Supplemental Information

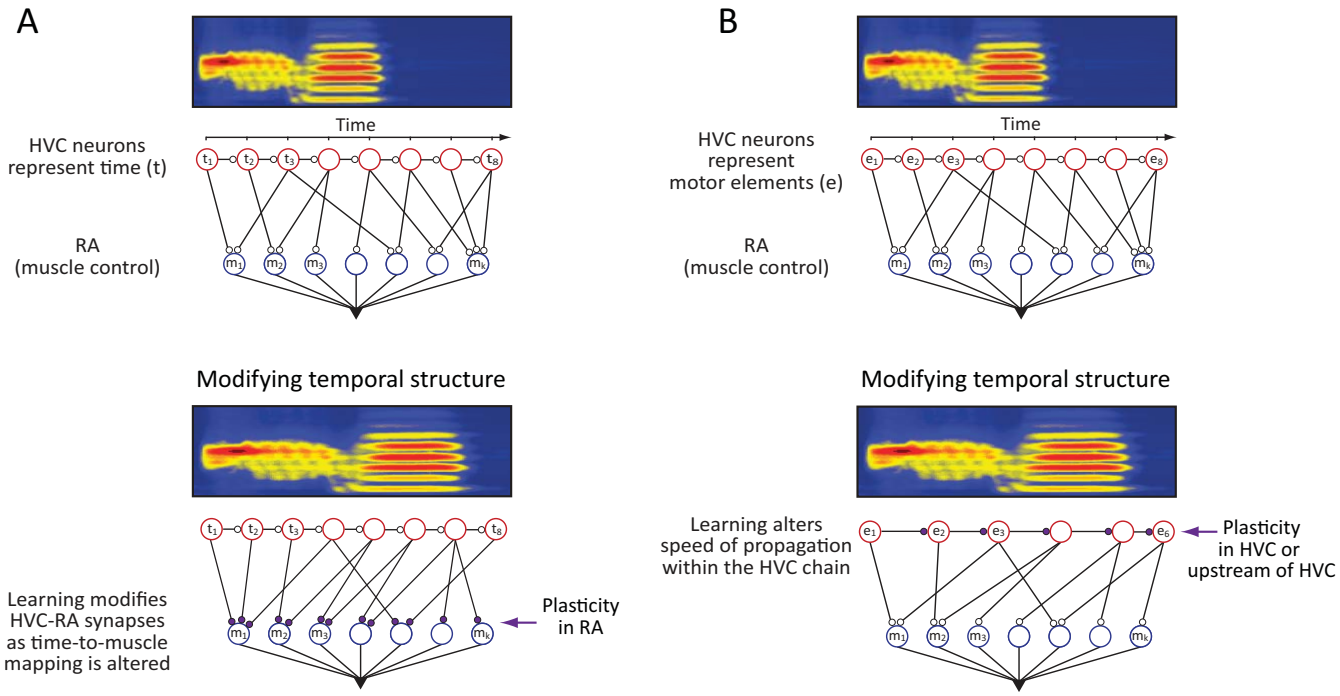

Figure S1 (related to Figure 1). Conceptual models for how temporal structure of birdsong can be modified, involving plasticity in RA (A) and HVC (or upstream of HVC) (B) respectively. (A) *(top)* Presumed functional organization of the motor pathway underlying song. A synaptic chain network in HVC serves as a generic time-keeper. The unary time code produced by RA-projecting HVC neurons is transformed into a specific motor program through connections to neurons in motor cortex analogue RA, which control vocal musculature. *(bottom)* Within this framework, changes to temporal structure (e.g. lengthening of a song segment) can be implemented by reorganizing the connections between time-keeper neurons in HVC and muscle-related neurons in RA. Additional plasticity within RA (Sizemore and Perkel, 2011) could also serve to transform the motor program in RA. After learning each time-keeper neuron would drive a different set of RA neurons than before. **(B)** *(top)* Alternatively, HVC neurons could encode specific motor (i.e. song) elements, the specific timings of which are amenable to change. *(bottom)* Within this framework, the duration of a song segment can be altered by changing the speed of propagation in the part of the HVC network encoding the segment, leaving the connection between HVC neurons and RA neurons intact. Plasticity underlying this form of learning could plausibly be implemented within the HVC network or at the level of its inputs.

