2025 28th International Conference on Computer and Information Technology (ICCIT)
19–21 December 2025, Cox's Bazar, Bangladesh

Vocal Tract Resonance Estimation: An Energy Weighted Singular Spectrum Approach

Mohaimenul Islam
Department of Electrical and Electronic Engineering, Khulna University of Engineering & Technology (KUET),
Khulna-9203, Bangladesh.
Email: mohaimenul.work.2001@gmail.com

Md Rakibul Hasan
School of Electrical Engineering, Computing and Mathematical Sciences Curtin University
Bentley, WA 6102, Australia
Email: rakibul.hasan@curtin.edu.au

Md Mahbub Hasan
Department of Electrical and Electronic Engineering, Khulna University of Engineering & Technology (KUET),
Khulna-9203, Bangladesh.
Email: mahbub01@eee.kuet.ac.bd

2025 28th International Conference on Computer and Information Technology (ICCIT) | 979-8-3315-7867-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICCIT68739.2025.11490505

***Abstract*—Formant extraction, therefore, the recovery of vocal-tract information is central to acoustics-based phonetics, and most vocal tract analysis algorithms operate on spectral envelopes or pole domains. However, the rich time-domain structure and its distribution of energy remain underused. We examine whether singular values, containing energy distribution information, derived from Hankel matrix, which contain temporal information, do in fact encode formant related information or not. In this work, ground-truth formants and bandwidths are obtained from four isolated bengali vowels (অ /ɔ/, আ /a/, ই /i/, and উ /u/) using polynomial decomposition LPC (PDLPC), one of the archetype methods, and formant candidates from the spectral energy domain are calculated from energy-weighted component-spectrum peaks ("weighted peaks," WP). We then fit linear maps in both directions to test mutual predictability and see whether there is a relation between. Upon application across 160 vowel tokens from 20 speakers, it is observed that Weighted Peaks show strong, monotonic relations with formants, suggesting that they do contain formant information. Theoretically, the findings should provide us with concrete evidence that a few dominant patterns extracted from a short slice of sound (Roughly Singular Pairs 1 - 10) are enough to predict the mouth's resonant tones (Formants F1 and F2 mainly). Practically, this pipeline offers a lightweight feature for estimating or stabilizing formants and enables singular value decomposition based formant tracking directly from raw waveforms. The procedure is reproducible, language-portable, and easy to integrate into analysis or learning systems.**

***Index Terms*—Formants, PDLPC, Hankel matrix, SVD, Bandwidth, Bengali Vowels, Weighted SVD Peaks**

I. Introduction

In everyday life, whether coordinating work, accessing services, or interacting with technology, spoken communication is indispensable. But the human vocal tract is prone to many complications that hamper the ability to speak [1]. As a result, researchers have toiled long to analyze the vocal tract to its depths, and some scientists have also been working on non-invasive ways to investigate issues relating to the vocal tract [1]. One of the key markers of any vocal tract are the formants. Formants are the broad resonant bands that arise when airflow from the vibrating vocal folds is shaped by the cavities of the vocal tract.The first two or three resonances (F1, F2, F3) are especially informative for distinguishing vowels [2]. Estimating these resonances from short recordings matters to many audiences. A widely used approach is classical linear predictive coding (LPC). In practice, LPC predicts each sample from a weighted sum of past samples; the resulting filter is *all–pole*, and the formant frequencies and bandwidths are read from the complex roots ("poles") of the prediction polynomial [3]. Although LPC is fast and interpretable, especially with Burg's variant [4], its estimated spectral envelope can shift with the chosen model order, windowing/tapering, noise level (SNR), and the placement of source harmonics. As a result, the tallest peaks in a short–time spectrum do not always coincide with the tract's true resonances [5]. Merger–aware PDLPC improves stability by explicitly handling clusters of nearby poles [6], yet it still reasons primarily in the envelope/pole domain.

Isolated vowel segments are rich in low–dimensional regularities that can be exposed by a time–delay (Hankel) embedding followed by singular value decomposition (SVD) which reveals the energy distribution. Furthermore, from their spectra it is expected that resonance evidence can be found. Although typical formant regions for vowels are well documented, the use of such short–time subspace structure as *direct* evidence for formants and bandwidths is comparatively underexplored. This motivates a focused question: *does temporal information procured from Hankel embedding of vowel waveforms, and its energy statistics derived from singular value components, encode sufficient information to predict vocal tract resonances across speakers?* In more field specific terms, the question sums up to: *Can SVD be used to estimate formants?* Our hypothesis is in the affirmative and we will present the evidences necessary to answer the question above. Although, this presumption is debatable as temporal subspace markers may or may not align with resonances in short windows, but it is very much testable and verifiable.

