

A Structural Dynamics Based Visual Microphone

David Pinheiro de Sousa, Moisés Felipe Silva, Victor H. R. Cardoso and João C. W. A. Costa

Abstract— Visual Microphone enables passive acoustic recovery from subtle object vibrations captured in video sequences. Conventional approaches rely on multiscale phase decompositions with high computational cost. This work proposes a structural dynamics Visual Microphone framework operating directly on vectorized temporal pixel signals, avoiding phase based transformations. The method combines Principal Component Analysis and Complexity Pursuit within a temporally chunked architecture that estimates projection matrices once and reuses them throughout the sequence. Experiments on benchmark datasets achieved an 83% reduction in execution time and reduced pitch RMSE by 31% relative to traditional phase based methods, demonstrating efficient acoustic recovery from video.

Keywords— Visual Microphone, Blind Source Separation, Complexity Pursuit, PCA, Passive Audio Recovery.

I. INTRODUCTION

Visual Microphone is a technique that enables the passive recovery of audio signals from submillimeter mechanical vibrations captured by high-speed cameras [1]. The potential of this technique to estimate material properties from imperceptible movements was first explored in the work on Visual Vibrometry [2].

Historically, the isolation of these micromovements has relied on motion decomposition methods, such as phase-based processing [3] and video magnification techniques in the presence of large displacements [4], and modal identification from high-speed video using motion magnification [9]. Although robust against noise, these approaches rely on multiscale transformations, such as complex pyramids, which impose a high computational cost and hinder real-time processing [1], [3]. Recently, the use of computational imaging with Single-Pixel Imaging (SPI) has been proposed to reduce the volume of data during capture [5]. Unlike this approach, which focuses on acquisition hardware, this work aims to optimize the processing of conventional videos.

The rationale for optimization emerges from recent advances in video based structural dynamics and full field modal analysis. In these fields, latent dynamic responses are extracted directly from temporal pixel intensity variations through matrix factorization and blind source separation techniques [8], avoiding explicit geometric motion reconstruction. Methods based on traveling wave identification [6] and nonnegative matrix factorization for full field modal decomposition [7] have demonstrated that vibration signatures and dynamic behaviors can be recovered directly from video measurements with high temporal and spatial resolution. These developments suggest that passive acoustic recovery can similarly be interpreted as a latent vibration separation problem operating on temporal pixel responses.

Building on this foundation, this paper proposes a computationally efficient Visual Microphone pipeline that

operates directly on vectorized pixel intensity signals extracted from video frames. Unlike traditional approaches that rely on multiscale phase decompositions and complex steerable pyramids, the proposed method analyzes temporal pixel variations through matrix projections in a reduced latent subspace. The pipeline combines Principal Component Analysis (PCA) and Complexity Pursuit (CP) within a temporally chunked architecture. By estimating the projection and mixing matrices only during the first processing block and reusing them throughout the sequence, the method avoids iterative recalculations and substantially reduces computational latency.

The main contributions of this work are fourfold. First, we introduce a Visual Microphone formulation based on direct analysis of vectorized pixel intensity signals, avoiding complex phase-based pyramidal decompositions. Second, we propose a temporally chunked projection strategy that enables the reuse of PCA and CP subspaces across video batches, reducing computational cost. Third, we demonstrate that the proposed PCA+CP framework improves harmonic frequency recovery, reducing pitch RMSE relative to traditional phase-based approaches. Finally, we characterize the resulting space-time trade-off, achieving approximately 83% lower execution time at the expense of increased RAM utilization.

II. THEORETICAL BACKGROUND

The recovery of acoustic signals from video is modeled as a blind source separation problem applied to time series of luminance. Unlike geometric phase-based approaches that require iterative multiscale decompositions, the proposed spatial formulation vectorizes the sequence of frames into an observation matrix $X \in \mathbb{R}^{T \times N}$, where T represents the number of frames and N the total number of pixels in the Region of Interest (ROI).

