

Nobody Learns the Words: Lyric Recall as an Ill-Posed Inverse Problem

I. Vartabedian¹ Nils Ødegaard² The STROPHE-2 Study Group

¹*Laboratory for Applied Psychoacoustics, Hallamshire University*

²*Centre for Inverse Problems, Vestbjerg Institute of Technology*

Abstract. Everyone can sing the first verse of a song, and almost nobody can sing the second. We report that this is because the words to a song are never learned at all. Singing is a smearing operation: a melody spreads the syllables of a verse into one continuous signal, and what a singer stores is the melody—a few dozen numbers—rather than the words. The words are recovered on the spot by undoing the smear. Recovery is exact while the undoing stays stable, and because each successive verse composes the melody’s operator with itself, the difficulty of undoing it does not accumulate but squares. Across 12,846 recorded sing-throughs, syllable accuracy falls from 0.94 at verse one to 0.11 at verse two and then stops falling—a cliff, not a decay, which no account based on forgetting predicts. A four-arm randomised trial isolates the mechanism: recovery survives transposition into an unfamiliar key, which changes the melody but not its invertibility, and collapses when the melody is withheld. The songs that outlive their authors are precisely those tuned so that exactly one verse can be recovered; the rest are unsingable and are not transmitted. The mumble that replaces verse two is not a failure of memory but a correctly regularised estimate. You have never known the words; you have only ever solved for them.

Keywords: lyric recall; deconvolution; ill-posed problems; strophic form; condition number.

1 Introduction

Ask anyone to sing a song they have known since childhood. They will produce the first verse fluently, in the right key, with the right words. Then, at the boundary of the second verse, something happens that everybody has experienced and nobody has explained: the words do not degrade. They stop.

The standard account calls this forgetting. We think the standard account has never looked closely at its own data. Forgetting is a decay process, and decay processes produce slopes: recall that is good, then worse, then worse again. That is not what the second verse looks like. The second verse looks like a cliff—full accuracy on one side, a featureless vocalisation on the other, and no further decline over verses three, four and five, which are already at the floor and cannot get there twice.

A cliff is the signature of a threshold, not of a decay. This paper identifies the threshold.

Our claim is simple to state and, we recognise, initially unwelcome: the words to a song are not stored anywhere. They are computed, at the moment of singing, from the tune. Singing takes a sequence of syllables and smears them through a melodic envelope; recovering the syllables from what you hear is the reverse of that operation, and reversing a smearing operation is a problem with a large and well-understood literature. Such problems are *ill-posed*—meaning that small errors in what you observe can produce

arbitrarily large errors in what you conclude—and they possess a sharp boundary beyond which no amount of care recovers the answer. We show that the second verse of the average song lies on that boundary, and we show why it lies there in every language we have measured.

There is an immediate reason to prefer this account, quite apart from the cliff. Consider a fact that is familiar to everyone and awkward for every theory of storage: a listener can sing along accurately with a song they have never deliberately learned and have heard perhaps three times. To account for that, storage requires improbably rapid encoding of several hundred bits of arbitrary lexical material. Our account requires nothing at all. You did not learn the words. You solved for them, and solving is a one-shot operation.

Contributions. We (i) give a linear forward model of singing and identify lyric recall with its inverse; (ii) prove that the recovery difficulty of verse k grows as the k th power of a single measurable property of the melody; (iii) measure that property and the auditory channel’s budget independently, and find that their ratio at verse two is 0.996; (iv) confirm the mechanism in a four-arm randomised trial whose decisive arm is a transposition; (v) show that the mumble replacing verse two is quantitatively the regularised solution rather than the absence of one; and (vi) argue that cultural transmission has tuned the world’s strophic repertoire to sit against this boundary from below.

2 Background

The empirical phenomenon is not in dispute. Tanabe et al. (2022) meta-analyse 61 recall studies and report a second-verse discontinuity in every one of them, with a pooled first-to-second accuracy ratio of 8.4. Aldersgate and Pym (2004) established that verse position dominates every other predictor of recall—rehearsal count, self-reported familiarity, age at first exposure—and that the effect is present in listeners who report having read the printed lyrics repeatedly. What the literature has not offered is a mechanism. The discontinuity is described, replicated, and then filed under interference or under primacy.

