To appear in the Proceedings of the1994 International Conference of Auditory Display (Santa Fe, NM, Nov. 7–9)

# Efficient listening with two ears: Dichotic time compression and spatialization

Barry Arons
MIT Media Laboratory
20 Ames Street
Cambridge MA, 02139
barons@media.mit.edu

**Abstract**

To increase the amount of information we can collect in a given amount of time, it is possible to employ signal processing techniques to speed up the rate at which recorded sounds are presented to the ears. Besides simply speeding up the playback, it is possible to auditorily display the signals in a way that allows us to process and interpret the signals more efficiently by exploiting the use of our two ears.

This paper first reviews time compression techniques for increasing the amount of information that can be presented to a listener, with an emphasis on techniques that use two ears. The paper then describes a new technique that integrates these dichotic time compression techniques into a spatial audio display system.

## 1 Introduction

Auditory information is collected through our ears at a fixed rate and processed in our brain. To increase the amount of information we can collect in a given period of time, it is possible to employ signal processing techniques to speed up the rate at which recorded sounds are presented to a listener. These “time compression” or “time scale modification” algorithms have primarily been used on speech recordings. Besides simply speeding up the playback, it is possible to auditorily display the signals in a way that allows us to process and interpret the signals more efficiently by exploiting the use of our two ears. These “dichotic” time compression techniques present different portions of the audio signal to each ear, increasing intelligibility.[1]

Current spatial audio display systems attempt to take advantage of the fact that human listeners have two ears by creating virtual sound sources that are synthesized over headphones. However, one of the fundamental design premises of a spatial audio system conflicts with the presentation needs of a dichotic time compression algorithm. This prevents the use of the dichotic time compression technique with a conventional spatial audio system.

This paper first reviews time compression algorithms for increasing the amount of information that can be presented to a listener, with an emphasis on methods that use two ears. The paper

[1]*Dichotic* refers to two different signals that are presented to the ears over headphones.

then describes a new technique that integrates these dichotic time compression techniques into a spatial audio display system to further increase the bandwidth of the listener.

## 2 Time compression

One technique for increasing listening capacity is by time compressing an auditory signal—to play back an audio recording in less time than it took to record. A wide variety of time compression techniques have been developed that allow audio recordings (the primary focus has been on speech) to be presented at a faster rate without seriously degrading the audio quality. Spontaneous, or conversational, speech can be time compressed by a factor of about two and still remain intelligible and comprehensible [1, 2, 3, 4].[2] Time compression techniques rely on the temporal redundancy of speech as demonstrated by Miller and Licklider—the intelligibility of speech recordings interrupted by periods of silence remains high if the number of interruptions per second and the portion of time the speech is on are properly selected [5].

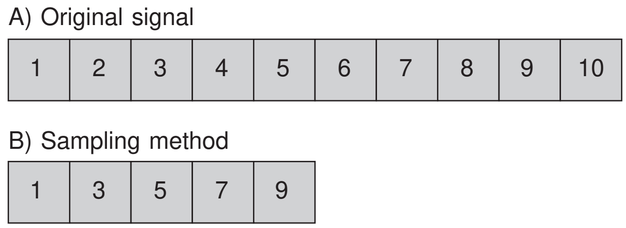


Figure 1: Part (A) shows the original signal divided into short (e.g., 50 ms segments). Part (B) shows the signal time compressed by the sampling method with every other segment removed. The amount of time compression can be varied by changing the relative lengths of the retained and discarded segments.

One of the simplest techniques to time compress a recording is the sampling, or Fairbanks, method [6]. This technique consists of removing short segments of the signal at regular time intervals (figure 1). For speech recordings these segments are usually longer than a pitch period (> ~10 ms) and shorter than a phoneme (< ~100 ms), and are often 30–60 ms. The perceived quality of a signal time compressed by this method can be improved by performing a short cross fade[3] (figure 2) rather than simply abutting the segments (as shown in figure 1B).

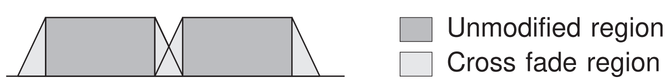


Figure 2: A linear cross fade between segments of the sampling method to reduce distortions.

