DOI: 10.61782/fa.2025.0881

FORUM ACUSTICUM
EURONOISE 2025

THE REVOLUTION IS HERE: DEEP AUDIO PROCESSING REDEFINES OFFLINE AND REAL-TIME AUDIO EXPERIENCES

Bar-Yosef Yossi[1*] **Thor Yami**[1] **Fink Nir**[2,1]

[1] BETTEAR – Accessibility Technologies Development LTD, 33 HaBarzel, Tel-Aviv, Israel

[2] Department of Communication Disorders, Acoustics and Noise Research Lab in the Name of Laurent Levy, Ariel University, Ariel 40700, Israel.

ABSTRACT

Deep Audio Processing (DAP) introduces a novel approach to personalized audio optimization, bridging established techniques such as time stretching, remixing, and audio channel separation with a groundbreaking dual-mode application for offline and real-time scenarios. While traditional audio processing methods excel in specific domains, their adaptation to dynamic real-time environments or offline pre-processing with synchronization constraints has been limited. DAP addresses this gap by integrating adaptive latency management and user-specific audio adjustments to deliver an enhanced listening experience tailored to diverse hearing capabilities and preferences. Key innovations in DAP include latency reduction techniques leveraging silent or non-speech segments and extended temporal flexibility for optimizing speech rates, singer-to-instrument ratios, and dynamic loudness ranges. Additionally, machine learning–driven auditory scene classification dynamically adjusts acoustic weights and parameters, optimizing intelligibility and comfort. This unified framework supports applications ranging from live-streaming accessibility enhancements to personalized audio delivery in cinema, teleconferencing, and music playback.

Keywords: *deep audio processing, source separation, time scale modification, real-time processing, accessibility*

1. INTRODUCTION

Audio experiences are central to communication, entertainment, and education. Yet common consumer systems lack flexible, adaptive approaches to personalize audio for listeners with diverse hearing capabilities-particularly those requiring specialized real-time adjustments. Traditional settings such as "Movie Mode" or "Rock Mode" provide fixed equalization filters without dynamically adapting to listener needs or acoustic conditions.

Recent progress in deep learning has enhanced the ability to analyze and manipulate complex audio signals, particularly through advanced source separation [1,2] and non-uniform time scale modification [3]. However, fully deploying these technologies in everyday consumer applications remains challenging, as two main issues persist:

1. **User-specific personalization:** Beyond simple equalization, listeners may require adjustments of signal-to-noise ratio (SNR) and speech tempo that reflect their hearing needs or preferences.
2. **Synchronization constraints:** Maintaining accurate sync across events is essential in real-time or time-sensitive contexts (e.g., live streams or cinematic audio). Excessive time-stretching can introduce objectionable delays and user disapproval.

Deep Audio Processing (DAP) addresses these limitations by integrating classical audio methods (channel separation, time stretching, remixing) with modern machine learning algorithms for source separation and content recognition. A central contribution is DAP's non-uniform time scale modification (NU-TSM) procedure, which balances

[*] ***Corresponding author***: yossi@bettear.com.

the user's desired speech rate with the need to remain within allowable latency limits. The main contributions of this paper are:

1. **Personalized Deep Audio Processing System:** We propose an assistive architecture that seamlessly unifies audio source separation, content classification, and user-driven mixing/time-stretch control.
2. **Time-Scale Modification Under Latency Constraints:** We formulate a synchronization problem that constrains how speech segments can be stretched without violating lip-sync or real-time requirements.
3. **Speech Rate Preference Experiment:** We present an empirical user study measuring preferred speech rates in noisy backgrounds for listeners with varying degrees of hearing loss, informing the default target speech rates presets for our DAP system.

2. METHODS

2.1 System Architecture

Figure 1 provides a high-level depiction of the DAP system. The system operates in two layers:

- **Processing Pane (Upper):** Source separation, mixing, and time-scaling.
- **Analysis & Control Pane (Lower):** Content classification, user preference retrieval, and adaptive logic for mixing and stretching.