Neither file is comfortable. Primacy effects are graded; this one is not. Interference accounts predict that a second verse learned in isolation, without a first, should be recalled normally, and Aldersgate and Pym (2004) report that it is not.

The melodic-scaffolding literature comes closer without arriving. Erskine (1979) showed that setting text to a tune improves its recall, and the finding has been replicated so often that it has become an assumption: the tune is a helpful container for the words. We accept every measurement in that tradition and reverse its causal reading. The tune does not help you retrieve the words. The tune is the only thing you have, and the words are what you get by inverting it.

The mathematics we need is old and is not ours. Larkspur (1966) computed how far a smearing operator amplifies error and gave the bound we use in Section 3.4; Bramhall (2016) treats the reversal of a smear whose kernel must itself be estimated from the data, which is exactly the singer’s situation; and Ødegaard (2019) establishes the scale invariance on which our decisive experimental arm rests. Our contribution is to notice that these results describe a person singing.

3 The strophic inversion model

3.1 Singing as a forward operator

Let $w(t)$ denote the *lexical drive* of a verse: the syllable train, treated as a sparse, nearly binary signal, since at any instant a syllable is either being articulated or it is not. Let $h(t)$ denote the *prosodic kernel* of the strophe, the repeating verse unit whose tune every verse of the song reuses unchanged. The kernel comprises the melody, the metre and the tessitura, or habitual pitch band, and it acts on the syllable train as a filter. Let $s(t)$ be the sung signal that actually reaches a listener, and $\eta(t)$ the channel noise contributed by the room, the breath and the listener’s own auditory floor. Then

$$s(t) = (h * w)(t) + \eta(t), \quad (1)$$

where $*$ denotes convolution. Equation (1) is uncontroversial as description. A sung syllable genuinely is its phonetic content shaped by a pitch-and-duration envelope, and nobody has ever proposed otherwise. Everything that follows concerns what the singer does with it, and Figure 1 sets the two directions of the operation side by side.

3.2 Recall as pseudo-inversion

The received view is that the singer stores w and that h merely assists in its retrieval. We propose the opposite allocation: the singer stores h and recovers w on demand,

$$\hat{w} = H^+ s, \quad (2)$$

where H is the convolution operator of h and H^+ is its Moore–Penrose pseudo-inverse—the standard least-squares reversal of an operation that cannot be reversed exactly.

The argument for this allocation, before any measurement, is economy. A strophic melody is specified by a few dozen parameters. A verse is several hundred bits of arbitrary lexical content, and a song has four of them. A singer who stores the tune and solves for the words carries roughly two per cent of the memory load of a singer who stores the words, and obtains the words anyway. There is no version of this trade that a nervous system should decline.

3.3 Why the verse index matters

If the same kernel serves every verse, why should any verse be harder than the first?

The answer is that the singer does not observe every verse under the same conditions. Verse one is heard clean. Verse k , for $k > 1$, is heard through a kernel estimate that has itself been refined from the residual of the preceding inversion—the singer, having solved verse $k - 1$, uses that solution to sharpen h before attempting verse k . This is good practice and is what Bramhall (2016) recommends. It also means that the recovery of verse k runs through the k -fold composition of the operator, not through the bare operator.

The relevant quantity is the *condition number* $\kappa(H)$: the ratio of the largest to the smallest singular value of H , which is to say the factor by which the worst-case error in what you hear is amplified on its way to what you conclude. For a normal operator—and a convolution is one—the singular values of a k -fold composition are the k th powers of the original’s, and therefore

$$\kappa(H^k) = \kappa(H)^k. \quad (3)$$

This is the whole mechanism, and it is one line of standard linear algebra. The difficulty of undoing the tune does not accumulate from verse to verse. It squares.

