Performance Investigation of Joint Source-Channel Coding of Speech for Persian Dialectic, using MELP Vocoder and MAP Decoding

Ahmad Kasaeyan[§], Mohammad Hossein Moghaddam[§] and Ghasem Assarzadeh[‡]
[§]Department of Electrical Engineering, K.N. Toosi University of Technology, Tehran, Iran
Email: {ahmadkasaeyan@ee.kntu.ac.ir mhmoghadam@ee.kntu.ac.ir}
[‡]School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran.
Email: {gh.assar@ut.ac.ir}

***Abstract*—Mixed-excitation linear prediction (MELP) is a United States department of defense speech coding standard used mainly in military applications and satellite communications, secure voice, and secure radio devices. MELP vocoder could provide an acceptable voice quality over prone erroneous channels in various environments by exploiting redundancy statistics of voice data. This paper explains performance investigation of the MELP vocoding algorithm in conjunction with convolutional codes, and maximum a posteriori techniques in both hard and soft decoding regimes which utilize speech redundancy statistics of Persian language. We have evaluated our results according to weighted spectral distortion (WSD) and ITU criterion, perceptual evaluation of speech quality (PESQ).**

***Keywords*—Mixed-excitation linear prediction, redundancy statistics, convolutional codes, maximum a posteriori algorithm.**

I. Introduction

One fundamental trade-off in telecommunication has always existed between high data rate and high error protection. This fundamental trade-off deals with source coding (compression) and channel coding (error control) in separate or joint scenarios. According to Shannon's separation principle [1], the performance achievable via jointly designed sourcechannel code, can also be achieved with a source code designed separately with regard to the source description and a channel code designed solely with regard to the channel description. However, the separation principle could be utmost beneficial when there is no concern about unlimited complexity and delay in the encoding and decoding operations. In another words, for a fixed degree of delay or complexity, one could achieve a jointly designed source channel code which could outperform the best separately designed pair, and this fact made such an enthusiasm among scholars to work on joint source-channel coding designs during past half a century.

Joint source-channel coding algorithms could generally considered in two main categories: source-centric and channel centric. In source-centric algorithms such as channel-optimized vector quantizers [2], [3], source codes are designed to be robust in case of channel errors. On the other hand, in channel-centric algorithms [4], [5], channel decoders are designed in a way to exploit the known characteristics of the source code which could leads to a better error resilience in prone channels. According to the structure of channel-centric algorithms, these algorithms are dependent to source statistical characteristics, and for each type of these statistical characteristics, the decoder should be amended to exploit each individual characteristic for each type of source. This paper could be categorized as the latter (channel-centric) approach, with the speech, considered as source type.

There are several speech coding algorithms in the literature and over the past three decades, a series of vocoding algorithms have been developed specifically for application in U.S. government communications equipments with strict interoperability requirements [6]. These include FS1015 linear predictive coding (LPC) at 2400 bps, FS1016 code excited linear prediction (CELP) at 4800 bps, and mixed-excitation linear prediction (MELP) vocoder at 2400 bps. Among them, MELP could significantly surpass other standards in case of intelligibility, voice quality, talker recognizability, and communicability. After a multi-year extensive testing program, in March 1996, the US governments digital voice processing consortium (DDVPC) selected the 2400 bps MELP speech-coding algorithm to be the standard for narrow band secure voice coding products and applications. The algorithm then expanded by Microsoft Corp. and AT&T in 2002 to rate 1200 bps [7]. Nowadays, MELP standard could be found in various military, civil and multimedia communication devices with different rates of 600/800/1200/2400 bps [8]–[11].

As it stated above, channel-centric algorithms deal with statistical characteristics of the source which needed to be extracted from a generic source data according to different source types via some training steps and in case of speech, training results could be different for each language. In this paper, the results of experiments on using MELP for Persian language are presented. The statistical characteristics of Persian language are extracted through a training phase (which could be made just one time) and then exploited through maximum a posteriori (MAP) decoding of convolutional coded speech data for both MELP 1200 and 2400 bps in both soft and hard decoding regimes. The results are evaluated by weighted spectral distortion (WSP) [12] and ITU perceptual evaluation of speech quality (PESQ) measure [13], as standard criterions.

