PROBABILISTIC EXTRACTION AND DISCOVERY OF FUNDAMENTAL UNITS IN DOLPHIN WHISTLES

Daniel Kohlsdorf[⋆] *Celeste Mason*[⋆] *Denise Herzing*[†] *Thad Starner*[⋆]

[⋆] {dkohlsdorf6, celeste.m, thad}@gatech.edu, 85 5th Street NW, 30308 Atlanta, GA, US
[†] wdpdenise@wilddolphinproject.org, P.O. Box 8436, 33468 Jupiter, FL, US

ABSTRACT

The study of dolphin cognition involves intensive research of animal vocalizations. Marine mammalogists commonly study a specific sound type known as the whistle found in dolphin communication. However, one of the main problems arises from noisy underwater environments. Often waves and splash noises will partially distort the whistle making analysis or extraction difficult. Another problem is discovering fundamental units that allow research of the composition of whistles. We propose a method for whistle extraction from noisy underwater recordings using a probabilistic approach. Furthermore, we investigate discovery algorithms for fundamental units using a mixture of hidden Markov models. We evaluate our findings with a marine mammalogist on data collected in the field. Furthermore, we have evidence that our algorithms enable researchers to form hypotheses about the composition of whistles.

Index Terms— Probabilistic Modeling, Bio Acoustics, Marine Mammals, Dolphin Behavior

1. INTRODUCTION

One part of marine mammalogy is the study of dolphin cognition and communication. Communication and vocalizations of animals can give valuable insight into the social structure of groups of animals. One of the research goals in dolphin behavior research is the association of behavior with vocalizations by correlating sound with observed social clues from video. Therefore, researchers collect large video and audio databases in the field containing long-term, continuous observations of the complex social behavior of dolphins and their communications. While the physical structure and the production of dolphin vocalizations is known, there is still the potential for unknown patterns and structures in their communication. Herzing [1] provides an overview of current tools and techniques in aquatic mammal research. For example, researchers know that dolphins use frequency modulated whistles to identify each other (Figure 1).

One question is if there are fundamental units in these whistles and how these units occur in context. Currently researchers evaluate such patterns by manual measurements of magnitudes in frequency space and by visual inspection. One common problem is that whistles can be occluded by other underwater noise making extraction even harder. Despite recent attempts to automate the exploration process and develop scores of similarity, current analysis of dolphin whistles is still based on visual inspection of a spectrogram. Furthermore, to our knowledge, research on dolphin whistles focuses on clustering or detecting whole whistles. However, the composition of dolphin whistles and combinatoric analysis of these can help researchers to understand how whistles change in a group of dolphins and how parts of whistles are reused.

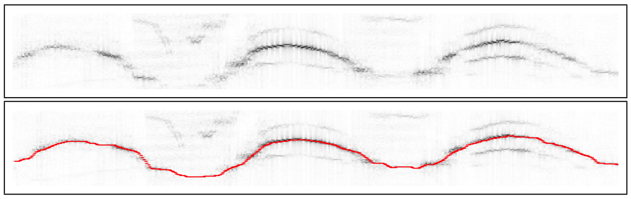

Fig. 1. Top: A dolphin whistle occluded by underwater noise. Bottom: The dotted red line shows a desired trace or contour of the whistle.

We present a system that automates parts of the exploration effort in dolphin communication using techniques from speech recognition. Our system uses a probabilistic filter to extract a whistle from a spectrogram. We also implement a novel pattern discovery algorithm based on mixtures of hidden Markov models that searches for fundamental units in a whistle database. We evaluate our findings on a database of whistles recorded from spotted dolphins in the field by marine mammalogists.

2. RELATED WORK

A survey of algorithmic methods used in detection and analysis of underwater acoustic signals by Lampert et al. [2] identified the three main areas of concentration to include image processing, neural networks, and statistical models. Their evaluation was based on comparison of methods according to a range of criteria. Those criteria applicable to dolphin whistle detection include: handling of high signal-to-noise ratios, noise variation, whistle shape variability (with no *a priori* assumption), multiple whistles, between-whistle proximity/crossing, initial/endpoints of whistles, and computational resource use and process time. They conclude that hidden Markov models are currently the most prevalent, promising methods in research literature for use in cetacean vocalization spectrum analysis.

