VOL. 14, NO. 1, JANUARY 2019 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

PERFORMANCE IMPROVEMENT IN SPEECH ENABLED IVR SYSTEMS USING ARTIFICIAL BAND WIDTH EXTENSION

Mohan D.[1, 2] and K. Anitha Sheela[2]
[1]Department of Electronics and Communication Engineering, Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India
[2]Department of Electronics and Communication Engineering, JNTUH, Hyderabad, Telangana, India
E-Mail: mohan.aryan19@gmail.com

ABSTRACT

New blends of Speech Enabled Interactive Voice Response (SEIVR) Systems have been replacing the existing time consuming menu driven IVRs. But there is significant reduction in the quality and intelligibility of the speech signal when transmitted through the telecom networks which use narrowband codecs. Providing wideband quality signal without much modification of the existing network infrastructure can only be possible with a novel technique of Artificial Band Width Extension (ABWE). In this paper, ABWE is implemented using a QMF filter bank which band split the input speech into LF and HF components and further the HF components are compressed and encoded using a novel data hiding technique. In the reconstruction phase, an artificial wideband speech signal is generated through a QMF synthesis filter bank. For implementation of the proposed model, a client-server approach with socket programming on a single machine has been used assuming no noise and no transmission errors. A comparative analysis has also been done to find out the root cause for degradation in performances of SEIVR systems. The simulations on the proposed SEIVR model with ABWE have shown significant improvement in the speech recognition accuracy and overall performance.

Keywords: speech recognition, speech enabled IVR artificial bandwidth extension, narrow band coding.

1. INTRODUCTION

In today's world, a new blend of Speech Enabled Interactive Voice Response Systems (SEIVR) systems are gradually replacing the existing time consuming menu driven IVRs in most of the day-to-day services including customer care and phone banking. But even these SEIVRs have been failing due to their lesser accuracy in recognizing speech when transmitted over a communication line. Traditional telecom networks use a narrowband (NB) codec with a limited frequency range of about 200-3400 Hz only and hence cannot accurately transmit the speech signal generated by a human with the frequencies ranges from 0 Hz to 7000 Hz. This consequently results in significant reduction in the quality and intelligibility of the speech signal [1].

The ultimate solution to improve this accuracy would be to replace the NB codecs with Wide band (WB) speech codecs which support 50-7000 Hz. Since majority of the service providers are already using NB codec, providing wideband quality signal without much modification of the existing network infrastructure can only be possible with a novel technique of Artificial Band Width Extension (ABWE). This paper addresses the issues of existing SEIVR systems and proposes a new model to overcome the same by investigating the performance with recognition accuracy as an evaluation parameter. Better accuracy is achieved using the ABWE method wherein the WB spectral components are extracted as side information from the WB input speech of 16-KHz sampling rate [2].

The WB speech signal of 16-KHz sampling rate is fed as input to a two-channel QMF bank, which consists of a low-pass filter (LPF) and a high-pass filter (HPF). The output from the QMF bank is down-sampled or decimated into LF and HF components of 8-KHz each. The extracted LF components are encoded by the GSM-EFR Speech encoder whereas the HF components are converted into binary format and given as input to a Data hiding module, where, the HF parameters are compressed and encoded and the resultant bit stream is transmitted to the receiver through a narrowband communication network. At the receiving terminal, the LF component is decoded with GSM-EFR decoder, while the HF parameters are extracted using reversible data-hiding technique and thereby recovering the HF speech. After reconstruction, the LF and HF signals are interpolated and the artificial wideband speech signal is generated through a QMF synthesis filter bank [1, 2].

In this paper, Section 2 highlights the implementation of a speech enabled IVR system using Novel techniques of Artificial Band width Extension. Section 3 discusses BWE with side information for existing NB networks and Section 4 addresses the SEIVR performance with ASR Accuracy.

2. IMPLEMENTATION OF A SPEECH ENABLED IVR SYSTEM

The SEIVR system used in this paper has been implemented with a novel method of bandwidth extension of NB speech to provide a better wideband speech signal. To achieve this, a data hiding technique has been employed. In this approach, the high frequency components are extracted from the interpolated version of the high frequency speech signal. At the receiver, wideband speech signal is reconstructed and given as input to the Automated Speech Recognition (ASR) engine and later to a Text-To-Speech (TTS) engine which can speak back to the user in a voice that resembles closely with the original speech signal. Resources in the form of pronunciation lexicon and dictionary assist ASR and TTS

for better recognition accuracy. The following subsections discuss about the modules mentioned in Figure-1.

