Cumulative Occurrences of Notes as Basis for Key Classification

Alexczar Dela Torre
College of Computer Studies
De La Salle University
Manila, Philippines
alexczar_delatorre@dlsu.edu.ph

Jose Mikhael Uriel Promentilla
College of Computer Studies
De La Salle University
Manila, Philippines
jose_mikhael_promentilla@dlsu.edu.ph

***Abstract*—Key classification is a important component in music analysis. The study broadly identifies two key classification methods. Profile based methods use key profiles or pitch class profiles to classify keys. Neural network based methods use require extensive training to become genre-agnostic key classifiers. The proposed algorithm aims to reach a compromise between the two methods by being genre-agnostic without the extensive training. The proposed algorithm classifies keys through matching the top note occurrences with the note components of the key signature and iteratively eliminates key candidates. The proposed algorithm is able to achieve a reasonable performance of 45.03% from the GiantSteps Key dataset.**

Keywords-component; key classification; key-profiles;

I. INTRODUCTION *(HEADING 1)*

Music is a series of sounds holds an important part in human culture. In the modern age, these sounds are arbitrarily labeled into musical notes. Notes are generally represented into 7 letters: A, B, C, D, E, F and G. Notes signifies the highness or lowness of a sound. In other words, a particular note corresponds to a certain pitch or also known as the fundamental frequency. An example is the note A4 corresponding to a value of 440 Hz.

The fundamental frequency can be measured in Hertz (Hz), a unit of frequency. The fundamental frequency predominantly determines the highness and lowness of the sound even if it is not the only frequency present in the sound. There are other frequencies in the sound, and the group of frequencies that are a positive integer multiple of the fundamental frequency is defined as the harmonics. The first fundamental frequency is defined as the first harmonic. If the first harmonic was 100 Hz, then the following harmonics would be 200 Hz (second harmonic), 300 Hz (third harmonic), and so on and so forth.

Notes can be ordered into increasing or decreasing order of pitch, and this ordered notes is known as the scale. Finally, a scaled group of notes can be categorized into key signatures. In Western music, there are primarily 24 key signatures with each key signature composed of seven notes in scaled order. One of the key signatures that exist is the C major and has the notes C, D, E, F, G, A, B in its signature. Each note is separated by a gap known as the tone or semitone. The first note of the key signature is the labeled as the tonic and serves as the reference for the succeeding notes. The tonic of the C major key signature would be the musical note C.

Key classification is an important aspect in music analysis. Pauws [1] even mentions that musical keys can be applied to mood induction. Key classification also contributes to music composition. In order to extend a music or compose a new one that would be compatible with other music, knowing the keys would be of importance. Korzeniowski and Widmer [2] also says that key classification is of importance because it is important to mix music harmonically well. Korzeniowski and Widmer [2] also observes that deriving a key is a demanding task and that it would be costly even for experts to annotate large amounts of music classification. Therefore, there is an opportunity for research on key classification..

II. REVIEW OF RELATED LITERATURE

A. *Profile Based Classification*

Temperley [3] developed one of the first models to key classification based on the key-profile model. The Bayes method was applied to the key profile model. By comparing the distributions of the pitch classes and the distributions of the key profiles, the key profile is matched to the pitch classes. The key profile with the highest correlation value with the pitch classes is classified as the key of the signal. The Bayesian model had a 77.1% accuracy rate with the Kostka-Payne corpus.

Pauws [1] develops a key extraction algorithm through chromagrams. The extracted signal comes from the range of A0 to A6 (27.5 Hz to 1760 Hz) and segmenting them into a 100 msecs time frames. Chromas are extracted and structured into a chromagram by assigning each bin a chroma value and mapping each frame into one of the 12 ideal chroma values. Each frame is transformed into a spectral representation and the likelihood of each pitch is aggregated into a chromagram. Key profiles are also used to classify keys in this study. The key profile with the highest correlation with the chromagram is classified as the key. Pauws [1] then shows that the algorithm classified with a 75.1% accuracy rate on piano sonatas.