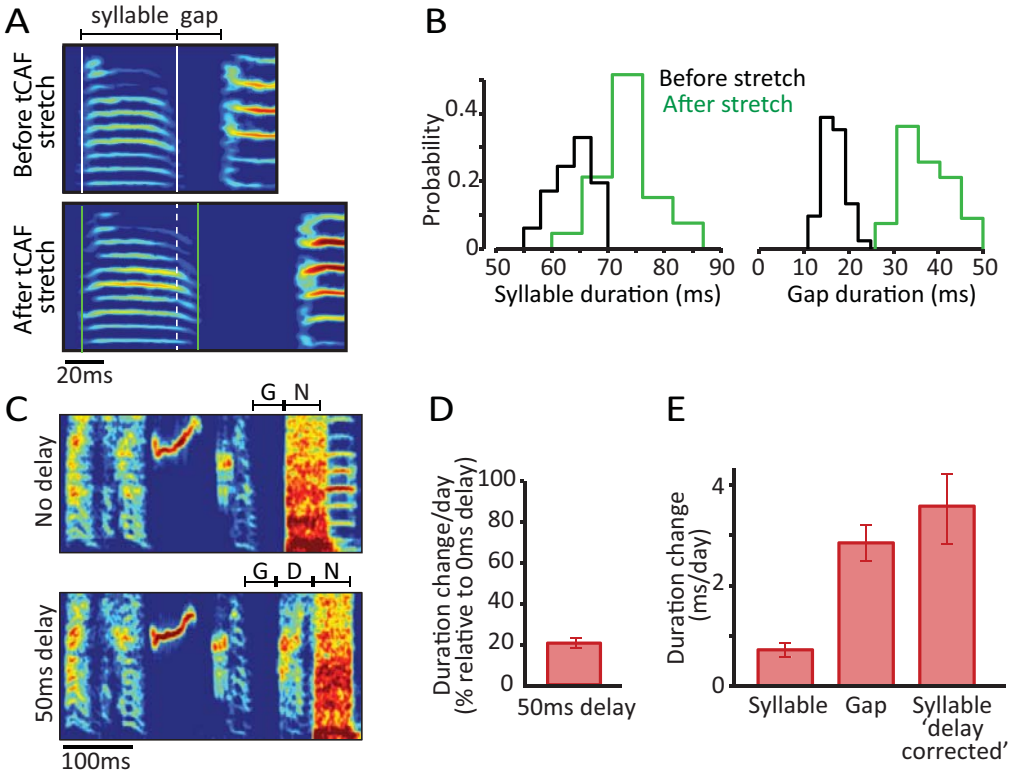

Figure S2 (related to Figure 2). The slower learning rate for syllables relative to inter-syllable gaps in our tCAF paradigm can be explained by the syllables being temporally further removed from the reinforcer. (A) Example of a 'syllable + gap' segment that stretched after 4 days of tCAF. Though gaps change more than syllables, note the significant change also in syllable duration, marked by white lines before the stretch and green lines after (see also Figure S3C). **(B)** Syllable and gap duration distributions before and after tCAF for the example in A. The mean stretch in the syllable and gap over the 4 days was 9.8 ms ($p = 10^{-28}$) and 20.5 ms ($p = 10^{-62}$) respectively. **(C)** Example spectrograms showing no delay in the noise-feedback (N) (top) and 50 ms delay (D+N) relative to the end of the targeted gap (G). **(D)** Summary of the delayed noise-feedback experiments showing a 79.7% reduction in learning rates of the targeted gaps in the 50ms delay condition (n = 3 birds). **(E)** Gap and syllable learning rates across 21 birds. After correcting for the difference in the reinforcement delay between syllables and gaps (see Supplemental Experimental Procedures for details), the capacity for syllables to change was not significantly different from that of gaps (Gap vs. Syllable 'delay corrected', p = 0.35).