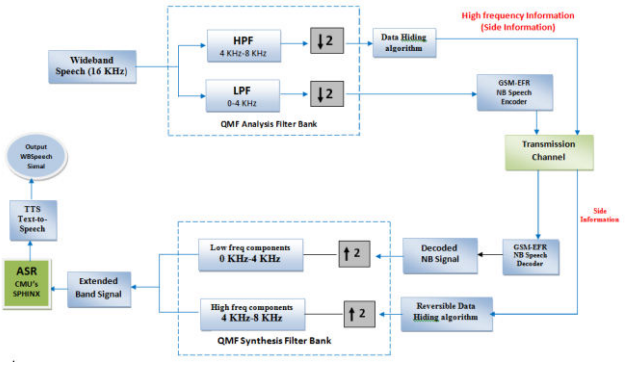


Figure-1. Speech enabled IVR using ABWE.

2.1 QMF filter bank

In general, a Quadrature mirror filter (QMF) bank splits the input signal into two equal bandwidth signals, using the low-pass and high-pass analysis filters. During this process, to achieve signal compression, the sub signals are decimated by a factor of two and later at the receiver, the decimated signals are interpolated by a factor of two and finally given as input to low-pass and high-pass synthesis filters respectively. The final outputs from the synthesis filters are combined and signal is reconstructed [3].

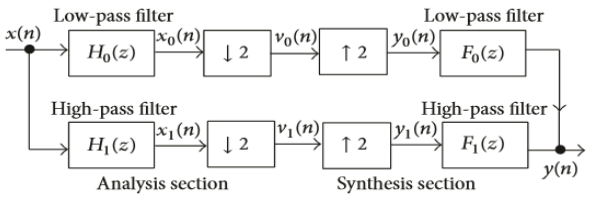


Figure-2. Two-Channel quadrature mirror filter bank.

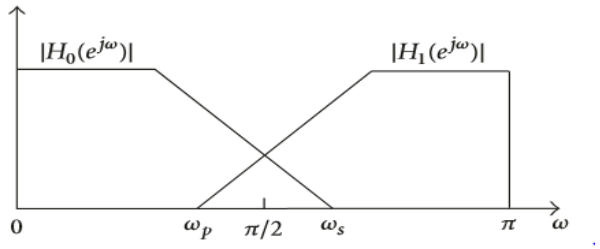


Figure-3. Frequency response of the analysis filters $H_0(z)$ and $H_1(z)$.

Magnitude response $|H_0(e^{j\omega})|$ is a mirror image of

$|H_1(e^{j\omega})|$ with respect to the quadrature frequency $\pi/2$ this has given rise to the name quadrature mirror filter bank.

2.2 Data hiding algorithms (Reforming and deforming)

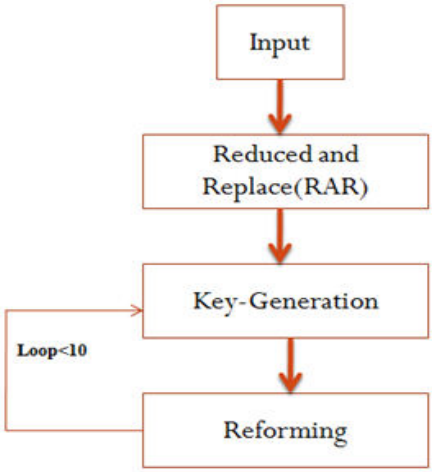
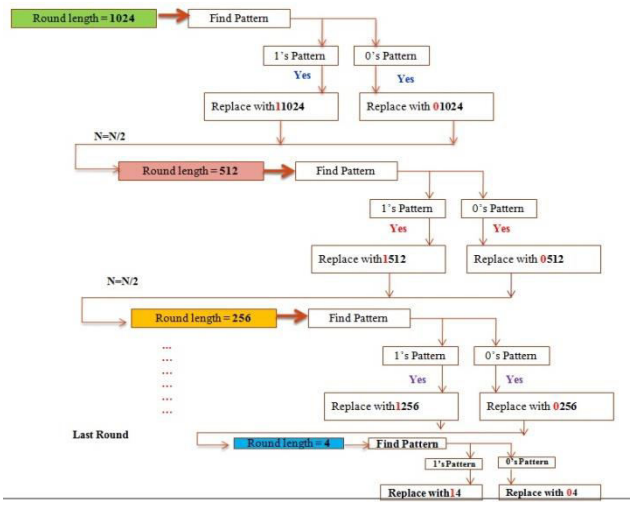


Figure-4. Proposed l data encoding technique.

