VOICE-ACTIVATED AGC FOR TELECONFERENCING

Peter L. Chu

PictureTel Corporation, MS 635
222 Rosewood Dr.
Danvers, MA 01923, USA
chu@pictel.com

ABSTRACT

In a group teleconference, sound pressure levels at the microphone vary due to variations in speaking distances and loudness. The conventional Automatic Gain Control (AGC) appropriately adjusts the microphone gain so as to achieve a constant level. However, due to the lack of discrimination between speech and nonspeech sounds, nonspeech sounds can cause the gain to increase, giving rise to "noise pumping" or big increases in the background noise when no one is talking.

In this paper, an AGC is described which uses a speech detector that examines stationarity of pitch and energy to discriminate speech against sounds like air conditioning, doors closing, footsteps, etc. For any gain increase to occur, a positive speech detection must be declared, thereby eliminating any "noise pumping" artifacts.

1. INTRODUCTION

This paper deals with an algorithm which automatically adjusts the gain of an audio channel so that the speech is the same volume all the time. Existing AGC (Automatic Gain Control) algorithms are defective because they try to equalize the volume for all sounds- speech, chairs moving, background noise, walking sounds, etc. What is really desired from an AGC in a telecommunications environment is constant volume on speech only. Other sounds have no significance to the parties participating in the call.

Conventional AGC will increase the gain when a person stops talking, resulting in amplification of the ambient noise. When a person starts talking again, the gain is decreased. The constant change in background noise level is quite annoying and is called "pumping".

In section 2 the voiced segment detector will be described. The voiced segment detector detects human voiced speech (vowels sounds, as opposed to consonants like "sh" which have no periodicity), discriminating against non-human sounds like doors closing, footsteps, finger snaps, paper shuffling, etc. The voiced segment detector must yield a positive detection for any gain increase to occur. In [1], examination of the smoothness of pitch contours was similarly used to discriminate speech against nonspeech sounds for speech recognition purposes. In section 3 the control logic of the AGC gain control is discussed. Finally, in section 4, results are summarized.

2. VOICED SEGMENT DETECTOR

The voiced segment detector works by examination of variances in pitch and energy versus time. It is assumed that voiced speech will have a smaller variance in these two parameters over time than nonspeech sounds. Some nonspeech sounds clearly will violate this rule, such as a musical instrument or a ringing telephone, but it is assumed that the instances of these nonspeech sounds occuring in teleconferencing is small enough so that their false detection is permissible.

Nonspeech sounds can be classified into two broad categories:

- Constant background noise from fans, computer hard disk drives, electronic circuit noise. The spectra of such sounds doesn't change appreciably over time.
- Noise whose spectra changes in nature over time, such as finger snaps, footsteps, paper shuffling, doors opening, etc.

When the signal consists only of the first type of noise, its presence can be determined by simply comparing the spectra of the signal with that derived from a background noise estimator.

When the signal consists of the second type of noise, the lack of similar periodicity in the spectra and energy over time is a good discriminator of these sounds against real speech.

0-7803-3192-3/96 $5.00©1996 IEEE

The signal is first split up into blocks of data whose length must be long enough to include at least two pitch periods of speech. Within each block, an FFT is taken. Via the FFT, the spectral shape, the degree of periodicity, periodic frequency, and energy is found. If a sufficiently periodic sequence of frames occurs with sufficiently small variance of pitch and energy for the frames, voiced speech is declared as present, and the gain is allowed to increase.

The blocks in figure 1 corresponding to the Voiced Segment Detector shall now be described individually in each of the following subsections.

2.1 Window

The signal is sectioned into 20 millisecond frames and smoothly windowed so that no artifacts will arise from discontinuities introduced at the beginning and end of the frame. The resulting windowed frame must be long enough to encompass two speech pitch periods. The window's length is 32 milliseconds (512 samples, 16 kHz sampling rate used, 7 kHz audio bandwidth), requiring 12 milliseconds of signal from the previous frame. This length window will allow the detection of human speech with a pitch period as low as 62.5 Hz. Most males have a pitch somewhat higher than 80 Hz, with the mean at about 100 Hz.

2.2 FFT

The FFT is taken of the frame, producing spectra of the frame. The implementation's FFT size is 512 points.

2.3 Background Noise Estimator

The Background Estimator finds an estimate of the underlying, stationary, constant background noise signal in the signal. A number of different schemes could work here, but a simple one would involve finding the energy in 1 kHz bands for each frame, and then checking to see if these energies do not change appreciably over a second or so of time. If there is little change, the averaged spectrum of these frames is declared an estimate of the background noise.

2.4 Signal vs. Noise Detector

The spectrum of the current frame is compared against that of the background noise estimate. If the spectrum of the current frame is sufficiently different from that of the background noise, then signal is declared present. Otherwise, background noise only is declared present, and the Voiced Segment Detector declares this frame as not having human speech in it, eliminating the need to execute any remaining blocks of the Voiced Segment Detector.

