Method for manipulating listener perceived distance to voice source in speech signals

Table of Contents

Table of Figures

Background of the Invention

Distribution of information between audio and visual modalities is a way to reduce the cognitive overload on users of complex systems, especially if they need to simultaneously monitor and respond to multiple changes in the system. One example is the fighter jet cockpit where the pilot has to monitor many gauges and displays simultaneously and can easily be overwhelmed.

Auditory displays have advanced significantly in recent years, especially in delivering spatial information about the direction from which any sound source emanates.

Many individuals, as well as, commercial and research groups, including the inventor has designed and implemented spatial sound rendering software, for creating three dimensional sound effects for the listener who is either wearing stereo headphones, or is listening to multiple loud speakers. Sound sources used in this rendering can be synthetic or natural voices, music, environmental sounds and noises [1].

Yet, the ability of spatial audio systems to provide robust information about the distance between a sound source and the listener remains limited. In ordinary, free-field conditions, acoustic distance information can be vital, enabling the listener to distinguish among multiple acoustic objects. When headphones or other audio devices are used, this information is usually lost or degraded.

To emulate a distance cue in voice signals, one needs to manipulate the voice to vary the perceived "vocal effort" for that voice. Vocal effort, in this context, can be defined as "the quantity that ordinary speakers vary when they adapt their speech to the demands of an increased or decreased communication distance." Speakers also react to the increased or decreased environmental noise (Lombard effect) similarly to the way they react to increased or decreased communication distance, although one should not assume complete equivalence [3].

The following properties of speech are known to depend on vocal effort level (VEL):

1. Mean and range of the fundamental frequency of speech, F0
2. Formant frequencies, F1 and F3
3. Sound pressure level (SPL)
4. Duration of vowels and consonants
5. Pausing
6. Occurrence of creaky voice
7. Spectral emphasis

Thus, if one can modulate these properties in the correct manner and amounts, one can change the vocal effort level and consequently, the perceived distance to the sound of the voice apparent to the listener.

Each of the properties and how they vary with vocal effort level (VEL) are explored below. The information comes from the Traunmüller and Eriksson study [3]:

1. Mean and range of the fundamental frequency of speech, F0: This value increases with increasing distance. The actual rate of increase depends upon whether a speaker is a male (increases faster) female or child (both increase slower).
2. Formant frequencies, F1 and F3: F1 value increases with increasing distance. The actual rate of increase depends upon whether a speaker is a male (increases slower) female or child (both increase faster).
3. Sound pressure level (SPL): Vocal effort is not equivalent to the sound pressure level. The increase in vocal effort does not fully compensate for the actual distance. For instance, a 6 dB increase is needed for compensating for doubling of distance whereas a 4.6 dB increase is observed. The SPL rule is not fixed, but a function of the desired distance, foremost dependent on whether the speech is conversational or shouted.
4. Duration of vowels and consonants: Vowel durations increase with increasing vocal effort, faster for adults than for children. Consonants don't seem to vary much.
5. Pausing: According to the Traunmüller and Eriksson study, pauses increase significantly for adults for distances beyond 37.5m. Children don't seem to pause for increased distances.
6. Occurrence of creaky voice: For women, occurrence of creaky voice decreases with distance, measured as a percentage of the total duration of the voiced signals. For men, the occurrence increases from very close to 1.5m and decreases beyond that.
7. Spectral emphasis: This is a measure of the energy of the speech below and above 1.5F0mean. This measure is equal to zero when partials above the first are totally absent and it is 3dB when there are equal amounts of energy below and above 1.5 *F*0mean.

Distance Cueing Method of this Invention

There are two categories of approaches to adding distance cues to a particular type of sound source, namely voice: One can either (1) build a speech synthesizer that makes distance cues part of the speech generation process, or (2) develop algorithms capable of adding the distance cue to natural or synthesized speech.

This invention follows the second approach. This alternative is preferable because the marketability of the product using this invention will increase if it can work with actual voice communications or tag this to an existing off-the-shelf speech synthesizer or communications system.

The particular approach taken, involves exploiting the speaker's vocal effort level (VEL).

Vocal effort level zones

One observation one can immediately note regarding vocal effort levels (VEL) of everyday speech is that there are four main zones of vocal effort:

(1) Whispered speech for very close range, e.g., 0-0.3m
(2) Conversational speech for intermediate range, e.g., 0.3m-3m
(3) Loud / shouted speech for increased distance range, e.g., 3-30 m., and
(4) Yelled / screamed speech for a far-off range, e.g., 30 m – 100 m.

