LOCALIZATION, LOUDNESS, AND PROXIMITY

D Griesinger David Griesinger Acoustics, Cambridge, Massachusetts USA

1 INTRODUCTION

Lokki et al. have found that "proximity," the perception of being acoustically close to the sound sources, is a major contributor to preference. [1] We have found that in most cases the perception of “being close” occurs when the azimuth of individual sound sources can be distinctly perceived. We believe the two perceptions are closely linked, and may share a common mechanism. In this paper we will describe some of the current knowledge of how the ear and brain achieve both source separation and localization for different types of signals. We will also present ways to calculate how these processes are affected by room reflections and reverberation.

An important aspect of this work is that we find the ability to sharply detect azimuth, and thus proximity, is discrete – you can ether sharply localize a source, or you cannot localize it at all. In most venues there is a distinct distance from sound sources where both proximity and localization suddenly become impossible to detect. We call this distance the Limit of Localization Distance, or LLD. In our work with multiple listeners there has been general agreement on where the LLD is found.

Considerable work has gone into the mechanisms for detecting localization, largely based on detecting azimuth through Interaural Level Differences (ILDs) and Interaural Time Differences (ITDs). There are a number of models that graph speech signals from one or more talkers in azimuth and time. [2] But the ability of these models to function in the presence of reflections has not been extensively studied. In this paper we are interested in understanding as closely as possible how the human ear performs this feat. We want to be able to predict when and how the mechanisms will fail.

We find that sharp localization in rooms is only possible when the ear and brain can separate the direct sound component from a particular instrument from the reflected sound that often quickly overwhelms it. If there are multiple sound sources the ear and brain must also separate the direct sound of one or more particular instruments from the sound of other instruments.

In our work and that of Homann [3] the periodic pulses in the amplitude envelope of speech and most musical instruments are a critical part of both processes. Homann has used the word “Periodicity” to describe the aspect of signals that makes them possible to separate. For speech and many musical instruments the ear detects the direct sound component of a complex signal by recognizing this periodicity, and then using the pitch information that periodicity provides to separate multiple sources into independent streams. Reflections from any direction randomize the phases of harmonics and reduce periodicity. When the sum of the reflected energy is strong enough to mask a significant part of the onset of sounds it is no longer possible to detect the direct sound as separate from the reflections.

2 PERIODICITY – SOURCE SEPARATION BY PITCH

The fundamental frequencies of speech vary from about 100Hz to 350Hz. The formants that carry the vowel information are much higher, in the range of 800Hz to 3000Hz. For both males and females formants are formed from the harmonics of low frequency fundamentals. Spectrograms that show the average frequency and strength of formants have been widely studied, but they show nothing of the phase of the individual harmonics that compose them. Detecting sound depends on more than the frequencies of the fundamentals and the harmonics. The waveform of sound – the ups and downs of the pressure envelope - is detectable over the entire audio frequency range. And the waveform depends on the phases of individual harmonics.

For human speech and most musical instruments the phase of harmonics is special. Speech harmonics are created by pulses or impacts. Pulses form when vocal cords open, a reed vibrates, or a hammer hits a string. In all these sounds the phases of the harmonics must re-align once in every fundamental period to re-create the original pulse.

Although the hair cells in the basilar membrane are insensitive to phase above about 1000Hz, and fire completely randomly above 1500Hz, they respond easily to the pulses in the amplitude envelope of speech and music. The effect can be easily heard. Tones where the phases align to form a sharp pulse can be louder and richer than tones where phases align symmetrically.

Both symmetric and asymmetric pulses cut through noise and are easier to hear than tones where the harmonic phase is random. But when there is high periodicity at the vocal formant frequencies a comb filter or an auto correlator can separate formants of an individual instrument or speaker from noise or other instruments if the filter is tuned precisely to the fundamental period. Humans are capable of simultaneously separating at least three such sources, and once separated each source is independently localizable.

