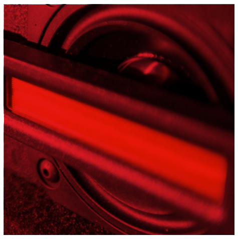
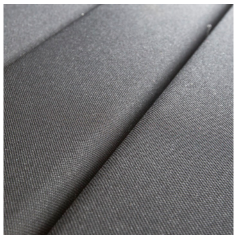
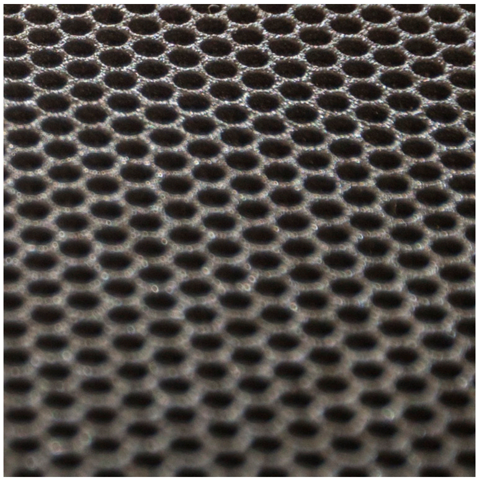
New Best Practices for Speech Intelligibility

Prepared by
SDA Consulting Inc.







innovox
be understood

New Best Practices for Speech Intelligibility

Introduction

For the last 20 years researchers have been working to develop a computer interface that could recognize speech and replace the keyboard. This effort has been funded by many sources, including the National Science Foundation, the Department of Defense and Microsoft. The recent crop of voice recognition applications (Siri, Alexa, Google Assistant and Cortana) are a direct result of this research.

Initially computer attempts to recognize speech in a serial progression were unsuccessful. Even with supercomputers, speech recognition algorithms were unable to handle the variations in a talker's accent, pronunciation and rate of speech, which human listeners can do with relative ease.

Performance was even poorer in the presence of competing noise and reverberation. To solve these problems a new approach was undertaken. Research was re-focused toward "reverse engineering" how humans process and comprehend speech. An unexpected side benefit of this "reverse engineering" was the discovery of special processes in our brains, which can be leveraged by a sound system to improve the ease of speech comprehension, especially in acoustically difficult spaces.

Previously, it has been thought that mental processing affecting speech comprehension occurred in the higher brain centers. But brain researchers have recently discovered processes in the mid-brain which have a major effect on our ability to quickly comprehend speech. These discoveries suggest that a new design approach is needed for sound systems, which must deliver the best possible speech intelligibility.

"Best Sound System Design Practices" have traditionally focused on signal quality and uniformity of coverage. This has been predicted using computer modeling and measured using test equipment, without also considering the listener's mental processes. It has been assumed that a sound system with wide frequency response, uniform loudness and low distortion, would be highly intelligible.

We now know that signal quality and uniform coverage are only a portion of what a sound system must deliver for the best speech intelligibility. New multi-sensory research has uncovered a complex relationship between what we see, what we hear and when we hear it, which affects the time we have available to comprehend speech.

The Task of Decoding and Comprehending Speech

The task of decoding and comprehending speech has been compared to passing a timed multiple choice test. That is because speech is a semi-continuous stream of data, delivered at a rate of three to five syllables per second, while the listener has a short (approximately 4 to 7 second) working memory available for mental processing. We have just enough time to decode the information contained in a typical sentence and move it to longer term memory, before new incoming data replaces it.

Often other senses are competing for the same working memory. If, due to this competition, the listener does not have sufficient working memory available for complete speech decoding and comprehension, he is forced to guess. The role of a "Best Practices Sound System Design" should include techniques for reducing this cognitive load and resulting confusion.

Reconciling the Speed of Seeing with the Speed of Hearing

Humans are "wired" to simultaneously use seeing and hearing to improve their accuracy in knowing where things are. This presents a challenge, because seeing and hearing do not operate at the same speed. The speed of light is some 800,000 times faster than the speed of sound in the outside world. But once visual and aural information is inside the brain, aural impulses are processed at about five times the speed of visual.

