December 2021

Normalizing Non-Linear Speech Speed for Maintaining Listener Comprehension at Increased Playback Speeds

Malcolm Slaney

Normalizing Non-Linear Speech Speed for Maintaining Listener Comprehension at Increased Playback Speeds

Abstract:

This publication describes methods of normalizing the speed of non-linear speech by applying an algorithm to allow for improved listener comprehension at increased playback speeds. The algorithm computes an amount of tension for a given audio file and subsequently computes a running average of the tension. A high-pass filter is then applied to the tension to remove the average tension. The resulting audio file allows a listener to increase playback speed or maintain a desired average speed while retaining comprehension.

Keywords:

tension, non-linear speech, speech pattern, low-pass filter, high-pass filter, normalize, consonant, bandpass filter, low-cutoff filter, compression rate, playback, comprehension

Background:

Speech "tension" refers to a combination of speed of change in speech and the amount of energy in a particular speech pattern. For example, consonant sounds often have high tension and are difficult to comprehend when the speech rate increases, but certain vowel sounds have low tension and may be sped up while still allowing for listener comprehension.

The playback rate of an audio file may be slowed down or sped up by a listener (e.g., a user listening to a podcast on their smartphone may increase the playback rate to 2.5 times the default speed). There are various criteria that can be used, including speech tension, for modifying the playback rate. The problem with using speech tension as criteria for speeding up or slowing down

a given audio file is that maintaining listener comprehension at increased speed rates depends on the type of speech. For example, speech with a slower cadence (e.g., a southern accent) is easier to speed up while maintaining listener comprehension than speech with a faster cadence (e.g., a New York City accent). Therefore, it is difficult to speed up a given speech pattern to a set playback rate (e.g., 2.5x speed) while maintaining a listener's ability to comprehend the given speech pattern.

As is known in the literature, the tension for an audio file may be computed using a linear formula:

$$g(t) = a\,((E(t) - ME) + b(S(t) - MS))$$

Where:

- $g(t)$ represents an audio-tension value
- $E(t)$ represents a local-emphasis estimate for the speech utterance being compressed
- $S(t)$ represents a relative speaking-rate estimate
- ME and MS are the mean values of the local-emphasis and speaking-rate estimates, respectively
- $a > 0$ is a constant-valued coefficient
- b is a constant-valued coefficient, with $b > 0$ for time compression and $b < 0$ for time expansion

Description:

The following describes methods of normalizing the speed of non-linear speech by applying an algorithm to allow for improved listener comprehension at increased playback speeds. The disclosed speech-processing algorithm utilizes the inhomogeneous properties of an input

speech (e.g., audiobooks, classroom lectures, videos, vlogs) for normalizing the speed of speech on a computing device (e.g., smartphone) to allow for improved listener comprehension at increased playback speeds. For example, particular types of speech (e.g., a pause in speech) may be sped up while other parts of speech (e.g., consonant sounds) remain at an original speed because increasing the speed would otherwise result in decreased listener comprehension of the speech.

Normalizing the speed of non-linear speech may be accomplished by applying the described algorithm to an audio file. The algorithm first computes the tension of the audio file as a measure of speech "tightness." The speed of speech with inherently high tension cannot be increased without decreasing listener comprehension, while the speed of speech with inherently low tension can be increased while maintaining listener comprehension.

After the algorithm computes tension, a running average of tension for the audio file is calculated at specific average intervals (e.g., three seconds, thirty seconds). A high-pass filter is applied to the audio file to remove the average tension in the signal by subtracting the running average of the tension from the computed tension. The result is a tension signal that averages to zero. A listener may then increase the playback speed of the audio file while still comprehending the speech at increased playback speeds. For example, a user of a smartphone desires to play an audio file at 2.5x playback speed, the algorithm produces a tension signal that averages to zero and allows for the specific rate of 2.5x playback speed with maintaining listener comprehensibility.

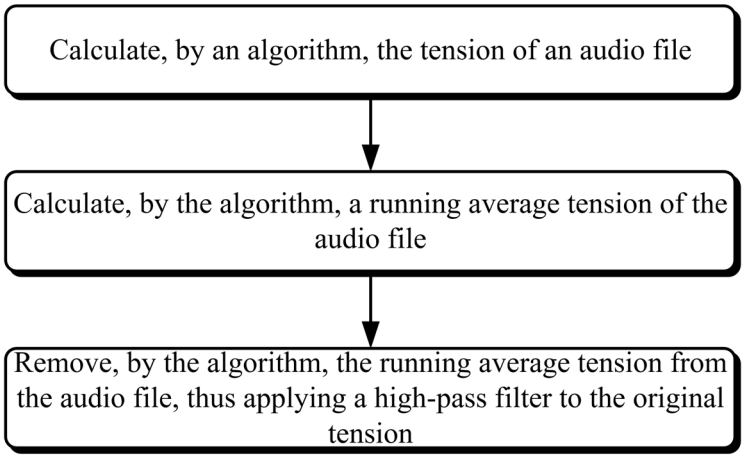

Figure 1

Figure 1 is a flow chart that illustrates a method for removing average tension from an audio file. First, tension is identified for an audio file by the algorithm. Subsequently, a running average of the local tension is calculated by the algorithm. The running average tension is then removed from the audio file by subtracting the running average to maintain listener comprehension at increased playback speeds.

