AUTOMATIC PARAMETER ESTIMATION AND DETECTION OF SAIMAA RINGED SEAL KNOCKING VOCALIZATIONS

Adrià Solana[1], *Ludwig Houegnigan*[1], *Climent Nadeu*[1], *Mairi Young*[2], *Mervi Kunnasranta*[2,3]

[1] Polytechnic University of Catalonia, Signal Theory and Communications Dept. Barcelona, Spain
adria.solana@estudiantat.upc.edu, ludwig.houegnigan@upc.edu, climent.nadeu@upc.edu
[2] University of Eastern Finland
[3] Natural Resources Institute Finland

ABSTRACT

The Saimaa ringed seal (*Pusa hispida saimensis*) is an endangered subspecies of ringed seal that inhabits Finland's Lake Saimaa. Many efforts have been put into studying their ecology; however, these initiatives heavily rely on human intervention, making them costly. This study first analyzes an extensive dataset of acoustic recordings from Lake Saimaa with a focus on "knocking" vocalizations, the most commonly found Saimaa ringed seal call type. Then, the dataset is used to train and test a binary deep learning classification system to detect these vocalizations. Out of the 8996 annotated knocking events, the model is trained and tuned with 8096 samples and tested with the remaining 900 events. The system achieves a 97% F1-Score in the test set, demonstrating its capacity to identify knocking segments from noise and other events.

***Index Terms*—** Convolutional Neural Network, Support Vector Machine, fundamental frequency, spectrogram

1. INTRODUCTION

For the last decades, the Saimaa ringed seal population has been critically low, from nearly a hundred individuals in 1984 [1] to over 400 individuals [2]. Even though the population has increased in the last years, this growth may not be enough to overcome sudden changes and threats, mainly related to human activity and climate change [3], [4], [5]. In order to understand these impacts, it is important that the population is carefully monitored.

In light of the above, many efforts have been dedicated to the research and conservation of these seals. There are some commonly used methods such as photo-ID [6] and snow lair censuses [7] that aim to estimate breeding success and overall population, as well as identify known individuals. However, just detecting the individuals is a costly task that heavily relies on humans for many of the currently followed approaches. Although recent literature offers automated systems based on image identification [8], [9], [10], automated acoustic systems remain unexplored for this subspecies. The literature shows that automated systems have been used for other pinnipeds [11], [12], and some of these strategies exploit recent advances in Machine Learning, but no such automated developments have been made for Saimaa ringed seals. Our proposed solution tackles this by exploring an automatic method to detect vocalizations in audio recordings.

The study utilized a detection by classification approach to determine whether an audio recording contained knocking calls, a common vocalization among ringed seals [13], [14], [15]. While the focus was on optimizing the classification of audio segments, a complete detection mask could be easily obtained by retrieving the onsets and offsets of knocking segments. Knocking vocalizations consist of repetitions of short pulses with a consistent fundamental frequency (Figure 1), which we aimed to estimate in this study along with other characteristics. While past research investigated the periodicity of knocking calls [15] and relevant characteristics of the same calls uttered by other subspecies of ringed seals [13], an automated estimation of those parameters is also proposed here. Even though some strategies were investigated to increase the robustness of these estimates, no ground truth was available to assess them.

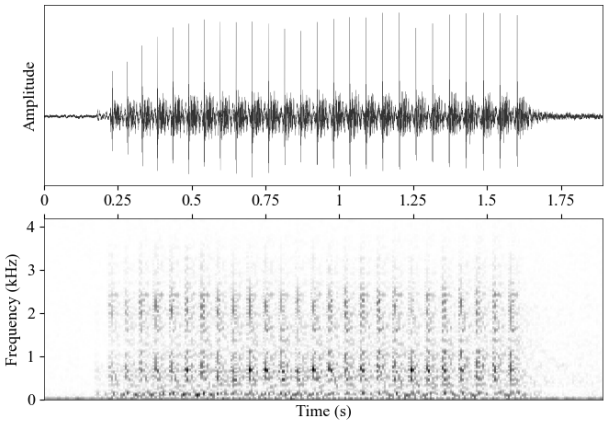

Figure 1: Knocking signal in the time domain (top) and associated spectrogram (bottom).

