Faculty of science

spectral Analysis

Bird Syllable Detection

Author:
Mohamed Abrash,
Paulina Ibek

March, 2024

Contents

1 Introduction **3**
1.1 Data Structure 3

2 Theory And Background **4**
2.1 Objectives And Challenges 4
2.2 Spectrogram Cross Correlation 5
2.3 The wavelet transform 7
2.4 Feature Extraction Using SVD 8

3 Results **10**
3.1 SPCC 10
3.2 Scalogram cross correlation 16
3.3 SVD 18

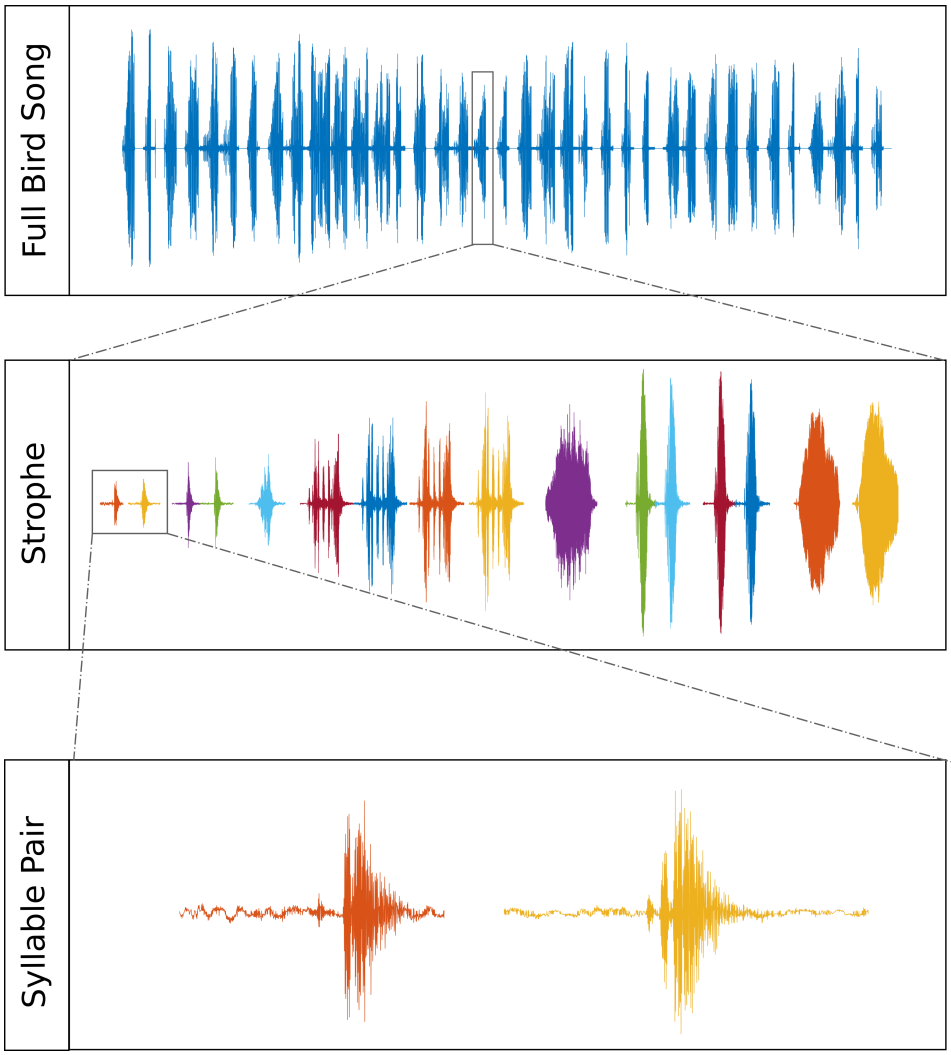


Figure 1: Six bird songs are provided where each song is segmented into smaller phrases called strophes. Each strophe consists of a series of syllables which are extracted from the original song and compared in pairs.

1 Introduction

At the intersection of ornithology and digital signal processing lies a captivating field where birds and researchers collaborate to deepen our understanding of bird biology and behavior while simultaneously refining our signal processing methodologies. For biologists, the cataloging of bird vocalizations serves as a vital tool in understanding avian behavior and helps unravel exciting connections to other biological aspects such as avian evolution and socialization. Historically, biologists have undertaken the tedious task of classifying bird songs and calls for the purpose of monitoring the evolution of a bird's vocal repertoire. Traditionally, this was done by recording bird calls of a given species and cataloging them, relying heavily on manual work by leveraging human sound perception with rudimentary digital signal inspection. However, these conventional methodologies are tedious and time-consuming. With the advent of technological advancements, an array of digital signal processing techniques has emerged, offering the capacity to automate some parts of the avian vocalization analysis. For scientists and engineers, the study of bird vocalizations serves as fertile ground for refining and testing various signal processing techniques. Moreover, avian vocalizations offer insights applicable to the development of natural language processing methodologies, as there are many parallels between avian communication and human language to draw upon.

In this project, we address the task of automating bird syllable detection by devising a similarity measure that helps determine if two bird syllables are similar. We begin with a short exploration of anticipated challenges and formal requirements that the similarity measure must fulfill. Subsequently, we introduce three distinct similarity measures: spectrogram cross-correlation utilizing spectrograms, the wavelet transform, as well as a feature extraction method that relies on singular value decomposition (SVD). In the methodology section, we present our procedure which we use to tune and test the different similarity measures. Finally, the results and analysis are shown in the results section.

1.1 Data Structure

The data used in this project consists of 6 song recordings gathered from two birds in three consecutive years 1989,1990, and 1991. Song labels are found in table 1. Each song was divided into smaller phrases or stophes which are further divided into syllables as shown in figure 1.

Table 1: This table shows the labels for the bird song data used during this project.

Song label	13B04	14A03	19A04	19A01	23B05	23A01
Bird label	H7-3	V-99	H7-3	V-99	H7-3	V-99
Year of recording	1989	1989	1990	1990	1991	1991

2 Theory And Background

2.1 Objectives And Challenges

The objective of this project is to develop and assess the robustness of various stochastic methods for determining whether two given syllables are sufficiently similar to be considered identical. Given the dynamic nature of these syllables, they cannot be treated as stationary signals. Therefore, an emphasis on temporal structure must be incorporated into our analysis. To achieve this, we utilize the spectrogram.

