Lorenzo Cioni

Saturation and time domain parameters

(oral presentation, 17th International Congress on Acoustics, Rome, September 2-7, 2001)

1 Introduction

1.1 The topic of my talk will be the saturation of speech signals and its effect on time domain parameters.
1.2 I will present some preliminary (mainly) theoretic results.
1.3 The basic idea is very simple: starting from some synthesized and natural speech signals artificially saturated deduce quantitative criteria (the so called “diagnostic cues”) so to determine if a speech signal underwent saturation, which kind of saturation and the degree of the saturation.
1.4 The main aim is to quantify the reliability of the time domain parameters that can be extracted from such speech signals.

2 The practical side of the problem

2.1 Saturation can occur in many points during the acquisition/analysis process of speech signals:
- during the recording phase (map task and other linguistic games, interviews, conversations);
- during the acquisition phase (bad settings on the audio board);
- during analysis (wrong operations on the signals).

2.2 Saturation of speech material is an unsolvable problem whenever the material cannot be recovered:
- unavailability of the speakers (death, distance);
- difficulty of finding the right speakers (variety of language i.e. dialect, age, cultural backgrounds, family and personal characteristics);
- single copy speech recordings;
- original materials corrupted or lost.

3 The theoretic side of the problem

3.1 Mathematical models

Three mathematical models of saturation as non linearities transforming an input signal x=x(nT) into another signal y=y(nT) are proposed

clipping
$y = kx\,[u(x+x_0) - u(x-x_0)] + y_0\,[u(x-x_0) - u(-x-x_0)]$

two’s complement
$y = k\,(x-jx_0)$ if $(j-1)x_0 \le x \le (j+1)x_0$ $j = 2m$, $m \in Z$

zeroing
$y = kx\,[u(x+x_0) - u(x-x_0)]$

T = sampling period;
x_0, y_0 define the range of linearity;
u(x) = 1 if x ≥ 0 and u(x) = 0 if x < 0;
k=y_0/x_0 (usually we have k=1);

3.2 Description

Clipping occurs whenever speech samples are substituted with a (positive/negative) constant value and has physical plausibility.

Two’s complement occurs whenever there is an overflow that cannot be managed by the system. For instance if you represent integers in two’s complement using three bits the following situations can arise:
100 + 111 = 1011 but with three bits you have 011 and so you get -4-1 = 3 instead of -5,
011 + 010 = 101 since a leading 0 is lost on three bits and so you get -3 instead of 5. Also two’s complement has physical plausibility.

Zeroing occurs whenever samples that exceeds a certain threshold are substituted with null values. It is a more theoretic model and I never found it in practice, at least within the scope of speech processing.

3.3 Time domain parameters

RMS

$$RMS(x) = \frac{\sqrt{\sum_{i=1}^{N} x_i^2}}{N}$$

N is the amplitude of the frame, the frame shift is N/2 and we use low values of N (about one pitch period) because we do not want the graph be too smooth.

ZCR

$$ZCR = \frac{Zf_s}{2N}$$

Z is the number of times a signal crosses the x-axis, f_s is the sampling frequency and N is amplitude of the frame (see above for comments).
Short time autocorrelation

$$R[k] = \sum_{n=0}^{N-k-1} x[n]x[n+k]$$

k is known as lag and varies from 0 to a value that includes at least two pitch periods and N is again the frame amplitude. R[k] is usually represented normalized by dividing it with R[0] so that it varies between -1 and 1.

VOT

Voice Onset Time is the duration of the burst from the offset of an occlusion to the periodic onset of the following vowel.
VOT can be estimated using a spectrogram together with the graph of the speech signal.

RT

Rise Time has been defined as the time interval from the onset of frication to the maximum amplitude of frication.

Both VOT and RT are correlated with measurements between significant cues on the graph of the speech waveform.

3.4 The use of the time domain parameters

The starting point is a set of normal/artificially saturated synthetic signals, normal/artificially saturated natural signals from which extract the time domain parameters.

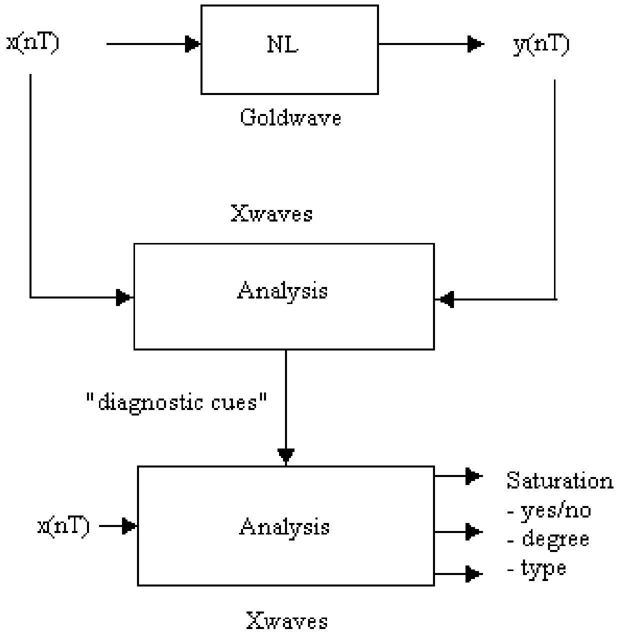
The next step is a comparison among such parameters so to define a set of “diagnostic cues” and, finally, apply such of “diagnostic cues” to new signals to estimate the reliability of the time domain parameters that we can extract from such signals.