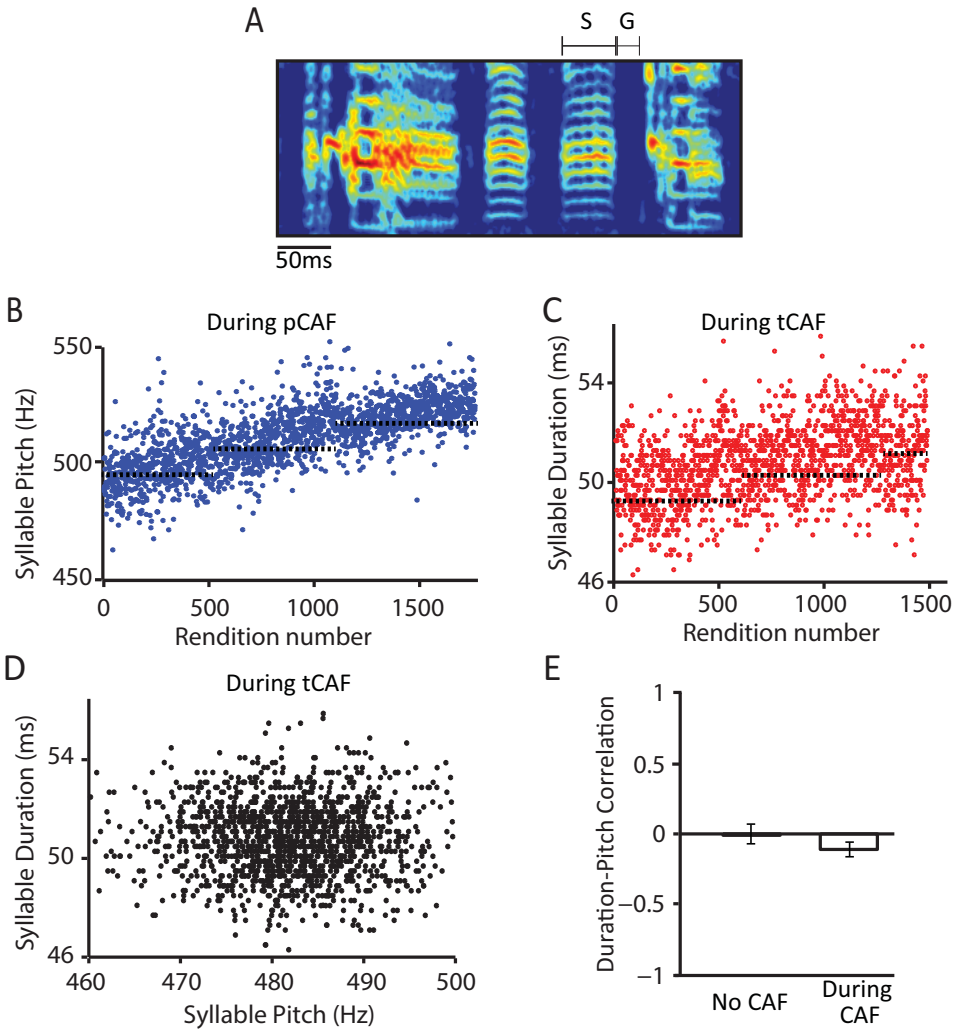

Figure S3 (related to Figure 2). Rendition-by-rendition estimates of pitch and duration for a targeted syllable during normal singing and during CAF. (A) Song spectrogram from a bird that underwent pCAF and tCAF targeting the same syllable (S) (tCAF target also included the ensuing gap (G))**. (B)** Rendition-by-rendition estimates of syllable pitch for a day of pCAF. Difference in mean syllable pitch over the course of the day was 32.9 Hz ($p = 10^{-71}$; using first and last 100 songs of the day for comparison). **(C)** Rendition-by-rendition estimates of syllable duration during a day of tCAF in the same bird. Increase in mean syllable duration over the course of the day was 1.6 ms ($p = 10^{-10}$). Dashed horizontal lines in (B) and (C) indicate the thresholds for white noise (for tCAF, threshold is estimated as actual online threshold was applied to S+G). **(D)** Scatterplot of duration vs. pitch for the targeted syllable in (C) ($r = 0.004$, $p = 0.88$). **(E)** Summary statistics for birds that did both tCAF and pCAF ($n = 5$ birds) showing no significant correlation between pitch and duration for targeted syllables either in baseline (no CAF) or during CAF ($p = 0.89$ and 0.09 respectively). Summary statistics computed only on catch trials without white noise interference (see Experimental Procedures).