The classification method is based on a Convolutional Neural Network (CNN) that was able to achieve a remarkably high classification accuracy. Additionally, other systems were explored, such as predictors working in the time domain or classifiers based on the features obtained by estimating the fundamental frequencies of the audio segments. Our investigation shares common practices with recent works that explore detecting and classifying audio events in polyphonic environments using deep learning, namely [16], [17] and [18].

Finally, this study has two goals: contributing to the knowledge we have on Saimaa ringed seals' underwater vocalization repertoire, especially from a signal processing point of view, and providing a reliable method to simplify the task of researchers that work on de-

tecting the presence of these seals from sound recordings. With these objectives in mind, we aim to support the conservation of Saimaa ringed seals and any other marine mammal sharing calls of similar features.

2. METHODOLOGY

2.1. Dataset definition

The initial dataset consisted of sound recordings from Lake Saimaa, gathered and annotated by the University of Eastern Finland (UEF) over 23 weeks, for a total of 3678 hours. The annotations included the onsets and offsets of significant audio events from both seals and external agents, totaling 9717 events. The majority of the annotations corresponded to knocking calls, a smaller portion of them were sounds of seals scraping the ice to maintain breathing holes. The rest of them were human-produced sounds or events with unknown origin. All the sound files were recorded with a sampling frequency of 96 kHz. However, after analyzing the bandwidth of several knockings, the files were downsampled to 19.2 kHz to safely reduce the overall computational cost.

The knocking annotations did not necessarily span the exact duration of the periodic part of the event, as some of them included some margins to the left and right. Furthermore, the power of the knockings ranged from -88 dB to -42 dB, with a mean of -76 dB. In general, the high power knockings had a more distinct periodicity, as the low power knockings lacked resolution. Finally, their duration averaged 1.78 seconds, with a minimum of 0.54 seconds and rarely exceeding 2 seconds. Considering the looser bounds in the annotations, this is reasonably consistent with [15], which states that the duration of the studied knocks could be found in a range from 0.216 to 1.73 seconds.

2.2. Fundamental frequency estimation

To estimate the fundamental frequency in an automated, unsupervised manner, we implemented several widely used time domain-based estimators in an effort to have more robust estimates. We used the autocorrelation function, the average magnitude difference function (AMDF) and the YIN fundamental frequency estimator [19] to find extrema located at a period distance. In addition, the same techniques were used but preceded by a prefiltering step with the non-linear Teager-Kaiser energy operator (1), in order to achieve more distinctive extrema, especially in noisier or lower amplitude knockings.

$$\Phi[x[n]] = x^2[n] - x[n-1]x[n+1] \tag{1}$$

The seals studied in [15] exhibited knocking vocalizations with different characteristics. However, the fundamental frequencies of all knocking vocalizations could be found in the range of 10 Hz – 30 Hz, as estimated from the mean duration between knocking pulses. Following these guidelines, we implemented the algorithms to find fundamental frequencies in the range of 8 Hz – 60 Hz. To implement YIN, we used the code from [20], and a harmonic threshold of 0.6 was set based on heuristics. Moreover, as the annotations did not fit tightly the knocking events, a workaround had to be implemented to avoid the possible influence of noise. In particular, the knocking sounds were split into frames of 125 ms, then some frames to the left and the right were discarded and finally, the remaining central frames had their fundamental frequencies computed and averaged. The frame length was selected so that it matched the largest possible period, corresponding to a frequency of 8 Hz. A 25% of frames were dropped to the left and right, so only 50% of the frames remained in the middle. Thus, the periodicity of each annotated knocking sound was estimated by averaging the fundamental frequency found at the central segments of the event.

2.3. Dataset for classification

The dataset used to train the classifiers consisted of frames belonging to knocking events and non-annotated events (”noise”, from now on), extracted with a procedure similar to the one for estimating the fundamental frequency. The knocking annotations were loaded, split into equal-length frames, and cropped to disregard the most extreme segments. In this case, and responding to the fundamental frequency estimation results, the frames were 83ms long, being able to capture frequencies equal to or larger than 12 Hz. Also, the number of discarded margin frames was reduced to 20% from the left and the same amount from the right. This less extreme trimming provided a fair number of knocking frames, although with the risk that some of them contained noise. After this process, the new dataset consisted of 53721 frames belonging to knocking calls and 50286 frames belonging to noise. The latter was created by selecting 2-second long frames that did not belong to any annotation in the original dataset, so they could contain some non-annotated activity rather than just background noise.

