In Proceedings of the International Conference on Advanced Infocomm Technology, Da Meisha, China, July 2008. 7 pages.

Energy Efficient Implementation of G.729 for Wireless VoIP Application

Aihong Yao
School of Computer Science and Technology and
Institute for Computer Architecture
Harbin Engineering University
Harbin, Heilongjiang, China
yaoaihong@hrbeu.edu.cn

Junjun Gu, Gang Qu, and
Shuvra Bhattacharyya
Electrical and Computer Engineering Department and
Institute for Advanced Computer Studies
University of Maryland, College Park, Maryland, USA
{alicegu, gangqu,ssb}@umd.edu

ABSTRACT

In this paper, we describe the formatting guidelines for ACM SIG Proceedings. Traditionally silence in VoIP applications is detected by a VAD algorithm after G.729 compression, both have high computational and energy cost. However, such cost on silence frames will be completely wasted because the detected silence frames will not be sent to receiver. We propose to use a silence pre-detection (SPD) module to detect silence frames directly from the voice sample in order to reduce energy. Our adaptive SPD algorithm has little hardware requirement and low computation cost. However, it can detect 59% of the silence, which results in 34% energy saving. The impact to quality of speech is almost unnoticeable.

Categories and Subject Descriptors

C.3 **[Special-purpose and Application-based Systems]:** *Signal processing systems, Real-time and embedded systems.* C.2.0 **[Computer-Communication Networks]:** General – *data communication*

General Terms

Design, Performance.

Keywords

embedded system design, VoIP, speech signal codec

1. INTRODUCTION

Wireless internet telephony (WIT) plays an increasing important role in today's mobile phone market. Cell phones (such as Apple's iPhone and Google's Gphone) are now equipped with new features such as multimedia playback, web browser, video game, GPS, and email, they are changing the way how people live and stay connected. With the addition of these variety and computation intensive applications, power and energy efficiency of the processing unit quickly becomes one of the most critical design challenges. SPD algorithm has little hardware requirement and low computation cost. However, it can detect 59% of the silence, which results in 34% energy saving. The impact to quality of speech is almost unnoticeable.

Voice over IP (VoIP) protocol defines the way that voice signals are carried over the IP network. Tremendous amount of research efforts, mainly from the communication and networking society, have been put to provide high quality speech under constraints such as network bandwidth, latency, jitter, packet loss, and most G.729, an audio data compression standard [5] widely used in VoIP applications. The annex B defines how silence can be compressed to save system resource without causing large speech quality degradation. It has a voice activity detection (VAD) algorithm that can detect the silence in the speech and disable the transmission of silence packets. On the receiver's end, a discontinuous transmission (DTX) module updates the background noise parameters at the absence of the (silence) data frames.

Although the VAD algorithm and DTX module achieve the goal of improving G.729 standard's network bandwidth utilization with little quality of speech degradation, they, particularly VAD and G.729 compression algorithm, are computation intensive operations and consume a lot of energy on the processing unit. In this paper, we propose a silence pre-detection (SPD) method that can significantly reduces this part of the energy consumption.

The rationale behind our approach is as follows: since the silence frame, once detected by the VAD algorithm, will not be transmitted, we can save the energy consumed to perform the G.729 compression and VAD algorithm on such frames if we can predict the silence during the speech. Normal conversation in telecommunication scenario contains an average of 60% silence with many of them last long period [5]. If we can predict 50% of these silences, we can save about one-third of the energy consumed on G.729 compression and VAD module. At the same time, on the receiver end, the codec can adjust its execution speed according to the packet mark.

We propose to add an adaptive SPD module before the G.729 compression module. It will directly estimate the voice sample. If the estimation shows a silence, G.729 compression and VAD algorithm will be bypassed; if the estimation shows the sample is a speech, the voice sample will go through G.729 compression and VAD following the G.729 standard. For energy efficiency, we use a simple estimation method to judge whether a frame is a silence or a speech. To increase the accuracy of the SPD module, we use a feedback loop to update the estimation criteria whenever a silence missed by SPD module is detected by the VAD algorithm. We test our approach on 16 audio clips, on average, the SPD module can detect 59% of the total silence which results in 34% energy saving on the processing unit. The PESQ test shows that the degradation of speech quality using our implementation of G.729 is only 3.92% comparing with the standard G.729 approach which is hardly perceivable.