Our aim is to devise a function on the provided data that takes two syllables, X_1 and X_2, as input and returns a measure denoted by $\mu(X_1, X_2)$ indicating their similarity.

However, several challenges and requirements must be addressed for this measure to be effective. Firstly, bird songs are typically recorded in the birds' natural habitat, where background noise is prevalent. Even in controlled recording environments, some level of background noise is expected. Thus, our similarity measure must exhibit robustness against such noise.

Additionally, as birds produce sounds using their syrinx, natural imperfections and variations may exist. Even when the same syllable is repeated, slight deviations are anticipated. Consequently, these signals must still be classified as similar or identical despite these natural fluctuations. Moreover, the same phrase or syllable may be repeated at slightly different frequencies while still retaining its identity. Therefore, our similarity measure cannot simply compare signals on a one-to-one basis; rather, it must adopt a statistical approach that accounts for these natural variations as noise. Formally, to accommodate small natural deviations in the signals, a degree of smoothness is required in the similarity measure for each argument. This ensures that small deviations in one syllable correspond to minor changes in similarity.

Lastly, it is essential for the similarity measure to be translation invariant in time. In other words, the measure should yield consistent results regardless of the temporal alignment of the signals being compared.

Below, we summarize the key requirements:

- Robustness against background noise: The measure should remain relatively unaffected by the addition of background noise.

  $$\mu(X_1, X_2 + \epsilon) \approx \mu(X_1, X_2)$$

  where ϵ is white noise of amplitude much smaller than the signal X_i.

- Invariance under argument transposition: The measure should be unaf-

fected by the order in which the syllables are compared.

$$\mu(X_1, X_2) = \mu(X_2, X_1)$$

- Translation invariance in time: The measure should produce consistent results irrespective of temporal shifts in the signals.

$$\mu(X_1(t), X_2(t)) = \mu(X_1(t), X_2(t+s))$$

- Smoothness in each argument: Small deviations in time and frequency should result in proportionally small changes in similarity.

$$\mu(X_1, .), \mu(., X_2) \in \mathrm{C}^1$$

2.2 Spectrogram Cross Correlation

The spectrogram is a tool used for representing a signal in both the time and frequency dimensions. The spectrogram is defined as the absolute square of the short-time Fourier transform (STFT) of a time-varying signal, which is defined as

$$X(t,f) = \int_{-\infty}^{+\infty} x(t_1)h^*(t_1 - t)e^{-i2\pi f t_1} dt_1, \qquad -\infty < t, f < \infty, \tag{1}$$

where $h(t)$ is the window function. The spectrogram is thus defined as

$$S_x(t,f) = |X(t,f)|^2, \qquad -\infty < t, f < \infty. \tag{2}$$

In this project we used the fast Fourier transform to generate the spectrogram, where the number of FFT's was set to 1024.
The first algorithm used in this project for detecting the similarity between two syllables is the spectrogram cross correlation (SPCC), which utilizes the time and frequency marginals of the spectrogram. The spectrogram image can be denoted by a matrix $A(i,j)$, where the first index, i, corresponds to the frequency and the the second index, j, to the time axis. By taking the sum over all frequencies and times respectively, the spectrogram can be reduced to two one-dimensional signals called the time and frequency marginals, which are defined as

$$m_f(i) = \sum_j A(i,j) \tag{3}$$

$$m_t(j) = \sum_i A(i,j). \tag{4}$$

The marginals facilitate the comparison between two syllable pairs. Instead of analyzing two spectrograms, one can calculate the cross-correlation between the time marginals of two neighbouring syllables, as well as the cross correlation

between the frequency marginals. The cross correlation between two signals x and y is defined as

$$xcorr_{x,y}(\tau) = \frac{1}{N}\sum_{t}\frac{(x(t)-\bar{x})}{s_x}\frac{(y(t+\tau)-\bar{y})}{s_y}, \tag{5}$$

where τ is the lag, $\bar{x}$ and $\bar{y}$ are the mean values of the signals respectively, s_x and s_y are the sample means and N is the length of the signals. The cross correlation is a measure of similarity between two signals for different relative displacements between them. In this method we calculated the cross correlation for time and frequency lags up to 300. The reason for comparing the marginals at lags other than lag zero is due to inaccuracies when dividing the bird song into syllables. The shift in time between two syllables should not change the classification; the shift in frequency should affect the similarity between syllables. It is not a good idea, however, to only compare the frequency marginals at lag zero, since that set up would be very sensitive. Instead, a new Gaussian signal is created, with a peak at lag zero, which is then multiplied with the cross correlation between the frequency marginals. This approach gives more weight to similarities near lag zero, and acts as a penalty for increasing frequency lags. The total similarity for two syllables is then taken as the maximum value of the cross correlation between the time marginals times the maximum cross correlation of the frequency marginals and the Gaussian signal:

$$SPCC_{x,y} = \sup_{\tau}\left(xcorr_{m_t^x,m_t^y}(\tau)\right)\cdot\sup_{\nu}\left(xcorr_{m_f^x,m_f^y}(\nu) * \frac{1}{\sigma_f\sqrt{2\pi}}e^{-\frac{\nu^2}{2\sigma_f^2}}\right), \tag{6}$$

$$:= Xmt_{x,y}\cdot Xmf_{x,y} \tag{7}$$

where σ_f is the variance of the Gaussian. Finally, a cut-off level $L_{\text{cut-off}}$ is determined to help us assign a binary classification for the syllable pair such that:

$$\begin{cases} SPCC_{x,y} \leq L_{\text{cut-off}} \implies x \text{ and } y \text{ are deemed dissimilar.} \\ SPCC_{x,y} > L_{\text{cut-off}} \implies x \text{ and } y \text{ are deemed similar.} \end{cases}$$

