TUNING YOUR TALENT'S VOICE
TUTORIAL ON SPEECH PROCESSING FOR BROADCASTING

Martin Wolters
Cutting Edge
Cleveland, Ohio

ABSTRACT

Recent developments in digital speech processing have led to major improvements in the quality of speech signals, specifically in a broadcast facility. Technology is now available that combines all speech processing functions into a single unit. This not only allows improved performance due to interaction of algorithms but allows personalized parameter settings for each talent. These settings can then easily be recalled. But individual adjustments of a variety of algorithms are new in a broadcast environment. Therefore this presentation explains in detail the different steps in adjusting today's speech processors and describes the effect on recorded voice.

STEP 1: SETUP OF THE MICROPHONE

The setup and placement of a microphone is not directly related to the adjustment of a digital speech processor. Nevertheless, the basic microphone techniques will be described first. The physical placement of the microphone and the connection to a microphone preamplifier/processor have a significant impact on the sound quality. Unfortunately even digital signal processing cannot always countermand the negative effects introduced in this early part of the signal chain. Therefore, one has to provide a correct setup and placement of the chosen microphone to guarantee the best sound quality.

The microphone should be connected using a balanced connection. This reduces the noise in the audio signal specifically in a broadcast environment. To reduce the occurrence of pop sounds, acoustical filters can be used (*pop screens*). The noise introduced by the microphone stand can be reduced by using an *elastic suspension* and by preventing a direct connection between the microphone and the work table.

Once the microphone is connected, the analog signal needs to be amplified. The amount of amplification depends mostly on the kind of microphone and the person's voice. The necessary gain typically varies between +20 and +60dB. Today, the digital speech processors will take care of a detailed gain adjustment. (See the paragraph about automatic gain control). Usually only a course adjustment of the analog gain is necessary. In order to protect the AD converter from clipping, analog limiters are used. These limiters do not need to be adjusted on a per speaker basis and their operation is inaudible.

Once those analog adjustments are done, one can start to adjust the functions and parameters in a digital signal processor. A state of the art speech processor provides the following functions: highpass filter (sometimes called *rumble-filter*), expander/noise gate, phase rotator, automatic gain control (AGC), equalizer, compressor, de-esser and reverb. Figure 1 gives an overview of a common signal chain.

In order to simplify the setup process, one should first turn off all processing functions.

STEP 2: REDUCING BACKGROUND NOISE WITH FILTERING AND EXPANDING

The common background noises in a broadcast environment are hum, noise from air conditioning, and noise from people in the recording room (moving papers, etc.). Whereas hum is introduced by bad cabling and could be prohibited by a correct connection between the different sound processing equipment, filtering of the audio signal can also significantly reduce hum. The fundamental frequency of hum is usually between 50 and 70 Hz. The lowest frequencies in human voices are typically above 100 Hz. Therefore filtering with cut-off frequencies below 100 Hz reduces the unwanted hum without affecting the voice signal. Notch filters result in the best reduction of this category of noise. Air conditioning noise can also have a strong fundamental frequency,

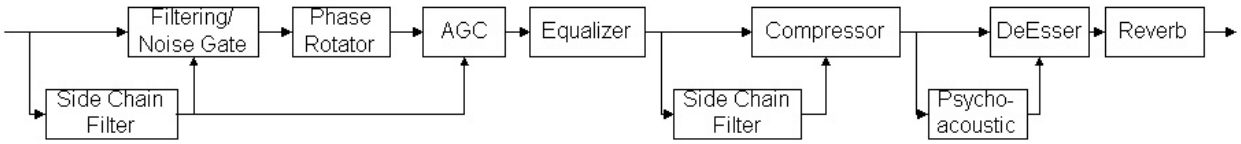

Figure 1: Typical signal chain in a digital speech processor for broadcasting

but usually has a broader spectrum. In the latter case, a highpass filter may result in better noise reduction.

