(3) Hybrid Synthesizers which combine both of the above techniques. HLSyn falls into this category.

The scope of this paper is limited to just synthesis of few words of Urdu, using HLSyn and analyzing the quality of non-aspirated consonants of Urdu.

3. LITERATURE REVIEW & PROBLEM STATEMENT

HLSyn is a high-level front end to Klatt-type synthesizer.

3.1. Overview of HLSyn

The design of the HLSyn is based on the observation that the values of the 40 parameters used to control Klatt-type synthesizers are not independent, but are subject to inter-parameter constraints. The constraints arise because speech production, as a physical process. This constraint permits only certain combinations of synthesis parameters to arise and also limits the rates at which the parameter values can change with time. To express these constraints, a small set of 10 high-level (HL) parameters was originally proposed. This set has now been expanded to include 13 HL parameters. The HL parameters are more closely related to the actual states and articulatory movements in the vocal tract than are lower-level (Klatt) parameters. The principle employed in HLSyn is simple: a set of mapping relations within HLSyn transforms the HL parameters into the values of the corresponding lower-level parameters (HLSyn User Manual, p.1,4,5).

3.2. HLSyn parameters

The functions of these parameters can be described in terms of three broad classes:

Class 1 parameters control the first four natural frequencies of the vocal tract (f1, f2, f3, f4); these parameters specify acoustically the vocal tract configuration and slow movements of articulators. The f0 parameter specifies the fundamental frequency.

Class 2 parameters control cross-sectional areas of local constrictions formed by the lips (al) and the tongue tip/blade (ab). They specify the fast movements of primary articulators that rapidly decrease/increase airflow within the oral tract.

Class 3 parameters control cross-sectional areas of the glottal orifice (ag) and velo-pharyngeal port (an) and the pharyngeal volume (ue). These parameters specify opening/closing movements of the glottis and velum and active expansion or contraction of the pharynx respectively.

3.3. HLSyn mapping relations

HLSyn parameters can be divided into two broad categories, namely Speech controlling parameters and Speaker dependent parameters. Twelve of them are speaker dependent parameters while thirteen of them are speech dependent (See fig.1).

After the HLSyn parameter values have been specified, the first step in determining values for the low level (LL) parameters is to calculate the pressures and flows at the supra-glottal and glottal orifices, (shown on the bottom left side of fig.1) using an aerodynamic model. The output of this model along with the HL parameters under goes to mapping equations to calculate the 25 LL parameters. Twelve of the LL parameters have been fixed and twenty-three of them uses a default value, thus completing the formation of all LL parameters (HLSyn User Manual, p.6-23).

The design of Klatt-type synthesizer (LL synthesizer) is based on the source-filter theory of speech production, presented by Fant (1960) and is summarized in the Fig. 2.

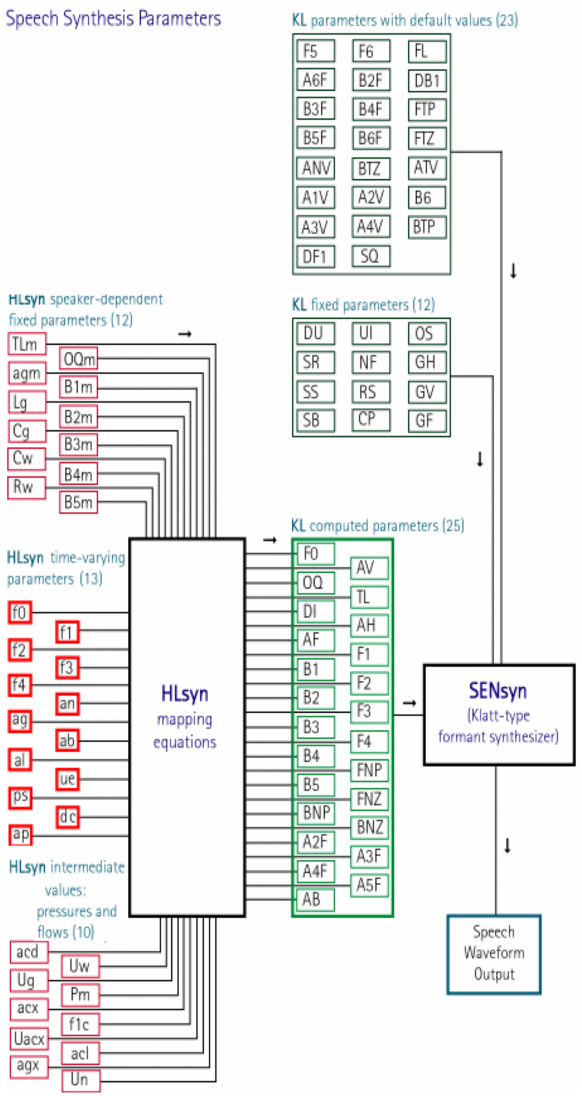


FIGURE 1 Block Diagram of Mapping Relations. Reprinted with minute modifications, Owned by Sensimetrics Corp. HLSyn

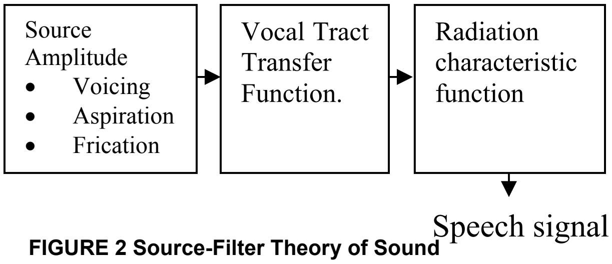
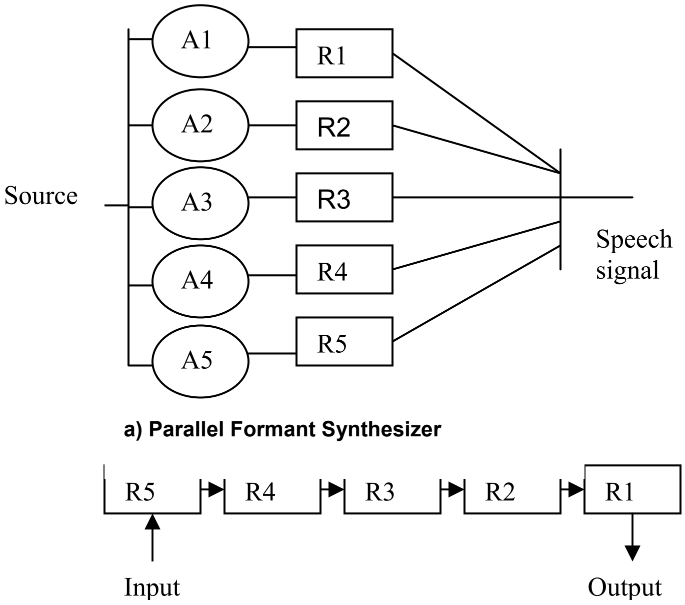