Noland and Sandler [4] furthermore builds on top of key profiles by using a Hidden Markov Model (HMM). There are 24 states and each state is represented as one of the major or minor key. Chord transitions served as the observations of the model. The training was done on a certain sequence of chord transitions. This resulted to a 91% accuracy rate on the 110 Beatles songs.

The previous studies show the past key classification systems. The past systems generally follow the same method: base the classification on the template of key profiles, process the signal or train the model, and then classify based on correlations. The key profiles can be further modified such as the Bayesian method. Signals are processed by extracting the chroma feature from each frame, and aggregating them into a chromatogram. HMM models can be trained based on chord transitions. And in the end, using the chromatogram or models, the key profiles with the highest correlation are selected as the classified key. The problem with this, as noted by Korzeniowski and Widmer [2], is that key templates are specific to musical style. This leads to key templates being biased depending on what musical style it came from.

This is exactly what O 'brien and Lerch [5] evaluated. O 'brien and Lerch [5] evaluated key profiles under 9 different genres. The study used the GTZAN dataset composed of 1000 songs from different genres such as classical, rock, disco, jazz, etc. When a pitch chroma feature was extracted and classified using key profiles, the accuracy rate was 35.04% with a standard deviation of 1.97. This performance indicates that each profile had genre-specific information. Templates-based classification will then have limited applications if it is tailored to it's music genre.

B. Neural Network Based Classification

Korzeniowski and Widmer [6] propose an end-to-end system on key classification through convolutional neural networks. The advantage of training an end-to-end neural network is that it does not require pre-processing steps and feature design. The system bypasses the need for a template and the bias associated with it. Due to the nature of neural networks, massive amounts of annotated data is a prerequisite. The following are the datasets used:

1. GiantSteps Key Dataset.
2. GiantSteps MTG Key dataset.
3. McGill Billboard Dataset.

GiantSteps Key Dataset is comprised of 2 minutes audio segments from www.beatport.com. All 604 audio recordings had a corresponding ground truth value for the global key. This dataset was gathered based from the study of Knees et al. (2015).

GiantSteps MTG Key dataset are also a set of 2 minute audio segments and counts up to 1486 recordings. This is also compiled by Knees et al. [7] and is retrieved from www.beatport.com. Each recording has an annotation for its global key.

McGill Billboard Dataset is comprised of pop and rock music and counts up to 742 unique songs. The problem is that only the tonic is annotated for each piece. The mode of the audio recording was then estimated through the tonic and chord annotations.

Two separate test sets were selected: the GiantSteps Key Dataset and the McGill Billboard Dataset. The training sets were also separate, utilizing all the datasets. With the test set of GiantSteps Key Dataset and training set of GiantSteps MTG Key, the system performed 67.9% accuracy. Meanwhile, with the test set of the McGill Billboard Dataset, and training set of GiantSteps MTG Key, the system performed 62.5% accuracy. The drop in accuracy indicates that different musical styles have biases in key classification. But even if the model was trained on a different dataset with the test dataset, the system did not greatly drop in accuracy. Korzeniowski and Widmer [6] observes based from the results that the model is able to adapt to different musical styles without sacrificing much accuracy. The study demonstrates that the CNN system has state-of-the-art performance on electronic music and pop or rock music.

Korzeniowski and Widmer [2] continues their research by proposing a genre agnostic key classification system. Training with certain musical style optimizes the system to such classify keys according to that style which introduces bias and loss on accuracy on other datasets. The neural networks was modified as to not be biased during the training phase of a specific musical style. Furthermore, the study shows that the system performs well not despite to being exposed to different musical styles, but because of it.

The system was trained on 3 different datasets with each corresponding to a different music genre or style. The following are the music genre and the dataset used.

