# Robust Speech Processing in EW Environment

Akella Amarendra Babu
Progressive Engineering College, Hyderabad, AP, India

Ramadevi Yellasiri
CBIT Osmania University Hyderabad, AP, India

Nagaratna P. Hegde
Vasavi College of Engineering Ibrahimbag, Hyderabad AP, India

## ABSTRACT
Speech communication in Electronic Warfare (EW) environment should be resistant to interception, masquerade and tolerant to communication channel errors.

In this paper, we described an algorithm which provides speech compression, strong encryption, error tolerance and speaker authentication features. This Robust Speech Coder (RSC) is backward compatible with the existing codecs with capability to opt for additional features as and when required.

## General Terms
Speech Compression, Triple DES, Robust Secure Coder (RSC) algorithm, Secure Hash Algorithm SHA-512

## Keywords
Robust Speech Coder (RSC), Electronic Warfare (EW), Authentication algorithm, Mixed Excitation Linear Prediction (MELP)

## 1. INTRODUCTION
Objectives of any speech coding algorithm are low bit-rate, high quality speech and low coding delay.

### 1.1 Low Bit-Rate
Compression saves bandwidth in communication channels or memory space for storing. A Pulse Code Modulated speech sampled at 8 KHz and coded at 16 bits per sample needs a 128 Kbps bit rate per each speech channel whereas MELPe coded speech channel works at 1.2 Kbps bit rate producing savings of 106.6 times.

### 1.2 High Speech Quality
The decoded speech should have quality acceptable for target application. There are many dimensions in quality perception, including intelligibility, naturalness, and pleasantness and speaker recognition.

### 1.3 Low Coding Delay
Coding delay is measured as the time shift between the input speech signal of the encoder with respect to the output speech signal of the decoder. Coding delay above 150 ms will affect the ability to hold a conversation.

Military communications in EW environment have additional requirements. They are encryption, authentication and robustness in the presence of channel errors.

### 1.4 Encryption
In the battlefield, the value of the information is incalculable because the lives of many soldiers depend on battlefield information. Therefore, the users in the battlefield should be given a facility to encrypt as and when need arises.

A block cipher processes the plain text input in fixed size blocks and produces a block of cipher text of equal size for each plain text block. Block symmetric encryption is suitable for use with parametric speech coders because both buffer and process the input data frame by frame.

### 1.5 Authentication
Authentication provides protection against active attacks like masquerade. A masquerade takes place when one person pretends to be a different person. Although human speech can be recognized by the recipient, it is foolproof only when speech is synthesized using high quality speech coders, typically coders like ITU-T G.711 PCM, ITU-T G.726 ADPCM, ETSI GSM 6.10 RPE-LTP etc. Although vocoders like MELP, CELP etc., attempt to retain the naturalness of the synthetic speech, human auditory system cannot authenticate the speaker when subjected to pressures of battlefield disturbances. Therefore, authentication is a security pre-requisite for military communications. Speaker identification algorithms or secure hash functions may be used for authentication.

### 1.6 Robustness in the Presence of Channel Errors
This is crucial for military communications where harsh acoustic disturbances introduce channel errors which will have a negative impact on the speech quality.

## 2. SPEECH COMPRESSION
Mixed Excitation Linear Prediction (MELP) is one of the speech coding algorithms used in military communication equipment. 180 samples of input speech signal are buffered into frames and 2880 bits corresponding to 180 samples of original speech frame are compressed to 54 bits per frame using MELP algorithm, thus achieving a compression ratio of 53.3. Salient features of MELP coder are described in the next section.

### 2.1. MELP [1]
A block diagram to implement **MELP** coding algorithm is shown in Figure 1. A randomly generated period jitter is used to perturb the value of the pitch period so as to generate an aperiodic impulse train. The MELP coder extends the number of classes into three: unvoiced, voiced, and jittery voiced. Jittery voiced state corresponds to the case when the excitation is aperiodic but not completely random, which is often encountered in voicing transitions. This jittery voiced state is controlled in the MELP model by the pitch jitter parameter and is essentially a random number. A period jitter uniformly distributed up to +/- 25% of the pitch period produced good results. The short isolated tones, often encountered in Linear Prediction coded speech due to misclassification of voicing state, are reduced to a minimum.