Source
Amplitude
• Voicing
• Aspiration
• Frication
Vocal Tract Transfer Function.
Radiation characteristic function
Speech signal

FIGURE 2 Source-Filter Theory of Sound Production

According to this theory, one or more sources (voicing, Aspiration, Frication) are activated by the build-up of pressure in lungs. When a sound source excites the vocal tract, it acts like a resonating system.

So firstly the vocal tract and then the radiation function, filters the waveform generated by source.

Initially two kinds of configurations were used by synthesizers to model the source filtering effect namely **Parallel Formant Synthesizers** and **Cascade Formant Synthesizers**. In parallel configuration transfer functions of resonators are connected in parallel (as shown in fig. 3a). Each resonator is preceded by an amplitude control, which determines the amplitude of the output spectrum for both voiced and unvoiced sources. Cascade formant synthesizers simulate the formant resonators by connecting transfer functions in cascade fashion (See fig.3b).



a) Parallel Formant Synthesizer

b) Cascade Formant Resonator

FIGURE 3 Configuration of Parallel and Cascade Synthesizer.

The main advantage of using cascade configuration is the automatic calculation of the amplitudes of formants. Moreover it is more accurate in results than parallel resonator. Its disadvantage is its inability to produce the sounds like fricatives and plosives (Dennis H. Klatt, 1980, p.971-973).

Klatt (1980) presented a new type of synthesizer formed by the combination of both, parallel and cascade configuration, to get the accuracy of cascade resonators (for sonorant sounds) and flexibility of parallel synthesizer (for non-sonorant sounds) (Dennis H. Klatt, 1980, p.971-973).

3.4. Parameter Setting

Functionally HLSyn parameters could be divided into three categories namely, oral constriction parameters, Glottal constriction parameter and Formant parameters. Time varying trajectories of these parameters specify the production different kinds of sounds.

3.4.1. Formant Parameters Setting

Formant parameters (f1,f2,f3,f4,f5) are calculated by **copy-synthesis.** This is the approach of calculating parameters by analyzing the recorded speech, with the help of the spectrograms (Jenolan Caves, 1998, p1-24).

Sampling rate is one of the major factors for the determination of the quality of synthesized speech. For vowels variable length sampling rate is used i.e. obtaining the sample at every point where formant frequencies (specially f3) has local maxima or minima. And for constants, fixed sampling rate of about 10ms is used.
Oral constriction parameters (al, ab) and Glottal constriction parameters (ag) are then estimated by mimicking the articulation (place and manner) and phonation (voicing and aspiration) respectively, for the particular sounds to be synthesized.

3.4.2. Oral Constriction Parameter Setting

Oral constriction parameters (al and ab) specify the type of articulator and the manner of sound to be produced. Place of articulation is implicitly calculated by HLSyn, with the help of the coordination of formant and articulatory parameters.

3.4.2.1 Parameter Setting For Stops

Production of stops includes a complete closure caused by the rapid movement of articulator followed by the release. The closure of **Labial** Consonants is modeled by the HL parameter ‘al’ (cross sectional area of Lips). **Dental, alveolar and alveo-palatal** consonants are formed by the blade of tongue—specified by HL parameter ‘ab’ (cross sectional area of tongue blade). To uniquely specify the place of tongue blade, formant parameters describe the configuration of oral cavity, and ‘ab’ specifies the area of the constriction formed by the tongue blade. There is no explicit parameter for the control of tongue body, so the constriction formed at the place of **velum** is implicitly calculated by HLSyn from formant parameters using the relation

$$acd = 12.5 * \left((1080 – f1) / 180 \right)^2 - 12.5$$
For male speakers

$$acd = 8.8 * \left((1280 – f1) / 180 \right)^2 - 8.8$$
For female speakers

Where acd = Area of constriction formed at the place of velum (HLSyn User Manual,p.8).

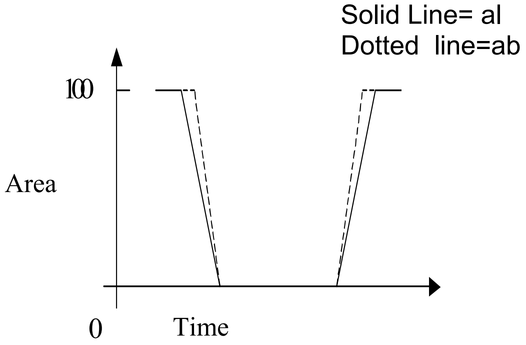


FIGURE 4 Trajectory of al and ab for labial and oral consonants.

At the beginning of stop when there is no constriction, the value of ‘ab’ is set to 100 sqr-mm (i.e open). Then there comes the closure part and constriction area is rapidly set to 0 sqr-mm. The rate of change of the movement of lips in case of labial consonants is observed to be lesser that the rate of change of tongue as in the case of oral consonants. The rates used for lips closure is 100 sqr mm / 10ms and that of tongue is 100 sqr mm / 20ms (Williams, D.R, 1996, p.2219-2222). Followed by the closure is the release, in which the closure areas are increased to 100 sqr-mm, using the same rates as described above.