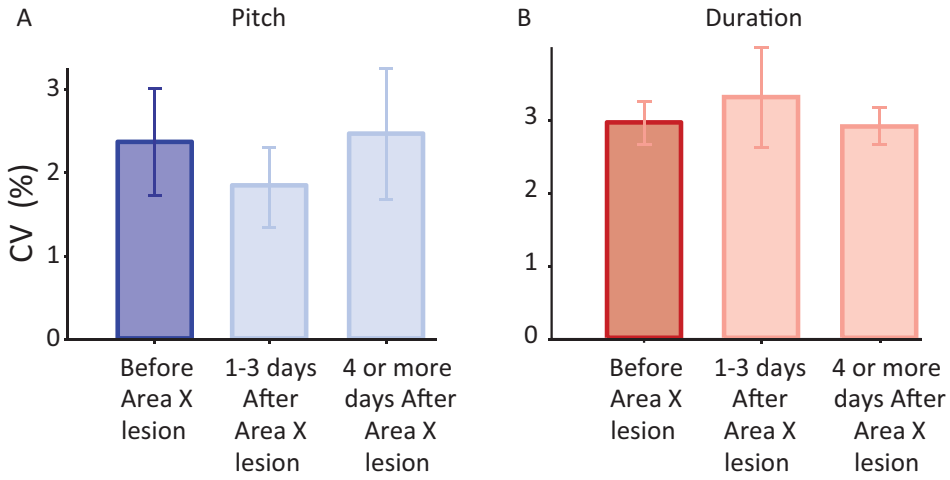

Figure S4 (related to Figure 3). Acute and persistent effects of Area X lesions on spectral and temporal variability. (A) Variability in the pitch of harmonic stack syllables showed a small, but significant decrease immediately following Area X lesions (1 - 3 days post-lesion), but recovered to pre-lesion levels after 3 days. The coefficient of variation (CV) for pitch went from 2.4 ± 1.7 % before lesion to 1.8 ± 1.3 % immediately after (n = 7 birds, average reduction of 22.1%; p = 0.001), consistent with ref. (Kojima et al., 2013). Pitch CV measured 4 or more days after lesion, however, was 2.5 ± 2.1% (p = 0.90 compared to before lesion), indicating that the effect of Area X lesions on pitch variability is transient. **(B)** Variability in the duration of syllables and gaps, however, remained unchanged after Area X lesions. CV of interval durations was 3.0 ± 0.8% before lesions and 3.3 ± 1.8% (p = 0.65 compared to pre-lesion) in the first three days after lesions and 2.9 ± 0.6 (p = 0.91 compared to pre-lesion) 4 or more days after lesion.