This dataset was transformed in different ways to experiment with several features and classifiers. Firstly, 16 features related to fundamental frequency estimation and voicing detection were extracted with the aforementioned methods, adding also common voicing detection features such as power and zero crossing rate. In addition to the classical features, CREPE pitch estimator [21] was used to obtain the confidence of the network in the presence of pitch and the average activation along the last layer. Even if CREPE is not trained to detect the range of pitches that the knockings exhibited, these two metrics contributed positively to the separability between the knocking and noise cases.

Besides raw audio frames and fundamental frequency-related features, time-frequency representations were computed. In the first place, spectrograms were obtained using a Hanning window with a 2048-point Discrete Fourier Transform (DFT) and a 75% overlapping. On top of that, the mel spectrograms were computed with 20 energy bands. Finally, the Mel Frequency Cepstral Coefficients (MFCC) and their first and second-order time derivatives were obtained with 20 coefficients and no liftering.

2.4. Classification

Experiments were conducted with large architectures such as fully connected neural networks with four hidden layers or convolutional neural networks with two convolutional layers and two linear layers. These architectures were used to predict the audio segments in the time and frequency domains, respectively. Additionally, a Support Vector Machine (SVM) was tested with the estimated fundamental frequency features, since the dimensionality and variability of these data allowed for a simpler model.

To optimize the models, a grid search-like approach was followed. The goal was to achieve the best results without overfitting the classifiers excessively or creating architectures with a disproportionate number of parameters. Commonly utilized metrics were used to assess the performance of the classifiers, paying special attention to the number of false negatives, as it is crucial for systems

with our kinds of goals to avoid losing any potential instances of knocking calls.

3. EXPERIMENTS

3.1. Fundamental frequency estimation

Even considering the guidelines in [15] and [13], a broader spectrum of fundamental frequencies (10 Hz – 600 Hz) was tested to avoid missing any potential higher range of frequencies. Then, the final, narrower spectrum (8 Hz – 60 Hz) that agreed with the previous results was explored to obtain more fitting estimates.

3.2. Classification

In the first place, a Support Vector Machine was trained and optimized to classify the 16 features extracted by estimating the fundamental frequency. Grid-search cross-validation was approached by using 80% of the data to find the best parameters, and the remaining 20% was used to test the best classifier. The SVM used a Radial Basis Function (RBF) kernel with different penalty parameter (C) values and several kernel coefficients (γ). The tested parameters were in a logarithmically spaced range between 0.001 and 100. In addition, the number of features times the variance of the data was added as a possible value for γ.

Thereafter, some experiments were conducted by classifying time-domain frames with a Multi-Layer Perceptron (MLP). A first set of tests studied the utilization of raw frames to fit the model. Then, another set of experiments explored using the autocorrelation sequences, with and without Teager-Kaiser prefiltering, to lead the network to focus on periodicity. The architectures were selected by tuning their sizes and the training hyperparameters manually to obtain results coherent with our delimitations. To train these models and the ones to come, a train/validation/test split of 70%/20%/10% was utilized.

Finally, several tests were done in which the input features were obtained by means of different time-frequency representations. These representations included spectrograms, mel spectrograms, and MFCCs by themselves, with their first-order time derivatives, and with their first and second-order time derivatives. All of them were classified with convolutional neural networks, the architectures of which were optimized using a fairly superficial grid-search evaluation. The optimization criterion was to find architectures that didn't overfit the data excessively and that kept the number of parameters in a reasonable range.

4. RESULTS

4.1. Fundamental frequency estimation

The initial estimations made on a broad range (10 Hz – 600 Hz) showed how the large majority of knocking calls were distributed around 19 Hz, with some low power instances at the tail of the distribution. This tail was less significant when the signals were prefiltered with the Teager-Kaiser operator. This behavior was common for all the frequency estimation methods. Furthermore, when narrowing the range consequently to (8 Hz – 60 Hz), the fundamental frequencies estimated by all methods were consistent with each other, as seen in Figure 2. Additionally, the tails of the distributions were also characterized by low power knockings, although these tails were mitigated by the Teager-Kaiser operator, as can be observed at the bottom of Figure 2.

