ON VARIABLE RATE FRAME INDEPENDENT PREDICTIVE SPEECH CODING: RE-ENGINEERING ILBC

Christopher M. Garrido[†]*, Manohar N. Murthi*[†]*, and Søren Vang Andersen*[‡]

[†] Dept. of Electrical and Computer Engineering
University of Miami, USA
c.garrido@umiami.edu, mmurthi@miami.edu

[‡] Department of Communication Technology
Aalborg University, Denmark
sva@kom.auc.dk

ABSTRACT

The Internet Low Bit-rate Coder (iLBC) is now widely used for Voice over Internet Protocol (VoIP) applications. Unlike speech coders such as those based on Code Excited Linear Prediction (CELP), the iLBC achieves superior robustness to packet loss by avoiding inter-frame coding dependencies. While robustness to packet loss is essential, a VoIP codec should also possess the flexibility to change its source coding rate in order to counter network congestion and facilitate joint source channel coding for wireless channels. Previously, we presented a new variation of the iLBC encoding procedure which yielded a more efficient, rate-flexible result. In an effort to improve performance at lower source rates, we present various improvements to the original framework. Specifically, we reallocate bits from the Adaptive Codebook) procedure; reduce the length of the start state vector; utilize an adaptive pulse gain quantization scheme; and extend the use of entropy coding. Overall, the various combined improvements result in the modified iLBC (with entropy coding) achieving a rate reduction of 2.0 to 2.9 kbps when compared to the original fixed-rate iLBC without any loss in quality. In comparisons with Adaptive Muiti-Rate (AMR), the modified iLBC coder remarkably exhibits equivalent Perceptual Evaluation of Speech Quality (PESQ) scores as the AMR coder at 10.2 and 12.2 kbps, and out-performs AMR for all packet loss rates. This is a significant result as the modified iLBC performs equivalent to AMR without exploiting inter-frame redundancies.

1. INTRODUCTION

As Voice over IP (VoIP) becomes more widespread, it is necessary to develop and refine speech coding technologies to provide enhanced flexibility and robustness. In particular, flexibility is required in order to react to constantly changing channel characteristics. For example, it should facilitate better end-to-end Quality of Service (QoS) by allowing the use of Joint Source-Channel Coding (JSCC)[1] in wireless IP networks or effect TCP friendly rate/congestion control[2] which helps support the existence of heterogenous Internet applications.

Although providing good robustness to packet loss by utilizing frame independent coding, the internet Low Bitrate Coder (iLBC) [3] features an inflexible fixed-rate coding scheme. In [4], we proposed a framework with which to achieve a more rate flexible iLBC [3] speech coding solution. Specifically, we introduced a non-square synthesis matrix which captured how the iLBC builds up an approximation of the LP excitation vector from a much smaller 'start state' vector through its forward/backward Adaptive Codebook (ACB) procedure. The search and quantization of the start state was then framed in terms of an Analysis-by-Synthesis (AbS) matching problem which was solved using a Multi-Pulse (MP) approach to effect a variable rate speech coding solution. The start state was reformulated as a sparse vector of non-uniformly spaced pulses.

While speech quality was good at higher rates, the quality degraded rapidly as the rate decreased. Therefore, we present several improvements. Specifically, we reallocate bits from the ACB procedure to the MP state quantization and reduce the length of the start state vector in order to dramatically extend higher speech quality to lower rates. We also introduce an adaptive pulse gain quantization scheme to squeeze more efficiency at very low rates and lastly expand upon the entropy coding of the MP parameters. Overall, the various combined improvements result in the modified ILBC (with entropy coding) achieving a rate reduction of 2.0 to 2.9 kbps when compared to the original fixed-rate iLBC without any loss in quality. Remarkably, when operating without packet loss, the modified iLBC coder provides equivalent performance to AMR in terms of PESQ[9] at the 10.2 and 12.2 kbps rates. In the presence of packet loss, the modified iLBC coder performs better than Adaptive Multi-rate (AMR) on average at all loss packet loss rates when operating at source rates of 10.2 and 12.2 kbps. For lower bit-rates, for example 6.7, 7.4, and 7.95 kbps, the modified iLBC outperforms AMR when the packet loss percentages are greater than 7.5, 3.5, and 2.5%[5], respectively. It is vital to underscore the fact that our frame-independent coder under lossless conditions at rates above 10.2 kbps can achieve performance equivalent to industry standard codecs such as AMR.