[2]There are many factors that influence the maximum practical time compression including: the listener's familiarity with the material, the rate and content of the original speech, the compression technique, and the listener's prior experience with time compressed speech.

[3]Decreasing the amplitude of the end of one segment while increasing the amplitude of the beginning of the next segment.

The synchronized overlap-and-add (SOLA) method of time compression further improves the quality of the speech by ensuring that the segments are optimally aligned before performing the cross fade [7, 8]. This is done by checking different amounts of overlap between the end of one speech segment and the beginning of the next to find where the signals are the most similar (i.e., by computing the cross correlation). This technique requires more computation, but it effectively removes entire pitch periods, and produces better sounding speech than the sampling method.

## 3 Integrating information between the ears

There are a variety of psychoacoustical phenomena that illustrate the human ability to integrate information presented to both ears (e.g., localization, lateralization, binaural masking level differences, and binaural beats—see [9]).

Speech signals are treated differently than tones or noise in higher levels of human auditory processing, and are grouped more cohesively than other sounds [10]. For example, the continuity of pitch helps control attention when speech signals are presented to both ears. Gray and Wedderburn showed that although there is a tendency to group signals according to the ear they are presented to, this can be overcome if there are strong cues that favor a different grouping [11, 12]. Their study showed a preference for grouping by meaning rather than by ear for digits presented dichotically with words or syllables.

## 4 Dichotic presentation of time compressed speech (DTCS)

The sampling time compression technique illustrated in figure 1 reduces the listening time by discarding a portion of the original signal. The time compressed speech is typically presented over a loudspeaker or diotically[4] over headphones. However, rather than simply removing the material, it is possible to play the portion of the signal that would normally be discarded to the other ear (figure 3).

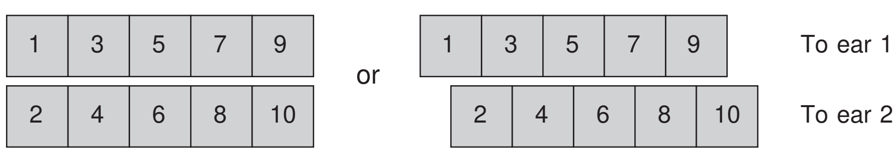


Figure 3: Presenting speech that has been time compressed with the sampling method to both ears. Segments can be completely overlapped (left), or offset by half of a sampling period (right).

This style of presentation of dichotic time compressed speech (hereafter DTCS) was first described by Scott in the mid 1960s [13]. Scott reports that subjects found the dichotic presentation to be more intelligible than a diotic presentation. The dichotic speech sounds a bit annoying at first, as most listeners switched attention between their ears, but this unusual sensation became less noticeable over time. Scott says "although … there is a temporal mismatch of the two speech signals when presented dichotically, a fusing of information at both ears must take place to in-

[4]*Diotic* presentation is when the same signal is presented to both ears over headphones.

crease the intelligibility" [13, p. 64]. Gerber showed that under a variety of different configurations intelligibility of time compressed speech was always better for dichotic presentation than with diotic presentation [14, 15]. With a properly selected discard interval (the length of audio segment removed from the signal and played to the opposite ear), word intelligibility errors decreased 49% for a 2:1 time compression under dichotic conditions [15].[5]

It is also possible to create an analogous dichotic SOLA signal by processing speech through the SOLA algorithm a second time with an offset in the starting point. Note that because the algorithm shifts the segments to minimize irregularities between segments, a dichotic signal produced with this technique may not contain all of the information contained in the original recording.

## 5 Presenting DTCS spatially

In spatial audio display systems one or more channels of audio are presented to the ears based on the head related transfer function (HRTF) and the spatial location of the source relative to each ear [16, 17]. For example, in figure 4A, a real sound source S is filtered based on the reflective characteristics of the head, body, and ears (pinna) and the interaural time delay due to the path length difference to the ears to produce a virtual sound S′ when presented over headphones (figure 4B).

