

Investigations of the Approximate Percentage of Noise Required to Perceive Hindi Phonemes using HNM

Padmini Rajput* and Parveen Lehana

Department of Electronics, University of Jammu, Jammu – 180006, Jammu and Kashmir, India;
gloryvirgo24.rajput@gmail.com, pklehana@gmail.com

Abstract

Objectives: Harmonic plus Noise Model (HNM) analysis model has been found to be one of the best methods of speech production in terms of important characteristics like naturalness, intelligibility, and pleasantness which are of pre-requisite in any speech synthesiser. Present study explores the approximate percentage of noise required to perceive some phonemes of Hindi language. **Method / Analysis:** HNM assumes speech as a combination of both periodic and aperiodic signals, so the effect of each part may be individually measured on the quality and intelligibility of different phonemes using HNM. HNM has been employed as the analysis-synthesis platform and the quality of the synthesized speech is tested with the ITU-T standard PESQ measure (perceptual evaluation of speech quality and MOS (mean opinion score)). **Findings:** Objective results suggest that the percentage of the noise serves as a significant constituent in the quality of synthesized speech. **Novelty:** Investigations suggest that each individual phoneme requires different noise and voice percentage for clear perception. Further, the optimum percentage of the noise part for good speech quality has been found speaker and phoneme dependent.

Keywords: Analysis-synthesis Models, Speech Processing, HNM, PESQ

1. Introduction

Verbal communication is the ability to convey one's thought by means of a set of signs, whether graphical, acoustic, gestural, or even musical. Among these, speech is an incomparable feature of human beings which states intent, ideas, and desires¹. Human speech production is a complicated sensory organization which requires the assimilation of diverse information sources and intricate patterns of muscle activations^{2,3}. Thought process in the brain initiates the excitation of the vocal tract required for the production of an utterance from the oral tract. The source of vibration called vocal tract, vibrates with the fre-

quency called the pitch frequency (for periodic sounds) normally in the range of 75 Hz to 300 Hz⁴. Vocal folds in case of males are usually longer (17.5-25 mm) than that in females (12.5-17.55 mm), causing a lower pitch and a deeper voice. This difference in the size of vocal chords causes a difference in vocal pitch⁵. In human beings, different sections of brain have different capabilities, for example in 97% of the right handed people, the left hemisphere is responsible for production, and about 19% of the left handed people right hemisphere is responsible for speech production however for about 68% of people both the left and right hemispheres play the part^{6,7}. Young children, constantly track the development course despite

*Author for correspondence

of culture and begin to learn quickly, transitioning from babbling at 6 months of age to full sentences by the age of three.

Knowledge of generation of various speech sounds helps in understanding its spectral and temporal attributes, which in turn helps in classifying it on a broader scale. Speech can be classified into voiced sounds, unvoiced sounds, vowels, consonants, nasal sounds, continuants, stops fricatives, syllables, diphthongs, and monothongs. Slight amendment of the shape of the vocal tract, different types of sounds ranging from vowels to consonants may be produced⁴. The unvoiced sounds, normally represented as a random white noise source don't show any periodicity as a result they do not have a straight relationship with the pitch⁸. The information that is communicated through the speech is intrinsically of a distinct nature as the constituents of the speech are discrete phonemes⁵. Phonemes: also referred to as speech units, show critical importance not because they are indestructible, but because they comprise the smallest distinction among minimal pairs such as let and lit, pat or bat. Wide range

of phonemes occupies a region in the articulator space⁹. To utter vowels, the tip of the tongue is positioned in the middle region of the oral cavity (vocal area). The lowering of the velum with the simultaneous constriction of the oral cavity so as to let the air flow through the nasal area generates nasal sounds which show quite broader spectral response similar to that of vowels^{8,10}. Stops define that category of sounds which require complete closure of the vocal tract⁸. Noise is produced as a result of partial restriction in the vocal tract, while fricatives are the result of this kind of constriction. Another class of speech called syllable is a vowel surrounded by consonants¹¹. Vowels are quasi-periodic in nature¹². All Indian languages use different scripts consisting of dissimilar graphemes, but there is a wide-ranging linguistic uniformity at the micro-level¹³. India has 22 official languages and about 1652 dialects/native tongues consisting of 10-12 major scripts. Indian languages have a refined notation of a character unit (or akshara).

An akshara is an essential linguistic entity which includes 0, 1, 2, or 3 consonants and a vowel. A word com-

Table 1. IPA symbols for phonemes of Hindi language

	Stops				Nasals				
	Unvoiced		Voiced						
	Unaspirated	Aspirated	Unasiprated	Aspirated					
Velar	[kə]	[kʰə]	[gə]	[gʰə]	[ŋə]				
Palatal	[[tʃə]	[tʃʰə]	[dʒə]	[dʒʰə]	-				
Retroflex	[ʈə]	[ʈʰə]	[ɖə]	[ɖʰə]	[ɳə]				
Dental	[t̪ə]	[t̪ʰə]	[d̪ə]	[d̪ʰə]	[nə]				
Labial	[pə]	[pʰə]	[bə]	[bʰə]	[mə]				
Semi Vowels			Fricatives						
[jə]	[rə]	[lə]	[ʋə]	[ʃə]	[ʒə]	[sə]	[hə]		
Short Vowels				Long Vowels					
[ə̃]	[ɦə]	[ʊ]	[o:]	[e:]	[a:]	[i:]	[u:]	[ɛ:]	[ɔ:]

prises of one or more aksharas and since the languages are exclusively phonetic, each akshara can be expressed independently¹⁴. Aksharas with more than one consonant are termed as samyuktaksharas or combo-characters. The consonant at the end is the foremost in a samyuktakshara. Unique characters in Indian languages scripts are near to syllable and can be, in general, structured as: C, V, CV, VCV, CVC, and CCV, where C signifies a consonant and V stands for a vowel¹⁵. Hindi is an official language of India spoken by 33% of the total inhabitants. I.P.A symbols for different classes of Hindi phonemes are shown in Table 1, which are further categorized into short vowels, long vowels, semivowels, nasals, fricatives, and stop sounds.

Restoration and aspiration are two phonemic features which acquire a vital place in Hindi. There are 8 aspirated plosives and 2 aspirated fricatives in addition to their unaspirated counterparts¹⁶. It has been stated¹⁶ that when the duration of the short vowel pair is essentially high (i.e., in case of vowels ओ and ऐ), native speakers appear to stress on the duration of vowels for preserving the phonemic distribution between members of a vowel pair more than non-native Hindi speakers. In case of stop sounds, release durations are essentially more significant than closure durations¹⁷. Stop sounds illustrate that release durations have statistical significance in case of unvoiced unaspirated, unvoiced aspirated and voiced aspirated phonemes. In Hindi, the durations of unvoiced aspirated are twice the durations of unvoiced unaspirated.