3.4.2.2 Parameter Setting for Fricatives

Forming partial constriction at certain place of the oral tract forms fricatives. Again the place of articulation is specified by formant parameters as in the case of consonants and constriction by ‘ab’. The value of constriction in the segment of fricative is dependent upon its context. Let us consider the context, in which fricative is preceded by vowel and proceeded by consonant. At the vowel the constriction of ab is 100 sqr-mm, which is decreased to about 25 sqr-mm at the beginning of the segment of fricative. During the segment of fricative it continuously decreases to about 10 sqr-mm. After this segment, as there is a consonant, so this constriction is decreases to 0 sqr-mm as shown in fig.5.

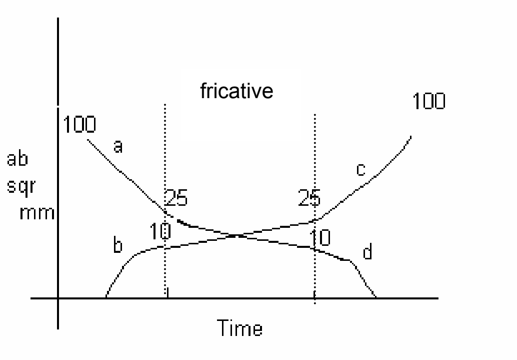


FIGURE 5 Trajectories of ‘ab’ for fricatives in different contexts

- **a) Vowel to Fricative**
- **b) Sop to Fricative**
- **c) Fricative to Vowel**
- **d) Fricative to Stop**

1.1.1.1 Parameter Setting For Affricates

Affricates are produced by the complete closure at first, followed by a frictional noise, just like fricatives. The closure is specified by setting ‘ab’ to 0 sqr-mm and then it is moved to 40 sqr-mm, to represent the frictional noise. After the noise, trajectory of ab follows the same analogy as in the case of fricatives. Fig. 6 shows the trajectory of affricate followed by vowel.

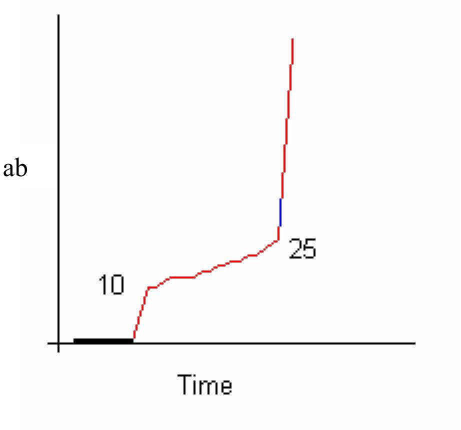


FIGURE 6 Trajectory of ab for Affricate

1.1.1.2 Parameter Setting For Liquids and Glides

In the present version of HLSyn, liquids and glides are synthesized by specifying only formant parameters. This approach ignores the minute (but not negligible) modifications on formants, caused by tongue blade. Next versions are expected to have this change in this approach (HLSyn User Manual, p.28).

1.1.1.3 Parameter Setting For Vowels

Vowels are produced with no constriction on the vocal tract, so they are synthesized by specifying only formant parameters.

1.1.2. Glottal Constriction Parameter Setting

Voicing and Aspiration is controlled by the trajectory of 'ag' (Cross-sectional area of glottal opening). For aspirated voiceless stops, ag increases from a default value of 4 sqr-mm to just above 15 sqr-mm at stop closure and then to about 28 sqr-mm at stop release and voice-onset-time, is the time between stop release and the point at which ag falls below 15 mm sqr-mm (Williams, D.R, 1996, p.2219-2222). For voiced stops when there is a closure, the volume of the vocal tract is increased to provide room for voicing; this is controlled by increasing the value of parameter 'ue' during the time of closure (HLSyn User Manual, p.61).

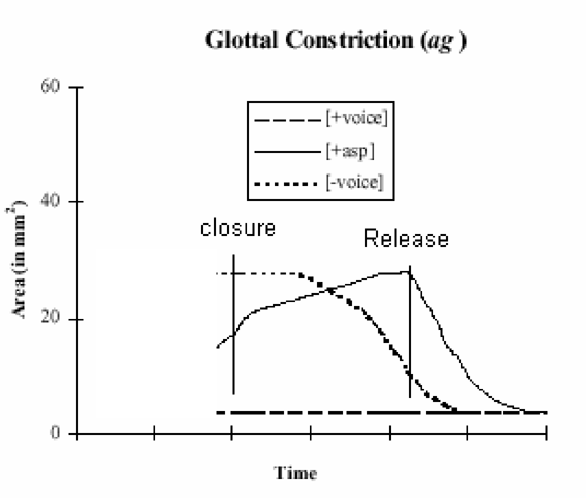


FIGURE 7 Control of voicing and Aspiration via 'ag'

The aim of this paper is to analyze the quality of non-nasal and non-aspirated stops of Urdu, synthesized using HLSyn.

4. METHODOLOGY

Synthesizing speech using HLSyn means to convert the acoustic and articulatory features of given utterance of speech into HL parameters. Following the set of principles outlined in above section, HL parameters of first four Urdu numbers ("ek", "d̲o", "t̲in", "tʃar") are calculated and is subject to a simple perceptual test. Fifteen persons, who have Urdu as there first language, are asked to identify the words along with the question "Does it seem strange"? The aim is to get an idea about the quality of synthesized words.

Another perceptual experiment is designed to test the quality of non-aspirated Urdu stops. Non-aspirated stops are characterized by 'place of articulation' and the presence or absence of 'voicing'. So the test data consists of voiced and non-voiced stops forming closure at labial (b, p), dental (d̲, t̲), alveolar (t, d) and velar (g, k) positions. These stops are synthesized by combining them with some vowel (chosen at random) to assist the perceptual understanding, as isolated stop cannot be perceived. Again fifteen persons (having Urdu as their first language) are asked to identify the stops.

5. RESULTS

Table 1 shows the results of first perceptual test in which different persons were asked to identify the Urdu numbers “ek”, “d̲o”, “t̲in”, “tʃar”.

