Volume 1,Issue-1,December-2011

ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

A Monthly Scientific Journal of Computer Science

www.ijarcsse.com

Research paper

Available online at www.ijarcsse.com

IMPROVED SPEECH QUALITY FOR VMR - WB SPEECH CODING USING EFFICIENT NOISE ESTIMATION ALGORITHM

Mr. M. Mathivanan
Associate Professor/ECE
Selvam College of Technology
Namakkal, Tamilnadu, India

Dr. S.Chenthur Pandian
Principal
Dr. Mahalingam College of Engineering and Technology
Pollachi, Tamilnadu, India

Mr. D. Palmani
II M.E., Department of ECE,
Paavai College of Engineering,
Namakkal, Tamilnadu, India

ABSTRACT

The recent development of mobile communication has the great challenge in transparent communication (better speech quality and intelligibility) under non-stationary environment. The VMR-WB speech coding technique selected by 3GPP faces degraded speech quality due to the use of noise reduction algorithm which is suited only for stationary environment. The Voice Activity Detection(VAD) Algorithm used in VMR-WB will not work well in more realistic environments. This paper presents novel noise estimation algorithm for wideband speech coding under realistic environment conditions. The MCRA (Minima Controlled Recursive Averaging) algorithm will give the improvement of speech quality and updates the noise estimation by tracking the noise region of noisy speech spectrum. The noise estimate is given by averaging past spectral power values, using a smoothing parameter that is adjusted by the signal presence probability in subbands. Presence of speech in subbands is determined by the ratio between the local energy of the noisy speech and its minimum within a specifed time window. The noise estimate is computationally effcient, robust with respect to the input signal-to-noise ratio and type of underlying additive noise, and characterized by the ability to quickly follow abrupt changes in the noise spectrum.

Keywords

VMR-WB (Variable Rate Multi mode – Wide Band), CELP, VAD, MCRA, AMR (Adaptive Multi Rate).

1. INTRODUCTION

A noise estimation algorithm plays an important role in speech enhancement. Speech enhancement for automatic speaker recognition system, Man–Machine communication, Voice recognition systems, speech coders, Hearing aids, Video conferencing and many applications are related to speech processing. VMR-WB[1] was originally designed as a cdma2000 native codec for wideband or narrowband voice and multimedia services. The cdma2000 standards are developed by Third-Generation Partnership Project (3GPP). Compared to the traditional narrowband (NB) telephony bandwidth of 200–3400 Hz, the wideband (WB) speech signal of 50–7000 Hz provides substantially improved speech quality, naturalness, and adds a feeling of transparent communication. The importance of wide band speech for 3G mobile communications has been recognized within 3GPP by adopting the Adaptive Multi Rate Wideband (AMR-WB) speech codec [2]. The same codec has been subsequently adopted by the International Telecommunication Union (ITU). The operation of VMR-WB is controlled by speech signal characteristics (i.e., source-controlled) and by the traffic condition of the network through selection of the mode of operation. The VMR-WB core technology is based on AMR-WB codec.

Applications of VAD include speech recognition, voice compression, noise estimation/cancellation, and echo cancellation. The accuracy of the VAD [4] has a large impact on the performance of the algorithms that depend on the VAD decisions, thus many approaches have been developed including energy level detection, zero crossing rates, periodicity, LPC distance, spectral energy distribution, timing, pitch, zero crossings, cepstral features and adaptive noise modeling. An important consideration for the VAD algorithm is the processing power required. This paper shows that an initial VAD decision and spectral subtraction can be used to produce more accurate VAD for the purposes of noise reduction. This paper shows that performance of voice activity detection (VAD) on the output of a spectral subtraction noise reduced signal increases the accuracy of the VAD and reduces the VAD sensitivity to fixed thresholds. An initial VAD decision is used to control the noise estimate update in the spectral subtraction algorithm. The more accurate VAD after the first spectral subtraction is then used to

reprocess the original noisy speech again via spectral subtraction to reduce the noise while not attenuating the speech. Auditory masking thresholds were used to weight the spectral subtraction to avoid the introduction of musical noise artifacts. Section I introduction, Section II explains simple approach of wide band speech coding techniques (VMR-WB), Section III explains about the existing noise estimation algorithms, Section IV explains performance of proposed noise estimation algorithms, Section V conclusion.