The rest of the paper is organized as follows: in section II, we give a brief review of the related work on VoIP, G.729, and energy efficient multimedia device design. In section III, we present the preliminary on G.729 and elaborate our silence pre-detection based implementation. Section IV reports the simulation results and section V concludes.

2. RELATED WORK

As we have already surveyed in the introduction section, most of the up-to-date research focus on network aspects of VoIP applications. Here we only discuss those that are relevant to our work.

Due to the imperfectness of human auditory system, certain level of frame loss might not be noticed. There are several papers measuring the perceptual impact caused by frame losses. The general belief is that such impact depends directly on the content of the lost frames and the level of impact varies a lot.

De Martin et. al.[4] presented a source-driven approach to mark a packet as premium or regular after the frames were encoded and encapsulated in a packet. To make it simple, they assumed that each packet contains exactly one G.729 output frame, which is one data sample of 10 ms. They marked a coded frame to be perceptual critical (or premium) by computing the distortion between original parameters and corresponding estimates. Their formal listening tests showed that the source-driven packet marking enhances the speech quality from MOS (Mean Opinion Score) 3.4 to 3.7 in case of a 5%loss rate when 19% of all packets are marked as premium.

Similarly, [4] presented a classification of AMR frames. His analysis-by-synthesis distortion evaluation calculates the spectral distortion in dB for the LP (Linear Prediction) coefficients, the percentage difference for the long-term prediction coefficients and the difference in dB for the codebook gains. If any of these values is above a given threshold, an AMR frame is marked as premium.

C. Hoene described a quality metric based on ITU-T P.862 PESQ (Perceptual Evaluation of Speech Quality) to describe the importance of speech frames or VoIP packets. They presented an aggregation function to quantize the impact of multiple losses, e.g. burst losses, using the importance of single frame loss. Their simulations showed that most speech frames during voice activity were not important at all and different frame dropping strategies can increase the MOS from 1.8 to 3.7 in the case of a 40% frame loss rate.

The above work study random frame loss. The 60% silence frames during VoIP applications have also been studied, most to improve network utilization and bandwidth. [ITU-T G.729 Annex B] defined a VAD module to detect inactive voice frames, also called silence or background noise frames, and defined a discontinuous transmission (DTX) module to decide whether a compressed low-bit-rate silence frame should be sent. At the receiving end, a comfort noise generator module synthesized background noise from those compressed silence frame parameters.

[3] proposed a hybrid power saving mechanism for VoIP applications under the IEEE 802.16e standard. They used power saving class II during talk periods and power saving class I during silence periods. This gave about 20% energy saving on the radio communication unit while meeting the packet drop probability required by VoIP services.

Similar to drop the silence frames, there have been several research on energy saving by intentional missing execution deadlines for soft real-time systems. W. Yuan et. al. [9] implemented an energy-efficient soft real-time CPU scheduler, named GRACE-OS, for mobile multimedia systems. Through a cycle counter added into the process control block of each task, the GRACE-OS kept track of the number of cycles consumed by n jobs (e.g., the last 100 jobs) of the task. And according to the histogram approximates of the cumulative distribution function of a task's cycle demand, it could make a prediction of the time demand in the future at certain probability.

S. Hua[8] et. al. proposed a new design method, called probabilistic design, to expand the traditional design space of the multimedia embedded systems while occasional deadline misses can be tolerated. They developed algorithms for quick exploration of the enlarged design space at early design stage. They assumed a given execution time distribution of each task and the tolerance to deadline misses, therefore, they estimated the probabilistic timing performance and managed system resources, e.g. energy consumption, one of the most critical resources for multimedia embedded system.

Wireless internet telephone and VoIP applications are also soft real-time applications, where the streamlined frames need to be processed in a timely fashion and the frame/packet drop can be tolerated to certain degrees. Our proposed work belongs to the category of probabilistic design because we deliberately bypass G.729 encoding and VAD algorithm when silence frame is detected. However, since we are dropping only silence frames, there will be little impact to speech quality.