The remainder of the paper is organized as follows. Section 2 summarizes our previous work from [4]. Section 3 introduces the various areas of improvement. Section 4 presents some performance results, and Section 5 concludes the paper.

2. TOWARDS VARIABLE RATE ILBC

The iLBC is a narrowband linear predictive speech coder which utilizes block based coding of the linear prediction (LP) residual signal through a combination of scalar quantization and ACB operations. The ACB is used to represent the longer $M \times 1$ LP residual vector $\mathbf{t}_{res}$ with a much shorter $N \times 1$ 'start state' vector $\mathbf{v}_{ss}$, where $M = 240$ and $N = 58$ samples, respectively. The start state is identified though a constrained search of $\mathbf{t}_{res}$ to select the N contiguous samples with the highest energy. Note that $\mathbf{v}_{ss}$ is vital to the iLBC coding algorithm in that it attempts to capture a

This work was supported in part by the National Science Foundation via CAREER Award CCF-0347229.

Special thanks to Deepak Dwarakanath, Aalborg University, for his contributions to the study of the reduced start state length.

good representation of periodicity or high energy noise in voiced or unvoiced speech and is used to exploit long-term redundancies in the LP residual.

In [4], we attempted to add flexibility to the iLBC algorithm by reformulating the representation of $\mathbf{v}_{ss}$. We hypothesized that a more judicious quantization of $\mathbf{v}_{ss}$ was possible by analyzing how the ACB, starting solely with the short vector $\mathbf{v}_{ss}$, is used to approximate the larger vector $\mathbf{t}_{res}$. We introduced a non-square $M \times N$ synthesis matrix $\mathbf{H}$ which captures this relationship in the resulting system of linear equations

$$\mathbf{t}_{res} \approx \mathbf{H}\mathbf{v}_{ss}. \quad (1)$$

Once formed, we placed this relationship within the perceptually weighted domain, effectively reducing the task to solving an AbS matching problem for the vector $\hat{\mathbf{v}}$ which minimizes

$$\|\tilde{\mathbf{t}}_{pw} - \tilde{\mathbf{H}}\hat{\mathbf{v}}\|^2. \quad (2)$$

where $\tilde{\mathbf{t}}_{pw}$ and $\tilde{\mathbf{H}}$ are the perceptually weight LP target and synthesis matrix, respectively.

While many solutions were available within this AbS matching framework, we chose a Multi-Pulse (MP) excitations approach [6][7]. Using MP, we achieved a more judicious quantization of $\mathbf{v}_{ss}$ by spending bits on the most perceptually important elements and altogether neglecting those deemed less important. We were able to vary the number of pulses P thereby effecting the rate flexibility we sought to add. The matching problem was of reasonable complexity as we had to choose P pulse locations out of N possible positions to match an M dimension target. The P pulses were found sequentially and their position and gains subsequently quantized. The gains were quantized using 4-bit scalar quantizers trained from speech taken from the TIMIT database. The positions were quantized in two ways: using $\lceil \log_2(N) \rceil = 6$ bits per position when $1 \leq P \leq 6$ and for larger values of P, using a N-bit position vector with a '1' denoting the presence of a pulse and a '0' indicating no pulse. The position vectors were then entropy coded using an arithmetic coder. We denoted the new MP start state representation as $\hat{\mathbf{v}}_{mpss}$.

The work resulted in a multi-rate speech coder operating between 7.87 and 13.2 kbps in the 30ms frame mode. At the highest rate, we achieved identical speech quality as the standard iLBC while saving 100bps. However, as the number of pulses was reduce the reproduced speech quality fell off sharply. In comparisons with AMR, our multi-rate coder achieved better performance at rates of 12.2 and 10.2 kbps when the packet loss rates were greater than 4 and 5%, respectively.

Our work left some question to be solved. For instance, can we reallocate bits from elsewhere in the coder to use in quantizing our new MP start state representation? Can the length of $\mathbf{v}_{ss}$ be reduced in the original algorithm before applying the AbS matching solution? Can we adaptively vary the number of bits used on the non-zero pulses? Can we entropy code the pulse gains as well?

