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

Panagiotis Pantalos[1], George P. Kafentzis[1], Anna Sfakianaki[2], Yannis Stylianou[1]
[1]University of Crete, Greece
[2]University or Ioannina, Greece

https://doi.org/10.36505/TheLinguisticProceedings/2025/17/02/016/000702

Abstract

Time-scale compression enables faster speech playback but often reduces intelligibility, especially under high compression rates where non-stationary speech components are distorted. This work investigates improving intelligibility for visually impaired and sighted listeners by protecting non-stationary regions during time-compression. Using Waveform Similarity Overlap Add (WSOLA), we propose a protection method that adapts scale factors based on three non-stationarity criteria derived from Root-Mean-Square (RMS) energy and Line Spectrum Frequencies. Experiments with visually impaired and control participants evaluate intelligibility and listener preference across uniform and protected WSOLA variants. Results show that RMS-based-protected WSOLA improves intelligibility, while equal word per minute comparisons reveal smaller perceptual differences. Findings highlight the importance of preserving transient information for accessible high-speed speech.

Keywords: speech, transformations, intelligibility, visual-impairment, perception

Introduction

Time-scale compression (TSC) of speech enables listeners to consume spoken content significantly faster. This is especially beneficial for visually impaired individuals, who often depend on screen-reading software and routinely listen to speech at speeds far beyond those comfortable to sighted individuals. However, as playback speed increases, speech intelligibility typically declines. The primary reason is that non-stationary speech components – brief transients, plosives, fricatives, and rapid transitions – tend to be lost or heavily distorted when compression is too aggressive.

Time-domain TSC algorithms such as the Waveform Similarity Overlap-Add method (WSOLA) provide high-quality compression for many types of audio. WSOLA preserves local periodicity by adaptively shifting analysis windows to maintain waveform similarity. Despite its success, when high compression factors are applied, WSOLA still reduces or eliminates transient regions. This leads to a characteristic loss of clarity, particularly in languages containing many stop consonants or short syllables. As visually impaired individuals often rely on rapid audio consumption, the loss of intelligibility becomes a barrier to accessibility and efficient information processing.

To address this issue, we explore a method that selectively protects non-stationary regions during time-scale compression. Rather than applying a single uniform compression factor across the signal, we extract time- and frequency-domain features to identify portions of the waveform where intelligibility-critical content appears. These segments are then compressed less aggressively, allowing transient cues to remain intact while more stationary voiced regions undergo stronger compression. The goal is to balance faster playback with speech clarity.

This work proposes and evaluates three non-stationarity detection criteria. Integrated into a non-uniform WSOLA framework, these criteria dynamically adjust the local compression factor. We evaluate this system using listening tests with both sighted and visually impaired individuals to determine whether protecting non-stationary content yields measurable improvements in intelligibility and listener preference.

Methods

Materials

Our experiments utilized the GrHarvard corpus (Sfakianaki, 2021), a phonetically balanced Greek sentence dataset designed to parallel the structure of the classic Harvard/IEEE corpus. This dataset contains 720 sentences organized into 72 lists of 10 sentences each. The design ensures broad phonetic coverage, consistent sentence length, and controlled phonotactic structure across speakers. Each sentence includes exactly five keywords, making the corpus well-suited for intelligibility experiments where keyword recognition serves as a standardized performance metric.

For the purposes of this study, only a subset of the corpus was used. Specifically, we selected four lists: two for intelligibility tests and two for preference tests. This choice balanced the need for controlled experimental design with the time demands placed on participants, particularly visually impaired individuals.

Before applying time-scale compression, all recordings were preprocessed by removing leading and trailing silences. This ensured that time-scale factors were applied only to active speech content and not distorted by silent intervals.

Experimental design

The backbone of our approach is the Waveform Similarity Overlap-Add (WSOLA) algorithm, a widely used time-domain method for modifying speech rate without affecting pitch. WSOLA operates by splitting the signal into overlapping frames, then adaptively shifting these frames to maximize waveform similarity when reconstructing the compressed output. This mechanism helps preserve local periodicity, making WSOLA particularly effective for voiced speech. However, like other OLA-based methods, WSOLA can shrink or skip rapid transients when compression is high, reducing intelligibility.