3. PROPOSED ENERGY-EFFICIENT DESIGN OF WIT

3.1 Preliminary on WIT Device and G.729

There are three major components in a WIT device: controller, audio signal processor, and wireless network communication processor. The controller executes the network protocols handling, user interface controlling, and other tasks to collaborate with the other two parts of the system. The wireless network processor is used to process wireless network protocols, such as 802.11 series and Bluetooth, as well as to interact with the RF (Radio Frequency) module. The audio signal processor compresses the local input voice data and decompresses the encoded speech frames received by the wireless communication module.

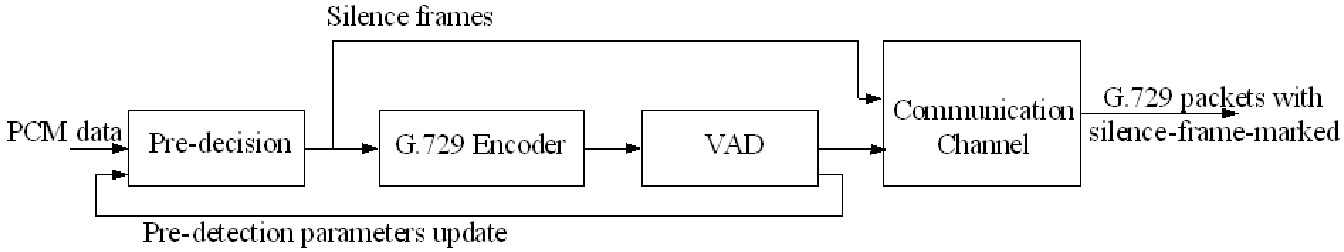

Figure 1. Overview of the proposed energy efficient G.729 implementation with the silence pre-detection module.

The complexity of coding process depends on the encoding algorithms that the WIT device supports. For mobile VoIP, ITU-T G.729, a toll-quality speech coding algorithm, is the main contender for the baseline codec. G.729, also known as CS-ACELP (Conjugate Structure Algebraic Code Excited Linear Prediction), compresses speech from 16-bit 8kHz samples (128kbps) to 8 kbps by taking advantage of the correlation between consecutive input frames. As we know, the higher compression ratio the codec provides, the more complicated the compression algorithm is. For the TI TMS320 C5X series that adopt G.729, the encoding algorithm requires approximately 21.47 MIPS (18.5 MIPS for encoding and 2.97 MIPS for decoding) for real-time applications.

3.2 Overview of the Energy Efficient G.729 Implementation

G.729 is an audio data compression algorithm for voice. It takes and compress voice sample in a 10-millisecond period. It is widely used in VoIP applications at 8 kbps. The annex B of G.729, or G.729B, includes a VAD algorithm to detect whether a frame is a speech frame or a silence frame. Because the speech/silence decision is a hypervolume of multi- dimensional Euclidean space constructed by speech parameters, exact decision can not be obtained just by one characteristic. The VAD algorithm needed some parameters to make the correct decision. These parameters include linear prediction spectrum, full-band energy, low-band energy, zero-crossing rate, etc. Most of them are obtained from or shared with G.729 encoder. The goal of VAD algorithm is to detect silence frames so they will not be sent in order to save network bandwidth and power. To increase the accuracy of VAD algorithm, many approaches, some are quite complex and computational intensive[2, 12], such as Hidden Markov Model, have been successfully proposed.

Figure 1 depicts the overview of our proposed energy efficient implementation of G.729 with a new silence pre-detection (SPD) module. The SPD module works on the PCM data sample to identify as many as possible silence frames. Once identified, a signal will be sent directly to communication channel where a mark is made for the silence frame. Thus the computational expensive G.729 encoding and VAD algorithm are bypassed, saving processing energy. On the other hand, if the SPD module estimates a frame to be a speech frame, the data will go through G.729 encoder and VAD module as usual.

As we mentioned that the speech/silence detection is a non-trivial task, we do not expect the SPD module to detect all the silence frames. The undetected silence frames will be detected by the VAD algorithm. Therefore our proposed architecture will not degrade the silence detection ratio. However, because a frame that SPD module identifies as silence will bypass G.729 and VAD, we want to eliminate or reduce the false positive, i.e. reporting speech as silence.

Our empirical study indicates that energy of input voice can be used as a simple criterion for the SPD module. If the energy of input sample frame is below a threshold value, we say it is a silence. This SPD module requires little computation. However, the determination of the threshold value is critical. A high threshold results in low silence detection ratio and a low threshold will cause false positive.