TABLE 1 Results for the identification of Urdu numbers

Total no. of persons = 15

Urdu Number	**Correctly Identified (No. of Persons)**	**Seems Strange? (No. of Persons)**
‘ek’	15	1
‘d̲o’	15	0
‘t̲in’	15	0
‘t̲ʃar’	15	2

Table 2 shows the results of second perceptual test in which persons were asked to identify the different stops associated with some vowel.

TABLE 2 Results of The Identification of Stops

Total No. of persons = 15
Vertical axis = stops synthesized
Horizontal axis = stops perceived

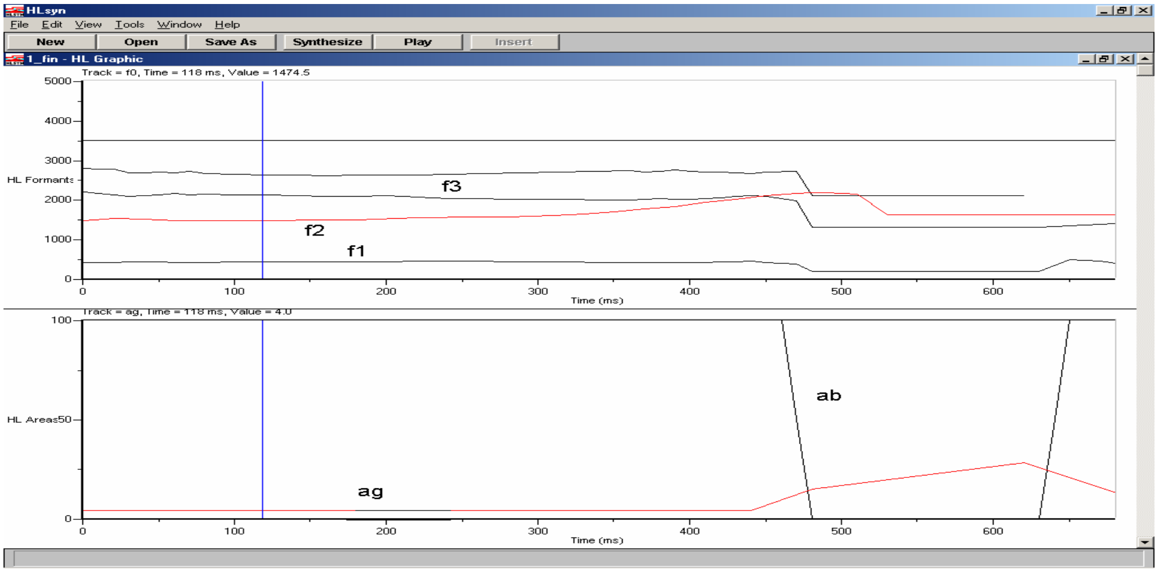
	p	**b**	**t̲**	**d̲**	**t**	**d**	**k**	**g**
p	14		1					
b	3	12						
t̲			13		2			
d̲				15				
t			3		12			
d				4		11		
k							15	
g							2	13

Figure 8 shows the Trajectories of the HL articulatory and formant parameters along with the spectrograms of originally recorded and synthesized Urdu numbers (See fig.7).

The Urdu sound 'ek' in Figure 8(a) starts with a vowel with no oral constriction, thus leading the parameter 'al' and 'ab' to 100 sq-mm (open state). At this point 'ag' is set to modal constriction for producing voiced sound. Next comes the consonant 'k'. There is no explicit HL parameter for specifying velar constriction, so this constriction is achieved by the coordination of formant parameters along with 'ab', i.e., by making downward transition of f2 and f3 of about 1000Hz and of f1 by 100 Hz. The value ag (constriction area of glottis) is increased sufficiently before entering the segment of 'k' because it is voiceless. The value of 'ag' is kept on increasing upto 28 until the time of closure, to prepare for VOT. Then trajectory of 'ag' is controlled for producing the desired VOT. The time of VOT is the time between the current state (where ag = 28 sqr-mm) and the state where ag will become less than 15 sq mm. So more abruptly the ag decreases, lesser the VOT and vice versa. As k is a voiceless stop and voiceless stops has VOT relatively greater than of voiced stops, so ag is decreased slowly.

Now consider the production of Urdu sound 'd̲o' in Figure 8(d). It starts with a voiced consonant, forming closure by the blade of tongue at the dental position. From the start, ab is set to zero, representing the closure of stop. Here 'ag' set to model constriction for the voicing of dental 'd' and then it is increased until the closure of stop. At this time the volume of the vocal tract starts to increases to make the room for voicing. This is modeled by increasing the value of HL parameter ue (volume of vocal tract) until the release. After the release, 'ag' is set back to its modal constriction very abruptly, to make the VOT of this voiced stop small.

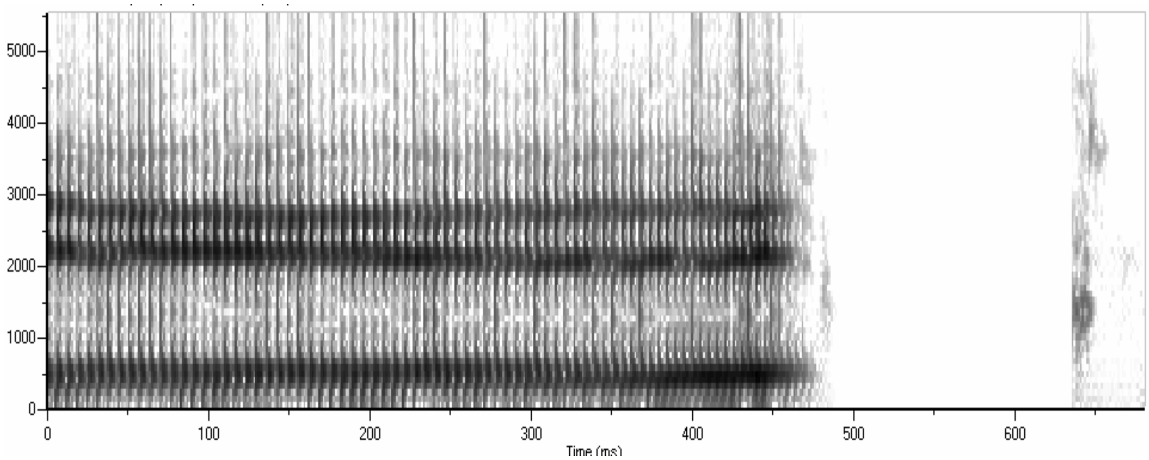
Urdu sound 't̲in' in Figure 8(g) is synthesized with the similar analogy as described above but with the difference of nasal stop /n/ at the end where 'an' is set to 40 sq-mm to couple the oral tract.