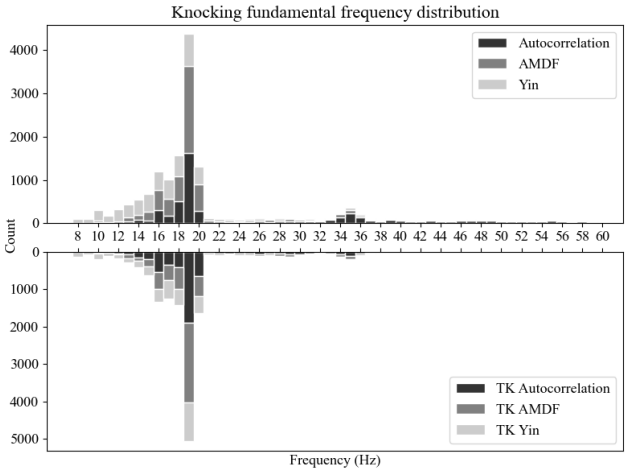

Figure 2: Distribution of the fundamental frequencies estimated with different autocorrelation-based techniques, with and without Teager-Kaiser (TK) prefiltering.

To select a representative fundamental frequency estimate, we repeated the experiments selecting only knocking vocalizations with a power higher than –80 dB, thus the 3327 knockings with higher power were selected amongst the total 8996. Out of all the fundamental frequency estimation strategies, the one that yielded the most central mean value and with nearly less standard deviation was AMDF prefiltered with Teager-Kaiser (Figure 3). The estimated values had a mean of 19.16 Hz with a standard deviation of 3.79 Hz. Although the distribution spanned from 8.63 Hz to 48.12 Hz, the first, second and third quartiles were clustered towards lower frequencies, exhibiting values of 18.18 Hz, 18.82 Hz and 19.23 Hz respectively.

4.2. Classification

The classification results evaluated on a test dataset unseen by the models can be found in Table 1. The presented evaluation corresponds to the models that yielded the most noteworthy results, in terms of evaluation metrics, features used and model dimensions altogether. It is worth noting that more complex classifiers, namely CNNs with larger architectures, attained slightly better scores. We found that CNNs that tripled the number of parameters listed in Table 1 generally outperformed their analogous smaller versions by less than 0.5% F1-score. However, we determined that the marginal improvement in performance did not justify the significant increase in size.

5. DISCUSSION

Overall, the knocking vocalizations seemed to have several characteristics that made them easily distinguishable from background noise and other events. Knocking events and noise were especially discernible when attending at both their time and frequency features, using representations such as spectrograms and MFCCs. Even if the lower power signals lacked a clear fundamental frequency, the energy distribution along their bandwidth proved to be a good descriptor.

			Evaluation metrics (%)			
Classifier	**Input feat.**	**Num. params**	**Precision**	**Recall**	**F1-score**	**FN ratio**
SVM	F_0 features	-	97.55	89.66	93.44	10.34
CNN	Mel spectrogram	1,053,633	99.49	94.87	97.12	5.13
CNN	MFCC	46,273	99.23	94.91	97.02	5.09
CNN	MFCC & D1	128,193	99.39	94.44	96.85	5.56

Table 1: Classification results for the different models. Evaluation metrics include precision, recall, f1-score and false negative ratio. The *Input feat.* column indicates the kinds of features used to train the models. Some abbreviations correspond to: F_0 *features* – fundamental frequency estimation-related features; *MFCC & D1* – mel frequency cepstral coefficients with their first-order time derivative appended.

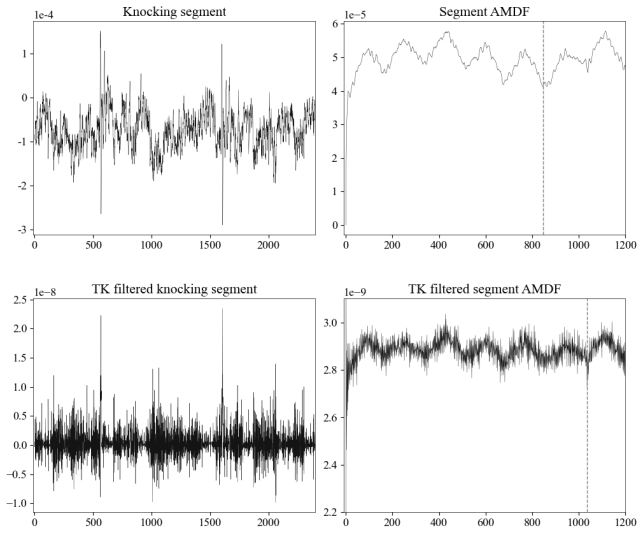

Figure 3: Periods of a knocking frame (top left) and its AMDF (top right). Right below, the effects of the Teager-Kaiser energy operator can be seen with the filtered signal (bottom left) and its AMDF (bottom right). The dashed lines show the period length in samples estimated by the AMDF. Even if the periods are not especially clear, a period of roughly 1000 samples can be observed in the left panels. In this example, the periodicity is correctly estimated only when the signal is prefiltered.