We choose the average energy of the silence frames as the threshold. These silence frames include those detected by the SPD module and those SPD fails to detect but reported by VAD algorithm. An adaptive method is used to determine this value as shown in Figure 1, where a feedback threshold update command can be sent to the SPD module if there are false positive is accumulating. The next two subsections elaborate how this method works.

3.3 Adaptive Silence Pre-detection Module

Speech/silence detection can be modeled as a Hidden Markov Model [2] or Probabilistic Finite State Machine [10]. If S0 denotes silence state, S1 denotes speech state, the state transition between speech and silence can be illustrated as in Figure 2. The state transition probability matrix T={t_{ij}}, where t_{ij} is the transition probability from state S_i to state S_j for i,j = 0,1. The following condition holds $\sum_j t_{ij} = 1$ for $i = 0,1$.

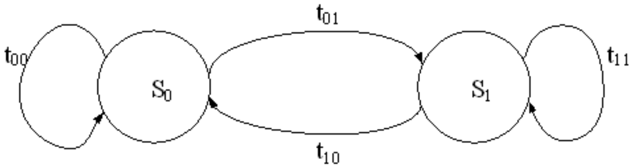

Figure 2. Probabilistic finite state machine of VAD.

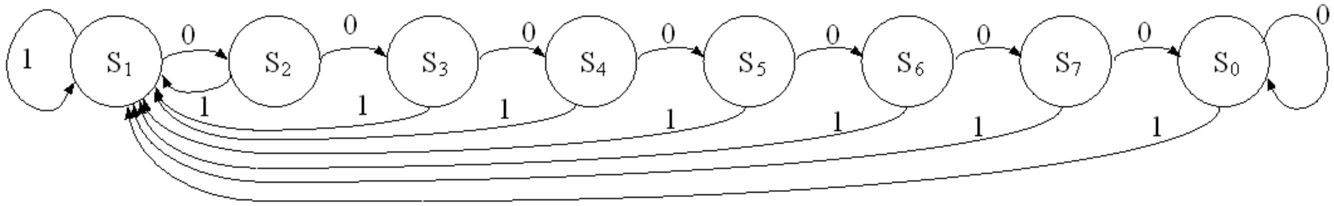

Figure 3. Finite state machine model for the silence pre-detection update procedure. The number 0/1 on edges from state S0 is the decision by the SPD module, and the number 0/1 on other edge indicates the feedback from the VAD algorithm. 0 for a silence and 1 for a speech in both cases.

The transition probabilities between speech to silence, t_{10} and t_{01}, are the most important for a communication system. They affect the silence detection accuracy and perceptual speech quality. The t_{10}, for example, affects the false negative ratio, i.e. misclassifying a silence frame as a speech frame, which the echo cancellation systems are normally sensitive to. On the other hand, the t_{01} can increase the clipping errors, which impacts the speech intelligibility. [2] proposed an exponential p-dimensional multivariate distribution for state Probability Density Function (PDF). And it can improve the average clipping rate and false detection rate, 72.21% and 72.37%, respectively under different SNR and background noise, such as babble, car, and white noise comparing to the G.729B.

We use an FSM (Finite State Machine) model for the silence pre-detection update procedure as shown in Figure 3. We use eight states, less or more states can be used based on different characteristic of the VoIP applications, to identify the detection status. S0 and S1 are silence and speech frames as before. The states S2 to S7 represent the hangover of speech frame. During the hangover states S2 to S6, we conservatively treat the frame as speech although SPD module may tell us it is a silence. These frames will be passed to G.729 encoder and VAD algorithm. If VAD algorithm finds the frame is indeed a speech, the state moves back to S1. Only when the VAD algorithm finds enough (in this case, seven) consecutive true silence frames, will the state move to the silence state 0. Meanwhile, SPD module will upgrade its threshold value if the current threshold is producing false negatives (i.e. failing to detect silences). In this way, we achieve the goal of eliminating or minimizing false positive to ensure speech quality.