2. INTRODUCTION TO VMR-WB SPEECH CODEC

VMR-WB has been designed for encoding or decoding wideband speech sampled at 16 kHz. However, VMR-WB also accepts NB input signals sampled at 8 kHz, and it can synthesize NB speech. Similar to AMR-WB, the internal sampling rate of VMR-WB is always 12.8 kHz. Overall, five VMR-WB modes of operation have been standardized. The process of standardizing a variable-rate multimode codec for wideband speech services in 3GPP2, compliant with cdma2000 Rate-Set II. Modes 0, 1, and 2 are Rate-Set II modes specific to CDMA systems with Mode 0 providing the highest quality and Mode 2 the lowest ABR. Mode 3 has been designed in Rate-Set II for direct interoperability with AMR-WB codec at 12.65, 8.85, and 6.60 kb/s. The performance of Mode 3 is slightly superior to the performance of the corresponding modes of AMR-WB, especially for lower rates. However, the enhancements do not affect the bit stream structure, and frames of Mode 3 contain a bit stream fully compatible with AMR-WB. It should be noted that given the limited rates available in CDMA systems, an full rate (FR) frame (13.3 kb/s) is always needed to encode active speech in Mode 3. Mode 4 is the only mode designed for Rate Set I, and its ABR is slightly lower than the ABR of Mode 2 [3].

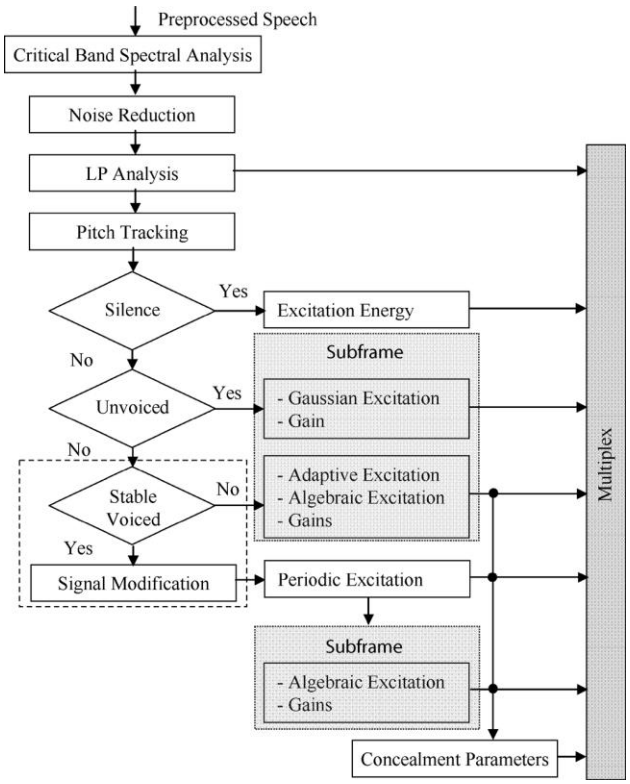

Fig. 1. Flow chart of the VMR-WB Encoder

An optimized signal classification and rate selection mechanism is perhaps the most important part of any variable-rate codec. In VMR-WB, a speech frame can be roughly classified into one of four following speech classes:

• Inactive frames are characterized by the absence of speech activity.

• Unvoiced speech frames are characterized by an aperiodic structure and energy concentration toward higher frequencies.

• Voiced speech frames have a clear quasi periodic nature with energy concentrated mainly in low frequencies.

• Any other frame is classified as a transition having rapidly varying characteristics.