Noise is mostly recognized during speech pauses. In addition, environmental noise such as the shifting of papers, tends to occur during speech pauses. With the help of an *expander* (or *noise gate*) the audio signal is reduced during these pauses. These speech pauses are generally detected by monitoring the input level. If the level drops below a certain threshold, the expander will begin applying attenuation to the audio. This threshold should be chosen carefully, based on the surroundings in which the voice talent is speaking (for example, a studio versus a dance club) and the speaker's style of speech. A high threshold, although useful for isolating the speaker's voice from ambience, can cause a "chopping" effect. Should the talent be speaking quietly, this effect often cuts off the end of sentences, or completely mutes the voice. The amount of attenuation can be adjustable (expander) or infinite (noise gate).

Care should be taken when adjusting the time constants of the control signal. These constants determine how fast the input signal is reduced after a speech pause is detected (*release time*) and how fast the original gain is recovered after a speech signal is detected (*attack time*). The optimal attack time should be as fast as possible (only a few milliseconds). If the attack is too fast, the expander may improperly release the attenuation when ambient noise suddenly exceeds the expander threshold. In addition, disturbing noise will be introduced by the expander. If the attack time is too slow, the beginning of sentences and phrases may be missed, as the attenuation applied by the expander is still in the process of muting the audio.

If the release time is too slow, ambient noise may not be attenuated quickly enough. A fast release time can make the audio sound like a half-duplex speaker phone, with the sudden muting of the audio after each phrase. Generally speaking, a fast attack time and medium-slow release time will achieve the desired effect. Usually multiple iterations of the threshold, the time constants and the amount of reduction are necessary to get the desired result.

The performance of an expander/noise gate can be increased when the side chain signal (the signal that is used to calculate the control signal) is filtered. Since the frequency range of the human voice is limited, the side chain signal can be reduced for example to 150 - 3000 Hz. This allows a more accurate separation of speech pauses.

STEP 4: AGC - AUTOMATIC GAIN CONTROL

The AGC is designed to change the input gain in a way that the overall signal level of the recorded voice is nearly constant. The AGC can reduce or increase the signal level. In order to be inaudible, the AGC uses long time constants (up to a few seconds). Its operation can be compared to the work of a recording engineer who, among other things, monitors the input level and adjusts the input gain slowly in case the level changes.

To setup the AGC first switch off the expander.

To reduce the possibility of clipping, the attack time should be faster than the release time. Fast attack and release times can be useful for creating a very uniform output level, but may cause an unpleasant "sea-sickness" effect on the listener, as the level constantly rides up and down. The slower time constants will reduce the audible artifacts of the AGC, but will obviously work slower, resulting in inconsistent levels.

One of the inherent problems with a typical AGC is that when the audio level drops very low (during pauses or periods of silence), the AGC will increase the gain to its maximum. When the audio comes back, the huge boost will cause the audio to be incredibly loud, and often clipped. For this reason, a *freeze gate* as part of the AGC algorithm is necessary. The detection of the speech pauses is similar to the pause detector in the expander. If the signal level drops under a certain threshold, a speech pause is detected and the last gain setting is freezed. Usually this *freeze threshold* is about 6 to 10db higher than the expander threshold.

A high freeze threshold is good for speakers who maintain a constant level, but for dynamic speakers a high threshold will not adequately boost the level during quiet speech periods. A low threshold is much better in that case because it will not disturb the normal speech, but will still prevent the AGC from