This observation leads one to define concentric circles around the listener, as shown in Figure 1, within which a different method is used for distance cueing. Within each circle, one can first render the voice as the appropriate type for the zone (e.g., make it a whisper) and then use the published observations of other experts to vary the listed VEL dependent speech properties. For example, for SPL of the overall voice, one can use the finding cited above that "a 6 dB increase is needed for compensating for doubling of distance whereas a 4.6 dB increase is observed."

The distances above that define the four zones are horizontal distances. Ranges are approximate and can change significantly depending on distances in the vertical direction (namely elevation differences between listener and voice source). Other factors that need to be considered include, but are not limited to:

1. Wind and weather conditions.
2. Acoustic environment, reverberations and reflections,
3. Noise level and type, and
4. Terrain upon which the listener and the sound source are located.

Experimentally determining speech characteristics per VEL

Once approximate VEL zones are defined, the next step is to determine speech characteristics per VEL.

This can be accomplished experimentally. Figure 2 illustrates the experimental process of this invention, which can be used to identify the relationships between the speech signal produced by the voice source (production voice) and speech signal heard by the listener, as the distance between the voice source and listener varies. Moreover, the changes in the characteristics of speech signals (for both the production voice and what is heard by the listener) can be identified both within and in-between VEL zones, using the same procedure. The boundaries of the VEL zones can be refined or the number of zones can be modified using the data from the same process.

The process begins with making recordings in a setup illustrated in Figure 3. Both the production voice and the resulting sound heard by the listener need to be recorded at varying distances of listener-voice source separation. Minimally monoaural recordings

need to be made of the production voice and the signal that the listener hears. Binaural recordings can be made with microphones positioned around the listener's ears.

Note that the production voice is most preferably natural and from a live speaker to capture the VEL natural for the current distance. The person at the production recording site, on the other hand, needs to get a sense of the distance to the listener, which can be established by one or a combination of:

1. Visual distance cues, e.g., relative size of objects observed by the person producing the voice,
2. Auditory cues, e.g., listener or someone standing very near the listener (another live speaker) prompting the person producing the voice.

The next step is to analyze the listener/producer recordings and identify the differences between the two and to identify how each one changes with changing distance between listener and producer.

Analysis of recordings will be in terms of the characteristics that are known to vary with VEL. Seven of these characteristics were itemized in the Background section.

This analysis will yield how VEL-dependent characteristics vary, both qualitatively and quantitatively with listener-source distance. These characteristics can be compared and relationships established between

1. The production voice and the corresponding sound (or in case of binaural recordings, sounds) heard by the listener, and
2. The production voice at distance X and production voice at distance Y to the listener
3. The listener rendition of the production voice while the source is at distance X and at distance Y.

The characteristics can so be established as a function of distance and VEL zone.

Relationships mentioned above can be formulated in terms of transfer functions familiar to those skilled in the art of signal processing. The transfer functions will define how a voices recorded can be transformed from one VEL to another. Transfer functions will be of two primary types, frequency domain and time domain.

Modulating speech characteristics to vary VEL

Once the frequency and time domain transfer functions for each VEL zone and distance within each zone are determined, each characteristic can be manipulated applying one or both types of transfer functions.

That is, one can manipulate an audio stream or file to sound like it is coming from a different VEL zone then it was recorded at. Figure 4 illustrates this process. First the

approximate VEL of an incoming audio stream or file is determined. Next, the desired VEL is used to define the parameters of the frequency and time domain transfer functions corresponding to the desired VEL zone. The transfer functions are then applied to the original audio stream to modulate it to the desired VEL zone.

Potential Applications of the Invention

Adding distance cues can be useful for speech synthesis (or text-to-speech TTS engines). While speech synthesis is very useful for delivering information over phone lines and in a hands-and-eyes-free manner, most synthesized voices, after listening to them for a period of time, become boring and at time annoying, mostly due to their monotonous nature. Adding distance cues can break this monotony.

Another application area is that of communication systems, especially ones where multiple speakers are talking to one listener. For example in a virtual reality program, not only would the direction of which a person is speaking from be known but also the distance the speaker is from the user. This will also help lessen the listener's cognitive overload.