The general flow chart of the VMR-WB encoder is shown in Fig. 1. Following sampling rate conversion, the input signal is preprocessed. A fast Fourier transform (FFT)-based spectral analysis is then performed twice per frame for use in background noise estimation, noise reduction, voice activity detection (VAD), and rate selection algorithms. The signal energy is computed for each perceptual critical band. Linear prediction analysis and open-loop (OL) pitch analysis are performed on a frame basis on the denoised signal. The linear prediction (LP) [3] filter, modeling the human vocal tract, is estimated similar to AMR-WB. The OL pitch analysis determines the period of the fundamental frequency (pitch) of the speech signal given by the vibration of the vocal cords during voiced speech. A new OL pitch-tracker is used in VMR-WB to improve the smoothness of the pitch contour by exploiting adjacent values.

The rate determination for VMR-WB is implicitly performed during the selection of a particular encoding scheme for the current frame. It is dependent on the mode of operation and the class of input speech. For each frame, the signal classification is initially performed to detect inactive frames, then continues with unvoiced, voiced, and transition frame detection (Fig. 1). Inactive frames are encoded using the lowest possible bit rate to roughly capture the characteristics of the background noise. The characteristics are then smoothed over time, and noise is regenerated in the decoder so that the user does not have the impression of interrupted communication during silence intervals. This technique is known as comfort noise generation (CNG). If the frame is classified as active speech by the VAD algorithm, unvoiced signal classification is applied.

Frames not classified as unvoiced are processed by a transparent signal modification algorithm based on the generalized analysis-by- synthesis or relaxation code Excitation Linear Prediction (CELP) paradigm[2]. The VMR-WB signal modification algorithm comprises an inherent classifier of voiced frames. The remaining frames are likely to contain a non-stationary segment such as voiced onset or rapidly evolving voiced speech. These frames typically require a general-purpose coding model at a high bit rate for maintaining good speech quality; FR coding is mainly used. Only frames with very low energy can be encoded using generic HR in order to further reduce the ABR. Following the frame-based processing stage, the frame is divided into four subframes and the signal is encoded on a subframe basis in order to find the adaptive and fixed-codebook indices and gains. The

encoding model used for generic or voiced frames is based on the Algebraic CELP (ACELP) paradigm [3]. Generic and voiced frames are processed similarly with the exception that for voiced frames the model is applied to the modified signal. Unvoiced frames exploit LP synthesis filter excited by a Gaussian noise with appropriately scaled energy. The information transmitted through the communication channel to the decoder comprises all or some of the following parameters like the coding-scheme selection bits, the quantized parameters of the LP synthesis filter, the adaptive and fixed-codebook indices and gains, and the information for improved frame-erasure protection — pitch-synchronous energy, the glottal pulse position, and classification information.

The frame structure of VMR-WB for all encoding types is comprehensively specified in Section 8 of VMR-WB 3GPP2 specification. The rate selection mechanism is then used to determine the bit rate suitable for encoding each speech frame based on the signal classification and the operating mode. The source-coding ABRs are summarized in Table I, measured on active-speech frames only. The table is shown as

3. EXISTING NOISE ESTIMATION TECHNIQUE

There are several classes of noise estimation algorithms like Minimal tracking Algorithms, Time Recursive Algorithms and Histogram based Algorithms[6]. All algorithms operate in the following fashion. First the signal is analyzed using short time spectra computed from short overlapping frames, typically 20-30 msec. Windows with 50% overlap between adjacent frames. Then several consecutive frames called analysis segment are used in the computation of the noise spectrum. Typical time span of this segment may range from 400 msec. to 1 sec. The noise estimation algorithms are based on the assumptions that the analysis segment is too long enough to contain speech pauses and low energy signals segments and the noise present in the analysis segment is more stationary than speech, new assumption is that noise changes at a relatively slower rate than speech. The analysis segment has to be long enough to encompass speech pauses and low energy segments, but it also has to be short enough to track fast changes in the noise level, hence the chosen duration of the analysis segment will result from a track-off between these two restrictions.