The remainder of this paper is organized as follows: Section II makes a review on joint source-channel coding, explaining MELP algorithm and MAP decoding. In section III, the details

of conducted experiments and simulation results are explained, and performance of different algorithms are compared to each other, and finally Section IV makes some conclusions based on conducted experiments.

II. Joint Source-Channel Coding/Decoding

A. MELP

MELP is a frame-oriented parametric voice coding algorithm with Each frame consists of 54 bits and represents 22.5 ms of speech. The MELP parameters include Pitch, Gain, a 4 stage MSVQ (Multi Stage Vector Quantizer) which characterizes the LPC coefficient line spectral frequencies (LSFs), fourier magnitude (FM), bandpass voicing (BP), Aperiodic Flag (AF), and a frame Synchronization bit as illustrated in Table 1 [6]. In critical situations where channel capacity is limited, for instance situations in which only half of the MELP bits (27 of 54 bits per frame) could be protected, one would only protect most significant frame bits. Fortunately, due to statistical characteristics of speech data exploited by MELP, performance results for this condition approach that of full protection, indicating the considerably varying importance of parameter bits even for such a low rate vocoding algorithm [6].

According to the nature of human speech signal, and structure of MELP, the MELP bitstream could not be considered equiprobable, memoryless, and uniformly distributed which could help the designer to exploit the residual redundancies of its parameters for designing efficient compression and decoding algorithms. MELP interprets short segments of speech as the output of a linear filter with an appropriate excitation signal. In the transmitter side, the encoder is programmed to design the proper filter and select the excitation signal and then represent both with a frame of binary data. In the receiver side, the decoder, use the encoded description to synthesize the filter and apply the excitation signal, thereby generating the transmitted speech segment. In this way, it is required that the residual redundancy statistics of source data, be extracted from a training sequence to form the transition probabilities which then could be used in decoder. For doing this, we have selected 3 high-order bits of 4 MELP parameters pitch, gain 1, gain 2, and MVSQ 1 according to table 1, and maintained the training phase for about one hour (16000 frames) training sequence of persian speech to extract the transition probabilities. After training phase, we could create a first-order 8-state Markov model and a 8*8 conditional distribution matrix for each of those 4 parameters.

To find the residual redundancies of MELP parameters, we need to estimate the entropy rate of each parameter. Consider a stationary discrete-time random process X with $X\{X_i : i = 1, 2, ...\}$. The entropy rate of this process is defined as [4]:

$$H_X = \lim_{n\to\infty} H(X_n|\, X_1, ..., X_{n-1}) \tag{1}$$

with

$$\begin{aligned} H(X_n|\, X_1, ..., X_{n-1}) = \\ - \sum_{[x_1, ..., x_n] \in X^n} p(X_1 = x_1, ..., X_n = x_n) \\ \times log_2[p(X_n = x_n|\, X_1 = x_1, ..., X_{n-1} = x_{n-1})] \end{aligned} \tag{2}$$

TABLE I. MELP parameters

MELP Parameter	Number of Bits	MSBs in critical situation (27 of 57)
Pitch	7	7
Gain 1	5	4
Gain 2	3	-
MSVQ 1	7	7
MSVQ 2	6	4
MSVQ 3	6	1
MSVQ 4	6	1
FM Fourier	8	-
BP Bandpass	4	1
AF Aperiodic	1	1
Sync Bit	1	1

where H_X represents the minimum rate at which the process $\{X_i\}$ can be encoded without distortion. On the other hand, if the process is encoded at a rate R bits/letter, then the quantity $\rho = R - H_x$ is the residual redundancy incurred by the encoding [4]. According to equations (1) and (2), to estimate the residual redundancy of MELP parameters we could model MELP parameters with stationary first-order Markov chain, i.e. for second term of equation (2) we could say:

$$\begin{aligned} p(X_n = x_n|\, X_1 = x_1, ..., X_{n-1} = x_{n-1}) = \\ p(X_n = x_n|\, X_{n-1} = x_{n-1}) \end{aligned} \tag{3}$$

B. Convolutional coding & using transition probabilities in MAP decoding

After processing the speech data stream with MELP procedure, the MELP data stream is transmitted over a noisy channel with a rate k/n convolutional code which in case of using 3 high-order bits in MELP parameters, k would be 3 and for implementing a rate 0.5 convolutional code we should choose $n = 6$ in our system model.According to convolutional coding, each transition through the trellis corresponds to three data bits and six channel bits. There are eight outgoing edges from each state with regards to 8-state Markov model of 3 bit data, and each edge corresponds to a sequence of three edges in the usual binary implementation which is called a super-trellis since contains three one-bit state trellises.

