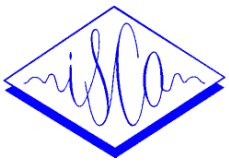

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

TOWARDS SYNTHESIS OF HINDI CONSONANTS USING KLSYN88

Shyam S. Agrawal

**Central Electronics Engineering Research Institute Centre,
CSIR Complex, Hill Side Road, New Delhi 110 012**

Kenneth Stevens

Research Laboratory of Electronics, MIT Cambridge, MA 02139

ABSTRACT

This paper presents results of synthesis of Hindi consonants using KLSYN88 speech synthesizer. All frequently occuring 29 consonants of Hindi were synthesised in the initial position of CVC syllables. The central vowel /a/ and the final consonant /l/ was always used to make the syllables into meaningful Hindi words. The words spoken by a standard Hindi male speaker were digitised at 10K samples per second using a VAX 750 computer system. Techniques employed for analysis include short term DFT magnitude spectrum, variations in formant frequencies, fundamental frequency and amplitude, and display of digital spectrograms. Quantitative acoustic parameters required for synthesis of phonetic features of phonemes were determined. The consonants were synthesised in combination with vowel /a/ to generate CV syllables and concatenated with the syllable /al/ to form CVC type synthetic words. A number of synthsiser control parameters were interactively varied for each sound till a satisfactory quality of synthetic speech and distinction among all the consonants was achieved. Special attention was paid to the synthesis of stops and affricates with various voicing and aspiration features. These sounds required careful selection and timing of source parameters. The spectral characteristics of synthesised and original sound segments were also compared to further improve the quality of synthetic speech. The results indicate that all the consonants of Hindi speech can be synthesised with natural quality. The perception tests indicate that the intelligibility scores for both types of speech are nearly the same. It is envisaged that further improvements are possible by careful control of synthesizer parameters and improvements in the aspiration source.

10.21437/ICSLP.1992-33

I. INTRODUCTION

High quality synthesis of speech using computers has been a subject of research for the speech engineers for the past several decades [1,2]. Two major techniques i.e. articulatory synthesis and the formant synthesis have been developed for this purpose. In articulatory synthesis an attempt is made to model faithfully the mechanical motions of the articulators and the resulting volume velocity and sound pressure in lungs, larynx, vocal tract and nasal tract. The formant synthesizers attempt to approximate directly the speech waveform and spectrum by a simpler model formulated in acoustic domains. Several versions of the articulatory and formant synthesizers have been developed to synthesize different spoken languages.

In the recent past an improved version of Klatt's synthesizer, called KLSYN88 has been successfully used for the synthesis of voice quality variations among the female and male talkers of American English and other languages also [3,4,5]. Since the KLSYN88 possess the flexibility of simulating the sound generating system i.e the vocal tract and also designed configuration of the vocal source, it is very convenient to control the acoustic parameters by an experimenter. It was therefore envisaged that with the help of KLSYN88 it is possisble to synthesize highly intelligible and natural Hindi speech. The special features of Hindi speech sounds such as distinction between voiced and unvoiced, aspirated and unaspirated stops and affricates etc.could be faithfully reproduced.

This paper presents a study to synthesize Hindi speech sounds (all consonants in vowel context /a/) using KLSYN88.

II. THE CONSONANTS OF HINDI SPEECH

Hindi speech possess a very rich set of about 35 consonants. Out of these, 29 consonants are very frequently used in all positions of a syllable or a word [6]. The articulatory classification of these consonants are shown in Table I. It may be observed that the consonants of Hindi are different from English and any other European languages. The most significant differences are in stops and affricates which use both voicing and aspiration to distinguish them [6,7,8]. The features provide a four way manner contrast of the stop and affricate consonants i.e. unvoiced unaspirated, unvoiced aspirated, voiced unaspsirated and voiced aspsirated.

Table I. Classification of Hindi Consonants.