Let y(n)=x(n)+d(n), where y(n) is the noisy speech signal, x(n) is the clean signal and d(n) is the additive noise. The smoothed power spectrum of the noisy speech signal can be estimated using a first-order recursive formula as follows:

$$P(\lambda , k) = \eta P(\lambda - 1, k) + (1 - \eta) |Y(\lambda , k)|^2 \quad (1)$$

where $|Y(\lambda , k)|^2$ is an estimate of the short-time power spectrum of y(n) obtained by wavelet-thresholding the multitaper spectrum of y(n) [6], η is a smoothing constant, λ is the frame index and k is the frequency bin index. Since the noisy speech power spectrum in the speech absent frames is equal to the power spectrum of the noise, we can update the estimate of the noise spectrum by tracking the speech-absent frames. To do that, we compute the ratio of the energy of the noisy speech power spectrum in three different frequency bands to the energy of the corresponding frequency band in the previous noise estimate.

3.1 Voice Activity Detection (VAD)

The purpose of Voice Activity Detection (VAD) [3] is to determine whether a frame of the captured signal represents voiced, unvoiced, or silent data. Voice activity detection ideally is aware of the human speech production system, so it can differentiate between silence, unvoiced, and voiced sounds. Voiced sounds are periodic in nature and tend to contain more energy than unvoiced sounds, while unvoiced sounds are more noise-like and have more energy than silence. Silence has the least amount of energy and is a representation of the background noise of the environment. Simple approach to estimate and update the noise spectrum during the silent segments of the signal is using a Voice Activity Detector (VAD). The process of discriminating between the voice activity that is speech presence and silence that is speech absence is called voice activity detection. VAD algorithms typically extract some type of feature (e.g. short time energy, zero crossing etc.) from the input signal and compared against threshold value, usually determined during speech absent period. Generally output of VAD algorithms is binary decision on a frame-by-frame basis having frame duration 20-30 msec. Several VAD algorithms were proposed based on various types of features extracted from the signal. Noise estimation can have major impact on the quality and Intelligibility of speech signal. The early VAD Algorithms decisions were based on energy levels and zero crossing, ceptral features, and the periodicity measures. Some of VAD algorithms are used in (GSM) System, cellular networks, and digital cordless telephone systems. VAD Algorithms are suitable for discontinues transmission in voice communication systems as they can be used to save the battery life of cellular phones.

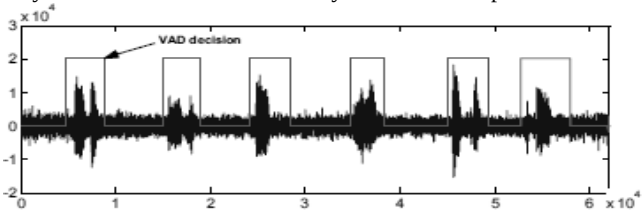

Fig. 2. VAD output result for a sample speech.

4. PROPOSED NOISE ESTIMATION METHOD

Majority of the VAD Algorithms encounter problems in low SNR conditions, particularly when the noise is non-stationary. Having an accurate VAD Algorithm in a non-stationary environment might not be sufficient in speech enhancement. Applications, as on accurate noise estimation is required at all times, even during speech activity. In case of Noise estimation algorithms they continuously track the noise spectrum therefore more suited for speech enhancement applications in non-stationary Scenarios.

The proposed VMR-WB encoder is shown in Fig-2 has similar operations as in existing method here the MCRA algorithm replaces the VAD algorithm. a minima controlled recursive algorithm (MCRA) which updates the noise estimate by tracking the noise-only regions of the noisy speech spectrum. These regions are found by comparing the ratio of the noisy speech to the local minimum against a threshold. The noise estimate, however, lags by at most twice that window length when the noise spectrum increases abruptly. In the improved MCRA approach , a different method was used to track the noise-only

regions of the spectrum based on the estimated speech-presence probability. This probability, however, is also controlled by the minima, and therefore the algorithm incurs roughly the same delay as the MCRA algorithm for increasing noise levels.

