applied sciences

MDPI

Article

Application of Pitch Derived Parameters to Speech and Monophonic Singing Classification †

Xabier Sarasola *, Eva Navas *, David Tavarez, Luis Serrano, Ibon Saratxaga and Inma Hernaez

Aholab Signal Processing Laboratory, University of the Basque Country (UPV/EHU), 48013 Bilbao, Spain

* Correspondence: xsarasola@aholab.ehu.eus (X.S.); eva@aholab.ehu.eus (E.N.)

† This paper is an extended version of our paper published in IberSPEECH2018.

Received: 8 July 2019; Accepted: 30 July 2019; Published: 2 August 2019

check for updates

Abstract: Speech and singing voice discrimination is an important task in the speech processing area given that each type of voice requires different information retrieval and signal processing techniques. This discrimination task is hard even for humans depending on the length of voice segments. In this article, we present an automatic speech and singing voice classification method using pitch parameters derived from musical note information and f_0 stability analysis. We applied our method to a database containing speech and a capella singing and compared the results with other discrimination techniques based on information derived from pitch and spectral envelope. Our method obtains good results discriminating both voice types, is efficient, has good generalisation capabilities and is computationally fast. In the process, we have also created a note detection algorithm with parametric control of the characteristics of the notes it detects. We compared the agreement of this algorithm with a state-of-the-art note detection algorithm and performed an experiment that proves that speech and singing discrimination parameters can represent generic information about the music style of the singing voice.

Keywords: audio segmentation; voice discrimination; singing voice; pitch

1. Introduction

Discrimination of speech and singing is not an easy task even for humans, who need approximately one-second-long segments to discriminate singing and speaking voices with more than 95% accuracy [1]. When speech is repeated, rhythm patterns appear and repeated spoken segments are perceived as singing [2,3]. Even if singing and speech are closely related and difficult to be distinguished by humans, speech technologies developed for spoken speech are not directly applicable to singing speech. In general, results deteriorate heavily when using classic automatic speech recognition [4] or phonetic alignment [5] methods with singing speech. Therefore, when dealing with recordings that contain both types of speech, some tool must be designed to identify them in order to apply the technique suitable for each type of voice. Many works have addressed the problem of speech and music discrimination [6–9], but these techniques are not directly applicable in the case of a capella singing because they exploit the presence of music.

There are two different tasks related to the automatic discrimination of singing and speech: classification, when each segment (or file) belongs to only one class and segmentation where both classes are present in the same file and have to be first separated and then classified. For the classification step, short-term and long-term features extracted from the audio signal have been traditionally used. With the use of short-term features, the signal is classified at frame level and then decisions for different frames must be combined to obtain a single decision for each segment. Most works, however, rely on long-term parameters related to pitch to discriminate speech and singing. For instance, distribution of different pitch based features was tested in [10] and most of them achieved correctly classifying more than half the database. The best results in this work were obtained for autocorrelation of pitch between syllables. Pitch and energy related features were also fed to a multi-layer Support Vector Machine (SVM) in [11] and the pitch related features contributed the most to the discrimination. Thomson [12] proposed to use the discrete Fourier transform of the pitch histogram to estimate the distribution of pitch deviations that should have lower variance for singing than for speech and obtained very good results.

Some works combine short-term features related with the spectral envelope and long-term features related with prosody to train Gaussian Mixture Models (GMM) and distinguish speech and singing. In [1], the authors found that short-term features work better for segments shorter than 1 s and pitch related features obtain the best results for segments longer than 1 s. On the contrary, the work in [13] finds that spectral features work better than pitch based ones when used alone in their database composed of segments between 17 and 26 s. Regardless, they obtained the best results when combining both types of features. In [14], a large set of 276 attributes related with spectral envelope, pitch, harmonic to noise ratio and other characteristics and a ensemble of classifiers are proposed to classify singing voice, speech and polyphonic music and get very good results.