Recently, Kershenbaum et al. [3] published a method on similarity of dolphin whistles using dynamic time warping[4]. They extract the whistles manually by following the trace of the whistle in a spectrogram using an interactive tool. We try to ease this process by extracting or tracing the whistle in the spectrogram automatically. Our algorithm is a single frequency pitch tracker similar to the formulation by Lee and Ellis [5].

Other work in classification and clustering of dolphin whistles uses neural networks [6, 7] and clustering based on hidden Markov models in a K-means framework [8]. Both approaches filter the data first and use mel-cepstral coefficients or other measurements from the spectrogram as features. Halkias and Ellis [9] extract the whistle from the spectrogram using a Bayesian approach on each frame.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

Lampert and O'Keefe use Kalman filters for contour track detection [10]. Shapiro and Wang apply pitch detection designed for human telephone speech to whale vocalizations [11]. Since most of the frequency bins in a spectrogram do not contain any information about a dolphin whistle, we use a whistle extraction approach inferring the frequency of the whistle at every time step, ignoring all other frequency bands. The result of our filtering is a single frequency changing over time.

Most discovery algorithms in the literature use distance-based approaches and approximations or discretization [12, 13, 14, 15, 16]. Our discovery algorithm is based on Smyth's and Minnen's work of learning a mixture of hidden Markov models [17, 18]. Smyth represents a mixture of hidden Markov models as a single, larger model [17] and performs estimation of the mixture's parameters using the Baum Welch algorithm [19]. The initialization procedure uses agglomerative clustering with the likelihood as a distance metric. The number of components is estimated using Monte Carlo Cross Validation on the mixture's likelihood. Minnen's greedy mixture learning algorithm [18] adds components to a mixture until the information gain does not change. The initialization is based on a neighborhood estimate using the k-nearest neighbors under Dynamic Time Warping (DTW) distance. In contrast to Smyth, the mixture is used for continuous recognition instead of pattern spotting (see Figure 2).

3. A PATTERN DISCOVERY SYSTEM FOR DOLPHIN WHISTLES

The goal of our pattern discovery system is to extract whistles automatically from a continuous recording and find small patterns shared among multiple whistles. In that way we can provide visual guidance for the whistle analysis work of marine mammalogists. An example of such a visualization is shown in Figure 4.

In the following we describe a probabilistic approach for extraction of whistles. Furthermore, we describe a clustering algorithm that is capable of discovering fundamental units in extracted dolphin whistles.

3.1. Noise-Robust Extraction of Dolphin Whistles

Given a continuous audio stream our algorithm starts by computing its spectrogram. We refer to a sample $x_t(f)$ from the spectrogram X of length T as representing the magnitude of frequency f at time t. Since a whistle is produced by a single oscillator we model a whistle as a sequence $F = f_1 ... f_T$ with f_i representing the frequency of the whistle at time t. Furthermore, dolphins create whistles by pushing air between two air sacks (so as to avoid bubble noise underwater). The physics of such a system allow the assumption that there will be no large jumps in frequency in a continuous whistle. A naive approach to extraction of a whistle could be to take the maximum frequency at every sample x_t. However, as one can see in Figure 1, ambient underwater noise as produced by waves, splashes and bubbles will lead to a distorted whistle. We account for the noise by probabilistically inferring the true, undistorted trace from noisy measurements. We model the measurements from the observed magnitudes of the spectrogram as the conditional probability $P(x_t|f_t)$. We compute the measurement probabilities by normalizing each sample from the spectrogram:

$$P(x_t|f_t) = \frac{x_t(f_t)}{\sum_{i=0}^{D} x_t(i)} \tag{1}$$

This probability model assigns higher probabilities to frequencies with higher magnitude in the spectrogram. We model the state transition probabilities using a Gaussian centered at the predicted frequency. We assume a linear motion model of the whistle such that the predicted frequency is $f_{tpredict} = v_{\hat{f}}(t-1)f_{t-1}$ where $v_{\hat{f}}(t-1)$ is the change in frequency trace between steps. In other words we assume Gaussian noise with some chosen variance σ^2 on the current position estimate: $N(f_t|v_{t-1}f_{t-1}, \sigma^2)$. Now we can compute a maximum *a posteriori* trace of the whistle. We can compute this trace given our model similarly to the Viterbi algorithm for hidden Markov model decoding [19]. We compute the likelihood $\delta_f(t)$ of frequency f belonging to the trace at time t as:

$$\delta_f(t) = P(x_t|f_t) \max_{\hat{f}=1}^{N} \delta_{\hat{f}}(t-1) N(f_t|v_{t-1}\hat{f}_{t-1}, \sigma^2) \tag{2}$$

Furthermore, we store the maximum frequency up to time t for backtracking purposes which also results in our velocity estimation:

$$\begin{aligned} ptr_f(t) &= argmax_{\hat{f}=1}^{N} \delta_{\hat{f}}(t-1) N(f_t|v_{t-1}\hat{f}_{t-1}, \sigma^2) && (3) \\ v_f(t) &= f - ptr_f(t) && (4) \end{aligned}$$

Then we backtrack from the most likely frequency at time T and follow the backtracking pointer backwards in order to extract the trace.