The resulted binary sequence is modulated by BPSK as $X = [X_0, X_1, ...]$ with $X_k \in \{-\sqrt{E_s}, +\sqrt{E_s}\}$ and after passing through AWGN channel the received signal is $Y = X + Z$, where Z is a sequence of i.i.d. Gaussian random variables with mean zero and variance $N_0/2$. The received signal is demodulated and decoded by Viterbi decoder with Maximum A Posteriori algorithm using those parameters containing significant residual redundancy. Through using MAP metric we have used a priori transition probabilities according to 3 higher-order bits of pitch, gain 1, gain 2, and MVSQ1 in MELP data stream.

Each MELP frame contains $54/3 = 18$ 3-bit MELP words with sequence of words for each frame as $U = [U_0, U_1, ..., U_{17}]$. By considering first four words in data stream as words with most significant residual redundancy, we could state that for

$$p(U_i = u_i | U_{i-18} = u_{i-18}) \quad for \ (i \bmod 18) = 0, 1, 2, 3 \tag{4}$$

distributions on U_i could be described by the stationary distribution of the appropriate Markov chain and for others, U_i could be considered equiprobable:

$$\begin{aligned} p(U_i = u_i | U_{i-18} = u_{i-18}) = 1/8 \\ for \ (i \bmod 18) \neq 0, 1, 2, 3 \end{aligned} \tag{5}$$

Now, Assume that the sequence of MELP words is coded by rate $3/6$ convolutional code and modulated by BPSK. Each word U_i after coding and modulation could be represented as $X_i = [X_{6i}, X_{6i+1}, ..., X_{6i+5}]$ and the whole signal stream could be represented as $X = [X_0, ..., X_{6N-1}]$. By considering the corresponding received sequence $Y = X + Z$, the MAP select the transmitted sequence as

$$\begin{aligned} \hat{U} &= \arg\max\{p(U = u | Y = y) : \\ &\quad u = [u_0, u_1, ..., u_{N-1}], u_i \in \{0, 1\}^3\} \\ &= \arg\max\{p_Y(y|X = x).p(U = u) : \\ &\quad u = [u_0, u_1, ..., u_{N-1}], u_i \in \{0, 1\}^3\} \end{aligned} \tag{6}$$

where in memoryless AWGN channel we have:

$$\begin{aligned} p_Y(y|X = x) &= \prod_{i=0}^{N-1} \prod_{j=0}^{5} p(y_{6i+j} | x_{6i+j}) \\ &= \prod_{i=0}^{N-1} \prod_{j=0}^{5} \frac{1}{\sqrt{\pi N_0}} \exp[\frac{-(y_{6i+j} - x_{6i+j})^2}{N_0}] \end{aligned} \tag{7}$$

and according to Markov structure of U we have:

$$p(U = u) = \prod_{i=0}^{N-1} P_A(U_i = u_i | U_{i-18} = u_{i-18}) \tag{8}$$

in which $P_A(U_i = u_i | U_{i-18} = u_{i-18})$ is the a priori probability derived from the training sequence.

Now, according to equations (6) to (8), we have:

$$\begin{aligned} \hat{U} = \arg\max \Bigg\{ &\frac{1}{(\pi N_0)^{\frac{N}{3}}} \prod_{i=0}^{N-1} \prod_{j=0}^{5} \exp\left[\frac{-(y_{6i+j} - x_{6i+j})^2}{N_0}\right] \\ &.P_A(U_i = u_i | U_{i-18} = u_{i-18})\} \end{aligned} \tag{9}$$

After taking logarithm from equation (9), The MAP metric used for soft-decoding by considering euclidian distance could be obtained as:

$$\begin{aligned} &\sum_{i=0}^{N-1} \left[\sum_{j=0}^{5} |y_{6i+j} - x_{6i+j}| \right]^2 \\ &\quad -N_0 \log P_A(U_i = u_i | U_{i-18} = u_{i-18}) \end{aligned} \tag{10}$$

and for hard-decoding, by using Hamming distance d_H the MAP metric could be obtained as:

$$\begin{aligned} &\sum_{i=0}^{N-1} \left[\sum_{j=0}^{5} d_H(y_{6i+j}, x_{6i+j}) \right] \\ &\quad -N_0 \log P_A(U_i = u_i | U_{i-18} = u_{i-18}) \end{aligned} \tag{11}$$