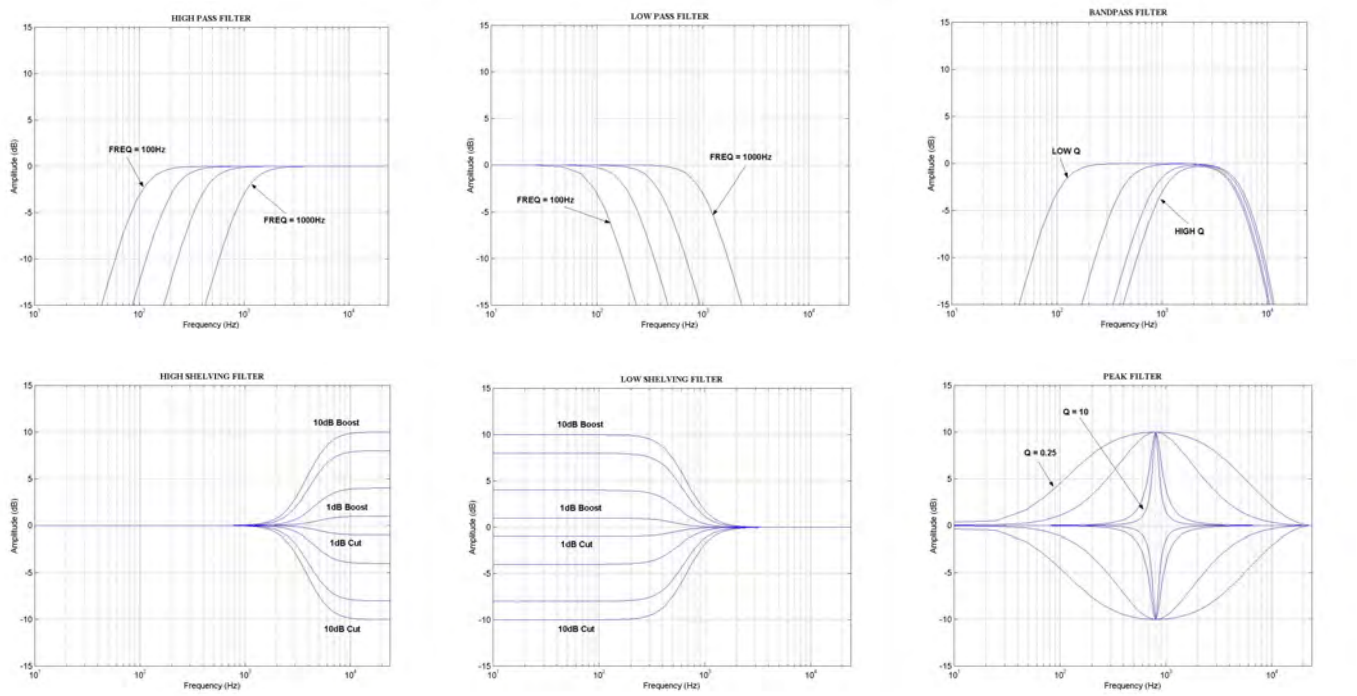

Figure 2: Common filter characteristics

boosting the noise floor and then distorting the normal audio.

As for the expander, a bandpass filter in the detector chain can increase the performance of the freeze function.

STEP 5: EQUALIZATION

Adjusting an equalizer section is a very subjective operation. Other than the highpass or notch filter in the input section, these equalizers are used to create a certain sound. State of the art speech processors should provide at least three independent, full parametric[1] equalizer sections with a variety of selectable filter characteristics. Figure 2 gives an overview of the most common filter characteristics.

It might be very difficult to adjust an equalizer section in a way that a pour microphone sounds good, but it is not impossible. It certainly is much easier to make a good sounding microphone sound bad. Therefore one should be careful using extreme parameter adjustments, unless a certain effect such as a telephone voice[2] is desired.

A note on the use of the different filters: a common misconception among users attempting to “fix” their sound is to constantly *boost* certain frequencies. A better technique in certain cases is to *cut* the audio level in a frequency range. For example, a user who complains that their audio signal sounds “muddy” will usually boost the high frequencies with a high-shelving filter. This only serves to partially mask the problem. A better solution would be to cut out the troublesome low end with the low-shelving filter. This will reduce the amount of energy in the low-end, rather than push the whole audio signal closer to the clip-point, and possibly run into trouble later in the chain.

[1] A parametric equalizer allows the user to adjust all parameters, namely cut-off frequency, gain, and Q-factor.