The input is an original audio signal plus user-specific settings (e.g., target SNR and preferred speech rate). The output is a newly reconstructed signal, where speech intelligibility or vocals are enhanced according to personalized criteria.

Next, we briefly describe the basic building blocks of the DAP system.

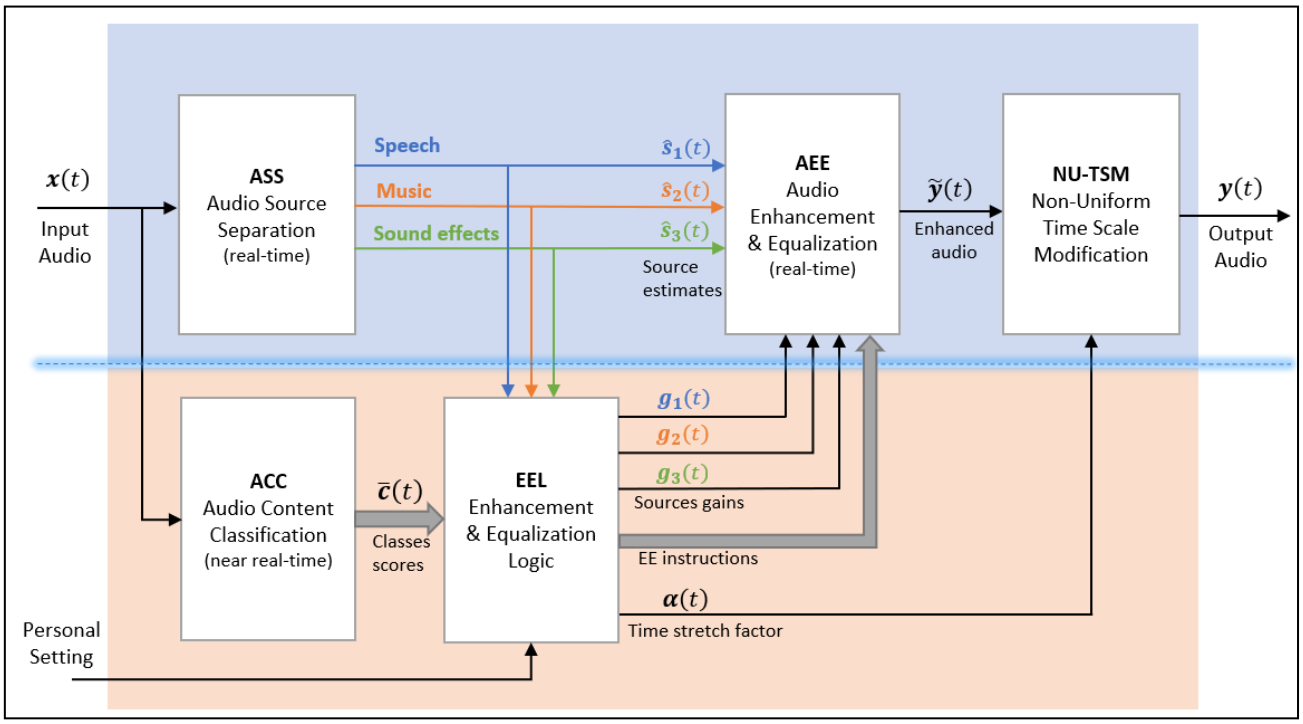

Figure 1. High level architecture of the Deep Audio Processing (DAP) system for assistive audio applications, illustrating the integration of audio capturing, source separation, content classification, enhancement logic, and audio enhancement modules. The upper pane is the "processing pane" and the lower is the "analysis & control pane".