Due to this relative latency in the brain's internal visual processing, audio can lead what is being seen. Since the brain is "wired" to expend processing time trying to reconcile these differences, it is possible to reduce its cognitive load by adding the correct amount of signal delay into a sound system's signal path.

The Ears Point the Eyes

Most mammals (humans included) possess a special aural/visual space-mapping function which is continually being updated to keep track of where things are in the local environment. Working in the background, it manages situational awareness and commands attention when a change in the sound field may signal danger. When there is an abrupt change in the sound field (like the snap of a twig), the eyes are automatically directed to search for its location.

This function is triggered when special "novelty detector" neurons, located in the mid-brain, sense a change or the sudden cessation of a repetitive background sound. The "potential danger" must then be reconciled with what is being seen. Priority and focus is given to any sound which matches the perceived location of the "danger," while sounds with a lower priority are either ignored or wait in short term memory.

Everything that we hear passes through this mid-brain process as it travels to higher brain centers. This "automatic protection function", which is always operating in the background, may either facilitate or impede the upstream transmission of a sound impulse (including speech), based on how closely its origin compares with what is being seen.

For instance, when the locations of the listener's visual and aural stimuli are closely matched, the time needed to sort out and ignore the effects of noise, reverberation and other competing interference is greatly reduced. But when visual and aural stimuli do not match, additional processing delay is introduced. This can rob higher brain functions of critical time needed for speech comprehension, and result in increased cognitive load.

The listener's natural ability to selectively ignore noise and reverberation, and focus on the person who is speaking, can be significantly enhanced when a sound system design clarifies and reinforces the apparent location of the presenter.

Establishing an Aural / Visual Anchor (Tø)

A new "Best Practice" Sound System Design should incorporate an "anchor loudspeaker" in close proximity to the talker. This will provide a unified point of reference for nearby listeners, and a timing reference (Tø) with which to synchronize and match the sound arriving from any remote loudspeakers. It is very important that the direct time of flight from the talker to the listener be exactly duplicated by a remote loudspeaker.

There are special voice-lift situations, like a courtroom or legislative forum, where there will be multiple "Aural/Visual Anchors" (one associated with the location of each talker's microphone). In these cases, the image of the talker and his voice, serve as the "Aural/visual Anchor".

When using a remote downstream "repeater" loudspeaker, it will need a unique signal path, and a separate synchronizing delay for sound arriving from each microphone in the sound system. Use of a digital signal processor is essential for coordinating and synchronizing all of these multiple signal paths with their appropriate talker location.

Musical Envelopment vs Speech Engagement

The experiences of listening to music and listening to speech are very different. Music is best experienced in an environment where it surrounds or "envelopes" the listener. Sound quality, in combination with an appropriate amount of reverberation are the prime factors needed to support a good musical experience. Knowing the exact origin of the music is secondary.

Speech, however, is most easily comprehended when a "distinct" source can be quickly identified and separated from the ambient sound environment. Speech is more "engaging" when the brain can quickly comprehend it, and is not diverted by trying to process ambiguous data.

Role of the Precedence Effect

In his 1949 paper "The Precedence Effect in Sound Localization" Hans Wallach describes a frequency dependent integration window in which separate sounds will appears to have the same origin. It varies from less than five milliseconds for single clicks to more than 40-milliseconds for sounds with a complex character like music. This is important because many of the transient components of speech have a duration in the range of a single click.

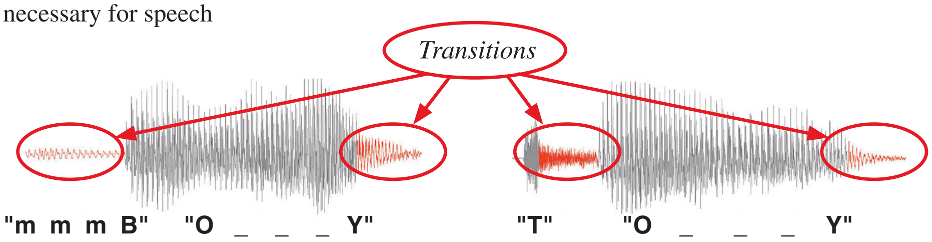
The Precedence Effect has been confused with the Haas Effect (also described in a 1949 publication) which identifies an integration window of about 35 milliseconds but without a frequency distinction