Yet another application area is in speech delivery to high noise environments. Here, one needs to consider that speakers also react to the increased or decreased environmental noise (Lombard effect) similarly to the way they react to increased or decreased communication distance, although one should not assume complete equivalence [3]. One can investigate the usefulness of transforms implemented in improving the intelligibility of speech in high noise environments and modify some transforms as needed.

References

[1] http://www.ic-tech.com/3d_audio/3d_audio_demo.html
[2] *Signals Sounds Sensation*. W.M. Hartmann. AIP Series in Modern Acoustics and Signal Processing. Springer (1998)
[3] Traunmüller, H. and Eriksson, A. (2000) Acoustic effects of variation in vocal effort by men, women, and children. Journal of the Acoustical Society of America, **107** (6)
[4] Brungart, D.S. and Scott, K.R. (2001) The effects of production and presentation level on the auditory distance perception of speech. Journal of the Acoustical Society of America. 110(1), 425-440.
[5] Brungart, D.S. (2000) A Speech-Based Auditory Distance Display. Proceedings of the 109th Convention of the Audio Engineering Society, Los Angeles, CA, September 22-25, 2000.
[6] Brungart, D.S., Kordik, A.J., Das, K., and Shaw, A.K. (2002) The effects of F0 manipulation on the perceived distance of speech. Proceedings of ICSLP 2002, Denver, CO, September 16-20, 2002, pp. 1641-1644.
[7] Naguib, M., and Wiley, R. H. (2001) Estimating the distance to a source of sound: mechanisms and adaptations for long-range communication. Animal Behavior. 62, 825-837.
[8] Blauert, J. (1983) Spatial Hearing. MIT Press, Cambridge, MA.