2.1.1 Audio Source Separation (ASS)

The ASS module splits incoming audio into constituent sources such as speech/vocals, music, and sound effects. It uses a deep neural network architecture inspired by the MRX model [1]. Separated sources are crucial for fine-grained manipulation - particularly for adjusting gain or stretching specific elements (e.g., speech) without affecting background signals. Given an input signal $\boldsymbol{x}(t)$, the goal is to decompose it into constituent sources. This step is critical for enabling individualized control over each source's gain and temporal characteristics. Our current implementation follows the "cocktail-fork" concept [1], that separates the audio into three sources, speech, music, and sound effects (sfx). In the more general case, the output of the ASS module is a set of N source estimates,

$$\{\hat{\boldsymbol{s}}_1(t), \hat{\boldsymbol{s}}_2(t), \dots, \hat{\boldsymbol{s}}_N(t)\} \tag{1}$$

fed into subsequent processing blocks for enhancement and personalization.

2.1.2 Audio Content Classification (ACC)

In parallel, a convolutional neural network (CNN) classifies audio content into categories (inspired by [4]) (e.g., conversation, music, ambient noise). This "pseudo" real-time classification produces a time-varying score vector indicating the probability of each category.

$$\bar{\boldsymbol{c}}(t) = [\, c_1(t), \dots, c_K(t) \,]^T \tag{2}$$

where each component $c_k(t)$ reflects the likelihood of a specific content category k, and K is the total number of contents categories. These scores guide the system's adaptive mixing and time-scaling decisions. The index t refers to the corresponding audio frame starting time[1].

2.1.3 Personalized Targets

User preferences for (i) speech/vocal SNR relative to accompaniment and (ii) speech rate in various noise conditions is stored in an external module. These preferences can be specified manually (e.g., through an interface) or derived from user data. The system references these targets to dynamically apply mixing gains and stretching factors. These targets guide the subsequent processing steps within the EEL module.

[1] For simplicity we use the same notation for time indexing, while in all analysis modules and in most processing modules it applies to the index of the audio frame

2.1.4 Enhancement & Equalization Logic (EEL)

The EEL module interprets classification scores and user targets to generate control logic signals to subsequent processing blocks. It carries out:

2.1.4.1 Mixing Logic for SNR Target

In this module gains are computed for each source to match the desired speech-over-background ratio. The time-varying gain parameters are denoted as

$$\{\boldsymbol{g}_1(t), \boldsymbol{g}_2(t), \dots, \boldsymbol{g}_N(t)\}\,. \tag{3}$$

Practically, the SNR calculation is performed frame-by-frame. The calculated gain parameters are then passed to the mixer, which applies the adjustments to ensure that the desired SNR is maintained throughout the audio signal.

2.1.4.2 Speech Rate Estimation and Time Scale Logic

The time scale logic module focuses on the dynamic aspects of speech. It first estimates the local speech rate from the separated speech signal $\hat{\boldsymbol{s}}_1(t)$, to quantify the speech tempo in terms of phonemes-per-second (PPS). There are various approaches for approximating the PPS. The simple ones may consider spectral transition differences between time frames, while the more complicated may use a phonetic recognizer or speech recognition (more applicable to offline mode). This measurement is critical for determining how to adjust the audio timescale. Based on this estimation and the user's preferred speech rate, the module computes an adaptive time-scale factor $\boldsymbol{\alpha}(t)$ to achieve the desired speech rate,

$$\boldsymbol{\alpha}(t) = \boldsymbol{\alpha}\big(\bar{\boldsymbol{c}}(t), \mathrm{PPS}(t)\big)\,, \tag{4}$$

where $\bar{\boldsymbol{c}}(t)$ represents the classification scores from the ACC block, and $\mathrm{PPS}(t)$ represent the estimated momentary speech rate at time t. Eqn. 4, emphasizes that the scale factor is not only adjusted according to the momentary PPS, but is also oriented to the auditory scene, whether it contains merely speech (and may need to be stretched) or if it contains a sound effect (for instance the sound of a gunshot) that needs to maintain the same ratio (factor of 1 indicating no need for stretching), or whether it is a non-speech or a pause section that may be shortened.

The $\boldsymbol{\alpha}(t)$ factor guides the NU-TSM process, allowing for localized adjustments that slow down or speed up audio

segments while preserving the natural timing of non-speech elements.
Together, these two components enable the EEL module to output two sets of control signals: the gain parameters for achieving the desired SNR and the adaptive time-scale factor for targeted speech rate modification. This logic-based stage lays the groundwork for the subsequent re-mixing and time-stretching that occur in the Audio Enhancement & Equalization and NU-TSM modules respectively, ensuring that personalized audio processing meets user preferences.

