

State of the Art Speech Processing for Broadcasting

Martin Wolters
Cutting Edge
Cleveland, Ohio

ABSTRACT

Many algorithms for processing speech have been developed over the past few decades including compression, automatic gain control, de-essing and equalization. Today, equipment is available providing only a subset of these functions (compressor, de-esser) or providing a combination of many functions (microphone processor). Sometimes, the same devices are used by engineers in recording and broadcasting, although there are different objectives in each application and different considerations have to be taken into account. Historically, limitations of analog equipment and limited budgets often led to workarounds and very inefficient use of available algorithms. This paper explores state-of-the-art processing of speech signals in a broadcast environment. The advantages of digital processing are described taking into account the interaction between speech processing and commonly used audio processing. Finally, different ways of integrating a digital microphone processor into a broadcast studio are illustrated.

INTRODUCTION

Creating the "sound of the station" has become an important issue in the broadcast industry over the past decades. A number of factors make the aesthetics of sound a key point in a station's format and success. These include, for example, increased competition, improved quality of alternative transmission systems (e.g. cable, DAB), the high quality of new receivers and stereo systems (even for car radios) and the higher expectations of listeners for good sound quality. Using purpose-built audio processing equipment — usually inserted at the very end of the audio chain — is a common technique to create the specific "sound of the station". Most of the time, this audio processing is optimized to improve the sound of the radio station's music format — obviously a very important, sometimes the most important part of

the program. Since the music within a format and therefore the sound of the different songs within a program tends to be quite consistent, one can find that the application of certain processing parameters suffices to establish a station's on-air sound. In this case the raw material fed into a sound processor consists of more or less carefully produced recordings with a certain standard of quality with regard to leveling and equalization.

But announcers', talents' and DJs' voices are also an essential component of most formats. Much of this raw material is produced live, and very often there is no way to maintain the same standard that you can find in the above mentioned recordings. A specific processing of speech becomes necessary and is part of most modern broadcasting facilities. Nevertheless, it seems that the development of speech processing specifically for broadcasting has been neglected during the past decades. The result is little knowledge about how to use digital signal processing most effectively for such applications.

Based on knowledge about broadcasting and sound processing, combined with new scientific approaches about the properties of speech signals and the utilization of digital signal processing, new investigations toward the development of microphone processing products have recently been made. Some of the results are presented in this paper.

ALGORITHMS AND FUNCTIONS

The algorithms used in processing speech are automatic gain control (AGC), equalization (EQ), dynamic range control (DRC), de-essing, phase rotation (PR) and reverberation. Each of these algorithms has a specific task and the order in which

these functions are arranged should not be arbitrary. Figure 1 shows an optimized signal path.¹

Each function will be discussed in the following paragraphs focusing on the specific requirements in a broadcast studio, the advantages of digital signal processing and the benefits of combining these functions into a single unit.

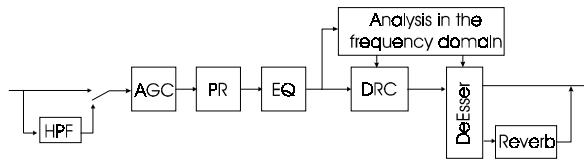


FIGURE 1

Automatic Gain Control (AGC)

One important issue in processing speech signals is level control. In a recording studio, the sound engineer usually takes care of the correct gain settings. The necessary gain is dependent on the room, the choice and position of the microphone and, of course, the person's voice. In a broadcast environment, the same room and the same microphone is used most of the time, so one could adjust the gain by taking into account these two factors. But the person's voice and the person's position might change. This is why an AGC is necessary. From an engineer's point of view, an AGC is a gain controller with a slow attack and release time. Another point of view is to consider the AGC as a replacement for the sound engineer. This latter concept might be more helpful, because one can visualize some important issues:

1. A good sound engineer carefully monitors the input level to make sure that it is nominally at 0 dB (the reference level). A compressor does a similar task; however it changes the level more frequently and more quickly.
2. A sound engineer can "detect" if a person is speaking or not and tries to maintain the desired level of operation when the person is speaking. When the person is not speaking, the sound engineer "freezes" the last gain setting. Therefore, the AGC must take into account the operation of a noise gate² which detects voice activity.