To address this question, first, we use a Bengali vowel dataset. The Bengali language has an abundance of phonetic features, especially the vowel portion [7], [8]. Bengali people are used to pronouncing lots of vowels with depth which gives a more detailed structure to vocal tract resonances. Second, a whole methodology is devised to test the question and

find out whether the claim is true. We desire to keep the pipeline design deliberately lightweight and for it to not require dynamic trackers. As a result, the findings might not only be used as a standalone estimator but also serve as a bandwise post–processor to correct pioneer LPC pipelines.

This line of inquiry matters to multiple audiences. Such as, engineers use them in automatic speech recognition and speaker normalization, and clinicians track atypical speech or therapy outcomes [9] [10] [11]. Also, WPs offers an inexpensive stabilizer for formant estimates and a plug–in correction layer compatible with existing tools (e.g., Praat) and temporal trackers [12], [13].

To summarize the main contribution of this work, we will demonstrate that the majority of the useful information for the first two formants is carried by a small number of singular value pairs, whereas the singular component structure of the other formants is more dispersed. This allows for the usage of these pairs to estimate formants more accurately and more quickly.

II. Method

We seek a compact, SVD–based summary of spectral evidence that is maximally aligned with the *acoustic* resonances (formants) obtained from a strong LPC baseline. The pipeline has four stages: (i) precise segmenting and conditioning of the vowel nucleus; (ii) PDLPC for ground-truth (F_i, B_i); (iii) time–delay (Hankel) embedding followed by SVD; and (iv) bandwise aggregation into *weighted peaks* (WPs) and (v) linear maps between the weighted peaks and formants. So, the overall work flow is 1. $SoundSegment \rightarrow PDLPC \rightarrow Formants$, 2. $SoundSegment \rightarrow SSA \rightarrow WeightedPeaks$, 3. Linear Mapping

A. Segment, Preprocessing, and PDLPC

Let $x[n]$ be a single vowel token sampled at $f_s = 44.1\,\text{kHz}$. We center a steady-state nucleus of $T = 60\,\text{ms}$ (so $N = \lfloor T f_s \rfloor$ samples), normalize its peak to unity, and apply a Hamming taper $w[n]$:

$$y[n] = w[n]\,\frac{x[n] - \mu_x}{\max_m |x[m] - \mu_x|}, \quad n = 0, \ldots, N-1, \quad (1)$$

with μ_x the mean over the nucleus. To temper spectral tilt we pre-emphasize with transfer

$$H_{\text{pre}}(z) = 1 - \alpha z^{-1}, \quad \alpha = 0.97, \quad (2)$$

and feed $y_{\text{pre}} = \mathcal{Z}^{-1}\{H_{\text{pre}}(z)Y(z)\}$ into PDLPC.

For an LPC order $p = 16$, the predictor polynomial is

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}. \quad (3)$$

PDLPC explicitly handles *pole clusters* by decomposing $A(z)$ into lower-degree polynomials before root-finding, improving stability around mergers. For each complex-conjugate pole pair $z_i, \bar{z}_i$ with modulus r_i and angle θ_i, the corresponding formant frequency and bandwidth are

$$F_i = \frac{f_s}{2\pi}\,\theta_i, \qquad B_i = -\frac{f_s}{\pi}\ln r_i, \quad (4)$$

Where the formant frequency and bandwidth are read from the complex roots of the prediction polynomial, following standard LPC practices [6] [4] [3].

To summarize, the main target of this phase was to isolate a steady segment that has consistent repeating patterns so that upon applying LPC correct formants and bandwidths could be extracted and accurate SVD components can be calculated. The accuracy of the chosen segment can be viewed in Fig. 1 where the LPC spectrum peaks matches with the formant locations, which is near to ideal.

