SPEECH INTELLIGIBILITY PREDICTION AND NEAR END LISTENING ENHANCEMENT FOR MOBILE APPLICIATION

V.Shruthi[1], R.Senthamizhselvi[2], G.R.Suresh[3]

[1]PG scholar, Department of Electronics and Communication
Easwari Engineering College Chennai, (India)
[2]Associate professor, Department of Electronics and Communication
Easwari Engineering College Chennai, (India)
[3]Professor, Department of Electronics and Communication
Rajalakshmi Institute of Technology Chennai, (India)

ABSTRACT

Mobile telephony is often performed in the presence of background noise, such as traffic noise or murmur. In this situation, the near-end listener perceives a mix of clean speech from the far end and ambient noise coming from the near-end side, which entails a greater listening effort and possibly an intelligibility of the lower speech. This article deals with the problem of predicting the average intelligibility of noisy and potentially processed vocal signals towards the end, as observed by a group of listeners with normal hearing. The proposed model can make a short-term prediction based on the hypothesis that the intelligibility is monotonic correlated with the mutual information between the amplitude envelopes of the critical cleaning signal band and the corresponding noise signal. The resulting intelligibility predictor is a simple function of the mean square error(MSE) that occurs when an amplitude of the clean critical band is estimated using a minimum mean square error(MMSE) estimator based on the noise amplitude. The proposed model predicts that speech intelligibility will be improved by processing the cochlear filter of noisy critical bandwidths.

Keywords- ; Mean square error (MSE), Minimum mean square error (MMSE).

I.INTRODUCTION

Cell phones can work incredibly well, such as music players, web browsers and email clients, but it's easy to overlook their main function: allowing two people to converse from a distance. Even the most modern and feature-rich smartphone becomes useless if the user can not hear the voices of the callers clearly[11]. And because ambient noise is one of the factors that can reduce speech intelligibility in mobile voice calls, mobile phone manufacturers now employ a variety of software-based techniques to mitigate their effects. Because everyone who uses a cell phone in a busy train station, airport, or sports stadium knows these software techniques typically leave the user with unsatisfactory sound quality in voice calls in noisy environments. This is because these techniques of

digital signal enhancement and automatic volume control are approximate methods that can only mitigate the effects of ambient noise and not eliminate it. Active Noise Canceling (ANC) technology, on the other hand, is a well-known method for attenuating unwanted ambient noise very effectively, but is currently only found in high-end stereo headphones. Here, it is much appreciated by wealthy consumers, for example, frequent travelers who wish to stop the irritating roar of jet engines.

The NELE algorithm was the one that maximizes the speech intelligibility index (SII) and, therefore, the speech intelligibility through the selective increase in the frequency of the vocal signal strength [12-13]. The filtering in the time domain with the filter coefficients adapted in the frequency domain was carried out using a frequency deformed filter bench equalizer. This allows processing with a spectral resolution of the Bark scale according to the human auditory system and a low signal delay.

Applications where the loudspeaker signal strength is considered are limited to the original signal strength [8]. A recursive optimization of the closed-form solution of the spectral vocal signal power allocation is obtained that maximizes the SII under this restriction. However, for small speakers used in mobile phones, the thermal load during continuous playback is an important limitation. Therefore, most mobile phone applications limit the overall power of the speaker signal to a constant maximum power instead of the original signal strength.

II. METHODOLOGIES USED

A. Short time objective intelligibility prediction

The basic structure of STOI is illustrated in Fig.1. It is a clean and degraded language function, indicated by and, respectively. The STOI output is a scalar value that is expected to have a monotone relationship with the average intelligibility of (for example, the percentage of words correctly understood mediated in a group of users). A sampling rate of 10 kHz is used to capture a frequency range related to speech intelligibility [9-10]. First, both signals are decomposed by TF to obtain a simplified internal representation that resembles the transformation properties of the auditory system. This is achieved by segmenting both signals into the 50% overlap, the frames with a Hann window with a length of 256 samples, in which each frame has zero fills up to 512 samples. Before evaluation, silent regions that do not contribute to speech intelligibility are eliminated. This is done first by finding the box with the maximum energy of the clean voice signal. Both signals are then reconstructed, excluding all frames in which the energy of the clean voice is less than 40 dB compared to this frame of maximum clean speech energy[15-18]. Thus, a one third of octave band analysis is performed by grouping the DFT-bins. In total, 15 one-third of octave bands are used, where the lowest central frequency is set at 150 Hz and the highest one-third of octave band has a central frequency of about 4.3 kHz.