3.2. Segmented Clustering of Dolphin Whistles

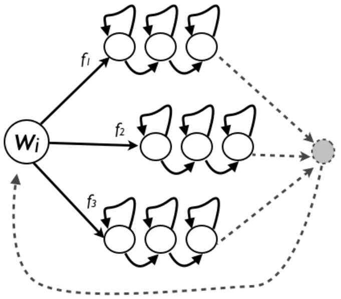

Fig. 2. A loose mixture of hidden Markov models (bold lines) normally used for word spotting and a connected mixture of Hidden Markov Models (dashed lines) used for continuous recognition.

In order to discover fundamental units, we cluster sliding windows extracted from a set of extracted whistles. We obtain such a clustering by learning a finite mixture of hidden Markov models (see Figure 2). Our algorithm is a variant of Smyth's algorithm fit to the purpose of discovering short patterns that occur in sequence. For a set of extracted whistles we run several iterations of Monte Carlo Cross Validation. For each fold we extract short sliding windows from each whistle in the training set. Then we run agglomerative clustering iteratively using the average linkage between groups of sliding windows with dynamic time warping distance as the distance measure. We decided to use dynamic time warping instead of the likelihood dissimilarity described by Smyth [17] since the patterns can be very short and there might not be enough data to fit a hidden Markov model to just one example. Every iteration reduces the number of clusters by one. For every number of clusters k greater than one and smaller than the maximum allowed k we fit one hidden Markov model per cluster using multiple iterations of Baum-Welch.

We then combine the resulting models into a loose mixture of hidden Markov models by rewriting the transition matrices $A_1 ... A_k$ into a single one with no connection between each of these matrices [17]:

$$A_{mix} = \begin{pmatrix} A_1 & & & \\ & A_2 & & \\ & & ... & \\ & & & A_k \end{pmatrix}$$

We re-estimate its parameters by multiple iterations of Baum-Welch again. We then compute a score for each model using Viterbi decoding and keep the best overall folds. The main difference between our algorithm and the one proposed by Smyth is the scoring. Since our goal is to cluster short patterns in sequence instead of complete time series, the algorithm will overfit the model. Windows containing one pattern in the beginning and a neighboring pattern in the end might be assigned to different components increasing the model size. Our new score accounts for this fact by connecting all the end states of each component to every initial state of each component resulting in a connected mixture. In this way the evaluation can be seen as connected speech recognition instead of a loose mixture as found in word spotting systems. The difference between a loose and connected mixture is visualized in Figure 2. We then evaluate the score by decoding complete whistles from the test set using the connected mixture and the token passing algorithm [20] similar to the score described by Minnen [18]. In that way the window overlap does not matter since the beginning of the window will be aligned to an end state and the end of the window to an appropriate start state. Furthermore, we penalized the resulting likelihood using the Bayesian information criterion in order to avoid overfitting in general [21]. The complete algorithm is shown in Algorithm 1.

Algorithm 1: Segmented Clustering algorithm

```
Data: A set of whistles $W$
Result: A mixture of hidden Markov models $M$
for n folds do
    Split $W$ into $W_{train}$ and $W_{test}$ using MCCV;
    Extract sliding windows $\forall w \in W_{train}$;
    $C :=$ Agglomerative Clustering($W_{train}$);
    for $i \leftarrow 1$ to $|W_{train}|$ do
        $C \leftarrow merge(C)$ based on DTW;
        if $i > 1$ AND $i <= max_k$ then
            Train $hmm_i$ from $C_i$ for $i \in 1...max_k$;
            Train mixture HMM $M$ from all $hmm_i$;
            Connect all end states of $M$ to all initial states;
            Evaluate Log-Likelihood $LL(M|W_{test})$;
            $BIC = LL(M|W_{test}) - \frac{k}{2} log(|W_{test}|)$;
        end
    end
end
return mixture $M$ with best $BIC$ score;
```

Such a mixture can be seen as a pattern code book. In order to analyze the composition of whistles we decode every whistle using our continuous mixture. Backtracking then produces a string of patterns for each whistle.