B. Hankel SVD and Component Spectra

This phase is concerned with making a Hankel matrix from time series data and deriving the SVD components. This whole process is also called Singular Spectrum Analysis [14] [15]. We form a near-square Hankel matrix of lag $L{+}1$ and shift $K{+}1$ with $L{+}K{+}1 = N$:

$$H = \begin{bmatrix} y[0] & y[1] & \cdots & y[K] \\ y[1] & y[2] & \cdots & y[K{+}1] \\ \vdots & \vdots & \ddots & \vdots \\ y[L] & y[L{+}1] & \cdots & y[N{-}1] \end{bmatrix}, \quad L = \lfloor \rho N \rfloor,\ \rho \approx 0.5. \quad (5)$$

After column-mean removal, we compute the full SVD

$$H = U\,\Sigma\,V^{\top} = \sum_{k=1}^{\min(L+1,K+1)} \sigma_k\,\mathbf{u}_k\,\mathbf{v}_k^{\top}. \quad (6)$$

The construction follows the usual formulation of singular spectrum analysis for time series data [14] [15] The right singular vectors $\mathbf{v}_k \in \mathbb{R}^{K+1}$ serve as time-domain *components*. For each k, we take a zero-padded DFT to obtain magnitude spectra $S_k(f)$. A sharp scree decay (rapid drop of σ_k) indicates that the leading components capture coherent, quasi-periodic structure due to voicing (cf. Fig. 1).

Sinusoidal view.: If $y[n] \approx \sum_{q=1}^{Q} a_q e^{(j\omega_q - \lambda_q)n} + \eta[n]$ (damped complex sinusoids + noise), then H is approximately low-rank and the dominant $\mathbf{v}_k$ span the same subspace as $\{e^{j\omega_q n}\}$. Thus peaks in $|S_k(f)|$ tend to concentrate near resonances.

C. Bandwise Weighted Peaks (WPs)

This step is concerned with finding the formant candidates from the energy domain. The idea is to find the FFT of each SVD component within a range and then checked which of those component's FFT peak appears within the bandwidth of each formants. Those components are selected and their weighted average is calculated based on their energy or component index. These weighted peaks or WPs are the formant candidates. Let $B_i = [F_i^-, F_i^+]$ denote a frequency gate

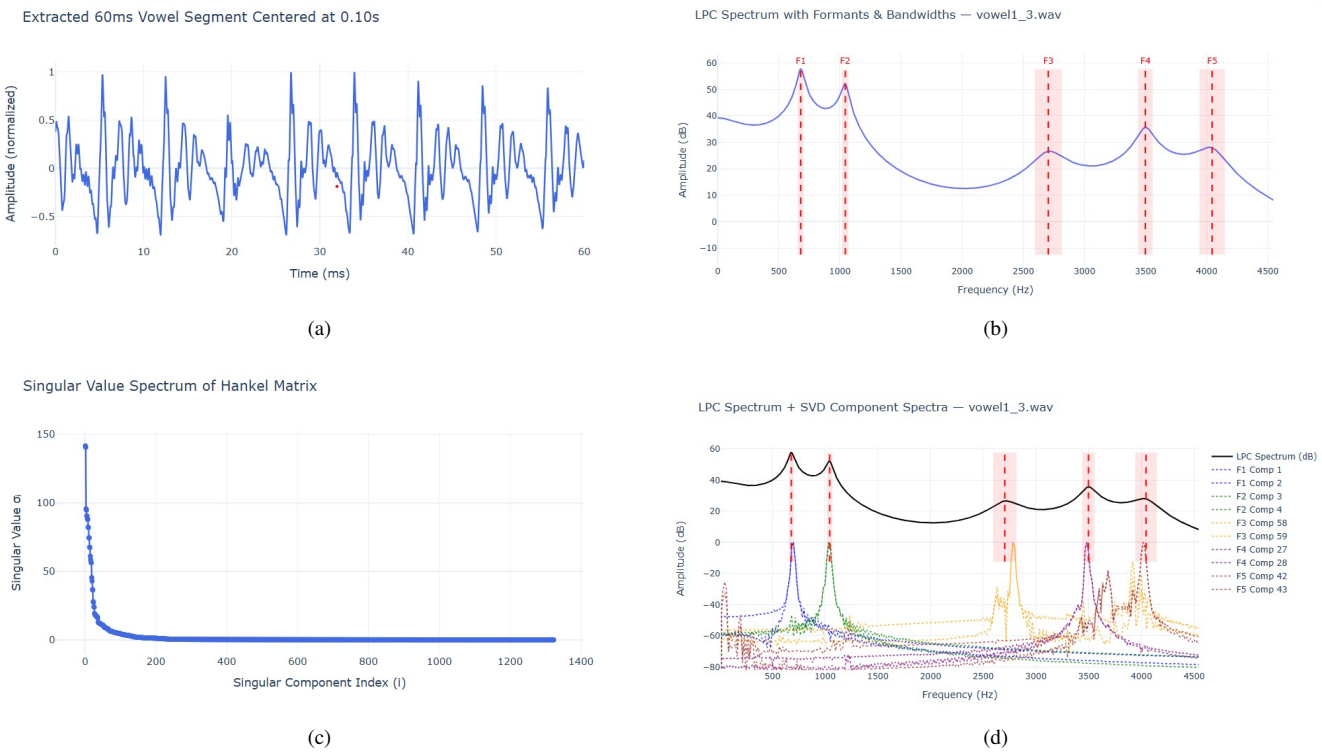


Fig. 1. Method snapshots: (a) 60 ms isolated vowel nucleus containing repeating patterns, (b) LPC spectrum showing derived formant candidates from LPC, (c) Singular Component energy distribution derived from SVD, (d) Singular component spectra superimposed on LPC spectra for visual affirmation of relation between formant and SVD components.