Stops/Affr	Bil	Den	Ret	Pal	Vel
UV UAs	p	t	ʈ	t ʃ	k
UV As	p^h	t^h	$ʈ^h$	t ʃ ^{h}	k^h
V UAs	b	d	ɖ	dz	g
V As	b^h	d^h	$ɖ^h$	dz^h	g^h
Nasals	m	n			
SemiVowels	w			y	
Liquids		l	r		
Fricatives		s		ʃ	h_{GL}

III. PROCEDURE FOR ANALYSIS/SYNTHESIS

The consonants were used in the initial position of CVC syllables. The central vowel /a/ and the final consonant /l/ was always used to make syllables into meaningful CVC type Hindi words. These were recorded by a male speaker having standard Hindi speaking background and mother tongue. The recorded samples were low pass filtered at 4.8 KHz and then digitized at 10K samples per second using a VAX 750 computer system attached with suitable A/D, D/A conversion systems and audio equipment. The speech samples were analyzed using spectral analysis techniques to study the audio waveform, digital spectrograms, formants and F0 variations in time, short-term DFT magnitude spectrum of a windowed segment and the average amplitude spectrum. A number of acoustic parameters necessary for synthesis of different phonetic features for the Hindi phonemes were quantified.

For the purpose of synthesis, the default configuration of the KLSYN88 synthesizer was used as the basic source of parameters. It uses 60 constants and variables. Depending on the nature of a sound its individual parameters describing the source and tract behaviour were varied in 5 msec steps of time. For generating the voice source parameters, the KLGLOTT88 voicing source model was used. The source control parameters which were frequently used for generating Hindi consonants include AV (amplitude of voicing), F0 (fundamental frequency), OQ (open quotient of the glottal waveform), TL (tilt of the voicing source spectrum) and AH (amplitude of aspiration noise or breathiness). The frication source was used for the parallel synthesizer. A very careful control and timing of these parameters was required in relation to the dynamically changing vocal tract parameters and to generate

mixed excitation sources for sounds such as voiced aspirated consonants. In the cascade vocal tract the parameters which were varied mostly include first five formant frequencies, their bandwidth and amplitude and a nasal pole-zero pair. In the case of a parallel vocal tract model a bypass amplitude path AB was used along with five formant resonators.

IV. RESULTS

a) Stop and Affricate Consonants: As shown in Table I these consonants in Hindi could be classified into four categories viz. unvoiced unaspirated, unvoiced aspirated , voiced unaspirated, and voiced aspirated. Acoustically, these classes could be distinguished in CVC syllables by the presence or absence of features such as the plosive burst, Silence, VOT, aspirated noise, voice bar, vowel transitions etc. For example, the sequence of features seen in the spectrogram of the word /d^{h}al/ using a voiced aspirated plosive (see Fig. 1) are the presence of a voice bar, silence (VOT), a plosive spike, voiced aspirated noise and the formant transitions. Differences among these consonants across the places of articulation are primarily reflected in the spectra of the burst and formant transitions of the target vowel (see Table II). In affricates, a short duration frication noise is produced immediately after the burst in addition to all those features that are required for stop consonants. In addition to these features there are some differences in the spectra of the aspiration noise as well as in VOT. For generating natural quality synthetic speech, it is necessary to study the differences in the spectral characteristics of these features. In the following paras these aspects have been discussed.

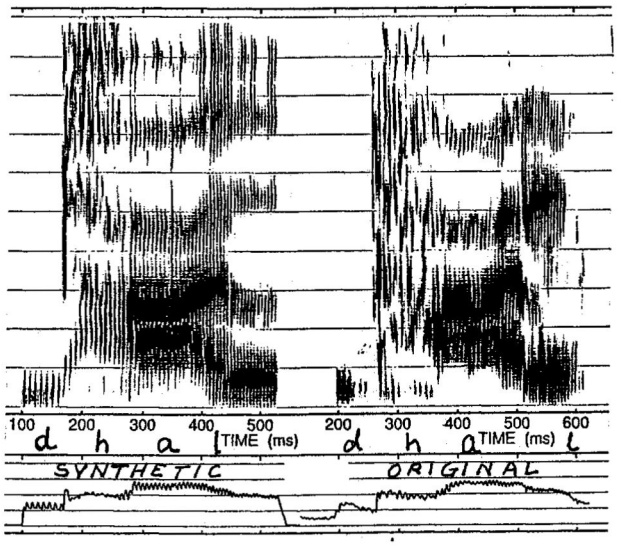

Fig. 1. Wideband spectrograms of the Hindi vowel /d^{h}al/.

TABLE II. Initial values of formant frequencies (in Hz) used to distinguish the places of articulation of stop and affricate sounds.