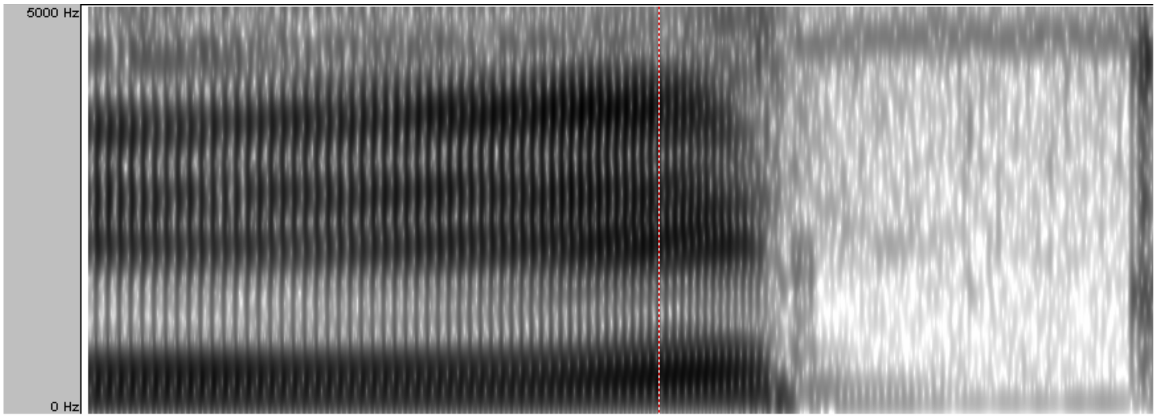
Track = f0, Time = 118 ms, Value = 1474.5

Track = ag, Time = 118 ms, Value = 4.0

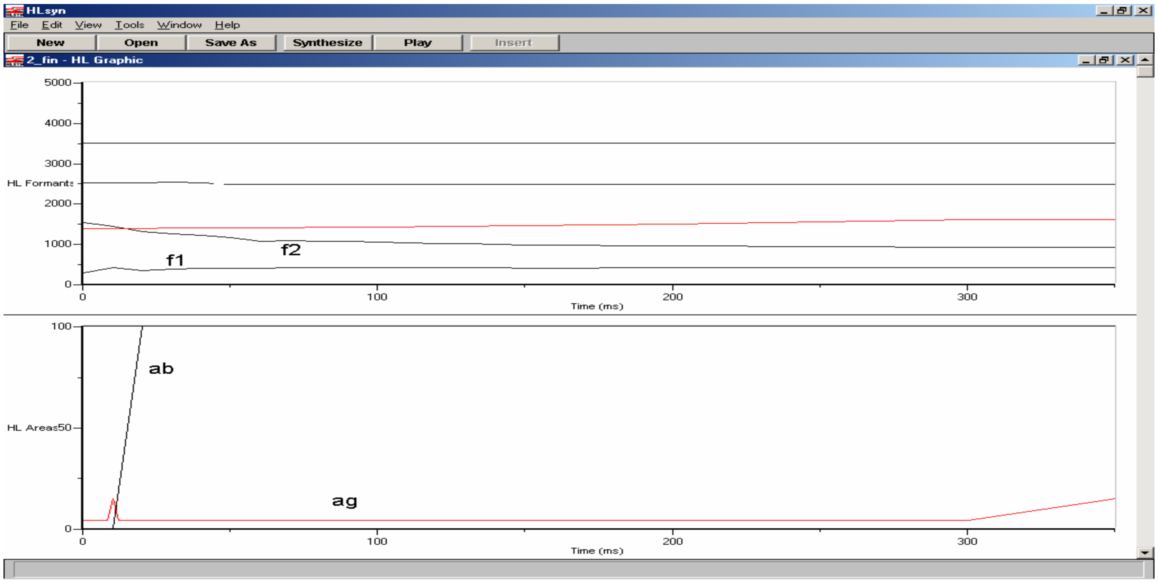
a) Trajectory of HL Parameters of Urdu Word ‘ek’



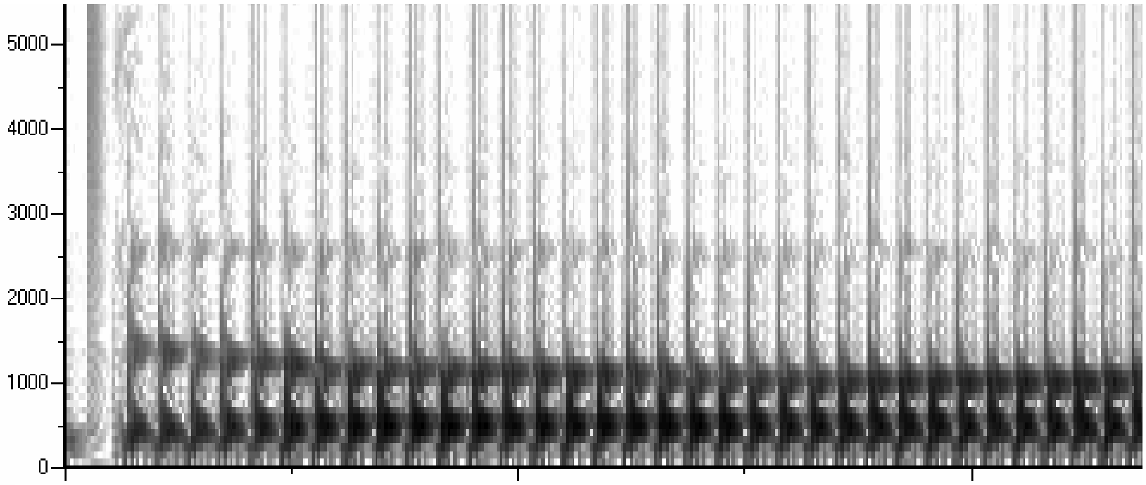
b) Spectrogram of Synthesized Urdu word ‘ek’