Replace and reduce

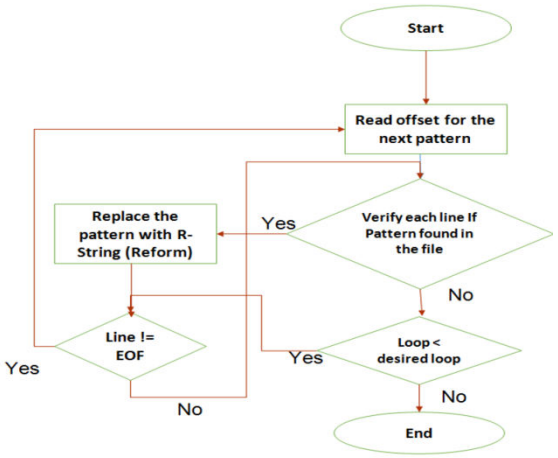
Voice data is given to Replace and Reduce (RAR)

- Pattern finding starts from the length 1024 (2^{10}) to 8 (2^3).
- This patterns are replaced with smaller piece of code ex: 1024 number of 0's are replaced with '01024'
- This way 1024 data length is reduced to data of length 5
- File size is reduced from R to R1.



Reforming

- Here user defined offset is considered to find the max occurrence of the pattern in the whole file.
- This pattern is replaced with 'R-String'
- Each time this replacement happens, we generate a metadata or Key file. This help us in deforming
- This is an iterative process, this can be done minimum of 2 power 3 times. This is optional number; this can be extended to any number of times from 8 to 1024.
- File size is reduced from R1to R2 (R2<R1)



Deforming

- Here the input is "file with R2 length" and "respective key file".
- From the above mentioned information, R-String is replaced with Key values which eventually give the original file.

This process is also reverse iterative process of Reforming and RAR. Such that we het the original file R from R2 without missing single bit of data.

3. GSM COMMUNICATION SYSTEM

The NB speech encoding and decoding in the proposed SEIVR model has been carried out using GSM-EFR speech and channel encoder of 260 bit capacity. The principle of GSM-EFR codec is based on Algebraic Code Excited Linear Predictive Coding (ACELP) with a bit rate of 12.2 kbps and 8 kHz sampling rate [4, 5]. Based on these specifications, 244 bits out of 260 represent the signal and the remaining 16 bits depict error control coding. A detailed flow of this GSM system is given in Figure-5.

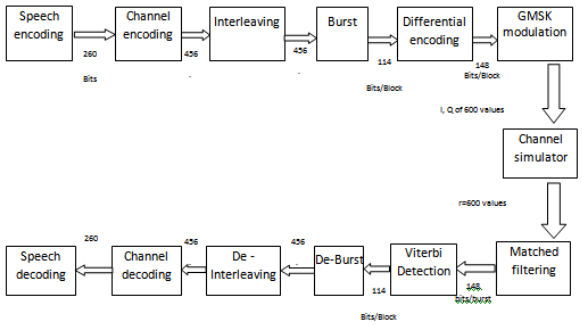


Figure-5. Functional elements of GSM System.

From the encoded speech signal, every 260 bits are grouped as a block. Every 260 bits are further divided into Class I with 182 bits and Class II with 78 bits. Class I (182 bits) is further divided into Class IA with 50 bits and Class IB with 132 bits. The 50 bits in Class IA pass through a Parity encoder and then combined with Class IB. Further, 179 bits from this combination pass through a Convolution encoder to double the bits resulting in 378 bits. Finally these 378 bits are augmented with Class II to get an output of 456 bits.

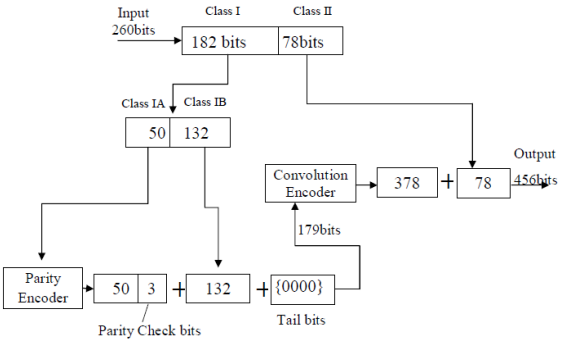


Figure-6. Classifications of bits in channel encoding.