[2] The frequency response of a telephone connection is typically limited to a frequency range from 100 to 4000 Hz. Since this is the main characteristic of a "telephone voice" this effect can be simulated by a simple bandpass with the mid frequency at 2000 Hz and a very low Q-factor.

STEP 6: COMPRESSION AS A CREATIVE TOOL

Another function to create a certain sound is compression. The *compressor* is used to increase the density of the audio signal by reducing the level of the audio above a certain, adjustable threshold. While the AGC slowly rides the gain of the audio up and down to keep a constant level, the compressor intentionally decreases the audio level that exceeds the chosen threshold, and then applies gain. This has the effect of compressing the dynamic range of the audio, giving a more controlled, tighter feel to the processed audio.

The amount of reduction is determined by the *ratio* parameter. The compressor again has attack and release time constants to control the dynamic operation. A low threshold with a high ratio and fast attack and release time constants results in a more aggressive and most likely audible operation. A very compressed speech signal sounds dense, as if the speaker would sit right next to the listener. Artifacts introduced by the compressor are often described as "pumping" and "breathing". The compressor might reduce the naturalness of the speech in that mode.

The compressor reduces the peak values and thus decreases the dynamic range of a speech signal. Therefore additional gain (often referred to as *boost*) can be added after the compression.

The adjustment of the parameters will certainly need several iterations. The quality of the resulting sound highly depends on the experience and expertise of the sound engineer.

In the past, compressors were usually implemented as wide-band compressors (the same gain value is applied to all frequencies). Modern equipment uses multi-band compression to increase the sound quality. For a speech processor, a two-band compressor seems to be most suitable. The frequencies of the human voice can be subdivided into two major regions. The lower region (up to 1500-2500 Hz) consists mainly of fundamental frequencies and their harmonics (formants[3]). These frequencies characterize vowels and voiced consonants (e.g. /z/ in zero, zoo), which are the loudest voice signals. The upper region is dominated by noise during the production of consonants. Frequencies in this region are much less intense than frequencies in the lower range. The crossover frequency between the two regions depends on the speaker.

[3] *Formants* are the characteristic frequencies of a vowel or voiced consonant.

A wide-band compressor reduces vowels and voiced consonants more than unvoiced speech signals. Specifically after the boost-operation this will result in very loud sounding consonants. The voice will sound harsh. This harshness can be reduced by a de-esser (see next paragraph), but can also be prevented with a two-band compressor. By lowering the threshold in the higher band, the compressor reduces unvoiced sounds as much as the voiced sounds. Thus, the balance between these sounds is not changed, the voice still sounds natural. Even better results can be created with a combined wide-band/multi-band compressor (Figure 3). In addition to possible in-band compression, this algorithm will also slightly reduce the complete audio signal.

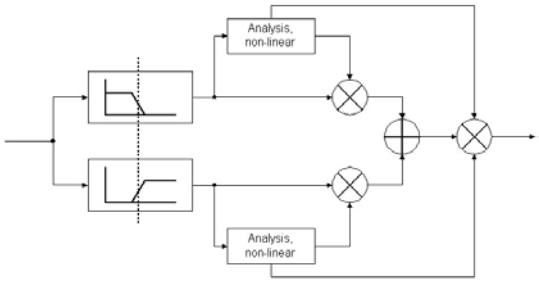

Figure 3: Sophisticated two-band compressor

Another kind of compression algorithm incorporates a filter in the side chain. One example is the use of a highpass filter which will create a very simple de-esser.

STEP 7: DE-ESSING REDUCING DISTURBING SIBILANTS

The *de-esser* is a sophisticated piece of the audio processing chain specifically designed to detect and reduce the amount of *sibilance* in the audio signal. *Sibilance* in speech is typically caused by sounds called *sibilants*. Examples of these sounds are /s/ (e.g. *c*ease), as well as /tʃ/ (e.g. *ch*eese, Tea*ch***)** and /ʃ/ (e.g. *sh*oot, di*sh***)**. In the broadcast environment, the large amount of energy contained in sibilance (because it resembles something close to white noise) can cause undesired effects in the air chain (and are subject to the most distortion during transmission). One problem in the broadcast environment is the *pre-emphasis* applied to the air signal. This can drive the high frequencies contained in sibilance into clipping sooner than the rest of the audio. Multipath distortion can also cause phasing and other distortion to sibilance. This distortion is easily recognized by the listener (as are most distortions applied to speech).