As mentioned before, a related problem that has been studied in more depth is the separation of the singing voice from the background music. In this area, different strategies have been applied. Many works are based on the use of pitch related information. For instance, in [15] and with the goal of melody extraction, the use of a joint detection and classification network to simultaneously estimate pitch and detect the singing voice segments in a music signal is proposed. Pitch information is also exploited in an iterative way to detect singing voice in a music signal using a tandem algorithm in [16]. Other works apply voice activity detection, like [17] where the authors propose to separate the singing voice from the instrumental accompaniment using vocal activity information derived with robust principal component analysis. Spectral information has also been used, as in [18] where Long Short-Term Memory (LSTM) Networks are applied to separate singing from the musical accompaniment in an online way. Deep neural networks are also considered in [19] where a Bidirectional Long Short-Term Memory (BLSTM) Network is applied to enhanced vocal and music components separated by Harmonic/Percussive Source Separation [20]. However, these techniques usually exploit the differences between vocal signals and signals produced by music instruments. In general, they are not well suited for speech and singing separation [11,12], which is the problem we want to address in this work.

We are interested in singing and speech segmentation because it is a basic tool necessary to deal with recordings from bertsolarism. Bertsolarism (bertsolaritza in Basque) is the art of live improvising sung poems typical in the Basque Country. The host of the bertsolarism show proposes the topic for the improvised verses and introduces the singer. Then, the singer has to create a rhymed verse about the proposed topic and sing it with a melody that fits the defined metre. The singers are called bertsolari and they have to perform a capella, without the support of any musical instrument. The main goal is to produce good quality verses and not to sing on tune, so most bertsolaris are not professional singers. Bertsolarism associations provide access to many recorded live sessions of bertsolaritza together with the corresponding transcriptions and metadata. The recordings of the live shows include speech from the host, sung verses and overlapped applauses and, consequently, a good segmentation system is required to isolate the segments of interest.

In this paper, we propose a new technique to segment speech and singing using only two parameters derived from the pitch curve. The technique is applied to a Bertsolaritza database and a database of popular English songs and compared with other techniques proposed in the literature. Experiments show that the proposed segmentation algorithm is fast, robust and obtains good results. A novel algorithm to automatically label notes in the audio file is also proposed.

The rest of the paper is organized as follows. Section 2 presents the datasets and describes the proposed segmentation system, including the algorithm developed to assign a musical note label to each audio frame. Section 3 describes the experiments and results of both algorithms and compares them to other techniques for note labelling and speech/singing voice classification. Finally, Section 4 discusses our findings.

2. Materials and Methods

2.1. Datasets

There are very few publicly available datasets that contain monophonic singing (MIR-1K [21] and the Singing Voice Audio Dataset [22] for instance) and even fewer datasets that contain both spoken and singing speech. In fact, we have only found the NUS Sung and Spoken Lyrics Corpus [23] that is not completely suitable for our task because the files it contains are mono-class, i.e., they contain either speech or singing. Our goal is to separate speech segments from singing segments contained in the same recording, so we have used our Bertsolaritza database [24] to develop and train the algorithms. These algorithms have then been tested both in Bertsolaritza and NUS databases.

The Bertsolaritza database contains 2095 Basque audio files from 187 different singers and has a total duration of 59 h, 10 min and 40 s. The whole database has been manually labelled to separate singing from speech, resulting in 53 h, 22 min and 38 s of singing voice and 5 h, 48 min and 2 min of speech. The metadata that come with the recordings have information about all the singer identities, but the host identity is only annotated if he or she also sings to introduce the topic for the verses. It was not feasible to manually label the identities of the rest of hosts, therefore we decided to use an approximate method to create the host labels for the metadata that were missing. Recordings are separated in sessions that correspond to different places and dates. We decided to assign the same speaker to all the recordings of one session, as usually there is only one host in each bertsolaritza show. We also classified the genre of the hosts by applying a threshold to the average value of the f_0 in such a way that speakers with an average f_0 value higher than the threshold are considered female and male, otherwise. The optimal threshold to define the genre of a host has been defined using the hosts whose identity was present in the metadata (31 different speakers in 54 sessions, with a total duration of 15 min and 5 s). Table 1 shows the classification results obtained using different thresholds, where 250 Hz is the value that gets best F-score, so we used this value to automatically label the genre of hosts in the recordings that lacked this information. With metadata available for all the singers and hosts, the final database contents are shown in Table 2. In most cases, speakers either sing or act as host, but, as we have already commented, some hosts give the topic for the improvised verses singing as well. These hosts appear in the recordings both singing and speaking.