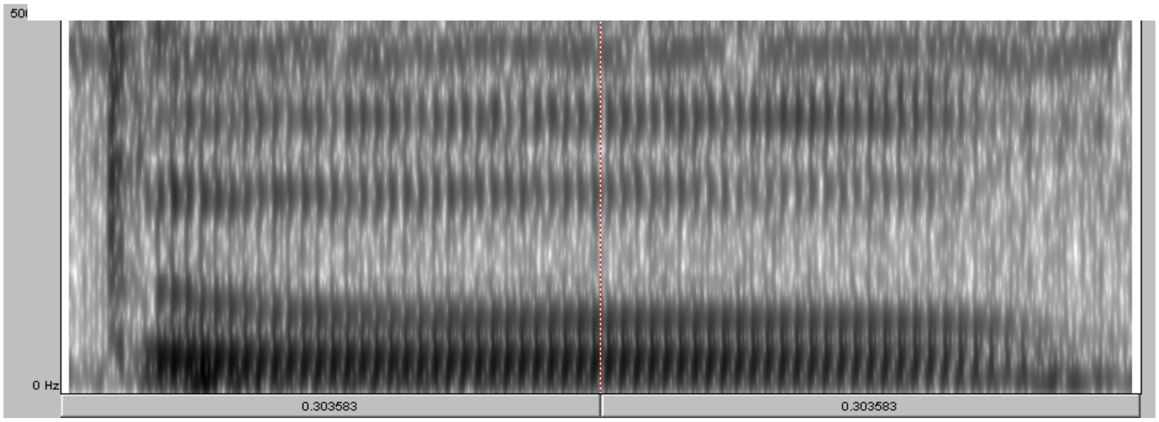
c) Spectrogram of recorded Urdu number ‘ek’
(Ignore the continuous noise during the time of closure)



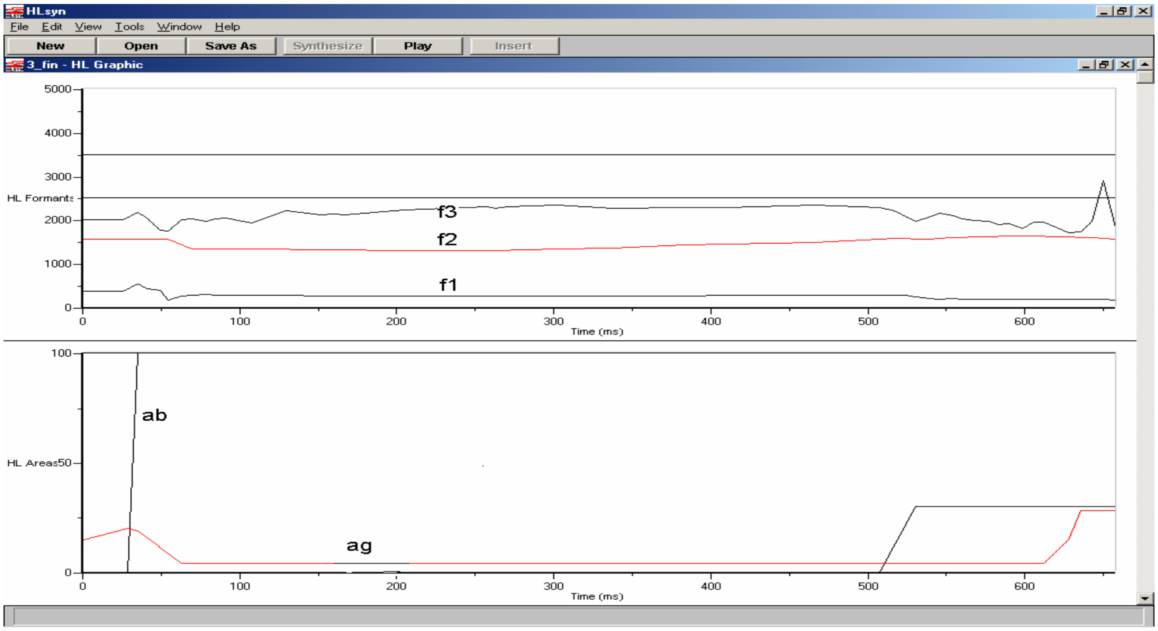
d) Trajectory of HL parameters of Urdu word 'do'



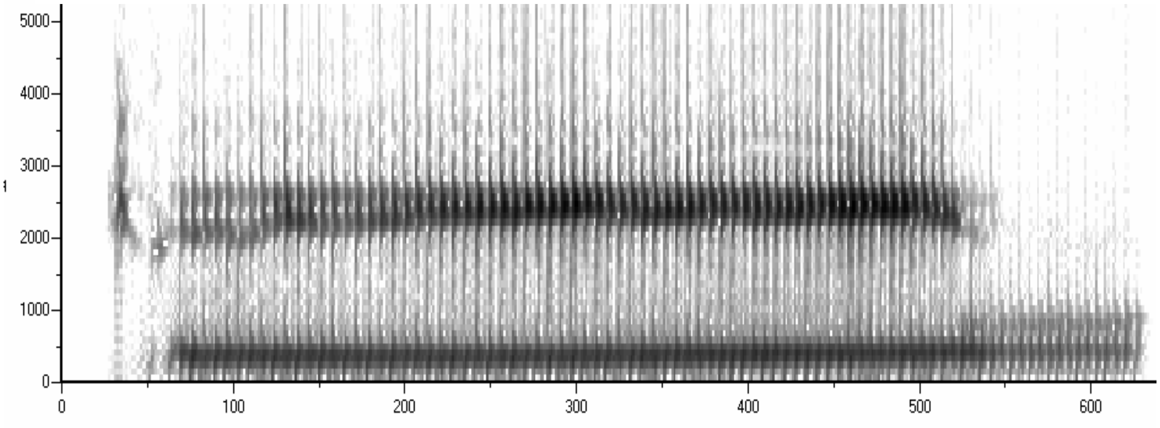
e) Spectrogram of Synthesized Urdu word 'do'