Let X (k, m) denote the DTH-bin frame of mth frame of a clean speech[1-5]. The standard jth of the third-octave band, called the TF unit, is defined as

$$\mathbf{x}_{\mathbf{j(m)}} = \sqrt{\sum\nolimits_{\mathbf{k=}}^{\mathbf{k2}}} \qquad \textbf{(1)}$$

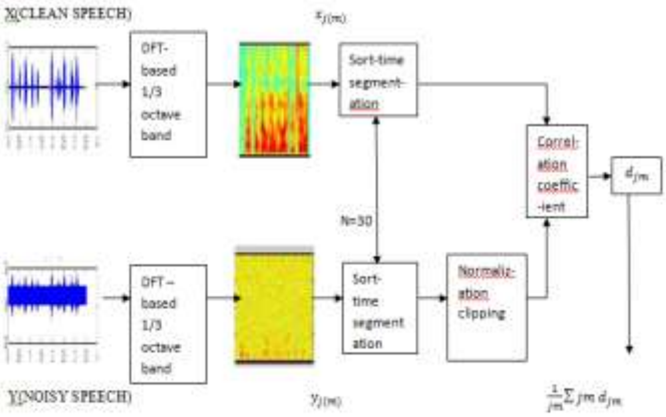


Figure 1:Block diagram of STOI

STOI is a function of the clean and degraded speech, which are first decomposed into DFT-based, one-third octave bands. Next, short-time (384 ms) temporal envelope segments of the clean and degraded speech are compared by means of a correlation coefficient. Before comparison, the short-time degraded speech temporal envelopes are first normalized and clipped (see text for more details). These short-time intermediate intelligibility measures are then averaged to one scalar value, which is expected to have a monotonic increasing relation with the speech intelligibility.

Where k1and k2 denote the one-third octave band edges, which are rounded to the nearest DFT-bin. The TF-representation of the processed speech is obtained similarly, and is denoted by Yj(m).

STOI is a function of a TF-dependent intermediate intelligibility measure, which compares the temporal envelopes of the clean and degraded speech in short time regions by means of a correlation coefficient. The following vector notation is used to denote the short-time temporal envelope of the clean speech

$$= \quad [\mathbf{xj(m-N+1), Xj(m-N+2), \ldots, Xj(m)}]^{\mathbf{T}}$$

B. Cochlear filter

A In this work, by exploiting the hybrid of Cochlear Filter and Short Time Objective Intelligibility Prediction algorithm improves the quality of speech and intelligibility which in turn perform this in the less duration and with low complexity[14-15]. As shown in fig 2, which explain about the cochlear filter based intelligibility prediction. The parameter extraction procedure for auditory-based spectral coefficients, consists of series of cochlear filter bank based on the auditory transform, hair cell function, nonlinearity and Discrete Cosine Transform (DCT).

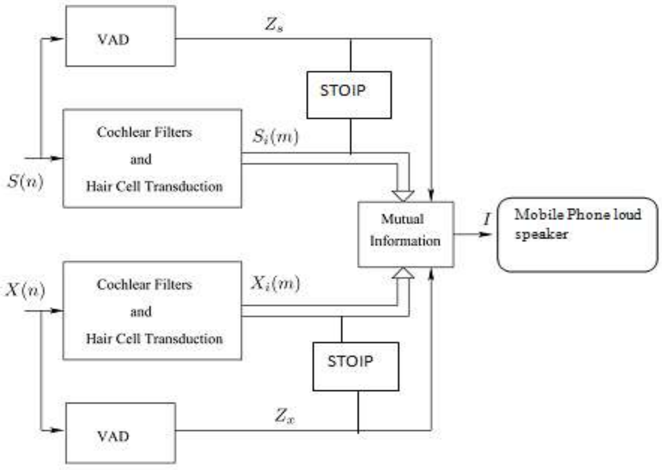

VAD
Z_s
STOIP
Cochlear Filters and Hair Cell Transduction
$S_i(m)$
$S(n)$
Mutual Information
I
Mobile Phone loud speaker
$X(n)$
Cochlear Filters and Hair Cell Transduction
$X_i(m)$
STOIP
VAD
Z_x