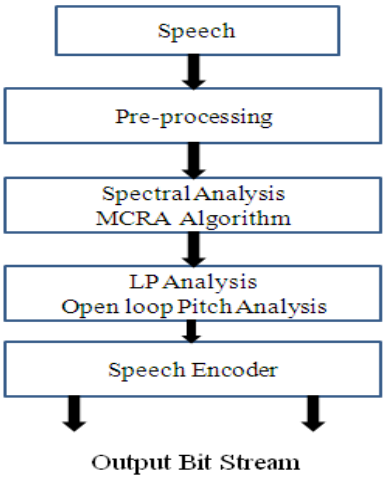

Fig. 2. Flow chart of the Proposed VMR-WB Encoder

4.1 Time Recursive Averaging for Noise Estimation

The time–recursive averaging Algorithms[7] exploit the observation that the noise signal typically has non uniform effect on the spectrum of speech in that some regions of the spectrum will typically have a different effective signal to noise ratio (SNR). As a result, different from bands in the spectrum will have effectively different SNRs. More generally, for any type of noise we can estimate and update individual frequency bands of the noise spectrum whenever the probability of speech being absent at a particular frequency band is high or whenever the effective SNR at a particular frequency band is extremely low. This observation led to the recursive arranging type of algorithms in which noise spectrum is estimated as a weighted average of past noise estimates and the present noisy speech spectrum. The weights change adaptively depending either on the effective SNR of each frequency bin or on the speech present probability.

4.1.1 Minima Controlled Recursive Averaging (MCRA) Algorithm

The minima controlled recursive averaging (MCRA)[6] was introduced for noise estimation. The noise estimate was updated by averaging the past spectral values of noisy speech which was controlled by a time and frequency dependent smoothing factors. These smoothing factors were calculated based on the signal presence probability in each frequency bin separately. This probability was in turn calculated using the ratio of the noisy speech power spectrum to its local minimum calculated over a fixed window time. We show that presence of speech in a given frame of a subband can be determined by the ratio between the local energy of the noisy speech and its minimum within a specified time window. The ratio is compared to a certain threshold value, where a smaller ratio indicates absence of speech. Subsequently, a temporal smoothing is carried out to reduce fluctuations between speech and non speech segments, thereby exploiting the strong correlation of speech presence in neighboring frames.

The resultant noise estimate is computationally efficient, robust with respect to the input SNR and type of underlying additive noise and characterized by the ability to quickly follow abrupt changes in the noise spectrum. According to method explained in [8], the conditional speech presence probability P^(λ, k) is computed by comparing the ratio of the noisy speech power spectrum to its local minimum against a threshold value. The probability estimate P^ (λ, k) and the time smoothing factor α (λ, k) is controlled by the estimate of spectral minimum and due to this reason this algorithm is called as Minima Controlled Recursive Averaging Algorithm (MCRA). This Algorithm is modified by researchers and some of them are MCRA-2 Algorithm explained in [6,7], improved MCRA Algorithm explained in [8]. The MCRA noise estimation algorithm is calculated as

1. Smooth noisy psd S(λ, k) as follows

$$S(\lambda, k) = \alpha_s S(\lambda - 1,k) + (1 - \alpha_s)| Y(\lambda, k)|^2 \quad (2)$$

Where α_s is smoothing constant.

2. Perform minimal tracking on S(λ,k) to obtain Smin (λ, k)

3. Determine P(λ, k) using equation(3)

If S (λ, k) > δ (threshold)

P^ (λ, k) = 1 speech present

else (3)

P^ (λ, k) = 0 speech absent

end.