In the NUS database, each participant sings and reads the lyrics of four different well-known songs in English. There are 20 different songs in the database that includes 6 male and 6 female participants. All the participants are professional singers. Each recording contains either speech or singing voice and we used a Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) based Voice Activity Detector (VAD) explained in Section 2.2.1 to obtain the voice segments and labelled them with the type of the recording.

Table 1. Results of the automatic genre classification.

Threshold (Hz)	Precision	Recall	F-Score
100	0.13	0.5	0.21
150	0.66	0.63	0.44
200	0.77	**0.84**	0.77
250	0.88	0.77	**0.81**
300	**0.9**	0.64	0.67
350	0.37	0.5	0.43

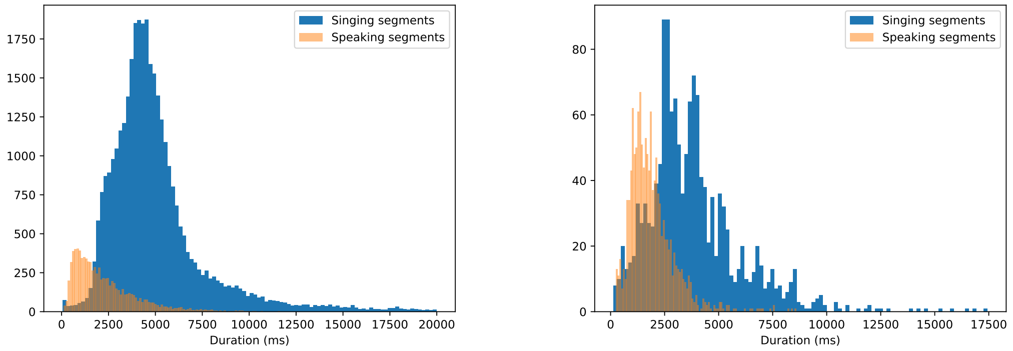
Table 2. Number of speakers and singers in the Bertsolaritza database.

	Singer	Host	Singers Who Host	Hosts Who Sing	Total
Female	33	43	1	9	86
Male	140	528	13	28	709
Total	173	571	14	37	

Both databases have 44.1 kHz sample rate and have been downsampled to 16 kHz and converted to Windows Pulse Coded Modulation (PCM) files. The database cannot be distributed, but the recordings and metadata are available online (Metadata search https://bdb.bertsozale.eus/en/web/bertsoa/bilaketa) (Examples of signals contained in the Bertsolaritza database can be accessed at https://bdb.bertsozale.eus/en/web/bertsoa/view/14sdqg).

In the Bertsolaritza database, there are no mixed segments, i.e., all the frames in each segment identified by the VAD belong to the same class. As stated above, in the NUS database, each segment contains frames of only one class also.

As expected, singing speech segments have longer durations than spoken speech segments in both databases (5.03 ± 2.67 and 2.30 ± 1.74 s, respectively, in the Bertsolaritza database and 3.83 ± 2.21 and 1.92 ± 1.06 s in the NUS database). The distributions of segment durations in both databases are shown in Figure 1a,b. In the NUS database, the linguistic content is the same in singing and speech segments. On the contrary, in the Bertsolaritza database, linguistic content is different and both classes are heavily unbalanced. Nevertheless, both databases exhibit the similar behaviour with respect to duration distribution.



(**a**) (**b**)

Figure 1. Distribution of singing and speech segment durations. (**a**) Bertsolaritza database; (**b**) NUS database.

2.2. Proposed Segmentation System

The scheme of our speech/singing segmentation system is shown in Figure 2. In the first step, a GMM-HMM VAD is applied to the Mel Frequency Cepstral Coefficient (MFCC) features to locate voice segments. Then, the pitch contour is extracted and a smoothing process is applied on it to remove vibrato effects. Using the smoothed contour, we detect note areas using our algorithm described in Section 2.2.2 and we calculate the voicing and note percentages in each segment. Finally, a statistical classifier is used to assign the correct class to these percentage parameters.

