Noise as a Tool for Spoken Language Identification

.

Sunita Maithani
Scientific Analysis Group,
Defense Research & Development Organization,
Metcalfe House, Delhi-110054, India.
E-mail: ysmaithani58@yahoo.com

J. S. Rawat
.Scientific Analysis Group,
Metcalfe House, Delhi-110054, India.
Defense Research & Development Organization,

Abstract

Segmental SNR (Signal to Noise Ratio) is considered to be a reasonable measure of perceptual quality of speech. However it only reflects the distortion in time dependent contour of the signal due to noise. Objective Measures such as Log Area Ratio (LAR), Itakura-Saitio Distortion (IS), Log-Likelihood Ratio (LLR) and Weighted Spectral Slope (WSS) are better measures of perceptual speech quality as they represent deviation in the spectrum. Noise affects the speech time contour and the corresponding frequency content. Different languages have some peculiar characteristics due to variation in the phonetic content and their distribution. Distortion introduced by noise and application of enhancement algorithm varies for different phonemes. In this paper a novel idea of using noise and speech enhancement as means of identifying a language is presented, using objective measures of speech quality. Study is done on three spoken Indian regional languages namely Kashmiri, Bangla and Manipuri, when corrupted by white noise. It is found that the objective measures of noisy speech, when determined using corresponding clear and enhanced speech are different for different languages over a range of SNR, giving clue to the type of the language in use.

1. Introduction

Speech is a signal which easily gets corrupted as it comes in contact with the environment. Except in sound-proof rooms used in studios, it is not possible to find such ideal noise free conditions in practice. Although a large number of noises exist in environment, broadly they can be classified into Factory, Babble, Engine, White and Channel noises etc. However most common kind of noise encountered is white noise, may it be in communication systems due to channel or generated in the equipment due to thermal or other electronic sources or combination of noises due to Central Limit Theorem (Aleksandr Lyapunov, 1901). Noise thus corrupts the speech, causing listener's fatigue and deteriorating performance of speech systems. Application of Speech enhancement or noise cancellation algorithms alleviates such problems to some extent. In literature several speech enhancement techniques exist. Though most traditional algorithms are based on optimizing mathematical criteria, they are not well correlated with speech perception and have not been as successful in preserving or improving quality in all regions of speech, especially transitional and unvoiced. Performance is also influenced by the specific type of noise, specific SNR, noise estimate updates and algorithm parameter settings. Spectral Subtraction technique of speech enhancement is popular and is still widely used as front end to speech systems for its simplistic nature and high quality performance except at very low SNRs (J. Lim, 1983).

Variety of languages exists in Indian region, with Dravidian, Tibeto-Burman, Indo-European, Indo-Aryan and Indo Iranian background. Mostly Indian languages are phonetic in nature that is there is one to one correspondence between sounds and the representative alphabet, and combining them creates similar kind of sounds. However different languages vary in its perceptibility due to differences in its phonetic contents and variations in distribution of different phonemes, stress level distribution among phonemes and of course intonation pattern, nasality usage, allophonic variants, contextual, phonotactic, or coarticulatory constraints etc.

Introduction of noise in speech distorts the speech spectrum and affects its phonetic perceptibility differently, due to the factors mentioned above. Enhancement of noisy speech though reduces the noise and subsequent irritation, but generally results in distortion of the speech spectrum. The kind and amount of distortion in the spectrum of enhanced speech will depend on the particular enhancement technique applied, and the SNR of the noisy speech. Therefore different types of speech units will get affected differently by the noise and subsequent enhancement.

In this paper, a novel work on identification of spoken languages, based on effect of distortion introduced by white noise in the phonetic contents of different Indian Regional languages namely Kashmiri, Bangla and Manipuri is reported. This kind of approach is not found in the literature for any other language as well. Effect of Speech enhancement technique namely spectral subtraction on noisy speech of these languages is also studied at different levels of segmental SNR. White noise has been considered for noisy spoken language, as it affects all frequency components of speech uniformly. The distortion introduced in the resulting speech is measured by estimating objective measures of perceptual speech quality such as LLR, LAR, IS and WSS (Hansen and Pellom, 1998). The variation of these estimated objective measures of the spectral distortion, with regard to a particular language, is studied and analyzed, to see language specific effects of the noise and enhancement algorithm, in order to provide clue to the identity of language in use.