Shape of the excitation pulse for periodic excitation is extracted from the input speech signal and transmitted as information on the frame. The shape of the pulse contains important information and is captured by the MELP coder through Fourier magnitudes of the prediction error. These quantities are used to generate the impulse response of the pulse generation filter (Figure 1), responsible for the synthesis of periodic excitation.

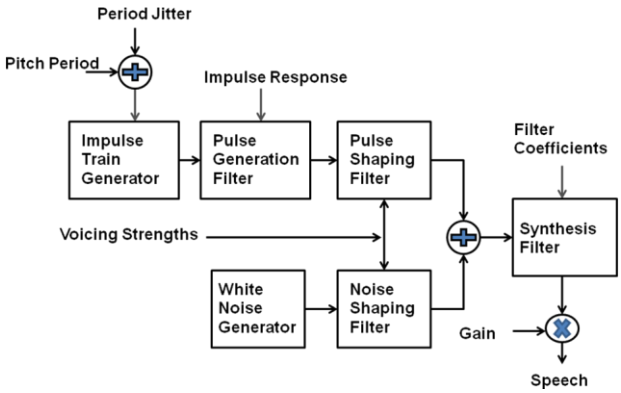


Figure1: The MELP speech production model

Periodic excitation is filtered using pulse shaping filter and noise excitation is filtered using noise shaping filter. Filters' outputs are added together to form the total excitation, known as the mixed excitation, since portions of the noise and pulse train are mixed together.

Voiced strengths are the parameters which is a measure of voicedness in the input signal. Voiced strengths are used to control the frequency responses of the shaping filters. The responses of these filters are variable with time, with their parameters estimated from the input speech signal, and transmitted as information on the frame.

## 2.3. Bit Allocation

The allocation scheme of FS MELP [1] is summarized in Figure 2. A total of 54 bits are transmitted per frame, at a frame length of 22.5 ms. 2.4 kbps bit-rate is required to transmit 54 bits per frame.

| Parameter | Voiced | Unvoiced |
|---|---|---|
| LPC | 25 | 25 |
| Pitch period/low-band voicing strength | 7 | 7 |
| Band pass voicing strength | 4 | 0 |
| First gain | 3 | 3 |
| Second gain | 5 | 5 |
| Aperiodic flag | 1 | 0 |
| Fourier magnitudes | 8 | 0 |
| Synchronization | 1 | 1 |
| Error protection | 0 | 13 |
| Total | 54 | 54 |

Figure 2: Bit allocation for the FS MELP Coder

# 3. ALGORITHMS

## 3.1. Data Encryption Algorithm (DEA)

An encryption scheme computationally secure if the cost of breaking the cipher text generated by the scheme exceeds the value of the encrypted information and the time required to break the cipher exceeds the useful lifetime of the information. In the battlefield, the value of the information is incalculable because the lives of many soldiers depend on this information. However, the useful lifetime of the information is known and the time required to break the cipher can be calculated.

Assuming that there are no inherent mathematical weaknesses in the algorithm, brute-force approach makes reasonable estimates about the time. Brute-force approach involves trying every possible key until intelligible translation of the cipher text into plain text is obtained. Assuming that it takes 1 micro second to perform single decryption, 10.01 hours [3] are required to break a 56-bit key size DES and 5.4 x 10**30 years to break a 168-bit key size DES.

In 1999, Triple DES (3DES) was incorporated as part of the Data Encryption Standard and published as FIPS PUB 46-3. 3DES uses three different keys and three executions of the DES algorithm. 3DES is very resistant to cryptanalysis and makes the system robust.