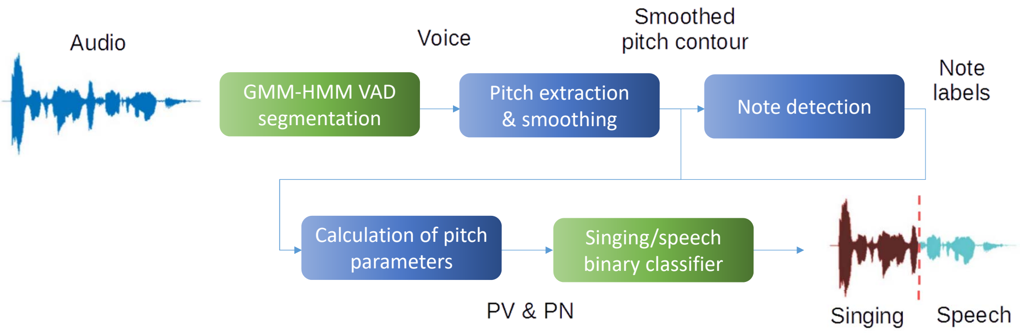


Figure 2. Structure of the proposed speech/singing voice segmentation system.

2.2.1. GMM-HMM Based VAD

In the VAD of the proposed system, three possible classes are defined as output in each acoustic frame: voice, applause or silence. We decided to split applause and silence in two separate classes because applause is common in our Bertsolaritza database and the acoustic nature of both classes is very different. The classification of recordings is made using a frame-level GMM with an HMM post-smoothing [25]. We used 13 MFCC values with Δ and Δ^2 values calculated applying a 25 ms window and 10 ms frame period. For the initial frame classification, independent GMMs are trained per each class using Expectation Maximization [26]. Frames are classified using these GMMs, but this frame level classification can create fast label changes that do not fit well to the data. This is why we used an HMM to smooth the resulting sequences. The HMM has fixed transition probabilities and forces the frame labels to remain in the same class for minimum durations depending on the likelihood values of the GMMs. We used a probability of 0.0001 outside the transition matrix diagonal for this purpose. To classify the segments, the likelihood of observation provided by each model is calculated using expression (1):

$$P(o|s_i) = \sum_{j=1}^{M} w_{ij} N(o|\mu_{ij}, \Sigma_{ij}), \tag{1}$$

where o is the MFCC vector, w_{ij}, μ_{ij}, Σ_{ij} are the weight, mean and diagonal covariance of the component j of the state s_i and M is the number of Gaussian components.

2.2.2. Note Detection Algorithm

Our algorithm to detect note areas uses the premise that music is a sequence of tones that need minimum stability and duration to be noticed as such. Compared with a state-of-the-art algorithm like Tony [27], our method is much simpler, but, as we do not have any annotated data, we needed to devise a method that required the minimum supervision.

The first task is to map the pitch curve to cent scale [28,29] taking as origin the lowest note that we consider may appear in the recordings. We use expression (2) to do the mapping:

$$f_{0c} = 1200 log_2(\frac{f_o}{f_{ref}}) + 5800, \tag{2}$$

where f_{ref} is 440 Hz, the frequency of A_4 note.

In a cent scale, a smoothing is applied to the pitch curve to neutralise vibrato variation within the notes. The smoothed curve is created averaging interpolation curves of local maxima and minima. Then, the smoothed curve is discretised to the closest ideal semitone as shown in Figure 3. As a result of this process, the real smoothed f_0 curve (in blue) is substituted by the discretised curve (in orange).