- Electronic Dance Music from the GiantSteps MTG Key dataset.
- With GiantSteps Key Dataset as the test set.
- Pop/Rock Music from the McGill Billboard dataset.
- Classical Music from their internal database.

The GiantSteps Key Dataset, GiantSteps MTG Key dataset, and McGill Billboard dataset remain as the same dataset as in the previous study.

The dataset for the classical music is a collection of 1504 pieces of music. Because classical music change their keys (modulation) during the piece, only the first 30s of the piece is used. Other datasets from KeyFinder, Isophonics, R. Williams, and Rock were also used as test datasets in the study.

The results show that with the GiantSteps dataset, they system performed with a 67.9% accuracy rate. Although the results for the GiantSteps dataset remain the same from the previous study, for the McGill Billboard dataset, the system performed with a 79.9% accuracy rate. There is a great increase in performance compared to the 62.5% accuracy of the previous study. Korzeniowski and Widmer [2] conclude that models need to consider the harmonic coherence for key classifications of short segmented audio.

TABLE 1: SUMMARY OF PERFORMANCES FOR EACH REVIEWED LITERATURE

Author	Method	Dataset	Performance
Temperley (2002)	Key Profile Bayesian Model	Kostka-Payne Corpus	77.1%
Pauws (2004)	Key Profile Chromatograms	Piano Sonatas	75.1%
Noland and Sandler (2006)	Key Profile HMM	Beatles Songs	91%
O 'brien and Lerch (2015)	Key Profile	GTZAN Dataset Different Music Genre	35.04%
Korzeniowski and Widmer (2017)	Neural Networks (CNN)	GiantSteps Key Dataset Electro Music	67.9% (62.5%)
Korzeniowski and Widmer (2018)	Neural Networks (CNN)	GiantSteps Key Dataset	67.9% (79.9%)

The previous studies demonstrate that state-of-the-art key classifications involves training convolutional neural networks (CNN) in different datasets. But CNN requires massive amounts of data and takes a fair amount of training time. The proposed algorithm requires no training set and training time. The algorithm also addresses the overfitting of a profile to a music genre and is genre-agnostic. The algorithm aims to achieve quick key classification and reasonable accuracy.

III. EXPERIMENTAL SETUP

A. Input Signal Dataset

The database to be used is the GiantSteps Key Dataset. The dataset is comprised of 604 audio recordings each one with a duration of 2 minutes. The sampling rate is of standard audio quality which is 44100 samples per second or 44.1 kHz. Each recording has a corresponding ground truth value for the global key. Knees et al. [7] extracted the global keys through expert sources online, key information based on online forum, and beats per minute features. Experimentation on this database also serves as a platform for direct comparison with neural networks-based key classification.

The GTZAN Dataset was also used to test the proposed algorithm The GTZAN dataset was used by O 'brien and Lerch [5] and would provide a platform for comparison. The GTZAN dataset is composed of 1000 songs, with different music genres. There are ten different music genre with each having 100 songs. Each song lasts for 30 seconds. This dataset will be used to confirm the genre-agnostic quality of the proposed algorithm.

B. Keys Signatures

Each input signal will be classified into one of the 12 key signatures. Each key signature is composed of seven note components. These seven note components are matched to classify the input signal.

C major or A minor	C	D	E	F	G	A	B
C# major or A# minor	C#	D#	F	F#	G#	A#	C
D major or B minor	D	E	F#	G	A	B	C#
D# major or C minor	D#	F	G	G#	A#	C	D
E major or C# minor	E	F#	G#	A	B	C#	D#
F major or D minor	F	G	A	A#	C	D	E
F# major or D# minor	F#	G#	A#	B	C#	D#	F
G major or E minor	G	A	B	C	D	E	F#
G# major or F minor	G#	A#	C	C#	D#	F	G
A major or F# minor	A	B	C#	D	E	F#	G#
A# major or G minor	A#	C	D	D#	F	G	A
B major or G# minor	B	C#	D#	E	F#	G#	A#

Figure 1. Key Signatures and their corresponding key components

C. Note Frequencies

There are 12 notes and each note has a corresponding frequency that can be matched to it. Octaves are factors of its frequency. For example, the 2nd octave is simply a frequency of the first octave scaled by 2. Any frequency or octave is classified according to its note.