Figure 2: Cochlear filter based intelligibility prediction scheme

C. Auditory transform

The auditory transform has well defined wavelet properties with an existing inverse transform. It converts the time domain signal into a set of filter bank output with frequency responses similar to those in the BM of the cochlea. Let *s*(*t*) be the speech signal and the cochlear filter be _(*t*). Thus, the auditory transform of *s*(*t*) (i.e., *W*(*a*,*b*)), with respect to _(*t*) as the impulse response of BM in the cochlea is defined as follows

$$\mathbf{W\ (a,b) = s(t) *\ \ {}_{a,b}\,(t)} \qquad \mathbf{(3)}$$

$$\mathbf{W\ (a,b) = \qquad *_{a,b}\,(t\text{-}T)dT} \qquad \mathbf{(4)}$$

$$\mathbf{{}_{a,b}\,(t) = \ 1 /} \qquad \mathbf{(5)}$$

where in eq. (1), * indicates convolution operation, *a* _ R+ and *b* _ R, *s*(*t*) and _(*t*) belongs to Hilbert space *L2*(R) and *W*(*a*,*b*) represents traveling waves in the BM. The factor *a* is the scale or dilation parameter, which allows changing the center frequency, *fc*, while factor *b* is the time shift or translation parameter. The energy remains equal for all *a* and *b*.

III. SIMULATION RESULTS

A. Noise database

NOIZEUS is a noisy speech corpus recorded at the Center for Robust Speech Systems, Department of Electrical Engineering, University of Texas, Dallas. The noisy database contains 30 IEEE sentences produced by three male and three female speakers (five sentences /speaker), and was corrupted by eight different real-world noises at different SNRs. Thirty sentences from the IEEE sentence database were recorded in a sound proof booth using Tucker Davis Technologies (TDT) recording equipment.

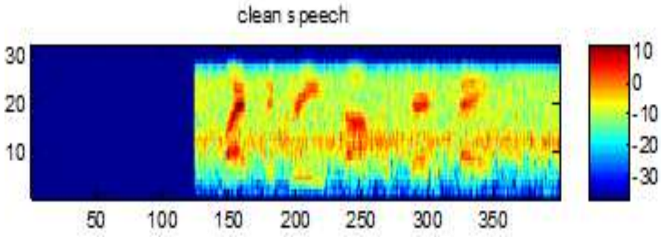
The sentences were originally sampled at 25 KHz and down sampled to 8 KHz. To simulate the receiving frequency characteristics of telephone handsets, the speech and noise signals were filtered by the modified Intermediate Reference System (IRS) filters used in ITU-T P.862 for evaluation of the PESQ measure. Noise signals were taken from the AURORA database and included the following recordings from different places: babble (crowd of people), car, exhibition hall, restaurant, street, airport, train, station, and train

The noise signals were added to the speech signals at SNRs of 0, 5, 10, and 15dB. From NOIZEUS database, different noise signals are added to the speech signal and are denoised using cochlear based STOIP modeling with different Signal to Noise Ratio (SNR) levels. The algorithm proposed in the previous chapter is implemented and tested with different database to analyze its performance

