Tutorial about the VQR (Voice Quality Restoration) technology

Ing Oscar Bonello, Solidyne
Fellow Audio Engineering Society, USA

INTRODUCTION

Telephone communications are the most widespread form of transport of audio signals. Its frequency range, however, is very restricted. At the beginning of this technology, it was due to technical limitations of carbon microphones and headphones with metal diaphragm. As technology improved and the transducers have good audio quality, the telephone communication did not include a significant increase in its frequency response. This was for economic reasons, to reduce the required bandwidth. While it is proven that a 100% of intelligibility needs 5.000 Hz of audio band (French-Steinberg, 1947) the telephone industry accepts a slight loss of intelligibility in order to reduce the bandwidth to 3400 Hz and thus be able to handle more channels in the same bandwidth.
Low frequencies (bass) are also cut below 300 Hz in order to reduce the cost of transducers and also eliminate the interference of ambient noise like traffic or machines.

How does this affect the transmission of news for broadcasting ? It is clear that the phone was not designed as a system of high quality audio. So the voices transmitted (even if it goes directly to the telephone line, bypassing the transducers), are doubly limited in frequency, both in bass as in treble. The transmission of frequencies between 300-3400 Hz is a very important limitation. The use of modern cellular phones is even worst due to the use of a reduced bandwidth.
The human voice has frequencies from 80 Hz to 10,000 Hz therefore lose two octaves on bass and two treble octaves. This produces a voice decidedly *metallic* (lack of bass). This forces many radios to invest in UHF links with big antennas and truck mounted facilities. This however takes away their mobility and prevents journalists of working quickly because they must travel followed by an expensive team of technicians.

But the limited frequency response is not the only problem, because phone communications have a reduced dynamic range. Recall that we call dynamic range of the relationship between the loudest sound and background noise. The human ear operates in normal life with 90 dB dynamic ranges, being acceptable for FM broadcasting a 60 to 70 dB. Telephone transmissions handle dynamic ranges between 40 dB and 50 dB at best.

EXISTING SOLUTIONS TODAY

The problem of transmitting audio over the phone line has had many attempts at a solution in the past. None of them were full satisfactory. A brief analysis is as follows;

a) Multiple frequency shifters. One encoder divides the audio band into three sub-bands 300-3400 Hz and transmits them using three fixed telephone lines. At the other end (Studies) the decoder shifts it to original band. You get an acceptable sound (50-7500 Hz) but the high cost and difficulty to maintain three simultaneous lines for each event and the inability to use with mobile phones, has made the system fall into disuse.

b) Frequency extenders. The principle is similar to above but uses a single line. Requires a encoder in transmission that shifts all frequencies above 250 Hz (or a similar value) Then a voice at 100 Hz is transmitted at 250 + 100 = 350 Hz At the studio, the decoder does the reverse task running down 250 Hz in order to recover the original band. Please note that we lost 250 Hz at the treble band.

Disadvantages: You need to use two computers (coder + decoder) / not usable for telephone interviews (where the interviewee speaks from its own telephone set) / not correct the treble / no better dynamic range

c) ISDN transmission systems to send phone service for high-speed data, available in Europe. The quality is excellent and this is their main advantage.
Disadvantages: Only usable in Europe and very few places in the world, in major cities / Not 100% portable because it requires physical line / High equipment cost and time of transmission / need to use two computers / no use for phone interviews (where the interviewee speaks from its own telephone set)

d) CODEC Digital Systems In this category falls the Musicam, Patriot Tieline, TELOS, etc. It is an encoder system that converts the digital signal to the MPEG or other compression system, and transmits it through a telephone modem, using a land telephone line or a GSM Modem through the cellular network. The frequency response and sound quality is good. But it has latency (sound delay) that in some cases, such as remote interviews conducted from studio, can be annoying.

Disadvantages: High-cost equipment / It has delay, which prevents interviews from studies / It needs special decoders at studio side

e) Dual transmission systems. Falls in this category the Solidyne MB2400 that simultaneously transmits a digital MP3 streaming and analog cellular for processing VQR, eliminating the delay and drawbacks of conventional Digital Codec

Compare the previous systems with the new system VQR

The VQR encoder is a low cost system that does not require a computer. In fact, improve the quality of the interviews in which the remote person uses its own phone line or cell phone. The VQR has no delay. Restoring frequencies between 50 Hz - 10,000 Hz improves the bass and treble. Increases the dynamic range up to 70 dBA (similar to a good studio microphone).
95% of the listeners believe that the remote caller is at the radio study.

Disadvantages: The sound quality is good but not perfect as in the CODEC Digital. We recommend listening audio demos at Solidyne WEB site.