Place of Articulation	Formant 1	Formant 2	Formant 3	Formant 4
Bilabials	500	1000	2200	3500
Dentals	450	1600	2500	3700
Retroflex	450	1800	2700	3700
Palatals	400	2100	2800	4000
Velars	550	1500	2400	3600

(i) Differences in burst frequency: As it may be seen from Fig. 2, the burst spectra of bilabials and dentals is more flat as compared to those of retroflex and velar sounds. The burst of /p/ and /t/ is quite sharp (about 5 msec to 10 msec in duration) having a flat spectrum with slight variation in energy in the formant regions. The energy is roughly distributed as ($a_2f > = a_3f > ab$) for /p/ and as $a_5f > = a_4f > = a_2f > a_3f >$ ab for /t/. The major concentration of energy for /ʈ/ is in the upper middle region of the frequency scale ($a_3f > a_2f > a_4f$). In case of /k/ and other velar sounds, double or sometimes triple burst is obtained. The energy is concentrated mostly in the middle region of the frequency spectra ($a_2f > a_4f > a_3f$). In the case of voiced stops, the overall intensity of the burst is reduced.

As shown in Table III the burst frequency for these sounds was generated by exciting the fricative noise source for the desired duration (5 to 10 msec) in the parallel synthesizer. The amplitude values of the formants were varied in the ratio's described in the preceding paragraph. OQ and TL were kept high. Voicing and aspiration noise was added for the respective categories of sounds.

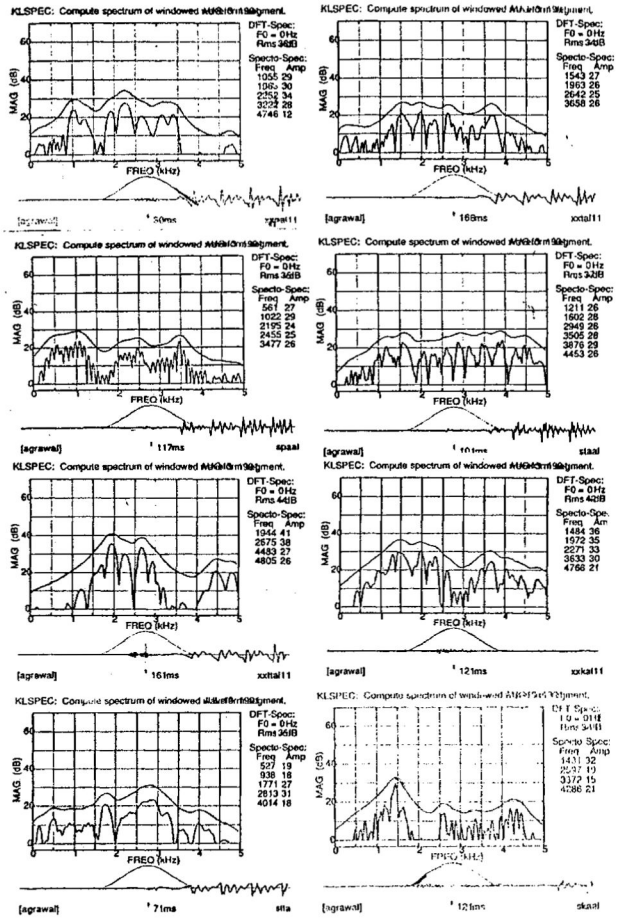

Fig. 2. Burst of Hindi stop consonants (upper diagram spectrum of synthetically generated burst, lower diagram spectrum of original burst).

(ii) Voice bar: The voice bar is a low frequency resonance and indicates pre-voicing during the closure period of the voiced unaspirated and voiced aspirated stop and affricate sounds. During the production of these consonants, the glottis remains closed. Maintainance of pressure drop across the vocal folds leads to voicing throughout the closure duration. The center frequency of voice bar varies between 200-300 Hz and the amplitude of its first resonance is high after which it decays very fast having very low or no higher resonance. There is a tendency of increasing the F0 of voice bar frequency from front to back sounds. For voiced aspirated plosives there is decay of amplitude and break in the voice bar just before the aspiration starts. This may be due to heavy building up of air pressure for the release of the burst as well as aspiration.

The voice bar has been produced by exciting the voice source at a given fundamental frequency (it is in general lower than the F0 of the following vowel). As shown in Table III, AV is kept low (around 50) and OQ and TL high. The bandwidth of all the higher formants is kept high. The duration of the voice bar also changes depending upon the manner and place of production of the consonant. (see Table IV).

(iii) VOT: The voice onset time for unvoiced as well as voiced consonants belonging to different places of articulation is different. For voiced plosives VOT may be defined as duration between the end of the initial voice bar and beginning of the vowel. The VOT appears to increase along

the continuum from bilabials to velar place of articulation. It is greater for aspirated sounds as compared to unaspirated sounds (Table IV).