3.4 Threshold Updating Algorithm

The threshold, Th, used in the SPD module is defines as the average energy of the silence frames. The following pseudo-code shows how this is updated according to the above adaptive method:

```
1.  Th=0; N=0; i=1;
2.  compute energy $E_i$ of sample i;
3.  if ($E_i$ <= Th)        {N++; Th = Th - (Th-$E_i$)/N;}
4.  else if (fb == 0) {N++; Th = Th - (Th-$E_i$)/N;}
5.  i++;
6.  goto 2;
```

Initially, we set Th=0 and N=0 silence frame and start with frame i=1. Setting Th=0 guarantees the speech frames will not be detected as silence When a silence frame is detected either by the SPD module (in line 3) of by the VAD algorithm (in line 4), we update Th as the new average silence energy with this newly detected silence frame:

$$Th_{new}= (Th_{old}*N_{old}+E_i)/(N_{old}+1) = Th_{old}-(Th_{old}-E_i)/N_{new} \quad (1)$$

Note this calculation requires two subtractions and one division. To further simplify this computation, we can replace Nnew by 2k, where $k=\lceil \log Nnew \rceil$. This will replace the division by a logic right shift operation.

3.5 Design Evaluation and Energy Efficiency Analysis

The hardware implementation of the SPD module is very low. We need a counter (to counter the number of silence frames, however, we will see in the simulation section that the threshold Th converges rapidly and the Th update can be stopped), two registers (one stores the threshold value Th, the other stores the current sample energy Ei). Both subtraction and logic shift are easy to implement. Furthermore, we can implement them in the controller. Trivial computation implies negligible delay and energy consumption, particularly when compared to G.729 encoding and VAD algorithm.

For the normal VoIP applications which has 60% silence. If our SPD module can detect α ($0<\alpha<1$) of the total silence frames, 0.6α of the total frames will bypass the G.729 encoder and VAD algorithm, which means a 0.6α saving of execution time and energy on the processing unit. On the receiver end, we can save the same amount of energy. By taking advantage of the network protocol, further saving is achievable. For example, in a DiffServ network [11], a packet marking strategy can be applied with the result of VAD. In the case of 40 ms packet duration, i.e. 4 G.729 frames per packet, four-bit mark can be added to the packet header to inform the receiver which frames are silences. It can be recognized while the controller depacketizes the speech packet. This knowledge can help the decoder decide how fast it performs the decoding tasks.

4. EXPERIMENT RESULT AND DISCUSSION

In this section, we first investigate the accuracy of the proposed SPD module. Then we report the energy saving, finally, we provide results on the measurements of the speech quality. Our experiment is based on ITU-T P.862 VoIP speech samples [6].

4.1 Accuracy of the Silence Pre-detection Module

We measure the silence frame detection accuracy, or silence detection ratio, of the SPD by the number of detected silence frames and the total number of silence frames, where the latter is obtained by running the G.729 compression and VAD algorithm for each voice sample. Figure 4 depicts the change of SPD's silence detection ratio with time under different signal-to-noise ratio (SNR) white noise for voice sample u_af1s01.wav. We observe that the detection accuracy is directly related to SNR. With a high SNR = 20dB, our simple SPD module is able to detect more than 60% of the silence. However, in the noisy environment with SNR=0dB, the accuracy is almost halved.

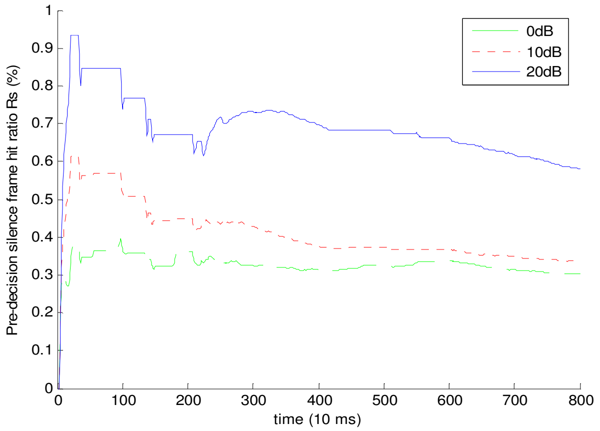

Figure 4. Silence detection ratio under different SNRs.