Given the high correlation between adjacent pixels under acoustic excitation, PCA is applied to project X onto a reduced subspace $Y \in \mathbb{R}^{T \times K}$ $K \ll N$, mitigating stochastic sensor noise. The extraction of the latent acoustic source $s = w^T Y^T$ is performed via CP, which seeks the projection vector w that maximizes the temporal predictability of the signal by minimizing the Rayleigh quotient:

$$J(w) = \frac{w^T R_L w}{w^T R_S w}$$

where R_L and R_S represent the long-term and short-term covariance matrices of the subspace Y . The solution converges to the generalized eigenvalue problem $R_L w = \lambda R_S w$, in which the eigenvector corresponding to the smallest eigenvalue isolates the component with the most ordered temporal

David Pinheiro de Sousa, Victor H. R. Cardoso and João C. W. A. Costa, affiliated with Applied Electromagnetism Laboratory (LEA) at the UFPA, located in Belém. Contact e-mail: davidpinheirosousa.ds@gmail.com. Moisés Felipe Silva, affiliated with the Bandelier Technologies, NM, EUA.

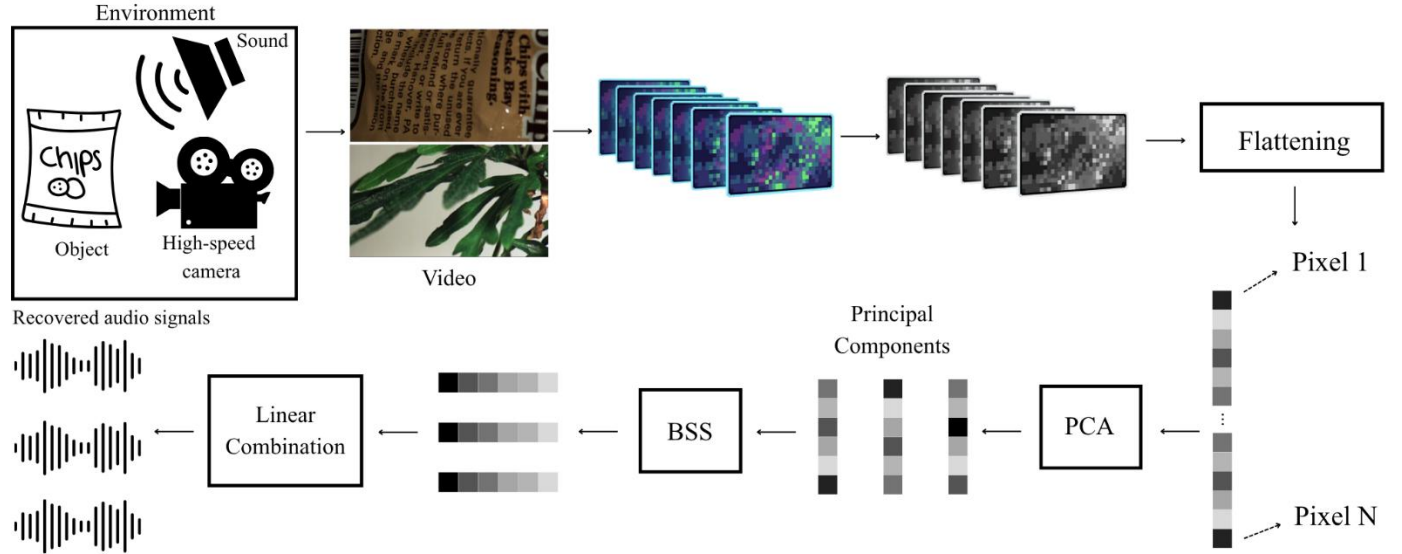


Fig 1. Diagram of the experimental setup and the proposed signal processing workflow. Sound pressure waves from an acoustic source induce submillimeter-scale micromotions on the target object, which are recorded by a high-speed camera. The resulting video data is then vectorized and processed directly as pixel-intensity signals, using Principal Component Analysis (PCA) for dimensionality reduction and Blind Source Separation (BSS) to isolate the latent acoustic signal.

structure, an intrinsic characteristic of the harmonic acoustic signal relative to background noise.