2.5 Magnitude Squaring and Spectral Noise Subtraction

Autocorrelation of a sequence is the most common method for detecting and estimating periodicities in a signal, and is the technique used in the Voiced Segment Detector.

One of the central theorems of signal processing is that convolution in the time domain is equivalent to multiplication in the frequency domain. Thus the autocorrelation of a sequence (which is equivalent to the convolution of a sequence with a time-reversed version of itself) is equivalent to multiplying the Fourier transform of the sequence times the complex conjugate of the same Fourier transform, and then taking the inverse Fourier transform. The self-multiplication operation of the Fourier transform is equivalent to squaring the magnitude of each frequency component,

$$S_k = \text{ real }^2(k) + \text{ imag }^2(k), \text{ for } k = 0,1,2,...256 \quad (1)$$

for the 512 point FFT, where S_k are the magnitude-squared components of the FFT. To eliminate any artifacts in the computation of periodicities due to the background noise, the background noise's estimated magnitude-squared components are subtracted out,

$$M_k = S_k - cN_k, k = 0,1,2,...,256 \quad (2)$$

where N_k are the magnitude-squared components of the Background Noise Estimator, and M_k must be non-negative so that if $M_k < 0$ then M_k is set to zero in value, and c is a fixed constant. If c is too large, some of the non-noise will be set to zero, whereas if c is too small, some of the noise will cause errors in the periodicity detection and estimation.

2.6 Highpass Filter to Emphasis Spectral Peaks

The periodic components of the signal will give rise to spectral peaks in the magnitude. By feeding the $M_k, k = 0,1,2,...,256$ through a highpass filter, these peaks will be emphasized. The slowly changing component of the spectra will be eliminated. This operation is analogous to "whitening" the spectra, a technique which generally has been found to be a useful preprocessing step before performing the autocorrelation in pitch estimation. The highpass filter operation is

$$H_k = M_k - .5(M_{k-2} + M_{k+2}), \text{ for } k = 2,3,...,254 \quad (3)$$

where $H_0, H_1, H_{255}, H_{256}$ are set equal to zero, and H_k must be non-negative, so that if $H_k < 0$ then H_k is set equal to zero in value.

2.7 IFFT

The inverse Fourier transform is taken of the H_k derived in section 2.6. The H_k represent the real part of the kth frequency component while the imaginary part of the kth frequency component is zero. The inverse Fourier Transform produces an approximation to the autocorrelation sequence of the highly processed signal frame. It is an approximation only because of the ommission of appended zeroes to the signal frame to correct for circular convolution artifacts. However, the windowing that occurs in section 2.1 greatly minimizes the circular convolution artifacts, eliminating the need for appending zeroes, which is a significant computational burden.

2.8 Peak search

The peak of the autocorrelation sequence is found for lags appropriate to the range of human pitch. If the ratio of the peak value to the autocorrelation for lag 0 is sufficiently high, a periodic frame with pitch period equal to the lag is declared. The autocorrelation with lag zero is simply the energy of the processed frame, and autocorrelation values of other lags will be smaller in value. The pitch is stored in a buffer.

2.9 Standard Deviation of Recent Pitch Periods

If the current frame contains pitch and a sufficient number of the last several frames contained pitch, the standard deviation of the pitch period values of all the pitch-containing frames is found. If this standard deviation is sufficiently small, then there is a declaration of voiced speech being present. The standard deviation of the energy of these frames must also be small, to guard against the false detection of resonant objects being tapped.

3. GAIN CONTROL LOGIC

The purpose of the gain control logic is to control the gain the same way a human would in a teleconference situation. The control logic is rather ad hoc but has been found to give satisfactory performance in most teleconferencing situations.

3.1 Find Peak Energy

The maximum energy of all 20 ms frames in the last 2 seconds is found.

3.2 Reduce Gain if Peak Energy High

If the peak energy is greater than that necessary for a comfortable volume, decrease the gain so that the output will be at a comfortable volume, regardless of any declaration of the Voiced Segment Detector.

3.3 Increase Gain if Peak Energy Low

If the peak energy is much less than that necessary for a comfortable volume, increase the gain so that the output will be at a comfortable volume, if and only if the Voiced Segment Detector indicates that the current signal is from human speech. Otherwise, do not increase the gain.

3.4 Limit Maximum Gain

At no time should the gain be such that the output noise will be at an uncomfortably loud level. The background noise level energy determines some maximum value for the gain.

3.5 Decrease Gain if No Speech

If there are no human sounds for several seconds, the gain should be decreased to the point where the background noise energy at the output is at a fairly low, unobtrusive loudness.

4. SUMMARY

The voice-activated AGC adds the concept of allowing increasing audio gain based on decisions of a voiced speech detector. The voiced speech detector is computationally efficient because of its reliance on FFT's for which highly optimized assembly code exists on nearly all processors.

The algorithm has been found to be useful in real products, and a patent is pending on the concepts.

5. REFERENCES

[1] Hidefumi Kobatake, Katsuhisa Tawa, Akira Ishida,"Speech / Nonspeech discrimination for Speech Recognition System under Real Life Noise Environments", **Proceedings ICASSP 1989**, Glasgow, Scotland, pp. 365-368.