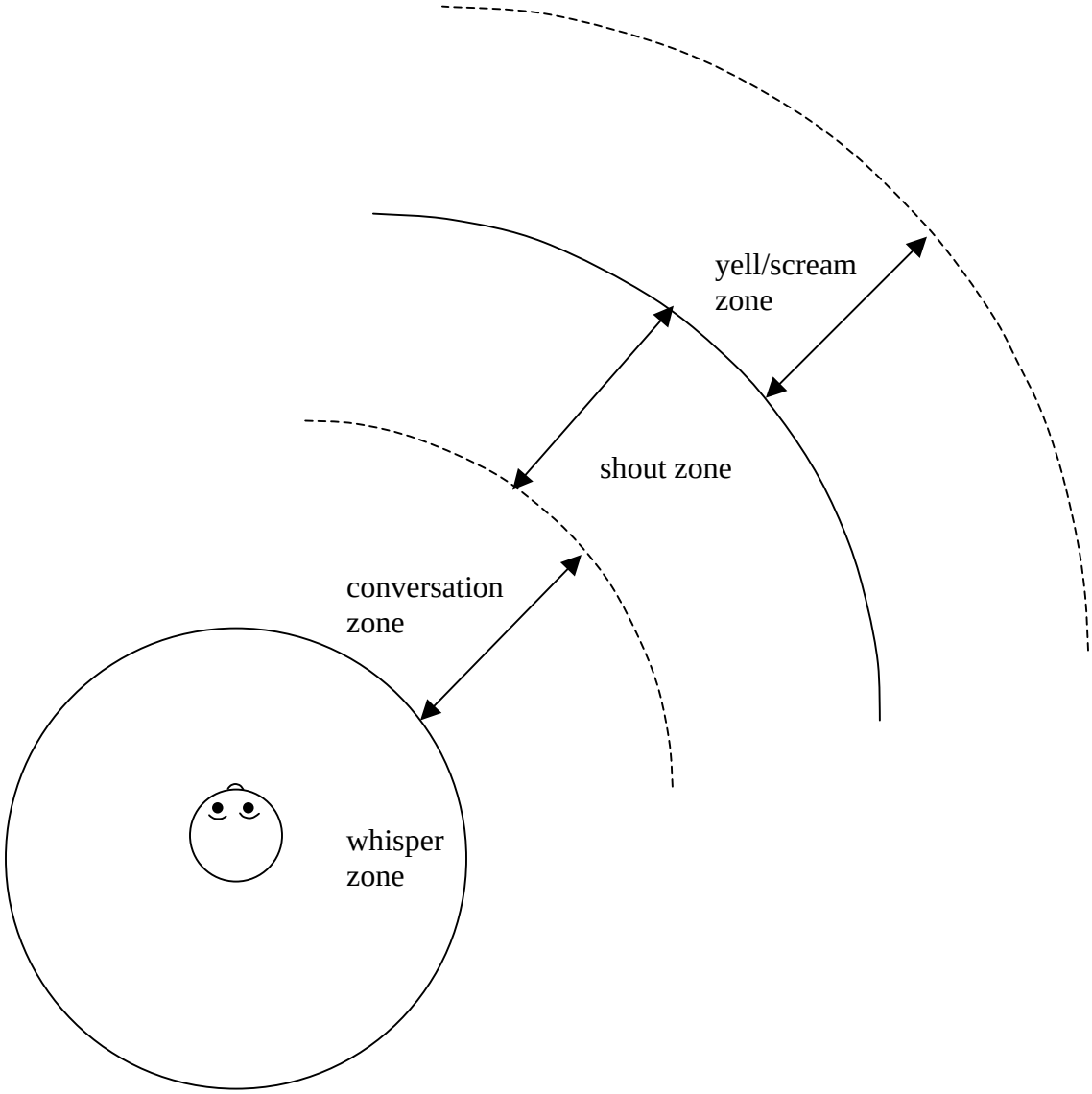

Figure 1 Defined VEL zones around the listener.

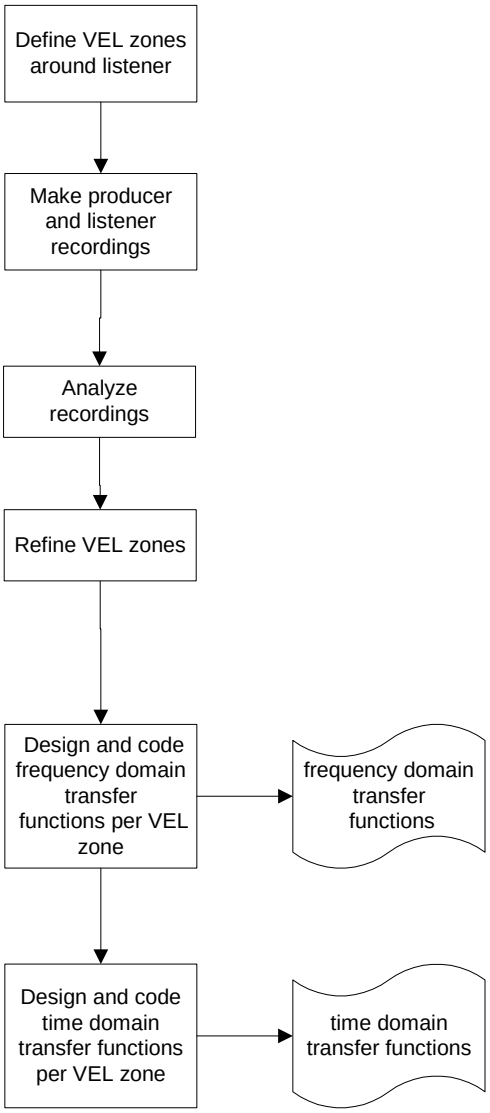

Figure 2 Process of identifying the transfer functions associated with each VEL zone.