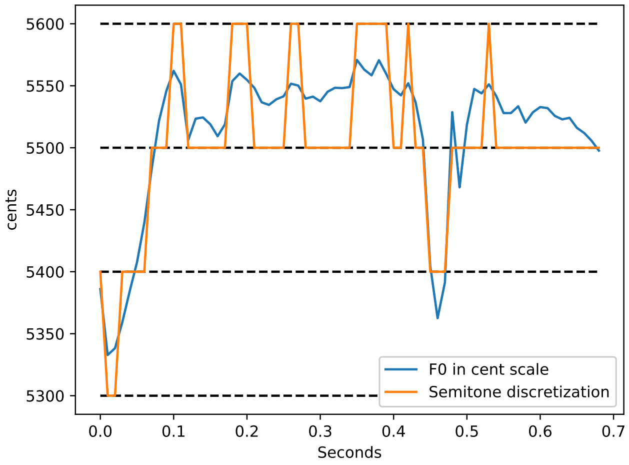


Figure 3. Discretisation of a f_0 curve.

To detect notes in the discretised curve, we apply an algorithm that uses subsequence search techniques [30,31]. We search sequences in the discrete curve that fulfil the minimum conditions of length and stability defined in expressions (3) and (4):

$$Len(s) \geq L, \tag{3}$$

$$max(s) - min(s) \leq R, \tag{4}$$

where s is the semitone subsequence, R is the maximum amplitude range and L is the minimum length. The detailed steps of the proposed algorithm are defined in Algorithm 22.

In Figure 4, we can see how a note detected by the algorithm would look (green line) and how the rest of the voice sequence is split into two smaller sequences. These two new sequences will be added to the sequence list to be analysed recursively if they meet the duration requirements.

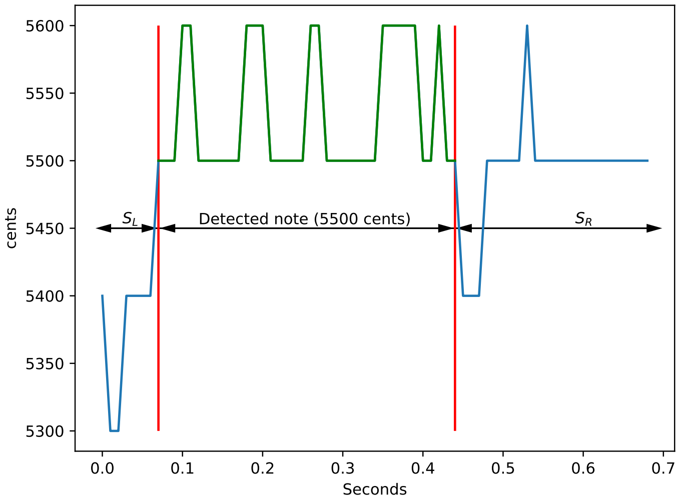


Figure 4. Detected musical note and new split sequences.

Algorithm 1: Note detection.

```
Data: Sequence of K voiced sequences of contiguous semitones
Result: Note list
begin
    $S \leftarrow$ Sequence of K voiced sequences of contiguous semitones $S = \{S_1, S_2, ..., S_K\}$;
    $L \leftarrow$ Minimum note length;
    $R \leftarrow$ Maximum note variation;
    $N \leftarrow$ Empty Note list;
    while $length(S) > 0$ do
        $S' \leftarrow$ empty New sequences list;
        for $S_i$ in $S$ do
            Find longest $s$ that fits $Len(s) \geq L$ and $max(s) - min(s) \leq R$;
            Save $s$ in N;
            $S_{Li} \leftarrow$ Sequence left to $s$ in $S_i$;
            $S_{Ri} \leftarrow$ Sequence right to $s$ in $S_i$;
            if $Len(S_{Li}) \geq L$ then
                Include $S_{Li}$ in $S'$
            end
            if $Len(S_{Ri}) \geq L$ then
                Include $S_{Ri}$ in $S'$
            end
        end
        $S \leftarrow S'$;
    end
end
```

2.2.3. Speech/Singing Classification

In our database, the separation of speech and singing voice segments is clear, i.e., there are no adjacent boundaries between the two classes. To create a singing/speech segmentation system, we would have to artificially create new recordings that fulfil this condition and we considered that this would create undesired artefacts. This is why we address the problem as a binary classification of the voice segments detected by the VAD. The pitch parameters we propose for the classification of each segment are: proportion of voiced segments (PV) and percentage of pitch labelled as a musical note (PN). The pitch curve has been calculated using PRAAT autocorrelation method [32] with a frame period of 10 ms.