The paper has been organized in the following form: Section 2 gives details of Spectral Subtraction technique of enhancement used. Section 3 gives a comparative study of phonotactics of the three languages i.e. Kashmiri, Bangla and Manipuri in brief. Section 4 introduces the objective measures used, namely LAR, IS, LLR and WSS. Section 5 describes the Results and discussion. Section 6 gives conclusions.

2. Spectral Subtraction

This technique of speech enhancement is computationally very efficient, particularly for stationary noise or slowly varying non-stationary noise. Spectral subtraction is a noise suppression technique used to reduce the effects of added noise in speech. It estimates the power of clean speech by explicitly subtracting the estimated noise power from the noisy speech power. This of course assumes that the noise and speech are uncorrelated and additive in the time domain. Also, as spectral subtraction based techniques necessitate estimation of noise during regions of non-speech activity, it is supposed that noise characteristics change slowly. However, because noise is estimated during speech pauses, this makes the method computationally efficient. Unfortunately, for these reasons, spectral subtraction is beset by a number of problems. First, because noise is estimated during pauses the performance of a spectral subtraction system relies upon a robust noise/speech classification system. If a misclassification occurs this may result in a misestimating of the noise model and thus a degradation of the speech estimate. Spectral subtraction may also result in negative power spectrum values, which are then reset to non-negative values. This results in residual noise known as musical noise. In a speech enhancement application it has been shown that, at 5 dB SNR, the quality of the speech signal is improved without decreasing intelligibility. However, at lower SNR speech this performance reduces rapidly. When used in Automatic Speech Recognition (ASR), the trade-off between SNR improvement and spectral distortion is important. To provide a mathematical description of the spectral subtraction technique, we write the spectrum of the noisy speech y (t) in terms of that of the clean speech x (t) and additive noise n (t) (the simplest acoustic distortion model):

$$y(t) = x(t) + n(t) \qquad -(1)$$

The enhancement is explained in the following formula (Berouti et al., 1979).

$$\hat{X}(w) = \left[\left|Y(w)\right|^{\lambda} - \alpha \left|\hat{N}(w)\right|^{\gamma} \right]^{1/\gamma} e^{j\theta_y(w)} \quad - (2)$$

$\hat{X}(w)$ and $Y(w)$ are DFT (discrete fourier transform) of the enhanced and noisy signal. N (w) is estimate of noise and θ_Y phase of original signal. λ is 2 for working in power spectrum domain and α is the over subtraction factor.

3. Characteristics of Manipuri, Bangla and Kashmiri spoken languages

Different Indian regional languages have certain linguistic background of their own and later have added certain foreign loan words. Their phonotactics and grammar is also quite distinct Following are features of above spoken languages:

Manipuri*:* It is a Tibeto-Burman language. Tone is used to convey phonemic distinction. Aspirates are present. High frequency of the velar nasal is particularly striking. Grammatical gender is missing. The normal order of words in a sentence is SOV-subject, object, verb, though this is not always and everywhere rigorously observed. Tibeto-Burman words are monosyllables. Phonological system of Manipuri can be categorized into two groups – segmental phonemes and supra-segmental phonemes. Segmental phoneme includes vowels and consonants and supra-segmental phoneme includes tone and juncture. All the six Manipuri vowels can occur in initial, medial and final position. There are six diphthong like sounds in Manipuri. They are

- (/əy/,/ay/, /əw/ ,/oy/, /uy/, /aw/)

There are 24 consonant phonemes in Manipuri p,t,k, ph,th,kh,m, n,ŋ,c,s,l, h,w,y,b d,g,bh, dh,gh,j, jh,r . Among these the last 9 voiced sounds are borrowed from other languages and they cannot occur in the initial and final position. Only four phonemes can occur in the second element of the cluster. They are w, y, r and l. It can occur only in the initial and medial position of a word. There are two types of tone in the language level and falling tone. Juncture, other than phonetic features, has a phonemic status.