3DES processes the input data in 64-bit blocks. Using MELP compression algorithm, a frame of input speech of 22.5 ms is encoded using 54 bits. These 54 bits are combined with another 10 bits to ruggedize the coder to make total of 64 bit block. These 10 bits are used for error correction and authentication. 9 bits are utilized for error correction to make the coder more robust in the presence of channel errors and one bit is utilized for authentication. This block of 64 bits is given data input to 3DES encryption.

## 3.2. Authentication

In terms of communication security issues, a masquerade is a type of attack where the attacker pretends to be an authorized user of a system in order to gain access to it or to gain greater privileges than they are authorized for. For example, enemy after gaining access to the net-centric communication links will be on listening mode while tactical operations progress and at a opportune critical time, he takes control of the speech channel connection and uses it for passing operational orders which are favorable to him. Consequently, a battle will be lost due to the above vulnerability. A security alert by the communication system in such situations will save a country from defeat.

Using speech recognition algorithms, the speaker is identified from the original speech input frames. Index to the speaker is obtained from the data store. A secure hashing algorithm SHA-512 is applied on this speaker index and 512 bit message digest is obtained. 512 bits are combined with MELP encoded speech frames at the rate of 1-bit per frame.

At the receiving end, the message digest is recovered. New speaker Id is calculated at the receiving end from the synthetic speech and index to the speaker is obtained from local database at the receiving end. Hashing algorithm is applied on the index and a new message digest is obtained. The new message digest and received message digest are compared. If there is mismatch, the user at the receiving end is alerted.

## 3.3. Error Protection [2]

In 1950, Hamming introduced the (7,4) code. It encodes 4 data bits into 7 bits by adding three parity bits. Hamming(7,4) can detect and correct single-bit errors. With the addition of an overall parity bit, it can also detect (but not correct) double-bit errors.

In MELP algorithm, Forward Error Correction (FEC) is implemented in the unvoiced mode only (Figure 2). The parameters that are not transmitted in the unvoiced mode are the Fourier magnitudes, band pass voicing and the aperiodic flag. FEC replaces these 13 bits with parity bits from three Hamming (7,4) codes and one Hamming (8,4) code. However, no error correction is provided for the voiced mode MELP coder. The DES/3DES encryption algorithms process input data in 64-bit blocks. 54 bits are allocation for MELP encoded speech frame. 1-bit per frame is added for authentication. Remaining 9 bits are utilised for FEC parity bits for voiced mode from three Hamming (7, 4) codes.

## 4. ROBUST SPEECH CODER (RSC) ALGORITHM

Figure 3 gives the block diagram of RSC voice coder with 3DES encryption scheme, SHA-512 authentication and 9-bit FEC incorporated in its algorithm.

**Step 1: Data Compression**
The original speech is buffered into 22.5 ms frames and passed through MELP coding filter. The 22.5 ms frame coded into 54 bits compressed speech frame.

**Step 2: Authentication**
The original speech corresponding to 512 frames is buffered and using speech recognition algorithm, the speaker is identified. All the authorised speakers' names are recorded in the local database. A replica of this database is loaded at all destination receiving stations. All the names are indexed. The index number corresponding to the speaker is retrieved from the local database. The index is hashed using SHA-512 and resulting 512-bit Message Digest (MD) is buffered and 1-bit per frame is added to 54-bit compressed speech frame.

In case an unauthorized person speaks, the database will return a special code corresponding unknown speaker and the destination recipient will get alert.

One bit per frame (22.5 ms) is added to the 54-bit compressed speech frame. In case SHA-1 hashing algorithm is used, the size of the message digest is 64-bits length and it takes 22.5 x 64 = 1440 ms. (1.5 seconds approximately) to buffer the original speech. Therefore the speaker is authenticated every 1.5 seconds. In case stronger secure hashing algorithm like SHA 512 is used, the authentication period would be 12 seconds.