TABLE 2: FREQUENCIES AND THEIR CORRESPONDING NOTE

Octave Note	1	2	3	4	5	6	7
C	32.703 (24)	65.406 (36)	130.81 (48)	261.63 (60)	523.25 (72)	1046.5 (84)	2093.0 (96)
C♯/D♭	34.648 (25)	69.296 (37)	138.59 (49)	277.18 (61)	554.37 (73)	1108.7 (85)	2217.5 (97)
D	36.708 (26)	73.416 (38)	146.83 (50)	293.66 (62)	587.33 (74)	1174.7 (86)	2349.3 (98)
E♭/D♯	38.891 (27)	77.782 (39)	155.56 (51)	311.13 (63)	622.25 (75)	1244.5 (87)	2489.0 (99)
E	41.203 (28)	82.407 (40)	164.81 (52)	329.63 (64)	659.26 (76)	1318.5 (88)	2637.0 (100)
F	43.654 (29)	87.307 (41)	174.61 (53)	349.23 (65)	698.46 (77)	1396.9 (89)	2793.8 (101)
F♯/G♭	46.249 (30)	92.499 (42)	185.00 (54)	369.99 (66)	739.99 (78)	1480.0 (90)	2960.0 (102)
G	48.999 (31)	97.999 (43)	196.00 (55)	392.00 (67)	783.99 (79)	1568.0 (91)	3136.0 (103)
A♭/G♯	51.913 (32)	103.83 (44)	207.65 (56)	415.30 (68)	830.61 (80)	1661.2 (92)	3322.4 (104)
A	55.000 (33)	110.00 (45)	220.00 (57)	440.00 (69)	880.00 (81)	1760.0 (93)	3520.0 (105)
B♭/A♯	58.270 (34)	116.54 (46)	233.08 (58)	466.16 (70)	932.33 (82)	1864.7 (94)	3729.3 (106)
B	61.735 (35)	123.47 (47)	246.94 (59)	493.88 (71)	987.77 (83)	1975.5 (95)	3951.1 (107)

D. Pitch Detection Methods

The main pitch detection used for the proposed algorithm is the Cepstrum Pitch Determination [8]. The reason for this selection is because the cepstrum pitch detection method is often used in other literatures and serves as an established baseline. Other pitch detection methods that are also included are the Normalized Correlation Function, Pitch Estimation Filter, Log-Harmonic Summation, and Summation of Residual Harmonics.

1. Cepstrum Pitch Determination (CEP) [8].
2. Normalized Correlation Function (NCF) [9].
3. Pitch Estimation Filter (PEF) [10].
4. Log-Harmonic Summation (LHS) [11].
5. Summation of Residual Harmonics (SRH) [12].

IV. PROPOSED ALGORITHM

The proposed algorithm involves counting the occurrences of each note from the input signal and matches the most occurring notes to seven component notes of each key signature. The key signature with the highest match will be the predicted key for that input signal. The proposed algorithm is as follows:

1. Initialize the databases for key signatures and frequencies for each note.
2. Downsample the input signal by a factor of 2.
3. Segment each frame into 52 ms (millisecond) windows with an overlap of 42ms.
4. Estimate fundamental frequency/pitch for each frame.
5. Classify each frequency into one of the 12 notes.
6. Count the occurrences of each note for the entire input signal.
7. Set a list of candidate keys (initially 12 keys).
8. Starting from 1, get the top n occurring notes.
9. Match the top n occurring notes with the 7 note components of each key signature.
10. Reset the list of candidate keys to include only those with the highest matches with the top n occurring notes.
11. Increment the top n occurring notes by 1.
12. Repeat until only 1 candidate key remains.
13. The remaining candidate key is selected to be the predicted key for classification.