In Tamil language (spoken in Tamil Nadu, India), durations of unvoiced aspirated are equal to durations of unvoiced unaspirated, this may be because of having single alphabet for the stop sound in the place of articulation¹⁷. In Telugu (spoken mainly in Andhra Pradesh, India), durations of voiced unaspirated are equal to that of voiced aspiration and durations of unvoiced aspiration are twice the durations of unvoiced unaspirated¹⁷. The vowels show major dissimilarity between Telugu and Hindi in short vowels and between Tamil and Telugu in long vowels. Vowels help in distinguishing Telugu lan-

guage, since Hindi and Tamil speakers speak both short and long vowels approximately at equivalent period rates.

Hindi has more duration compared to Tamil, while Tamil shows a lesser duration than Telugu in short and long vowels correspondingly. For nasals, Hindi language shows important durational attributes¹⁷ and features craft practical significance in differentiating Hindi language, consequently, Telugu can be classified using vowels and Hindi can be classified using nasals. Also singleton stop phoneme durations are imperative feature for the classification of these three languages¹⁷. It has also been investigated that most of Dogri (spoken in Jammu, northern India) vowels have shorter duration, in comparison to Hindi phonemes. The period in milli seconds of same vowel when spoken in Hindi is found to be of longer duration compared to when it is spoken by a person with Dogri as the mother tongue¹⁸. Thus phoneme durations play a considerable role in distinguishing Hindi, Dogri, Telugu and Tamil languages pertaining to stop sounds, vowels and nasals¹⁶⁻¹⁸ and the awareness of the durational characteristics of a language plays an essential role for building highly intelligible Text to Speech Systems (TTS). Research has shown that the people are sensitive not only to the words they hear but also to the manner they are spoken.

Speech synthesis finds wide application in text-to-speech (TTS) systems¹⁹ in speaker text-independent^{20,21} and de-dependent identification²² multimedia entertainment, speech recognition²³ and speaker transformation²⁴. Two ways of speech generation are model based and waveform based. Generally, source filter model, direction into velocity of articulator's model-DIVA model, and Level's models are used in model based synthesis. Unit selection, HNM²⁵ concatenative, Linear Predictive Coding (LPC), and formant synthesis techniques are employed in waveform based techniques^{26,27}. Hidden Markov Model (HMM) based speech synthesis system (HTS) version-2 has also been employed for speech synthesis²⁸. Taking consideration of the attestations based on the last ten-year of research HNM model seems to be more promising and

robust speech synthesis technique^{27,29-31}. The objective of this paper is to determine the minimum percentage of noise and harmonic parts required to perceive the sound of the phonemes. Different models for speech analysis and synthesis are described in the following section. The experimentation and estimation methods employed for the evaluation of the quality of synthesized speech are discussed in Section 3. Section 4 presents the results and discussions part and the conclusion is summarized in Section 5.

2. Analysis-synthesis Models

One of the most attractive areas of speech communication is the area of man-machine communication. Researchers, including linguists, psychologists, and neurologists, have made an attempt to throw some light on the development of speech production and perception, initially focusing the monolingual speaker, and then moving ahead to more complex situations, which takes account of bilingual and multilingual speakers³². Although for many years man-made speech has been completely lucid from a segmental point of view, but there are certain areas which still look forward to acceptable realization. During recent years much effort has been directed on increasing the intelligibility as the speech that is synthesized from arbitrary text still sounded unnatural³³. The mechanical sounding voices may be satisfying but only to a limited extent³⁴. The need for natural sounding voices led to the requirement of more innate and intelligible speech synthesizers. In 1791 Von Kempelen projected that the human speech production system can be represented using mechanical models. He demonstrated his idea by building a machine that could produce human voice^{35,36}. Imitating Von Kempelen's work another well-known scientist Wheatstone made a speaking machine. Much later, Riesz, Homer Dudley, Haskins Play Back, and many other contributed to this field of speech³⁷.

Commercial formant synthesizer DECTalk was the first emotionally expressive speech synthesis system^{38,39}. The use of talking machines provides flexibility for comprehensive vocabularies, which is essential for appli-

cations such as unlimited translation from written text to speech⁴⁰. Only few Indian languages like Hindi, Tamil, Kannad, Marathi, and Bangla have been employed for developing TTS systems. Synthetic speech can be generated by two approaches: waveform and model based⁴¹. Model based techniques make use of the natural model of speech generation in human beings for developing artificial models⁴². Phonemes, morphemes, diphones, triphones, or syllables as basic acoustic units required to build a speech synthesizer. Combination of these units is also used in some synthesizers²⁰.

DIVA model of speech production provides computationally and neuroanatomically comprehensive account of the network of brain sections required in speech acquisition and production^{19,20}. Level developed a model of human communication in steps, visualizing speech production as a sequence of diverse phases, in three foremost mechanisms, specifically, the conceptualizer, the formulator, and the articulator³². Articulation synthesis employs the use of simulations and models that initiate from the articulatory mechanism of human speech production system for generating more natural sounding voices^{35,43}. The automatic systems built by Von Kempelen and Wheatstone fit in to this class^{40,44}. Source-Filter synthesis also called formant synthesis makes use of spectral shaping of driving excitation that employs formants to characterize the spectral shape³⁵. Formants have a direct acoustic phonetic interpretation, and are computationally simple when compared to full articulatory models. The formant synthesis method makes use of an acoustic model for speech generation instead of a real recorded human speech⁴⁵.

LPC is amongst the most powerful and latest synthesis techniques used in signal processing for the demonstration of the spectral envelope of speech in compact form taking into consideration the information exploited in linear productive model⁴⁶. It is an important technique for precise, inexpensive⁴⁷ measurements of speech parameters like pitch, formal spectra, vocal tract area functions, and for the representation of speech for low rate transmission and storage. Such approaches are extensively used by

many systems, in which speech unit waveforms are stored and then later concatenated during synthesis⁴⁶.

The acoustic composition of the speech can be described as a combination of units called phonemes, put another way, this theory brings into mind that in a similar way the distinct phonemes can be blended together to generate a speech waveform and this is the basic principle of concatenative speech synthesizers³⁰. Most of the cases involve a mechanism in which voice sounds are compressed by manipulating a pitch period waveform to reduce the number of signal samples that are required to have a power spectrum that should be sufficiently close to the original. Concatenative speech synthesis may be categorized into unit selection and diphone synthesis. Unit selection synthesis uses richer variety of speech and simply cuts out speech and rearranges it. The main aim is to find more naturalness in the generated speech³⁵. It uses a large database of pre-recorded speech. Research has shown that the sound obtained as result of unit selection is often difficult to distinguish from the real one, however for maximum naturalness we need large database, it may be a recorded speech of hours^{48,49}. Diphone synthesis makes use of the notion that a little portion of the acoustic signal varies to minor amount, and is also less subjective by the phonetic context than others. The quality of sound obtained using diphone synthesis lies somewhat in between concatenative and formant synthesis. However, the resulting speech suffers from glitches and is even little mechanical to hear, similar to the quality of sound as obtained from formant synthesis⁵⁰. Sine wave syntheses operate by replacing the formants with the pure form whistles. Domain specific synthesis is a relatively uncomplicated technique which uses the principle of concatenation of the pre-recorded speech to generate a complete statement. For the reason of its trouble-free implementation, this method is extensively²⁹ used for commercial purposes since long. This technique finds application in the areas where the output text is limited to a particular domain like weather reports, transit schedule announcements.