Bangla: An Indo-Aryan language. Standard colloquial Bengali contains 35 essential phonemes. 5 non-essential phonemes which occur only as variants of other sounds or in borrowed foreign words & not used by all speakers. The ten aspirated stops and affricates are characteristics and essential sounds of the language. They are not simple but compounds.

Seven vowel phonemes occur with their opposite nasal phoneme. All may be long or short. Length is not considered to be phonemic. There is one 1st person pronoun, three 2nd person pronouns and three pairs of 3rd person pronouns with polite, informal, singular, plural discrimination. Pronoun and verb have no gender discriminatory word. Most of the sentences don't explicitly use verbs. Verbs are inflected in person (1st, 2nd, 3rd), in degrees of politeness (intimate, familiar, respectful), and in tense (past, present, future). Plural can be inflected by adding suffix –ra, -der, -era, -diger, -guli, -gulo, -gana. The dominant word order in Modern Bengali sentences is:

Subject + Indirect object + Direct object + Oblique object + Verb.

Kashmiri*:* All the vowels have a nasal counterpart. Nasalization is phonemic in Kashmiri. Palatalization is phonemic in Kashmiri. All the non-palatal consonants in Kashmiri can be palatalized. There are eight pairs of short and long vowels. Kashmiri is a syllable-timed language, sometimes; individual words are stressed for emphasis. There are four major types of intonational patterns: (1) High - fall, (2) High - rise, (3) Rise &fall, (4) Mid - level. Intonations have syntactic rather than emotional content. Vowels /ə/, /o/, /ɔ:/ do not occur in the word final position. The short vowels /ɨ/, /e/, /u/, and / ɔ/ do not occur in the word-initial position. Usually the semi-vowel /y/ is added in the initial position of the words beginning with /i/, /i:/, /e/ and /e:/. Similarly, the semi-vowel /v/ is added to the words beginning with /u/, and /u:/. Vowel sequences usually do not occur in Kashmiri. Word initial consonant clusters are not as frequent as the word medial consonant clusters. Kashmiri has (C)(C)V(C)(C) syllable structure.

4. Objective methods of speech quality measure

In general speech enhancement or noise reduction is measured in terms of improvement in SNR, but in reality, this may not be the most appropriate performance criteria for improvement of perceptual speech quality. Humans do have an intuitive understanding of spoken language quality, however this may not be easy to quantify. In a number of studies, it has been shown that impact of noise on degradation of speech quality is non uniform. An objective speech quality measure shows, the level of distortion for each frame, across time. Since speech frequency content varies, across time, due to sequence of phonemes, needed to produce the sentence, impact of background distortion will also vary, causing some phone classes to get more effected than others, when produced in a noisy environment. Objective methods rely on mathematically based measure between reference signal and the signal under consideration. The objective measures are based on different parametric representation of the speech, and differ due to inclusion or non-inclusion of various parameters and the different weightage given to them, in order to imitate auditory model and perception as closely as possible. The details of each one is given below.

Itakura-Saitio Distortion Measure (IS): If for an original clean frame of speech, linear prediction (LP) coefficient vector is $\vec{a}_{\Phi}$, correlation matrix is R_{Φ}. And for processed speech LP coefficient vector is $\vec{a}_d$, correlation matrix is R_d, then Itakura-Satio distortion measure is given by,

$$d_{IS}\left(\vec{a}_d\ ,\vec{a}_\phi\right)=\left[\frac{\sigma_\phi^2}{\sigma_d^2}\right]\left[\frac{\vec{a}_d R_\phi \vec{a}_d^T}{\vec{a}_\phi R_\phi \vec{a}_\phi^T}\right]+\log\left[\frac{\sigma_d^2}{\sigma_\phi^2}\right]-1 \qquad - (3)$$

Where σ_d^2 and σ_Φ^2 represents the all-pole gains for the processed and clean speech frame respectively.