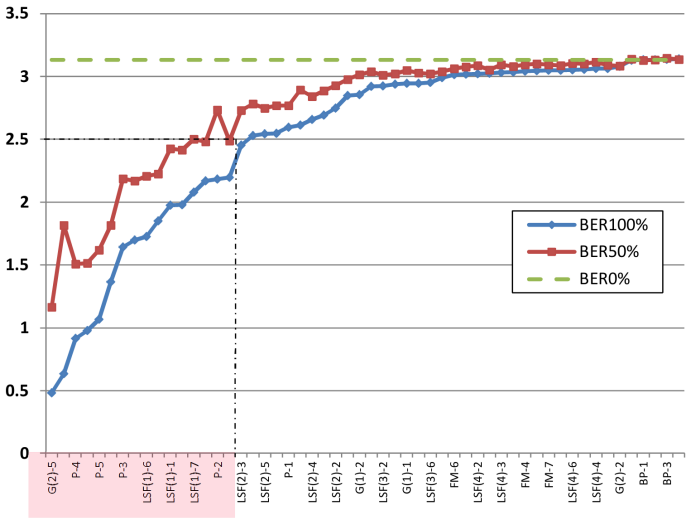

Fig. 1. PESQ metric for MELP 2400

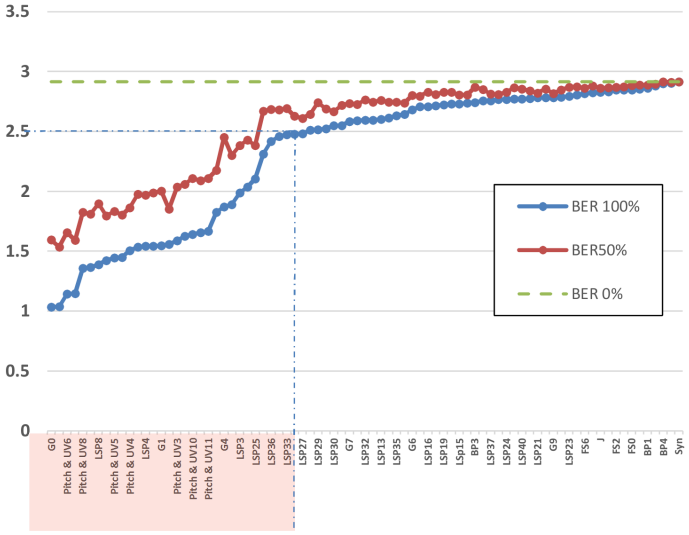

Fig. 2. PESQ metric for MELP 1200

Maximizing these metrics over all valid input sequences corresponds to an enhanced MAP algorithm which uses a priori information for decoding.

III. Numerical Experiments

We have used two quality metrics for performance evaluation, Frequency-Weighted Spectral Distortion (WSD) proposed by McCree [12], and perceptual evaluation of speech quality (PESQ)recommended by ITU [13]. For WSD metric, the spectral distortion (SD) associated with T frames of speech is given by [12]

$$\frac{1}{T} \sum_{j=1}^{T} \left[\int_{-\pi}^{\pi} \frac{|A_B(\omega)|^2}{A_0} \left(10\log_{10} \frac{|S_j(\omega)|}{|\hat{S}_j(\omega)|} \frac{d\omega}{2\pi} \right) \right]^{\frac{1}{2}} \tag{12}$$

where $S_j(\omega)$ and $\hat{S}_j(\omega)$ are the original unquantized spectra and reconstructed spectra associated with frame j and $A_B(\omega)$

TABLE II. BIT PRIORITIES FOR MELP 2400

Group Index	Bit Index	Parameters	Modeled Subjective Evaluation
1	2,39;16,20,29,45,46,48; 30,33-35,49-52;54	BP;LSF4;FM/FEC;SYNC	No Audible Degradation
2	1;4,24,28,32,40;5,8,12,4 1,43,44;18;25,38;37	Gain2;LSF2;LSF3;LSF1; BP;Gain1	Moderate
3	3,14;9;19,22,23,26,31;3 6,53;42	Pitch;Gain2;LSF1;Gain1; LSF2	Poor
4	6,7,10;11;13,15,17,21;2 7;47	Gain2;Pitch;LSF1;AF	Intolerable