Speech analysis-synthesis techniques taking consideration of the characteristic and modification of diverse

models for the speech quality enrichment can provide more natural sounding and intelligible systems. It has been seen that the quality of the speech originated from the speech synthesizer depends upon the model being used by the synthesizer for this process. Harmonic models based concatenative techniques are widely used in TTS systems²². The fast generation of a harmonic signal is an important issue in reducing the complexity of TTS systems based on these models²². A good model generally requires virtues like intelligible synthesized speech, no difficulty of parameter extraction, ease of amendment of parameters, lesser number of parameters required, and lesser computation load.

A very versatile speech synthesizer called Festival is a concatenative TTS synthesis system developed by at the Centre for Speech Technology Research, University of Edinburgh with components supporting front-end processing of the input text^{47,51}. HNM model performs analysis and synthesis of speech signal and is basically a modification of sinusoidal models²⁶. HNM decomposes speech signal into quasi periodic: a lower harmonic” part, and non-periodic part represented by an upper “noise” part^{20,21}. This breakdown method⁵² which represents the upper harmonic part as the voice part and lower stochastic as the noise part of speech signal, employed in HNM permits more naturalness in synthesized speech.

HNM has reduced database, and provides a direct technique for smoothing discontinuities of acoustic units around concatenation points^{53,54} produced as a result of different distributions of the system phase around the points of concatenation. This incoherence generates noise in between the harmonic peaks thus destroying the harmonic structure of a periodic sound henceforth degrading the voice quality. Analysis shows that all vowels and syllables can be produced with a better quality syllables by the implementation of HNM²⁰ also HNM is a pitch-synchronous system⁵⁵ and unlike TD-PSOLA and other concatenative approaches hence it eliminates the problem of synchronization of speech frames and shows the capabilities of providing high-quality prosodic

modifications without business when compared to other methods⁵⁶.

HNM framework is also used in a low bit rate speech coder to increase naturalness as well as intelligibility⁵⁷. Harmonics plus noise model has also been used for the development of a high-quality vocoder applicable in statistical frameworks, particularly in modern speech synthesizers^{58,59}. Speaker transformation and voice conversion method techniques can also be implemented using HNM system^{20,46,47} and since HNM appears to be more promising than all existing models, thus in the present research this technique has been employed to determine the effect of noise and voice part separately on the speech signal.

3. Methodology

3.1 Recording and Segmentation

For data collection, six speakers (3 males and 3 females) in the age group of 18-25 years were selected for recording in Hindi language. It is desirable that the speakers belong to same group in terms of language and education. The speakers participating in our experiments were university students and they had Hindi as their first language.

The script for recording as shown in Table 2 consists 35 phonemes (consonants) with first and the last phoneme (v) being a: for all the vcvs combinations. Speech was recorded in an acoustically treated room using Sony ICD-PX820 audio recorder with 16 kHz sampling and 16-bit quantization. The recorded speech was manually segmented and labeled into vcvs which were considered to be correctly articulated by all the speakers were selected for use as the speech utterances for the experiment. The duration of each utterance was 5–9 s.

3.2 Speech modification using HNM

Figure 1 shows the basic scheme used for HNM-based modification of speech. The parameters of the speech signal are obtained from HNM analysis and modified for the spectral modification. The synthesis axis is prepared according to the given pitch and time scaling factors. HNM parameters are estimated at the instants on the synthesis axis using interpolation. The modified parameters on the synthesis axis are used for synthesizing the speech output. In the present experiment, in order to inspect the effect of noise percentage on speech quality, the analysis-synthesis of the speech using HNM has been performed. Keeping voice part fixed, the speech was synthesized by

Table 2. VCV (vowel consonant vowel) combinations for Hindi language

[a:kəa:]	[a:k ^h əa:]	[a:gəa:]	[a:g ^h əa:]	[a:tʃəa:]
[a:tʃ ⁿ ə a:]	[a:dʒəa:]	[a:dʒ ^h əa:]	[a:ʈəa:]	[a:ʈ ^h əa:]
[a:dəa:]	[a:d ^h əa:]	[a:ŋəa:]	[a:ʈəa:]	[a:ʈ ^h əa:]
[a:ɖəa:]	[a:ɖ ^h əa:]	[a:nəa:]	[a:pəa:]	[a:p ^h əa:]
[a:bəa:]	[a:b ^h əa:]	[a:məa:]	[a:jəa:]	[a:rəa:]
[a:ləa:]	[a:ʊəa:]	[a:ʃəa:]	[a:ʃəa:]	[a:səa:]
[a:həa:]	[a:ks ^h əa:]	[a:traa:]	[a:jnaa:]	[a:s ^h rəa:]

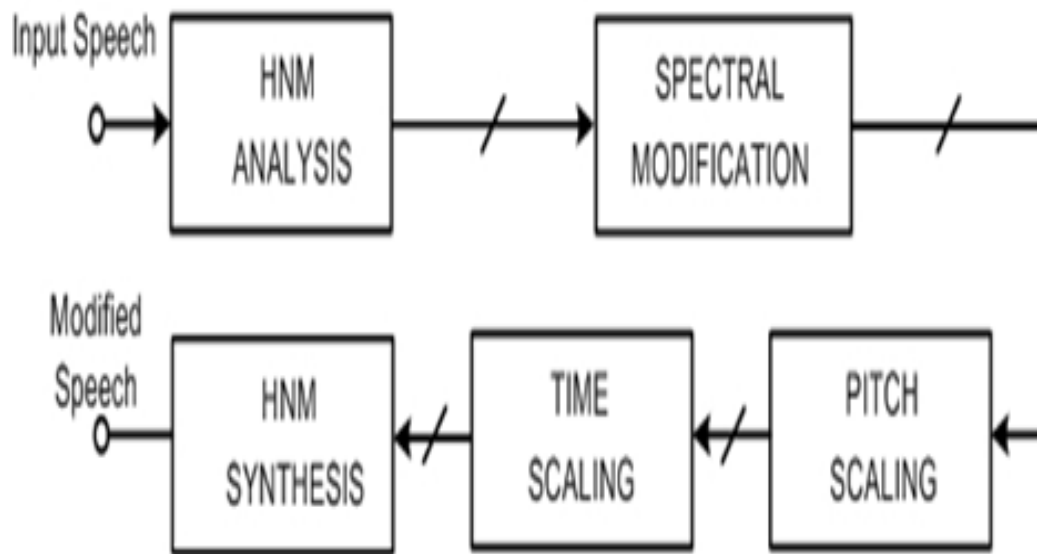


Figure 1. HNM based modification.

varying noise part (and vice-versa) at different percentage levels (10-100%).

3.3 Evaluation techniques

The methods used for the evaluation of can be broadly classified as subjective and objective. Subjective methods demand listening tests by means of human subjects. The results of the subjective evaluation may get affected by the test conditions, and hence these have to be standardized and consistently followed. The subjects should be adequately familiarized with the reference quality before the test. The subjective tests may be grouped in three categories: intelligibility, quality, and identity. Quality of the phrases is generally evaluated by-mean opinion score (MOS), degradation category rating (DCR), and preference tests.