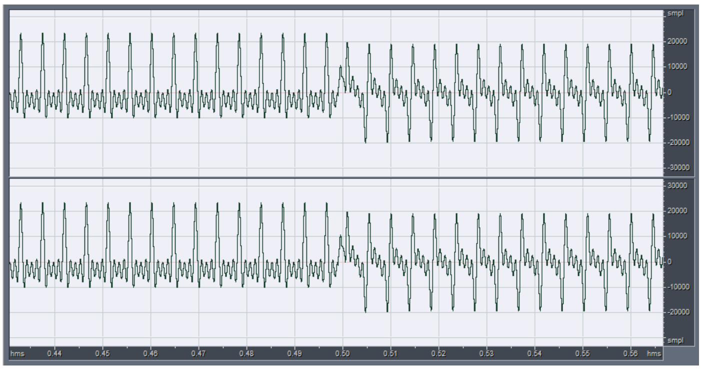

Figure one: 220Hz, 440Hz, 660Hz, and 880Hz sine tones combined, first with maximum phase alignment, and then simply added as sine waves. Both are heard as having the same pitch but they sound different.

The author has made hearing models that include comb filters that can separate simultaneous speech. The best ones separate two monotone talkers if they are at least a semitone different in pitch. There is still work to do. Human listeners separate monotone speech with a pitch difference of half a semitone. Normal non-monotonic speech can also be separated by pitch, but the fundamental pitches must be known, ideally with an accuracy of a third of a percent. Both Homann and the author find that the low frequency fundamental and the first few harmonics are essential for determining the precise frequency needed. If you filter out frequencies below 800Hz humans can no longer separate two monotone talkers unless the pitch difference is much greater than a semitone.

3 PITCH ACUITY

The basilar membrane is a mechanical filter fitted with a set of transducers (the outer hair cells) that enhance its sensitivity through positive mechanical feedback. The amount of feedback is controlled by the brain stem to make the well-known nearly logarithmic relationship between sound pressure and perceived loudness. A consequence of this feedback is that the quality factor – the sharpness of the resonance in the membrane – increases as the level of the acoustic signal decreases.

The number of hair cells that detect sound is known, as is their maximum firing rates. From the known number of hair cells in each critical band, and the known nearly logarithmic dependence of firing rates as a function of level we can estimate the error that can be expected in our ability to detect small changes in sound level. A shirtsleeve calculation for the just noticeable change in sound level matches the data in Zwicker rather well [4]. Our ability to perceive changes in amplitude is near the theoretical limit, and decreases predictably as level decreases.

The ability to perceive small changes in amplitude can be tested with standard clinical methods, and might be used to test for hair cell or spiral ganglia loss at normal speech levels. But to be clinically useful the test needs to be calibrated with a large number of subjects.

Our ability to hear small changes in the pitch of tones can be similarly estimated from the number of hair cells activated, the sound pressure level, and sharpness of the resonance of the membrane. We find that measuring pitch acuity as a function of level gives clues into the function of the basilar membrane and its associated neurology. In our experiments the acuity for hearing small changes in the pitch of sine tones does not decrease with decreasing level. The author’s personal pitch acuity for sine tones is about one percent over a broad frequency range at 60dB SPL. It is slightly better (and certainly not worse) at a sound pressure of 20dB SPL. But the pitch accuracy he can detect with sine tones is not sufficient to separate the vocal formants from two simultaneous talkers. The pitch accuracy needs to be higher, ideally around 0.3 percent. And it needs to be accurate, not just detectable. The pitch we perceive must be exactly the fundamental of the harmonics we need to separate.

The basilar membrane alone cannot reach this level of precision, and to make matters worse, the perceived pitch of sine tones is not stable. Because of the mechanical properties of the membrane perceived pitch decreases as the level increases. But if we add the three or four first harmonics to form a complex tone pitch becomes stable and precise. Once again the phase alignment of the harmonics is important. The best results come from the harmonics created by periodic pulses. Tones with downward peaks in pressure work better than tones with upward peaks, but we do not believe this asymmetry is important in practice. We have not found much up/down asymmetry in the pulses from male speech, female speech, or most instruments.

The effect of harmonics on pitch stability and accuracy may also give clues into the neurology of the basilar ligament which underlies the hair cells. We believe the pitch accuracy we observe can only be explained by a type of neural net that combines hair cell firings from regions of the membrane that are separated by harmonic intervals. This idea is not new. Neural networks that precisely determine pitch at low frequencies were described by Licklider in 1951 [5]. He postulated a bank of autocorrelators in the basilar ligament.