VQR TECHNOLOGY (Voice Quality Restoration)

The comparison just made between VQR and other solutions seem to indicate that the VQR is a kind of magic solution. But of course this is not true, because it is based on scientific principles.

Solidyne has worked on the issue of telephone transmission for 35 years. We have analyzed several mathematical techniques that could solve this problem. Several prototypes were built and conducted numerous tests. But our solutions had the same problems as mentioned above and therefore we never commercialize it.

But from 2005 we face the problem from a radically different point of view. We use the Psychoacoustic point of view and the analysis of the human voice generation mechanisms.
We were encouraged by the success of the same idea in 1988, when we achieved the world's first system of recording and reproducing music in Hard Disk, based on the invention of data compression (bit compression) using psychoacoustic theory of Critical Masking bands. This allowed us to build the world's first system for data compression that was a precursor of MP3, ATRAC, Dolby and all other systems in use today by millions of people around the world.

But in 1988 it seemed magic.

VQR system, invented by Solidyne in 2005, operates exclusively on the human voice. Let's see then what we know about the speech generation. In Fig 1 we see the phonatory system

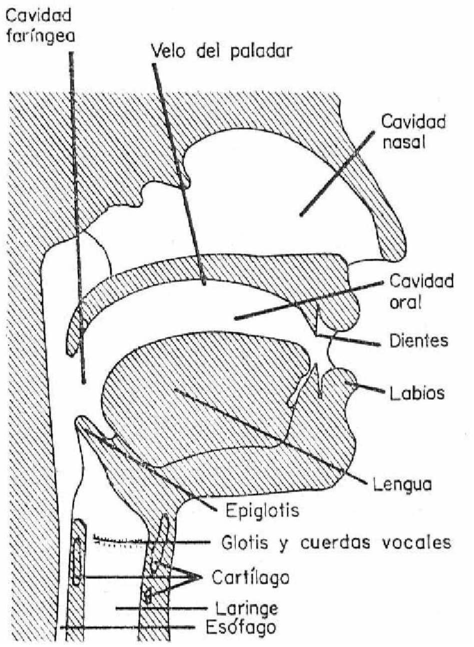

Figure-1

We can see that the vocal cords are at the beginning of the entire chain of generation, making a sound fundamental or pitch that is the base of the speaker voice. Note that the sound goes through three cavities (pharynx, oral and nasal) that behave like three resonators altering the harmonic components of the sound of the vocal cords. The pitch is rich in harmonics due to the fact that it is an asymmetrical triangular wave, with even and odd harmonics . The movement of the vocal cords in the sequence is shown in Fig-2

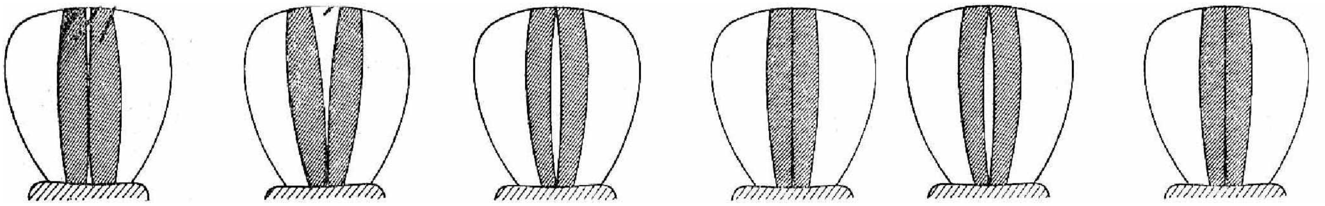

Figure 2 Vocal chord movement (After M.Guirao, 1980)

The acoustic action of the vocal cords has been simulated in the laboratory with two masses connected by a spring (Fant 1970, Veldhuis 1995) is evident from Fig 1 that the engineers can simulate cavities with tubular resonators, thus imitating the vocal tract (O'Saughnessy 1987). See Fig 3

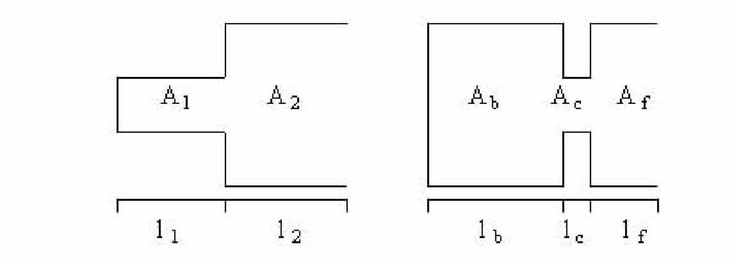

Figure 3 Modeling vocal tract with 2 or 3 resonator cavities

Glottis system (generation of pitch) and three cavities behave as a series equivalent circuit. This hypothesis is now fully accepted (McAulay 1983, Macon 1996, Klijn 1998)