Voiced/unvoiced segments are obtained directly in the pitch curve where the relative value of maximum autocorrelation is used to take this decision. Stable musical note segments are found using our algorithm explained in Section 2.2.2. The features for classification are calculated according to expressions (5) and (6):

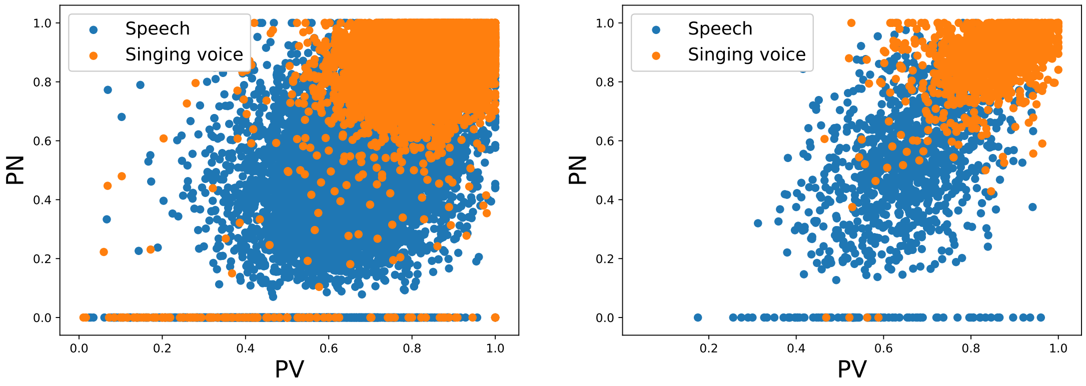
$$PV = \frac{N_{VF}}{N_T}, \tag{5}$$

$$PN = \frac{N_{NF}}{N_{VF}}, \tag{6}$$

where N_{VF} is the total number of voiced frames, N_{NF} is the total number of frames labelled as a musical note and N_T is the total number of frames, all of them calculated within the segment to be classified.

Figure 5a,b show the distribution of the proposed classifying features PV and PN in Bertsolaritza and NUS databases. In both cases, speech presents a more scattered distribution than a singing voice. However, good discrimination can be achieved when considering both parameters at the same time.

As a final step of the proposed algorithm, a classifier has to be trained with the vector containing these two parameters to obtain the final speech/singing classification.



(**a**) (**b**)

Figure 5. Distribution of PV and PN parameters. (**a**) Bertsolaritza database; (**b**) NUS database.

3. Results

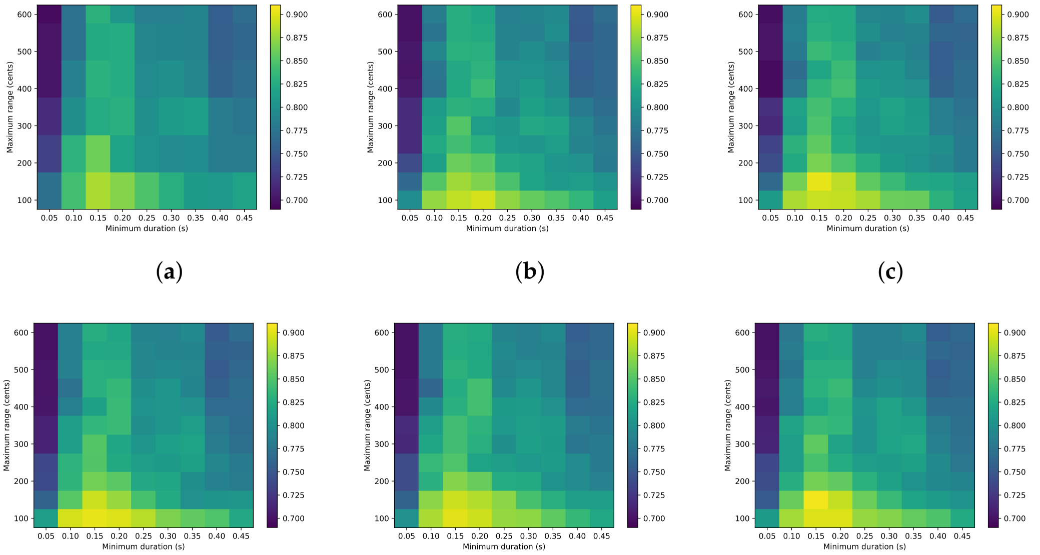
3.1. Parameter Sensitivity of the Note Detection Algorithm