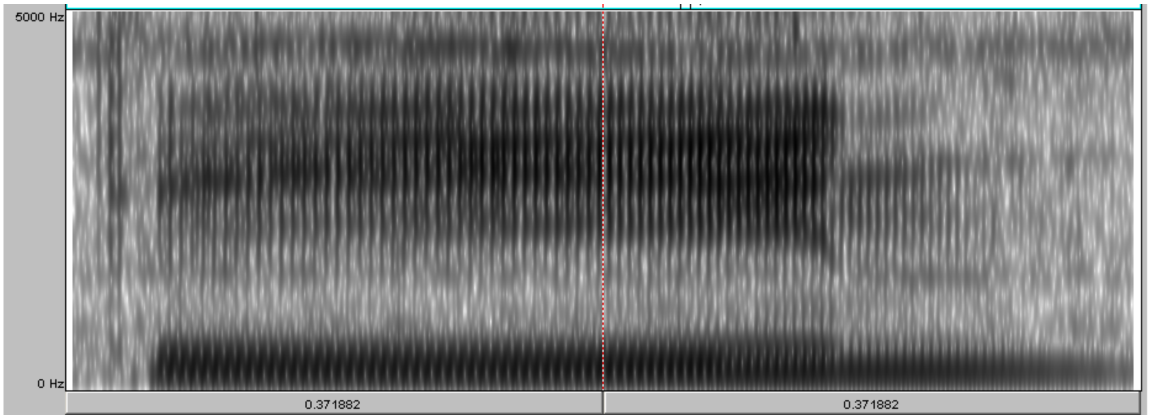
f) Spectrogram of recorded Urdu number 'do'
(Ignore the continuous noise during the time of closure)



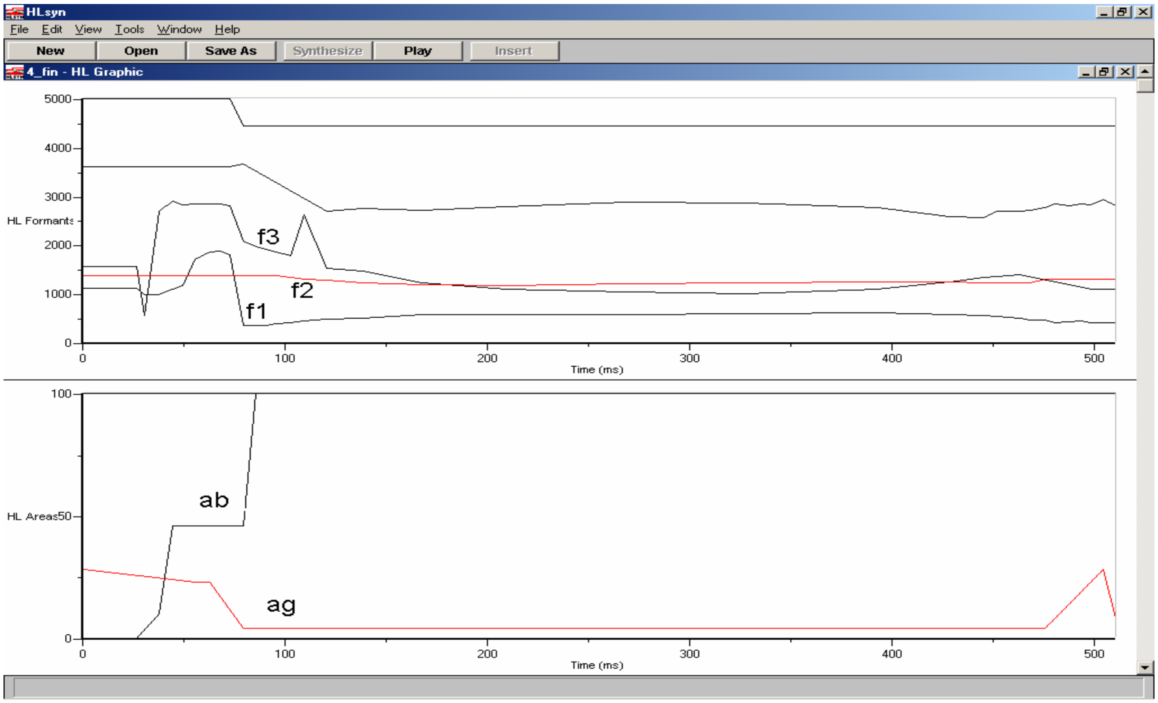
g) Trajectory of HL parameters of Urdu word 'ṯin'



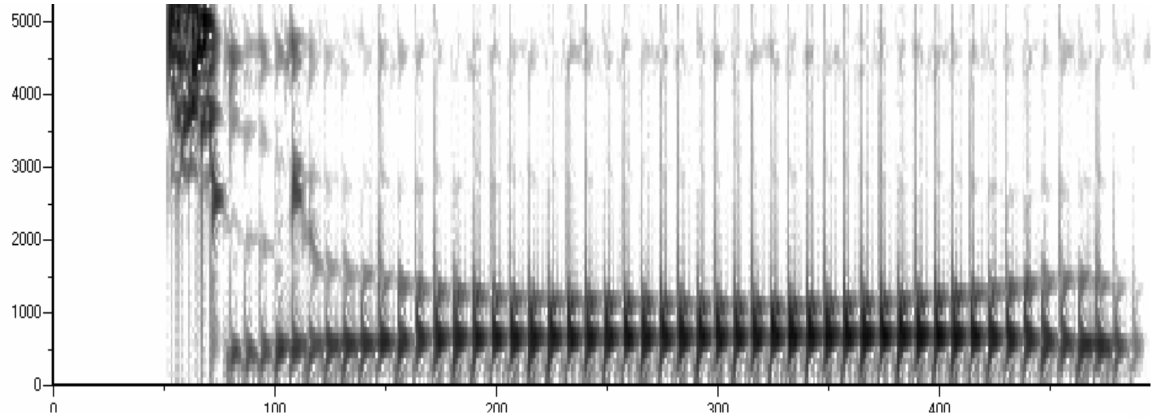
h) Spectrogram of Synthesized Urdu word 'ṯin'