3. CODEC REFINEMENTS

While our previous work was a bridge towards a practical implementation of a rate flexible iLBC, its performance and implementation left room for improvement, particularly at lower rates. In this section we will provide some answers to the various questions posed at the end of the previous section. Note that while both 20ms and 30ms modes were introduced in [4], the 30ms mode is preferred and will be solely discussed in the remainder of this paper. Also, performance evaluations of each proposed improvement will be presented in section 4.

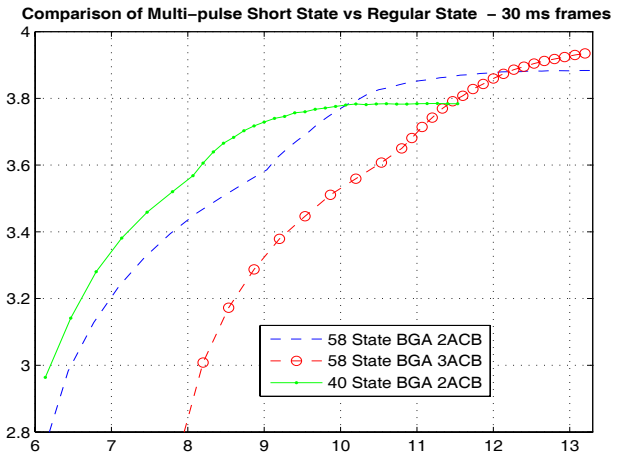

Fig. 1. Comparison of N=58 and N=40 state lengths in conjunction with the proposed modified iLBC without entropy coding. Here, the x and y axes correspond to rate and PESQ score, respectively. Curves of coder modes 3, 2, and 1 are ordered from left to right, respectively. Note that mode 1 was the best performance in [4].

3.1. Global Bit Reallocation and Fine Tuning

After examining our initial results, it was determined that a fine tuning process of reallocating bits from other portions of the iLBC might provide better performance at lower bit rates. This notion led to the idea of removing one of the ACB refinement stages and allocating that bandwidth towards state quantization. Therefore, less bits are used during the build up of $\mathbf{t}_{res}$, while more bits are used to represent $\hat{\mathbf{v}}_{mpss}$. This re-allocation essentially gives up a degree of refinement in one area, namely the ACB procedure, in order to add it to another, state quantization. By removing the third ACB stage, the maximum number of pulses/frame has increased while the base rate of the coder has decreased. Therefore, we use more pulses per frame at a given rate when compared to using all 3 ACB stages. It is important to note that the standard iLBC with 2 ACB stages does suffer a slight performance loss with a corresponding reduction in source rate. For example, at a fixed number of pulses P, the 3 ACB stage method will have better reproduced speech quality than the 2 ACB stage method; however, at that same P, the 2 ACB stage method operates at a source rate 1.77 kbps lower than the 3 ACB stage method. Lastly, as the distributions of the pulses change when using 2 ACB stages, we trained new 4-bit quantizers for use in these scenarios. For notation purposes, we will denote the modified iLBC coder with 3 and 2 ACB stages as modes 1 and 2, respectively.

3.2. Reducing The Length of the Start State

Using the multi-pulse approach to solve the proposed AbS matching problem yields viable results particularly when the number of pulses $P \rightarrow N$. However when P is small the average speech quality tends to be poor. This is mainly due to a disproportionate distribution of pulses in $\mathbf{v}_{ss}$. For example, the $N = 58$ sample vector is reduced to $P << N$ non-zero values. One method to solve this problem, is to reduce the dimension N of the vector $\mathbf{v}_{ss}$. By shrinking N, there exist fewer non-zero elements in $\hat{\mathbf{v}}_{mpss}$. A change in the dimension of N, however, requires some fundamental changes to the iLBC algorithm. First an algorithm to locate the $\mathbf{v}_{ss}$ is adapted from the original iLBC state search to fit a proposed dimension $N = 40$. In the new search, a constrained dimension $N = 40$ energy computation is performed over each of the six