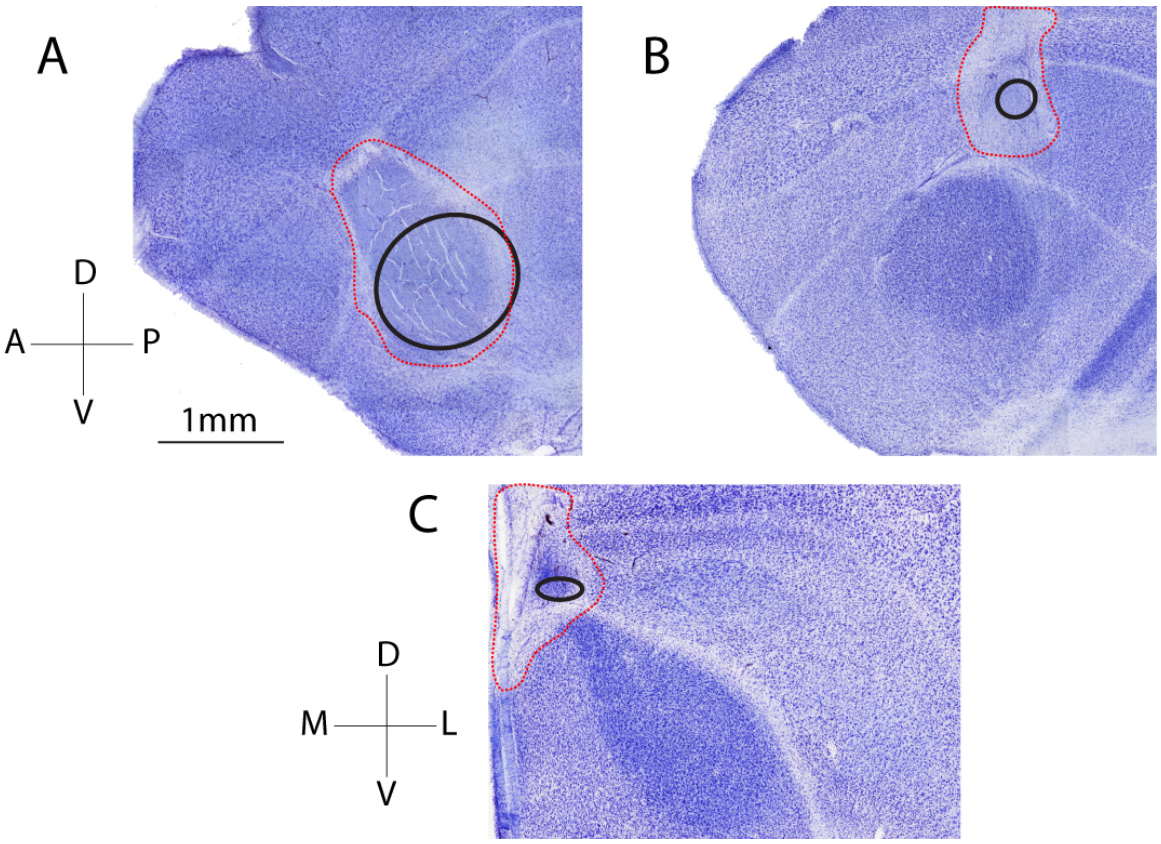

Figure S5 (related to Figures 3, 4, 5, and 6). Histology of Area X, LMAN, and MMAN lesions. (A-B) Sagittal brain slices showing chemical lesions targeting Area X (A) and LMAN (B) respectively. Lesion boundaries demarcated by dashed red lines and estimated from close inspection of viable cell bodies; estimated locations of the targeted nuclei is demarcated by black lines and based on the corresponding (medial-lateral) slices in zebra finch atlases. **(C)** Coronal brain slice showing lesions targeting MMAN. Same convention as in A and B.

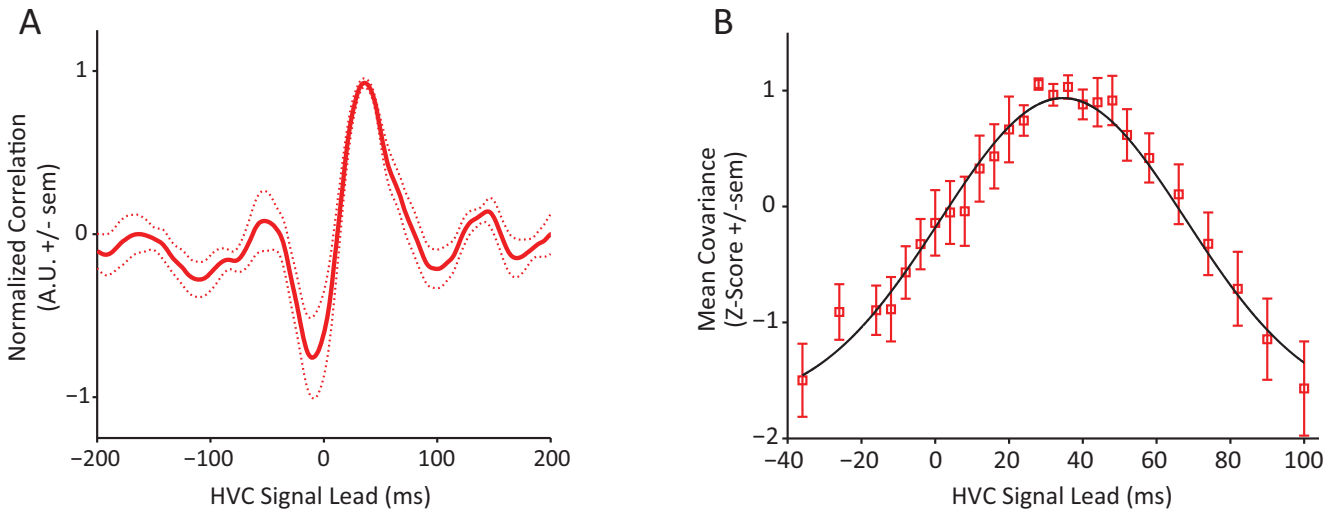

Figure S5 (related to Figure 7). Two independent measures of the temporal relationship between HVC activity and vocal output shows neural activity leading sound by ~35 ms. (A) Normalized cross-correlation function between HVC activity and sound amplitude averaged across 7 birds (solid line: mean, dashed line: s.e.m.; peak time = 35 ms; data from each bird is an average over 4 days (100 renditions/day)). The mean lag at peak correlation, across all recorded birds, was 34.9 ± 4.4 ms. **(B)** Estimated covariance between the temporal variability in HVC activity and song over a range of HVC signal lead times (see Supplemental Experimental Procedures). Red squares show the mean covariance z-scores, across n = 7 birds, each averaged over 4 days (~100 renditions/day), at the specified HVC signal lead times; error bars report s.e.m. A Gaussian fit to this data (μ = 34.7 ms, σ = 33.3 ms, R^2 = 0.97) is shown in black.

Table S1 (related to Figures 3, 4, 5, and 6). Coordinates and pharmacological doses for Area X, LMAN, and MMAN lesions. All coordinates are relative to the bifurcation of the midsagittal sinus. Head angle is relative to beak horizontal. Injections of 4% NMA at each site were in steps of 9.2 nL using a Nanoject II (Drummond Scientific) every 10 sec. In two birds, LMAN was lesioned electrolytically (200 μA, 60 sec at each of the 4 sites below).