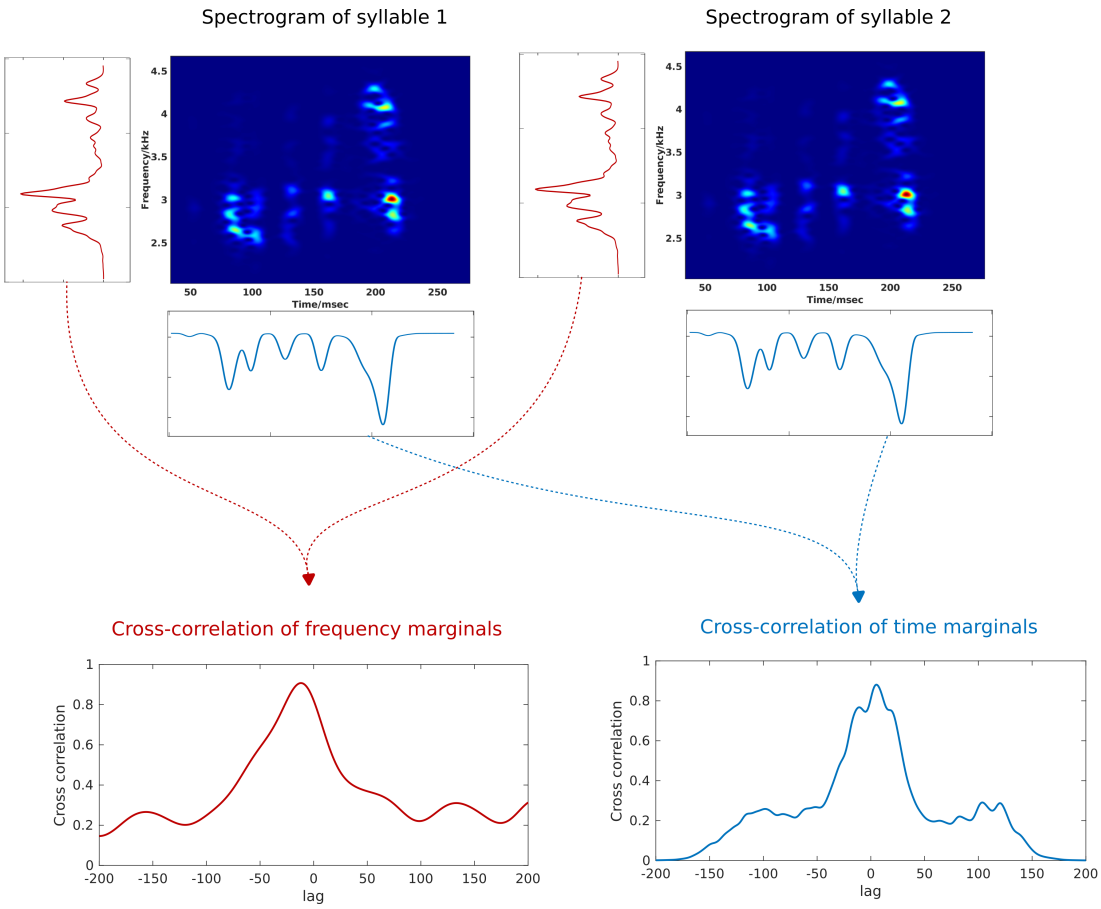


Figure 2: Visualization of SPCC Similarity Measure Computation. This diagram demonstrate the

2.3 The wavelet transform

The idea behind the SPCC method in not limited to the spectrogram but rather can be applied to any time-frequency representation of the signal such as the scalogram which is computed analogously to the spectrogram only with the continuous wavelet transfrom instead of the short time fourier transform. The continuous wavelet transform is a decomposition of the signal into basis of wavelets that are localized in time and frequency.

The continuous wavelet transform decomposes a signal into a set of basis functions known as wavelets, which are localized both in time and frequency. This decomposition allows for a detailed examination of signal features at various scales and positions. The concept of wavelets and their applications were extensively developed by the French mathematician Yves Meyer, who demonstrated

the fundamental principles underlying wavelet theory.

In wavelet analysis, a 'mother wavelet' function, denoted by $\psi(t)$, serves as the basis for generating a family of wavelets through translation and scaling operations. Specifically, an orthogonal basis is constructed by scaling and translating the mother wavelet, yielding basis functions $\psi_{a,b}(t)$ defined as:

$$\psi_{a,b}(t) = \frac{1}{|a|^{1/2}}\psi\left(\frac{t-b}{a}\right)$$

Here, 'a' represents the scale parameter, controlling the frequency content of the wavelet, while 'b' denotes the translation parameter, determining the temporal position of the wavelet. Given a signal $x(t)$, the wavelet coefficients in these basis functions are computed as:

$$X(a,b) = \int x(t)\psi^*_{a,b}(t)dt$$

Finally, the scalogram, denoted by $S_x(a,b)$, is obtained by computing the squared magnitude of the wavelet coefficients:

$$S_x(a,b) = |X(a,b)|^2$$

The scalogram provides a detailed representation of signal characteristics across different scales and positions, offering valuable insights into localized features and patterns within the signal. To construct a similarity measure from the scalogram, we can as before calculate the time and frequency marginals of two given signals and calculate the cross-correlations in identical ways to the SPCC as presented in figure 2. We will call this method the wavelet transform cross-correlation or simply WTCC.

2.4 Feature Extraction Using SVD

This method is based on a factorization method called singular value decomposition, which decomposes a matrix into $A = U\Sigma V^T$, where U and V are unitary matrices and Σ is a diagonal matrix, where the diagonal entries, called the singular values, of A are unique. One important note is that although the singular values for matrix A are uniquely determined, that is not the case for U and V. The number of non-zero entries in Σ is equal to the rank of A. If the SVD of A is sorted so that the singular values come in descending order, the columns of U and V form sets of orthonormal bases and the matrix can thus be written as

$$A = \sum_k^r \sigma_k u_k v_k^* = \sum_k \sigma_k A_k,$$

where σ_k is a measure of how much the matrix A_k contributes to the original matrix A. In this method we can either decompose the spectrogram image matrix, or the scalogram image matrix, which will be denoted as previously

by $A(i,j)$. Since we want to treat the matrices $A_k(i,j)$ as probability distributions, they must be normalized $\sum_i \sum_j \tilde{A}_k(i,j) = 1$, which is achieved through $\tilde{A}_k(i,j) = u_k(i)^2 v_k(j)^2$, since the vectors form an orthonormal base. The normalized vectors $u_k(i)^2$ and $v_k(j)^2$ are in this case the time and frequency marginals and since all matrices A_k are rank one, the marginals contain the same information as the matrix A_k. One approach of comparing two syllables could be to calculate the cross correlations of the marginals of all the matrices A_k, but this method would be ineffective due to the large number of A_k. Instead, the singular values together with the mean and the variance of the marginals are calculated and stored for the matrices A_k which contribute most to the overall spectrogram or scalogram image A. These features are then stored for each syllable and compared to the features of another syllable using the Hausdorff distance. The Hausdorff distance measures how for two sets are from each other, and for sets X and Y is defined as

$$d_H(X,Y) = \max\left\{ \sup_{x\in X} d(x,Y), \sup_{y\in Y} d(X,y) \right\}, \tag{8}$$

where $d(a,B) = \inf_{b\in B} d(a,b)$.