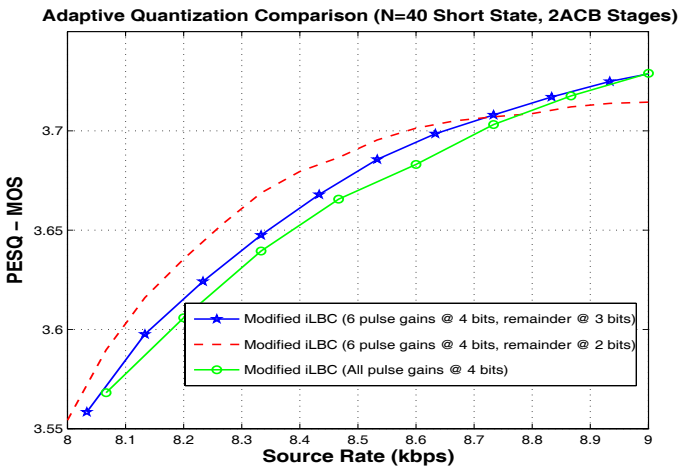

Fig. 2. Results of an adaptive pulse gain quantization scheme compared to the modified short state iLBC with MP start state. Both new curves (left and center) used a fixed 4bits/pulse for the first 6 pulses and 2 or 3 bits/pulse for the remaining pulses respectively. Here we save 200 bps at a fixed PESQ of 3.65.

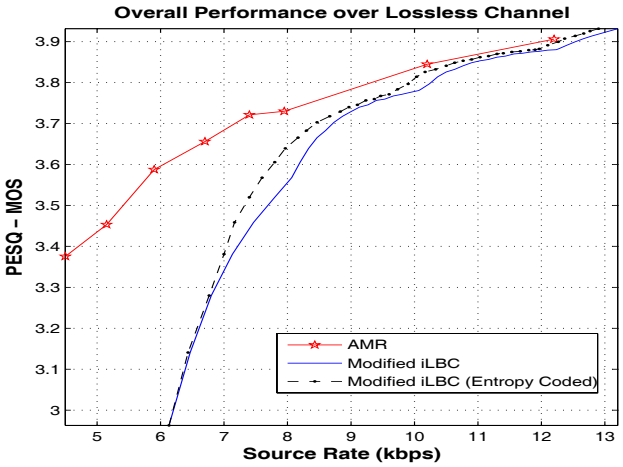

Fig. 3. Performance comparison between AMR and proposed modified iLBC over lossless channel.

subframes. The subframe with the most energy is selected as the 40 sample short start state vector $\mathbf{v}_{ss40}$. After selection, $\mathbf{v}_{ss40}$ is scalar quantized and used to populate the initial ACB memory. The ACB operations continue as normal; however, now the even smaller vector $\mathbf{v}_{ss40}$ is used to build up an approximation of $\mathbf{t}_{res}$. The resulting changes yield a version of the iLBC operating at a base rate of 11.6 kbps with 3ACB stages and 9.8kbps with 2ACB stages.

After altering the functionality of the fixed rate coder, we incorporated the MP approach into the short state coder design. The reduction in dimension of N lead to a reduction in the size of the synthesis matrix $\mathbf{H}$, a reduction in the maximum number of pulse search locations, and a new distribution of pulse gains for which we trained new scalar quantizers. Again for notation purposes, we denote the modified $N = 40$ iLBC with 2 ACB stages as coder mode 3.

3.3. Adaptive Gain Quantization

By introducing the new $N = 40$ short state iLBC and combining the MP approach, the overall range of the proposed multi-

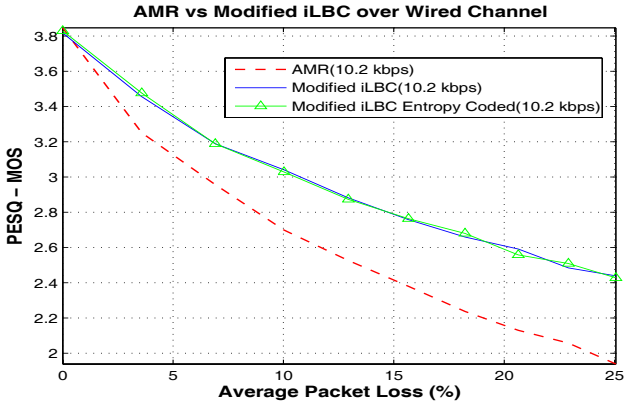

Fig. 4. AMR vs. modified iLBC @ 10.2kbps over lossy packet-switched network.

rate coder was enhanced, particularly at the lower source rates. While significant improvements were made, gains at even lower rates were desired. At very low values of P, $\mathbf{v}_{mpss}$ is still quite sparse even when its dimension is reduced to $N = 40$ samples. It should be clear that adding pulses improves quality within this framework. So how then can we squeeze more pulses out of the same amount of bits? One answer lay in our method of gain quantization. Previously, a fixed allocation of 4 bits/gain was utilized when quantizing the pulse gain parameters. This lead to fixed intervals in the usable source rates and more importantly assumed that all pulses were created equal. In a new adaptive pulse gain quantization scheme, not all gains were to be treated equally.