The reference frame is provided by a set of well known properties of time domain parameters so that any anomaly in their values and/or behavior can be seen as a cue that something went wrong with the source speech signal.

The underlying hypothesis is that what went wrong is that the source signal underwent saturation.

3.5 The experimental setting and the involved programs

Experimental setting



Programs

Goldwave.

Goldwave is used both to record natural speech utterances, generate synthetic signals and to change their dynamics so to simulate the effect of the three types of saturation.

Xwaves.

Xwaves is used to perform a time domain analysis of the various signals and of their saturated counterparts so to deduce the “diagnostic cues” of which at point 3.4.

4 Synthetic signals

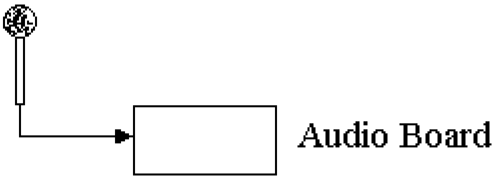
1) Fundamental frequency: $F_0 = 70$ Hz;
2) Sinusoidal waves at multiple integer frequencies of F_0;
3) Signals as linear combinations of sinusoidal waves at various frequencies and amplitudes;

Use of non linearities to change signal dynamics and simulate saturation.

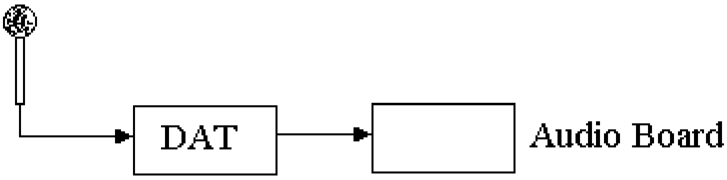
5 Natural signals

Modes of acquisition:

1) Direct connection of a microphone to an Audio Board



2) Acquisition chain:
- microphone → DAT
- DAT → Audio Board



Technical data:

Sampling frequency: 44.1 kHz
16 bit/sample
single channel (mono)

Materials

First utterance and sections

R7	R7sec1	R7sec2	R7sec3
R7cl	R7clsec1	R7clsec2	R7clsec3
R7tc	R7tcsec1	R7tcsec2	R7tcsec3
R7ze	R7zesec1	R7zesec2	R7zesec3
0-1.866	0.241-0.299	0.858-0.915	1.201-1.259

Second utterance and sections

R8	R8sec1	R8sec2	R8sec3
R8cl	R8clsec1	R8clsec2	R8clsec3
R8tc	R8tcsec1	R8tcsec2	R8tcsec3
R8ze	R8zesec1	R8zesec2	R8zesec3
0-0.99	0.030-0.087	0.301-0.359	0.8-0.858

Use of non linearities to change signal dynamics and simulate saturation.

6 *The diagnosis of the speech signals*

Once the “diagnostic cues” has been defined and quantified some way they can be used to diagnose new speech signals.

The diagnosis starts with an examination of the speech waveform and is carried on by listening sections of the signal, displaying the corresponding graph, evaluating the time domain parameters and taking measurements of such parameters till a decision can be taken by comparing the measured values with the reference frame of the “diagnostic cues”.

A major problem occurs whenever a signal has suffered a low degree of saturation so that by listening to it you cannot perceive any sort of distortion and by displaying the graph of the signal there is no evident cue that the graph has an anomalous behavior. In such a case the availability of quantitative criteria can be of help in deciding about the reliability of the signal.

7 *Conclusive remarks and future projects*

“So many things to do and so a short time..”:

7.1 Theoretic aspects must be examined more closely so to quantify the “diagnostic cues”;

7.2 The “data base” of both natural/synthetic normal/artificially saturated signals must be extended so to make more precise the “diagnostic cues”;

7.3 The relations drawn with Goldwave to modify signal dynamics are not perfect so that we are planning to design ad hoc filters to be used in conjunction with Xwaves;

7.4 Diagnosis and cure: the use of LPC techniques to patch saturated signals?