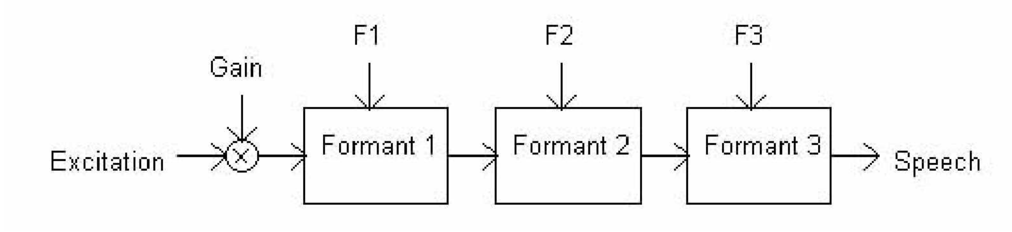

Figure 4 Equivalent circuit f vocal chords and resonators

Each of the filters (they are band-pass second-order) is quickly changed by the movement of the tongue, producing different resonances that are called *formants* of the voice. This is defined in the classic equation of Harvey Fletcher (Allen, 1995) which defines the relationship between the sound pressure for a harmonious Pk with respect to the fundamental P1

$$\frac{P_k}{P_1} = \frac{F_k}{F_1}\sqrt{\frac{\left(\frac{\Delta}{\pi f_1}\right)^2 + \left[1 - \left(\frac{f_0}{f_1}\right)^2\right]^2}{\left(\frac{\Delta}{\pi f_k}\right)^2 + \left[R - \frac{1}{R}\left(\frac{f_0}{f_R}\right)^2\right]^2}}$$

Where delta is the damping factor

Now let's see how looks the audio spectrum that the mechanism of phonation of the human voice presents (Fant (1960), Flanagan (1965), Fry (1979), Lieberman & Blumstein (1988).

Consider Fig 5

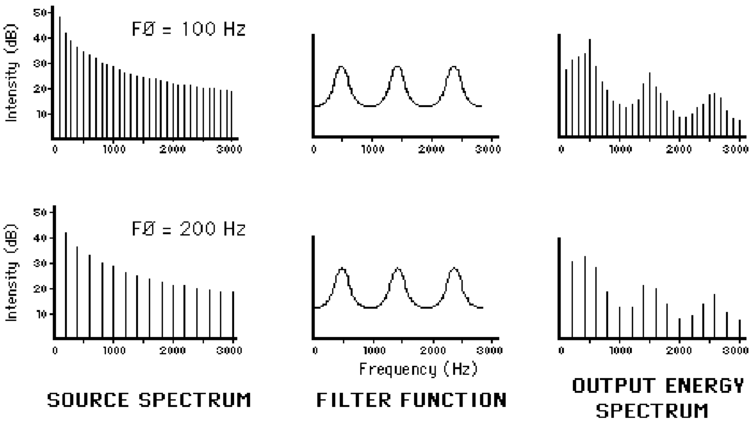

Figure 5 Audio spectrum for 100 & 200 Hz pitch

We see at the first graph the sound of the vocal cords for the case of a fundamental of 100 Hz (male voice) and a 200 Hz (female voice.) It is interesting to note that there are many harmonics that extend to over 3 KHz

In the second graph we see the three filters for the three cavities, or formants. Finally in the third graph we see the final result issued by the human voice. Contrary to what one might think this is a discrete spectrum and not a continuous one. That is, consists of single-spaced frequencies. We emphasize this detail because it is the technology base for restoration of serious VQR.

Let us see this phenomenon in greater detail. Graph each harmonic as a vertical bar and have the Fig 6 and Fig 7

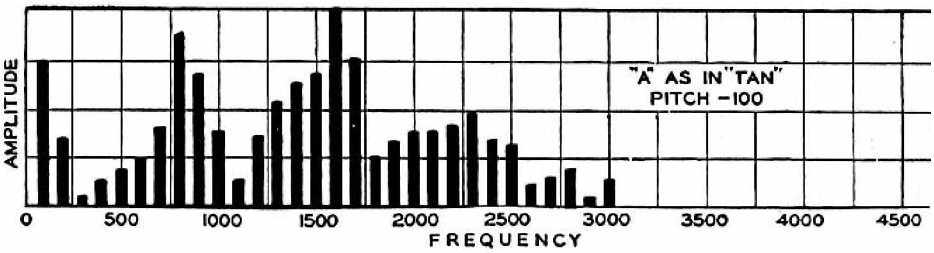

Figure 6; Harmonics of the “A” vocal with male voice (pitch = 100 Hz)