around the ith PDLPC band (e.g., half–bandwidth from (4)). The peak set in band i is

$$\mathcal{P}_i = \Big\{(k, r, f_{k,r}, a_{k,r}) : f_{k,r} \in B_i, \quad a_{k,r} \text{ is a local maximum of } |S_k|\Big\}. \tag{7}$$

We select components either by *energy* (Method A: top $\lfloor K/2 \rfloor$ by σ_k) or by *cardinality* (Method B: first two peaks in B_i). To fuse evidence, we use weights

$$w_{k,r} = \sigma_k^{\beta}\, a_{k,r}^{\gamma}\, \exp\!\left(-\frac{(f_{k,r} - F_i)^2}{2\tau_i^2}\right), \tag{8}$$
$$\beta \in [0,1],\ \gamma \in [1,2],\ \tau_i = \kappa\, \tfrac{B_i}{2},\ \kappa \in [0.8, 1.2].$$

The *weighted peak* (band centroid) is

$$\mathrm{WP}_i = \frac{\sum_{(k,r)\in\mathcal{P}_i} w_{k,r}\, f_{k,r}}{\sum_{(k,r)\in\mathcal{P}_i} w_{k,r}}, \tag{9}$$
$$\mathbf{p} = \big[\mathrm{WP}_1, \ldots, \mathrm{WP}_5\big]^{\mathsf{T}}.$$

Choosing $\gamma \approx 1.5$ sharpens centroids without over-weighting narrow spikes; Method A improves mid/high-band stability by aggregating more than two components.

D. Linear Maps, Calibration, and Uncertainty

We fit ordinary least squares (OLS) maps in both directions.
Inverse (WPs → formants):

$$\widehat{\mathbf{f}} = \mathbf{W}_F\, \mathbf{p} + \mathbf{c}_F, \tag{10}$$
$$\mathbf{f} = \big[F_1, \ldots, F_5\big]^{\mathsf{T}}.$$

Forward $\big((F,B) \to WPs\big)$:

$$\widehat{\mathbf{p}} = \mathbf{W}_P\, [\mathbf{f}; \mathbf{b}] + \mathbf{c}_P, \tag{11}$$
$$\mathbf{b} = \big[B_1, \ldots, B_5\big]^{\mathsf{T}}.$$

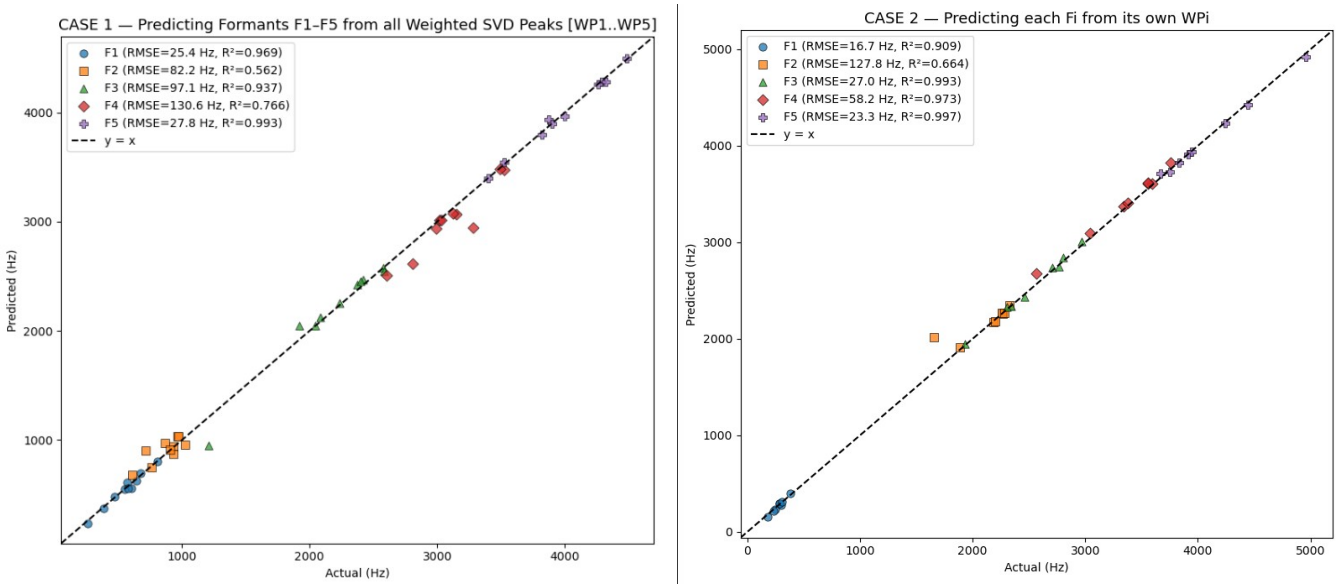
Error decomposition and covariance: We define error as