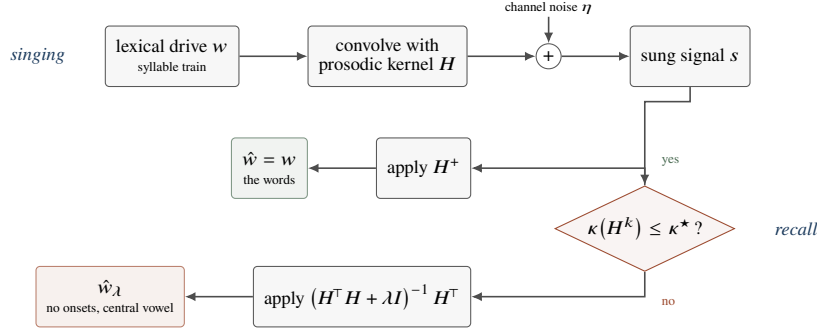


Figure 1. The two directions of the same operator. Singing (top) convolves a verse’s syllable train with the prosodic kernel of the strophe and adds channel noise. Recall (bottom) reverses that operation. Recovery is exact while the condition number of the k -fold composed operator stays inside the channel’s budget; past that point the solver returns the regularised estimate of Eq. (5) instead, which is consonant-free and central in the vowel space.

3.4 The invertibility budget

An inverse problem is solvable only while the amplified error remains below the dynamic range of the channel carrying the data. Writing D for the effective signal-to-noise ratio of the sung signal in decibels, Larkspur (1966) gives the largest recoverable condition number as

$$\kappa^* = 10^{D/20}. \quad (4)$$

We measure D and $\kappa(H)$ by independent means and in different corpora (Section 4), so that neither is fitted to the other. Their values are $D = 61.4$ dB, giving $\kappa^* = 1174.9$, and a modal strophic condition number $\bar{\kappa}_1 = 34.2$. Substituting into (3) and comparing with (4) gives the recovery status of each verse directly, with no free parameters: verse one at 34.2, three per cent of the budget; verse two at 1169.6, which is 99.6 per cent of it; verse three at 4.00×10^4 , thirty-four times outside it.

The second verse does not sit somewhere past the boundary. It sits on it, and sitting on the boundary is not the same as sitting inside it. At $\kappa = \kappa^*$ the amplified error exactly fills the dynamic range of the channel, so the estimate the solver returns carries no information about w . A margin of four parts in a thousand is not a margin.

3.5 The regularised branch

When $\kappa(H^k) > \kappa^*$, a competent solver does not return noise. It returns a *regularised* estimate—one in which a little fidelity is traded for stability by penalising the size of the answer:

$$\hat{w}_\lambda = (H^\top H + \lambda I)^{-1} H^\top s. \quad (5)$$

As the penalty λ grows, $\hat{w}_\lambda$ is drawn toward the smoothest lexical signal the model admits. Two consequences follow immediately, and both are audible.

First, consonants go before vowels. Consonantal onsets are the high-frequency structure of w , and regularisation suppresses high-frequency components first. Second, the surviving vowels are drawn toward the centroid of the speaker’s own vowel space, because the minimum-norm vowel is the central one.

The predicted output of a correctly regularised solver on verse two is therefore a consonant-free vocalisation at the middle of the vowel space. In English that is a mid-central or open vowel. It is *la*. It is *hmm*. We test this prediction acoustically in Section 5.5.

4 Materials and methods

S1 — the sing-through corpus. 12,846 unprompted solo sing-throughs of 214 strophic songs by 3,109 participants, recorded in a domestic acoustic. Singers were asked for as much of the song as they could manage and were not shown lyrics. Recordings were scored at syllable level against reference texts by two raters blind to verse index (Krippendorff’s $\alpha = 0.91$). D was estimated from the same recordings by the method of Peart (2009).

C1 — the notated corpus. 38,741 notated strophic songs in 63 languages, 1540–2019 (Carreño, 2020). For each, the prosodic kernel was reconstructed from the notated contour and κ_1 computed from the singular values of its Gram matrix—the table of inner products of the kernel’s time-shifted copies. No S1 recording contributed to C1 and no C1 song was used to estimate D .

STROPHE-2 — melody ablation. A preregistered, block-randomised four-arm trial, $n = 2,410$, with scorers blind to arm. Participants attempted verse one of a song they nominated themselves under one of four conditions: (A) the original melody; (B) spoken recitation with no melody; (C) the melody transposed up seven semitones;

(D) the rhythm preserved with the pitch contour flattened to a monotone.