To exemplify how this feature extraction method works, assume we investigate the spectrogram image of a syllable, A, which is decomposed using the SVD. In this method we do not want to investigate all submatrices, A_k, that the spectrogram is composed of, but only the ones which capture the main features. To do this, the largest singular value, σ_0, is found, and only submatrices with corresponding singular values $\sigma_k > \sigma_0/20$ are investigated. The matrices with smaller singular values contain less important information and are treated as noise in this case. Next, the normalized matrices $\tilde{A}_k$ and the time and frequency marginals are found, as well as the mean values and variances for the marginals. This procedure is then repeated for all the k submatrices which were deemed to contain important information and the values are stored:

$$\begin{bmatrix} \sigma_0, \bar{m}_t^0, \mathrm{var}(m_t^0), \bar{m}_f^0, \mathrm{var}(m_f^0) \\ \sigma_1, \bar{m}_t^1, \mathrm{var}(m_t^1), \bar{m}_f^1, \mathrm{var}(m_f^1) \\ \vdots \\ \sigma_k, \bar{m}_t^k, \mathrm{var}(m_t^k), \bar{m}_f^k, \mathrm{var}(m_f^k). \end{bmatrix}$$

This is also done for the next syllable, and the two feature matrices are compared, where each row corresponds to a point. The distance between the two sets (syllable features) is then calculated using the Hausdorff distance.

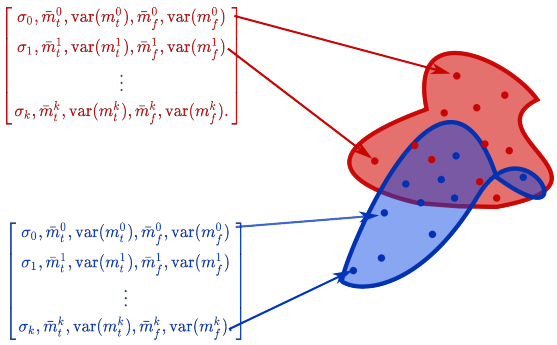


Figure 3: Two feature matrices represented as sets of points. The similarity is determined by calculating the Hausdorff distance between them.

The number of significant matrices A_k will vary between syllables and due to the SVD not being unique, the same syllable might have different number of significant matrices. However, since we extract important features, we still expect the two sets of points to be close, even if the number of points is different.

3 Results

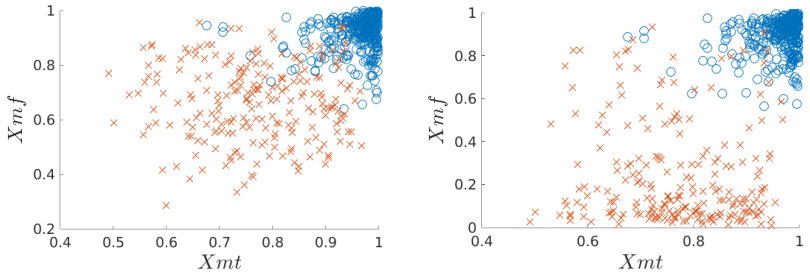
3.1 SPCC

The SPCC similarity measure was computed for each syllable pair in the song "13B04" to facilitate parameter tuning. Figure 4 illustrates scatter plots depicting the temporal and spectral similarity measures, represented as points (Xmt, Xmf), where Xmt and Xmf denote the maximum cross-correlation in time and frequency marginals, as defined in equation (7). In these plots, syllable pairs classified as similar (true label = 1) are denoted by blue circles, while dissimilar pairs (true label = 0) are indicated by red crosses.

These figures illustrate the separability of the syllable pairs into similar and dissimilar. Similar pairs are concentrated at high marginal frequency and time cross-correlations. However, figure (a) shows significant mixing between the two regions when the variance of the Gaussian window is chosen to be wide ($\sigma^2 = 100$ lag). This corresponds to the case where we allow the frequency marginals to shift to find the best alignment between the two syllables. Figure 4(b) demonstrates the influence of the superposed Gaussian window on the SPCC results. In this case, the variance of the Gaussian was chosen to be $\sigma^2 = 20$ lag, allowing for frequency adjustments. The result, as seen in figure 4(b), is a further separation of the two regions of syllable pairs into regions of similarity.

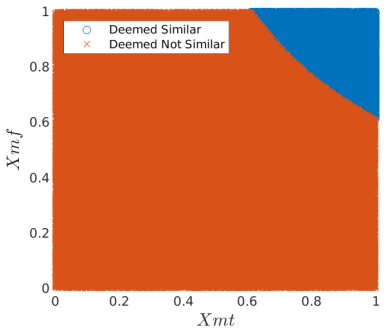
Figure 5(a) provides a visual of how the SPCC is used to produce a binary classification. It shows the sets $\{(mt, mf)\text{where } SPCC > L\}$ and its complement set.

The tuning of the cut-off level is depicted in figure 5(b), where we plot the success rate in classifying similar and dissimilar syllables versus the used cut-off level. $L^* = 0.62$ was found to be the optimal cut-off level, yielding a success rate of 97% on the training set (song "13B04").

