Using Next-Generation Non-Stationary Noise Suppression to Enhance Voice Quality

Emulating the way that the human hearing system characterizes and groups the audio stream into sound sources proves to be an extremely efficient and accurate way to suppress both stationary and non-stationary noise.

By Lloyd Watts, Founder and CTO, Audience

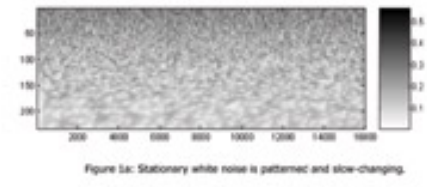


Figure 1a: Stationary white noise is patternec and slow-changing.

[1]

Higher voice quality in noisy environments through next-generation noise suppression technology is proving to be one of the key differentiators for handset manufacturers and service providers. Voice quality, in this case, refers to how well you can understand a conversation in high noise conditions. Environmental noise can make it difficult for a person to be heard, thus limiting where and when subscribers can reliably make calls. Until recently, however, noise suppression technology has focused primarily on removing stationary noise, leaving non-stationary noise sources to degrade perceived quality.

By understanding the difference between stationary and non-stationary noise, engineers will be able to more effectively protect voice quality, bringing substantial benefits to both subscribers and carriers. Subscribers will enjoy greater freedom in that they will be able to talk even in noisy places and not be asked to leave important conference calls or return a call later. In addition, they will have increased privacy by being able to speak softly and still be heard. Not to be underestimated is the increase in handset battery life that results from eliminating codec and bandwidth inefficiencies attributed to the presence of noise.

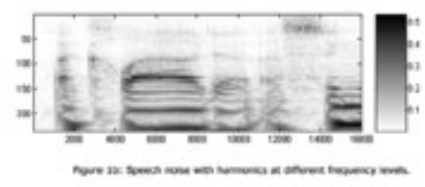


Figure 1b: Speech noise with harmonics at different frequency levels.

[2]

From a carrier perspective, next-generation suppression and higher voice quality will reduce handset returns and customer churn while increasing minutes of usage since subscribers can use their handsets reliably under a wider range of environmental conditions. Carriers will also see more efficient use of network bandwidth through the improved quality of low bitrate voice codecs, leading to significant capital and operational expense savings.

Stationary versus Non-stationary Noise

Traditional noise suppression techniques identify stationary noise sources that are static or slowly changing in frequency or loudness, such as a fan running in the background (see Figure 1a). Because of their constant nature, these sources can be subtracted from the voice of interest using conventional signal processing techniques.

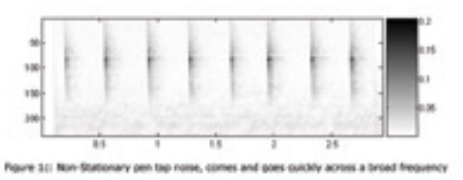


Figure 1c: Non-Stationary pen tap noise, comes and goes quickly across a broad frequency

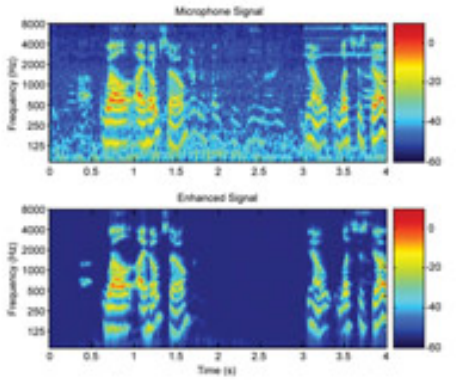
[3]

Non-stationary noise, in contrast, is characterized by rapid or random changes and frequently arises from multiple sources. Examples of non-stationary noise include a person talking (see Figure 1b), background music, and a pen tapping on a desk (see Figure 1c). By the time these sounds are recognized as noise, they have already changed or passed. As a consequence, more sophisticated noise suppression techniques are required.

Alternative Advanced Noise Suppression Technologies

There are three technologies that are addressing next-generation noise suppression, and all three employ two microphones to capture the audio signal. These technologies include Computational Auditory Scene Analysis (CASA), Beam Forming (BF), and Blind Source Separation (BSS).

Blind Source Separation (BSS) uses a simple linear un-mixing technique that assumes that there are at least as many microphones as there are sound sources. However, real-world conditions generally involve a large number of sound sources, requiring an impractical number of microphones for a successful BSS solution. In practice, BSS solution providers often use a simple time-domain Voice Activity Detector (VAD) to overcome this weakness, which limits their robustness and suppression consistency in noisy environments. Blind Source Separation solutions are also sensitive to reverberation and to motion of sound sources.