First we assumed that a certain number of pulses were essential for speech to be intelligible in our proposed coder. For example, one would not use rates where only 1 or 2 pulses are transmitted. To account for this our adaptive scheme maintains 4 bits/pulse gain for the first $P = 6$ pulses. After quantizing these 6 pulses, however, less bits are used for all remaining pulses. By decreasing the quantization resolution, lower source rates can be achieved. Also, more pulses can be used at a fixed rate. In one approach, after quantizing the first 6 pulses at 4 bits/pulse the remainder of the pulses were quantized using 2 bits/pulses. In the second approach, the remainder of the pulses were quantized using 3 bits/pulse. Note, that only a few simple schemes for the values for P and the number of bits were chosen as there are a large number of possibilities to be tested in such adaptive schemes.

3.4. Entropy Coding

In the previous work, we explored the idea of entropy coding the position parameters. For the purpose of this work, we wanted to investigate entropy coding of the pulse gain parameters. Initially using a first order analysis, we determined that the probability distributions were too uniform to achieve any lossless coding gains. Further analysis was conducted in which a second order approach was utilized. Now the probability of a symbol conditioned on the previous symbol is estimated. This makes sense in practice because there is a degree of correlation between the quantized pulses. For example, a large positive pulse is followed by another positive pulse more often than by a large negative pulse. In the second order analysis, the conditional probability system was modeled as a 16 state Markov model[8]. By estimating the transitional probabilities between each state and the stationary probability of being in each state, the entropy rate of the system was computed[8]. A second order Huffman coding strategy was implemented whereby 16 Huffman codebooks were built for each possible value of P,

one for each conditional state. Each codebook was built using the transitional probabilities from a particular state to all other states. Therefore each codebook represents a previous state (or symbol) and describes the codewords used to describe or encode the next state[8]. When using entropy coding, the total number of bits can vary from frame to frame.

4. EXPERIMENTAL RESULTS

As a benchmark for testing, we utilized over one hour of speech, including sentences from 100 male and female speakers, from the TIMIT database. All performance scores were obtained using the PESQ algorithm. In figure 1, we see comparisons of three coder modes 1, 2, and 3 respectively. Notice how each newly improved mode improves upon the next in terms of holding a quality level at a certain rate. For example, mode 3 holds a PESQ score of 3.4 at 7.2 kbps while modes 2 and 1 hold this score at 7.7 and 9.3 kbps, respectively. Figure 2 depicts the results of adaptive quantization for coder mode 3 for low values of P. Notice that while the gains are small, for example 100 bps when PESQ is fixed at 3.6, we tested only two simple schemes. This adaptive quantization area could be rigorously tested across all values of P using much more complex schemes.

In figure 3 we see several things. First, notice the overall combined curves of coder modes 1, 2, and 3 with and without entropy coding over lossless simulation channels. Note, the average rate reduction of 500 bps when using entropy coding at a fixed PESQ of 3.55. Also, note that adding entropy coding to the MP gains saved 7% extra on average. This figure also depicts a comparison of the improved modified iLBC with AMR. Notice that the two track each other closely at rates above 10 kbps. Overall this is a substantial improvement from the comparison in [4] where only coder mode 1 existed.

Finally, consider figures 5 and 4 where packet-loss simulations paralleling those in [4] for source rates of 10.2 and 12.2 kbps are depicted, respectively. Here the modified iLBC and AMR are compared in terms of robustness to packet loss (simulated using a Gilbert model with parameters, $q = P(loss|loss) = 0.7$ and $p = P(loss|noloss)$ varied to achieve average loss rate $\frac{p}{p+q}$). First, note that without packet loss, the modified iLBC performs equivalently to AMR at the same rate in terms of PESQ. This is a remarkable result as AMR is using the Adaptive Codebook across frames while iLBC remains a frame independent codec. Note that with the proposed improvements the modified iLBC performs as good or better than AMR at all loss rates. Note that in [4], AMR performed better until the loss percentage was above 5 and 4% for the respective figures. At lower bit-rates, for example 6.7, 7.4, and 7.95 kbps, the modified iLBC outperforms AMR when the packet loss percentages are greater than 7.5, 3.5, and 2.5%[5], respectively. It is important to note that the curves with and without entropy coding in figures 5 and 4 are nearly identical therefore concerns over unfair comparisons with AMR due to variance issues can be neglected. At the lower rates, however where entropy coding give greater gains these issue are still relevant.