¹ A discussion about the benefits of the particular order is beyond the scope of this paper.

² See the next section about dynamic range control that includes a description of noise gates.

3. A sound engineer not only watches the input level, but also watches the compressor's activity. Adjusting the parameters of a compressor and adjusting the input gain are not independent and, therefore, the AGC and the compressor should interact.

The AGC should be one of the first stages within the signal path. Only a high-pass filter with a very low cutoff frequency (often referred to as a "rumble filter") should be placed before this algorithm. This high-pass filter reduces a possible DC offset introduced by the analog input circuit and filters unwanted noises such as hum, low frequency disturbance from touching the microphone (stand), etc. These signals would otherwise affect the operation of an AGC.

Dynamic Range Control (DRC)

In a superficial view, AGC and DRC appear similar in some ways. This is the reason that compressors — one part of DRC — are sometimes used as AGCs by adjusting the threshold very low so that the compressor provides an almost constant output level, independent of the input level. The result is very poor gain control, since none of the above mentioned issues are taken into account.

There are three new issues addressed by DRC:

1. DRC is used to "optimally use the full amplitude range of a recording system"³. Unfortunately (from a sound engineer's point of view) there are a few high level peaks in speech which reduce the available headroom of a recording. These peaks do not increase the perceptual loudness of a signal because this is affected more by an average value⁴. Hence, reducing the peaks does not decrease the loudness, but increases the available headroom and allows additional gain, resulting in an overall increase in loudness. This is sometimes referred to as peak control and is a more technical aspect of DRC, especially compression/limiting. Carefully chosen parameters lead to inaudible compression, up to a certain amount of gain reduction.
2. Beyond this certain level of gain reduction, compression becomes audible. Fortunately, this "sound" imparted by a compressor — the increased density of the speech signal — can be considered pleasant and is sometimes used to create a specific "sound". This is the more art-related aspect of

³ From *Digital Audio Signal Processing* [1], page 207

⁴ See *Psychoacoustics* [2], page 471

compression; the compressor as a tool for creating the "sound of the station".

3. DRC consists of more than just compression/limiting. A second, lower threshold can be utilized to further reduce all signals below this value. This is called an expander and, if the ratio of the reduction is almost infinity, signals below that threshold are muted and would be referred to as a "noise gate". The idea is that signals below a certain threshold are generally non-speech signals (e.g. background noise, paper shuffling and so forth) and should be reduced. This is especially important during interviews with studio guests or in the case of multiple announcers where a person's microphone is open but that person is not speaking. New research in the field of speech detection (e.g. for applications like mobile phones) led to "intelligent" noise gate algorithms.[3] Rather than just monitoring the energy of a signal, these algorithms utilize zero crossing rate and analysis in the frequency domain to determine if a valid speech signal or a disturbing background noise is present. Digital audio processing allows the implementation of some of these ideas into a microphone processor, resulting in a more accurate noise gate.

De-essing

In the past DRC — especially compression — and de-essing were integrated. De-essing was an extension to a compressor. Since research during the last two years resulted in new information about the properties of sibilants and led to development of new algorithms based on psychoacoustic evaluations, the connection between de-essing and DRC needs to be re-evaluated. An overview of the algorithms and concepts used in the past and an overview about sibilants and the problems in recorded speech introduced by these sounds can be found in a research study from 1998 [4].

To summarize the new information, one should distinguish between detection and reduction of sibilants. Investigations on speech recordings in four different languages showed that a very good and reliable detector for unpleasant sibilants is the psychoacoustic unit sharpness.[5] Figure 2 shows the mean and standard deviation of sharpness calculated for 50 test sentences and for 141 sibilants within these test sentences which were marked as disturbing by at least three of four test persons (experts from the recording industry). A value of 1.2 acum can be utilized to safely detect unpleasant sounding sibilants. Based on a frequency analysis related to the human hearing system, sharpness can be calculated in today's digital signal processors (DSPs).[6]