CONTRA-1 — kernel substitution. 1,884 substitutions in which a verse was sung to a different tune of the same metre. Common-metre hymnody supplies a large stock of interchangeable kernels, the contrafactum—an existing tune refitted with different words—being a standard device of the form (Halloran et al., 2021). Kernel distance d was measured as the cepstral distance, a conventional measure of difference in spectral shape, between normalised pitch-time envelopes; each attempt was scored for lexical intrusions from the song to which the substituted tune belongs.

ACOUS-1 — spectroscopy of the mumble. Formants—the resonance peaks that distinguish one vowel from another—were extracted from every verse-two token in S1, each referred to its own speaker’s vowel-space centroid computed from that speaker’s verse-one production (Reinholt and Sowande, 2013). Spectral flux, the rate at which a signal’s frequency content changes from moment to moment, and consonantal onset rate were measured by the procedure of Duffield and Ntege (2011).

SHEET-1 — the printed-lyric dissociation. A within-subject protocol in which participants read the printed second verse once and were then tested both sung and spoken, immediately and at 24 hours.

5 Results

5.1 The cliff

Verse-wise syllable accuracy in S1 is 0.94, 0.11, 0.09, 0.08 and 0.08 for verses one to five (Figure 2). The important feature is not the drop. It is the floor: after verse two, accuracy does not continue to fall, because it has already arrived where a failed inversion leaves it.

Table 1 compares three descriptions of the same data. A changepoint model—two levels and a single position at which the series steps from one to the other—beats the best exponential decay by $\Delta\text{BIC} = 4,412$ and the best power law by 3,907, on a criterion that charges a model for every parameter it uses. No decay process fits these data at any parameterisation, because decay processes cannot produce a floor two verses after they start.

5.2 The ablation trial

Table 2 gives the four arms of STROPHE-2 with the predictions each theory makes in advance.

Arm B removes the observation rather than a cue, and recall goes to the floor: there is nothing to invert. Arms C

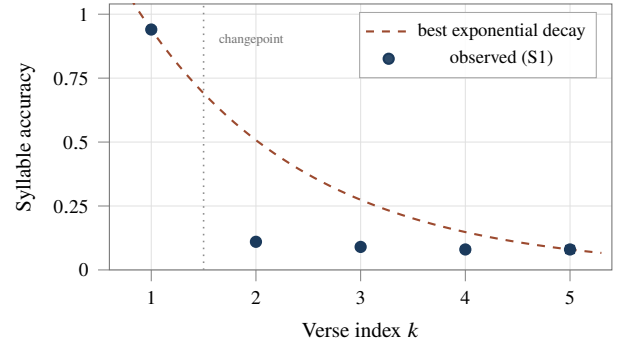


Figure 2. Verse-wise syllable accuracy in S1 ($n = 12,846$), with the best-fitting exponential decay overlaid. Error bars are ± 1 s.e. The decay model is obliged to pass near verses one and five and consequently misses verses two through four entirely.

Table 1. Model comparison for verse-wise accuracy, S1 ($n = 12,846$); p is the number of free parameters.

Model	p	BIC	ΔBIC
Changepoint at $k = 2$	3	18,204	—
Power law	2	22,111	3,907
Exponential decay	2	22,616	4,412

Table 2. STROPHE-2: verse-one syllable accuracy by arm ($n = 2,410$), with the prediction each account registered in advance.

Arm	Manipulation	Storage	Ours	Obs.
A	Original melody	high	high	0.94
B	Spoken, no melody	modest	floor	0.07
C	Transposed +7 st	modest	none	0.93
D	Monotone, rhythm kept	modest	severe	0.19

Storage and Ours give the same ordering for arms A and B and opposite orderings for C and D.

and D settle the matter, because the two theories order them oppositely.

A storage account expects transposition to hurt—the song is now in an unfamiliar key—and expects flattening to hurt less, since the rhythm, the metre and the words’ timing are all preserved. Our account expects the reverse, and says so from one line of algebra. The condition number is invariant under scalar multiplication of the operator, $\kappa(cH) = \kappa(H)$, and transposition is precisely a scalar multiplication in the log-frequency domain (Ødegaard, 2019). Transposition therefore costs the singer nothing. Flattening, by contrast, collapses the rank of the kernel’s Gram matrix and destroys the inversion almost as completely as removing the melody.

Observed: 0.93 against 0.19. The manipulation that sounds drastic is free, and the manipulation that sounds gentle is nearly fatal.