Figure 2 and Figure 3 illustrate plots of the computed tension and running average tension for a particular speech segment in an audio file. Time intervals in seconds are represented along the x-axis and tension is represented on the y-axis. In the figures, the computed tension is represented by a blue line and an orange line represents the running average tension relative to the computed tension, as computed by the algorithm. In Figure 2, the running average tension is illustrated for a three-second running average for a particular audio file. In Figure 3, the running average tension is illustrated with a thirty-second running average for a particular audio file. The running average tension over thirty-second intervals is closer to an average target tension of zero than the running average tension over three-second intervals.

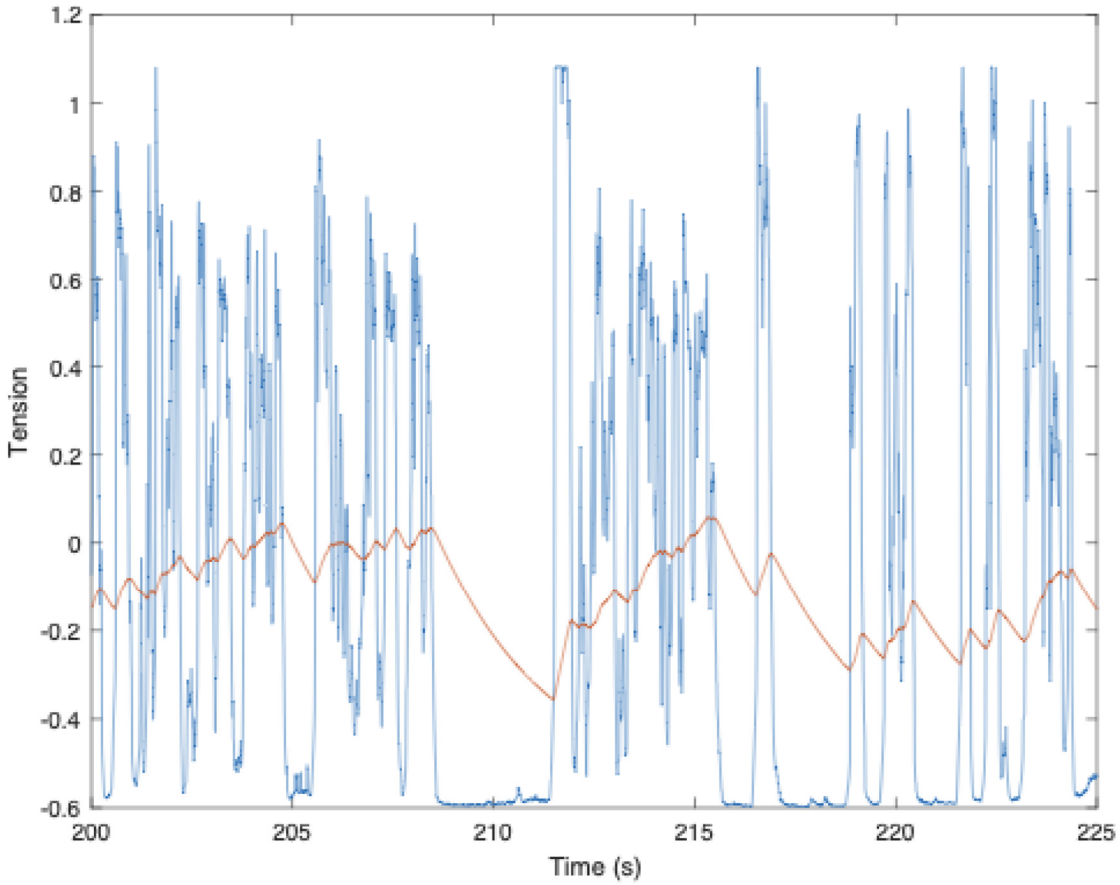

Figure 2

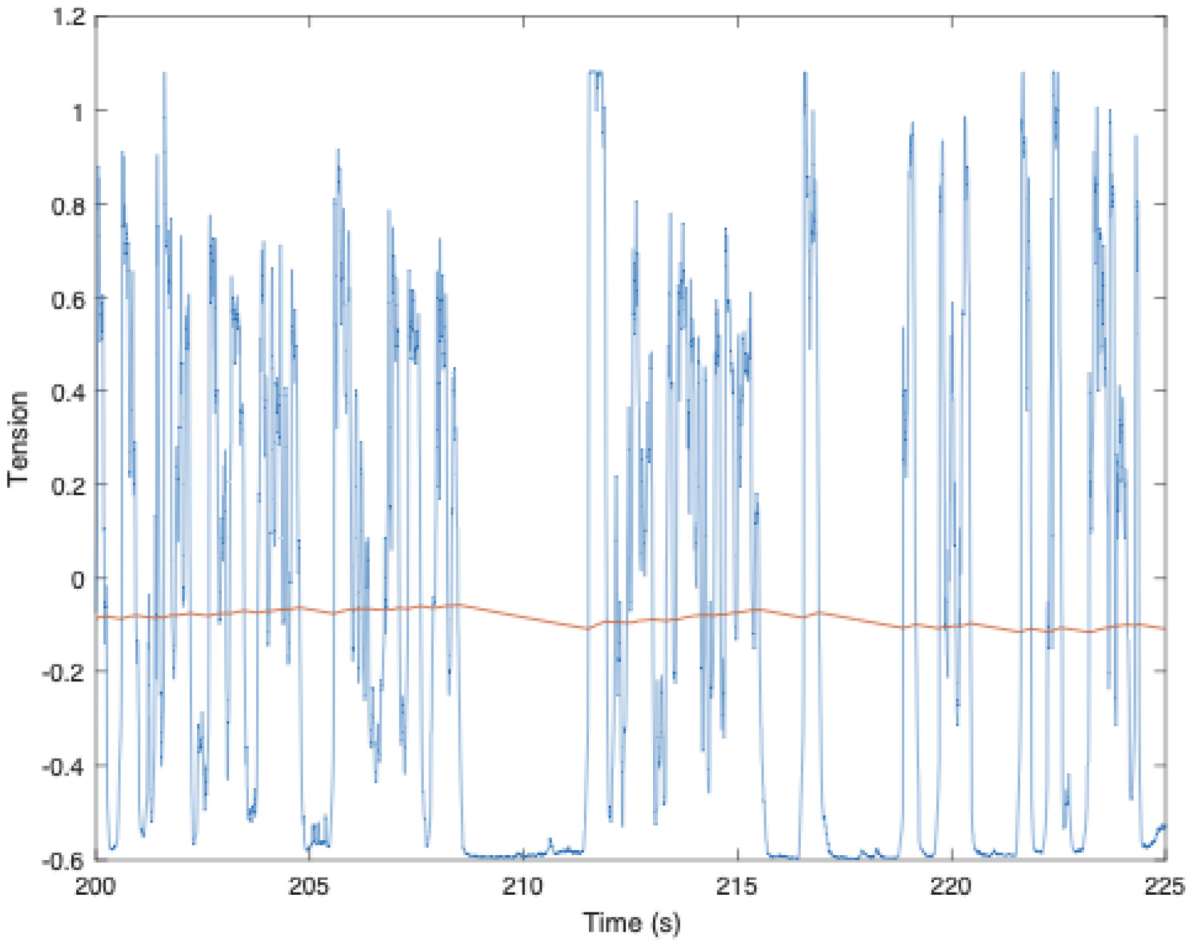

Figure 3

Further to the above descriptions, a user (e.g., a user of a computing device utilizing the disclosed speech-processing algorithm) may be provided with controls allowing the user to make an election as to both if and when systems, applications, and/or features described herein may enable collection of user information (e.g., information about a user's listening behaviors, a user's music selection, a user's voice patterns, a user's preferences, a user's current location), and if the user is sent content and/or communications from a server. In addition, certain data may be treated in one or more ways before it is stored and/or used, so that personally identifiable information is removed. For example, a user's identity may be treated so that no personally identifiable information can be determined for the user. In another example, a user's geographic location may