4. Compute the time-frequency dependent smoothing factor α_d (λ, k) using equation (4) and the smoothed Conditional probability P^(λ, k) from equation (5)

$$\alpha_d(\lambda, k) = \alpha + (1-\alpha)\, p(\lambda, k) \quad (4)$$

$$P^(\lambda, k) = \alpha\, p^(\lambda-1, k) + (1-\alpha p)\, p^(\lambda, k) \quad (5)$$

5. Update the noise psd ζ_{sd}^2 (λ, k) using equation (6)

$$\zeta_d^2(\lambda,k)=\alpha_d(\lambda, k)\ \zeta_d^2(\lambda-1,k) + [1-\alpha_d(\lambda, k)]|Y(\lambda, k)]|^2 \quad (7)$$

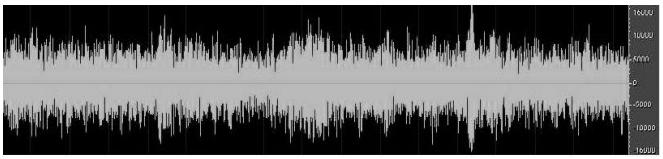

Fig. 3. Noisy Waveform

Fig. 4 Clean Speech Waveform

Novel techniques to enhance the MCRA noise estimation algorithm have been developed for speech enhancement in adverse environments. Our approach is to reduce the time delay for adapting to abrupt noise change while at the same time decreasing the speech leakage to avoid speech distortions. Figure 3 also shows that even with a larger window size, the speech leakage is clearly visible by the MCRA algorithm. The speech leakage is mostly visible only when the noise estimator is applied on clean speech signal with no noise. Because the noise spectrum should be always zero, any detected noise magnitude is erroneously produced from the speech component. In all the tests, an adaptive parametric Wiener filter has been used to perform the noise removal. The SNR indicates the global signal to noise ratio, the larger the better.

5. CONCLUSION

Recursive averaging is a commonly used procedure for estimating the noise power spectrum. However, rather than employing a voice activity detector and the noise estimator to periods of speech absence we adapt the smoothing parameter in time and frequency according to the speech presence probability. The speech presence probability is controlled by the minima values of a smoothed periodogram of the noisy measurement. Compared to a competitive method, the MCRA noise estimate responses more quickly to noise variations and, when integrated into a speech enhancement system, yields higher segmental SNR and a lower level of non-stationary environment. The proposed method of noise estimation method is used in the VMR-WB speech coding technique to establish better speech quality and intelligibility.

6. REFERENCES

[1]. Milan Jelínek, and Redwan Salami, "Wideband Speech Coding Advances in VMR-WB Standard", IEEE Trans. Speech Audio Process., vol. 15, no. 4, pp. 1167–1179, May 2007.

[2]. "AMR Wideband Speech Codec: Transcoding Functions" [Online]. Available: http://www.3gpp.org 3GPP Technical Specification TS 26.190.

[3]. "Source-Controlled Variable-Rate Multimode Wideband Speech Codec (VMR-WB), Service Option 62 for Spread Spectrum Systems" Jul. 2004 [Online]. Available: http://www.3gpp2.org, 3GPP2 Technical Specification C.S0052-0 v1.0

[4]. M. Jelínek and R. Salami, "Noise reduction method for Wideband Speech Coding," in Proc. Eusipco, Vienna, Austria, Sep. 2004, pp. 1959–1962.

[5]. Ningping Fan, Justinian Rosca, Radu Balan, "Speech Noise Estimation Using Enhanced Minima Controlled Recursive Averaging", in ICASSP 2007, no. 4, pp. 581– 584.

[6]. Loizou P. , Sundarajan R. "A Noise estimation Algorithm for highly non–stationary Environments", Speech Communication 48 (2006) Science direct pp. 220-231

[7]. Anuradha R. Fukane and Shashikant L. Sahare. " Noise estimation Algorithms for Speech Enhancement in Highly non-stationary Environments", IJCSI International Journal of Computer Science Issues, Vol.8, Issue 2, March 2011.

[8]. Cohen, I., 2002. "Noise estimation by minima controlled recursive averaging for robust speech enhancement", IEEE Signal Proc. Letter 9 (1), pp.12–15