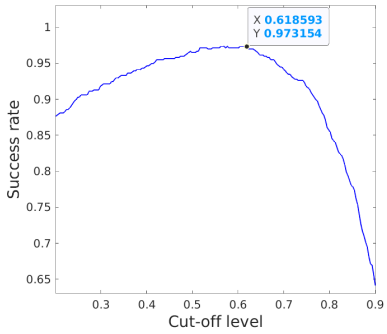


((a)) $\sigma^2 = 150$ ((b)) $\sigma^2 = 20$

Figure 4: Scatter plot of the SPCC outcomes for syllable pairs. Each point represent a syllable pair with the x- and the y- coordinates being the maximum cross-correlation between the two syllable's spectrogram marginals in time on the x-axis and in frequency on the y-axis. In frequency the marginal cross-correlation is multiplied with a gaussian window centered at zero lag and has variance of 100 in figure (a) and 20 in figure (b). Syllables pairs that have true label of 1 (similar) are plotted as a blue "o" while dissimilar syllabels are plotted as a red "x".



((a)) The SPCC algorithm classifies pairs of syllables as similar or dis-similar by dividing the plane $\{(Xmt, Xmf)\}$ into two regions. Xmt andXmf are the maximum cross-correlation of the time and frequency marginals generated by the syllable pairs. The blue region corresponds to pairs that have an SPCC measure of 0.62 or higher. These pairs are deemed similar. Pairs falling into the red region are deemed different.



((b)) Tuning of the cut-off parameter of the SPCC similarity measure. The highest success rate obtained on song "13B04" is 97% at a cut-off level of 0.62. That is, syllable pairs that get an SPCC measure higher than 0.62 are deemed similar while pairs that get a measure less than 0.62 are considered different.

The SPCC was then applied to the rest of the dataset, and the success rates were recorded in table 2. The SPCC performs almost identically on all songs, with an average success rate of 96.7

Table 2: SPCC performance on the entire data set comprised of six bird songs.

Song	13B04	19A04	23B05	14A03	19A01	23A01	Total
Success rate	97.15%	97.07%	96.90%	96.81%	94.70%	97.26%	96.71%
Number of evaluations	596	409	419	502	377	401	2704

To further verify the sensibility of the SPCC measure, we plotted some representative misclassified pairs. Figure 6 showcases the first type of errors that the SPCC results in. Syllable pairs with SPCC measures close to the cut-off level 0.62 are typically difficult to classify. As shown in Figure 6, the syllable pair can exhibit a great deal of similarity in one part, followed by some deviations in another part, making them intrinsically difficult to classify in a binary way. Although this pair received a similarity measure of $0.67 > 0.62$ deeming them similar, they were labeled as dissimilar.

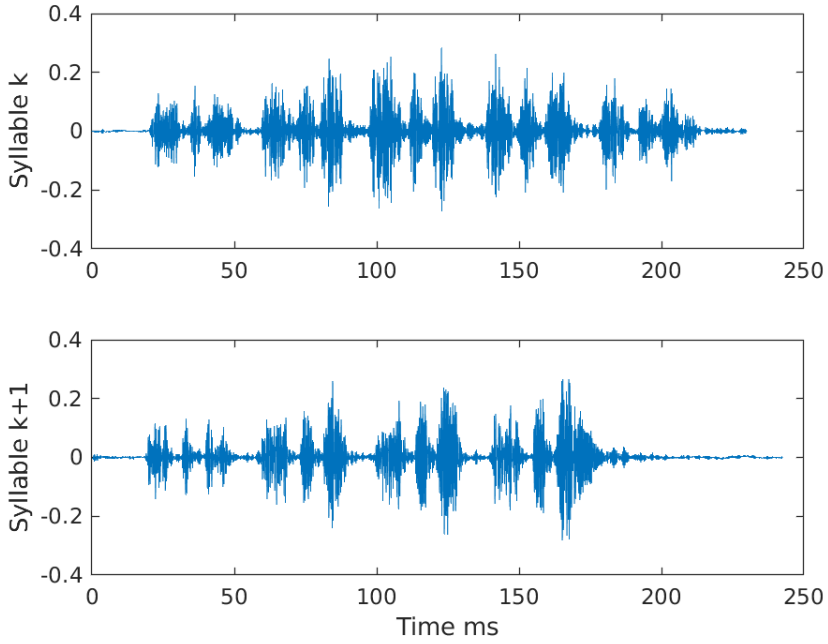


Figure 6: A syllable pair that was misclassified by the SPCC measure. This pair is an example of syllable pairs that are difficult to classify because they show strong similarity but still deviate in different aspects or parts of the syllables.

Another source of misclassifications arises from human error in the labeling of syllables, as illustrated in figure 7. This figure shows a pair of syllables that display strong similarity but were labeled as different. The SPCC, with a measure of $0.70 > 0.62$, agrees with the eye and deems the pair as similar.