Log-Likelihood Ratio Measure (LLR): The LLR measure is also referred to as the Itakura distance. The LLR measure is found as follows,

$$d_{LLR}\left(\vec{a}_d\ ,\vec{a}_\phi\right)=\log\left[\frac{\vec{a}_d R_\phi \vec{a}_d^T}{\vec{a}_\phi R_\phi \vec{a}_\phi^T}\right] \qquad - (4)$$

Log-Area-Ratio Measure (LAR): The LAR measure is also based on dissimilarity of LP coefficients between original and processed speech signals. The log-area-ratio parameters are obtained from the pth order LP reflection coefficients for the original $r_\Phi(j)$ and processed $r_d(j)$ signals for frame j. The objective measure is formed as follows,

$$d_{LAR}=\left|\frac{1}{M}\sum_{i=1}^{M}\left[\log\frac{1+r_\phi(j)}{1-r_\phi(j)}-\log\frac{1+\hat{r}_d(j)}{1-\hat{r}_d(j)}\right]\right|^{\frac{1}{2}} \qquad - (5)$$

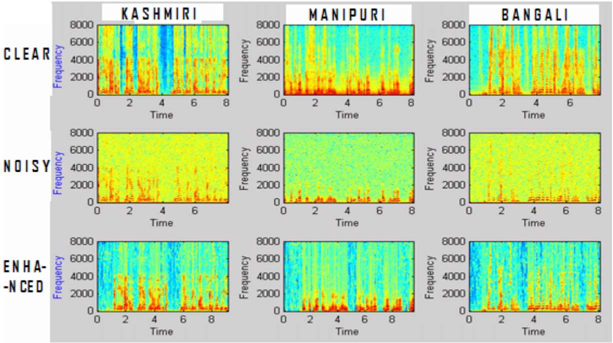
Weighted Spectral Slope Measure (WSS): The WSS measure by Klatt (1982) is based on an auditory model, in which 36 overlapping filters of progressively larger bandwidth are used, to estimate the smoothed short-time speech spectrum. The measure finds a weighted difference between the spectral slopes in each band. The magnitude of each weight reflects whether the band is near a spectral peak or valley, and whether the peak is the largest in the spectrum. A per-frame measure in decibel is found as

$$d_{WSS}(j)=K_{spl}\left(K-\hat{K}\right)+\sum_{k=1}^{36}w_a(k)\left(S(k)-\hat{S}(k)\right)^2 \qquad - (6)$$

where K, $\hat{K}$ are related to overall sound pressure level of the original and enhanced utterances, and K_{spl} is a parameter which can be varied to increase overall performance.

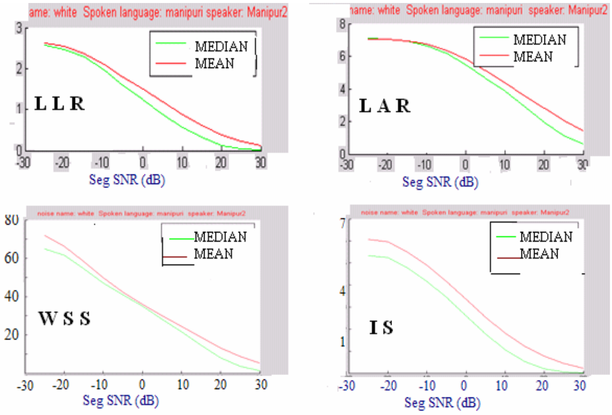
5. Results and Discussion

Sentences Spoken by 30 native speakers for each language namely Manipuri, Bangla and Kashmiri were recorded at 16 KHz. Noisy speech with white noise was simulated with segmental SNR from 30 dB to -20 dB. Objective measures i.e. IS, LAR, LLR and WSS are computed for each frame, with length ~ 512 samples. In first experiment these measures are computed for the noisy speech with reference to the corresponding clean speech sentence, whereas in second experiment the objective measures are computed using enhanced speech and the corresponding noisy speech for different sentences of the languages. Estimates of these measures are determined for the complete sentence using two methods, namely 5% trim mean and median of their values computed for each frame. Spectral subtraction method of enhancement is applied to obtain enhanced speech from the noisy speech sentences. For 10 dB SegSNR noisy speech, the spectrograms of the speech in three languages corresponding to Clean, Noisy and Enhanced, is shown in figure 1. It is observed through the spectrograms, that the noise has affected the three languages differently.