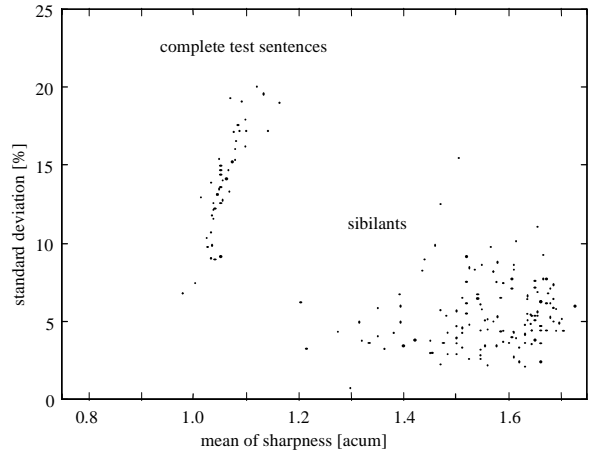


FIGURE 2

A very effective algorithm for the reduction of sibilants without many artifacts can be implemented using a combination of spectral subtraction with a time varying band-pass filter and broadband compression. Time varying means the band-pass adapts to the spectral properties of a specific sibilant. In addition the use of a small amount of broadband compression reduces the so called lisp-effect.[4]

Equalization

There are three different types of filters used in audio and speech processing:

1. High-pass/low-pass filters: As already mentioned, a high-pass filter with a very low cutoff frequency can be used to reduce a possible DC offset, low frequency hum, and background noise. Similarly, low-pass filters can eliminate high frequency noise. In general, these filters are used to limit the audio spectrum. They are less important in controlling the "sound of a station".
2. Shelving filters: These filters are used to weight (boost or cut) certain frequencies, in particular high frequencies above the cutoff frequency and low frequencies below the cutoff frequency respectively. One can create a specific sound of a station using these filters. But it may not be necessary to carefully adjust the parameters for each individual person. A more general approach (maybe separate for male and female announcers) can lead to a successful, good sounding timbre.
3. Peak filters: These filters allow very detailed changes within the frequency spectrum. They also allow changes of any desired frequency. Full parametric peak filters provide control of the center frequency, Q-factor and gain. Used with a low Q-factor, peak filters can be used as a general cut or

boost of the midrange, similar to the effect of shelving filters on high and low frequencies. Peak filters can be used for more detailed changes as well by utilizing a high Q-factor. But these kind of adjustments need to be made for each individual speaker.

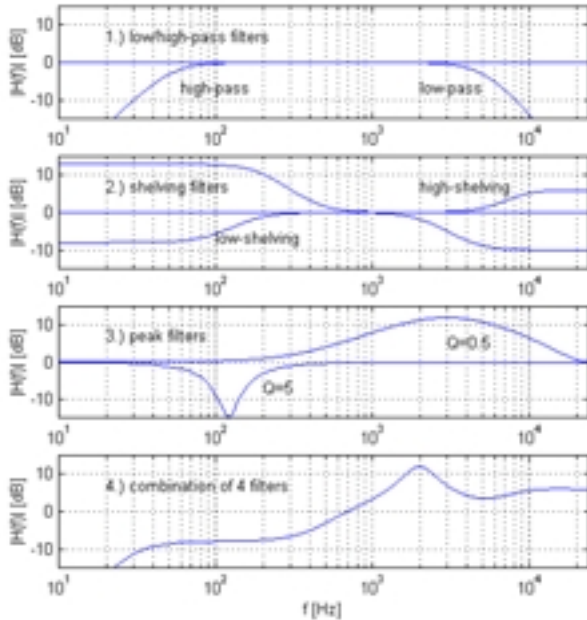


FIGURE 3

Figure 3 shows examples of the described audio filters. The graphic depicts second order low- and high-pass filters, four different shelving filters, peak filters with low and high Q-factors, and a combination of a high-pass, a low-shelving, a peak and a high-shelving filter. This combination simulates a possible EQ-chain in a microphone processor.