$$e_i^{(n)} \triangleq \hat{F}_i^{(n)} - F_i^{(n)}, \qquad n = 1, \ldots, N.$$

Then,

$$\mathrm{RMSE}_i^2 = \bar{e}_i^2 + \frac{1}{N-1}\sum_{n=1}^{N}\big(e_i^{(n)} - \bar{e}_i\big)^2, \tag{12}$$
$$\mathrm{cov}(\hat{\mathbf{f}}) \approx \mathbf{W}_F\, \mathrm{cov}(\mathbf{p})\, \mathbf{W}_F^{\mathsf{T}}. \tag{13}$$

This linear mapping is done for both method A and B. This is the deciding phase which tells whether there is a relation or not.



(a) (b)

Fig. 2. Output of linear regression: Predicted vs. actual. (a) Case 1: Mapping of weighted peaks to formants for all 5 formants together for one vowel. (b) Case 2: Mapping of weighted peaks to formants individually for each formant for one vowel.

E. Experimental Setup

All experiments use four isolated Bangla vowels অ /ɔ/, আ /a/, ই /i/, and উ /u/ from the "Isolated Bengali vowel and word speech sounds" corpus [16]. The dataset contains seven isolated vowels and seven isolated words, each produced 40 times by 20 native Bangla speakers (male and female, 20–26 years). In this work we retain the four vowels above, 40 tokens per class (total $N = 160$), and use a speaker-stratified 75%/25% train/test split with no speaker overlap.

The above mentioned methodology is applied in 2 ways. In method A the range of SVD components to search from was selected to be the first 50 percent of the total SVD components. This was based on the fact that theoretically audio can be reconstructed accurately using truncated SVD components and also remove noise [17], [18]. Method B takes the first SVD component pair that appears under each formant bandwidth. This is based on the theory that formant information is ideally contained in the first SVD pair or eigen pair that appears within their respective formant bandwidth [19]. Lastly, the approach assumes a clean, steady-state nucleus. This means coarticulation and nasalization can reduce agreement between LPC and SVD cues. The band priors are adult-oriented and may require retuning for children's speech and also we do not track through time.

III. Results

Fig. 1 summarizes the analysis stages on a representative token. The LPC spectrum provides interpretable candidates but exhibits the usual sensitivity to spectral fine structure. The spectrum plot in Fig. 1(b) shows that formants orient quite perfectly with the LPC spectrum peaks, which aligns with the classical dogma. The SVD magnitude plot in Fig. 1(c) reveals rapid energy concentration in the first few components, and their spectra tend to align near LPC-indicated resonances as shown in Fig. 1(d) .

A. Inverse: WPs → Formants (F,B)

Across vowels, it is observed that method A gives the best R^2 value and also good RMSE which can be observed in Table I. However, this method is very time consuming and inefficient due to the large amount of SVD components to surf through. Nevertheless, this method answers our main question, whether there is a relation between SVD components and formants, and the answer is yes. Method B tries to simplify this process by taking only the first pair of SVD component peaks that appear in the bandwidths. Theoretically, this should have given us the component pairs which correlate to the formants.In reality, the expected components are not the ones contributing to the formant making the RMSE values and R^2 values bad for certain vowels as seen in Table II.

B. Forward: (F,B) → WPs

The forward direction checks internal consistency: given ground-truth (F_i, B_i), the model recovers weighted peaks with small residuals in the outer bands and moderate residuals around F2/F4. This supports the view that the bandwise aggregation is well aligned with resonance structure, with bandwidths improving predictions at higher formants.