“Figure 1. Speech Spectrograms descriptions Rows: 1st-Clear, 2nd-Noisy, 3rd-Enhanced; Columns: 1st-Kashmiri, 2nd -Manipuri, 3rd -Bangla”

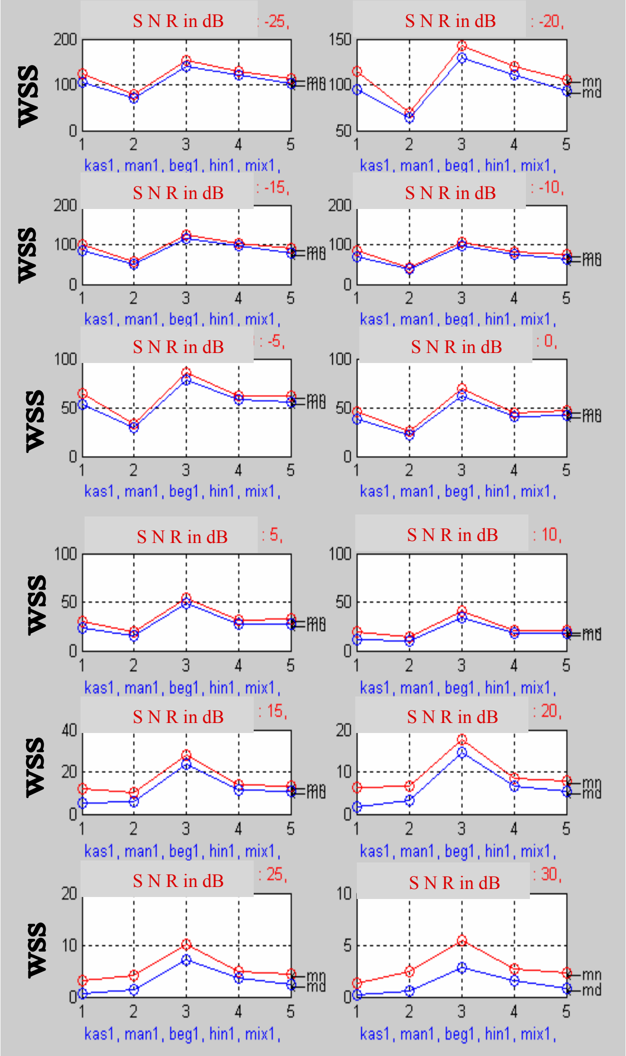
Estimates of LLR, LAR, IS and WSS are computed for SNR range 30 dB to -30 dB for different speech sentences in the three languages using noisy and clear and then enhanced and noisy speech. It is seen that WSS measure has the widest dynamic range almost 10 times the other measures



“Figure 2. LLR, LAR, IS and WSS estimates vs. SNR plots in Manipuri Speech with experiment-1.”

as shown in figure 2. of experiment -1, using Manipuri Speech. Thus it can be seen, that WSS is most suitable for the studies of distortion effects, of noise and enhancement algorithm, on different spoken languages.

WSS estimates of noisy speech, at different SNR are computed, as in experiment-1, and plotted in figure 3. It is observed that Manipuri is having lowest WSS estimate followed by Kashmiri and then highest for Bangla. This trend is more prominent particularly for low SNRs. The other points of plot are for Hindi and mixed languages.