be generalized where location information is obtained (such as to a city, ZIP code, or state level), so that a particular location of a user cannot be determined. Thus, the user may have control over what information is collected about the user, how that information is used, and what information is provided to the user.

References:

[1] Patent Publication: US20190318758A1. Adjusting Speed of Human Speech Playback. Priority Date: August 15, 2017.

[2] Patent Publication: WO199746999. Non-Uniform Time Scale Modification of Recorded Audio. Priority Date: June 05, 1996.

[3] Patent Publication: US20070033057A1. Time-Scale Modification of Data-Compressed Audio Information. Priority Date: December 17, 1999.

[4] Dellwo, Volker. "Choosing the Right Rate Normalization Method for Measurements of Speech Rhythm." Division of Psychology and Language Sciences, University College London. (January 2009) https://www.aisv.it/PubblicazioniAISV/V_AISV/Articoli/Dellwo.pdf.

[5] Covell, Michele, Withgott, Margaret, and Slaney, Malcolm. "Mach1 for Nonuniform Time-Scale Modification of Speech: Theory, Technique, and Comparisons." (1997) https://engineering.purdue.edu/~malcolm/interval/1997-061/.

[6] Covell, Michele, Withgott, Margaret, and Slaney, Malcolm. "Mach1 for Nonuniform Time-Scale Modification of Speech: Theory, Technique, and Comparisons." Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Seattle WA, May 12-15, 1998. https://engineering.purdue.edu/~malcolm/interval/1997-061/writeup.html.