(iv) Aspiration noise: Aspiration is an important feature in the distinction of Hindi stops and affricates. It is used during the production of unvoiced as well as voiced aspirated stop and affricate consonants of Hindi speech. It has been observed that glottis is widely open during unvoiced aspirated sounds whereas it is comparatively less open during the voiced aspirated sounds [7].

In the latter case, voicing is also present throughout their duration. The duration of the aspiration is about 50-70 msec. Fig. 3 shows the spectra of a typical unvoiced aspirated and voiced aspirated sounds. By comparing these diagrams one can see the presence of voicing components and better shaping of the formants during the voiced aspiration portions. Table III shows the value of aspiration source and some other parameters for synthesizing the two categories of aspirated sounds. When aspiration and voicing are both present in the case of voiced aspirated sounds, then the amplitude of aspiration noise is modulated by a 50% square wave amplitude modulation that increases the noise during the most open part of the cycle. Hence to achieve reasonable quality of voiced aspiration , the value of OQ, TL and BW's were also appropriately adjusted (see Table III).

(v) The noise source for affricate sounds: There are four affricate sounds of Hindi speech produced near the hard palate. The manner of production of these sounds is similar to those of plosives and in addition a noise source is used immediately after the closure period. For example, in the case of a voiced aspirated affricate /dz^h/, a voicebar is followed by a fricative noise. Its energy concentration is between 3 and 4.5 KHz,

TABLE III. Some parameters for synthesis.

(i) Stops and affricates.

Feature	F0	AV	OQ	TL	AH	AF
Burst (UV/V)	0	0	0/80	0/15	0/45	60
Voicebar	100	52	80	15	0	0
Asp. (UV/V)	0/105	0/50	80/60	15/12	0/55	0
Fric. (UV/V)	0/110	0/45	80	15	0	57
Vowel (tr)	110	55	70	10	0/50	0
Vowel	130	60	50	0/5	30	0

(ii) Fricatives.

Ph.	AH	AF	F1	F2	F3	F4	F5
/s/	0	50	500	1500	2550	3650	4500
/ /	0	60	550	1700	2000	3400	4500
/h/	50	0	1000	1700	2600	3500	4500

(iii) Nasals.

Ph.	F0	AV	F1	F2	F3	FNP	FNZ
/m/	120	55	250	1000	2500	1300	1500
/n/	125	55	250	1500	2600	1600	1800

(iv) Liquids.

Ph.	F0	AV	OQ	TL	AF	F1	F2	F3
/l/	125	55	50	0	0	350	1500	2500
/r/	115	55	80	15	55	700	1500	2500

(v) Semi-vowels.

Ph.	F0	AV	F1	F2	F3
/w/	115	55	400	900	2250
/y/	120	55	300	2400	3100

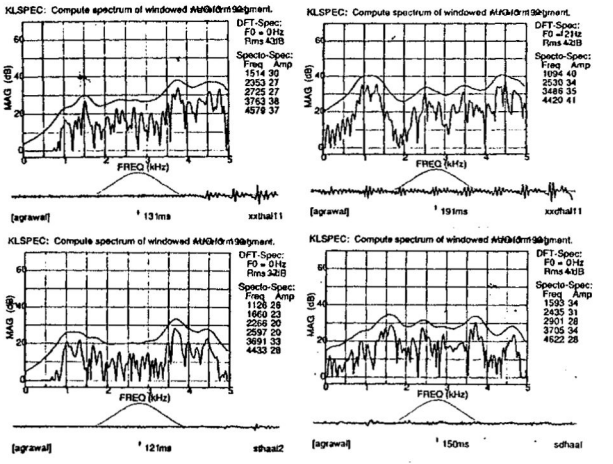

Fig. 3. DFT spectrum of unvoiced aspirated and voiced aspirated sounds /t^h/ and /d^h/ (upper diagram - synthetic, lower diagram - original).

having duration of about 20-30 msec. This is followed by aspiration noise with simultaneous voicing and the vowel transitions. The initial vowel target frequencies (F2 and F3) for these sounds are quite high (see Table II). It may be noted that in aspirated sounds most of the formant transitions takes place in the aspirated region.