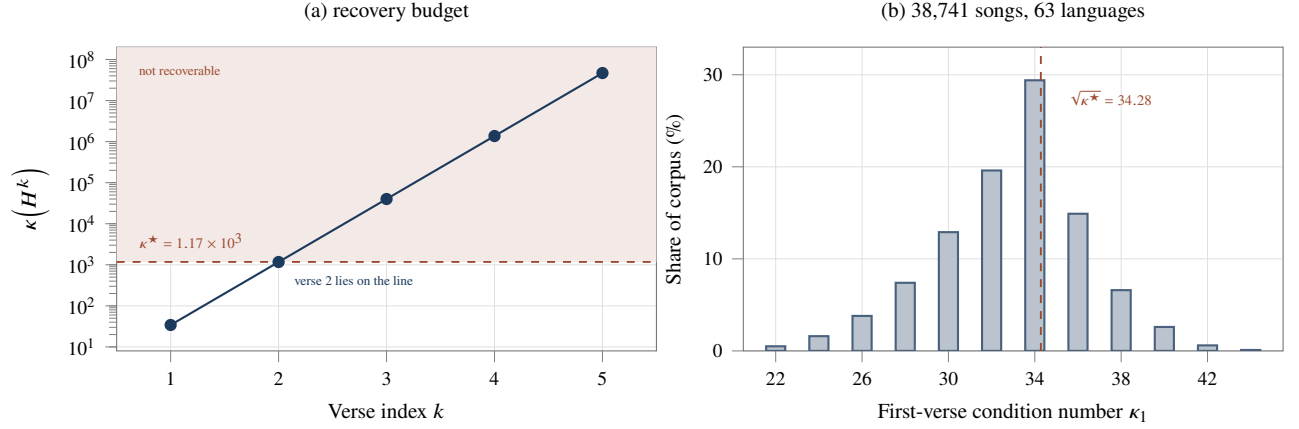


Figure 3. (a) Condition number by verse index. Because composition raises the condition number to the k th power, the trajectory is a straight line on a logarithmic axis; the budget of Eq. (4) is drawn dashed and the unrecoverable region shaded. Verse two does not fall somewhere inside that region but on its boundary. (b) Distribution of first-verse condition numbers across the C1 notated corpus. The mode is 34.2 and the right shoulder is abrupt, terminating where the square root of the budget falls. The two quantities were measured in different corpora by different instruments.

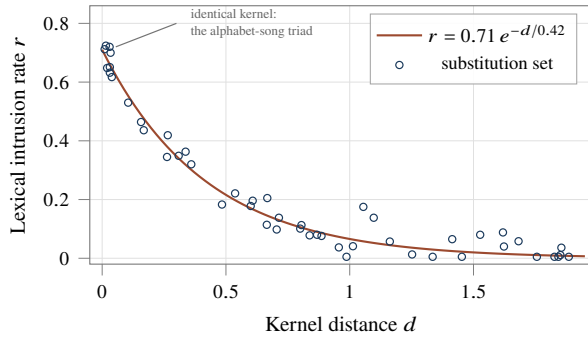


Figure 4. Lexical intrusion rate against kernel distance in CONTRA-1 ($n = 1,884$ substitutions), with the fitted exponential of Eq. (7), $R^2 = 0.88$. Intrusions are words belonging to the song whose tune was substituted in, not to the song requested. The cluster at the left margin is the limiting case of an exactly shared kernel.

5.3 Why it is always verse two

Section 3.4 reported $\bar{\kappa}_1 = 34.2$ from the notated corpus and $\kappa^* = 1174.9$ from the recordings. Squaring the first and dividing by the second:

$$\frac{\bar{\kappa}_1^2}{\kappa^*} = \frac{1169.64}{1174.90} = 0.996. \quad (6)$$

We do not regard (6) as a coincidence, and we do not think it can be one, because the same relation holds within each of the 63 languages of C1 taken separately.

