Improving intelligibility of time-scale compressed speech for visually impaired and sighted listeners

Panagiotis Pantalos[1], George Kafentzis[1], Anna Sfakianaki[2], Yannis Stylianou[1]
[1]University of Crete, [2]University of Ioannina
panoswork67@gmail.com, kafentz@csd.uoc.gr, asfakianaki@uoi.gr, yannis@csd.uoc.gr

In today's fast-paced world, text on social media and online education platforms is being replaced or enhanced with speech recordings, podcasts, and audiobooks. Accelerating speech recordings is often necessary to facilitate quicker information absorption. This is particularly beneficial for visually impaired individuals who rely on screen readers on their mobile devices. However, rapid speech, or time-scale compressed speech, tends to be less intelligible because critical speech components—such as transient sounds, plosives, and fricatives—are often lost. These components are typically non-stationary and play a crucial role in distinguishing syllables and words.

This study focuses on implementing algorithms to preserve non-stationary speech features, aiming to improve the intelligibility of time-scale compression. The experiments are based on the Waveform Similarity Overlap-and-Add (WSOLA) method. Speech waveforms are analyzed for their non-stationarity using simple time-domain and frequency-domain criteria. The first criterion (C1) relies on frame-by-frame RMS energy analysis, while the second (C2) uses Line Spectral Frequency (LSF) analysis. A hybrid criterion (C3) combines both approaches. The algorithms were tested on the GrHarvard dataset, which contains 720 Greek sentences balanced for phonemic content and spoken by both genders.

Two experiments were conducted involving both sighted and visually impaired participants. The first experiment compared uniform WSOLA, non-uniform C1-based WSOLA, and non-uniform C3-based WSOLA to evaluate the impact of protective measures on speech intelligibility. Results indicated that C1-based WSOLA performed best in both intelligibility and user preference, followed by C3-based WSOLA, with uniform WSOLA ranking lowest. The observed differences were statistically significant in most cases.

The second experiment assessed the same methods under equal words-per-minute (WPM) conditions, making the distinction between methods more difficult. While the C1-based method generally delivered the best intelligibility, the differences between methods were less pronounced, primarily due to variations in the size of stationary and non-stationary speech components. This made it challenging to definitively determine the superior method in terms of both preference and intelligibility. In addition, preliminary tests showed that, counter to expectations, visually impaired listeners did not significantly outperform sighted listeners.

Future research could involve fine-tuning the stationarity detection algorithm, exploring alternative time-scale compression models (such as the Harmonic+Noise model), and conducting additional experiments to gain further insights into the performance of the proposed methods.