It is useful to be able to present time compressed speech in a virtual acoustic display, such as in user interfaces that allow skimming or browsing of recorded audio material [18, 19], or systems that attempt to present multiple streams of recorded speech simultaneously [20, 21]. Presenting speech that has been time compressed using the basic sampling or SOLA techniques in a spatial audio display system is straightforward, as it can be treated like any other audio source. However to exploit the improved intelligibility of dichotically presented time compressed speech within a spatial audio system a novel approach must be taken to spatialize DTCS.

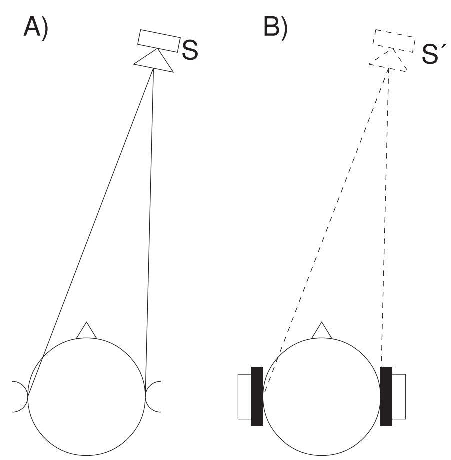


Figure 4: (A) Top view of a listener's head and sound source S (loudspeaker) located in space. (B) Virtual sound source S′ created by a spatial audio system.

---

[5]With discard intervals of 40, 50, 60, and 70 ms the intelligibility errors decreased 25, 49, 25, and 42% respectively (although the figure for a 70 ms discard interval was not found to be statistically significant).

The goal of DTCS is to explicitly present different signals to each ear, while a spatial audio system simulates a source at some spatial position by carefully controlling interaural time and intensity differences as well as the monaural spectral cues in the signals reaching the two ears. These two goals and their associated acoustic cues are thus seemingly in conflict. For example, if both channels of a DTCS signal are placed at the same location in a spatial audio system (e.g., at S´ in figure 4B), both ears will receive a portion of the signal from each channel. Unfortunately, this cross talk will degrade the DTCS signal, as Gerber notes "if one listens to both signals with both ears, the intelligibility is poorer than if one listens to one signal with one ear and the other signal with the other ear" [14, p. 459].

However, it is possible to create a virtual sound source where each ear only receives one channel of the DTCS signal. This can be achieved by placing two virtual sound sources at the same location, but only filtering each signal for one ear (figure 5A). One system configuration for creating this type of auditory display is shown in figure 5B.