In MOS test or absolute category test, the subject rates the quality of the speech stimuli on 1-5 scale (1: bad, 2: poor, 3: fair, 4: good, 5: excellent). The stimuli are presented in a randomized order, with three to five presentations of each stimulus. The average score calculated across stimuli and subjects is known as the mean opinion

score (MOS)⁶³. The test gives an assessment based on all the parameters affecting the quality. It is easy to conduct and does not need trained listeners, but its sensitivity for high quality speech is low. The objective methods ensure consistency in the evaluation and can be performed by means of computations⁶⁰. PESQ, one of the methods for objective evaluation evaluates one-way speech quality. The signal to be evaluated is introduced into the system under test and the synthesized output signal is matched with the input (reference) signal⁶¹. PESQ has been incorporated as the ITU-T P.862 recommendation⁶². PESQ is a narrow-band (3.2 kHz) speech quality assessment and has the capability of providing an admirable quality in a range of conditions together with background noise, analogue filtering, and variable delay.

Level alignment is required for signals like the reference speech signal, however the degraded signal is made parallel to the continuous power intensity. The signals are time aligned assuming the impediment initiated by the transmission system as piecewise constant. This transform eliminates those parts of the signal that are impossible to hear to the listener. Frame-by-frame interruptions are anticipated by means of envelope and fine correla-

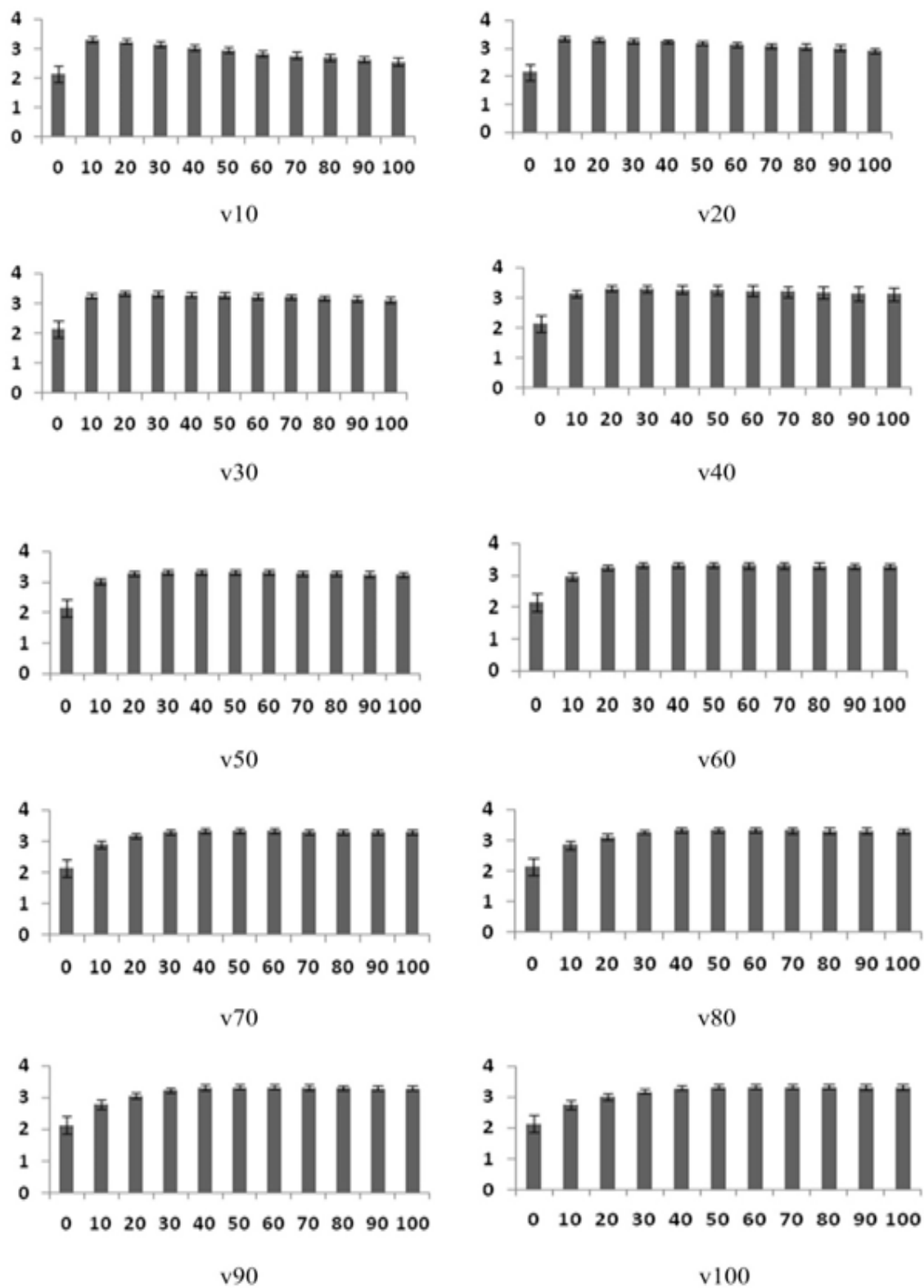


Figure 2. PESQ score obtained at different percentage of voice and noise part for six speakers, (3 males and 3 females).

tion histogram-based delay detection. PESQ assessment finds applications in wide range of applications. Because of its high-speed and repeatability PESQ make it feasible to execute extensive testing over undersized period and also facilitates the quality of time-varying conditions to be observed⁶¹. PESQ offers specific and repeatable estimation of speech quality.

4. Results and Discussion

Figure 2 shows different plots for different proportions of voice parts (v10-v100, ex. v10 indicates 10% voice part), of all the six speakers. In each histogram the horizontal axis shows the percentage of noise, and the vertical axis shows its corresponding PESQ score. With voice part of only 10%, in case of single female speaker it has been analyzed that, minimum quality (PESQ score of 2 approximately) is obtained with zero noise part, whereas adequate PESQ score (3 approximately) at which the speech quality is quite acceptable is obtained when only 10% of noise is added. As the proportion of voice part is gradually increased, the point at which the maxima (highest quality) occurs requires a greater percentage (40%) of noise for optimum speech quality, also the extent to which the

speech quality showed steep degradation (in case of v10 and v20) after the maxima was obtained is considerably reduced.