“Figure 3. Plots of WSS estimates (y axis) as in experiment 1. , for different SNRs in dB i.e. -25, -20, -15, -15, -10, 0, 5, 10, 15, 20, 25, 30 for Kashmiri, Manipuri, Bangla, Hindi and mixed languages (denoted in x axis by 1, 2, 3, 4, 5 respectively)”

In experiment 2, WSS and LAR estimates are computed for enhanced speech, with reference to the corresponding noisy speech, for the three languages namely Kashmiri, Manipuri, and Bangla. The enhanced speech is obtained after application of spectral subtraction algorithm on noisy speech of different SNRs, ranging from 30 dB to -20 dB in steps of 5 dB. The mean and median estimates of the WSS for the 2nd experiment are shown in table 1. Here also the WSS estimate is lowest for Manipuri, followed by Kashmiri and Bangla is the highest. This trend is more prominent for low SNRs.

SNR in dB	Language	WSS Estimates	
		Median	Mean
30	Kashmiri	36.22793	42.52874
	Manipuri	32.12041	36.83813
	Bangla	38.06589	42.30879
25	Kashmiri	40.25494	45.34821
	Manipuri	34.70880	39.67888
	Bangla	42.705033	48.92245
20	Kashmiri	46.72147	53.09616
	Manipuri	38.03188	42.95194
	Bangla	51.42718	57.53441
15	Kashmiri	53.70700	60.94677
	Manipuri	45.73857	51.09685
	Bangla	60.85805	67.17440
10	Kashmiri	65.43084	71.24645
	Manipuri	58.61265	71.94426
	Bangla	70.73388	77.70258
0	Kashmiri	87.72349	92.32025
	Manipuri	71.26169	78.03224
	Bangla	92.23746	97.70964
-5	Kashmiri	94.50976	97.43755
	Manipuri	78.38978	83.14540
	Bangla	101.4064	104.6625
-10	Kashmiri	98.70403	100.8097
	Manipuri	85.42304	91.05538
	Bangla	105.2050	109.2263
-20	Kashmiri	101.8426	106.9107
	Manipuri	96.24993	101.5472
	Bangla	109.1610	112.9643

"Table 1. Median and Mean estimates of WSS for Enhanced speech in Kashmiri, Manipuri and Bangla for SNRs -30 dB to 20 dB as in Experiment 2."

6. Conclusion

In this paper a study is done for possibility of using LLR, LAR, IS and WSS as objective measures of speech quality, for discrimination of Indian regional languages namely Kashmiri, Manipuri and Bangla. This is done by computing estimates of ofthese objective measures for noisy speech with white noise for the above spoken languages and at SNRs -30 dB to 30 dB. First these measures are computed for noisy speech with reference to corresponding clear speech and then for the enhanced speech with reference to the corresponding noisy speech. WSS has proved to be the most useful measure used due to its wider dynamic range. The two estimates of WSS do provide clue to the type of language in use due to differences in its phonetic content. The discrimination provided is highest at lower SNRs. The estimate being lowest for Manipuri, and highest for Bangla. The reason could be attributed to the presence of weaker speech units in relatively higher concentration, in the language with higher WSS estimates compared to others; as the speech parameters under consideration for them, would undergo higher distortion under the influence of noise.

7. Acknowledgement

The authors are thankful to Director, Dr. P K Saxena and Dr S S Bedi for encouraging us to carry out this work and allowing presentation of the paper.

References

A.F Martin F.J. Godman and R.E.Wohlford. 1989. *Improved automatic language identification in noisy speech.* Proc Int Conf Acoust. Speech, and Signal Processing . May. 528-531

John H.L. Hansen and Bryan L. Pellom. Nov 1998. *Speech Enhancement and Quality Assessment:* A Survey, IEEE Signal Processing magazine

J. Lim. 1983. *Speech Enhancement.* Prentice Hall Englewood Cliffs, NJ.

Klatt, D.1982. *Prediction of perceived phonetic distance from critical-band spectra.* Proc. of IEEE Int. Conf on ASSP, 1278-1281.

M. Berouti, R. Schwartz and J. Mahoul.1979. *Enhancement of Speech corrupted by acoustic noise.* ICASSP, 208-211