2.1.5 Audio Enhancement & Equalization (AEE)

The AEE module rebuilds the final audio from the separated sources using the gains calculated by the EEL. Additional filters or effects may also be applied to each separated source as desired. The remixing is performed by:

$$\tilde{\boldsymbol{y}}(t) = \sum_{i=1}^{N} \boldsymbol{g}_i(t)\hat{\boldsymbol{s}}_i(t). \quad (5)$$

2.1.6 Non-Uniform Time Scale Modification (NU-TSM)

The final component, NU-TSM, applies the adaptive stretch factor to slow down high-rate speech segments and optionally compress silence or non-speech segments. A key challenge here is ensuring overall synchronization. Excessive stretching can cause noticeable latency or lip-sync mismatches, so the NU-TSM module operates within strict deviation bounds.

2.1.6.1 NU-TSM in Sync – Problem Derivation

Time scale modification in sync poses additional limitations on latency or on "run ahead" in offline scenarios. Typically, TSM algorithms relate to two parameters – the analysis frame step (which is a constant) and the synthesis frame step which is set according to the required stretch factor $\boldsymbol{\alpha}(t)$. when adding sync constraints, we wish to force the synthesis frames to be located inside a certain shift from the original location of the corresponding analysis frame, and this poses a new optimization problem, which is derived next. In the following derivation we use a discrete notation for the stretch factor $\alpha_t \equiv \boldsymbol{\alpha}(t)$, suggesting that t is a running frame index.

Problem derivation:
- Assume a given audio signal has T analysis steps, and a series of stretch factors (per frame) is given by $\{\alpha_0, \alpha_1, \dots, \alpha_t, \dots, \alpha_T\}$ where $\alpha_0 = 0$.
- Let $\boldsymbol{a}$ be the constant analysis step (samples or time units).
- Hence, a requested synthesis step at time t would be $\boldsymbol{a}\alpha_t$.
- The analysis frame position at time t is defined by:

$$x_t = x_0 + \boldsymbol{a}t \quad (6)$$

where x_0 is the initial position of the analysis.
- The "requested" synthesis position at time t follows the recursion

$$z_{t+1} = z_t + \boldsymbol{a}\,\alpha_{t+1} \quad , \quad (7)$$

where z_0 is the initial position of the synthesis (typically $x_0 = z_0$), and the momentary stretch time between two consecutive frames is defined by

$$z_{t+1} - z_t = \boldsymbol{a}\,\alpha_{t+1} \quad . \quad (8)$$

- The terms z_t are the requested synthesis positions without sync limitations. We seek a new set of synthesis positions, $\{y_0, \dots, y_t, \dots, y_T\}$ that aim to maintain the momentary stretch time of all steps, $z_{t+1} - z_t$, while satisfying strict sync conditions to the original analysis positions, defining L to be the allowed time deviation from the original position, the optimization problem is formulated as follows:

$$\begin{cases} \text{minimize} & \sum_{t=0}^{T-1} |(y_{t+1} - y_t) - (z_{t+1} - z_t)| \\ s.t. & |y_t - x_t| \le L \quad \forall \quad t \in \{0, \dots, T\} \\ & y_{t+1} - y_t \ge 0 \quad \forall \quad t \in \{0, \dots, T\} \end{cases} \quad (9)$$

Simplifying by replacing variables produces the following optimization problem,

$$\begin{cases} \text{minimize} & \sum_{t=0}^{T-1} |y_{t+1} - y_t - \boldsymbol{a}\alpha_{t+1}| \\ s.t. & |y_t - x_0 - \boldsymbol{a}t| \le L \quad \forall\ t \in \{0, \dots, T\} \\ & y_{t+1} - y_t \ge 0 \quad \forall\ t \in \{0, \dots, T\} \end{cases} \quad (10)$$