C. Deep Analysis of F1 and F2, Method B in depth

Because *F1* and *F2* are the principal cues for vowel quality and vowel-space structure, we performed a targeted analysis of their *SVD structure* [2]. We tried to find which component *pairs* carry the information, how deep the SVD search must

TABLE I

METHOD A STATISTICS ARE SHOWN HERE. FOR EACH VOWEL RMSE(Hz) / R^2 ARE SHOWN WITH RESPECT TO THE FORMANTS AND WEIGHTED PEAKS BEING ANALYZED.

Band →	/ɔ/	/a/	/i/	/u/
F1	27.74 / 0.9636	38.77 / 0.9602	5.48 / 0.9529	19.47 / 0.8539
F2	29.16 / 0.9449	36.83 / 0.9571	51.62 / 0.9955	43.10 / 0.8697
F3	108.19 / 0.9223	102.31 / 0.9425	78.87 / 0.9426	53.67 / 0.9557
F4	153.70 / 0.6753	134.73 / 0.8955	52.96 / 0.9791	139.37 / 0.8840
F5	33.69 / 0.9900	130.36 / 0.7815	76.34 / 0.9501	74.86 / 0.9681

(a) Inverse mapping (WP→F)

Band →	/ɔ/	/a/	/i/	/u/
WP1	43.52 / 0.9203	33.59 / 0.9651	8.04 / 0.9360	34.95 / 0.7026
WP2	46.93 / 0.8734	30.14 / 0.9681	63.12 / 0.9936	39.19 / 0.8891
WP3	123.54 / 0.9424	113.80 / 0.9369	67.58 / 0.9553	47.84 / 0.9736
WP4	107.46 / 0.8879	83.88 / 0.9695	124.11 / 0.8884	150.55 / 0.9134
WP5	32.43 / 0.9902	93.49 / 0.8773	37.08 / 0.9906	55.36 / 0.9815

(b) Forward mapping ((F,B)→WP)

TABLE II

METHOD B STATISTICS ARE SHOWN HERE. FOR EACH VOWEL RMSE(Hz) / R^2 ARE SHOWN WITH RESPECT TO THE FORMANTS AND WEIGHTED PEAKS BEING ANALYZED.

Band →	/ɔ/	/a/	/i/	/u/
F1	23.96 / 0.9338	40.97 / 0.9556	5.51 / 0.9523	31.50 / 0.6178
F2	98.05 / -0.3267	77.47 / 0.8102	150.80 / 0.9618	77.89 / 0.5744
F3	131.61 / 0.9159	173.99 / 0.8338	112.39 / 0.8835	194.31 / 0.4195
F4	182.58 / 0.7212	247.91 / 0.6461	145.88 / 0.8416	428.54 / -0.0965
F5	228.73 / 0.4682	183.14 / 0.5688	151.71 / 0.8030	193.07 / 0.7881

(a) Inverse mapping (WP→F)

Band →	/ɔ/	/a/	/i/	/u/
WP1	22.13 / 0.9531	48.03 / 0.9505	7.81 / 0.9401	52.32 / 0.5430
WP2	174.15 / -0.1923	78.79 / 0.7647	208.26 / 0.9391	88.52 / 0.5118
WP3	206.74 / 0.8856	244.95 / 0.7254	114.49 / 0.8803	206.40 / 0.7392
WP4	117.32 / 0.8967	169.79 / 0.9092	88.84 / 0.9498	327.52 / 0.7978
WP5	159.31 / 0.9156	199.34 / 0.5407	75.77 / 0.9688	112.34 / 0.9241

(b) Forward mapping ((F,B)→WP)

TABLE III

F1 (VOWEL 1) VS. SVD DEPTH (COMPONENTS $\leq$ CAP).

Cap (comps)	RMSE [Hz]	R^2	Coverage
6	53.56	0.714	32/39
10	18.74	0.951	34/39
20	**15.41**	**0.973**	36/39
50	29.82	0.917	37/39

TABLE IV

F1 ACROSS VOWELS (RANGE=20): INVERSE FIT ACCURACY.

Vowel	RMSE [Hz]	R^2	Coverage
/ɔ/	15.41	0.973	36/39
/a/	25.61	0.821	32/39
/i/	4.11	0.966	19/39
/u/	20.68	0.857	28/38

TABLE V

F2 (VOWEL 1) VS. SVD DEPTH (COMPONENTS $\leq$ CAP).

Cap (comps)	RMSE [Hz]	R^2	Coverage
6	101.27	0.405	23/39
10	78.77	0.458	29/39
20	**40.33**	**0.927**	36/39
30	43.31	0.851	37/39

TABLE VI

F2 ACROSS VOWELS (RANGE=20): INVERSE FIT ACCURACY.