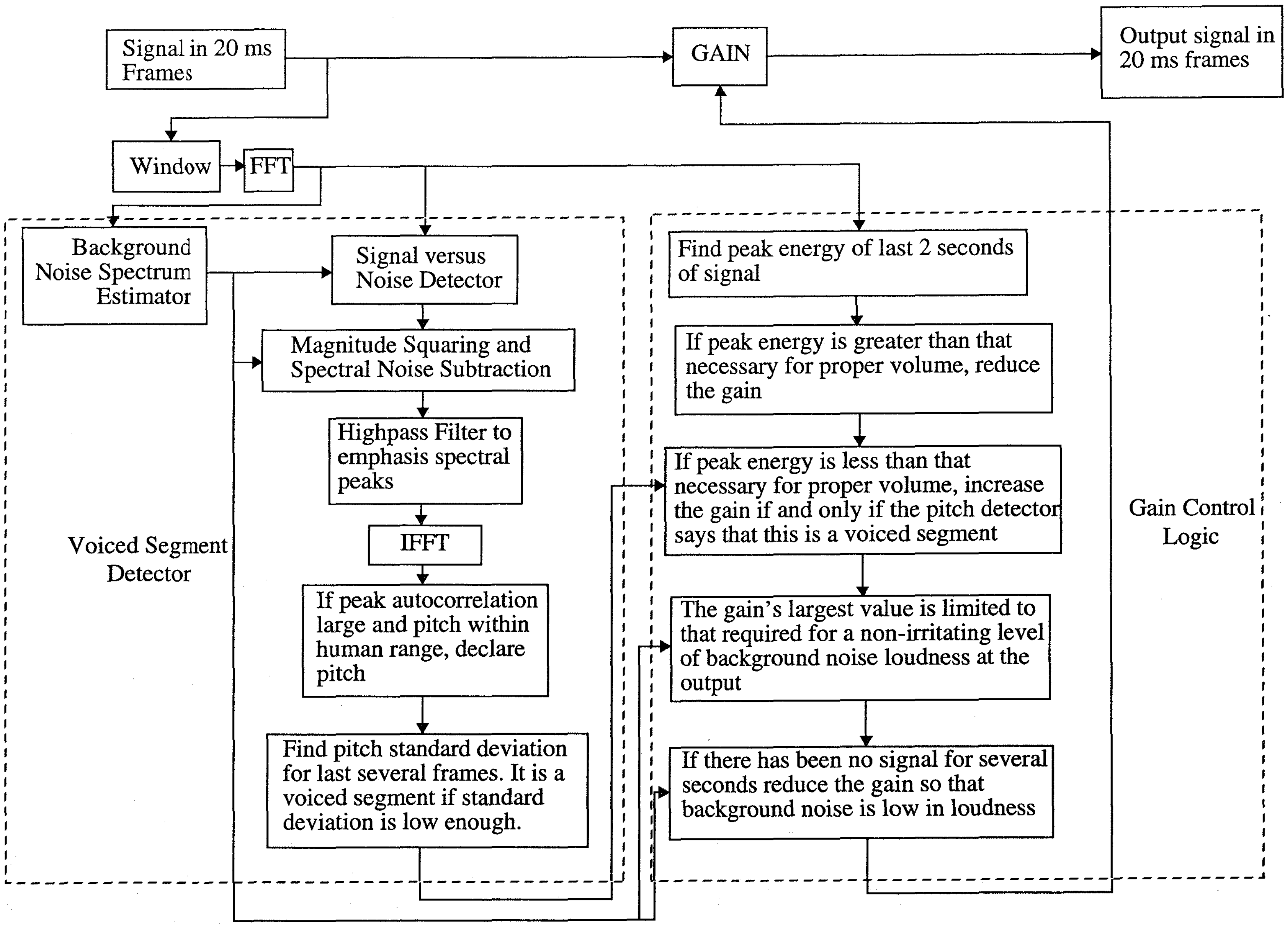
Signal in 20 ms Frames
GAIN
Output signal in 20 ms frames
Window
FFT
Background Noise Spectrum Estimator
Signal versus Noise Detector
Magnitude Squaring and Spectral Noise Subtraction
Highpass Filter to emphasis spectral peaks
IFFT
If peak autocorrelation large and pitch within human range, declare pitch
Find pitch standard deviation for last several frames. It is a voiced segment if standard deviation is low enough.
Voiced Segment Detector
Find peak energy of last 2 seconds of signal
If peak energy is greater than that necessary for proper volume, reduce the gain
If peak energy is less than that necessary for proper volume, increase the gain if and only if the pitch detector says that this is a voiced segment
The gain's largest value is limited to that required for a non-irritating level of background noise loudness at the output
If there has been no signal for several seconds reduce the gain so that background noise is low in loudness
Gain Control Logic

FIGURE 1. BLOCK DIAGRAM OF VOICE-ACTIVATED AGC