Figure 5 shows the normalized amplitude of the input voice sample u_af1s01.wav and how the threshold changes as time goes on. In the bottom half of Figure 5, the energy of noise is shown and the dark dotted line gives the threshold value from frame to frame. This threshold value is used by the SPD module to determine whether a voice sample is silence or speech. Note that we have an adaptive algorithm to update this threshold value. We see that the threshold converges quickly, in about 250 frames or 2.5 seconds, to the level of noise energy. We mention in most of the voice samples we used, the threshold converges within 2 seconds.

4.2 Energy Saving by Silence Pre-detection

In our implementation of G.729, once the SPD module detects a silence frame, the frame will not go through the G.729 compression and VAD algorithm. On the other hand, if the SPD detects a speech frame based on the data sample, the sample will be sent to the G.729 compression as usual. Therefore, the energy reduction by our approach comes from the detected silence frames. Hence, it is directly affected by the SPD module's silence detection accuracy. Recall that the SPD algorithm and threshold update takes only several simple arithmetic or logic operation. Therefore we ignore their energy overhead comparing to the computational intensive G.729 compression and VAD algorithm.

Table 1 shows the detailed result on 16 voice samples. The second column gives the length of each sample. The third column gives the percept of silence frames in each sample. This percentage is obtained from the simulation of G.729 compression followed by VAD algorithm, which detects all the silence frames. We see that on average there is 57.83% of silence in the samples as reported by [5].

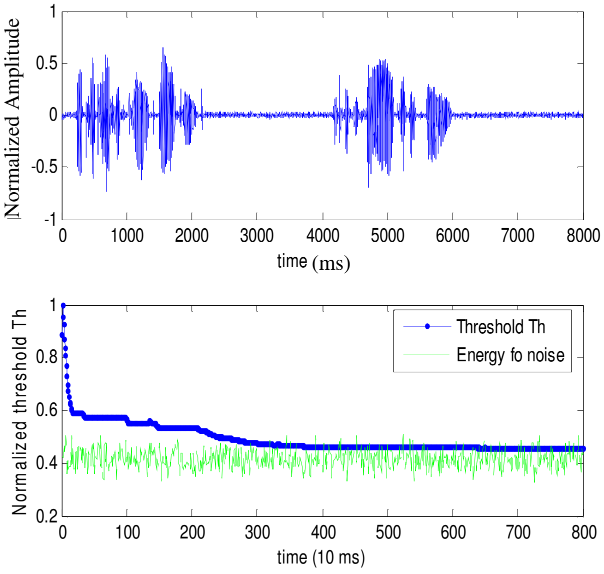

Figure 5. Threshold Th versus time

The next column shows the silence detection ratio of the SPD module. For example, It detects 73.28% of the silence in sample u_af1s01.wav, where there are in total 63.62% silence. Therefore, the energy saving (at the transmitter's side of the signal processor) is 46.62%. We see that on average, the SPD module's 59.4% silence detection accuracy brings us 34.3% energy saving.

4.3 Perceptual Speech Quality

Two popular metrics are being used to measure the speech quality. Subjective listening only test, such as MOS (Mean Opinion Score), can investigate the exact intelligibility of human, but it has some drawbacks, such as financial cost, time cost, and the inability in real-time applications. Objective metrics for perceptual speech quality have also been proposed to assess speech quality automatically. P.862 PESQ (Perceptual Evaluation of Speech Quality) is recently developed and has been adopted by many commercially available testing devices and monitoring systems.

Tabel 1. The performance of the proposed pre-decision design against the performance of the G.729.

Sample	Length (ms)	Silence (%)	α (%)	Energy* Saving (%)	PESQ (G.729)	PESQ (SPD)	Degradation (%)
u_af1s01.wav	8000	63.62	73.28	46.62	2.97	2.77	6.77
u_af1s02.wav	8000	63.25	68.77	43.50	3.50	3.34	4.71
u_af1s03.wav	8000	56.62	73.51	41.62	3.42	3.21	6.16
or105.wav	8402	60.00	51.98	31.19	3.77	3.65	3.00
or109.wav	8039	56.91	50.77	28.89	3.73	3.65	2.26
or114.wav	8591	62.28	62.24	38.76	3.77	3.62	2.95
or129.wav	7231	60.58	55.71	33.75	3.81	3.69	3.31
or134.wav	8075	64.31	57.23	36.80	3.78	3.55	6.19
or137.wav	7059	59.57	60.95	36.31	3.74	3.62	3.34
or145.wav	8251	52.00	52.91	27.51	3.71	3.57	3.86
or149.wav	8185	57.33	59.49	34.11	3.78	3.57	5.58
or152.wav	7193	41.03	68.47	28.09	3.71	3.56	4.15
or154.wav	7217	60.89	40.55	24.69	3.73	3.66	1.90
or155.wav	7407	62.03	56.21	34.87	3.84	3.75	2.37
or161.wav	7893	54.88	64.67	35.49	3.74	3.60	3.82
or164.wav	7606	50.00	54.21	27.11	3.72	3.64	2.34
Average	--	57.83	59.43	34.33	3.67	3.53	3.92