These filters, in general, are probably the best known sound processing tools. Rather than review the fundamentals of filters, there are two properties of digital signal processing related to the implementation of filters that will be discussed:

1. Without explaining the reasons and effects in detail, one should know that if DSPs with fixed point arithmetic are used, the quality of filters with low cutoff frequencies can be inferior. Even if this is a problem of fixed point arithmetic in general, there are good sounding, low noise algorithms available. These problems are less prevalent in a DSP with a floating point arithmetic, but even this approach may yield poor filter performance. This means a digital microphone processor that uses floating point arithmetic does not necessarily sound better than a

unit that uses fixed point arithmetic; the best way to test such units is to tune them to low cutoff frequencies.⁵

2. The number of algorithms that can be used at the same time within a digital processor is limited by the computational power of the DSP used. This means, for example, that the number of filters that can be used at the same time is limited. Traditionally, three filters have been a reasonable number for a broadcast microphone processor. However, there are no restrictions to the number of types of filters. Since there is no drawback, a digital microphone processor allows one to use all of the above mentioned types of filters in any combination. Assuming there are three filters available the following combinations could be useful: a) Three peak filters (this combination might need careful adjustments on a per person basis); b) an adjustable high-pass filter followed by peak and/or shelving filters; c) a low shelving, a peak and a high shelving filter⁶ or any other combination.

Artificial Reverberation

Although digital signal processing makes high quality reverberation possible there are still huge differences in the quality of artificial reverberation. This depends significantly on the computational power available — more than any other function described in this paper — and therefore directly impacts the price of a unit. In broadcasting, where artificial reverberation is infrequently used, the highest quality products are not required. For example a detailed adjustment of reverb parameters — such as the kind of surface, size of a room or absorption of higher frequencies — might not be necessary.

However, for broadcasters desiring artificial reverberation, there are two significant advantages in integrating artificial reverberation into a broadcast microphone processor:

1. A specific microphone preset (e.g. the personal preset of an announcer) would contain all parameters, including the settings for reverberation. Anticipating a later discussion in this presentation, it should be mentioned that it is very important to be able to restore settings of all parameters in a quick and easy way. It seems not to be very applicable to store

⁵ See *Digital Audio Signal Processing* [1] for a discussion of these topics.

⁶ In case one uses the peak filter with a low Q-factor this combination might be a good starting point for a general approach.

parameters for a separate reverb processor within a microphone processor. Integrating microphone processing into a broadcast facility includes controlling of reverb parameters and can be accomplished more easily by a built-in reverberation algorithm.

2. The combination of de-essing and reverberation increases sound quality. Some sound engineers in recording studios realized that the unpleasant sound of sibilants in recorded speech is significantly increased by artificial reverberation. They discovered that the problem could be mitigated by the use of two different de-essing units: One that controls the sibilants of the main signal and a second one that controls the sibilants of the signal used by the reverb processor. Integrating a de-esser and a reverb processor into a single microphone processor allows the use of this idea without increasing the cost of the unit. Since the detector for sibilants has to be implemented only once, a specific, advanced reduction of sibilants in the signal used by the reverb algorithm does not require much more computational power.⁷

Phase Rotation (PR)

A function unique to microphone processing for broadcasting is phase rotation. It was invented a couple decades ago to minimize artifacts created by general sound processing in broadcast facilities, especially during clipping. The reason for these artifacts is the asymmetric nature of some human voices.

Figure 4 shows, in the upper left corner, the waveform of a typical asymmetric voice. Although the average over time of this signal is zero (meaning that there is no DC offset), one can see that the peak values above zero are much smaller than the peak values below zero. A clipper limits a signal to an absolute value. The dashed lines in Figure 4 indicate a possible clip threshold. Clipping would affect the two halves of the signal differently. In such a case, clipping produces a more disturbing sound than clipping of a symmetric signal.