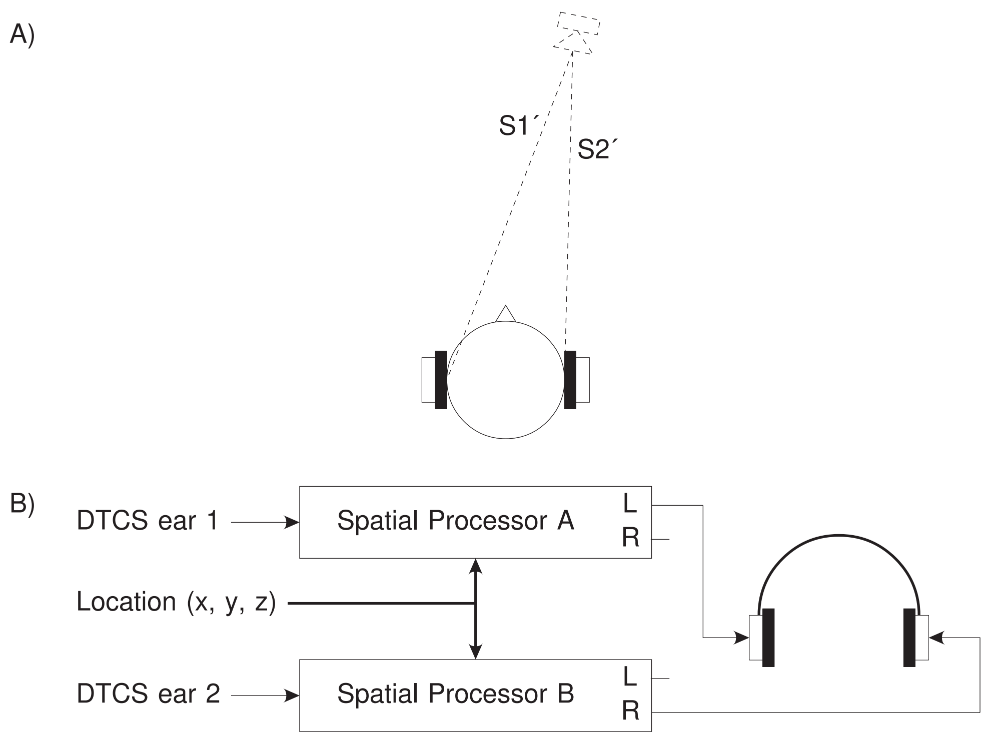


Figure 5: (A) Technique to present spatialized DTCS. S1´ and S2´ are two channels of DTCS originating from a single virtual sound location, but each is only presented to a single ear. (B) System configuration for spatializing DTCS using two Beachtron boards [22].

Moore says that "while the most reliable cues used in the localization of sounds depends upon a comparison of the signals reaching the two ears, there are also phenomena of auditory space perception which result from monaural processing of the signals" [9, p. 194]. Note however, that the spatialized DTCS technique described here is not strictly presenting two monaural channels, rather there are still a variety of rich interaural cues. The HRTF cues, including interaural intensity differences and monaural spectral cues, are all present. The only cue that is missing is interaural time difference, since the signals received by the ears do not originate from a single audio signal (however the two DTCS signals are still are highly correlated). Speech sounds in particular are very rich in familiar information. These common speech spectral cues make it easier for us to perceive these two channels as a single auditory stream [10].

This spatialized DTCS technique was informally found to produce an externalized virtual image. As with the DTCS technique, the speech sounds a bit choppy, however the speech was intelligible and comprehensible and could be localized about as well as a spatialized version of the original speech recording (time compressed, but not dichotic).

## 6 Issues

The work presented in this paper has only scratched the surface of the spatialized DTCS technique. Further development of the underlying technique is needed as well as a formal evaluation to test the efficacy of this method of auditory display. Specific areas of research include: optimizing the time difference between the DTCS channels; exploring the perceived spatial and comprehension effects of permitting a small amount of cross talk between channels (thus adding interaural time differences); and modifying the underlying spatial audio system architecture to allow DTCS to be presented spatially without requiring the use of two separate sound processing channels. The system also needs to be tested to perceptually evaluate: if the sounds can be localized and externalized; if the maximum preferred time compression is degraded when the DTCS is spatialized; and if presenting spatialized DTCS enhances or hinders a listener's ability to listen to multiple audio streams.

The dichotic SOLA technique also needs to be explored in greater detail. With knowledge of, and access to, the internal details of the SOLA algorithm it should be possible to generate a dichotic SOLA signal by calculating only one set of cross correlations. This will be a computational improvement over using the algorithm on a separate sound segment for each ear, as well as making it possible to maximize the amount of information presented in the two signals.

## 7 Conclusions

Auditory displays can exploit the fact that we have two ears in a variety of ways. This paper has discussed techniques for dichotically presenting time compressed speech to enable high speed listening. These two-eared presentation styles can be used to (1) increase the time efficiency of the listener, (2) increase the intelligibility and comprehension of the material, or (3) a combination of the two.

The decreasing costs of audio hardware and spatialization systems, along with their increasing power and accuracy are enabling audio to be considered in a variety of new application environments. Combining DTCS with spatialization may increase the power, usefulness, and acceptance of both technologies. These and related methods of dichotic auditory display will lead to more efficient listening through the use of two ears.

## Acknowledgments

Marc Davis asked challenging questions that led to the development of the spatialized DTCS technique. Atty Mullins re-cabled his hardware configuration and wrote software for the Beachtron that demonstrated spatialized DTCS. Thanks to Lisa Stifelman for discussing dichotic listening studies and helping edit this paper. Barbara Shinn-Cunningham and Chris Schmandt also provided valuable feedback on a draft of this paper.