In our experiments, we perform both subjective and objective tests for each sample, which is processed once by the normal G.729 implementation and another time by our energy efficient implementation. During the subjective test, the audience cannot tell any difference.

The last three columns of Table 1 list the PESQ test data for the two implementations of G.729. We see that on average, our implementation is able to maintain a score of 3.53. Note that according to the PESQ, the scores are in general acceptable. The last column shows that the speech quality degradation of our approach compared to the normal implementation is only 3.92%, which is negligible.

5. CONCLUSION

In this paper, we propose an energy efficient implementation of the G.729 standard. We use a silence pre-detection module to detect silence frame based on the sample data directly. When we detect silence fame, it will bypass the energy costly G.729 compression and VAD algorithm and thus saves energy. We show that our simple SPD module is quire accurate in detect silence, catching 59.4% of the silence on average. This results in a 34.3% energy saving with a merely 3.92% of speech quality degradation

6. ACKNOWLEDGMENTS

This work is partially supported by grant CNS0615222 from the National Science Foundation. Dr. Yao's work is done during her visit to the University of Maryland, College Park.

7. REFERENCES

[1] Bur Goode, “Voice over Internet Protocol (VoIP),” Proceedings of the IEEE, Vol. 90, No. 9, pp. 1495-1517, September 2002.

[2] H. Othman and T. Aboulnasr, “A semi-continuous state transition probability HMM-based voice activity detection,” EURASIP Journal on Audio, Speech, and Music Vol. 2007, Issue 1, pp.1-7, January 2007.

[3] Hyun-Ho Choi, Jung-Ryun Lee, and Dong-Ho Cho, “Hybrid Power Saving Mechanism for VoIP Services with Silence Supression in IEEE 802.16e Systems,” IEEE Comm. Letters, pp.455-457, May 2007.

[4] J. C. De Martin, “Source-Driven Packet Marking for Speech Transmission over Differentiated-Services. Networks,” in Proceedings of ICASSP, 2001.

[5] Adil Benyassine, Eyal Shlomot, and Huan-Yu Su, “ITU Recommendation G.729 Annex B: A Silence Compression Scheme for Use with G.729 Optimized for V.70 Digital Simultaneous Voice and Data Applications,” IEEE Comm. Mag., pp.64-73, September 1997.

[6] ITU-T Rec. P.862, “Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.”

[7] C. Hoene, H. Karl, and A. Wolisz, “A Perceptual Quality Model Intended Adaptive VoIP Applications,” Special issue Performance Evaluation of Wireless Networks and Communications of the Computer Communications Journal, 2005.

[8] S. Hua, G. Qu, and S. S. Bhattacharyya, “Probabilistic design of multimedia embedded systems,” ACM Trans. Embedded Comput. Syst. Vol. 6, No. 3, pp.1-24, July 2007.

[9] W Yuan, and K. Nahrstedt, “Energy-efficient soft real-Time CPU scheduling for mobile multimedia systems,” Proc. of the 19th ACM Symposium on Operating Systems Principles (SOSP’ 03), Bolton Landing, NY, October 2003.

[10] E. Vidal, F. thollard, C. de la Higuera, F. Casacuberta, and R. C. Carrasco, “Probabilistic finite-state machines-part I, ” IEEE Trans. Pattern analysis and machine intelligence, Vol. 27, No. 7, pp.1013-1039, July 2005.

[11] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, “An architecture for Differentiated Services,” RFC 2475, December 1998.

[12] O-W. Kwon, and T-W. Lee, “Optimizing speech/non-speech classifier design using Adaboost,” IEEE International Conference on Acoustics, Speech and Signal Processing, Apiral 2003.