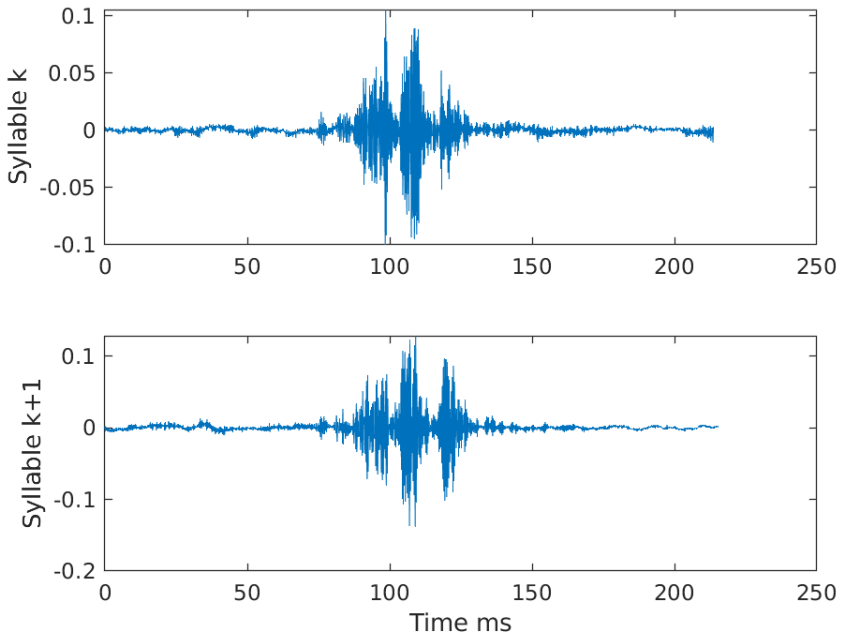
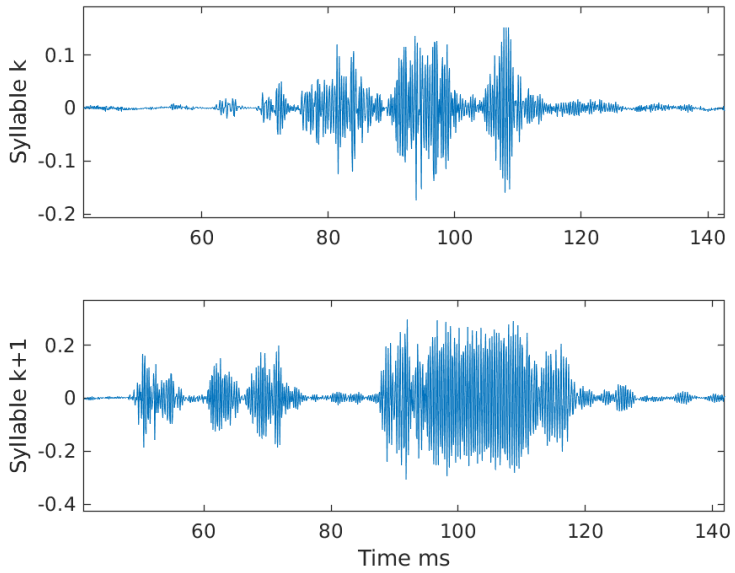


Figure 7: A syllable pair that was classified by the SPCC as similar with score 0.7 ¿ 0.62. However, this pair is labeled as dissimilar. This discrepancy is likely due to human error in labeling the data since the syllables indeed seem very similar.

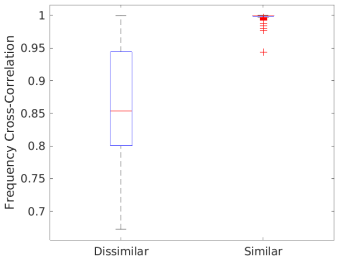
The final source of misclassifications discussed is due to the construction of the SPCC. Although cross-correlation in time and frequency is a good proxy for similarity most of the time, there are instances where it is not. Further analysis, as shown in Figure 8(a), reveals a pair that received an SPCC similarity rating of 0.72, deeming them similar. However, despite showing small similarities, these syllables are different.



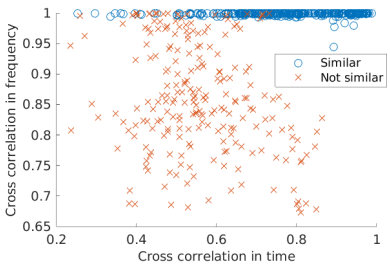
((a)) A syllable pair that was missclassified by the SPCC measure as similar. However, the pair is clearly not identical.

3.2 Scalogram cross correlation

The WTCC (Wavelet Transform Cross-Correlation) was applied to all pairs in the training set (song "13B04") using the Morlet wavelet. The results are visualized in Figures 9(b) and 9(a), where similarity measures in time and frequency are represented as points (Xmt, Xmf). Similar pairs are depicted as blue circles, while dissimilar pairs are indicated by red crosses. A notable pattern emerges, with similar pairs clustering at high cross-correlations in frequency, while time cross-correlations appear to have less influence on determining similarity.



((a)) This plot shows box plots of the distribution of similar and dissimilar syllable pairs with respect to the maximum cross-correlation in frequency marginals. This figure shows that the similar pairs cluster at high cross-correlations while dissimialr ones seem to have lower cross-correlation.



((b)) This figure shows a scatterplot of the WTCC outcome when applied to the training set (song "13B04"). Each syllable pair gives a cross-correlation in time shown on the x-axis and in frequency as shown on the y-axis. Similar pairs are plotted as blue "o" and dissimilar pairs are plotted as red "x". This figure shows that the similar pairs cluster at high cross-correlations while dissimialr ones seem to have lower cross-correlation.

To investigate why the cross-correlation in time for two similar syllables may not be as high as observed in the SPCC method, we selected a similar pair and examined their scalogram time marginals, as shown in Figures 10. These figures reveal intriguing insights into our analysis. Firstly, the temporal representation of the signal exhibits high resolution, allowing for detailed examination. Secondly, the two syllables appear to be aligned and highly correlated in the initial part, yet exhibit slight desynchronization in the latter part, resulting in decreased cross-correlation. This underscores the challenge of determining similarity, as syllables may exhibit differences while still being considered similar. Natural syllables often consist of sub-syllabic structures, and minor deviations, such as delays between these structures, should not significantly alter perceived similarity. However, in our current WTCC construction, such deviations lead to substantial portions of syllables being out of sync during comparison due to the high resolution of the wavelet transform.

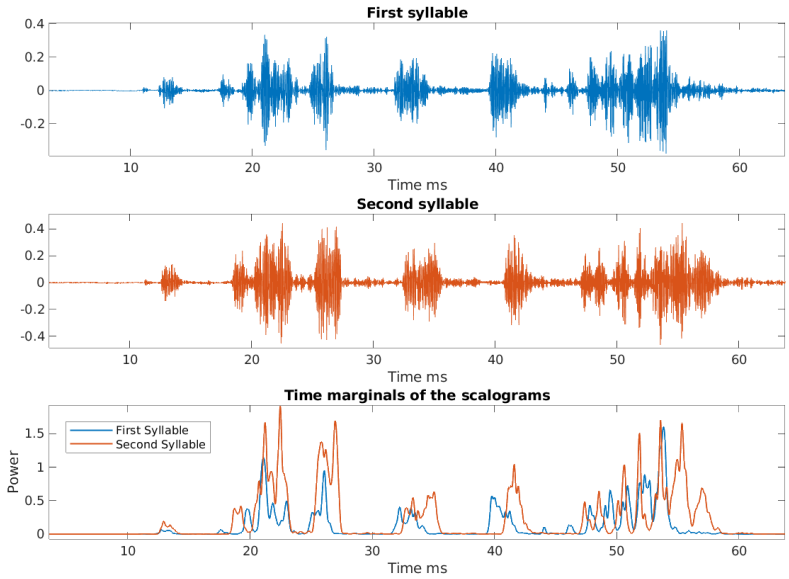


Figure 10: Illustrates how small time delays in sub-syllabic structures can lead to reduced cross-correlations in the time marginals when they are produced by a high resolution wavelet transform.