Vowel	RMSE [Hz]	R^2	Coverage
/ɔ/	40.33	0.927	36/39
/a/	63.20	0.899	38/39
/i/	101.71	<0	19/39
/u/	47.68	0.821	34/38

go, and how accuracy trades off against coverage. All results below use the same single 60 ms nucleus and the protocol of Sec. II.

F1 (vowel 1, single frame).: Evidence for F1 concentrates in the first SVD pair (1–2) and stabilizes when we include mid pairs up to (9–10). Table III shows accuracy as we increase an *even* component cap (complete pairs only).

At range 20 (pairs 1–10), the inverse calibration is nearly identity, $\widehat{F}_1 \approx 0.968\,\mathrm{WP}_1 + 9.64$, indicating minimal bias. Longer ranges slightly raise coverage but start to include weak, high-rank pairs that dilute the weighted peak.

F1 across vowels (range=20 components).: Using the same range across vowels yields strong or good accuracy (Table IV). Empirically, pair (1–2) remains dominant while pairs (3–10) provide robustness and speaker-specific fine-tuning.

Takeaway for F1.: A small, fixed SVD depth with complete pairs **1–10** (cap=20) is sufficient as pair (1–2) carries most of the F1 information and pairs (3–10) stabilize the estimate. Higher pairs add little and can hurt unless strongly down-weighted.

F2 (vowel 1).: F2 information is *distributed* across mid pairs (roughly 3–10), not concentrated in pair (1–2) or (3–4). Increasing the component range from 6 to 20 improves both accuracy and coverage substantially; pushing to 30 begins to erode R^2 (Table V).

At range 20 the inverse line is $\widehat{F}_2 \approx 0.841\,\mathrm{WP}_2 + 171$ Hz, reflecting a modest scale/offset bias that is easily corrected.

F2 across vowels (range=20) components.: It can be seen from (Table VI) that vowels $1, 2, 4$ show strong to good accuracy; vowel 3 has low coverage and poor R^2 without light gating, consistent with weaker/dispersed mid-band evidence.

Takeaway for F2.: It is best to search complete pairs **1–10** (range=20) and *prioritize pairs 3–10*. This produces

accurate, speaker-robust estimates with a tiny linear calibration per vowel.

D. Discussion

It can be suggested that SVD components 1 and 2 can accurately model formant 1 and components 3-10 can model formant 2. To encode the finer structures present in the vocal tracts more components can be added. But going beyond component 20 seems to reduce the accuracy of the algorithm. This implies the addition of unnecessary noise. For formants 3 to 5, it is much more difficult to find the accurate pairs that represent them because lots of SVD components come in to play as we reach those high formants. The best strategy is method A for them.

For context, we briefly compare our energy-weighted SVD front end with the quasi-closed-phase forward–backward (QCP–FB) plus deep neural network (DNN) tracker of Gowda et al. [20]. Both approaches combine a model-based spectral front end with a mapping from short-time frames to formants, but Gowda et al. track continuous sentences from the VTR-TIMIT database at 8 kHz using a DNN, whereas we operate on clean, isolated Bangla vowel nuclei at 44.1 kHz using a single linear regression from SVD-derived weighted peaks. In this controlled vowel setting our low-band errors (F1 RMSE typically ≈ 5–40 Hz and F2 ≈ 30–65 Hz with high R^2) are numerically smaller than the reported F1/F2 errors of the DNN–QCP–FB tracker (about 57 Hz and 105 Hz), indicating that a lightweight SVD-based front end can match or exceed the accuracy of more complex continuous-tracking pipelines.

IV. Conclusion

This article reveals a lightweight and physically significant approach to estimating formants. Our analysis suggests that singular components do have formant information within them. A focused analysis of formants **F1** and **F2** was done and it was revealed that a small fixed set of *complete SVD pairs* carries most of the usable information. However, adding too many singular components to the aggregated sum for these two formants seems to hinder the accuracy. We found that only observing the first 10-20 singular components are enough to track and estimate F1 and F2. For other formants, specifically for F3 to F5, aggregating components seems to increase accuracy, suggesting a more distributed energy structure across the whole range of singular components.