Role of the Syllable

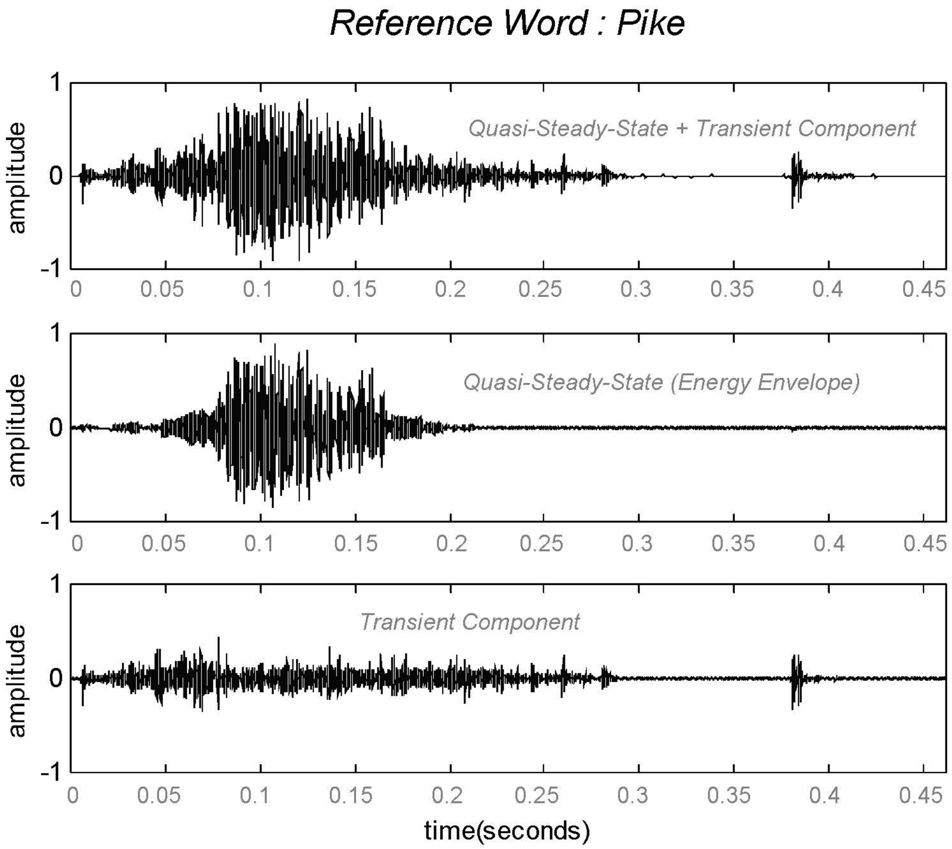
The old “phoneme” model of how we make sense of speech -- "That we process a serial flow of individual speech-stream components (vowels and consonants) like beads on a string" -- has been challenged by recent science.

The syllable is now thought to be the basic organizational unit of speech. Within the syllable, particular components are more critical to intelligibility, while others serve as place holder. Together they form a data packet with an overall energy envelope that is the rhythmic foundation for an intelligible speech-stream.

Equally important are the transient components occurring within the syllable. These short, decisive components serve as “markers” which are used by the brain to help parse syllables with similar sounding energy envelopes (as with “boy” and “toy”). In order to be clearly delineated, they require much better transient response from the sound system than is normally thought necessary for speech



New research finds that one half of speech intelligibility is contributed by the overall energy envelope of the syllable, while the second half is defined by transient components. Though equal in importance, on average, the energy envelope contains up to 90% of the total sound energy, while the "Transient Markers” contain the remaining 10%. (See example below)



Sungyub Yoo, et alia, “Signal Processing Relative Energy and Intelligibility of Transient Speech Components,” 2004 12th European Conference, pp. 1031-1034, IEEE Conferences, .1109/ICASSP.2005.1415052

The challenge for “best practice” design is how to effectively preserve and convey these transient components, thereby enhancing the articulation needed by the brain for best speech comprehension.