Similar are the results obtained in case of males, with only one significant change. The value of minima (minimum quality) is obtained at a PESQ score of 2.3 approximately, little higher than that found in case of females. Pictorial representation of the experimental results is represented in histograms shown in Figure 2, which depicts the consequence of the change in voice and noise percentage on the quality of the synthesized speech of all the six speakers (3 males and 3 females) taken all together. It may be analyzed that for a constant voice part, as the percentage of noise is increased, PESQ score shows a substantial increase till it attains a peak value, after which, there occurs a gradual decrease in the quality of the synthesized speech. However, as percentage of voice part is increased, although the value of PESQ score for which minima is obtained remains approximately the same, but the PESQ score at which the quality of speech was found to be maximum, requires a greater percentage of noise part. However as soon as this maxima is attained any further increase in noise proportion doesn't appear to affect the speech quality i.e. the quality becomes quite sta-

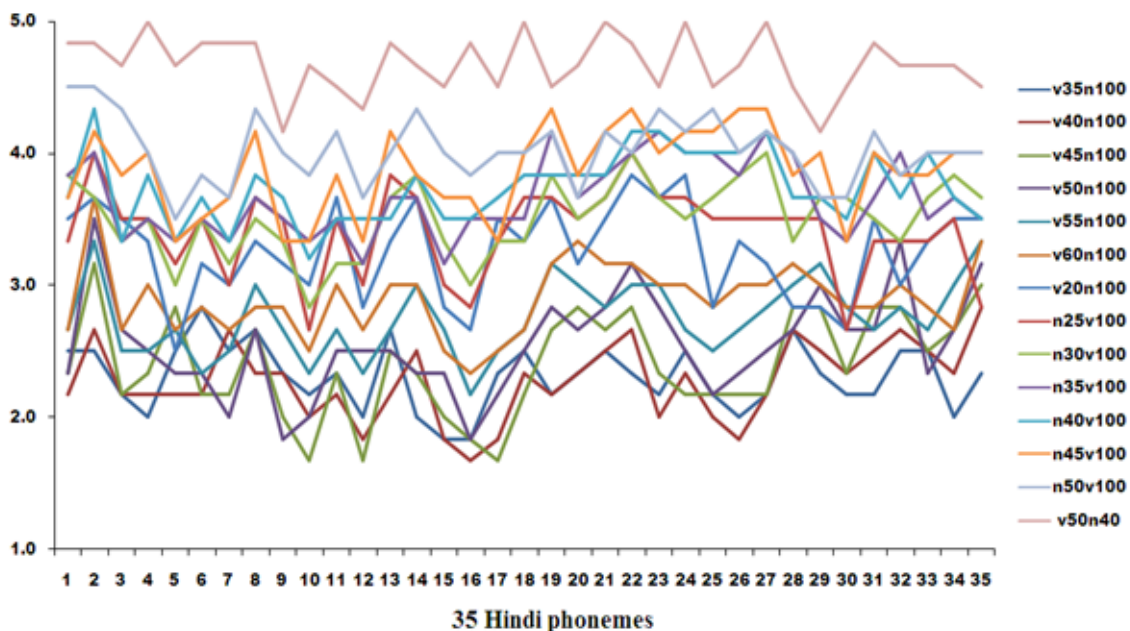


Figure 3. PESQ score for 35 VCV at different voice level and noise level in case of male speaker obtained from listening test.

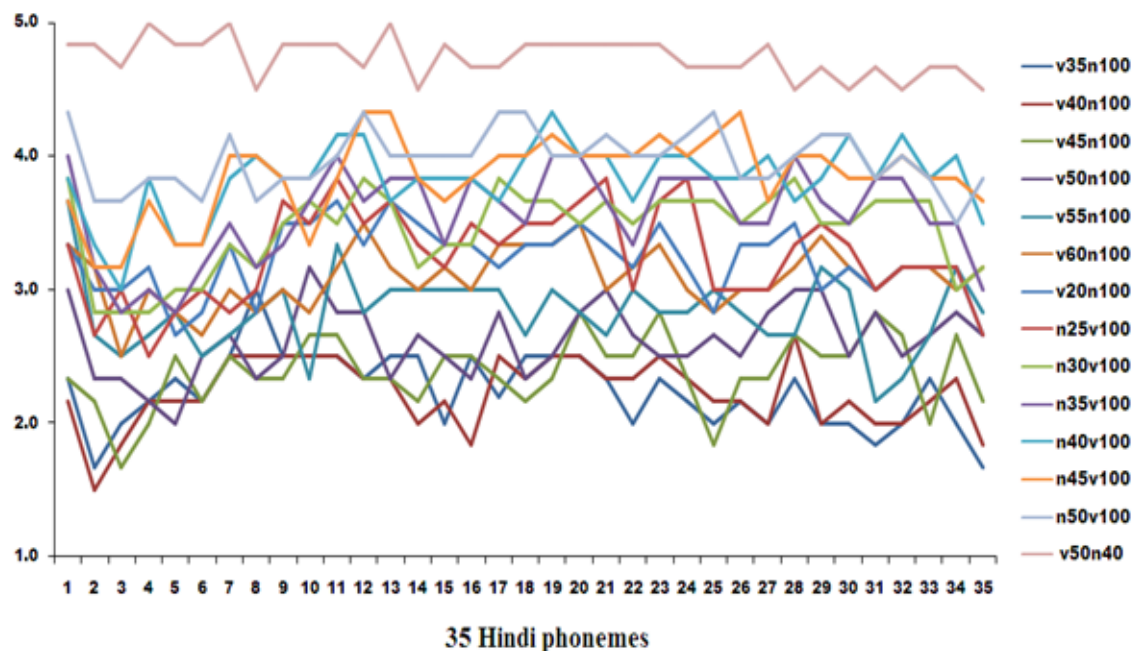


Figure 4. PESQ score for 35 VCV at different voice level and noise level in case of female speaker obtained from listening test.

Table 3. MOS scores for the sets of 35 VCV utterances for male speaker in Hindi language

Voice-Noise level	MOS Score							
	Sb1	Sb2	Sb3	Sb4	Sb5	Sb6	Mean	Std
v35n100	1.91	2.80	2.26	2.20	2.09	2.17	2.2	0.3
v40n100	1.94	2.74	2.17	2.29	2.14	2.09	2.2	0.3
v45n100	2.31	2.43	2.51	2.51	2.26	2.29	2.4	0.1
v50n100	2.83	2.94	2.43	2.97	2.26	2.29	2.6	0.3
v55n100	3.11	3.06	2.66	2.97	2.63	2.57	2.8	0.2
v60n100	3.14	3.23	2.80	3.09	2.86	3.43	3.1	0.2
v50n40	4.89	4.51	4.60	4.77	4.77	4.89	4.7	0.2
n20v100	3.23	3.23	3.34	2.69	3.40	3.46	3.2	0.3
n25v100	3.31	3.00	3.46	3.20	3.46	3.51	3.3	0.2
n30v100	3.63	3.20	3.69	3.49	3.49	3.51	3.5	0.2
n35v100	3.60	3.26	3.89	3.71	3.63	3.51	3.6	0.2
n40v100	3.69	3.77	4.14	3.71	3.91	3.74	3.8	0.2
n45v100	3.60	3.57	4.49	4.00	4.06	3.69	3.9	0.4
n50v100	3.86	3.63	4.51	4.77	4.06	3.83	4.1	0.4

Table 4. MOS scores for the sets of 35 VCV utterances for female speaker in Hindi language