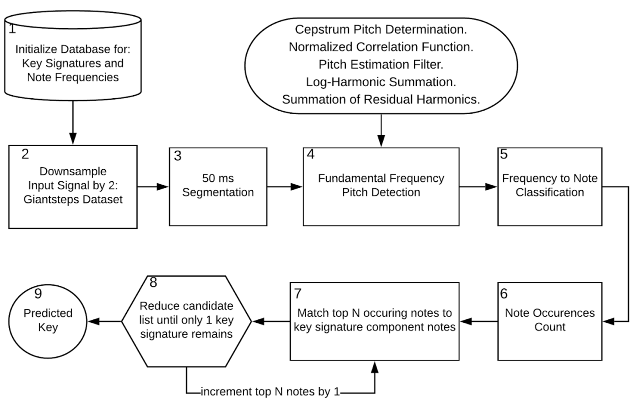

Figure 2: Proposed Algorithm Flow

Due to hardware limitations and the high sampling rate reaching 44100 Hz, the input signal was downsampled by a factor of 2. The Nyquist frequency is then 22050 Hz, which still serves as a sufficient upper limit. Initially, the top occurring notes to be matched was set to seven notes, the same number as the seven component notes each key signature has. But it was observed that the algorithm is left with multiple candidate keys, often three candidates, due to the similarity of the key signatures. There are multiple key signatures that differ by only one note component. Therefore, when even one note component has occurred more than it should have in input signal, similar key signatures remain as key candidates. Despite this, the correct key was always among the key candidates. To resolve this, the initial algorithm was modified to start the matching of the key candidates from the top one occurring note, and to discard iteratively other key candidates that have different note components.

V. RESULTS

For the initial experimentation, 32 audio recordings were selected from the GiantSteps Key dataset and was tested under different conditions such as pitch detection methods and segmentation parameters with the proposed algorithm.

A. Pitch Detection Methods

Due to the intensive computations on certain pitch detection methods such as PEF, SHS, and SRH, in order to perform those methods, the overlapping window was set to be none.

TABLE 3: PERFORMANCE OF DIFFERENT PITCH DETECTION METHODS

Pitch Detection Methods:	Accuracy Rate:
PEF	0%
LHS	12.5%
NCF	18.75%
SRH	34.375%
CEP	46.875%

The cepstrum pitch determination performs the best among the pitch detection methods in terms of accuracy. It is also observed that NCF and CEP gave quicker key classification results than PEF, NCF, and SRH. Based from the previous results, CEP was experimented under different parameters to improve performance.

B. Cepstrum Pitch Determination Parameters

TABLE 4: PERFORMANCE OF DIFFERENT CEPSTRUM SEGMENTATION PARAMETERS

Cepstrum Parameters	Accuracy Rate
Default Cepstrum Parameters • Segmented 52 ms • Overlap 42 ms	43.75%
Setting Frequency Range from 20 Hz ~ 4000 Hz	15.625%
Segmented by 100 ms	46.875%
Segmented by 200 ms	46.875%
Overlap 0 ms / No Overlap	46.875%
Segmented by 100 ms and No Overlap	40.625%

The default parameters for CEP is a segmentation window of 52 ms and an overlapping window of 42 ms. Increasing the segmentation or analysis window or removing the overlaps showed an increase in performance. But the combination of increasing the segmented window and removing the overlaps worsens the performance.

C. GiantSteps Key Dataset

Based from the previous results, the entire GiantSteps Key dataset will be experimented with the cepstrum pitch determination as the pitch detection method and with different parameters such as the overlapping window.