III. METHODOLOGY

The proposed approach aims to perform passive acoustic signal recovery from high-speed video sequences by exploiting micro-movements on the surface of various materials (such as flexible packaging and foliage). The computational pipeline departs from computationally expensive multiscale spatial decompositions [1], operating directly in the direct spatial-intensity domain via frame flattening, as an efficient adaptation of the modal analysis framework [6]. To enable the space-time trade-off and manage the high main memory consumption, data processing is segmented into temporal chunks (chunking). Figure 2 presents the methodology's flowchart, detailing the data estimation and matrix projection workflow.

A. Preprocessing and Temporal Filtering

The workflow begins by converting the 2200 fps video frames to grayscale. This sequence is vectorized via spatial row-major concatenation into an observation matrix $X \in \mathbb{R}^{T \times N}$ (T frames, N pixels). To isolate dynamic components, the mean frame (stationary background) is subtracted from the time series.

The matrix X is then subjected to a rigorous temporal filtering step. A 5th-order Butterworth bandpass filter is employed along the temporal dimension (the rows of X), with cutoff frequencies typically between 135 Hz and 1000 Hz, to isolate the band of acoustic interest. Concurrently, a cascade of notch filters is applied, designed to attenuate mechanical or lighting interference, mitigating the flicker effect induced by the 120 Hz power grid and its harmonics. To preserve the harmonic structure and avoid distortions in audio recovery, all temporal filtering is performed with zero phase delay.

B. Spatial Decomposition and Blind Source Separation

Direct source extraction from X is computationally intractable due to its high dimensionality. Thus, PCA [6]

concentrates the micro-movements' covariance, projecting the signal onto a lower-dimensional subspace retaining only the 8 principal components with highest variance. This empirically captures dominant structural modes and the acoustic signature while discarding stochastic sensor noise.

The extraction of the acoustic source in the reduced subspace is modeled as a BSS problem and solved using the CP algorithm [7]. By solving a generalized eigenvalue problem for short- and long-term covariance matrices, the CP algorithm isolates latent components based on their temporal predictability. The method proves to be robust for isolating the acoustic excitation, whose structured temporal dependence contrasts sharply with the projected non-acoustic residuals.

A key methodological aspect for the computational efficiency of this pipeline is matrix processing by temporal blocks. Since the PCA decomposition and the eigenvalue problem of the CP algorithm involve high memory allocation complexity, the spatial projection and mixing matrices are estimated exclusively during the processing of the first batch of frames. Assuming the spatial stationarity of the acoustic source and the stable physical structural behavior of the object throughout the short duration of the video, subsequent frame blocks are projected directly into this previously learned subspace. This logical architecture avoids iterative recalculations and is primarily responsible for the feasibility of the method's space-time trade-off.

C. Audio Reconstruction and Post-Processing

After extracting and temporally concatenating the latent sources across all video frames, the reconstruction of the acoustic signal does not use the global sum or linear combination of the subspace. This restriction is adopted because superimposing all components would tend to reintroduce structural and thermal noise that had already been segregated by the algorithm. Instead, a manual analytical inspection is currently performed, guided by variance, harmonic structure, and auditory coherence, to identify and

isolate the specific one-dimensional component that most purely encapsulates the mechanical-acoustic stimulus. While this manual step represents a limitation for fully automated or real-time implementations, future work will aim to automate this selection using algorithmic criteria based on these same features.

This isolated signal is then subjected to the conditioning stage (post-processing). First, a first-order pre-emphasis filter is applied, defined by the difference equation $y[n] = x[n] - 0.95x[n-1]$. The purpose of this filter is to equalize the spectral density of the reconstructed signal, compensating for the natural energy decay (characteristic of $1/f$ noise) at high frequencies and attenuating any residual low-frequency drifts.

Finally, an Automatic Gain Control (AGC) algorithm, operating with 0.5-second time windows, adjusts the local amplitude dynamics of the audio. This process evens out intensity fluctuations over time and prevents clipping, followed by peak normalization. This final processing ensures signal stability, enabling both auditory analysis and rigorous quantitative evaluation (e.g., SSNR and pitch error) relative to the reference audio (ground truth).