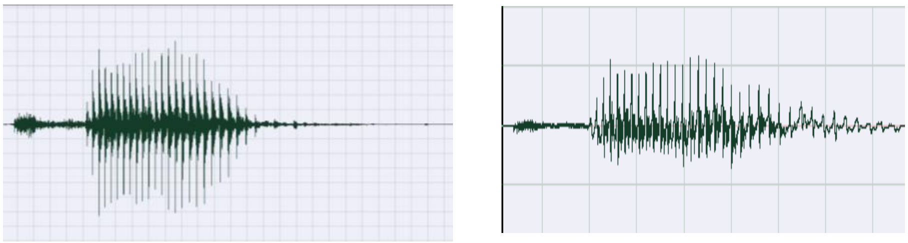

Figure 2: The syllable “ten” from the author after being high pass filtered and low pass filtered at 1000Hz. Note the clear periodic pulses in the waveform throughout the audible range.

4 LOC

We have developed a method intended to predict whether a source will be localizable or not from a binaural impulse response. We called it LOC, which is short for Localization. LOC was developed from a data set obtained using male speech with pauses between each word. We modeled a room with two loudspeakers placed at +-20 degrees in front of the listener in a space with two different reverberation times and four different pre-delays. The odd numbers from one to ten were presented from the speaker on the left, and the even numbers from the speaker on the right.

The subject was asked to vary the direct to reverberant ratio until it was just possible to hear the numbers at an azimuth of +- ten degrees or less. The room was modeled with 32 independent decaying noise signals convolved with 32 of the author's personal three dimensional HRTFs measured at the eardrums. The headphones were then equalized to match the author's frontal HRTFs at the eardrums. The reverberant field was generated from a three-dimensional decaying noise with a 5ms rise-time starting 5ms after the impulse for the direct sound. The sense of being in a three dimensional space was very good, and the sources localized in front, at least for the author.

In practice it was difficult to determine just when +-10 degrees was found, as the transition between full width and no width was so sudden. But it was easy and repeatable to find a value of D/R that caused the image to collapse within the space of one decibel. Figure one shows the data obtained.

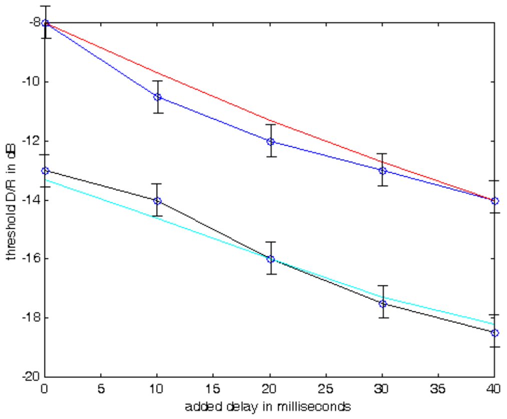

Figure 3: The localization data used to develop the measure LOC. The data indicates the direct to reverberant ratio, D/R, at which the ability to localize the words one to ten spoken by the author with sufficient pauses between words that reverberation does not mask the onset of the following word. The upper circles are for a reverberation time of one second, and the lower circles are for a reverberation time of two seconds. The colored lines indicate the fit to the data by LOC. Localization is easier to perceive with the longer reverberation time because more of the reverberant energy is later than the onset of the words.

Frequencies from 1000Hz to 4000Hz are principally responsible for the data we obtained. We developed a formula for LOC based on the idea that when the integrated loudness of the direct sound in that frequency range is stronger than the integrated loudness of the reflections in the first 100ms of the onset of a speech syllable, then the whole syllable can be localized and perceived as close to the listener. A very important aspect of the formula is that the "loudness" of the reflections is proportional to the integrated *logarithm* of the sound and not the integrated sound energy or pressure. [6]

LOC has been tested in several spaces, mostly large enough to hold 300 to 1500 seats. We typically used a loudspeaker array similar to Tapio Lokki's on stage playing his anechoic recordings of an electronic string quartet with a soprano voice playing a Mozart aria. We recorded impulse responses with my personal dummy head and a three dimensional microphone from each speaker array. We also listened to music playing through the array to find the LLD in various parts of the venues. Values of LOC above 3dB typically predicted good localization of the ensemble. Values below 2dB predicted poor localization. In most of these tests the reflections were more or less equal on both sides.