This formulation guarantees that the solution's time sync will not deviate from the preset boundaries however, the requested target scale is not guaranteed, it only tries to optimize it to the best effort.
Various real-time or offline approximation methods can solve this optimization efficiently. One approximation uses a greedy regressive procedure that may fit real-time scenarios due to its simplicity. In some conditions, the solution may be further simplified. For example, when assuming $\alpha_t \ge 1$ for all $t = 1, \dots, T$; the problem reduces to a convex problem with an analytical solution. In an offline scenario over an audio part of original length T_s where the requested total length is calculated by $z_T = \boldsymbol{a}\sum_{n=1}^{T} \alpha_n$ (following Eqn. (7)),

and the total margin is L before and after the speech part; setting $z_0 = x_0$, leads to a simple solution,

$$y_t = \begin{cases} z_t & \text{, if } z_T \le T_s + L \\ -L + z_t\left(\frac{T_s+2L}{z_T}\right) & \text{, if } z_T \ge T_s + 2L \\ -\left(\frac{z_T}{T_s} - 1\right) + z_t\left(\frac{T_s+L}{z_T}\right) & \text{else} \end{cases} \quad (11)$$

This offline solution obtains a "run ahead" time on y_0 (having a non-speech margin before the referenced part), allowing to compensate for the latency accumulated in the stretched part. This simplified result (Eqn. (11)) provides satisfactory outcomes in many situations.

2.2 User Experiment: Speech Rate Preferences Noise

2.2.1 Participants

A total of 44 participants with bilateral hearing impairment took part in the study. Participants were categorized by the severity of their hearing loss (HL) [5]; mild HL (n = 23), moderate HL (n = 8), severe HL (n = 6), and profound HL (n = 7). All provided informed consent and completed an online auditory profile questionnaire prior to testing, including hearing loss characteristics, onset, stability, and type of hearing impairment. The experiments were conducted remotely, with participants using their own listening devices in a quiet home environment. Stimuli and Equipment

2.2.2 Stimuli and Equipment

Speech stimuli consisted of recordings of a professional female narrator in a studio environment. The narrator read aloud multiple pages from a novel, ensuring natural prosody and articulation. The speech stimuli were presented simultaneously with a separately recorded police siren. The speech and siren were mixed to create a controlled background noise condition, ensuring consistency across all participants. All audio files were presented in MP3 format (44,100 Hz, 32-bit sampling rate). Participants were instructed to use headphones or their regular hearing aids/cochlear implants (if applicable) and to avoid adjusting device settings during the experiment. For further information, see [5].

2.2.3 Procedure

The experiment consisted of two phases. In the preliminary task, each participant determined their most comfortable level (MCL) for listening to speech in the presence of background noise. This was done using two pre-prepared audio files: one containing the narrated recordings and the other containing a constant-level police siren noise. The noise was played at a fixed intensity, while participants used a slider to adjust the intensity of the narrator's voice until they reached the minimal SNR that allowed for comfortable speech understanding.
Once the participant's individual SNR was determined, they proceeded to the main task, where they adjusted the speech rate of the narrator while listening under the same background noise conditions and using the SNR established in the preliminary phase. Participants used a sliding control to modify the recorded speech rate to the maximum speed at which they could still comprehend the story content.

2.2.4 Data Collection and Analysis

Data collection was conducted remotely via an online interface as described in [5]. Only participants who fully completed the experimental tasks were included in the analysis. An SPSS software was used for statistical analysis. Preferred speech rates (mean ± standard deviation) were calculated for each participant and group. A one-way analysis of variance (ANOVA) was conducted to assess whether speech rate preferences differed significantly between groups. Post hoc tests (Tukey's HSD) were performed to evaluate pairwise differences. Effect sizes (η^2) were calculated to quantify the magnitude of observed differences.