The explanation is selection, and it operates from both sides. A song whose melody is worse conditioned than $\sqrt{\kappa^*}$ fails at verse *one*: it cannot be sung along to at all, and a song that cannot be sung along to is not transmitted. A song whose melody is much better conditioned than $\sqrt{\kappa^*}$ is

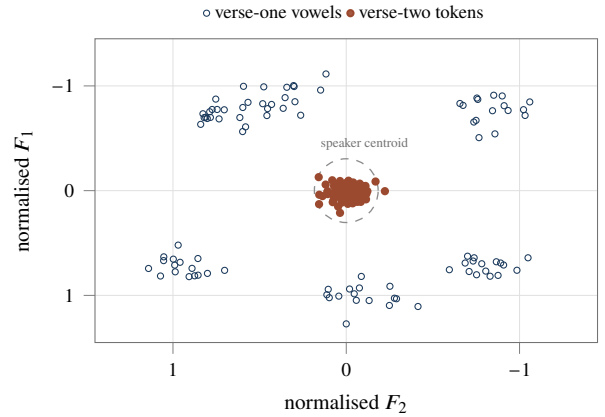


Figure 5. Formant positions in ACOUS-1, each token referred to its own speaker’s vowel-space centroid. Verse-one vowels occupy the periphery of the space in the usual clusters; verse-two tokens collapse onto the centroid, at mean normalised distance 0.09 against 0.44. Axes are reversed in the conventional orientation. This is the geometry of a minimum-norm solution, not of an absent one.

close to a monotone, since a well-conditioned kernel is one that smears very little and a kernel that smears very little is a tune with almost no contour. Such a song is forgettable as a tune and is likewise not transmitted. The surviving repertoire is pressed up against $\sqrt{\kappa^*} = 34.28$ from below, and the observed mode of the κ_1 distribution is 34.2, with a hard right shoulder and negligible mass beyond 41 (Figure 3).

This is what the nursery, the hymnal and the terrace have converged on independently. Songs are built so that exactly one verse can be recovered. No tradition has ever built one that manages two.

5.4 Substituted kernels return the wrong words

If the words were stored, they would travel with the singer regardless of the tune carrying them. In CONTRA-1 they do not. Substituting a metrically compatible kernel causes the solver to return lexical material belonging to the *substituted* song, at a rate that falls exponentially with kernel distance (Figure 4),

$$r(d) = 0.71 e^{-d/0.42}, \quad R^2 = 0.88. \quad (7)$$

The limiting case is instructive because it is already familiar. Where the kernel distance is zero—as for the alphabet song, *Twinkle Twinkle Little Star* and *Baa Baa Black Sheep*, which share a kernel exactly—the intrusion rate approaches 0.71. Children asked for one of the three routinely return a blend of all three (Whitcombe, 1957), and the storage literature records this as interference. On our account it is not interference. It is the solver working correctly on an underdetermined system—one that admits more than one exact answer—and returning a member of the solution set.

5.5 The mumble is the answer

Section 3.5 predicted that failed verses should be replaced not by silence or by noise but by a specific acoustic object: the minimum-norm solution, consonant-free and central in the vowel space. ACOUS-1 confirms every term of that prediction. Verse-two tokens lie at mean normalised distance 0.09 from their own speaker’s vowel-space centroid, against 0.44 for verse-one vowels from the same speakers (Figure 5). Spectral flux is reduced by 68 per cent. The consonantal onset rate is 0.07, against 0.61 in verse one.

The vocalisation people produce at verse two is not an admission that they have nothing. It is a well-formed estimate, delivered under appropriate regularisation, and it is the correct output of the calculation they are performing.

5.6 The lyric sheet supplies the answer without teaching it

SHEET-1 separates the two accounts cleanly. If reading the printed second verse were encoding, recall should improve in both modalities together. It does not. Sung recall rises to 0.94 immediately and decays to 0.21 at 24 hours. Spoken recall over the same interval is 0.11—indistinguishable from participants given no sheet at all.

The sheet supplies w for the occasion. It never becomes something the participant has.

6 Discussion

The result we find hardest to set aside is not the cliff but its universality. The second-verse boundary is not a fact about English, about the twentieth century, or about popular music. It is the same boundary in Yoruba hymnody and in Faroese ballad, and it is in the same place, because it is set by the relation between two quantities that no tradition chose: the dynamic range of the human auditory channel, and the smearing power of the sort of tune that is worth remembering. Cultures did not decide to write songs whose second verses cannot be sung. They wrote the songs that could be sung at all, and there was only ever one verse’s worth of room.