5. CONCLUSIONS

By adding multiple source rates to the iLBC's inherent packet loss robustness, we effect a speech coder with the flexibility needed to compete today's communication networks. However, the original design could not maintain good speech quality at medium to low rates. In this paper, we presented a series of improvements to our modified iLBC speech coder. By evaluating tradeoffs and exploring differing operational choices in the coder design, such as

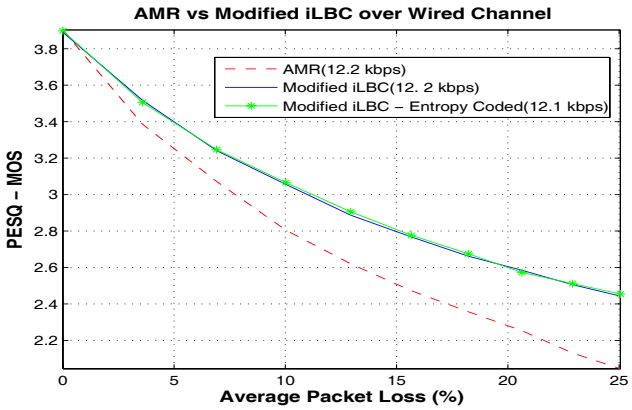

Fig. 5. AMR vs. modified iLBC @ 12.2kbps over lossy packet-switched network.

number of ACB stages, initial length of start state, adaptive pulse gain bits allocation and pulse gain entropy coding, in conjunction with the MP AbS matching framework, we pushed the practical range of the coder to much lower rates. For example, the modified iLBC (with entropy coding) achieves a rate reduction of 2.0 to 2.9 kbps when compared to the original fixed-rate iLBC without any loss in quality. In comparisons with AMR at source rates of 10.2 and 12.2 kbps, the iLBC now performs as good or better in packet loss situations at all loss rates. To the extent to which PESQ-MOS is accurate in assessing speech quality, it is remarkable that our enhanced iLBC, being a frame independent codec, obtained similar PESQ scores as a state-of-the-art CELP codec such as AMR while at the same time giving superior robustness to packet loss.

6. REFERENCES

[1] J. Hagenauer, and T. Stockhammer, "Channel coding and transmission aspects for wireless multimedia," In *Proceedings of the IEEE*, vol. 87, no. 10, Oct. 1999.

[2] S. Floyd, M. Handley, and J. Padhye, "Equation-Based Congestion Control for Unicast Applications: the Extended Version" ICSI Report Number TR-00-003, March 2000.

[3] Andersen, S.V., W.B. Kleijn, R. Hagen, J. Linden, M.N. Murthi, and J. Skoglund, "iLBC-A Linear Predictive Coder with Robustness to Packet Losses," In *2002 IEEE Speech Coding Workshop Proceedings*, pp.23-25.

[4] C.M. Garrido, M.N. Murthi, S. V. Andersen, " Towards iLBC Speech Coding at Lower Rates Through A New Formulation of the Start State Search," In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Philadelphia, USA. March 2005.

[5] Garrido, Christopher M. "Multi-Rate Speech Coding Through New iLBC Start State Formulation." Master's Thesis. University of Miami, 2005.

[6] B.S. Atal "High-quality speech at low bit rates: multi-pulse and stochastically excited linear predictive coders" In *Proceedings of ICASSP 1986*.

[7] A.M. Kondoz, *Digital Speech: Coding for Low Bit Rate Communication Systems*, Wiley.

[8] Thomas M. Cover and Joy A. Thomas. Elements of Information Theory. Wiley Series in Telecommunications. John Wiley & Sons, New York, NY,USA, 1991.

[9] ITU-T P.862 "Perceptual Evaluation of Speech Quality (PESQ)."

[10] Ekudden, E., R. Hagen, I. Johansson, and J. Svedberg, "The adaptive multi-rate speech coder" In *1999 IEEE Speech Coding Workshop Proceedings*, pp. 117-119.