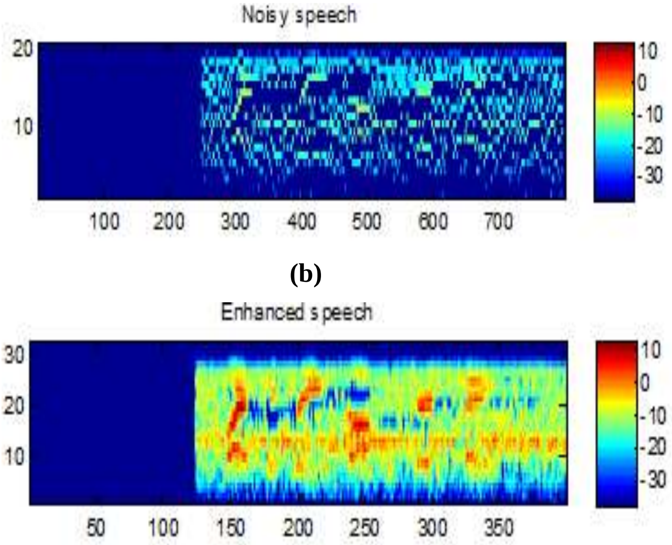
B. Short time objective speech intelligibility and quality prediction

The significant standardization efforts have been made by the International Telecommunications Union (ITU) for standardizing both intrusive and nonintrusive algorithms using NH listeners and mobile speaker[6,7]. On the other hand, only a handful of algorithms that are proposed are specifically tuned to assistive listening devices. In the following sections, the choice of measures used was guided only by the applicability to the task in Mobile Speaker, but also by the availability of publicly available source code licensed at a reasonable cost. The performance evaluation of this database contains IEEE sentences produced by male and 3 female speakers and was corrupted by 8 different real time noises at various levels of SNR at the input level to the Mobile Speakers Noise signals from the AURORA database is taken as input, also including the recordings from different environments such as: babble (multi talker), car, restaurant, exhibition hall, street and airport, station. The noisy signals were interpreted with the speech signals at SNRs of 0, 10,and 15dB. The clean signal which is subjective to different noisy signals is given as input to the Cochlear Implants, which is then processed with the noise suppression Algorithm. This process is evaluated using Signal to Noise Ratio (SNR) and Perceptual Evaluation of Speech Quality (PESQ) metrics.

C. *Spectrogram plot*



(a)



(b)

(c)

Figure 3:Spectrogram of (a) Clean speech (b) Noisy speech (c) Enhanced speech

D. SNR estimation

The signal-to-noise ratio (SNR) is one of the oldest and most used objective measures. It is mathematically simple to calculate, but requires distorted and non-distorted (clean) speech samples. Where, x (n) is a clean speech, x (n) a distorted speech and N the number of samples. This classic definition of SNR is not well correlated with the quality of speech for a wide range of distortions. Therefore, there are several variations of classic SNR that show a much higher correlation with subjective quality. It has been observed that classic SNR is not well correlated with voice quality because although the voice is not a stationary signal, SNR averages the relationship in the whole signal. The energy of the speech fluctuates over time, so the parts where the speech energy is large and the relatively inaudible noise should not be washed from other parts where the speech energy is small and the noise can be heard from the speech . Therefore, the SNR was calculated in short squares and then calculated as an average. This measure is called segmental SNR and can be defined as where *L* is the frame length (number of samples), and *M* the number of frames in the signal *(N = ML)*. The frame length is normally set between 15 and 20 ms. Since, the logarithm of the ratio is calculated before averaging, the frames with an exceptionally large ratio is somewhat weighed less, while frames with low ratio is weighed somewhat higher. It can be observed that this matches the perceptual quality well, i.e., frames with large speech and no audible noise does not dominate the overall perceptual quality, but the existence of noisy frames stands out and will drive the over all quality lower.

However, if the speech sample contains excessive silence, the overall *SNR*seg values will decrease significantly since silent frames generally show large negative *SNR*seg values. In this case, silent portions should be excluded from the averaging using speech activity detectors. In the same manner, exclusion of frames with excessively large or small values from averaging generally results in *SNR*seg values that agree well with the subjective quality. A typical value for the upper and the lower ratio limit is 35 and −10 dB. These ranges are also used for *SNR*seg calculation throughout this book. Another variation to the SNR is the frequency-weighed SNR (fwSNRseg). This is essentially a weighted SNRseg within a frequency band proportional to the critical band. The fwSNRseg can be defined as follows

where *W(j,m)* is the weight on the *j*th sub band in the *m*th frame, *K* is the number of sub bands, *X(j,m)* is the spectrum magnitude of the *j*th sub band in the *m*th frame, and ˆ*X(j,m)* its distorted spectrum magnitude.