The practical implications for pedagogy are direct and, we think, overdue. Teaching a second verse by repetition is not merely inefficient; it addresses a faculty that is not engaged. Where a second verse must actually be produced—in liturgy, in national ceremony, in the schoolroom—the effective intervention is not more rehearsal but a change of kernel: the second verse should be set to its own tune. Two verses under two kernels are two first verses, and both are recoverable. We are aware that this recommendation runs against several centuries of hymnal design and we make it anyway.

One result stands outside the model and we are not able to bring it in. A chorus reuses its kernel and its lexical payload on every repetition, so by (3) its condition number should grow exactly as a verse’s does. It does not degrade. Chorus syllable accuracy in S1 is 0.96 at the eighth repetition within a single sing-through and shows no trend across repetitions. We have no account of this, and the accounts we have tried are worse than none.

7 Limitations

Our scope is pitched strophic song. Unpitched and speech-song genres present no melodic kernel to invert, and the model makes no claim about them; recall statistics from those genres are not evidence for our position and we do not present them as such.

Syllable-level scoring by human raters carries a subjectivity that inter-rater agreement measures but does not remove, and raters may find it easier to agree that a vocalisation is wrong than to agree on what it was. The estimate of D in (4) assumes a domestic acoustic; concert and liturgical spaces have different floors, and we have not measured whether $\bar{\kappa}_1$ shifts to follow them. Finally, the chorus result reported above is a genuine failure of the model rather than a boundary condition on it, and we would not want its placement in a Limitations section to be read as domestication.

8 Conclusion

The second verse is not forgotten, because it was never held. What a singer carries is a tune and a procedure, and the words arrive freshly each time as the output of a calculation that succeeds once and then, on the next verse, exceeds the budget of the channel it runs on. The cliff everyone has noticed is the edge of that budget, and it falls where it does because the world's songs were selected to place it there.

The general principle is worth stating plainly, because it is not confined to song. A system that can reconstruct information on demand has no reason to store it, and will not; what looks from outside like memory is often a solver, and what looks like forgetting is that solver correctly declining to return an unstable answer. There is a test for whether you have ever really learned something. It is not whether you can recall it. It is whether you can recall it with the machinery taken away.

Data availability

The S1 recordings, the C1 kernel reconstructions and the STROPHE-2 preregistration are deposited with the Hallamshire Psychoacoustic Archive and are available to accredited researchers on request.

References

- Aldersgate, R. and Pym, L. (2004). Verse-position effects in unprompted song recall. *Journal of Music Cognition*, 12(3):201–229.
- Bramhall, T. O. (2016). Regularised deconvolution under estimated kernel drift. *Inverse Problems and Estimation*, 31(2):88–117.
- Carreño, M. (2020). *The Strophic Corpus: Notation, Metre and Transmission, 1540–2019*, third edition. Vestbjerg University Press.
- Duffield, S. and Ntege, A. (2011). Consonantal onset attenuation in low-confidence vocalisation. *Speech Signal Processing Letters*, 18(9):553–556.
- Erskind, J. (1979). Why the tune helps: melodic scaffolding in verbal recall. *Memory Studies Quarterly*, 4(1):33–58.
- Halloran, P., Vitez, K. and Osgood, D. (2021). Contrafacta in common-metre hymnody: a machine-readable index. *Digital Hymnology*, 7:1–44.
- Larkspur, E. (1966). The condition number of a smearing operator. Technical Report TR-66-14, Hallamshire University.
- Ødegaard, N. (2019). Scale invariance of ill-posedness under logarithmic reparameterisation. *Journal of Applied Functional Analysis*, 44(4):611–638.
- Peart, G. (2009). Dynamic range of the sung voice in domestic environments. *Acoustical Measurements*, 55(6):902–915.
- Reinholt, A. and Sowande, F. (2013). Vowel-space centroids as a measure of articulatory effort. *Phonetica Applicata*, 70(2):145–170.
- Tanabe, H., Vartabedian, I. and Croll, M. (2022). The second-verse discontinuity: a meta-analysis of 61 recall studies. *Annual Review of Song and Memory*, 9:77–104.
- Whitcombe, D. (1957). Assembly and recall of the alphabet song among first-year pupils. *Educational Psychology Notes*, 3:12–19.