4. EXPERIMENTS

In order to ground our work we test the discovery system in two experiments. In the first experiment we explore the capabilities of our algorithm on a set of known whistles and show how we can use the algorithm's results to compare whistle similarities. In a second experiment we run our algorithm on a set of unknown whistles and evaluate our findings qualitatively with a marine mammalogist. In the end, we show common failures of our tracing algorithm and describe in which scenarios it will most likely fail.

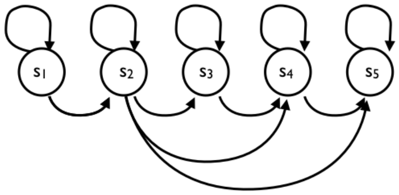

Fig. 3. A left-right hidden Markov model with 2 skip states as used in the experiments as the base model for the mixture components.

All data in our experiments consists of underwater recordings sampled at $44kHz$. We calculate all spectrograms in our experiments using the short term Fourier transform on a sliding window of $11ms$ (512 samples) with $2ms$ (128 samples) skips with a Hamming window applied. We set the variance for the tracing algorithm to $\sigma^2 = 1$. The sliding window for the discovery algorithm extracting patterns from the whistles is 80 samples long with a skip of 20 samples. The hidden Markov model topology for each mixture component is a left-right HMM with five states and two skip states.

4.1. Discovery Performance on Known Whistles

In this small experiment we evaluate the performance of our algorithm. The data set consists of 18 dolphin vocalizations of three different, known whistles. The experiment is inspired by error analysis for text input devices [22]. For every whistle we compute the strings for each vocalization using our algorithm. We then compute the Levenshtein distance [23] for each whistle's examples. In that way we get an idea of how many insertion, deletion and replacement errors we find for each whistle. These errors are important for deeper analysis of the composition of dolphin whistles since too many of these errors can lead to false claims when analyzing real world data. We found that the average Levenshtein distance is between zero and two while the length of the strings varies between three and six. Furthermore, we found that all errors occur at the beginning of the whistle.

4.2. Real World Discovery Experiment

In this experiment we explore the capabilities of our discovery algorithm in a real world setting. We use recordings from field studies of wild Atlantic spotted dolphins (Stenella frontalis). The marine mammalogists extracted the audio from video recordings with behavioral annotations. While we have no ground truth in this experiment, such annotated videos can help analyzing our discovery results. For our discovery experiment we collect 73 examples from the continuous underwater recordings. We use our whistle tracing algorithm to extract the contours of the whistles as described in Section 3.1. We then run the discovery algorithm on our whistles. Our algorithm returned a mixture with eight components. Decoding every whistle using the mixture resulted in strings of patterns with the length varying between one and eight patterns. Short strings with one or two patterns made up approximately 80% of the 73 whistles. We also found every pattern more than once across whistles. Figure 4 shows six whistles and the patterns from our experiment.

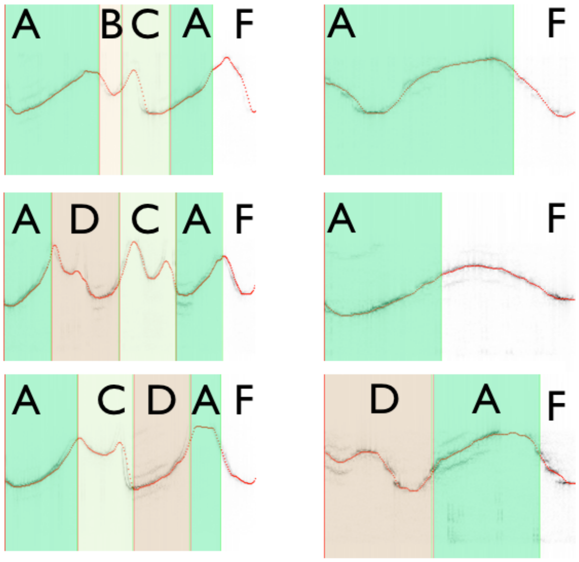

Fig. 4. Segmented dolphin whistles (best viewed in color): The dotted red line shows the extracted whistle plotted on its spectrogram (background). The highlights show the responsibility of the mixture's components obtained by continuous Viterbi decoding. The letters represent the units $\{A...F\}$ found.