To transmit these output bits, GMSK modulation is used and later Viterbi algorithm is used for demodulation. To recover the original speech signal, de-interleaving and de-convolution techniques are used.

4. EVALUATING SEIVR PERFORMANCE WITH ASR ACCURACY

To improve the performance of SEIVR systems, there has to be a drastic improvement in Speech recognition accuracy. This is possible if the input signal is a WB speech signal. Using the proposed model in this paper, we have generated an extended NB signal which resembles the WB quality. Now, using the NB and Extended NB (ENB) speech signals, we need to measure the recognition accuracy. For this investigation, we have used SPHINX-3 ASR toolkit, Speech database of TIMIT and NB, WB codec executable from the standard organizations like ITU-T, 3GPP, ETSI.

4.1. Utilization of narrow-band speech codec for ASR

In TIMIT Speech database, the original speech files are sampled at 16-kHz. But for our work, we have down sampled the data at 8-kHz since NB codec works only at 8-kHz sampling rate.

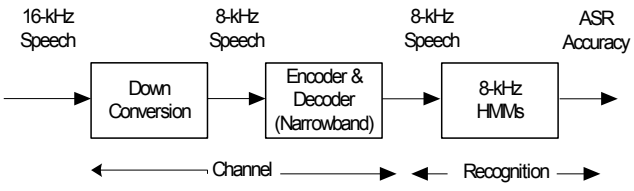


Figure-7. Speech recognition setup.

ASR Word Recognition Accuracy (A) is computed using the following relation:

$$A = [H - (I+S+D) / H] * 100\ \%$$

Where

- H - Total number of words,

- I - Total number of insertions,
- S - Total number of substitutions,
- D -Total number of deletions and

ASR Word Error Rate (WER) is obtained by
WER = (100-A) %

4.2 Result analysis

The proposed system assumes a client-server approach with socket programming on a single machine assuming no noise and no transmission errors. The performance of SEIVR system has been evaluated using Signal-Noise Ratio (SNR), Bit Error Rate (BER), ASR accuracy and Mean Opinion Score (MOS) as metrics.

Table-1. Performance of SEIVR for NB and ENB speech signals with NB codecs.

	NB codec with NB signal				NB codec with ENB signal			
Channel noise with gaussian distribution and variance	**SNR due to channel noise** (Avg)	**Bit error rate After channel decoding** (Avg)	**ASR Accuracy with EFR channel coding %**	**MOS with EFR channel coding**	**SNR due to channel noise** (Avg)	**Bit error rate After channel decoding** (Avg)	**ASR Accuracy with EFR channel coding %**	**MOS with EFR channel coding**
0.0001	36.989	0	92.9	3.982	38.7551	0	94.67	4.1066
0.0002	33.9794	0.000072	92.9	3.9817	35.4932	0.000062	94.41	4.1063
0.0003	32.2182	0.0037	92.8	3.9753	34.2366	0.003	94.82	4.0999
0.0004	30.9691	0.0198	92.9	3.9362	32.4829	0.0108	94.41	4.0608
0.0005	30.0001	0.05235	93.2	3.871	31.7662	0.04235	94.97	3.9333
0.0006	29.2082	0.1026	92.8	3.7752	30.722	0.0926	94.31	3.8375
0.0007	28.5388	0.1715	92.6	3.6364	30.0526	0.1092	94.11	3.6987
0.0008	27.9588	0.2552	92.2	3.4923	29.4726	0.1306	93.71	3.5546
0.0009	27.4475	0.3556	91.8	3.3166	28.9613	0.231	93.31	3.4412
0.001	26.9893	0.4777	91.3	3.1119	28.5031	0.4154	92.81	3.2365
0.0011	26.5757	0.6219	90.2	2.9028	28.0895	0.4973	91.71	3.0274
0.0012	26.1977	0.791	88.6	2.6842	28.2161	0.6664	90.62	2.8088
0.0013	25.8499	0.9911	86.5	2.4639	27.616	0.8665	88.27	2.5885
0.0014	25.5284	1.197	84	2.263	27.0422	1.1347	85.51	2.3876
0.0015	25.2288	1.4361	81.7	2.0814	26.9949	1.3738	83.47	2.206
0.0016	24.9482	1.7319	77.2	1.9125	26.9666	1.6073	79.22	2.0371
0.0017	24.6852	2.007	73	1.7688	26.4513	1.9447	74.77	1.8311
0.0018	24.4368	2.3602	66.9	1.6419	26.2029	2.2356	68.67	1.7665