Target	Amount of NMA (nL)	Anterior (mm)	Lateral (mm)	Depth (mm)	Head angle (degrees)
Area X					
Site 1	83	5.5	1.5	2.9	0
Site 2	83	5.5	2.0	2.9	0
Site 3	83	6.1	1.5	2.9	0
Site 4	83	6.1	2.0	2.9	0
LMAN					
Site 1	74	4.7	1.75	1.8	20
Site 2	74	5.3	1.75	1.8	20
Site 3*	-	5.3	1.2	1.8	20
Site 4*	-	5.3	2.2	1.8	20
MMAN					
Site 1	37	5.3	0.3	1.8	20
Site 2	37	5.3	0.7	1.8	20
Site 3	28	4.9	0.3	1.8	20
Site 4	28	4.9	0.7	1.8	20

* Only for the 2 birds lesioned electrolytically

Table S2 (related to Figures 3, 4, 5, and 6). Extent of Area X, LMAN, and MMAN lesions. Estimate of the lesions relative to age-matched intact brains. See Supplemental Experimental Procedures for detailed description of quantification methods used.

Area X lesions	Bird 1[i]	Bird 2[ii,iv]	Bird 3[i,iii]	Bird 4[i,ii,iv]	Bird 5[i,iii]	Bird 6[ii]	Bird 7[i,ii]
% lesioned	98	97	96	96	94	94	90
	Bird 8[i,ii,iv]	Bird 9[ii]	Bird 10[i,ii,iv]	Bird 11[i,iii]	Bird 12[i,ii]		
	84	83	83	82	72		

LMAN lesions	Bird 13	Bird 14[ii]	Bird 15[ii]	Bird 16	Bird 17[i]	Bird 18[i]	Bird 19[i,ii]
% lesioned	100	100	95	93	90	89	87
	Bird 20[i,ii]	Bird 21[i,ii]	Bird 22[ii]	Bird23[ii]			
	85	82	81	80			

MMAN lesions	Bird 23[ii]	Bird 24[ii]	Bird 25[ii]
% lesioned	100	100	75

[i] pCAF

[ii] tCAF

[iii] Return to baseline after pCAF

[iv] Return to baseline after tCAF

Supplemental Experimental Procedures

Lesion quantification. We measured the volumes of intact LMAN and Area X across three adult control birds by tracing the boundaries of Area X and LMAN in 100 µm thick Nissl-stained sagittal brain slices. LMAN and Area X appear as regions with a higher density of Nissl-stained cell bodies or stronger staining than surrounding areas. In identifying Area X and LMAN, we were also guided by published brain atlases of the zebra finch brain that provide well-established landmarks relative to where LMAN and Area X can be located (e.g. lamina pallio-subpallialis and lamina mesopallialis) (Karten et al.). The areas traced were then summed and multiplied by 0.1mm (section thickness) to obtain volume. The average volumes we obtained for LMAN (0.22 mm^3) and Area X (2.28 mm^3) in control birds were very similar to those reported in published studies (Airey et al., 2000; Thompson et al., 2011). For experimental birds, lesion areas were typically distinct in their lack of neuron cell body staining, the presence of gliosis, and other obvious tissue damage when viewed under a 10x microscope objective. The boundaries of unlesioned portions of LMAN and Area X with healthy-looking cells were traced, summed, multiplied by 0.1 mm and then divided by the average Area X and LMAN volumes in control birds (see Figure S5 for examples of boundaries, Table S2 for quantification). Lesions of Area X did not extend to LMAN as verified by careful examination of cell bodies in all slices with LMAN. Moreover, analysis of songs (blind to lesion treatments) showed no permanent effect on vocal variability after Area X lesions (unlike actual LMAN lesions that significantly reduced variability), providing further functional confirmation of the fact that LMAN remained intact in the Area X lesioned birds.