The current study is purposefully limited to isolated vowel nuclei from adult speakers analyzed at a fixed window position of 60 ms. Consonant-vowel transitions, continuous speech, children's voices, noisy recording conditions, etc are not taken into account. In subsequent work, we intend to: (i) expand the analysis to sliding-window tracking so that SVD features can follow time-varying formant trajectories; (ii) examine how the weighted peaks behave under additive noise and reverberation; and (iii) use publicly available corpora with hand-corrected reference formants to compare the suggested representation with neural formant trackers and other contemporary estimators.

To conclude, SVD components do in fact contain formant information and can be a fast alternative to calculate formants and also the finer vocal tract structures.

References

[1] G. Rodrigues, F. Zambon, L. Mathieson, and M. Behlau, "Vocal tract discomfort in teachers: Its relationship to self-reported voice disorders," *Journal of Voice*, vol. 27, no. 4, pp. 473–480, 2013.

[2] R. Swanepoel, D. J. J. Oosthuizen, and J. J. Hanekom, "The relative importance of spectral cues for vowel recognition in severe noise," *The Journal of the Acoustical Society of America*, vol. 132, no. 4, pp. 2652–2662, 2012.

[3] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *The Journal of the Acoustical Society of America*, vol. 50, no. 2B, pp. 637–655, 1971.

[4] J. D. Markel and J. Gray, A. H., "The burg algorithm for LPC speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 5, pp. 584–587, 1980.

[5] L. Welling and H. Ney, "Formant estimation for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 36–48, 1998.

[6] M. Fu, X. Wang, and J. Wang, "Polynomial-decomposition-based LPC for formant estimation," *IEEE Signal Processing Letters*, vol. 29, pp. 1392–1396, 2022.

[7] A. K. Datta, *Acoustics of Bangla Speech Sounds*. Singapore: Springer, 2018.

[8] M. A. K. Pramanik and K. Kido, "Formant frequencies of bengali vowels uttered in isolation," *The Journal of the Acoustical Society of Japan*, vol. 32, no. 8, pp. 488–489, 1976.

[9] W. R. Rodríguez and E. Lleida, "Formant estimation in children's speech and its application for a spanish speech therapy tool," in *Proc. Workshop on Speech and Language Technology in Education (SLaTE)*, 2009, pp. 81–84.

[10] A. Ali, S. Bhatti, and M. S. Mian, "Formants based analysis for speech recognition," in *2006 IEEE International Conference on Engineering of Intelligent Systems*. IEEE, 2006, pp. 1–3.

[11] M. Lincoln, S. Cox, and S. Ringland, "A fast method of speaker normalisation using formant estimation," in *Proc. 5th European Conference on Speech Communication and Technology (EUROSPEECH)*, Rhodes, Greece, 1997, pp. 2095–2098.

[12] K. Mustafa and I. C. Bruce, "Robust formant tracking for continuous speech with speaker variability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 435–446, 2006.

[13] L. Deng, L. J. Lee, H. Attias, and A. Acero, "Adaptive kalman filtering and smoothing for tracking vocal tract resonances using a continuous-valued hidden dynamic model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 13–23, 2007.

[14] N. Golyandina, "Particularities and commonalities of singular spectrum analysis as a method of time series analysis and signal processing," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 12, no. 4, p. e1487, 2020.

[15] S. Wadekar, A. Mahalkari, A. Ali, and A. Gupta, "A review on singular spectrum analysis," in *2022 IEEE International Conference on Current Development in Engineering and Technology (CCET)*. IEEE, 2022, pp. 1–6.

[16] M. R. Hasan and M. M. Hasan, "Isolated bengali vowel and word speech sounds," Mendeley Data, doi: 10.17632/2h6975kdsx.1, 2021.

[17] T. Albiges, Z. Sabeur, and B. Arbab-Zavar, "Complex audio signal data compression and reconstruction: A benchmark data pre-processing approach for machine classification of chronic respiratory diseases," *Digital Health*, vol. 10, p. 20552076241302234, 2024.

[18] D. M. Lyra-Leite, J. P. C. L. da Costa, and J. L. A. de Carvalho, "Improved MRI reconstruction and denoising using SVD-based low-rank approximation," in *2012 Workshop on Engineering Applications*. IEEE, 2012, pp. 1–6.

[19] A. Islam, M. R. Hasan, M. Z. Hossain, and M. M. Hasan, "The eigenvalue distribution of Hankel matrix: A tool for spectral estimation from noisy data," in *2021 24th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2021, pp. 1–6.

[20] D. N. Gowda, B. Bollepalli, S. R. Kadiri, and P. Alku, "Formant tracking using quasi-closed phase forward-backward linear prediction analysis and deep neural networks," *IEEE Access*, vol. 9, pp. 151 631–151 640, 2021.