We provided such images as well as the strings to the marine mammalogist for further analysis. Based on our results the researcher was able to form an hypothesis about the whistle construction. For example from the video annotations we learned that the strings "ADCAF", "ACDAF" and "ABCAF" happened during a mother-calf interaction. As one can see from the strings and from the spectrograms in Figure 4 on the right side the strings are very similar, differing by one pattern only. These could potentially lend support for the hypothesis that part of a mother's call whistle becomes part of the calf's whistle [24]. Furthermore, we found an interaction between two dolphins where we observed a specific whistle composed of only one pattern first, followed by another whistle composed of two patterns. Furthermore, we observed these whistles across multiple dolphins. Another hypothesis formed by the mammalogist is that these are generic call whistles that do not particularly name a dolphin.

4.3. Common Errors During Tracing

During our experiments we found multiple examples of noise leading to failures during tracing. Often these happen when the whistle becomes nearly indistinguishable in the spectrogram and the motion prior is much lower than the magnitude of another frequency. The obvious example is that low frequency boat noise is louder than the whistle over a longer period of time. Another example is when a harmonic is much louder than the actual whistle. Some example traces with errors are shown in Figure 5.

5. DISCUSSION

The main drawback of our system is the speed and memory used by the tracing and clustering algorithm. We conducted our experiments on a workstation with an Intel i7 processor and $16GB$ of RAM and the implementation is based on Java. When running the system from whistle detection to producing visual results, we observe memory usages of up to $5GB$ with run times of up to one hour. However, the clustering speed can be improved using a lower bounding approximation on the dtw distance [25]. Furthermore, re-estimation speed and memory usage can be improved using Viterbi training or even faster approximate inference methods such as variational inference. Furthermore, we observed some unintuitive groupings of patterns, however these were rare and can be identified visually by a researcher. Most of such errors occur for badly traced whistles. Another drawback of our method is that the size of the sliding window still influences the boundaries of our results. During our experiment we found that the algorithm is able to produce mixtures that are capable of recognizing its components across multiple instances of a whistle.

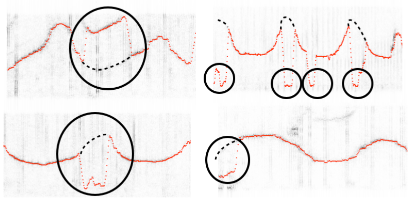

Fig. 5. Some failures observed during tracing. Circles indicate error regions. The dashed lines follow the hypothesized actual trace.

6. FUTURE WORK

Our experiments gave some technical insights into the performance of our algorithm on data collected in the field and showed a potential for the researchers to form an hypothesis about dolphin whistle composition and their meaning in a social context. However, we have to apply our methods to larger data sets in the future. Therefore, we have to improve the run-time of our algorithm using techniques described in Section 5. We also plan to investigate other dolphin vocalizations. The quantity of dolphin whistles observed in the wild is fairly low compared to other sounds such as dolphin burst pulses. However, these sounds have different characteristics than whistles and need other algorithms and assumptions for analysis tools. For example, burst pulses appear as packages of broadband clicks. We plan to develop methods to perform a similar analysis for other dolphin vocalizations. Furthermore, we will work on a method that is using multiple sizes of sliding windows.

7. CONCLUSION

We presented a novel system for whistle extraction using a probabilistic algorithm and a novel pattern detector algorithm. Instead of clustering whistles, the algorithm is made for discovery of smaller, fundamental units in whistles. This approach allows researchers to not only study the similarity of whistles but also the composition of whistles. In an experiment with a marine mammalogist, we showed how the fundamental units and their composition can help to form hypothesis about social interactions of dolphins. Furthermore, we evaluated our algorithm on three known whistles and pointed out typical failures of our tracing algorithm.

8. REFERENCES

[1] Denise Herzing, "The currency of cognition: Assessing tools, techniques, and media for complex behaviour analysis," *Aquatic Mammals*, 2006.

[2] Thomas A. Lampert and Simon E.M. OKeefe, "A survey of spectrogram track detection algorithms," *Applied Acoustics*, vol. 71, no. 2, pp. 87 – 100, 2010.

[3] Arik Kershenbaum, Laela S. Sayigh, and Vincent M. Janik, "The encoding of individual identity in dolphin signature whistles: How much information is needed?," *PLoS ONE*, vol. 8, 10 2013.