Therefore, the data allowed for a wide variety of classifiers to yield good results, even if the most complex ones pushed the evaluation ceiling slightly further. As mentioned in the previous section, far less complex architectures proved to be good alternatives, but as computational capabilities allowed, we decided to choose larger models. This can be especially seen in the models using MFCC features, that perform similarly to the mel spectrograms with significantly fewer parameters. Nevertheless, maintaining a low false negative ratio constituted a great challenge for all the tested systems. Moreover, classifiers working in the time domain -namely Fully Connected Neural Networks predicting the raw frame or any signal derived from the autocorrelation- were left out because they did not achieve results as good as the ones provided by the SVM or the CNNs, even when having a larger number of parameters.

Concerning the unsupervised estimation of the fundamental frequency, the results were arguably coherent with the ones inferred from [15]. However, as we do not have any means to identify the emitters of the vocalizations, our estimates may not be representative of the plurality of the seals. Nevertheless, our estimates should confidently capture the fundamental frequencies present in the knockings available in our dataset.

6. CONCLUSIONS

To help monitor Saimaa ringed seals, this study aims to provide a system capable of detecting instances of knocking vocalizations amongst underwater sound recordings. Our solution uses the time-frequency representations of audio segments to detect possible knocking vocalizations by classifying such frames in a deep learning framework. This proved to be the best alternative among using pitch estimation-related features and other representations in the time domain, achieving f1-scores as high as 97.12% when utilizing mel spectrograms. With the classified frames, the system can be easily expanded to provide the onsets and offsets of the detected events.

Additionally, this study explored the characteristics of Saimaa ringed seals' knocking vocalizations. Although not representative of all the seals, our study concludes that the fundamental frequencies of the studied knockings can be confidently found in a range between 12 Hz and 20 Hz, with occasional occurrences in higher frequencies. The statistics of such calls have been obtained by autocorrelation-based fundamental frequency estimators, namely AMDF, YIN and autocorrelation function extrema locators. As the lower signal-to-noise ratio (SNR) vocalizations proved to be a challenge, some of these estimations were preceded by a filtering based on the Teager-Kaiser energy operator, that emphasized the extrema.

Finally, we recommend further research based on the findings of this study. One potential approach is to use a Deep Neural Network (DNN) architecture to directly detect onsets and offsets without relying on classification. Additionally, other architectures such as Recurrent Neural Networks (RNN) or other CNNs that can observe more time context could be used to classify frames. A more advanced approach could involve utilizing a Transformer-based architecture, harnessing the power of attention mechanisms to detect knocking calls, following, for example, the methodology outlined by [22]. Additionally, other potential experiments could involve applying these architectures to other subspecies of seals that produce similar calls, or expanding the presented classifiers by incorporating "scraping" events.

7. ACKNOWLEDGMENTS

We would like to thank Professor Francisco Javier Hernando Pericas for bringing such a worthy and commendable topic to the attention of his UPC students.

8. REFERENCES

[1] T. Sipilä *et al.*, "Conservation biology of saimaa ringed seal (phoca hispida saimensis) with reference to other european seal populations," 2003.

[2] M. Kunnasranta, M. Niemi, M. Auttila, M. Valtonen, J. Kammonen, and T. Nyman, "Sealed in a lake—biology and conservation of the endangered saimaa ringed seal: A review," *Biological Conservation*, vol. 253, p. 108908, 2021.

[3] S. Bell, K. Hampshire, and M. Tonder, "Person, place, and knowledge in the conservation of the saimaa ringed seal," *Society and Natural Resources*, vol. 21, no. 4, pp. 277–293, 2008.