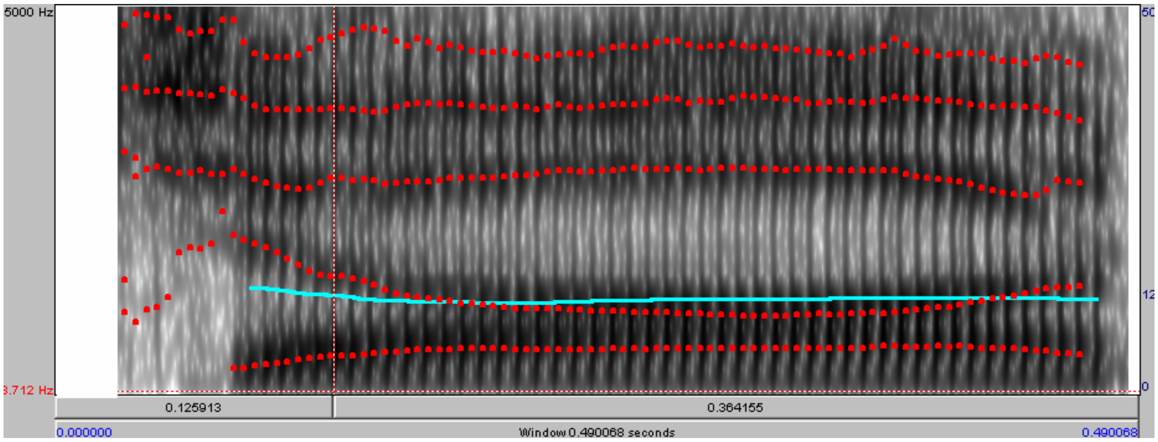
i) Spectrogram of recorded Urdu number 'ṯin'



j) Trajectory of HL parameters of Urdu word ‘tʃar’



k) Spectrogram of Synthesized Urdu word ‘tʃar’



l) Spectrogram of recorded Urdu number 'tʃar'
(Ignore the continuous noise during the time of closure)

FIGURE 8 Showing the HL trajectories, Synthesized and Original Spectrograms (a-l) of Urdu numbers "ek", "do", "tin", "tʃar" respectively.

6. DISCUSSION

Results of first perceptual test in which Urdu numbers "ek", "do", "tin", "tʃar" was given for identification, yields excellent results with the 'Identification accuracy' of 100 %. Although some persons say that it feels little strange for sounds "ek" and "tʃar" (See table 1).

However second perceptual test has produced relatively less accuracy, although the same method was used to synthesize sounds for both tests. The accuracy of stops in percentages is given in table 3.

Table 3 Accuracy of Perceived Stops

Sound Synthesized	Accuracy
pl	93%
bl	80%
ti	87%
do	100%
t shwa	80%
d shwa	73%
ek	100%
eg	87%

An interesting point to note in the above table is that the stop 'k' and 'd' which were associated with vowels 'e' and 'o' respectively, forms legal syllable of Urdu

Language and thus have perceived very accurately. While the others who do not form the legal syllables are perceived less accurately.

This observation suggests that the accuracy of perceived stops in legal syllables of the language is more than that of non-legal syllable.

This deduction even enhances the possibility of achieving better quality of speech when this methodology would be used in text-to-speech engine, for producing Urdu sounds.

Comparing the spectrograms of original and the synthesized voice also give us the idea that how correctly we have estimated the configuration vocal tract (Pamela Jean, 1995, p. 359-361). Figure 8 shows that the spectrograms of both originally recorded and synthesized speech are similar except few minute changes.

7. CONCLUSION

Accuracy achieved by this Hybrid-approach i.e., Formants parameter calculation by copy-synthesis and Articulatory parameter estimation by mimicking the articulatory configuration, yields fair quality for non-nasal and non-aspirated Urdu stops.

8. FUTURE WORK

The main hindrance for the automated generation of HLSyn parameters lies in the approach of calculating formant parameters using copy-synthesis.

Because of the change in place of articulation, formant transitions occur at boundaries of vowel and stops. Synthesizing this boundary successfully with HLSyn requires the values of formant parameters during the transition. So the rule engine will have states representing the formant parameter values of current segment and rules for converting formant parameters of one segment to another. The formation of these rules along with the calculation of formant parameters of each segment would be the next step in the development of text-to-speech Engine (I.H Witten, 1982, p.169-177).

9. REFERENCES

Bagshaw P.C. 1998. “Phonemic transcription by analogy in text-to-speech synthesis: Novel word pronunciation and lexicon compression.”. Computer Speech & Language vol. 12.

Dennis H. Klatt . 1980. “Software for a Cascade/Parallel Formant Synthesizer.” Acoustical Society of America.

HLSyn (High Level Speech Synthesizer) Reference Manual.1999. Sensimetrics Corporation.

I.H. Witten .1982. *Principles of Computer Speech.* Academic Press.

Jenolan Caves. Nov 1998. “A Hybrid Approach to High Quality Formant Synthesis using HLSyn”. International Workshop on Speech Synthesis Australia.

Pamela Jean. 1995 . “ Voice Quality Analysis of Male and Female Spanish Speakers”. Speech Communication Volume 16.

Williams, D.R. 1996. “Synthesis of initial (/S/-) stop-liquid clusters using HLSyn.” IEEE Catalogue No. 96TH06. Fourth International Conference on Spoken Language vol. 4.