The reason for the asymmetry of a voice signal can be found by observing the relation in time of the different formants of a specific phoneme.⁸ The two bottom plots on the left of Figure 4 show the major frequency components resulting in the asymmetric

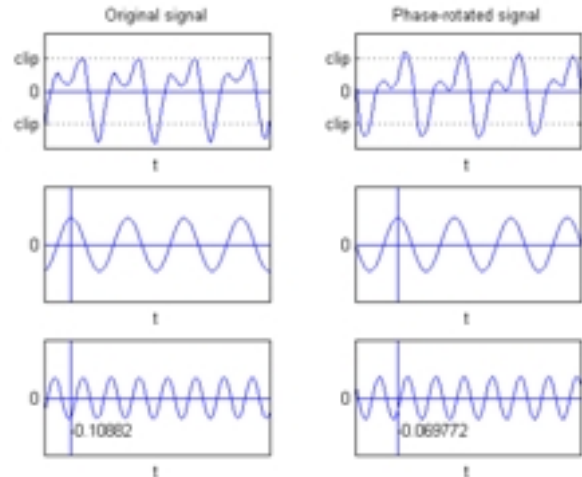


FIGURE 4

waveform. This relation in time is formed by the human vocal tract which can be modeled as an acoustical system of tubes with different lengths and sizes. The dimensions of these tubes are different for different individuals and different phonemes. This can cause an "unfavorable" phase of the frequency components resulting in an asymmetric signal.

Changing the phase of these signals more or less randomly (with an "all-pass" filter) is called "phase rotation" and results in reestablishing a symmetric signal. The right side of Figure 4 shows the processed signal and the changed relation in time of the two formants. Figure 5 clarifies the effect of an all-pass filter on the phase. One can see that the amplitudes of the signal are not affected. These changes of phase are usually not audible, except in the case of very

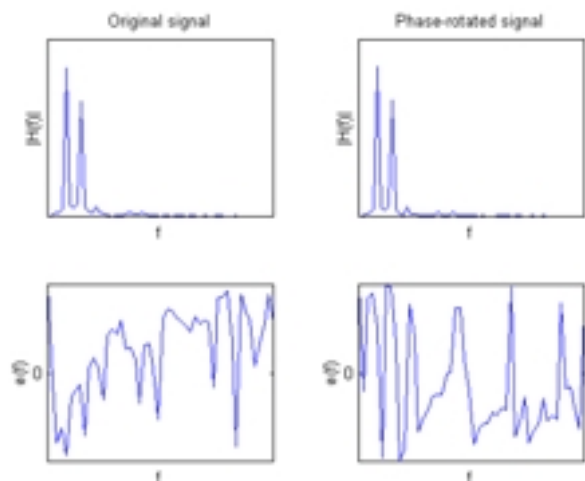


FIGURE 5

⁷ See Figure 1 also.

⁸ The relation in time can be studied by watching the relationship of the phase of each frequency too.

transient signals. The best solution is to implement a phase rotator in a microphone processor and adjust this function for each person individually. In this way, the music programming is not affected and the phase rotator is only used when desired.

The challenge for the sound engineer is how to determine whether to use a phase rotator for a specific person and, if so, how much phase rotation is necessary. One could simply trust his ears. In that case, limiting the voice using a clipper can aid in adjustment. This might not be very accurate but in the end is the most important detector. One could add an oscilloscope to visualize the signal making asymmetric voices easier to analyze. The most accurate and elegant method would be an indicator within the microphone processor. By measuring the peak-to-average level of the positive and negative signal values and comparing these values, a simple but highly effective indicator would help the sound engineer to adjust phase rotation for a specific person.

USER INTERFACE AND INTEGRATION OF A MICROPHONE PROCESSOR IN A BROADCAST FACILITY

Requirements

There are some important requirements on how to integrate a microphone processor in a broadcast facility which affect the user interface of such a device and which are different from requirements in a recording studio. Besides the differing algorithms and functions described in the first part of this paper, the requirements of the user interface are an important reason to design specific microphone processors for broadcast facilities:

- There are generally several on-air and production studios within a broadcast facility. Once the parameters are adjusted for a specific person, it should be possible to use these settings in every studio.
- There is often no technician available. Selecting the correct preset must be very simple so non-technical persons can perform that task.
- Radio stations take their sound very seriously. In most cases the talent should not have access to change parameters capriciously.
- The unit should assist a technician in troubleshooting. Live broadcast requires reliability and, in case of technical problems, a quick way to detect and fix problems.