TABLE III. BIT PRIORITIES FOR MELP 1200

Group Index	Bit Index	Parameters	Modeled Subjective Evaluation
1	1;67-72;74-80;73	SYNC;BP;[FS0…FS6];jitt er	No Audible Degradation
2	62-66;23-56	[G5…G9];LSP(9…42)	Moderate
3	14,15,17,18,20,21;81;58 -61;3,4,5,12,13	LSP(0,1,3,4,6,7); FS7;[G1…G4];P(1,2,3,10, 11)	Poor
4	2;6…11;16;19;22;57	[P0,P4…P9];LSP2;LSP5; LSP8;G0	Intolerable

is called the Bark weighting [12] and is equal to

$$A_B(\omega) = \frac{1}{25 + 75(1 + 1.4(\frac{\omega}{2000\pi})^2)^{0.69}} \qquad (13)$$

$A_0 = 196.725$ is a normalization factor. PESQ is a family of standards comprising a test methodology for automated assessment of the speech quality as experienced by a user of a telephony system and is standardised as ITU-T recommendation P.862. PESQ was particularly developed to model subjective tests commonly used in telecommunications (e.g. ITU-T P.800) to assess the voice quality by human beings. PESQ is a worldwide applied industry standard for objective voice quality testing used by phone manufacturers, network equipment vendors and telecom operators and its usage requires a license. The details of using PESQ for speech quality measurement could be found in reference [13].

We have made our experiments on MELP 2400&1200 bps and rate 0.5 convolutional code with enhanced MAP decoding using residual redundancy. For training phase, a sequence of speech in persian language with a duration of about one hour

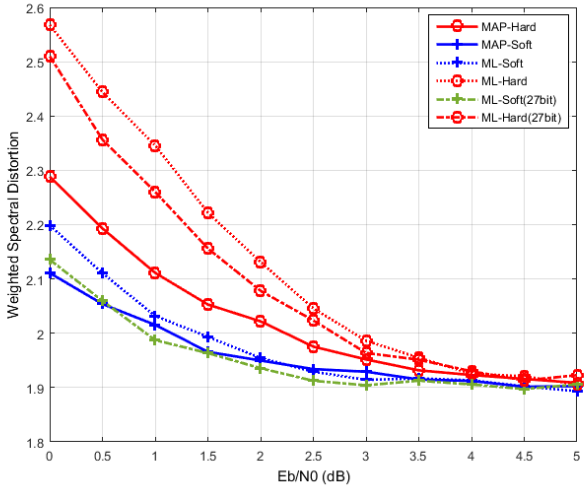

Fig. 3. WSD metric for MELP 2400

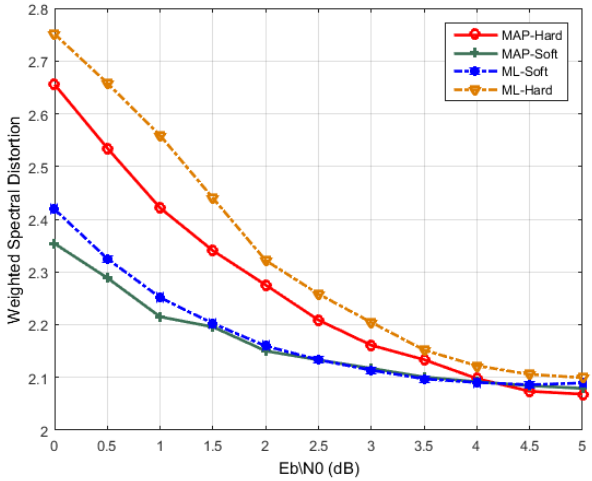

Fig. 4. WSD metric for MELP 1200

with 16000 frames is used to extract the residual redundancy of three bits with higher-order significance for 4 aforementioned MELP parameters. Figure 1 and 2 illustrate the PESQ metric for MELP 2400 and 1200 bps. PESQ experiments are conducted for two different scenarios with BER 50% and BER 100%. In BER 100% experiment, to model a 100% noisy channel each time one bit is reversed (not) and the others are kept unchanged [13]. In BER 50%,each time, some selected bits are exchanged with a random bit sequence [13]. PESQ level 2.5 is known for acceptable speech quality and according to this criterion, the priority of different bits are categorized in table 2 and 3 for MELP 2400 and 1200 bps.