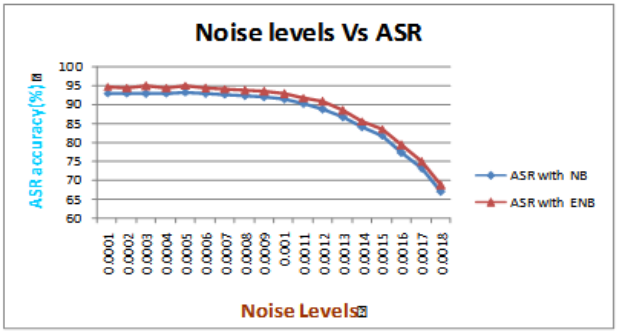


Figure-8. Noise levels Vs ASR.

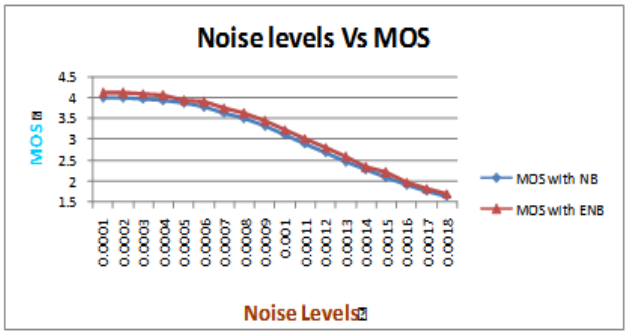


Figure-9. Noise levels Vs MOS.

Mean opinion score suddenly decreases with SNR, there is slightly degraded ASR accuracy due to the speech and channel coding.

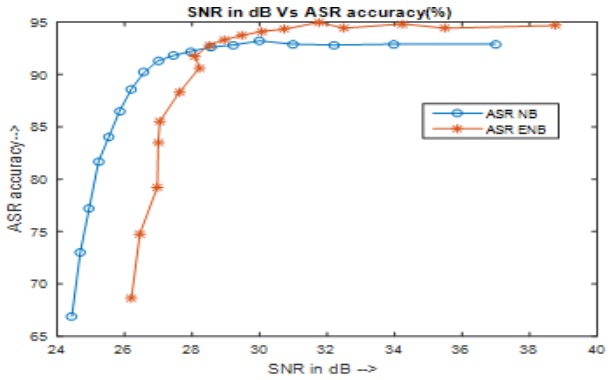


Figure-10. SNR in dB Vs accuracy.

The figures shows, how the ASR accuracy, MOS are affecting with respect to signal to noise ratio (SNR) due to different channel noises. The performance of ASR does not have the drastic change until the signal to noise ratio reaches to 30dB, after this drastically decreased with respect to signal to noise ratio.

CONCLUSIONS

Based on the simulated results it can be concluded that the recognition accuracy with artificially extended NB speech signal is far better than NB speech signal and close to that of a WB speech signal. Also, performance of the SEIVR system employing ABWE technique is much better than a traditional SEIVR system.

REFERENCES

[1] Bernd geiser, Peter Jax and Peter Vary. Robust wideband enhancement of speech by combined coding and artificial bandwidth extension. Institute of Communication Systems and Data Processing RWTH Aachen University, Templergraben 55, Aachen, Germany.

[2] Chen and H. Leung. 2005. Artificial bandwidth extension of telephony speech by data hiding. in Proc. of Intl. Symp. on Circuits and Systems (ISCAS), Kobe, Japan, May.

[3] 2006. ITU-TRecommendationG.729.1:G.729-based embedded variable bit-rate coder: an8-32kbit/s scalable wideband coder bit stream interoperable with G.729, (ITU-T).

[4] N. N. Katugampala, K.T. Al-Naimi, S. Villette and A. Kondoz. 2004. Real time data transmission over GSM voice channel for secure voice and data applications. In: Proceedings of the 2nd IEE Secure Mobile Communications Forum: Exploring the Technical Challenges in Secure GSM and WLAN, London, UK, September.

[5] ETSI Standards, http://www.etsi.org.