A sound system design which provides wide frequency response and uniform coverage for music can be expected to also effectively convey the energy envelope of the syllable. But this does not mean that it will automatically deliver optimum speech intelligibility. To do that, "Transient Markers” must arrive intact and at the correct time, consistent with the perceived location of the talker.

Delivering Face-to-Face Speech Clarity

Face-to-face conversation provides a benchmark for optimum speech clarity. When the talker and listener are only five to six feet apart, all of the requirements that are critical for best speech intelligibility are easily met. As the distance between the talker and listener increases, articulation begins to degrade.

In addition to competing noise, reverberation and specular or “mirror-like” reflections, the enemies of articulation are (1) the non-linear suppression of high frequency transients by air as distance increases, and (2) the masking of "Transient Markers” by lower frequency sound energy (upward masking).

These articulation deficiencies can be managed by using special techniques which “refresh” the "Transient Markers” for distant listeners while maintaining the perceived location of the talker. This starts with the introduction of special downstream “articulation repeater” loudspeakers located along the path between the presenter and the listener. Audio fed to them should be both frequency and amplitude-shaded to compensate for the non-linear suppression of transients by air.

The signals fed to these “repeater loudspeakers” must be delayed to closely match the flight-time of sound originating at Tø. This precision signal alignment is very critical.

An integration window of 5 milliseconds or less should be maintained to lock in the perceived location of the talker. This is much shorter than the 35 millisecond window which is commonly accepted for music program.

Loudspeaker directivity is also a critical factor, because sound energy radiating from the rear of a downstream “articulation repeater” loudspeaker, may not be correctly synchronized with Tø for listeners who hear it coming from behind them. Since this is primarily a problem for frequencies below 700 Hz, it can be dealt with in one of three ways: (1) high- passing the low frequency signals being sent to the extension loudspeaker, (2) using a cancellation element to remove the backward radiation of sound, or (3) using a loudspeaker that has the inherent directivity necessary.

The popular practice of using line array loudspeakers to cover listeners who are more than fifty feet away, works well for music program and can project the energy envelope of syllables for spoken word, but it does not adequately convey the "Transient Markers.” In longer rooms where speech intelligibility is critical, consider adding downstream articulation “repeater” arrays which have been amplitude and frequency shaded.

Transducer Considerations

The transient response of a sound system’s microphones and loudspeakers plays a critical role in capturing and preserving “transient markers”. In a “best practices” application, effective capture of the “Transient Markers” requires a microphone with flat frequency response in excess of 20 kHz, even though 8 kHz is generally considered “good enough” for spoken word. To achieve the most effective reintroduction of the “Transient Markers”, a loudspeaker system should have the rise time and settling time that is currently only delivered by ribbon or air-motion-transformer technology.

Where Loudspeakers Need to Be

The “Anchor” and downstream “Repeater Loudspeakers" need to be in the correct locations in order to establish the "Aural/Visual Anchor", and preserve the signal alignment of the speech-stream elements. This can often conflict with the architect or interior designer’s aesthetic preferences. Choosing loudspeakers of minimal size and unobtrusive shapes will help to appeal to architects and designers and allow speakers to be located where they need to be.

When Are “Best Practices” Appropriate? It Depends

There are basically two different levels of performance, distinguished by the complexity of the verbal content, which must be decoded and comprehended. Simple conversational words of one or two syllables, using basic, familiar vocabulary, can usually be conveyed by properly delivering the energy envelope of the syllable (provided that reverberation and noise are not excessive). This is consistent with traditional signal-quality-based sound system design practices.

Speech in acoustically difficult spaces, involving more complex messages with unfamiliar words, often requires more time for comprehension. To manage this added cognitive load, special attention must be paid to the relationship between what we see and what we hear, and the newly discovered role of transient markers.

Previously discussed “Best Practice Techniques” to enhance articulation, are the preferred solution when complex messages must be understood in acoustically difficult spaces. When clear, unambiguous transients are delivered in coordination with a well-defined aural/visual anchor, cognitive load can be reduced to a level where rapid comprehension is possible.