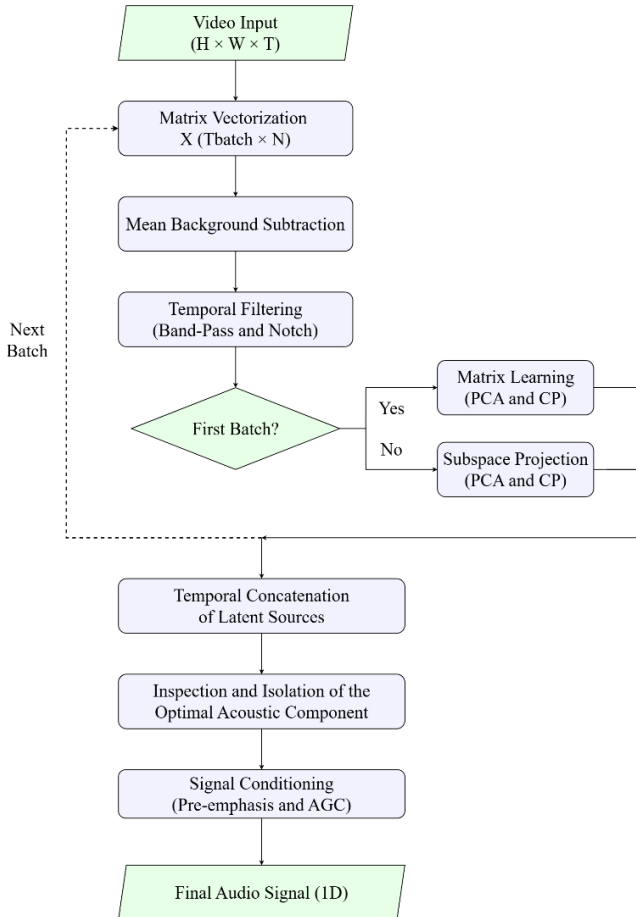


Fig 2. Computational Pipeline Flowchart.

IV. RESULTS

A. Qualitative Analysis and Component Selection

Figure 3 illustrates the comparison of waveforms and spectrograms, confirming the method's ability to recover the melody's frequency patterns.

The spectral and temporal analysis of the separated sources demonstrated how acoustic interaction behaves with different material properties. Visual inspection of the spectrograms and waveforms revealed that the acoustic signature manifests itself distinctly in the principal components extracted by the CP algorithm.

For the snack pack video, a rigid, lightweight, and highly reflective material, acoustic excitation acts as the dominant micromovement on the surface. Consequently, the acoustic signature emerged with high fidelity in the very first components of the separation, allowing the signal of interest to be isolated from a single optimal component. In contrast, the plant has a complex and flexible organic structure, subject to natural low-frequency micromovements and non-acoustic disturbances. In this scenario, the robustness of the CP algorithm was confirmed; the primary dominant (non-acoustic) motion was retained in the first components, while the pure audio signal was isolated and reconstructed exclusively from the fourth component, eliminating the need to sum it with the other residuals.

B. Computational Efficiency and Memory Trade-off

The evaluation of the computational execution profile was conducted using a 17-second video sequence with a resolution of 704 x 400 pixels. The proposed spatial pipeline was executed on a single 2 GHz processor, completing the entire extraction and digital conditioning sequence in 25 minutes. During execution, the system recorded a prominent increase in temporary RAM allocation, which peaked at approximately 28 GB.

This behavior characterizes the deliberate space-time trade-off introduced by transitioning to a direct spatial formulation. As outlined in the methodology, the high computational complexity associated with Singular Value Decomposition (SVD) and the generalized eigenvalue problem is restricted exclusively to the processing of the first batch of frames. By storing and vectorizing the learned projection and mixing matrices from this initial block, all subsequent temporal batches are projected directly into the established latent subspace. This architectural choice effectively bypasses the iterative frame-by-frame recalculations, preventing a continuous memory ramp and stabilizing hardware utilization across long video files.