TABLE 5: ENTIRE GIANTSTEPS KEY DATASET PERFORMANCE

CEP Conditions	Accuracy Rate
Downsampled by 2 Segmentation 52 ms Overlap 42 ms	45.03%
Downsampled by 4 Segmentation 52 ms Overlap 42 ms	34.60%
Downsampled by 2 Segmentation 52 ms Overlap 0 ms / No Overlap	42.55%

Downsampling by a factor of 4, although shortens the key classification time, removes too much samples to justify the tradeoff between the accuracy rate and performance time. In total, key classification for 1 input takes 5.781 seconds.

D. GTZAN Dataset

Testing the proposed algorithm under the GTZAN dataset with 837 songs with different music genre resulted in to an accuracy rate of 42.29%. The reason the number of songs decreased from 1000 songs because only 837 of the songs has its ground key value identified. The ground keys for the classical music all have unidentified keys therefore, it has 0 songs for experimentation.

TABLE 6: PERFORMANCE UNDER DIFFERENT MUSIC GENRE

Dataset	Number of Songs	Accuracy Rate
Total	837	42.29%
Blues	98	10.20%
Classical	0	NaN
Country	99	77.78%
Disco	98	56.12%
Hiphop	81	17.29%
Jazz	79	37.97%
Metal	93	25.51%
Pop	94	60.64%
Reggae	97	54.64%
Rock	98	56.12%

VI. DISCUSSION

Different key classification methods were reviewed under this study. There are generally two categories for key classification, namely profile based methods and neural networks based methods. The problem identified with profile based methods is that profiles can be overfitted to its music genre resulting to a degradation in performance when tested under other music genre from its profile. Under its music genre, profile based methods has an accuracy rate of 75.1% [1]. When tested under different music genre, performance was degraded to 35.04% [5]. Key classification that use neural networks performed with a 67.9% accuracy rate [2]. Although lower than profile based methods, the neural networks were able to generalize from its training data as to retain its performance throughout different music genres. The main disadvantage of neural networks is that is that it requires training time and a vast amount of training data. The proposed algorithm aims to reach a compromise between the two methods by being genre-agnostic in performance and would not require vast amounts of data and training time.

TABLE 7: COMPARATIVE RESULTS AMONG KEY CLASSIFICATION METHODS AND PROPOSED ALGORITHM

Methods	Accuracy Rate
Profile Based Method	75.1%
Profile Under Different Music Genres	35.04%
Neural Network Based Method	67.9%
Proposed Algorithm	45.03%

With a total accuracy rate of 42.29% from the GTZAN dataset, the proposed algorithm retaining its performance, indicating that it is genre-agnostic. The proposed algorithm performed badly with Blues songs with a 10.2% accuracy rate. Meanwhile it performed very well in country songs with an accuracy rate of 77.78%. The used profile in the proposed algorithm is not rooted from any distinct music genre, avoiding the trap of overfitting to a single music genre like other profile based methods. Though our proposed algorithm is supposedly genre-agnostic by definition, since the templates used was not rooted in any specific music genre, the results reveal that certain genre of music are more compatible having a far higher accuracy rate such as Country music. Perhaps this is due to the rate of repeated notes in a certain music. Country music might have less modulations or simply less notes that are not a part of its key signature. In contrast, Blues music may have a significantly more modulations or notes that are not part of its key signature.

TABLE 8: DIFFERENT PUBLISHED KEY DETECTION ALGORITHMS ON THE GIANTSTEPS KEY DATASET

Algorithm	Accuracy Rate
Beatport (2014 website)	29.14
QM-Key	39.40
Essentia (2015)	30.46
Faraldo et al., ECIR 2016 – edmt	33.40
Faraldo et al., ECIR 2016 – edma	58.10
Faraldo et al., ECIR 2016 – edma	58.10
Faraldo et al., ECIR 2016 – edmm	64.20
KeyFinder	45.36
Mixed-In-Key 7	67.22
Pioneer Rekordbox v3.2.2	71.85

According to the website by Knees et al. [7] detailing their study and dataset, a table of the performances for published key classifiers, including commercial ones, on the GiantSteps Key dataset was shown. It can be observed that performance on the GiantSteps Key dataset is relatively low, indicating that perhaps the dataset itself has a higher difficulty in classifying its keys. The proposed algorithm is observed to have a better performance than four other algorithms; and is on par with the KeyFinder classifier.