**Step 3: Forward Error Correction**
In MELP algorithm, Forward Error Correction (FEC) is implemented in the unvoiced mode only. RSC algorithm uses 9 parity bits to provide error correction. It uses one Hamming (31,26) code and one Hamming (15,11) code. LPC parameters are coded with 25 bits (Refer Figure 2 above). Hamming (31,26) code is applied to 25 bits of LPC parameter bits and one MSB bit of band pass voicing parameter. Hamming (15,11) code is applied to 5 bits of second gain parameter, 3 bits of first gain parameter and three LSB bits of band pass voicing parameter. Total 2 bits are corrected over 39 bits of data which cover four parameters, that is, LPC, first gain, second gain and band pass voicing parameter. Thus 9 parity bits are used to correct 2 errors over 39 bits out of 55 bits. These 39 bits are covering most critical parameters in the RSC algorithm.

**Step 4: Encryption**
54 bits of compressed speech, 9 bits of forward error correction and 1 authentication bit are buffered into a 64 bits compressed speech frame. This block of 64 bits is encrypted with 3DES and resulting 64 bit encrypted compressed speech is transmitted to receiver end.

**Step 5: Decryption**
The 64 bits of encrypted speech is input to 3DES decryption. The resulting 64-bit decoded speech is passed to the next stage.

**Step 6: Application of FEC and Reassembly of MD**
9 parity bits are used to correct errors, if any. 54 bits of compressed speech is separated and given to MELP Decoding filter. 1 authentication bit per frame is buffered and the original 512 bits of original message digest (MD) is reassembled. The original MD is used to compare with the new MD which calculated from synthesized speech.

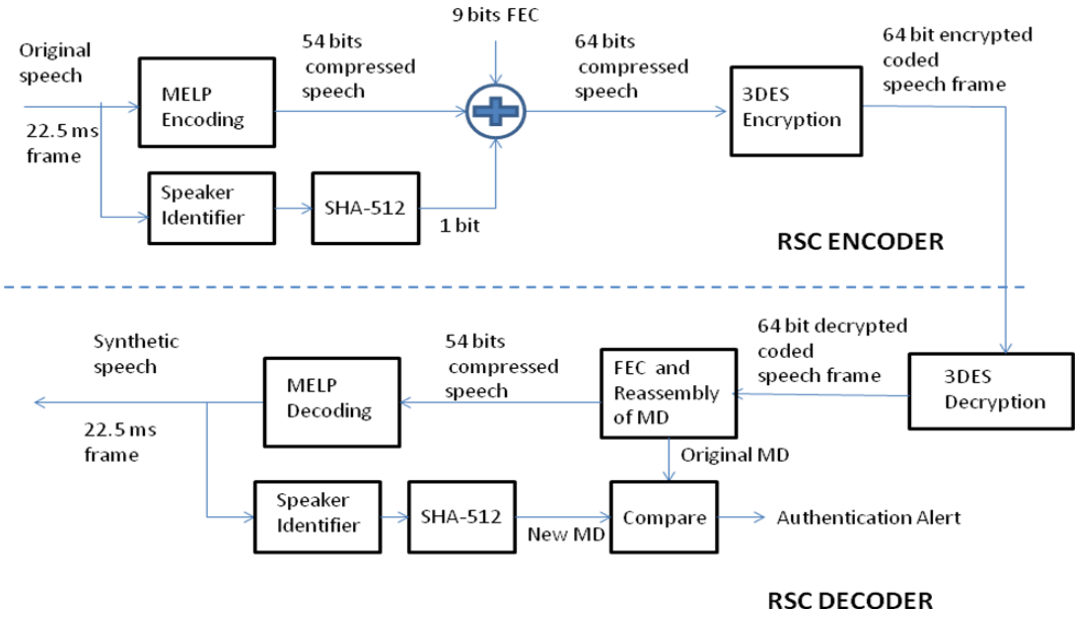


**Fig 3: RSC voice processor**

**Step 6: Application of FEC and Reassembly of MD**
9 parity bits are used to correct errors, if any. 54 bits of compressed speech is separated and given to MELP Decoding filter. 1 authentication bit per frame is buffered and the original 512 bits of original message digest (MD) is reassembled. The original MD is used to compare with the new MD which calculated from synthesized speech.