We also tested LOC with a data set from Boston Symphony Hall. All but one of the seats had LOC values of +3dB or more, and sounded good both on headphones and in person with a live orchestra. One seat had a strong reflection from the right side wall. LOC at that seat was below 2dB in the right ear and above 5dB in the left ear. The instruments were not localizable and the sound in that seat was muddy. Deleting the strong sidewall reflection from the impulse responses raised LOC in the right ear to +5dB. The sound improved greatly. We learned that to determine the ability to localize instruments in a particular seat we need to look at the minimum value of LOC from the two ears, not the maximum or average value.

5 LOCALIZING VIOLINS IN SMALL ROOMS

We tested LOC in an experiment at Rensselaer University using just the first and second violins from Lokki’s ensemble. Graduate students and two professors dropped a pencil on a table where they found the LLD. The agreement between the subjects was good, but LOC failed to predict the result. LOC predicted the LLD would be several feet farther back.

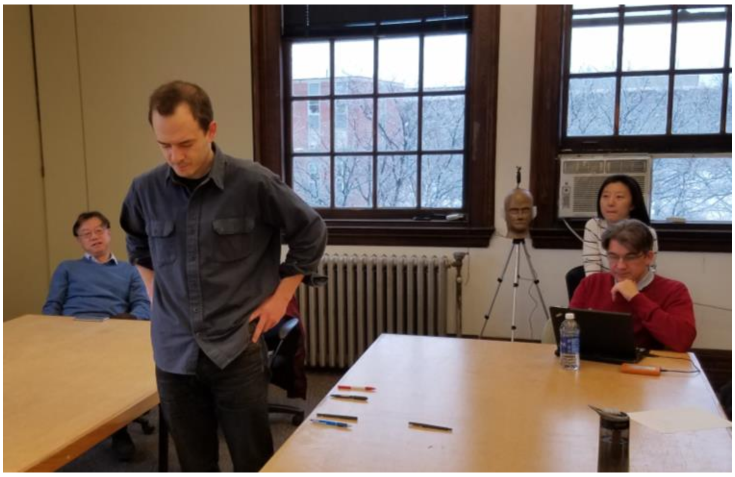

Figure 4: Finding the LLD in a small lecture room at Rensselaer University. Subjects put a pencil on the table where they could no longer localize two speakers in front playing separate anechoic violin parts. The room was small and reverberant. The dummy head and the 3D microphone used in the measurements can be seen in the background.

We would expect from the data in figure 3 and from considerable experience in large halls that a five part ensemble should be localizable even when the direct to reverberant ratio was much lower than one. In Boston symphony hall the critical distance, where the D/R is one, is only about 17 feet from an omnidirectional source. On the floor the best seats can be forty to fifty feet from the stage, and the great seats in the front of the first balcony are more than 110 feet from the stage. But in the small classroom the reflections come much sooner and a higher D/R is needed. LOC was not designed for this type of room.

There are several difficulties with the experiment in the small room. First – all the reflections come much sooner than they do in a concert hall. Also, the virtual room we used to develop LOC included a five millisecond delay before the onset of reverberation, followed by a 5 millisecond onset ramp. So there was an intrinsic ~15ms pre delay that is not included in the way we presented the data.

Violins typically play two octaves higher. That affects the density of harmonics in each critical band above 1000Hz, and we did not test the effect of that. But the major difficulty with using violins as sources for the localization experiment is that violin music does not resemble male speech. Violins start gradually. LOC was devised assuming the sounds we are trying to localize abruptly rise to full level. Violins can start notes abruptly, but usually they do not. Sound onsets can take 50 milliseconds or more to reach full level. By the time the note reaches full level reflections in a small room have substantial energy.

But with violins there is still a distinct point in a space where localization and proximity suddenly disappear. People agree on where that point is found. A modified version of LOC should be able to predict it.

We repeated the listening experiment that obtained the limit of localization data shown in figure 3, but used the first violin track from Lokki's Mozart recordings. Short phrases were extracted from the recording and presented alternately from the left and right virtual speakers. Figure 5 compares data obtained using violins to the data obtained with male speech.