For historical completeness and context within the literature, traditional phase-based architectures utilizing complex steerable pyramids are reported to require between 120 and 180 minutes (approximating 150 minutes) under a MATLAB environment, typically operating on specialized workstations equipped with dual 3.46 GHz processors.

However, the peak memory consumption of 28 GB for a short 17-second video indicates significant scalability limitations for the proposed approach. While the temporal chunking strategy mitigates continuous memory accumulation, processing longer video sequences or higher spatial resolutions under the current spatial formulation may still exceed the RAM capacity of standard hardware. Future optimizations will need to explore smaller batch sizes, out-of-core memory management, or more aggressive spatial downsampling to make the pipeline highly scalable.

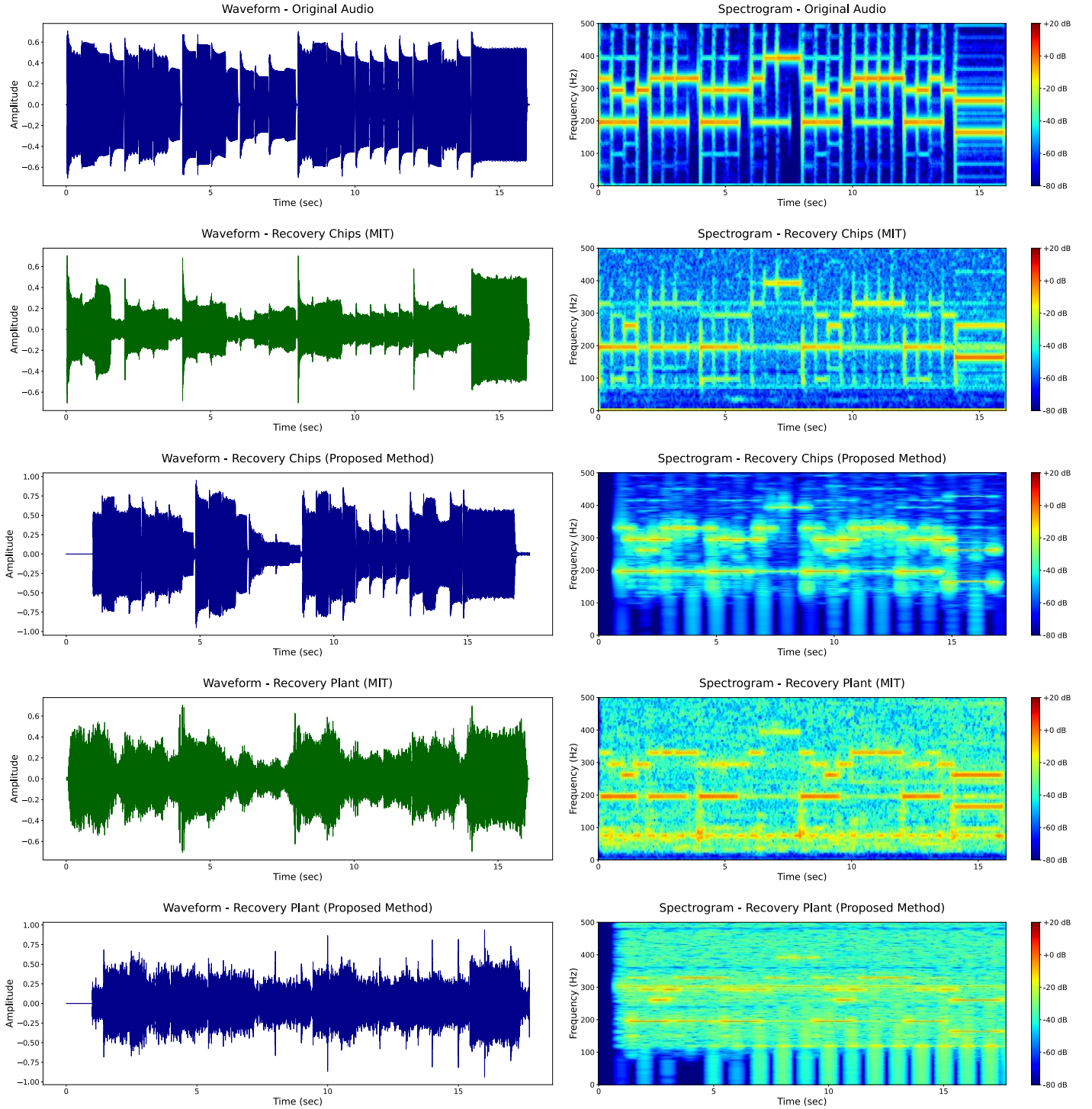


Fig 3. Qualitative comparison of acoustic reconstruction in the time and frequency domains. From top to bottom: original reference audio; chip packet recovered using the Complex Steerable Pyramid method (based on Davis et al.); chip packet recovered using PCA+CP; floor plan recovered using the base method; and floor plan recovered using PCA+CP. The spectrograms show that the proposed method preserves the harmonic structure with less background noise in the spectral envelope, outperforming the traditional phase transform followed by extraction via signal processing. Image adapted from [1].

C. Quantitative Evaluation of the Reconstruction

To quantitatively assess the intelligibility and fidelity of the recovered audio, three performance metrics were calculated: the Segment-Specific Signal-to-Noise Ratio (SSNR), Envelope Correlation, and the Root Mean Square Error of Pitch (RMSE in Hz). The comparative results are summarized in Table II.

A joint analysis of computational efficiency with Table II reveals the main advantage of the proposed approach. Although the baseline method shows advantages in SSNR and Envelope Correlation, the proposed algorithm achieved a significant reduction in Pitch Error (RMSE) for both materials, dropping from 109.39 Hz to 74.07 Hz (Chips) and from 157.16 Hz to 115.52 Hz (Plant).