For this reason, a de-esser is a valuable tool in any voice processor used in the broadcast arena. Outside the broadcast environment, a de-esser can be valuable as well. Because of the nature of human speech, sibilants are not as prominent when speaking face-to-face because of the spreading of the audio in space (the human mouth is far from a point source of audio). However, when speech is recorded and played through a loudspeaker (which *does* approximate a point source), the presence of sibilants can be much more unpleasant.

A psychoacoustic-based de-esser recognizes sibilance by taking an FFT (Fast Fourier Transform) of the audio, and makes judgments on the *sharpness* by applying a series of measurements to the level at each frequency (these measurements are based on a model of human hearing). When sibilance is detected, a broadband control signal is applied to the audio to reduce the level during periods of sibilance. This control signal is actually applied through an *adaptive* bandpass filter (meaning the frequency response of the filter changes according to the audio signal).

Recent publications have given a more detailed overview of the kind of sounds that produce disturbing sibilance, traditional de-essing algorithms, and new psychoacoustic-based concepts for reducing sibilance. [1], [2], [3]

Whereas adjusting a conventional de-esser is a rather tedious job, adjusting a psychoacoustic-based de-esser is a very simple task. In a special side-chain-listening mode one can easily recognize if all disturbing sibilance, and only disturbing sibilance, is detected. Thus a threshold value can be adjusted appropriately. A second parameter allows one to adjust the amount of reduction. A trade-off between well controlled audio and noticeable processing exists. If this value is too high, lisping can occur. In the case this value is too low, the sibilance can be disturbing.

STEP 8: ADDING REVERBERATION AND STEREO EFFECTS

Added reverberation gives the listener the impression, that the signal is recorded in a different room than the recording studio. However, depending on the adjustment, the artificial reverberation can sound very unnatural. Besides this more effect-oriented use of reverberation, this algorithm can also be used to create a "bigger", but still natural sounding, voice. A short reverberation time with a low reverberation level and damping of the high frequencies can increase the "volume" of a voice. This effect is greatened by the pseudo stereo effect created by modern reverberation algorithms.

A panorama slider in the output section can be used as part of the setup for a morning or a call-in show. One talent can be assigned more to the left channel and another to the right channel. This increases separation between both speakers and allows for ease in following a conversation.

AN EXCEPTION: DEALING WITH ASYMMETRIC VOICES

The phase rotator is an allpass filter (meaning no attenuation of the signal at any frequency) with a specific phase response designed to eliminate the asymmetry commonly found in speech.

Asymmetric voices are caused by certain physical dimensions of a person's vocal tract. An unfavorable relationship of the phases of different formants will lead to an asymmetric signal as shown on the left side of Figure 4. When an asymmetric signal is clipped, the results are far more unpleasant to the ear than symmetric clipping. Since clipping is often used in broadcast audio processing, a phase rotator is a useful tool in a microphone processor.

Figure 4 demonstrates the effect of a phase rotator on an asymmetric audio waveform. The left-hand figures are (in descending order):

1. An asymmetric waveform consisting of the sum of two sine waves of different amplitude, phase and frequency.
2. One of the sine waves.
3. The other sine wave (higher frequency with a phase shift).