For MMAN, as previously reported (Foster and Bottjer, 2001), the boundaries of the nucleus are not always clear in Nissl stained section, thus a reliable quantification based on tracing intact areas was not possible. We thus followed Foster & Bottjer's (2001) method of lesion assessment. Briefly, in the coronal plane, MMAN's location is well delineated by the two laminae (lamina mesopallialis and lamina pallio-subpallialis) on the dorsal-ventral axis; the lateral ventricle and the medial-most extent of LMAN identify the medial and lateral extent of MMAN respectively. The rostral-caudal axis of MMAN generally coincides with LMAN's (see Foster and Bottjer, 2001 for details). These landmarks provide the approximate location of MMAN along the three axes. We thus counted slices with LMAN present (as well as slices immediately before and after the rostral-caudal extent of LMAN) which were *not* lesioned in the expected MMAN location. Based on this criterion, two birds had complete lesions while the third had 3 partially lesioned slices out of a total of 12 in both hemispheres. We thus gave this bird a conservative lower-bound estimate of 75% lesion. All lesion quantifications were done blind to CAF outcomes.

Pitch estimation algorithm. The discrete Fourier transform of a 5 ms signal has a 200 Hz frequency resolution (time-frequency tradeoff) which is too coarse for pitch estimation. However, because we target harmonic stacks, we make the assumption that the data can be well approximated by a fundamental frequency and its harmonics. This allows us to obtain highly precise pitch estimates of even very short (i.e. 5 ms) syllable snippets. The algorithm is as follows: 1) Perform a fast Fourier transform (FFT) on the full 5 ms dataset (220 points given our 44.15 kHz sample rate), which has power at 0, 201, 402, …Hz, etc (rounded to nearest Hz). 2) Discard the first data point of the dataset to create a new FFT now with power at 0, 202, 404...Hz, etc. 3) Repeat the procedure of discarding the first data point in the diminishing data set to create up to 73 frequency based descriptions of the 5 ms sound, each offset relative to its neighbor by, on average, 2.7 Hz in the target fundamental range (typically estimated at 400 - 600 Hz). 4) Give the algorithm an estimate (accurate to within 100 Hz) of the fundamental frequency of the syllable (this 'prior' is updated as the syllable pitch shifts during pCAF). Sum the power in the frequency bins that best corresponds to the fundamental and first 19 harmonics of this 'prior' for each of the 73 FFT datasets. The fundamental of the FFT set with the highest sum (normalized to the power across all frequency bins) is used as the estimate of the pitch. When tested using playback of 5 ms sound with known pitch in a typical bird cage, the algorithm was able to accurately estimate pitch down to a 2-3 Hz resolution even with signal-to-noise ratio of 5:1 (proportion of white noise added to signal). Our algorithm gave estimates of pitch that were highly correlated with off-line estimates using a previously published algorithm ($r = 0.92 \pm 0.09$, $n = 10$ birds) (Andalman and Fee, 2009). Our algorithm was implemented in LabVIEW and estimated the pitch of a 5 ms sound in less than 3 ms.

Interval duration estimates. Song recordings were first segmented to contain only target motifs. The total power in each segment was normalized to 1 and converted to spectrograms (5 ms Hamming window with 1 ms advancement) and log-transformed. These processed signals were the ones used for all subsequent analysis. Interval (i.e. syllable or gap) durations were estimated using the following two-step procedure. First, we generated motif templates to which all renditions in a catch trial block were aligned. These templates were obtained as follows: (i) Motif renditions were visually inspected and up to 5 motifs were selected. All renditions were aligned to these motifs using a DTW algorithm (see below) and all the aligned renditions were averaged to obtain an average template for each starting motif. (ii) The average templates were summed across frequency bins to produce "power envelopes". For each such power envelope, the distribution of power was fitted to a mixture of two Gaussians, with the Gaussian with the lowest mean considered to be noise. (iii) A threshold on each power envelope was then set at 2-3 standard deviations above the mean of the noise Gaussian (kept

constant within a bird). Time points where this threshold was crossed were noted and labeled as interval onsets and offsets depending on whether it was an up-cross or down-cross respectively.

In the second step, interval durations (syllables and gaps) were estimated as follows. (i) Each rendition was aligned to the average templates. (ii) Time points in the DTW-aligned renditions that matched the interval onset and offset time points in the average template were used to estimate interval durations. As a result of this procedure, we obtained up to 5 different estimates (one for each average motif template) for the same interval duration. Statistics of the length of an interval across renditions of a catch trial block were calculated separately for each set of estimates (belonging to the same average template) and then averaged across different average templates to obtain a more robust estimate. We used the same average templates and onset and offset times for drives across multiple days (e.g., tCAF up and down) so as to minimize the differences in estimates due to use of different daily templates.