In addressing this challenge, we propose alternative solutions. One approach is to smooth the temporal marginals to mitigate small deviations. Alternatively, the WTCC method can solely rely on frequency marginals, ignoring the time aspect, as expressed by the inequality:

$$Xmf_{x,y} > L \implies \text{The pair } x, y \text{ are deemed similar}$$

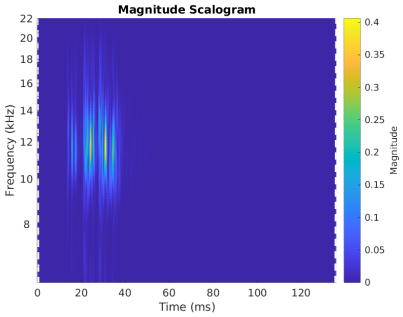
The optimal cut-off level L was determined using song "13B04" to maximize the success rate of the WTCC, giving $L^* = 0.98$. Subsequently, we evaluated all songs and recorded success rates in Table 3.

Table 3: WTCC performance on the entire data set comprised of six bird songs.

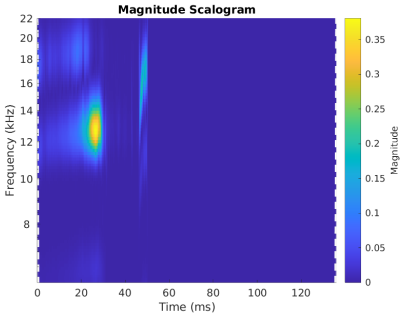
Song	13B04	19A04	23B05	14A03	19A01	23A01	Total
Success rate	0.9547	0.9609	0.9260	0.9422	0.9390	0.9352	
Number of evaluations	596	409	419	502	377	401	2704

Despite our method neglecting temporal information, we observed surprisingly high success rates with WTCC. However, missclassified pairs exhibit a distinct

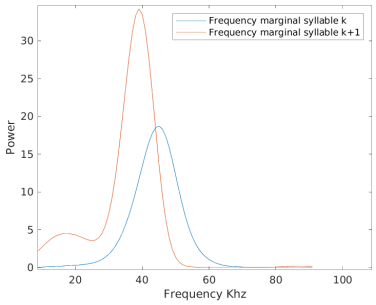
pattern, primarily comprising syllables sharing similar frequency profiles but differing substantially in time cross-correlations. An example of this phenomenon is illustrated in Figures 12(a), 11(a) and 11(a). These syllables display significant differences in their scalograms, yet exhibit similar frequency marginals, resulting in strong cross-correlation.



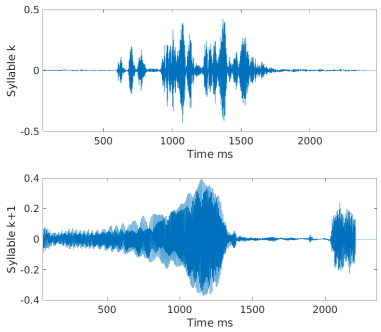
((a)) The scalogram of the first syllable.



((b)) The scalogram of the second syllable.



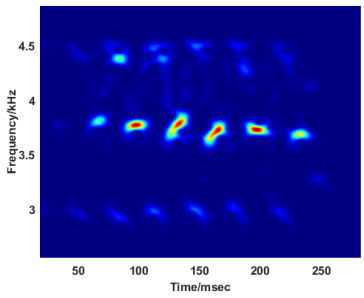
((a)) The frequency marginals of the syllable pair being compared. It shows similar frequency profile for the two syllables. This will result in strong cross-correlation between the pair even though they have different temporal structure.



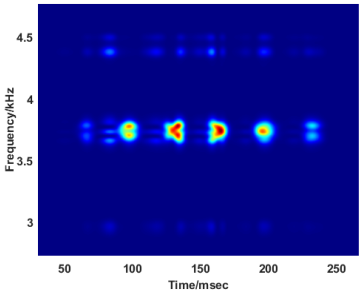
((b)) An example of a syllable pair that the WTCC can classify wrongly as similar. The reason for the missclassification is because the syllable pair have nearly identical frequency profile.

3.3 SVD

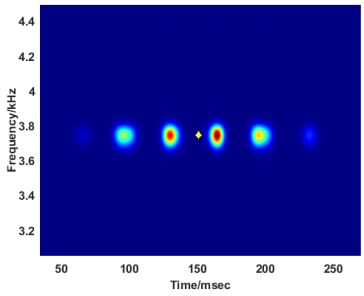
Finally, the SVD method was applied to the first song 13B05 for training. An example of the a syllable spectrogram, its four most important submatrices (the ones with the highest singular values) as well as their weighted sum can be seen in Figure 13.



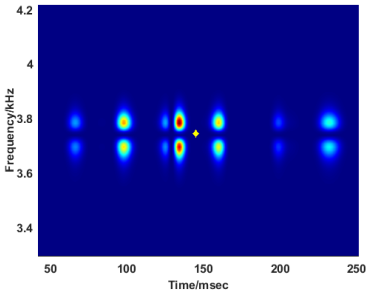
((a)) Spectrogram of a syllable.



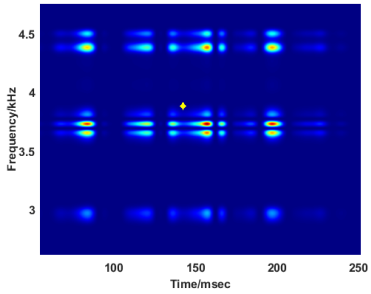
((b)) The weighted sum of the first four most important submatrices.