**Step 7: Speech Synthesis**
54 bits of compressed speech frame is passed through the MELP Decoder which produces the synthesized speech frame of 22.5 ms.

**Step 8: Calculation New MD and Alert Generation**
512 frames of synthetic speech is buffered and given as input to speaker identifier. The index of the speaker is retrieved from the local database. SHA 512 hashing is done on the index and new MD is produced. New is compared with the original MD reassembled from the received speech frames. In case both the MDs are the same, then there is no masquerade. In case they don't match, then an alert is given the receiver that there is a change in the speaker at the sending end.

## 5. DISTINCTIVE FEATURES

The speech coding algorithm described in this paper has additional features besides the common features of all other coding algorithms. The common features of any coding algorithm are low-bit rate, high quality of speech and low processing delay. The special features which additionally added are encryption to ensure confidentiality, authentication to guard against masquerade and error tolerance to enable the speech codec operate under harsh acoustic environment in battlefield.

The existing MELP compression technique was adopted as Federal Standard for military applications in 2002. However, it has some limitations from Information Security (IS) stand point. It is vulnerable to enemy Electronic Warfare attacks like interception, masquerade. Also, the error correction capabilities available in this algorithm are limited to unvoiced speech segments.

The special capabilities of the RSC algorithm described in the paper make the speech codec robust. The additional capabilities are available to the user with single press of the button. The algorithm is backward compatible with the existing MELP algorithm when the additional features are not used.

## 6. CONCLUSIONS

RSC voice processor uses 64-bit allocation scheme for 22.5 ms frame which would get translated to 2844 bps bit-rate. However, the RSC voice processor is interoperable with existing MELP based communication systems in non-encryption mode. The secure mode can be optionally switched over at the extra cost of 444 bps.

The encryption algorithm will introduce very small delay in processing time. Advances in microelectronics and the vast availability of low cost programmable processors and dedicated chips have enabled rapid technology transfer to product development. Assuming one microsecond for encryption / decryption, the delay added to the processing time is negligible and overall delay would be less than acceptable 150 ms from speaker to receiver and the conversation will not be impaired after switching to encryption mode.

Net-centric communications are accessed by large number of users and therefore, there is a need to provide protection against security attacks and suitable security systems should be introduced to match with the speed of migration to net-centric communications.

In this paper, we described RSC algorithm which provides speech compression, encryption, authentication against masquerade and forward error correction in the presence of errors produced by harsh acoustic noise disturbances produced in the battlefield. RSC algorithm described in this paper is suitable for secure net-centric communications in the battlefield and is a robust and secure voice processor with explicit encryption and error correction features.

## 7. ACKNOWLEDGMENTS

Our thanks to all who helped us to prepare this paper.

## 8. REFERENCES

[1] Wai C. Chu, 2003, Speech Coding Algorithms, Wiley Interscience.

[2] S. Collura, Diane F. Brandt, Douglas J. Rahikka, 2002, The 1.2Kbps/2.4Kbps MELP Speech Coding Suite with Integrated Noise Pre-Processing, National Security Agency.