Figure 3 and 4, depicted the WSD metric for MELP 2400 and 1200 bps. In figure 3, performance of MAP and ML decoder are compared to each other in hard and soft regimes for MELP 2400 bps. two additional curves for critical situation of MELP 2400 according to table 1 are depicted for comparison. In Figure 4, performance of ML and MAP decoding algorithms

are compared to each other for soft and hard regimes. From both figures it is obvious that soft decoding algorithms have better performance in comparison with hard algorithms. It could be inferred from Figure 4 that, the performance of MAP and ML algorithms are very close together which is logical according to smaller amount of residual redundancy in MELP 1200 bps.

IV. Conclusion

In this paper, we have investigated the performance of MELP 2400&1200 on Persian language using convolutional codes and MAP decoding for both Hard and soft regimes. The structure of MELP and its contribution to MAP decoding is explained and the role of residual redundancy in speech reconstruction is evaluated according to WSD and PESQ metrics. Simulation results express that the MAP decoding in soft regim by the aid of residual redundancy extracted from MELP data sequence as a joint source-channel scenario could outperform other separate source and channel scenarios in hard and soft regimes.

References

[1] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, no. 3-4, pp. 376–423, 1948.

[2] H. Jafarkhani and N. Farvardin, "Design of channel-optimized vector quantizers in the presence of channel mismatch," *Communications, IEEE Transactions on*, vol. 48, no. 1, pp. 118–124, 2000.

[3] F. Lahouti, A. K. Khandani, and A. Saleh, "Robust transmission of multistage vector quantized sources over noisy communication channelsapplications to melp speech codec," *Vehicular Technology, IEEE Transactions on*, vol. 55, no. 6, pp. 1805–1811, 2006.

[4] T. Fazel and T. Fuja, "Robust transmission of melp-compressed speech: An illustrative example of joint source-channel decoding," *Communications, IEEE Transactions on*, vol. 51, no. 6, pp. 973–982, 2003.

[5] C. Demiroglu, S. D. Kamath, and D. V. Anderson, "Segmentation-based speech enhancement for intelligibility improvement in melp coders using auxiliary sensors." in *ICASSP (1)*, 2005, pp. 797–800.

[6] D. J. Rahikka, J. S. Collura, T. E. Fuja, and T. Faze, "Us federal standard melp vocoder tactical performance enhancement via map error correction," in *Military Communications Conference Proceedings, 1999. MILCOM 1999. IEEE*, vol. 2. IEEE, 1999, pp. 1458–1462.

[7] T. Wang, K. Koishida, V. Cuperman, A. Gersho, and J. S. Collura, "A 1200/2400 bps coding suite based on melp," in *Speech Coding, 2002, IEEE Workshop Proceedings.* IEEE, 2002, pp. 90–92.

[8] M. W. Chamberlain, "Vocoder and associated method that transcodes between mixed excitation linear prediction (melp) vocoders with different speech frame rates," Nov. 19 2013, uS Patent 8,589,151.

[9] F. S. Best, D. A. McClintock, W. J. Lee, W. R. Hartwell, and E. Reed, "Four frequency band single gsm antenna," Oct. 2 2012, uS Patent 8,280,466.

[10] D. Heide, A. E. Cohen, Y. T. Lee, T. M. Moran *et al.*, "Variable data rate vocoder improvements for secure interoperable dod voice communication," in *Military Communications Conference, MILCOM 2013-2013 IEEE.* IEEE, 2013, pp. 702–707.

[11] F. S. Best, D. A. McClintock, W. J. Lee, W. R. Hartwell, and E. Reed, "Secure transmission over satellite phone network," Dec. 16 2014, uS Patent 8,913,989.

[12] A. McCree, K. Truong, E. B. George, T. P. Barnwell, and V. Viswanathan, "A 2.4 kbit/s melp coder candidate for the new us federal standard," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 1. IEEE, 1996, pp. 200–203.

[13] ITU. (2013) perceptual evaluation of speech quality (PESQ). [Online]. Available: http://www.itu.int/rec/T-REC-P.862