Delayed noise feedback experiments. When we targeted 'syllable+gap' segments the average delay of the white noise feedback relative to the center of the gaps was 15.6 ± 6.1 ms, whereas the average delay to the center of the syllables was 62.8 ± 17.0 ms. Thus syllables were, on average, 47.2 ms further removed from the reinforcer (i.e. white noise). To test the effect of delayed reinforcement on learning in the temporal domain, we compared learning rates in tCAF experiments with no delay between the end of the target gap and the reinforcer ($Gap_{no\ delay}$) to rates after experimentally adding a 50 ms delay ($Gap_{50\ ms\ delay}$) (Figure S2C). This 50 ms delay reduced the learning rate within a bird by, on average, 79.7 ± 4.1% (n = 3 birds) relative to the no delay condition. To estimate what syllable learning rates would be if the reinforcer was delivered right after the end of the syllable ($Syllable_{delay\ corrected}$), we applied the following formula (see Figure S2E):

$$Syllable_{delay\ corrected} = Syllable \times \frac{Gap_{no\ delay}}{Gap_{50\ ms\ delay}} \quad (1)$$

where '*Syllable'* is the learning rate in tCAF experiments targeting 'syllable + gap' segments.

Relationship between HVC activity and vocal output (Cross-Correlation Method). We estimated the time-lag between HVC activity and vocalization by cross-correlating sound amplitude and HVC activity and identifying the lag at which the function peaks. ~100 renditions of song and associated HVC traces from a morning catch-trial block were aligned to an average template (as described in the main text's Experimental Procedures). The mean neural traces and sound amplitudes were calculated (as described in the main text), normalized by their maxima, and mean subtracted. The two resultant time series were cross-correlated; the cross-correlation function was then normalized by its peak value. For each bird (n = 7), this procedure was repeated for each of 4

catch trial blocks. The normalized correlation functions were averaged across the 4 blocks for each bird. Over the 7 birds, we found the peak in the cross-correlation function to indicate a lag between HVC activity and sound of 34.9 ± 4.4 ms. In addition, we calculated the mean cross-correlation function across birds by averaging over each bird's mean cross-correlation function (Figure S6A); this function exhibited a clearly defined peak at 35 ms lag.

Relationship between HVC activity and vocal output (Covariance Method). We estimated the covariance between temporal variability in HVC activity and vocal output over a range of time lags. ~100 renditions of the song were aligned to an average template (as described in the main text's Experimental Procedures). The warping paths from these alignments were then applied to the corresponding neural traces with 26 different time lags (range: -36 ms to +100 ms; negative shifts imply that vocalization temporally precedes HVC activity). For each lag, we calculated the correlation coefficient for all possible neural trace pairs in the block. Because the warping path used to align each neural trace was derived from the corresponding song, the average correlation coefficient reflected the covariance between the temporal variability in the song and the corresponding HVC activity. For each bird (n = 7), the procedure was repeated for 4 catch trial blocks; the mean correlation coefficient calculated at each lag time, averaged across blocks, and then converted to a z-score. We then averaged these z-scores across birds to generate a mean 'covariance' profile (Figure S6B). A Gaussian fit to the data (μ = 34.7 ms, σ = 33.3 ms, R^2 = 0.97) indicated a 34.7 ms lag between HVC activity and sound.

Supplemental References

Airey, D.C., Castillo-Juarez, H., Casella, G., Pollak, E.J., and DeVoogd, T.J. (2000). Variation in the volume of zebra finch song control nuclei is heritable: developmental and evolutionary implications. Proceedings of the Royal Society B: Biological Sciences *267*, 2099.

Andalman, A.S., and Fee, M.S. (2009). A basal ganglia-forebrain circuit in the songbird biases motor output to avoid vocal errors. Proceedings of the National Academy of Sciences *106*, 12518–12523.

Foster, E.F., and Bottjer, S.W. (2001). Lesions of a telencephalic nucleus in male zebra finches: influences on vocal behavior in juveniles and adults. Journal of Neurobiology *46*, 142–165.

Karten, H., Brzozowska-Prechtl, A., Prechtl, J., Wang, H., and Mitra, P. Zebra Finch Brain Atlas.

Kojima, S., Kao, M.H., and Doupe, A.J. (2013). Task-related "cortical" bursting depends critically on basal ganglia input and is linked to vocal plasticity. PNAS *110*, 4756–4761.

Sizemore, M., and Perkel, D.J. (2011). Premotor synaptic plasticity limited to the critical period for song learning. Proc. Natl. Acad. Sci. U.S.A. *108*, 17492–17497.

Thompson, J.A., Basista, M.J., Wu, W., Bertram, R., and Johnson, F. (2011). Dual pre-motor contribution to songbird syllable variation. J. Neurosci. *31*, 322–330.