SNR COMPARISION

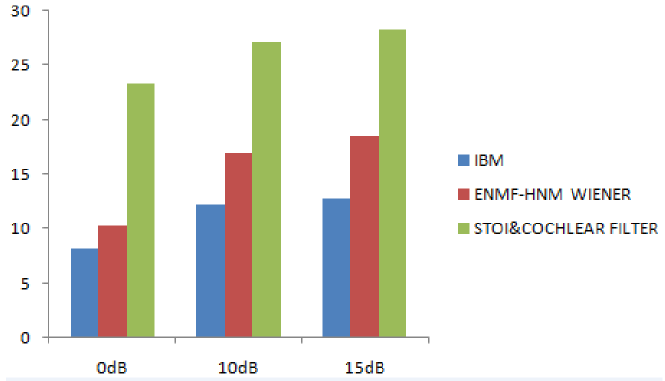


Figure4:Output SNR comparison for IBM,ENMF-HNM,STOI &Cochlear filter

From the fig:4 shows the SNR comparison in dB for 0dB Exhibition noise .It can be seen from the figure that compared
to IBM,ENMF-HNM wiener methods ,STOI and cochlear filter showed improved performance for various noise types and at various input SNR levels .

Table 1

Output signal to noise ratio result at different input SNR levels

NOISE	METHOD	SNR (0 dB)	SNR (10 dB)	SNR (15 dB)
Car	ENMF-HNM Wiener	10.0385	12.2582	19.5275
	IBM	8.1860	12.2537	12.7424
	STOI and Cochlear filter	24.4382	27.4981	28.1385

Exhibition	ENMF-HNM Wiener	10.3592	16.9273	18.5176
	IBM	8.1875	12.2636	12.7321
	STOI and Cochlear filter	**23.3754**	**27.1690**	**28.3633**

E. PESQ estimation

Perceptual assessment of speech quality (PESQ) is an international standard for estimating the average opinion score (MOS) of both the clean signal and its degraded signal. It has been developed from a number of previous MOS estimation attempts and is considered one of the most sophisticated and accurate estimation methods available today. PESQ has been officiallystandardized by the International Telecommunication Union Telecommunication Standardization Sector (ITU-T) as standard P.862 in February 2001PESQ uses a perceptive model to hide the degraded input and speech in an internal representation. The degraded entry is aligned over time with the original signal to compensate for the delay that may be associated with degradation. The difference in the internal representations of the two signals is used by the cognitive model to estimate the MOS. The PESQ values obtained using the cochlear filter and the STOI method and the same methods used separately are compared and the values are tabulated. PESQ scores were expressed using the mean auditory quality objective score scale (MOS LQO) and range from 1 (worst quality) to 5 (best quality)

.PESQ COMPARISON

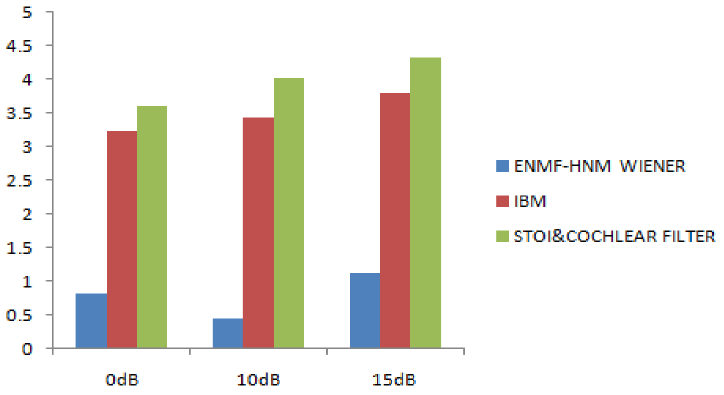


Figure5:Output PESQ Comparison for IBM,ENMF- HNM,STOI &Cochlear filter

Table 2

Result of objective measure(PESQ) with different input SNR(0dB,10dB,15dB)