Voice-Noise level	MOS Score							
	Sb1	Sb2	Sb3	Sb4	Sb5	Sb6	Mean	Std
v35n100	2.14	2.97	2.23	2.31	2.31	2.11	2.3	0.3
v40n100	2.23	2.86	2.09	1.94	2.40	2.11	2.3	0.3
v45n100	2.46	2.97	2.29	2.11	2.43	2.11	2.4	0.3
v50n100	2.89	2.97	2.43	2.31	2.51	2.17	2.5	0.3
v55n100	3.11	3.29	2.69	2.49	2.77	2.17	2.8	0.4
v60n100	3.09	3.49	2.77	2.74	3.00	2.29	2.9	0.4
v50n40	4.80	4.77	4.20	4.89	4.54	4.51	4.6	0.3
n20v100	3.46	3.57	3.34	3.11	3.31	2.77	3.3	0.3
n25v100	3.63	3.60	3.46	3.29	3.51	2.97	3.4	0.2
n30v100	3.71	3.71	3.69	3.26	3.57	3.11	3.5	0.3
n35v100	3.80	3.77	3.91	3.51	3.71	3.23	3.7	0.2
n40v100	3.94	3.77	4.03	3.69	3.74	3.26	3.7	0.3
n45v100	3.80	3.83	4.51	3.77	3.83	3.46	3.9	0.3
n50v100	4.03	3.89	4.54	4.00	4.11	3.57	4.0	0.3

ble even if the noise percentage is increased. Thus it may be concluded that noise part serves as an important part in the quality of synthesized speech. With no noise part added the speech quality is quite poor. Also, the percentage of noise part to be added for optimum voice quality depends strictly on the voice part.

Figure 3 and Figure 4 show the results obtained after quality assessment of HNM synthesized speech using MOS. The purpose of this experiment is to evaluate the quality and identity of the HNM synthesized speech using subjective listening tests and to verify the results of the objective tests (as obtained by PESQ evaluation) for male and female speakers. The set of 35 VCV utterances as listed in Table 2 were used as test material for one male and one female speaker for this experiment.

Subjective evaluation of the quality and intelligibility of the synthesized speech was carried out using MOS test. For each presentation subject could listen to the sounds one after the other in a sequence more than once before finalizing the response and proceeding to the next presentation. The average MOS scores for 35 VCV utterances at different voice and noise levels in case of male and female speaker are listed in Table 3 and Table 4 respectively. From objective evaluation it was found that a voice level of 50% and a noise level of 40% assure good quality for HNM synthesized speech. Above a voice level of 50% the quality doesn't degrade even if the noise percentage is further added. Mean score for male speaker at voice level 50% and noise level of 40% is 4.7. Mean scores for male speaker utterances at increasing voice level clearly

Table 5. Perception of Hindi phonemes at different noise percentage

Range Noise (%)	Speakers					
	sp1	sp2	sp3	sp4	sp5	sp6
10	दा,पा,प ^h , ब,ब ^h ,ज, र,ल	द, द ^h , त, प,ब व,ह	ग,त ^h ,त,प, म,ज	ड, ड ^h , द,त,द प, प ^h , म,ज,र ज,स, ज, त्रा	त ^h ,ड ^h ,त,न, द ^h ,न,प,र,ल, ज,व	ग ^h ,त,न,त ^h ,प ^h , ब,म,र,स, न,व,ज
20	न,ड,त,म, ज,क ^h , ज, त्रा	ग ^h , न,ड, ड ^h न, त ^h ,द ^h , प ^h ज, ल, क ^h , ज, त्रा	क ^h ,ग ^h ,ड, ड ^h , त ^h ,द ^h ,त ^h ,द ^h ,ब अ,र,व,ज,ज,ह	न,त ^h ,त,द, न, न, द ^h ,द ^h ,ह	क ^h ,त ^h ,त ^h ,द, म,ज,स,क ^h , द,ज,ज, त्रा	क ^h , ग, न, त ^h ,त, द,ज, क ^h , स ^h , ज,ड, द ^h
30	क,क ^h ,ग ^h ,ड, ह	ग,स	त,ब ^h ,स,क ^h , ज,ल	ग ^h ,स ^h	ग ^h ,न,त	
40	त ^h ,द ^h ,न,त ^h ,द ^h , ,त्रा,स ^h ,व	त ^h ,त,न,म ज,ज,त्रा,स ^h	क,द,न, स ^h	द ^h ,ब,ल, क,ज	क,स ^h	द,द ^h ,ह
50	त ^h ,द,न, त	त ^h ,द,ब ^h	न,द,न, त्रा	व		ज
60	स	क ^h		क ^h ,ज	द ^h ,ब ^h	क,त ^h ,प,ब ^h
70	ग	क		त ^h ,त ^h ,ब ^h	त्रा, ब	
80			प ^h	क	ग	ल,त्रा
90					प ^h	ड ^h
100				ग	ड,ह	द

show that the quality shows a gradual increase with the percentage of voice part and noise part and at a particular level of voice (50%) and noise (40%) the quality is maximum. The plot for quality as obtained by MOS score of six listeners has been presented in Figure 3 and Figure 4 for male and female speaker respectively at different voice and noise percentages (e.g. v35n100 denotes voice part 35% and noise part 100%). In case of males it can be analyzed that at v50n40 the quality is maximum. Similarly in case of females the MOS score at v50n40 is 4.6 and hence again appreciable quality has been obtained at this percentage of voice and noise part.

Table 5 shows the approximate percentage of noise required to perceive different Hindi phonemes for six speakers (sp1-sp6). It has also been analyzed that at 10%, quality dependence of phonemes pə, rə, p^hə, bə, jə, lə, t̪ə, və, mə is found to show least speaker dependency, while the phonemes d̪ə, t̪^hə, d̪^hə, fə, sə, d̪ə, tra, d̪ə, nə, ɳə, d̪^hə, t̪ə, sə, ɳə, fə, g^hə, b^hə, d̪^hə, d̪^hə, t̪ə, show speaker dependency. At 20% the phonemes ɳə, d̪ə, fə, fə, ks^h, k^hə, t̪^hə, d̪ə are speaker independent, but the phonemes g^hə, d̪^hə, d̪^hə, jna, ɳə, rə, jə, t̪^hə, t̪^hə, p^hə, lə, bə, və, fə, d̪^hə, nə, fə, t̪ə, sə, d̪ə, jna, s^hrə, gə, ɳə, t̪ə show speaker dependency.

At 30% only phoneme g^hə is found to be speaker independent while kə, k^hə, gə, d̪^hə, fə, sə, ks^hə, jna, lə, t̪ə, b^hə, s^hrə, ɳə, t̪ə are speaker dependent. At 40% only d̪^hə, s^hrə have been found to be speaker independent and the phonemes tra, ɳə, t̪^hə, və, t̪^hə, d̪^hə, tra, t̪^hə, t̪ə, nə, mə, fə, fə, d̪ə, bə, lə, ks^hə, kə, d̪^hə, d̪ə, fə are seen to be speaker dependent. At noise level of 50% the phonemes nə, t̪ə, d̪ə, t̪^hə, d̪ə, b^hə, t̪^hə, tra, ɳə, və, jə are least speaker independent. At 60% sə, k^hə, fə, d̪^hə, b^hə, kə, t̪^hə, pə, b^hə are speaker independent. At 70% gə, kə, t̪^hə, t̪^hə, b^hə, tra, bə are speaker independent. At 80% phonemes p^hə, kə, gə, lə, tra show speaker dependency. Also it may be analyzed that the required noise percentage for phoneme kə range from 30-80, k^hə 20-60, gə 10-100, g^hə 10-30, ɳə 20-50, t̪^hə 10-50, d̪ə 10-100, d̪^hə 10-90, t̪ə 10-50, t̪^hə 20-70, d̪ə 10-100, d̪^hə 10-60, ɳə 10-40, t̪ə 10-30, t̪^hə 10-70, d̪ə 10-40, d̪^hə 10-40, nə 10-50, pə 10-60, p^hə 10-90, bə 10-70, b^hə 10-70, mə 10-40, jə

10-50, rə 10-20, lə 10-80, və 10-50, fə 10-40, fə 10-60, sə 10-60, fə 10-100, ks^hə 20-40, tra 10-80, jna 10-30, s^hrə 20-40.

