A Description of the ITU G.729 Vocoder Variants

Adaptive Digital Technologies, Inc.

January 7, 2001
Issue 1

Source File: G729_White_Paper.doc

Revision History

Issue	Revision History	Date	Initials
1	Creation	1/7/2001	SDK

The ITU G.729 vocoder is used to compress digitized speech data. There are many variations of the G.729 vocoder. This paper discusses the most commonly used variations. These variations are the fixed point implementations that operate at 8000 bits per second.

The input (uncompressed) speech is sampled at 8000 samples per second. The samples are assumed to be quantized using 16 bits of resolution. The input bit rate is therefore 8000 * 16 = 128 kbps.

The G.729 vocoder operates on 10 millisecond frames. Its compressed bit rate is 8000 bps. The encoder therefore takes in 80 samples of speech data per frame and outputs 80 bits of compressed speech data. The decoder takes in 80 bits of compressed speech data per frame and outputs 80 samples of uncompressed synthesized speech.

This paper discusses the original G.729 vocoder as well as its first two Annexes (A and B). The basic vocoder employs an ACELP (Algebraically Excited Linear Prediction) speech coding algorithm. The details of the algorithm are beyond the scope of this paper. These details can be found in the ITU recommendations.

G.729 Annex A is a reduced complexity version of G.729. It requires about half the CPU power to operate G.729, with the drawback that it has slightly reduced speech quality. G.729 is rated at 3.9 on the MOS (Mean Opinion Score) scale, which is a subjective measure of voice quality. G.729 Annex A (or G.729A) achieves an MOS score of 3.7. This is usually considered an acceptable trade-off in order to achieve half the CPU requirements. The G.729 and G.729A bit streams are compatible and interoperable, but not identical.

Annex B of the G.729 vocoder (or G.729B) works in conjunction with both the standard G.729 vocoder and the its reduced complexity counterpart G.729A. G.729 B provides for voice activity detection, background noise modeling, comfort noise generation, and silence frame insertion. Before going into the overall operation of G.729B (from a black-box point of view), it is necessary to define a few terms.

Voice Activity Detection (VAD): G.729B uses a very sophisticated algorithm to distinguish between speech and non-speech signals. G.729B monitors statistics such as amplitude, zero crossing rate, and spectral characteristics. Based upon the current and past statistics, the VAD algorithm decides whether or not speech is present.

Background Noise Modeling: In a telephone system, it is not desirable to mute the background noise when speech is not present. On the other hand, it is not necessary to occupy 8 kbps to represent background noise. When the VAD algorithm decides that voice is not present, the G.729B background noise modeling algorithm is used to model the background noise in a G.729-like format, but with a reduced bit rate of 1500 bps. Furthermore, the background noise modeling algorithm determines if there has been a change in the noise characteristics when compared with that of the previous frame. If there has been no appreciable change, no data needs to be transmitted over the communication channel.

Silence Frame Insertion: The silence frame insertion is perhaps a misnomer, I'll use the term anyway. As stated in the previous paragraph, there are frames during which no data is sent across the channel. Under these circumstances, the G.729B decoder continues to generate comfort noise based upon the most recent noise model.

Comfort Noise Generation (CNG): The CNG is the algorithm that is used to generate comfort noise. The CNG expands the lower rate noise modeling data into a standard frame of G.729 data by filling in

some of the less significant parameters. It then performs G.729 synthesis to generate the comfort noise.

Here's how the whole thing works. The G.729 (or G.729A) encoder receives uncompressed PCM samples at a frame rate of 10 milliseconds. Given a sampling rate of 8000 samples/sec, a frame consists of 80 samples. The VAD determines whether or not speech is present.

If speech is present, the standard compression algorithm is used, resulting in an 80 bit output packet.

If speech is not present, the background noise modeler models the background noise. If the noise characteristics have changed, a 15-bit silence identifier frame (SID) is transmitted. If the noise characteristics have not changed, the background noise modeler indicates so, and no bits need to be sent across the channel.

The G.729 encoder therefore transmits one of the following frame types:

- Voice (80 bits)
- SID (15 bits)
- Null (0 bits)

The G.729 decoder receives one of the aforementioned frame types. If a voice frame is received, the G.729 synthesis algorithm regenerates the speech. If a SID frame is received, the G.729 decoder updates its noise generation parameters and generates comfort noise. If a Null frame is received, the comfort noise generator continues to generate noise based upon its most recent noise statistics.

Note that at least one SID frame must follow a speech spurt before the first null frame. This ensures that the correct noise model is used by the CNG at the decode side.