## References

[1] Foulke, E. "The Perception of Time Compressed Speech." Ch. 4 in *Perception of Language*, edited by Kjeldergaard, P. M., D. L. Horton, and J. J. Jenkins, 79–107. Columbus, OH: Merrill, 1971.

[2] Foulke, W. and T. G. Sticht. "Review of Research on the Intelligibility and Comprehension of Accelerated Speech." *Psychological Bulletin 72* (1969): 50–62.

[3] Beasley, D. S. and J. E. Maki. "Time- and Frequency-Altered Speech." Ch. 12 in *Contemporary Issues in Experimental Phonetics*, edited by Lass, N. J., 419–458. New York: Academic Press, 1976.

[4] Arons, B. Techniques, Perception, and Applications of Time-Compressed Speech. In *Proceedings of 1992 Conference,* American Voice I/O Society, Sep. 1992, 169–177.

[5] Miller, G. A. and J. C. R. Licklider. "The Intelligibility of Interrupted Speech." *Journal of the Acoustic Society of America 22* (1950): 167–173.

[6] Fairbanks, G., W. L. Everitt, and R. P. Jaeger. "Method for Time or Frequency Compression-Expansion of Speech." *Transactions of the Institute of Radio Engineers, Professional Group on Audio AU-2* (1954): 7–12. Reprinted in G. Fairbanks, Experimental Phonetics: Selected Articles, University of Illinois Press, 1966.

[7] Roucos, S. and A. M. Wilgus. High Quality Time-Scale Modification for Speech. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing,* IEEE, 1985, 493–496.

[8] Hejna Jr, D. J. "Real-Time Time-Scale Modification of Speech via the Synchronized Overlap-Add Algorithm." Master's thesis, MIT, 1990. Department of Electrical Engineering and Computer Science.

[9] Moore, B. C. J. *An Introduction to the Psychology of Hearing.* New York: Academic Press, 3d edition, 1989.

[10] Bregman, A. S. *Auditory Scene Analysis: The Perceptual Organization of Sound.* Cambridge, MA: MIT Press, 1990.

[11] Gray, J. A. and A. A. I. Wedderburn. "Grouping Strategies with Simultaneous Stimuli." *Quarterly Journal of Experimental Psychology 12* (1960): 180–184.

[12] Hede, A. J. "Awareness of List Organization and the Gray and Wedderburn Effect." *Psychological Reports 43* (1978): 371–374.

[13] Scott, R. J. "Time Adjustment in Speech Synthesis." *Journal of the Acoustic Society of America 41* (1967): 60–65.

[14] Gerber, S. E. "Limits of Speech Time Compression." In *Time-Compressed Speech*, edited by Duker, S., 456–465. Metuchen, NJ: Scarecrow, 1974.

[15] Gerber, S. E. and B. H. Wulfeck. "The Limiting Effect of Discard Interval on Time-Compressed Speech." *Language and Speech 20* (1977): 108–115.

[16] Wenzel, E. M. "Localization in Virtual Acoustic Displays." *Presence 1* (1992): 80–107.

[17] Wenzel, E. M. "Spatial Sound and Sonification." In *Auditory Display: Sonification, Audification, and Auditory Interfaces*, edited by Kramer, G., 127–150. Santa Fe Institute

Studies in the Sciences of Complexity in the Sciences of Complexity, Vol. XVII. Reading, MA: Addison-Wesley Publishing Company, Inc., 1994.

[18] Arons, B. SpeechSkimmer: Interactively Skimming Recorded Speech. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST),* ACM SIGGRAPH and ACM SIGCHI, ACM Press, Nov. 1993, 187–196.

[19] Arons, B. "Interactively Skimming Recorded Speech." Ph.D. dissertation, MIT, 1994.

[20] Arons, B. "A Review of the Cocktail Party Effect." *Journal of the American Voice I/O Society 12* (1992): 35–50.

[21] Mullins, A. T. "Audio Streamer: Browsing Concurrent Audio Streams." Master's thesis, MIT, 1995.

[22] Crystal River Engineering. *The Beachtron: Three-Dimensional Audio for PC-Compatibles*, Groveland, CA. 1993.