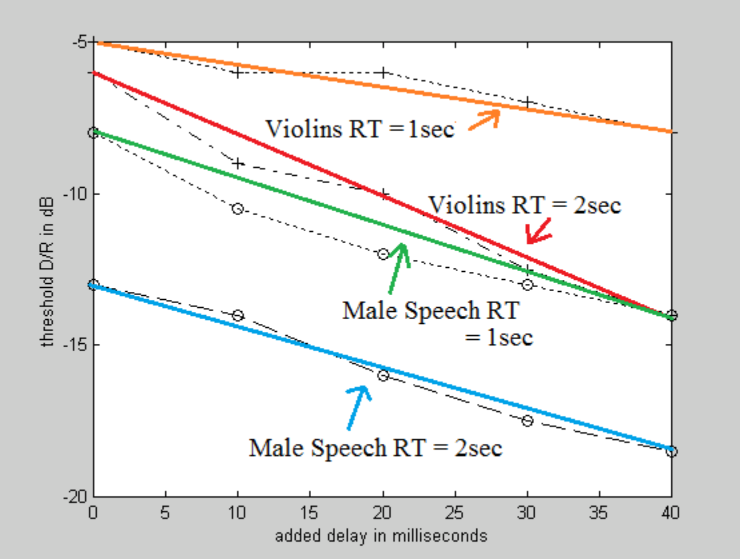

Figure 5: The thresholds of the Direct to Reverberant ratio for localizing a violin playing Mozart, compared to the thresholds for localizing male speech. In all cases the violins need a higher direct to reverberant ratio. The violin fundamentals varied from 350Hz to 630Hz. The male speech fundamentals varied from 120Hz to 140Hz. The onsets of the violin sounds varied from 30ms to over 60ms. The onsets of the male speech varied from 10ms to 30ms.

The direct to reverberant ratio needed to localize the violins is uniformly greater than for male speech. The data suggests that the major reason is the slower onset of each note. Reverberation with the one second RT builds up rapidly, and quickly masks the slow onset of the direct sound from the violin. A higher D/R is needed to overcome the masking. The slope of the one second RT as the predelay is also due to the slow onset of the notes. There is very little variation with pre-delay. The ear is not able to separate the direct sound and the reverberation unless the reverberation is nearly as strong as the loudness of the reflections inside the 100ms window.

With the two second reverberation and short values of pre-delay the D/R needed for localization for the violins is nearly the same as it is for the one second reverberation, but as pre-delay increases violins become easier to localize. The slope of the improvement in localizability is steeper than for male speech. In the theory behind the development of LOC the slope of the improvement with pre-delay is governed by the length of the comb filter or auto correlator that separates the direct sound from later reverberation. The data for violins suggests that the length of this filter is shorter for harmonics of violins than it is for male speech. This result was expected. High frequency fundamentals do not need as long a filter to obtain the same accuracy as lower frequency fundamentals.

6 LOCALIZING FEMALE SPEECH IN ROOMS

We also tested female speech using the same virtual room and the two second reverberation time. We tested two different speech segments, both use the syllables one to ten alternating from +- 20 degrees left and right. In one test the syllables were spoken slowly and carefully. The onset built up over a period of 50ms to 100ms. The other test was done with fast onsets, 20 to 50ms. In both cases there was sufficient pause between the syllables for the reverberation to decay.