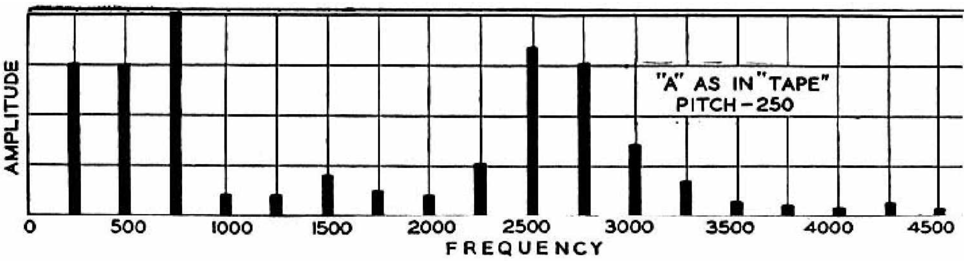

Figure 7; Harmonics of the “A” vocal with female voice (pitch = 250 Hz)

This is observed in great detail the discreet nature of the audio spectrum. This means that even cutting the bottom of the spectrum, below 300 Hz, the original bass frequencies can theoretically be reconstructed. Imagine that an observer examines two spectra between 1,000 and 2,000 Hz and he do not know anything below 300 Hz. Can our observer know what id the original pitch frequency? Clearly he could do it, measuring the frequency separation between two consecutive harmonics (in our example, 100 and 250 HZ).

This is the foundation of VQR restoration of bass frequencies

RECONSTRUCTION OF TREBLE
& DYNAMIC RANGE

As we have seen, human voice can be restored from the discrete components that pass through the telephone channel. This is a remarkable discovery that changes our perspectives on the quality of transmission by telephone or cellphone.

The reconstruction of the treble is also possible, but less accurate than the bass. This is because there are two mechanisms for producing high frequency sounds. One of them is given by high harmonics of voiced sounds that is produced by the vocal cords. But another very important group is produced by non-voiced sounds or fricatives. These are generated when the vocal tract is partially closed somewhere and the air pushes strongly in the same place producing turbulence.

Example of fricative sounds are the consonants f, s and j
Fortunately, these sounds have their fundamental between 2,000 and 3,000 Hz, inside the band of telephone systems. But its harmonics can reach up to 10,000 Hz. Then the telephone loss the brightness characteristic of high quality recordings.
However, a careful study of the harmonic components of the human voice above indicates that if excited with the components of the band of 2-3 KHz a harmonic generator based on asymmetric triangular waves (similar to those generated by the vocal cords) , yields a harmonic spectrum quite similar to actual sound. Then you get a high band complementary to the original sound that can be added to achieve the reconstruction of the high frequencies of the voice

To make the quality of transmitted phone voice as close as possible to the quality we would get in the radio studio, it is necessary to increase the dynamic range from 40 - 50 dBA of telephone calls to 60-70 dBA which is what we get in a good studio. We use a well known system used at modern recording studios.

The solution is psychoacoustic device (again!) It is based on the temporal masking that a strong signal produces in the human ear. The device that allows this is called Audio Expander.
Figure 8 shows the transfer curve of an Expander

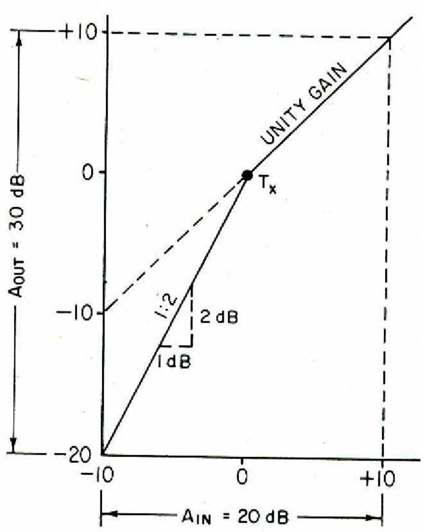

Figure 8 Input / Output graph of an Expander

This is a 1:2 slope expander. This device is absolutely linear. Its gain depends on the signal level. To be effective and meet the psychoacoustic conditions, the gain change when the input signal increases their level, should be very fast, typically less than 5 milliseconds. As the signal level decreases the response time should be about 50-100 mS.
Suppose that at the expander input we have a signal that changes its level 20 dB. At the output the same signal shows a variation of 30 dB, ie it increases the dynamic range.

This same principle is applied in VQR technology to reduce background noise of the telephone, to obtain a very clean audio signal with very low perceived noise.