3. RESULTS

3.1 DAP System Outcomes

Figures 2 and 3 illustrate qualitative examples of how DAP adjusts SNR and speech tempo over time. Figure 2 presents a snapshot of the DAP system in the time domain, over a short audio example. It shows the speech SNR adjustment obtained through the mixing gains, the speech rate measurement and the time-scale modification in term of accumulated latency. It is observed that the NU-TSM algorithm stretches the speech section (during the first second of the audio), followed by shortening the non-speech segment of the audio.
Figure 3 presents an example of an audio recording containing a heated debate between two people. One can observe the time-scaling applied non-uniformly, where the pre-treated higher speech rates receive a larger stretch factor, while the pauses in the speech sections are shortened to compensate for the latency caused by the stretching process.

Thus, the maximum latency does not exceed a predefined limit of $L = 240$ msec. This parameter can be adjusted according to user needs or even according to the audio content classification.

Clearly, the effectiveness of NU-TSM is highly influenced by the allowable sync deviation and the structure of the speech segments. When a conversation or cinematic audio track includes sufficient pauses, users can experience improved intelligibility without perceiving noticeable synchronization issues.

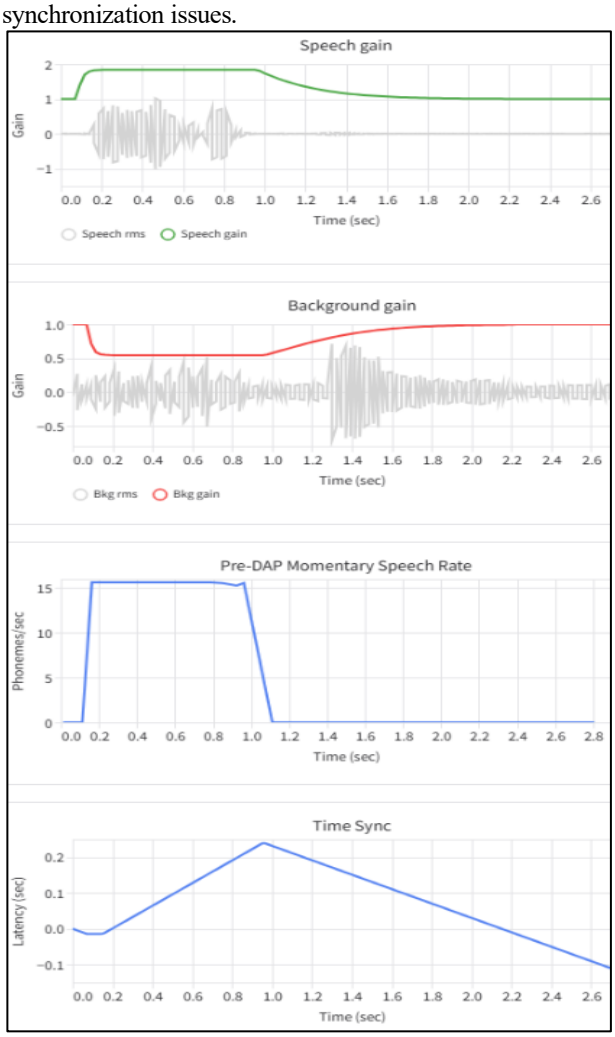

Figure 2. A snapshot of a processed audio example from the DAP system. The two upper plots describe the separated sources of 'speech' and 'background' (music and sound effects combined) with the gain factor (green and red gain curves) applied to emphasize speech. The third plot describes the approximated original speech rate, which is over 15 PPS, and the bottom plot describes the time scale modification in terms of latency. The speech part is stretched by almost 20%, allowing a latency of 240 msec, while the non-speech part is shortened to reduce the time latency. Notice that the line slope presents an effective scale factor.