[4] M. Auttila, M. Niemi, T. Skrzypczak, M. Viljanen, and M. Kunnasranta, "Estimating and mitigating perinatal mortality in the endangered saimaa ringed seal (phoca hispida saimensis) in a changing climate," in *Annales Zoologici Fennici*, vol. 51, no. 6. BioOne, 2014, pp. 526–534.

[5] R. Lande, "Genetics and demography in biological conservation," *Science*, vol. 241, no. 4872, pp. 1455–1460, 1988.

[6] M. Koivuniemi, M. Auttila, M. Niemi, R. Levänen, and M. Kunnasranta, "Photo-id as a tool for studying and monitoring the endangered saimaa ringed seal," *Endangered Species Research*, vol. 30, pp. 29–36, 2016.

[7] T. Sipilä, E. Helle, and H. Hyvärinen, "Distribution, population size and reproductivity of the saimaa ringed seal (phoca hispida saimensis nordq.) in finland, 1980–84," *Finnish Game Research*, vol. 47, pp. 3–10, 1990.

[8] E. Nepovinnykh, T. Eerola, V. Biard, P. Mutka, M. Niemi, M. Kunnasranta, and H. Kälviäinen, "Sealid: Saimaa ringed seal re-identification dataset," *Sensors*, vol. 22, no. 19, p. 7602, 2022.

[9] T. Chehrsimin, T. Eerola, M. Koivuniemi, M. Auttila, R. Levänen, M. Niemi, M. Kunnasranta, and H. Kälviäinen, "Automatic individual identification of saimaa ringed seals," *IET Computer Vision*, vol. 12, no. 2, pp. 146–152, 2018.

[10] E. Nepovinnykh, T. Eerola, H. Kälviäinen, and G. Radchenko, "Identification of saimaa ringed seal individuals using transfer learning," in *Advanced Concepts for Intelligent Vision Systems: 19th International Conference, ACIVS 2018, Poitiers, France, September 24–27, 2018, Proceedings 19*. Springer, 2018, pp. 211–222.

[11] C. D. Escobar-Amado, M. Badiey, and S. Pecknold, "Automatic detection and classification of bearded seal vocalizations in the northeastern chukchi sea using convolutional neural networks," *The Journal of the Acoustical Society of America*, vol. 151, no. 1, pp. 299–309, 2022.

[12] H. Klinck, L. Kindermann, and O. Boebel, "Detection of leopard seal (hydrurga leptonyx) vocalizations using the envelope-spectrogram technique (test) in combination with a hidden markov model," *Canadian Acoustics*, vol. 36, no. 1, pp. 118–124, 2008.

[13] D. Mizuguchi, M. Tsunokawa, M. Kawamoto, and S. Kohshima, "Underwater vocalizations and associated behavior in captive ringed seals (pusa hispida)," *Polar Biology*, vol. 39, no. 4, pp. 659–669, 2016.

[14] M. Kunnasranta, H. Hyvärinen, and J. Sorjonen, "Underwater vocalizations of ladoga ringed seals (phoca hispida ladogensis nordq.) in summertime," *Marine Mammal Science*, vol. 12, pp. 611 – 618, 08 2006.

[15] A. Rautio, M. Niemi, M. Kunnasranta, I. J. Holopainen, and H. Hyvärinen, "Vocal repertoire of the saimaa ringed seal (phoca hispida saimensis) during the breeding season," *Marine mammal science*, vol. 25, no. 4, pp. 920–930, 2009.

[16] J.-w. Jung, H.-j. Shim, J.-h. Kim, and H.-J. Yu, "Dcasenet: An integrated pretrained deep neural network for detecting and classifying acoustic scenes and events," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 621–625.

[17] K. Wang, L. Yang, and B. Yang, "Audio event detection and classification using extended r-fcn approach," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, 2017, pp. 128–132.

[18] V. Morfi and D. Stowell, "Deep learning for audio event detection and tagging on low-resource datasets," *Applied Sciences*, vol. 8, no. 8, p. 1397, 2018.

[19] A. De Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[20] P. Guyot, "Fast python implementation of the yin algorithm (v1.1.1)," https://doi.org/10.5281/zenodo.1220947, 2018.

[21] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 161–165.

[22] N. Moritz, G. Wichern, T. Hori, and J. Le Roux, "All-in-one transformer: Unifying speech recognition, audio tagging, and event detection." in *INTERSPEECH*, 2020, pp. 3112–3116.