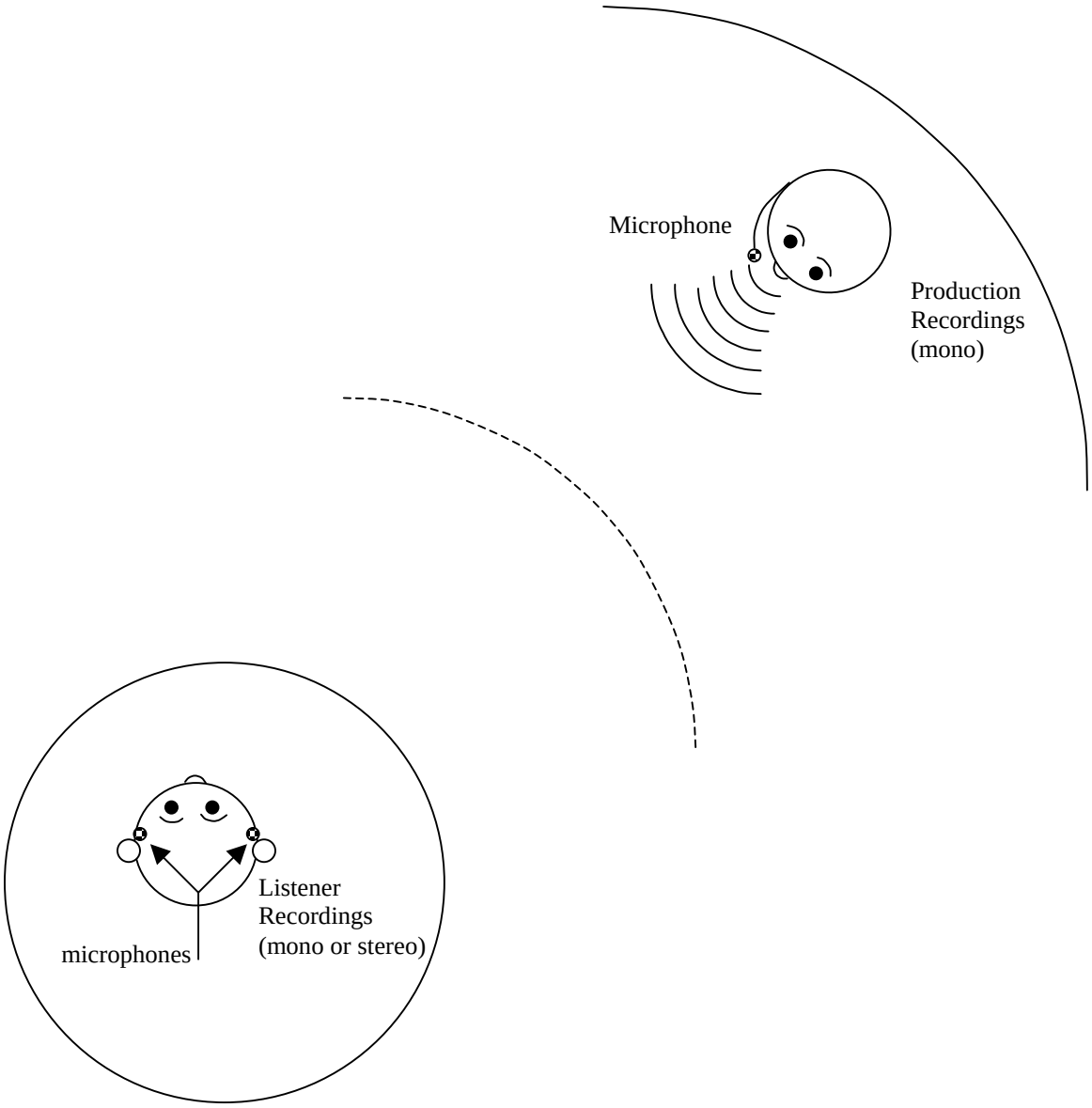

Figure 3 Example listener/producer recording set-up.

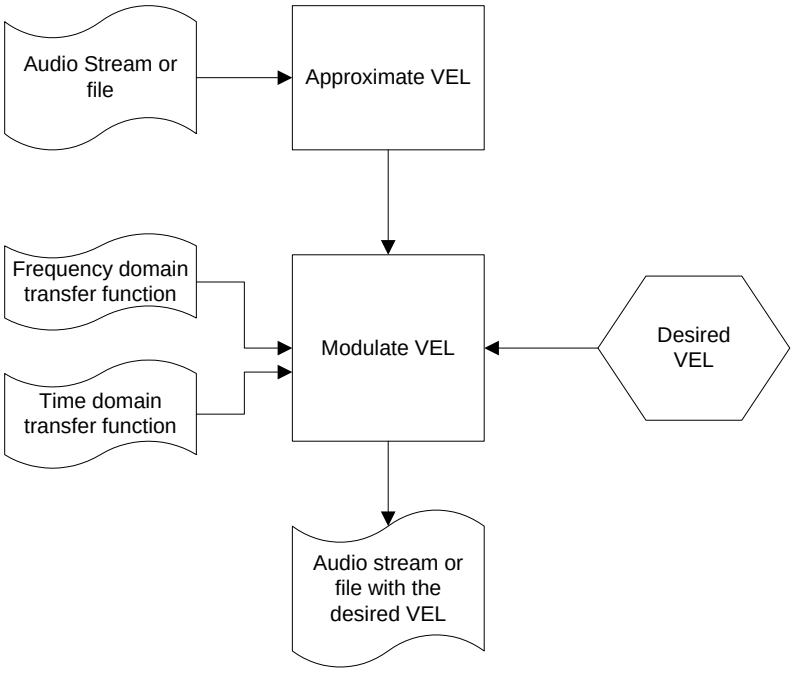

Figure 4 Process of modulating the VEL of an audio stream or file.