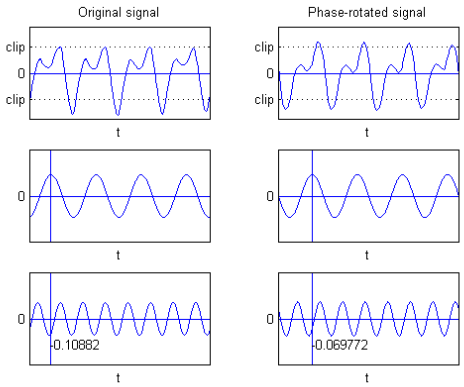

Figure 4: Phase rotation of a speech signal

The right column shows the resulting signals after the original audio passes through the phase rotator. The dotted lines in the top picture indicate a possible clip threshold; you can see the difference in the symmetry of the clipped audio. The bottom sine wave has been phase shifted with respect to the upper waveform.

APPENDIX: REFERENCE LEVELS IN THE DIGITAL DOMAIN

With the introduction of digital signal processing into the broadcast environment, the radio engineer faces a new world of reference levels (Figure 5). The knowledge about these levels is important for a successful use of modern processing equipment and the successful operation of an all digital broadcast facility. Like in an analog broadcast facility equipment connected using digital I/O needs to match a common reference level, otherwise the equipment might not operate as expected or clipping may occur. Clipping may also occur within a digital signal processor when the different reference levels are not taken into account.

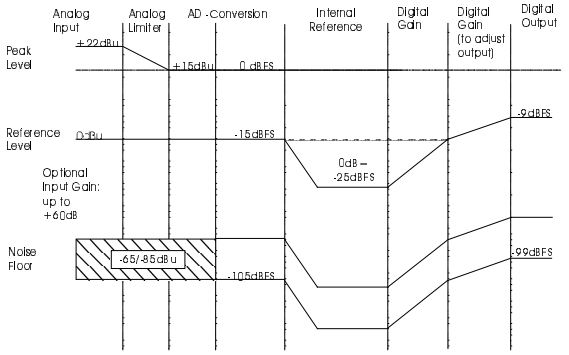

Figure 5: The different reference levels

The analog input section usually expects a nominal level of 0dBu, with a peak level of +22dBu. This can be attained by appropriately choosing the analog gain. The analog limiter will limit the input audio to a peak level of, for example, +15dBu.

After the A/D converter, the +15dBu peak level will become 0dBFS (the maximum audio level that can be represented in the digital domain). The 0dBu analog reference level will be represented by –15dBFS (referred to as the *I/O reference level*).

A certain amount of attenuation will be added to match the *internal reference level*. A typical internal reference level is –25dBFS, which will hereafter be simply referred to as 0dB. All thresholds in the processing chain are referenced to this level. The reason for the lower internal reference is to maintain adequate headroom for each of the processing blocks. The noise floor in a state of the art DSP is less than –140dB, so this low reference level will not affect the audio performance.

The AGC will attempt to keep the audio level at 0dB, assuming an AGC reference level of 0dB. Changing the AGC reference level will change the headroom within the DSP. An increased AGC reference level will decrease the headroom in the digital domain. Threshold parameters such as the compressor threshold are expressed in dB. Setting the compressor threshold to a value of 10dB will result in an effective threshold of -15dBFS in this example (0dB = -25dBFS).

The internal reference level is changed at the end of the processing chain to more closely match the I/O reference. The correction will depend on the processing parameters. The *output level* is adjustable (for both the analog and digital outputs) to provide the correct levels for your particular installation. Typical output reference levels in the analog domain are +4dBu and +6dBu. Reference levels for the digital output vary between -15dBFS and -9dBFS.

REFERENCES

[1] M. Wolters: *The Acoustical Properties of Sibilance and New Basic Approaches for De-essing Recorded Speech*, paper at the 20th Tonmeistertagung, Karlsruhe, Nov. 1998.

[2] M. Sapp, M. Wolters, J. Becker-Schweitzer: *Reducing Sibilants in Recorded Speech Using Psychoacoustic Models*, paper at the ICA/ASA-Meeting, Seattle, 1998.

[3] 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.

[4] Cutting Edge: *ToolVox User's Manual*, http://www.nogrunge.com, March 2000

[5] M. Wolters: *State of the Art Speech Processing for Broadcasting*, Proceedings of the 1999 Broadcast Engineering Conference, NAB, pp. 301-307.