NOISE	METHOD	PESQ with (0 dB)	PESQ with (10 dB)	PESQ with (15 dB)
Car	ENMF-HNM Wiener	0.6692	1.1825	1.1377
	IBM	3.0189	3.4402	3.6508
	STOI and Cochlear filter	3.8967	3.9429	4.2298
Exhibition	ENMF-HNM Wiener	0.8322	0.4529	1.1323
	IBM	3.2391	3.4363	3.7953
	STOI and Cochlear filter	**3.5980**	**4.0127**	**4.3279**

IV. CONCULSION

The proposed speech enhancement method combines STOI and cochlear filter. The combined technique reduces the near end noise and also increase the intelligibility of the speech signal. The combined algorithm shows the better results in the evaluation parameters such as SNR (dB) and PESQ (Out of 4.5) at various noise levels than the existing algorithms. The maximum SNR(28.3633) and PESQ (4.3279) achieved by the proposed method.

REFERENCES

[1] G. Kim, Y. Hu, and P. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," J. Acoust. Soc. Amer., vol. 126, no. 3, pp. 1486–1494, 2009.

[2] N.Madhu, A.Spriet, S.Jansen, R.Koning, and J.Wouters, "The potential for speech intelligibility improvement using the ideal binary mask and the ideal wiener filter in single channel noise reduction systems: Application to auditory prostheses," IEEE Trans. Audio, Speech, Lang. Process., vol. 21, no. 1, pp. 63–72, Jan. 2013

[3] N.Li and P.C.Loizou, "Factors influencing intelligibility of binary-masked speech: Implications for noise reduction," Journal of the Acoustical Society of America, vol. 123, pp. 1673–1682, 2008.

[4] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," J. Acoust. Soc. Amer., vol. 125, no. 4, pp. 2336–2347, 2009.

[5] M. C. Anzalone, L. Calandruccio, K. A. Doherty, and L. H. Carney, "Determination of the potential benefit of time-frequency gain manipulation," Ear Hearing, vol. 27, no. 5, pp. 480–492, 2006.

[6] E.W.Healy, S.E.Yoho, Y.Wang, and D.Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners,"J.Acoust.Soc. Amer., vol. 134, no. 4, pp. 3029–3038, 2013.

[7] Y. Hu and P. C. Loizou, "A comparative intelligibility study of singlemicrophone noise reduction algorithms," J. Acoust. Soc. Amer., vol. 122, pp. 1777–1786, 2007.

[9] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech perception of noise with binary gains," J. Acoust. Soc. Amer., vol. 124, no. 4, pp. 2303–2307, 2008.

[10] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," Speech Commun., vol. 49, pp. 588–601, 2007.

[11] I. Brons, R. Houben, and W. A. Dreschler, "Perceptual effects of noise reduction by time-frequency masking of noisy speech," J. Acoust. Soc. Amer., vol. 132, no. 4, pp. 2690–2699, 2012.

[12] R. J. van Hoesel and G. M. Clark, "Evaluation of a portable two microphone adaptive beamforming speech processor with cochlear implant patients," J. Acoust. Soc. Amer., vol. 97, no. 4, pp. 2498–2503, 1995.

[13] J. Wouters and J. Vanden Berghe,"Speech recognition in noise for cochlear implantees with a two microphone monaural adaptive noise reduction system," Ear Hearing, vol. 22, pp. 420–430, 2001.

[14] K. Kokkinakis, B. Behnam, Y. Hu, and D. R. Friedland, "Single and multiple microphone noise reduction strategies in cochlear implants," Trends Amplification, vol. 16, no. 2, pp. 102–116, 2012.

[15] S.J. Mauger, P. W. Dawson, and A. A. Hersbach, "Perceptually optimized gain function for cochlear implant signal-to-noise ratio based noise reduction," J. Acoust. Soc. Amer., vol. 131, no. 1, pp. 327–336, 2012.