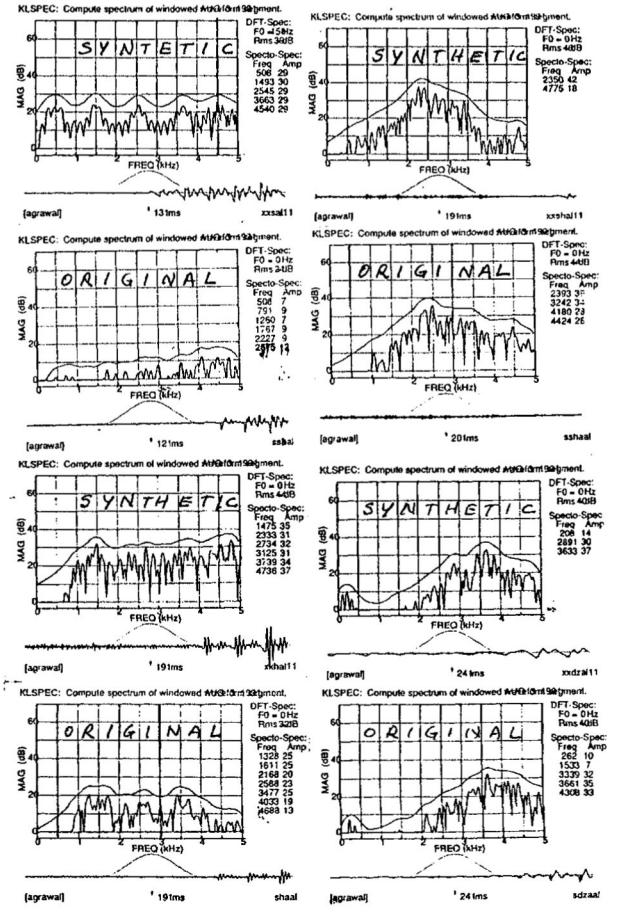

Fig. 4. Spectrum of noise source used for generating Hindi fricative sounds and affricative sounds.

(b) The nasals: The two most frequent nasal sounds of Hindi /m/ and /n/ have been synthesized using KLSYN88. The formant peaks and the nasal poles and zeros of these sounds are different. However the peaks are broader as compared to vowels. Although several pole-zero pairs (at least three) are observed in both /m/ and /n/, a satisfactory distinction among these sounds could be produced by introducing only one pole-zero pair at different frequencies near the F2 region.

(c) The fricatives: Hindi speech possesses a simple set of three fricative sounds only i.e. /s/, /ʃ/, and /h/. Frequency spectra of these sounds have peaks at different frequencies (see Fig. 4). The first two sounds are synthesized by using the fricative source and the parallel synthesizer whereas for /h/ the aspiration noise source was used.

(d) Semi vowels / Sonorant consonants: These sounds have similar characteristics as compared to that of a vowel, except that the formants are broader. AV is kept lower than those for the vowels. In the case of /r/ and /l/ additional poles and zeros occur as a consequence of more complex vocal tract shape. The formant frequencies of these sounds are different from each other. The major difference between /l/ and /r/ is in the first formant. Moreover for the generation of faithful Hindi trill /r/ fricative source has been used to generate impulses similar to the burst in plosives. At the same time OQ and TL were also kept high.

Table IV. Durational characteristics of Hindi stops

(i) Duration of Gap / V.B.

Unv Unasp > Unv Asp > Voiced Unasp > Voiced Asp

Bilabial > Velar > Dental ≥ Retroflex

(ii) Duration of Asp.

Voiced Asp >> Unvoiced Asp

(iii) Duration of Burst.

Voiced Asp ≥ Voiced ≥ Unvoiced Asp > Unv Unasp

Velar > Retroflex ≥ Dental > Bilabial

(iv) VOT

Asp > Unasp

Velar > Retroflex ≥ Dental > Bilabial

(v) Rate of Second Formant Transition

Bilabial > Velar > Dental > Retroflex

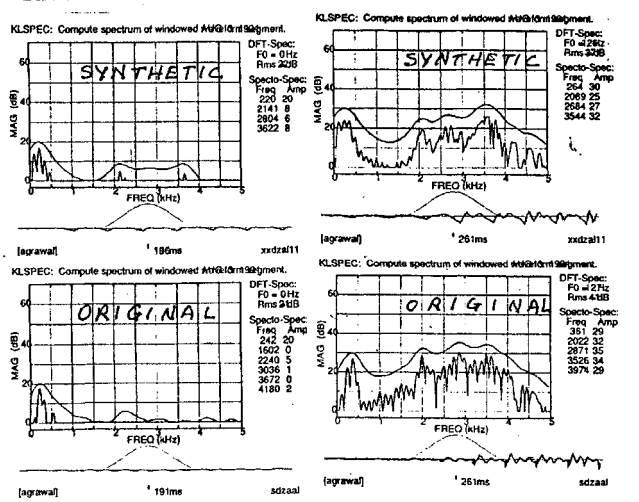

Fig. 5. DFT spectrum of voicebar and vowel transition in typical Hindi syllable.