5. Conclusion

Investigations were carried out to find the minimum percentage of periodic and aperiodic portion of speech for clear perception of different Hindi phonemes. HNM has been used for analysis and synthesis of speech HNM has been employed while PESQ method is used for objective evaluation and MOS for subjective evaluation of the speech quality. Investigations carried out by varying voice and noise parts of speech signal show that the quality and intelligibility is related to the relative percentage of noise and voice parts. The quality of the synthesized speech without adding any noise part gives almost same PESQ score independent of the percentage of the voice part with respect to the original amount of the voice part. With 50% voice part, the required noise percentage for acceptable speech quality is found to be around 40%. As the voice percentage is increased beyond 50%, the speech quality shows no degradation even if the noise percentage is further increased. Results obtained from MOS also verify the same. These values of noise proportions are approximately similar for male and female speakers.

6. References

1. Shetake PS, Patil SA, Jadhav PM. Review of text to speech conversion methods. Proceedings 10th IRF International Conference; 2014.
2. Stylianou Y, Cappe O, Moulines E. Continuous probabilistic transform for voice conversion. IEEE Transactions Speech and Audio Processing. 1998; 6(2):131–42. <https://doi.org/10.1109/89.661472>.
3. Yin H, Hohmann V, Nadeu C. Acoustic feature for speech recognition based on gammatonefilterbank and instantaneous frequency. ELSEVIER Speech Communication. 2011; 53(5):707–15. <https://doi.org/10.1016/j.specom.2010.04.008>.
4. Douglas OS. Speech communication: human and machine. Reading, Massachusetts: Addison-Wesley publishing company; 1987. p. 1–568.

5. Rabiner LR, Schafer RW. Digital processing of speech signals. Prentice-Hall Inc. Englewood Cliffs, New Jersey; 1978. p. 1–512.
6. Epstein CM, Lah JJ, Meador K, Weissman JD, Gaitan LE, Dihenia B. Optimum stimulus parameters for lateralized suppression of speech with magnetic brain stimulation. *The Official Journal of Neurology*. 1996; 47(6):1590–3. <https://doi.org/10.1212/WNL.47.6.1590>.
7. Guenther FH, Ghosh SS, Tourville JA. Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language*. 2006; 96(3):280–301. <https://doi.org/10.1016/j.bandl.2005.06.001>.
8. Furui S, Sondhi MM. Advance in speech signal processing. Marcel Dekker, New York; 1992.
9. Warner N, Smits R, McQueen JM, Culter A. Phonological and statistical effects on timing of speech perception: insights from a database of dutchdiphone perception. *Elsevier Speech Communication*. 2005; 46(1):53–72. <https://doi.org/10.1016/j.specom.2005.01.003>.
10. Gold B, Morgan N, and Ellis D. Speech and audio signal processing: Processing and perception of speech and music. 2nd Edition. Wiley-Blackwell; 2011. p. 1–688. <https://doi.org/10.1002/9781118142882>.
11. Dudley H. The carrier nature of speech. *Bell System Technical Journal*. 1940; 19(4):495–515. <https://doi.org/10.1002/j.1538-7305.1940.tb00843.x>.
12. Marwan AA. Fractal speech processing. Cambridge University Press; 2004.
13. Varalwar M, Nixon P. Characteristics of Indian languages. Bhrigus Inc. Hyderabad, India; 2013. p. 1–6.
14. Kumar PK. Speech synthesis based on sinusoidal modeling. Credit Seminar Report, Electronic Systems Group; 2004. p. 1–20.
15. Kishore SP, Kumar R, Sangal R. A data-driven synthesis approach for Indian languages using syllable as basic unit. *Proceedings International Conference Natural Language Processing (ICON)*; 2002. p. 1–7.
16. Rao PVS, Agarwal SS. Hindi speech database. Sixth International Conference on Spoken Language Processing; 2000.
17. Kanth LB, Keri V, Prahallad KS. Durational characteristics of Indian phonemes for language discrimination. *International Conference on Information Systems for Indian Languages*; 2011. p. 130–5. https://doi.org/10.1007/978-3-642-19403-0_21.
18. Rajput P, Lehana PK. Investigations of the distributions of phonemic durations in Hindi and Dogri. *International Journal on Natural Language Computing*. 2013; 2(1):1–14. <https://doi.org/10.5121/ijnlc.2013.2103>.
19. Wu CH, Hsia CC, Liu TH, Wang JF. Voice conversion using duration-embedded Bi-Hmms for expressive speech synthesis. *IEEE Transactions Audio, Speech, Language Processing*. 2006; 14(4):1109–16. <https://doi.org/10.1109/TASL.2006.876112>.
20. Lu H, Okamoto H, Nishida M, Horiuchi Y, Kuroiwa S. Text-independent speaker identification based on feature transformation to phoneme-independent subspace. *11th IEEE International Conference on Communication Technology*; 2008. p. 692–5.
21. Jin Q, Toth AR, Schultz T, Black AW. Voice converging: speaker de-identification by voice transformation. *International Conference on Acoustics Speech Signal Processing. (ICASSP)*. Carnegie Mellon University, Pittsburgh, USA; 2009. p. 1–4. <https://doi.org/10.1109/ICASSP.2009.4960482>.
22. Zhang M, Tao J, Nurminen J, Tian J, Wang X. Phoneme cluster based state mapping for text-dependent voice conversion. *9th International Conference on Signal Processing*; 2009. p. 723–6. <https://doi.org/10.1109/ICOSP.2008.4697232>.
23. Kain A, Macon MW. Spectral voice conversion for text-to-speech synthesis. *IEEE International Conference Acoustic, Speech and Signal Processing*. 1998; 1:285–8.
24. Xu N, Yang Z, Zhu WP. Modeling articulatory movements for voice conversion using statespace model. *Fifth International Conference Natural Computation Signal Processing*; 2009. p. 236–40. <https://doi.org/10.1109/ICNC.2009.491>.
25. Taylor P. Text to speech synthesis. Cambridge University Press; 2009. p. 1–149. <https://doi.org/10.1017/CBO9780511816338>.
26. Desai S, Raghavendra EV, Yegnanarayana B, Black AW, Prahallad K. Voice conversion using artificial neural networks. *International Conference on Acoustic Speech Signal Processing. (ICASSP)*; 2008. p. 1–4. <https://doi.org/10.1109/ICASSP.2009.4960478>.
27. Yoshimura T, Tokuda K, Masuko T, Kobayashi T, Kitamura T. Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis. *Proceedings Eurospeech*; 1999. p. 1–4.
28. Yamagishi J, Zen H, Toda T, Tokuda K. Speaker-independent hmm-based speech synthesis system-HTS, system for the blizzard challenge. *Proceedings of Blizzard Challenge*; 2007.
29. Stylianou Y. On the implementation of the harmonic plus noise model for concatenative speech synthesis. *IEEE International Conference on Acoustics, Speech, and Signal Processing*. 2000; 2:957–60.
30. Stylianou Y. Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Transactions on*