[4] M.K. Brown and L. Rabiner, "An adaptive, ordered, graph search technique for dynamic time warping for isolated word recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 30, no. 4, pp. 535–544, 1982.

[5] BS Lee and DPW Ellis, "Noise Robust Pitch Tracking by Subband Autocorrelation Classification," *Interspeech 2012*, 2012.

[6] Mahdi Esfahanian, Hanqi Zhuang, and Nurgun Erdol, "On contour-based classification of dolphin whistles by type," *Applied Acoustics*, vol. 76, no. 0, 2014.

[7] V. B. Deecke, J. K. B. Ford, and P. Spong, "Quantifying complex patterns of bioacoustic variation: Use of a neural network to compare killer whale (*Orcinus orca*) dialects," *Journal of the Acoustical Society of America*, vol. 105, no. 4, pp. 2499–2507, 1999.

[8] K. Adi, K.E. Sonstrom, P.M. Scheifele, and M.T. Johnson, "Unsupervised validity measures for vocalization clustering," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4377–4380.

[9] Xanadu Halkias and Dan Ellis, "Call detection and extraction using Bayesian inference," *Applied Acoustics*, 2006.

[10] Thomas A. Lampert and Simon E.M. O'Keefe, "An active contour algorithm for spectrogram track detection," *Pattern Recognition Letters*, vol. 31, no. 10, pp. 1201 – 1206, 2010.

[11] Ari D. Shapiro and Chao Wang, "A versatile pitch tracking algorithm: From human speech to killer whale vocalizations," *The Journal of the Acoustical Society of America*, vol. 126, no. 1, pp. 451–459, 2009.

[12] David Minnen, Charles L. Isbell, Irfan Essa, and Thad Starner, "Detecting subdimensional motifs: An efficient algorithm for generalized multivariate pattern discovery.," in *ICDM*. 2007, pp. 601–606, IEEE Computer Society.

[13] Bill Chiu, Eamonn Keogh, and Stefano Lonardi, "Probabilistic discovery of time series motifs," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, 2003, KDD '03, pp. 493–498, ACM.

[14] Abdullah Mueen, Eamonn J. Keogh, Qiang Zhu, Sydney Cash, and Brandon Westover, "Exact Discovery of Time Series Motifs," in *SIAM International Conference on Data Mining*. American Statistical Association (ASA), 2009.

[15] Yuan Li and Jessica Lin, "Approximate variable-length time series motif discovery using grammar inference," in *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, New York, NY, USA, 2010, MDMKDD '10, pp. 10:1–10:9, ACM.

[16] David Minnen, *Unsupervised Discovery of Activity Primitives from Multivariate Sensor Data*, Ph.D. thesis, Georgia Institute of Technology, 2008.

[17] Padhraic Smyth, "Clustering sequences with hidden Markov models," in *Advances in Neural Information Processing Systems*. 1997, pp. 648–654, MIT Press.

[18] David Minnen, Charles L. Isbell, Irfan Essa, and Thad Starner, "Discovering multivariate motifs using subsequence density estimation," in *In AAAI Conf. on Artificial Intelligence*, 2007.

[19] Lawrence R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in *Proceedings of the IEEE*, 1989, pp. 257–286.

[20] S.J. Young, N.H. Russell, and J.H.S Thornton, "Token passing: a simple conceptual model for connected speech recognition systems," Tech. Rep., Cambridge University Engineering Department, 1989.

[21] Gideon Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, 1978.

[22] Fred J. Damerau, "A technique for computer detection and correction of spelling errors," *Commun. ACM*, vol. 7, no. 3, Mar. 1964.

[23] V.I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Soviet Physics Doklady*, vol. 10, pp. 707, 1966.

[24] Stephanie L. King, Laela S. Sayigh, Randall S. Wells, Wendi Fellner, and Vincent M. Janik, "Vocal copying of individually distinctive signature whistles in bottlenose dolphins," *Proceedings of the Royal Society B: Biological Sciences*, vol. 280, no. 1757, 2013.

[25] Thanawin Rakthanmanon, Bilson Campana, Abdullah Mueen, Gustavo Batista, Brandon Westover, Qiang Zhu, Jesin Zakaria, and Eamonn Keogh, "Searching and mining trillions of time series subsequences under dynamic time warping," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*. 2012, KDD '12, pp. 262–270, ACM.