On-Line Audio Dilation for Human Interaction

John S. Novak, III[1], Jason Archer[2], Valeriy Shafiro[3], Robert Kenyon[1], Jason Leigh[1]

[1] Department of Computer Science, University of Illinois at Chicago, Chicago IL, USA
[2] Department of Communication, University of Illinois at Chicago, Chicago IL, USA
[3] Communication Disorders and Sciences, Rush University Medical Center, Chicago IL, USA

Jnovak5@uic.edu, Jarche2@uic.edu, Valeriy_Shafiro@rush.edu, Kenyon@uic.edu, Spiff@uic.edu

Abstract

On-line audio dilation is a technique that time stretches, or slows down the tempo of audio signals as they are generated. This paper presents an on-line dilation technique and our on-going research to assess the effects of audio dilation on speech communication using an interactive DIAPIX problem solving task.

Index Terms: Speech modification, audio dilation, Diapix task

1. Introduction

Slower speech is often perceived to be easier to understand: adults instructed to speak as if talking to a child, to a non-native speaker, or to a hearing-impaired person tend to reduce their speaking rate, producing "clear" speech that is perceived more accurately [9]. Slower speech rates also have a positive effect on memory recall [2][5], sentence comprehension[7][8], and intelligibility in attentionally demanding tasks [6].

Automatic audio dilation of speech is frequently suggested as a promising approach to improving the effectiveness of aural communication. Audio dilation through time stretching has now become quite common in everyday software applications and personal listening devices. It is often recommended in online educational settings that use lecture capture technology [10] and have been suggested as a future component for hearing aids [4]. However, existing algorithms that time-stretch audio are designed for off-line use, i.e., for the time-stretching of complete, existing audio files. In contrast, the approach described here performs "on-line" audio-stretching, where the processing occurs as audio signals are being acquired. The algorithm used here is in effect an on-line variant of existing algorithms, making it suitable for desktop and high-end smart phone class computing devices. In this paper, we assess its functionality using an interactive problem solving task in which two people must cooperate to find differences across a pair of pictures.

2. On-Line Audio Dilation Software

Audio time rescaling is typically accomplished with variants of a phase vocoder algorithm, implemented in the digital domain using a short time Fourier transform. This technique is a modification of one developed by Ellis [3]. Ellis' original technique processes complete audio files according to three parameters:

- N, a frame length,
- $R_a < N$, an analysis rate, and
- $D < 1$, a dilation factor.

The existing audio signal is decomposed into separate audio frames, each of length N. The start of each audio frame is offset from the start of the previous one by R_a samples. Since $R_a < N$, these frames necessarily overlap; in our work, R_a is set to N/4. Once decomposed into overlapping frames, these frames are individually transformed into the frequency domain with a Fast Fourier transform. Then entirely new frames of data are created by interpolating point-wise across these Fourier frames, according to the dilation factor, D, e.g., a dilation factor of ½ results in the creation of one new frame of spectral data between each two original frames of data. Care is taken to comply with phase wrapping constraints while interpolating phase data. Finally, the expanded set of spectral data frames are returned to the time domain through individual inverse Fourier transforms, and are summed together as frames at the original R_a overlap factor. (Note that dilation factors which do not reduce to values of 1/x for some integer value of x will result in some original frames being discarded.)

Our modifications to Ellis' algorithm for on-line audio dilation are as follows:

- Audio samples are read from a microphone, rather than an existing audio file.
- Overlapping frames of audio data are created as the audio signal is acquired from the microphone.
- As each frame of N samples is acquired, it is Fourier transformed and deposited into the input frame buffer.
- Whenever two frames exist in the input buffer, they are immediately used for interpolation, and the results placed in an output frame buffer.
- When all of the interpolation frames are complete, the earliest frame is discarded.
- If subsequent input frames are available, the process continues; otherwise, the software waits for additional input frames to appear.
- Finally, as frames accumulate in the output buffer, they are immediately transformed back to the time domain, overlapped, summed, and played through an output speaker.

Additionally, to prevent inter-speech pauses (especially those caused while listening to an interlocutor) from being dilated, we implemented a Voice Activation Detection protocol, which discards input frame buffers below a particular average user controlled audio energy level prior to processing. (See Figure 1.)

In our algorithm, we fixed the frame length at N = 1024 samples, and the analysis rate R_a, at 256 samples (frames overlapped by 75%). In our Diapix experiments, we fix the dilation factor D at 0.7, resulting in playback time that is 42% longer than the original speech.