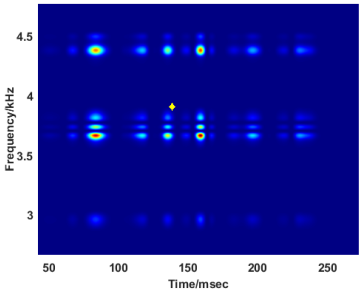
((c)) The submatrix with the highest singular value.



((d)) The submatrix with the second highest singular value.



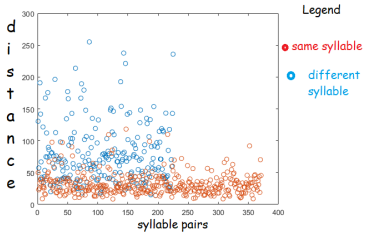
((e)) The submatrix with the third highest singular value.



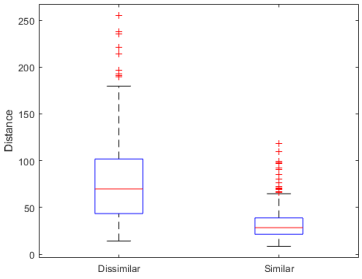
((f)) The submatrix with the fourth highest singular value.

Figure 13: Spectrogram and the submatrices generated by the SVD, together with their sum. The yellow dot which is present in subfigures 13(c) - 13(f) show the mean in time and frequency.

The Hausdorff distance between the features taken from the submatrices of all pairs of syllables are displayed in Figure 14(a), together with boxplots in Figure 14(b).



((a)) The Hausdorff distance of the different syllable pairs. Red color indicates that the pairs are similar, while blue color indicated dissimilarity between them.



((b)) Box plots of the distributions of the Hausdorff distances for dissimilar and similar syllable pairs.

We can see that the distance of similar pairs is on average smaller, but there is a very large variance in the Hausdorff distances of the dissimilar syllable pairs. If two syllables have different spectrograms, they could still "look" similar according to this measure if the mean values or the variances are similar, which could be the case for some syllables. Another thing that could affect the variance is the way the Hausdorff distance is calculated. In this method, the Hausdorff distance is set according to equation 8, which makes it sensitive to outliers which are very "far away". This could be rectified by calculating the average of the distances instead of the supremum. This improved distance measure was also tested, however, it didn't improve the success rate of the SVD method significantly.
Having found the Hausdorff distance, one has to decide on a threshold below and above which the syllable pairs will be classified as similar or dissimilar respectively. This threshold was tuned for the first song, where the suggested classifications for syllable pairs were compared to the predetermined labels. The results are displayed in Figure 15.

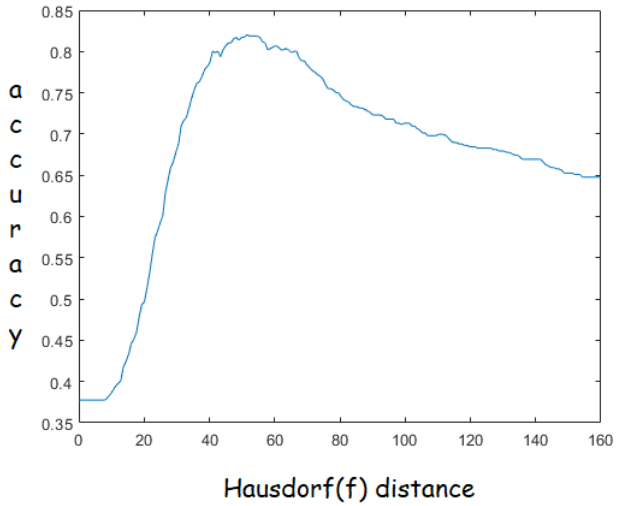


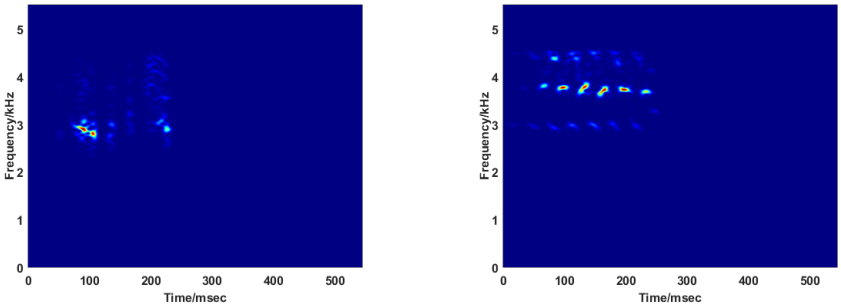
Figure 15: The accuracy of the SVD method as a function of the threshold Hausdorff distance.

From this tuning, the threshold was set to 63, which yielded the maximum accuracy of $\approx$ 80% for the first song. The SVD method together with this threshold was then used to classify the syllable pairs for the rest of the bird songs. The results are displayed in Table 4.

Table 4: Accuaracies for the SVD method

Song	13B05	19A04	23B05	14A03	19A01	23A01	Total
Success rate	0.8037	0.7995	0.8138	0.7470	0.7507	0.8329	
Number of evaluations	596	409	419	502	377	401	2704

From these results we see that the accuracies are much lower than for the other two methods, with an average of $\approx$ 79% across all evaluations. It seems that just calculating the mean and variance in time and frequency and having the singular value for the submatrices is not a good enough measure for determining similarity, since a lot of information from the sprectrogram is disregarded. One could perhaps try to include more submatrices with smaller singular values, but then there is a risk that very noisy submatrices will be included, which would add to the variance. One could also try to extract more information from the spectrograms other than just the mean and the variance. An example of a miss classification can be seen in figures 16(a) and 16(b), which had a Hausdorff distance of only 26, but where the syllables are different.



((a)) ((b))

Figure 16: Two syllables which were miss classified as being similar, although they are not. From the two spectrograms we see that the variance in both frequency and time are very similar for the syllables, and the mean values are also close.

The mean and variance in time and frequency look similar in both spectrograms, and that also holds for all of the extracted submatrices. The feature matrices for the syllables also happen to contain similar singular values. In this case the extracted features are misleading, and having more information would be beneficial for correct classification.
Finally, it is also worth mentioning that the SVD method was by far the most computationally expensive, and so it would only be worth using it if it yielded significantly better results than the alternative methods.

References

[1] Andrén, P.: *Quantification of similarity between dickcissel song dialects*, 2020