To address this limitation, we introduce a non-stationarity detection frontend that modifies WSOLA's local scale factors. Instead of prescribing a global time-scale ratio, our method generates a time-varying scale factor sequence where each value corresponds to the stationarity level of the underlying signal frame. Regions identified as non-stationary receive milder compression, preserving critical phonetic cues, while stationary voiced regions undergo stronger compression. This adaptive factor is then fed to the WSOLA backend, which adjusts frame selection accordingly. Non-stationarity detection is performed using three criteria derived from prior work on transient analysis and spectral dynamics. The first criterion (C1) tracks rapid changes in frame RMS amplitude, which highlights plosive bursts and other abrupt energy shifts. The second criterion (C2) uses the gradient of Line Spectral Frequencies (LSFs) fitted across time, capturing changes in spectral envelope shape associated with formant movement and transitions. The third criterion (C3) combines C1 and C2 to leverage both temporal and spectral cues for improved robustness.

Results and discussion

We conducted two sets of listening experiments involving both sighted participants and visually impaired participants, all of whom had prior experience listening to spoken Greek. The first experiment compared uniform WSOLA to the C1-protected and C3-protected versions under standard compression rates. Participants listened to each sentence only once, and intelligibility was measured as the percentage of correctly reported keywords. Results demonstrated a significant advantage for the C1-protected WSOLA method, which consistently yielded the highest intelligibility scores across most compression factors.

Sighted listeners showed clear and statistically significant preferences for the C1-protected method. They reported better preservation of plosives, clearer consonant–vowel transitions, and less smearing of rapid onsets. Preference test outcomes aligned with intelligibility scores, revealing that C1-based protection not only improved recognition but also created a more pleasant listening experience. C3-protected WSOLA also improved performance over uniform WSOLA but to a lesser degree, likely due to its noisier behaviour in stable voiced regions.

Visually impaired participants displayed similar trends, though statistical significance was limited due to their small sample size. These participants often listen to speech at extremely high playback speeds, and their auditory processing has adapted accordingly. In many cases, they demonstrated higher baseline intelligibility for compressed speech than sighted listeners. Nevertheless, they still showed improved recognition with non-stationarity-protected compression, especially at moderate compression values where transient cue preservation remained crucial.

A second experiment evaluated all methods under equal words-per-minute (WPM) constraints by adjusting output lengths. This reduced differences in listening duration but also minimized acoustic differences between methods. Across both participant groups, distinctions became more subtle and statistical significance weaker. Even so, C1-protected WSOLA maintained a slight intelligibility advantage, confirming its robustness even under constrained conditions. However, the reduced perceptual separation makes this scenario more challenging for listeners, and larger sample sizes would be needed to confirm significance in future studies.

Conclusions

This work demonstrates that protecting non-stationary regions during time-scale compression significantly improves the intelligibility of fast speech, particularly when using WSOLA as the underlying transformation. The RMS-based criterion (C1) emerged as the most effective detector of linguistically important transient events, leading to the best intelligibility and strongest listener preference across both sighted and visually impaired participants. Future work should refine the non-stationarity detection mechanism, explore pitch-synchronous segmentation, evaluate more advanced TSC models, and conduct larger-scale listening tests with visually impaired populations. Integrating these techniques into assistive technologies and screen-reading systems has strong potential to enhance accessibility for users who rely on rapid speech playback in daily life.

References

Choi, D., Kwak, D., Cho, M., Lee, S. 2020. "Nobody speaks that fast!" An empirical study of speech rate in conversational agents for people with vision impairments. CHI Conference on Human Factors in Computing Systems, 1–13.

Kapilow, D., Stylianou, Y., Schroeter, J. 1999. Detection of non-stationarity in speech signals and its application to time-scaling. Proc. 6th European Conference on Speech Communication and Technology, 2307-2310.

Pantalos, P. 2023. Exploration of non-stationary speech protection for highly intelligible time-scale compression (Master's thesis). University of Crete, Greece.

Sfakianaki, A. 2021. Designing a Modern Greek sentence corpus for audiological and speech technology research. Proc. 14th International Conference on Greek Linguistics (ICGL14), 1119-1129. University of Patras, Greece.

Verhelst, W., Roelands, M. 1993. An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech. Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2, 554–557.