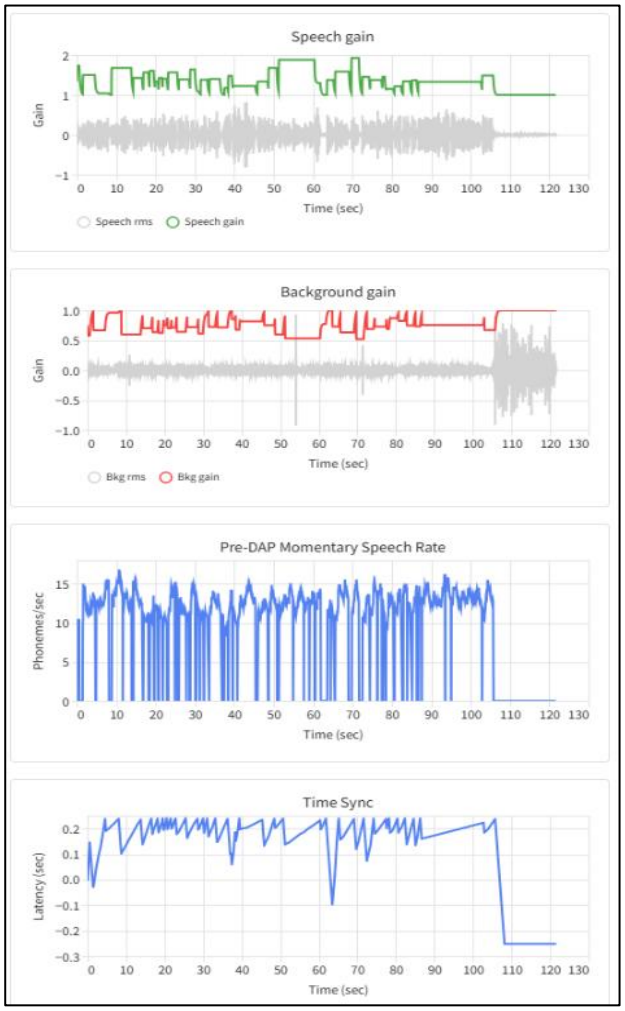

Figure 3. A snapshot of an audio recording of a heated debate. The target SNR is applied to most of the frames. Bursts of high speech rates exceed 15 PPS, and the time-scaling is applied non-uniformly. The maximum time deviation is 240 msec - as predefined in the system setting.

3.2 User Experiment Speech Rate Preference in Background Noise

Descriptive statistics for speech rate preferences in the presence of a background police siren are provided in Figure 4. Results show a general trend of decreasing preferred PPS with increasing hearing loss severity. Participants with mild HL preferred a mean PPS of 10.94, moderate HL preferred 10.42, severe HL preferred 9.30, whereas those with profound HL preferred a significantly slower rate of 8.51. On average, the preferred speech rate in background noise was 10.24 PPS.

A one-way ANOVA revealed a statistically significant effect of hearing loss severity on preferred PPS [$F(3,40) = 3.084$, $p = .038$, η^2=0.188]. Post hoc tests showed that the maximal preferred PPS for the perception of a female narrator speaking in background noise decreased significantly from mild to profound HL (-2.43 PPS, $p = 0.039$). No other significant group differences were observed.

These results suggest that individuals with greater hearing impairment prefer slower speech rates when background noise is present. Future studies should explore additional auditory factors influencing speech rate preference, such as spectral balance and dynamic range compression.

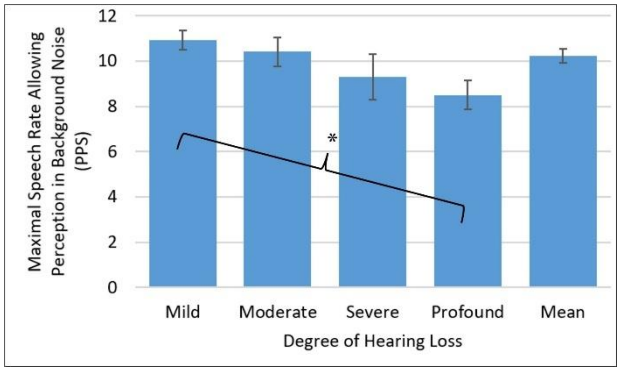

Figure 4. Maximal Speech Rate Allowing Perception (in PPS) of a female narrator in background noise across different hearing loss groups (Mild, Moderate, Severe, Profound and Mean). Error bars represent the standard error of the mean. ⋆ Indicates a statistically significant difference between the marked groups ($p < 0.05$).