- A microphone processor should be able to be integrated in an on-air scheduler. That way the selection of correct presets can be automated.
- The microphone processor can be inserted as an effects processor into a mixing console or can be used as a microphone preamplifier as the first component within the audio chain. In the case of a digital studio, the AES/EBU outputs should be able to be synchronized.

An Elegant Solution

Based on the premise that most radio stations are already equipped with a computer network, the following system was designed:

1. The microphone processor itself has a very easy-to-use user interface. The simplest design is appropriate — meaning that the user can only change the preset of the unit but no other parameter. He chooses from a list that is sorted by preset number, preset name or the most recently used presets, allowing a convenient and fast way to find a specific preset.
2. In the case where a fixed preset is required (e.g. guest microphone), the preset can be locked.
3. There are level meters and status LEDs to assist in case of technical problems.
4. A headphone jack allows monitoring without additional hardware (e.g. at a workstation) and assists during troubleshooting.
5. Parameters and presets can be edited using remote software running on a computer. A more sophisticated user interface on this remote application assists the sound engineer when adjusting parameters much better than a necessarily smaller display on the front panel of the unit. The remote software can use different physical connections to the microphone processor such as TCP/IP networks, RS232 ports or other serial connections.
6. In a broadcast facility with more than one microphone processor, the units are connected to the network. A preset management system integrated into the remote application allows for easy distribution of a new or changed preset to each unit. Bigger radio networks can administer microphone processors in different studios from a single place. A security system allows only certain people to change presets and protects the units against unauthorized tampering.

Figure 6 gives an example how the different units are connected to control parameters and presets. Whereas a computer in the production studio might primarily be used to adjust parameters for a specific person,

another computer (e.g. in the office of a station engineer) could run an application for the preset management and other administrative tasks.

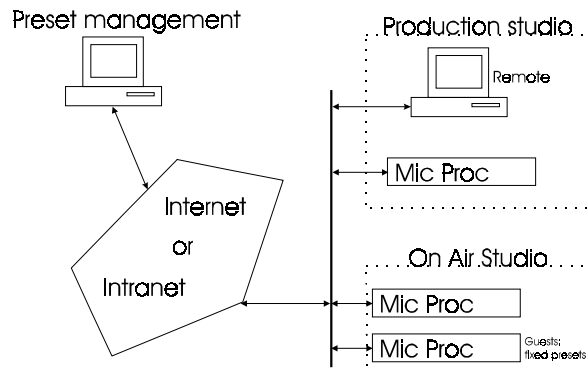


FIGURE 6

CONCLUSION

The algorithms and functions used in state of the art speech processing for broadcasting were summarized. Where digital signal processing can improve these functions, the necessary technical information was given. The benefits of combining all speech processing in a single unit were listed. A summary of the properties of speech signals was added where they explain the goals and reasons of a specific processing function. An overview of the requirements for the integration of a microphone processor into a broadcast facility led to a new approach for a specialized user interface.

REFERENCES

- [1] U. Zolzer: *Digital Audio Signal Processing*, John Wiley & Sons Ltd, Chichester, 1997
- [2] E. Zwicker: *Psychoacoustics*, Springer Verlag, Berlin, 1990
- [3] R.J. Santiago: *A Noise Robust Method for Detection of Endpoints of Speech Utterances*, Master Thesis, Marquette University, 1997
- [4] M. Wolters: *The Acoustical Properties of Sibilance and New Basic Approaches for De-essing Recorded Speech*, paper at the 20th Tonmeistertagung, Karlsruhe, Nov. 1998
- [5] M. Sapp, M. Wolters, J. Becker-Schweitzer: *Reducing Sibilants in Recorded Speech Using Psychoacoustic Models*, paper at the ICA/ASA-Meeting, Seattle, 1998
- [6] M. Wolters, M. Sapp, J. Becker-Schweitzer: *Adaptive Algorithm for Detecting and Reducing Sibilants in Recorded Speech*, 104th Convention of the AES, Amsterdam 1998, Preprint 4677