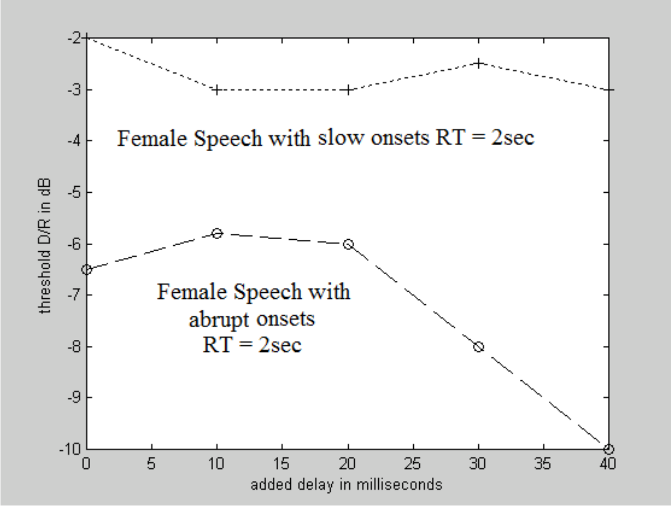

Figure 6: The threshold for localization of female speech with slow onsets compared to female speech with abrupt onsets. The speech with abrupt onsets, 20ms to 50ms, varied in fundamental frequency from 160Hz to 180Hz. The speech with slow onsets, 50ms to 100ms, varied from 170Hz to 240Hz.

The speech with slow onsets was more difficult to localize in this experiment than the violins in figure 5, even though in figure 6 the reverberation time is two seconds. This result is most likely due to the longer duration of the onsets. Reducing the onset time with a more aggressive speech signal improved localization considerably. The rapid improvement in localization after 20ms (4dB in 20ms) in the graph of the abrupt speech is almost twice as fast as the slope in the male speech data (2.5dB in 40ms). This is probably due primarily to the higher fundamental frequencies, which are a musical fifth higher than the male speech. Once again this implies that the ear employs a shorter autocorrelation filter at higher frequencies.

7 WHAT IS DIRECT SOUND?

When the author gave his first presentation on the subject of localization in the presence of reverberation [7] he received only one question from the audience. It was: “What is the direct sound?” The answer he gave was “The direct sound is the sound that travels directly from the source to the receiver – perhaps you could call it C5.” This was not a good answer. The question is profound.

We recently had the opportunity to make measurements and listen for the LLD in the EMPAC concert hall at Rensselaer University with a group of graduate students in acoustics. The result was surprising. Localization of Lokki's Mozart ensemble was at least plausible in a majority of the seats, with a few exceptions near the side walls. The impulse responses were unusual. In most seats there was a strong early reflection about 5ms after the direct sound. The 3D microphone revealed that these came from the floor. The geometry precluded the floor of the stage – it seemed the reflections were coming from the floor under the seats. The seats in this hall are very open. A great expanse of floor was visible under the seats in front of the test microphones.

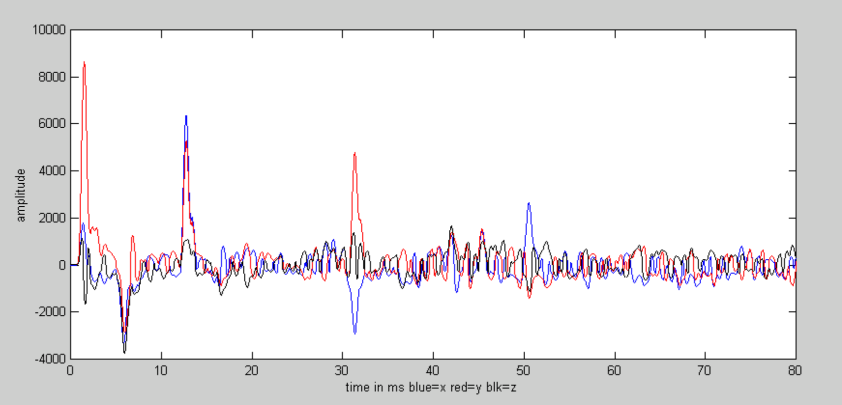

Figure 7: A three dimensional graph of a 3kHz octave band impulse response at seat F1 CII in the EMPAC Auditorium. Blue is left/right, Red is front/back, and black is up/down. Note the strong reflection 5ms after the direct sound.

LOC measured with the formula we presented in previous papers showed that in most seats including this one LOC was +3dB or above. The code for calculating LOC included the 5ms reflection as part of the direct sound. This raises two questions. First, if the seats were occupied by people, would that reflection still be present, or would it be blocked by bodies and feet? Second – what if the reflection arrived just a little bit later? Would it still be perceived as augmenting the direct sound?

We had done some preliminary thinking about this question while auralizing data from a new hall in Germany. There was a prompt reflection at 3ms from a side wall in a few seats near the side. It caused a shift in the direction of the sound, but did not affect the ability to separate instruments or the proximity.