Although the proposed algorithm does not achieve stellar performance, it reaches a compromise between the two methods by being genre-agnostic and not requiring extensive training.

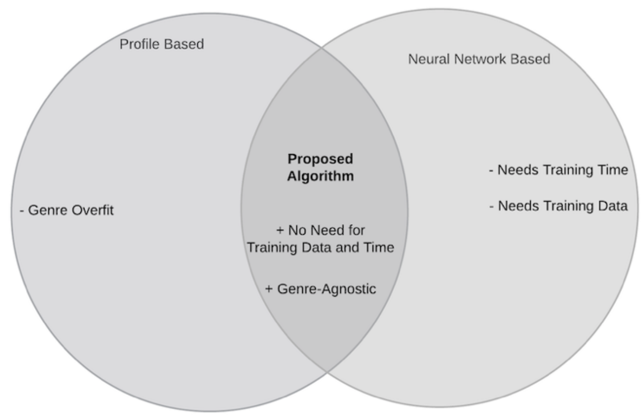

Figure 3: Proposed Algorithm Advantage

VII. CONCLUSION

Profile based methods suffer from overfitting its music genre while neural network based methods require training data and time. The proposed algorithm aims to reach a compromise between the two methods by being genre-agnostic and not requiring any sort of training. The proposed algorithm uses cepstrum pitch determination and the iterative elimination of candidate keys to classify the keys from the GiantSteps Key dataset. The proposed algorithm reaches a reasonable performance achieving its initial objective.

VIII. REFERENCES

[1] Pauws, S. (2004, 01). Musical key extraction from audio.

[2] Korzeniowski, F., & Widmer, G. (2018, 08). Genre-agnostic key classification with convolutional neural networks.

[3] Temperley, D. (2002, 01). A bayesian approach to key-finding. In (p. 149-155). doi: 10.1007/3-540-45722-418.

[4] Noland, K., & Sandler, M. (2006, 01). Key estimation using a hidden markov model. In (p. 121-126).

[5] O 'brien, C., & Lerch, A. (2015, 09). Genre-specific key profiles.

[6] Korzeniowski, F., & Widmer, G. (2017). End-to-end musical key estimation using a convolutional neural network.CoRR,abs/1706.02921. Retrieved from http://arxiv.org/abs/1706.02921.

[7] Knees, P., Faraldo, ́A., Herrera, P., Vogl, R., B ̈ock, S., H ̈orschl ̈ager, F., & Le Goff,M. (2015, October). Two data sets for tempo estimation and key detection in electronic dance music annotated from user corrections. In Proceedings of the 16th international society for music information retrieval conference(ISMIR'15). M ́alaga, Spain.

[8] Noll, Michael A. "Cepstrum Pitch Determination." The Journal of the Acoustical Society of America. Vol. 31, No. 2, 1967, pp. 293–309.

[9] Atal, B.S. "Automatic Speaker Recognition Based on Pitch Contours." The Journal of the Acoustical Society of America. Vol. 52, No. 6B, 1972, pp. 1687–1697.

[10] Gonzalez, Sira, and Mike Brookes. "A Pitch Estimation Filter robust to high levels of noise (PEFAC)." 19th European Signal Processing Conference. Barcelona, 2011, pp. 451–455.

[11] Hermes, Dik J. "Measurement of Pitch by Subharmonic Summation." The Journal of the Acoustical Society of America. Vol. 83, No. 1, 1988, pp. 257–264.

[12] Drugman, Thomas, and Abeer Alwan. "Joint Robust Voicing Detection and Pitch Estimation Based on Residual Harmonics." Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 2011, pp. 1973–1976.