- Speech and Audio Processing. 2001; 9(1):21–9. <https://doi.org/10.1109/89.890068>.
31. Lee KS, Cox RV. A very low bit rate speech coder based on a recognition/synthesis paradigm. *IEEE Transactions on Speech and Audio Processing*. 2001; 9(5):482–91. <https://doi.org/10.1109/89.928913>.
 32. Fernandes MHB, Brito KS. Speaking models: from Lavelet's monolingual to Williams and Hammarberg's polyglot. *Proceedings of the Fifth International Symposium on the Acquisition of Second Language Speech*; 2007. p. 1–8.
 33. VanSanten JPH. Combinatorial issues in text-to-speech synthesis. *Proceedings of the Fifth European Conference on Speech Communication and Technology, EUROSPEECH*; 1997. p. 1–4.
 34. Lehana PK, Pandey PC. Harmonic plus noise model based speech synthesis in Hindi and pitch modification. *Proceedings of the 18th International Congresses on Acoustics*; 2004. p. 1–4.
 35. Wouter J, Macon MW. Unit fusion for concatenative speech synthesis. *Proceedings International Conference on Spoken Language Processing, (ICSLP)*. 2000; 3:302–5.
 36. Eide E, Aaron A, Bakis R, Hamza W, Picheny M, Pitrelli J. A corpus-based approach to <AHem> expressive speech synthesis. *Proceedings ISCA SSW5*; 2004. p. 1–4.
 37. Black AW. Unit selection and emotional speech. *Proceedings Eurospeech*; 2003. p. 1649–52.
 38. Grimme B, Fuchs S, Perrier P, Schöner G. Limb versus speech motor control: A conceptual review. *Motor Control*. 2011; 15(1):5–33. <https://doi.org/10.1123/mcj.15.1.5>.
 39. Iwanska L, Shapiro SC. *Natural language processing and knowledge representation*. MIT Press; 2000. p. 1–480.
 40. Moulines E, Charpentier F. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*. 1990; 9(5–6):435–67. [https://doi.org/10.1016/0167-6393\(90\)90021-Z](https://doi.org/10.1016/0167-6393(90)90021-Z).
 41. Denes P, Pinson E. *The speech chain*. Bell Telephone labs, Murray Hill, NJ; 1963. p. 1–158.
 42. Manning CD, Schutze H. *Foundations of statistical natural language processing*. Massachusetts Institute of Technology. Cambridge, MA: The MIT Press; 1999. p. 1–680.
 43. Stylianou Y. Concatenative speech synthesis using a harmonic plus noise model. *Third ESCA/ COCOSDA, Workshop on Speech Synthesis*; 1998. p. 261–6.
 44. Honda M. *Human speech production Mechanisms*. NTT Technical Review. 2003; 1(2):1–6.
 45. Stylianou Y. Removing linear phase mismatches in concatenative speech synthesis. *IEEE Transactions Speech, Audio Processing*. 2001; 9(3):232–9. <https://doi.org/10.1109/89.905997>.
 46. Laroche J, Stylianou Y, Moulines E. HNM: A simple, efficient harmonic noise model for speech. *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*; 1993. p. 169–72.
 47. Black AW, Taylor P. *The festival speech synthesis system*. University of Edinburgh; 1997.
 48. Levelt WJM. *Producing spoken language: A blueprint of the speaker*. Brown C, Hagoort P, editors, *The Neurocognition of Language*. Oxford, University Press; 1999. p. 83–122. <https://doi.org/10.1093/acprof:oso/9780198507932.003.0004>.
 49. Levelt WJM. Monitoring and self-repair in speech. *Cognition*. 1983; 14(1):41–104. [https://doi.org/10.1016/0010-0277\(83\)90026-4](https://doi.org/10.1016/0010-0277(83)90026-4).
 50. Levelt WJM. *Speaking: From intention to articulation*. Cambridge, MA: MIT Press; 1989. p. 1–566.
 51. Black AW, Taylor P. Automatically clustering similar units for unit selection in speech synthesis. *Proceedings of the Eurospeech*. 1997; 2:601–4.
 52. Levelt WJM, Roelofs A, Meyer AS. A theory of lexical access in speech production. *Behavioral and brain science*. 1999; 22(1):1–38. <https://doi.org/10.1017/S0140525X99001776>.
 53. Levelt WJM. Models of word production. *Trends in Cognitive Sciences*. 1999; 3(6):223–32. [https://doi.org/10.1016/S1364-6613\(99\)01319-4](https://doi.org/10.1016/S1364-6613(99)01319-4).
 54. Levelt WJM. *Spoken words production: A theory of lexical access*. National Academy of Science. 2001; 98(23):13464–71. <https://doi.org/10.1073/pnas.231459498>.
 55. Ganchev T, Lazaridis A, Mporas I, Fakotakis N. Performance evaluation for voice conversion systems. *International Conference on Text, Speech and Dialogue*; 2008. p. 317–24. https://doi.org/10.1007/978-3-540-87391-4_41.
 56. Stylianou Y. A simple and fast way of generating a harmonic signal. *IEEE Signal Processing Letters*. 2000; 7(5):1–3. <https://doi.org/10.1109/97.841155>.
 57. Meister S, Wiggins RH. Quality comparison measure for linear predictive systems. *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*. 1976; 1:107–9.
 58. Sainz I, Navas E, Hernaez I. Harmonics plus noise model based vocoder for statistical parametric speech synthesis. *IEEE Journal of Selected Topics in Signal Processing*. 2014; 8(2):184–94. <https://doi.org/10.1109/JSTSP.2013.2283471>.
 59. Lehana PK, Chand PP. Effect of GCI perturbation on speech quality in Indian languages. *Proceedings TENCON, Conference on Convergent Technologies for Asia-Pacific Region*. 2003; 3:959–63.
 60. Falk TH, Chan WY. Objective speech quality assessment using Gaussian mixture models. *Proceedings 32nd Biennial Symposium on Communication*; 2004. p. 169–71.

61. Rix AW, Beerends JG, Hollier MP, Hekstra AP. Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs. IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221); 2001. p. 1–862.
62. Beerends JG, Hekstra AP, Rix AW, Hollier MP. Perceptual evaluation of speech quality, the new itu standard for end-to-end speech quality assessment part ii psychoacoustic model. Journal of the Audio Engineering Society. 2002; 50(10):765–78.
63. ITU-T. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. International Telecommunication Union, Geneva; 2001. p. 1–862.