That result was not entirely unexpected, as it is well known that a prompt early reflection can augment speech and musical instruments. Churches have put a wall and a ceiling behind and over pulpits for centuries. Often in a room where localization and proximity is poor a seat in the very last row, up against the back wall, will sound much better. But it has to be the very last row. The next to last row is no better than the others. In practice I had found that your ears had to be within two and a half feet of the wall for the trick to work.

We did some experiments using our virtual room and male speech. The results showed that a reflection at 5ms which was 6dB less strong than the direct sound did augment the loudness and the localizability of a source without detrimental effects on timbre. The current version of LOC appeared to work for this case. A reflection 6ms after the direct sound did not affect the localizability of the sound, and was beginning to alter the timbre. A reflection at 7ms reduced the ability to localize, and added an unpleasant timbre. We tried the same experiment with female speech and got the same result. We do not know why this effect occurs, but it is the same for both male and female speech.

As a result of these experiments we added a 2ms cross fade centered at 6ms to the calculation for LOC. Experiments showed that the new bit of code correctly predicted how very early reflections would affect the localization of both male and female speech.

8 DISCUSSION

We have presented a few simple experiments in the hope of better understanding how early reflections influence our abilities to localize speech and music in reverberant spaces. We found that a major difficulty in this ability lies in the duration of the onset of sounds. With violins and deliberately spoken speech the build-up of the voiced sound can be similar to the build-up of the reflected energy from that sound, and localization becomes difficult. However in practice, at least for the violin tracks in Lokki's Mozart examples, the task may not be as difficult as one would expect.

A possible reason is that both for speech and for violins the onsets of each note or syllable are not the same. In fact, while doing the listening that generated the data in figure 5 not all the syllables or notes could be localized. When there were a few that could be clearly localized these dominated the apparent azimuth and proximity of the whole stream. The design of the LOC measure was intended to predict the localizability of these more abrupt onsets. The predictions obtained from the few that are sufficiently abrupt seem to hold for the perception of the entire stream. In short – localization of a stream does not depend on being able to localize each sound in the stream.

Our experiments deliberately avoided source signals where the reverberation from a previous note or syllable masks the onset of a following note or syllable. In more realistic situations this masking can be a major problem for intelligibility and localization, especially for speech, since soft speech syllables can follow louder ones. The Speech Intelligibility Index, STI, predicts this kind of masking for speech. Both LOC and STI should be considered in a highly reverberant situation.

The duration of notes is problematic for music. Short notes do not excite long reverberation times very much, so masking by reverberation is principally a problem when notes are held for more than a quarter of the reverberation time. An additional complication is that sounds with fundamental frequencies different by a major fourth or more are not masked by reverberation from a previous sound. All these factors enter into our abilities to detect localization and proximity of speech and music in reverberant spaces. It is not news that marginal acoustics for speech can be overcome by articulation and tempo, and the same is true for music. Modifying LOC to accommodate all types of signals and all types of rooms is not likely to happen soon.

9 REFERENCES

1. T. Lokki, J. Pätynen, A. Kuusinen, & S. Tervo, 'Disentangling preference ratings of concert hall acoustics using subjective sensory profiles.' The Journal of the Acoustical Society of America, **132,** 3148-3161. (2012)
2. W. Hess 'Acoustical evaluation of virtual rooms by means of binaural activity patterns' Audio Engineering society Convention Paper 5864, (2003)
3. A. Josupeit, V. Hohmann, 'Modeling speech localization, talker identification, and word recognition in a multitalker setting' JASA 142, 35 (2017).
4. E. Zwicker, R. Feldtkeller, 'The Ear as a Communication Receiver' American Institute of Physics 1999
5. J. Licklider, A duplex theory of pitch perception. *Experientia,* Vol VII/4 128-134. (1951) Available on www.davidgriesinger.com
6. D. Griesinger, 'The relationship between audience engagement and the ability to perceive pitch, timbre, azimuth and envelopment of multiple sources' Proceedings of the IOA 2011. Available on www.davidgriesinger.com
7. D. Griesinger, 'Spatial perception of distance, azimuth, and envelopment when the direct to reverberant ratio (d/r) is below -6db' Preprint for the 19th ICA conference in Madrid, 2007. Available on www.davidgriesigner.com