[4]

With beam forming solutions, the primary speaker is identified within a cone-shaped area of interest emanating the microphone. As a consequence, any distracters that are located within the beam – such as someone sitting or standing behind the user – are not filtered out and will incorrectly be passed through as being part of the voice of interest. Further, beam forming systems tend to have very long convergence times, on the order of 5 to 15 seconds, before the noise suppression engages. Many Beam forming solutions also require a specialized, cardioid (unidirectional) primary microphone which increases system production cost. It has also been found that Beam Forming solutions are very sensitive to microphone mismatch, which can cause serious problems in high-volume production.

While both Blind Source Separation and Beam Forming techniques can suppress noise in some conditions, each has shortcomings which make them poorly suited to production quality cell phones.

CASA Computational Auditory Scene Analysis

Auditory neuroscience underlies a new approach to advanced noise suppression. The science of Computational Auditory Scene Analysis details a methodology based on the human auditory pathway – from the cochlea to the brain stem to the thalamus and cortex. Emulating the way that the human hearing system characterizes and groups the audio stream into sound sources proves to be an extremely efficient and accurate way to suppress both stationary and non-stationary noise.

CASA provides the necessary mechanisms to correctly distinguish a voice of interest from other voices or noise sources. Consider how a person at a cocktail party is able to follow a conversation even when the background is filled with other voices and loud music. What makes listening in such an environment difficult is that most sounds, including voice and music, are comprised of multiple frequencies which make up the perceived quality of the sound. When two or more sounds occur at the same time, the various components from each sound are received by the human ear at the same time or overlap each other.

To solve this problem, CASA grouped together acoustic energy to recreate the original sound sources, based on a diverse list of characteristics, or cues, such as pitch, onset/offset time, space and harmonicity. Pitch, which people use to

differentiate between male and female voices, distinguishes one sound from another based on the harmonics and distinct frequency patterns each sound source generates. Spatial location cues determine the relative placement of sound sources based on their distance and direction, enabling noise sources to be differentiated from the voice of interest. Onset time groups sounds using the principle that when two bursts of sound energy and their corresponding harmonics are synchronized, they are probably from the same source. Employing these cues enable accurate grouping of sounds, and avoids blending of sources that should be perceived as separate.

Instantaneous Noise Suppression

Because non-stationary noise sources can change so quickly, next-generation noise suppression techniques must be effectively instantaneous. Stationary noise suppression techniques fail to suppress non-stationary noise because they must first converge before they can suppress. Fast-acting cues characterize even instantaneous events such as a finger snap, enabling their removal.

A common tool for decomposing the frequency components of sound is the Fast Fourier Transform (FFT). FFTs, however, are based on a linear frequency scale that limits spectral resolution at low frequencies. They also employ a constant frame size and frequency-independent bandwidth which increase processing latency and limit the range of non-stationary noises that can be suppressed. New techniques, such as the Fast Cochlea Transform (FCT) which is based on the human cochlea and operates on a logarithmic frequency scale, provide a more efficient means for processing sounds. By operating continuously rather than in frames, latency is reduced.

CASA based solutions using the FCT have been measured at providing 25 dB of noise suppression across both stationary and non-stationary noise sources. They are much more robust than alternative technologies and deliver consistent levels of suppression regardless of the number of noise sources and whether active speech is present. Figure 2 illustrates the ability of CASA combined with FCT processing to provide non-stationary noise suppression in a severe noise environment.

Advanced Noise Suppression for Handsets

Suppressing both stationary and non-stationary noise effectively and consistently will be a critical differentiator for handset manufacturers and carriers. Through the implementation of next-generation noise suppression techniques, engineers will be able to maintain the highest voice quality in handsets over a wide range of environmental conditions.

Lloyd Watts is the founder of Audience, and as CTO provides ongoing guidance and impetus for the company's core technology direction, as well as the vision of neuro-morphic computing for voice systems. Audience is located in Mountain View, CA, 650-254-2800, www.audience.com.

Links:
[1] http://www.wirelessdesignmag.com/sites/wirelessdesignmag.com/files/legacyimages/0812/Fig1a_lrg.jpg
[2] http://www.wirelessdesignmag.com/sites/wirelessdesignmag.com/files/legacyimages/0812/Fig1b_lrg.jpg
[3] http://www.wirelessdesignmag.com/sites/wirelessdesignmag.com/files/legacyimages/0812/Fig1c_lrg.jpg
[4] http://www.wirelessdesignmag.com/sites/wirelessdesignmag.com/files/legacyimages/0812/Figure-3_lrg.jpg