V. PERCEPTION TESTS

The recordings of Hindi words (CVC type) of original speech and natural speech were presented to the listeners who had Hindi language background. The intelligibility scores of initial consonants for both types of speech are almost equal. However different types of mistakes are made by the listeners for the original and synthetic speech. For example, major mistakes in original speech occur for voiced aspirated and unvoiced aspirated sounds which are confused with similar sounds belonging to another place of articulation, such as retroflex sounds being confused with dentals. In synthetic speech major confusions are in weak fricative and aspirated sounds. Confusions also occur among the nasal sounds. However, the features of sounds are preserved. Generally the confusion are within single feature dimensions. It is envisaged that the above mistakes can be removed and the intelligibility and naturalness improved by carefully controlling the synthesizer parameters.

VI. CONCLUSIONS

The results of the above experiments indicate that the efforts to synthesize Hindi consonants using KLSYN 88 are quite encouraging. It is possible to achieve a clear perception of Hindi sounds and synthesize them with natural perceptual quality.

The quality and naturalness of the consonants can be further improved by optimizing the parametric descriptions. The sounds which need more attention are the aspirated, voiced aspirated, stops and affricates, the glottal fricative /h/ and nasals. The spectral analysis of aspiration source indicates that the distribution of aspiration energy at lower frequency region (500-1000 Hz) should be boosted. The recent approach proposed to do synthesis with large array of parameters (a set of high level parameters) that are related more closely to the articulatory parameters than the acoustically oriented parameters (KL parameters), is expected to be more useful and convenient [5]. With the introduction of correct values of poles and zeroes, the quality of nasals and liquids can be further improved. Future efforts include formulation of rules for concatenation of synthetic speech stimuli and develop a text to speech conversion system for Hindi.

ACKNOWLEDGEMENTS

The first author gratefully acknowledges the support of UNDP and DoE for providing the fellowship to work at MIT, Cambridge which made this work possible. He is thankful to his colleagues of his lab for assistance in analyzing the data and useful discussions. He is thankful to the Director, CEERI for providing encouragement and permission to publish this work.

REFERENCES

[1] D. H. Klatt. "Review of Text to Speech conversion for English," JASA, vol. 82, pp. 737-743, 1987.

[2] N. B. Pinto, D. G. Childers, and A. L. Lalwani. "Formant speech synthesis: Improving production quality," IEEE Trans. ASSP, vol. 37, pp. 1870-1887, 1989.

[3] D. H. Klatt and L. C. Klatt. "Analysis, synthesis and perception of voice quality variations among female and male talkers," JASA, vol. 87, pp. 820-857, 1990.

[4] K. N. Stevens and C.A. Bickley. "Constraints among parameters to simplify control of Klatt formant synthesizer," J.Phonetics , vol. 19, pp. 161-174, 1991.

[5] K. N. Stevens. "The contribution of speech synthesis to phonetics: Dennis Klatt's Legacy," Proc. XII ICPhS, Aix-en-Provence, Aug. 19-24, vol. 5, pp.28-37, 1991.

[6] S. S. Agrawal and K.D.Pavate. "On the perception of (Hindi) speech sounds and distrbution features in different phonetic contexts," Proc. International Symposium, Speech Processing, TIFR, Bombay, July 23-26, pp. 8.1-8.34, 1980.

[7] R. Dixit. "Inadequacies in phonetic specification of some laryngeal features: Evidence from Hindi," in *Current Issues in the Phonetic Sciences*, Ed. Harry and Patricia Hollien. Amsterdam-John Benjamin BV, 1979.

[8] M. Ohala. *Aspects of Hindi Phonology*, Motilal Banarsidas, Delhi, 1983.

[9] S. S. Agrawal. "Acoustic phonetic and prosodic correlates of Hindi stop consonants," Proc. XII ICPhS, Aix-en-Provence, Aug 19-24, vol.5, 1991.