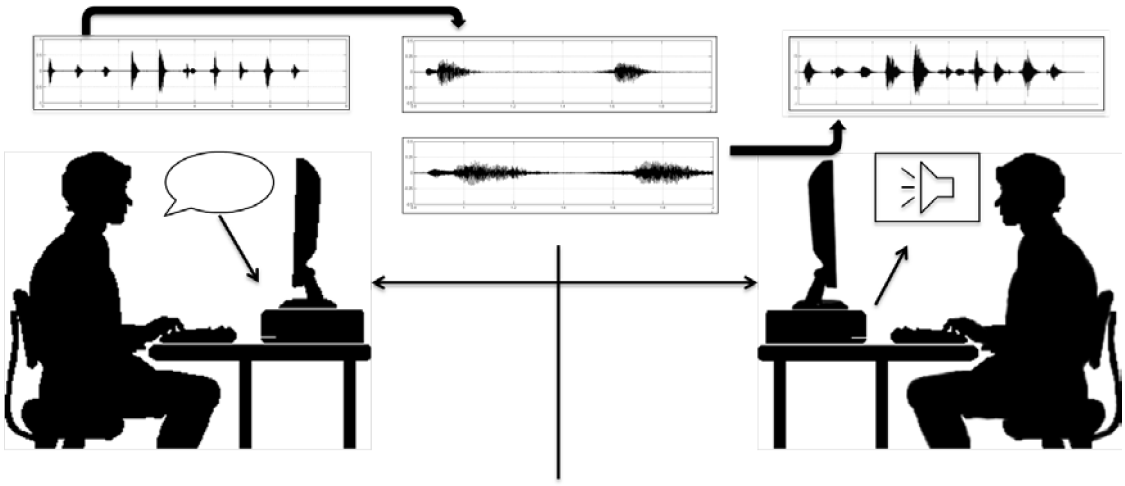

Figure 1: On-line Audio Dilation, with Voice Activation Detection

Preliminary experiments show that mid-range laptops (specifically a Dell Inspiron laptop with P8700 2.53 GHz processor) can perform the necessary transforms and inverse transforms swiftly enough to yield a seamless user experience. Preliminary estimates show that these computations are in the reach of high-end smartphones, as well.

3. Diapix Image Test

In order to assess the effects of on-line audio dilation on spoken interpersonal interaction, we employ the Diapix task to elicit spontaneous, cooperative, conversational speech from the participants[1]. Two test subjects are isolated from each other in separate rooms and allowed to communicate through audio headsets connected to laptop computers, simulating a telephone conversation. The subjects are then given two images representing similar scenes, but with twelve differences, and asked to identify the differences as quickly and as accurately as possible, by conversing through the headsets. Participants are given five minutes to complete each task. If the participants have not found all differences within five minutes, they are asked to stop. This task is repeated four times (with separate sets of difficulty-matched images) under the following conditions:

- Neither user dilated (baseline)
- Subject A dilated by 42%, Subject B non-dilated
- Subject A non-dilated, Subject B dilated by 42%
- Both users dilated by 42%

(Note that the isolation between the two rooms prevents the test subjects from hearing what their partners hear. Dilated speakers do not hear their own dilated utterances through their own headsets.)

Time to completion and task accuracy (i.e. number of differences found) are recorded, with the non-dilated condition serving as a control. Additionally, at the end of the experiments, the users are surveyed regarding the qualitative experience of conversing through the audio dilation program. Finally, all sessions are audio recorded to examine potential changes in interaction dynamics, speech production (i.e., speaking rate, pause duration, number of times speech is initiated, vocal effort, fundamental frequency changes, back channeling) of each participant in response to listening to dilated audio signals.

4. Preliminary Observations and Directions

Preliminary results obtained with adult native speakers indicate changes in speech communications and task solving interactions in response to audio dilation. When neither speaker is dilated, i.e., a standard Diapix test, participants take turns in conversation an approximately equal number of times. However, in tests where only one participant was dilated, the participant whose speech is being dilated (but who hears non-dilated speech) takes a subordinate role in task interactions. Furthermore, the dilated speaker, while not hearing his own dilated voice, nevertheless reported more frustration trying to coordinate actions during the interaction than the non-dilated speaker who hears the dilated voice of the other.

Additional application of online audio dilation can be found in interactions in which rapid flow of speech may hinder comprehension (e.g., with non native speakers of a language with reduced language proficiency or individuals with cognitive impairments (e.g. post Traumatic Brain Injury, older adults with Mild Cognitive Impairment.) Further work will examine the effects of audio dilation across participants with different levels of linguistic and cognitive abilities.

5. Audio Dilation at INTERSPEECH 2013

At INTERSPEECH 2013, we will demonstrate our prototype on-line audio dilation software on paired laptops with headsets. This software will be capable of dilating the speech on-line of one or two users with audio headsets, and modifying the rate of audio dilation during use with the Diapix interactive problem solving task.

6. References

[1] R. Baker, and V. Hazan, "DiapixUK: a task for the elicitation of spontaneous speech dialogs," *Behavior Research Methods,* vol. 43, no. 3, pp. 761-770, Mar. 2011.

[2] A. R. Bradlow et al., "Effects of talker, rate, and amplitude variation on recognition memory for spoken words," *Perception and Psychophysics*, vol. 61, no. 2, pp. 206-219, Feb 1999.

[3] D. Ellis, A Phase Vocoder in Matlab, *http://labrosa.ee.columbia.edu/matlab/pvoc/.*

[4] E. W. Foo, and G. F. Hughes, "Hearing assistance system for providing consistent human speech," U.S. Patent Application No. 20,120,215,532, Aug. 2012.

[5] S. D. Goldinger et al, "On the nature of talker variability effects on recall of spoken word lists," *Journal of Experimental Psychology: Learning, Memory and Cognition.* vol. 17, no. 1, pp. 152-162, 1991.

[6] B. Gygi, and V. Shafiro, "Spatial and temporal factors in a multitalker dual listening task," Acta Acoustica, vol. 98, no. 1, pp. 142-157, Jan. 2012.

[7] J. F. Schmitt and R. L. McCroskey, "Sentence comprehension in elderly listeners: The factor of rate," *Journal of Gerontology,* vol. 36, no. 4, pp. 441-445, 1981.

[8] J. F. Schmitt, "The effects of time compression and time expansion on passage comprehension by elderly listeners," *Journal of Speech and Hearing Research*, vol. 26, no. 3, pp. 373-377, 1983.

[9] R. M. Uchanski, "Clear speech," in D. B. Pisoni and R. E. Remez. *The Handbook of Speech Perception*. Oxford, UK: Blackwell, pp 207–235, 2005.

[10] E. Zhu, E, and I. Bergom, "Lecture Capture: A Guide for Effective Use. Center for Research Learning and Technology," University of Michigan, 2010.