4. DISCUSSION

This paper suggests a basic and general description for an assistive audio processing system. The practical system is more complicated and is not detailed here. Since the final output of the system depends on a series of different modules, source separator, content classification, auditory-scene oriented logic of time-scale modification, and time-constrained stretching algorithm, many independent variables may affect it. Because of this complexity, it is difficult to quantitatively validate the most influencing factors as they are entangled together. Future user experiments should address this issue. It also seems that to examine the quality of different algorithms, further comparative experiments should be conducted to provide empirical quantitative analysis.

While our study provides valuable primary insights about speech rate preferences in the presence of background noise, some limitations should be noted. First, hearing loss severity was self-reported rather than audiometrically verified, which may introduce variability in group classification. Future studies should incorporate objective hearing assessments to strengthen classification accuracy. Second, the experiment was limited to a single type of noise which in future experiments would be expanded to variable types of noise. The present study primarily examined speech intelligibility improvements. Additional research is needed to assess how DAP's strategies affect non-speech content such as music or sound effects, especially in scenarios where the artistic intent or emotional impact of the audio track must be preserved.

5. CONCLUSION

This paper has introduced a personalized DAP framework that combines advanced source separation, content-aware mixing, and NU-TSM to enhance audio in both offline and real-time contexts. By leveraging machine learning for audio content classification, DAP enables targeted improvements in speech intelligibility, adaptable speech-rate manipulation, and precise synchronization with visual media or other concurrent events. While the primary focus is on speech intelligibility, the underlying system can be adapted to enrich various user experiences - for example, by emphasizing a specific instrument or sound source.

Crucially, our NU-TSM approach shows how intelligibility can be improved without exceeding predefined latency constraints. We also introduced a user study on speech-rate preferences in noisy environments, establishing baseline presets that reflect listeners' auditory needs across different degrees of hearing loss. Although our findings suggest that DAP meets many user preferences, practical challenges persist - especially

regarding real-time processing demands and stringent latency requirements.
Moreover, fully understanding users' preferences and requirements necessitates subjective hearing experiments, due to the wide range of factors that shape listening experiences - hearing profiles, attention deficits, age, and content variability (e.g., music genres, cinematic audio, special effects). These complexities underscore the individualized nature of audio preferences and highlight the importance of extensive user-centric testing to capture diverse needs accurately.
Future research will address these constraints and further refine personalized audio processing by exploring more robust adaptation strategies, investigating broader content types, and incorporating formal audiometric evaluations. Ultimately, this work lays the groundwork for a versatile DAP system capable of serving a wide audience with varied listening preferences and challenges.

6. ACKNOWLEDGMENTS

The authors would like to express their gratitude to the participants who took part in this study and provided valuable insights and to Shaked Bigi for her assistance. This research was supported by a grant from The Israel Innovation Authority (Grant #83364) as part of the research and development project titled: "Development of Technology for the Accessibility of Auditory Content for Populations with Hearing Difficulties or Attention Deficit Disorders."

7. REFERENCES

[1] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[2] D. Petermann, G. Wichern, Z.-Q. Wang, and J. Le Roux, "The cocktail fork problem: Three-stem audio separation for real-world soundtracks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 876–880, 2022.

[3] A. Kupryjanow and A. Czyżewski, "A non-uniform real-time speech time-scale stretching method," in *Proceedings of the International Conference on Signal Processing and Multimedia Applications (SIGMAP)*, Seville, Spain, pp. 145–150, 2011.

[4] S. Soni, S. Dey, and M. S. Manikandan, "Automatic audio event recognition schemes for context-aware audio computing devices," in *2019 Seventh International Conference on Digital Information Processing and Communications (ICDIPC)*, pp. 23–28, 2019.

[5] Fink, Y. Thor, Y. Bar-Yosef. "Psychoacoustics of Remixing Music: Catering to Hard-of-Hearing and Normal-Hearing Audiences for Enhanced Experiences", in *Proceedings of Forum Acusticum (submitted). Malaga, Spain: European Acoustics Association (EAA),* 2025.