[3] William Stallings, 2009, Network Security Essentials Applications and Standards, Pearson Education

[4] Lann M Supplee, Ronald P Cohn, John S. Collura, Alan V McCree, 2002, MELP: The New Federal Standard at 2400 bps

[5] Arundhati S. Mehendale and M. R. Dixit, 2011, Speaker Identification, Signal & Image Processing: An International Journal (SIPIJ) Vol.2, No.2, June 2011

[6] Jelena NIKOLIC, Zoran PERIC, 2008, Lloyd–Max's Algorithm Implementation in Speech Coding Algorithm Based on Forward Adaptive Technique, INFORMATICA, 2008, Vol. 19, No. 2, 255–270

[7] Chetana Prakash, Dhananjaya N., and S. V. Gangashetty, 2011 "Detection of Glottal Closure Instants from Bessel Features using AM-M Signal," IWSSP 2011, pp 143-146

[8] Sri Harish Reddy M., Kishore Prahallad, S. V. Gangashetty, and B. Yagnanaryana, 2011, "Significance of Pitch Synchronous Analysis for Speaker Recognition using AANN Models", Eleventh INTERSPEECH 2010, pp 669-672

[9] A. Amarendra Babu, Suryakant V. Gangashetty, 2011, Algorithms for Speech Coders in Military Communications. In the proceedings of 3rd International Conference on Science Engineering and Technology (SET), 17- 18 Nov 2011,Vol 6 pp 754 – 760 at VIT, Vellore.

[10] A. Amarendra Babu, B. Sravan Kumar, K. K. Basheer, 2011, Algorithms for Secure Speech Coding in NetCentric Operations Environment. In the proceedings of 3rd International Conference on Science Engineering and Technology (SET), 17- 18 Nov 2011, Vol 4 pp 1203 – 1208 at VIT, Vellore.

[11] B. Yegnanarayana, R. Kumara Swamy and K. S. R. Murty, 2009 Determining mixing parameters from multispeaker data using speech-specific information,

IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, no. 6, pp. 1196-1207,Aug. 2009.

[12] K. S. R. Murty, B. Yegnanarayana and M. Anand Joseph, 2009, Characterization of glottal activity from speech signals, IEEE Signal Processing Letters, vol. 16, no. 6, pp. 469-472, June 2009.

[13] B. Yegnanarayana and K. S. R. Murty, 2009, Event-based instantaneous fundamental frequency estimation from speech signals, IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, no. 4, pp. 614-624, May 2009.

[14] Srinivas Desai, E. Veera Raghavendra, B. Yegnanarayana, Alan W. Black and S. Kishore Prahallad, 2009, Voice conversion using artificial neural networks, Proc. International Conference on Acoustics, Speech, and Signal Processing, 2009, Taipei, Taiwan, pp. 3893- 3896, Aptil 19-24, 2009.

[15] A. K. Sao and B. Yegnanarayana, 2009, Analytic phase-based representation for face recognition, Proc. International Conference on Advances in Pattern Recognition, 2009 Kolkata, India, pp. 453-456, Feb. 04-06, 2009.

[16] Anvita Bajpai and B. Yegnanarayana, 2008, Combining evidence from subsegmental and segmental features for audio clip classification, Proc. IEEE Region 10 Conference (TENCON) 2008, Hyderabad, India, pp. 1-5, Nov. 19-21, 2008.

[17] B. Yegnanarayana, S. Rajendran, Hussien Seid Worku and N. Dhananjaya, 2008, Analysis of glottal stops in speech signals, Proc. INTERSPEECH 2008, Brisbane, Australia, pp. 1481-1484, Sep. 22-26, 2008.

[18] N. Dhananjaya, S. Rajendran and B. Yegnanarayana, 2008, Features for automatic detection of voice bars in continuous speech, Proc. INTERSPEECH 2008, Brisbane, Australia, pp. 1321-1324, Sep. 22-26, 2008.

[19] E. Veera Raghavendra, B. Yegnanarayana, Alan W. Black and S. Kishore Prahallad, 2008, Building sleek synthesizers for multi-lingual screen reader, Proc. INTER-SPEECH 2008, Brisbane, Australia, pp. 1865-1868, Sep. 22-26, 2008.

[20] C. Krishna Mohan, N. Dhananjaya, B. Yegnanarayana, 2008, Video shot segmentation using late fusion technique, Proc. Seventh International Conference on Machine Learning and Applications, 2008, San Diego, California, USA, pp. 267-270, Dec. 11-13, 2008.