These results indicate that combining cascaded zero-order filtering with PCA+CP extraction more accurately preserves the frequency integrity of the fundamental structures. However, the substantial drop in envelope correlation, falling below 0.3 for the chips packet, must be critically acknowledged. This reduction indicates that the recovered signal loses a significant portion of its original amplitude dynamics, potentially resulting in a noisier or less intelligible reconstruction compared to the baseline approach. Therefore, the proposed method establishes a specific trade-off: it sacrifices acoustic amplitude fidelity and overall signal-to-noise ratio in exchange for a drastic reduction in computational latency and superior harmonic pitch extraction.

TABLE I. PERFORMANCE COMPARISON IN AUDIO RECOVERY.

Method	SSNR (dB)	Pitch Error (Hz)	Correlation
Davis <i>et al.</i> (Chips)	1.50	109.39	0.5998
Proposed (Chips)	1.04	74.07	0.2796
Davis <i>et al.</i> (Plant)	1.57	157.16	0.3915
Proposed (Plant)	1.18	115.52	0.3465

D. Comparative Synthesis of Methodological Features and Trade-offs

To clarify the practical contributions, the proposed pipeline is directly contrasted against the phase-based baseline [1], as synthesized in Table II. Operating on raw pixel-intensities with temporal chunking eliminates iterative recalculations, yielding an 83% reduction in processing latency. Additionally, CP extraction reduces pitch RMSE by 31%, preserving harmonic structures with superior accuracy.

Conversely, these gains introduce specific trade-offs. The spatial formulation requires high memory allocation (28 GB peak RAM), and bypassing pyramidal transformations lowers envelope correlation. Thus, the architecture establishes a clear methodological compromise, explicitly trading acoustic amplitude fidelity for reduced latency and superior harmonic pitch recovery.

TABLE II. METHODOLOGICAL PARADIGM COMPARISON.