Summary Conclusions

Successful comprehension of speech requires passing the timed multiple choice test.
The mind needs every chance to make a correct choice. In addition to signal quality,
three newly discovered factors should be considered:

1. Establishing an Aural/Visual Anchor
2. Enhancing and Preserving "Transient Markers”
3. Use of "Articulation Repeater Loudspeakers" in combination with
   Precision Signal Alignment

Direct face-to-face clarity is the ultimate paradigm. When you deliver that, you’ve delivered the speech-stream with the deck properly stacked for the listener.

Bibliography

1. Steven M. Chase and Eric D. Young, “Cues for Sound Localization Are Encoded in Multiple Aspects of Spike Trains in the Inferior Colliculus,” Journal of Neurophysiology, Vol. 99, pp. 1672-1682, April, 2008.

2. Steven Greenberg, “Understanding Speech Understanding: Towards a Unified Theory of Speech Perception,” Proceedings of the ESCA Workshop on the “Auditory Basis of Speech Perception,” Keele University, 1996. pp. 1-8

3. Steven Greenberg, “A Multi-tier Framework for Understanding Spoken Language,” Listening to Speech: An Auditory Perspective, Steven Greenberg and William Ainsworth, editors, Lawrence Erlbaum Associates Publisher, 2006.

4. David Griesinger, “Phase Coherence as a Measure of Acoustic Quality, part one: the … Neural Mechanism, part two: Perceiving Engagement, part three: Hall Design” Proceedings of 20th International Congress on Acoustics, ICA, 2010.

5. Jennifer M. Groh, Making Space: How the Brain Knows Where Things Are, Belknap Press, 2014.

6. Norbert Kopco, I-Fan Lin, Barbara G. Shinn-Cunningham, and Jennifer M. Groh, “Reference Frame of the Ventriloquism Aftereffect,” The Journal of Neu7roscience, Vol. 29, No. 44, pp. 13809-13814, 2009.

7. Dominic Massaro, “Reflections on the syllable as the perceptual unit in speech perception,” http://www.talkingbrains.org.

8. Dominic Massaro, “Tests of auditory/visual integration efficiency within the framework of the fuzzy logical model of perception,” Journal of the Acoustical Society of America, Vol. 108, pp. 784-789, 2000.

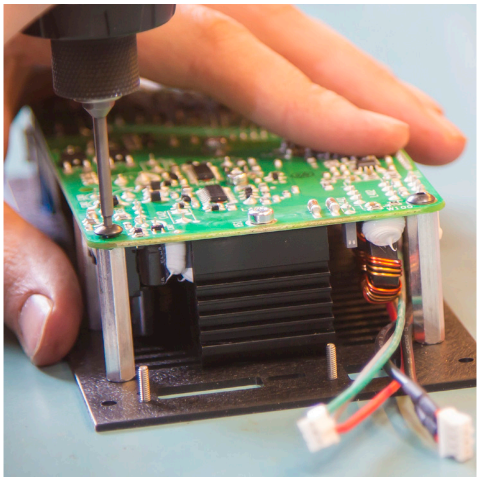
9. David Perez-Gonzalez, Manuel S. Malmierca and Ellen Covey, “Novelty detector neurons in the mammalian auditory midbrain,” European Journal of Neuroscience, Vol. 22, pp. 2879- 2885, 2005.

10. Bert C. Skottun, Trevor M. Shackleton, Robert H. Arnott, and Alan R. Palmer, “The ability of inferior colliculus neurons to signal differences in interaural delay,” PNAS (Proceedings of the National Academy of Sciences), Vol. 98, No. 24, pp. 14050-14054.

11. Hans Wallach, Edwin B. Newman and Mark R. Rosenzweig, “The Precedence Effect in Sound Localization,” The American Journal of Psychology, Vol. 62, No. 3, pp. 315-336, 1949.

12. Sungyub Yoo, "Relative Energy and Intelligibility of Transient Speech Components" 2004 12th European Signal Processing Conference, pp. 1031-1034.

Delivering face-to-face speech clarity.



innovoxaudio.com | 296 Pascal Street North St. Paul, MN 55104 | (651) 209-7660 | info@innovoxaudio.com

innovox
be understood