Ideal Sensor Characteristics	Phase-Based Baseline (Davis et al.)	Proposed Spatial Pipeline (PCA + CP)
Low Execution Latency	High latency: Requires 120–180 minutes of processing.	Low latency: Requires only 25 minutes (an 83% speedup).
High Frequency Fidelity	Moderate (Higher pitch error: 109.39 Hz / 157.16 Hz).	High fidelity: Lower pitch error (74.07 Hz / 115.52 Hz).
Low Memory Footprint	Low footprint: Stable baseline workstation utilization.	High footprint: Intensive RAM trade-off with a 28 GB spike.
No Complex Transformations	Uses transformations: Depends on multi-scale complex pyramids.	No complex transformations: Operates entirely in the pixel-intensity domain.
Targeted Signal Separation	Global approach: Performs global averaging of aligned local signals.	Targeted approach: Isolates a single optimal 1D acoustic component.

V. CONCLUSION

This work validated an architecture for the Visual Microphone operating directly on raw pixel-intensity signals via PCA and CP. The main conclusion is that replacing phase transformations with a block-segmented matrix model enables a highly efficient space-time trade-off. The method reduced processing latency by 83%, from 150 to 25 minutes, by avoiding the iterative recalculation of subspaces. In addition to speed, extraction via CP improved recovery of dominant harmonic frequency structures, reducing the pitch error (RMSE) by 31%. Thus, PCA in conjunction with BSS establishes itself as a robust and fast alternative for passive audio recovery. These computational and frequency extraction gains, however, come at the explicit cost of high peak memory consumption and a deterioration in the audio's temporal amplitude dynamics, highlighting areas for further refinement.

Future research will focus on overcoming the dependency on high-speed hardware by exploring the feasibility of the proposed pipeline with conventional 30 or 60 fps cameras. The primary strategy will involve investigating Compressive Sampling algorithms to reconstruct acoustic signals from under-sampled measurements, leveraging the inherent sparsity of harmonic signals in the frequency domain. Additionally, we intend to explore Active Frequency Transposition techniques, using low-frequency carriers to shift the original audio spectrum into the sensor's baseband to avoid destructive aliasing. This approach aims to restore the signal via digital demodulation while preserving the frequency integrity and the 83% reduction in computational latency already demonstrated by the current PCA and CP architecture.

REFERENCES

[1] A. Davis, M. Rubinstein, N. Wadhwa, G. J. Mysore, F. Durand e W. T. Freeman, "The Visual Microphone: Passive Recovery of Sound from Video", *ACM Transactions on Graphics (SIGGRAPH)*, v. 33, n. 4, pp. 79:1–79:10, July 2014.

[2] A. Davis, K. L. Bouman, J. G. Chen, M. Rubinstein, F. Durand e W. T. Freeman, "Visual Vibrometry: Estimating Material Properties from Small Motions in Video", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5335-5343, June 2015.

[3] N. Wadhwa, M. Rubinstein, F. Durand e W. T. Freeman, "Phase-Based Video Motion Processing", *ACM Transactions on Graphics*, v. 32, n. 4, pp. 80:1-80:10, July 2013.

[4] M. A. Elgharib, M. Hefeeda, F. Durand e W. T. Freeman, "Video Magnification in Presence of Large Motions", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4119-4127, June 2015.

[5] W. Zhang, C. Wang, S. Li, et al., "Visual microphone based on computational imaging", *Optics Express*, v. 33, n. 1, pp. 268-281, January 2025.

[6] M. F. Silva, A. Green, A. Cattaneo e D. Mascarenas, "Full-field blind identification of traveling waves from video measurements", *Mechanical Systems and Signal Processing*, v. 224, 111956, 2025.

[7] M. Silva, B. Martinez, E. Figueiredo, J. C. W. A. Costa, Y. Yang e D. Mascareñas, "Nonnegative matrix factorization-based blind source separation for full-field and high-resolution modal identification from video", *Journal of Sound and Vibration*, v. 477, 115326, 2020.

[8] A. Hyvärinen e E. Oja, "Independent component analysis: algorithms and applications", *Neural Networks*, v. 13, n. 4-5, pp. 411-430, June 2000.

[9] J. G. Chen, N. Wadhwa, Y.-J. Cha, F. Durand, W. T. Freeman e O. Buyukozturk, "Modal identification of simple structures with high-speed video using motion magnification", *Journal of Sound and Vibration*, v. 345, pp. 58-71, June 2015.