The note detection algorithm we explained in Section 2.2.2 has flexibility with respect to the characteristics of the notes it is detecting, namely different length and range values can be taken into account. In addition, different steps can be considered within each semitone. In our previous work [33], we used ideal semitones of a 440 A_4 tuned scale to discretise the f_0 curve because we think this procedure ensures enough definition to analyse the stability of the curve. However, the algorithm in Tony uses three steps per semitone for more precision in the detection of note levels [27]. This prompted us to include the number of semitone steps as a variable to be considered together with the minimum length and maximum pitch range. In order to check if these configuration parameters have an important impact in the detected notes, we devised an experiment to test the sensitivity of the results to these parameters.

In this experiment, we applied our note detection algorithm with different minimum length, maximum range and steps per semitone to later discriminate speech and singing using the detected notes. We considered maximum ranges from 100 to 600 cents with intervals of 50 cents and minimum lengths from 50 to 450 ms with intervals of 50 ms. The dataset used is a Bertsolaritza database excerpt that was also used in [24]. To evaluate the classification, we used a 10-fold cross-validation with a joint F-score test. The F-score results with different note detection parameters are shown in Figure 6. We also tested a method that does not discretise the pitch curve before sequence searching and finds sequences using directly the values of the f_0 curve. The F-scores obtained with this method are shown in Figure 7.

Figure 6 shows that, when the note algorithm parameters get closer to the values common in Western music [34] (minimum duration in the range of 100–200 ms and maximum pitch range between 100 and 150 cents), the discrimination between singing and speech improves. This means that the optimum value for the parameters has a strong relation with the style of singing that we want to discriminate from normal speech. The vertical resolution in Figure 6a is half of the resolution shown in the rest of the cases because one step per semitone does not allow for having different results between semitones.

We also tested the parameter setting where the possible semitone steps are infinite. This means that each sample in the f_0 curve is a possible semitone, creating a continuity in the sequence search that can completely ignore the discrete semitone scale. This option can help in the case of analysing out of tune singing. The results obtained with continuous semitone steps are shown in Figure 7, where a similar pattern to the one observed in the case of using the discretised f_0 curve can be seen: the best F-score is obtained for the common values of minimum duration and maximum pitch range in Western music.



(**a**) (**b**) (**c**)

(**d**) (**e**) (**f**)

Figure 6. Speech/singing classification F-score for different number of steps per semitone. (**a**) one step per semitone; (**b**) two steps per semitone; (**c**) three steps per semitone; (**d**) four steps per semitone; (**e**) five steps per semitone; (**f**) six steps per semitone.

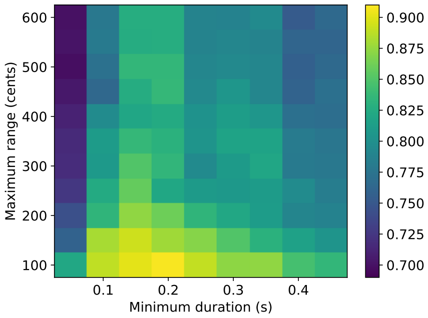


Figure 7. Speech/singing classification F-score for continuous f_0.

The parameter combination that provides the best discrimination results for each of the different semitone step number is presented in Table 3. F-score improves as we use more semitone steps and the best result is obtained with continuous f_0, which corresponds to a situation where infinite steps are considered. Nevertheless, the differences in F-score are slight and, providing that the maximum pitch range and minimum note length have values suitable for Western music, the number of steps considered has a small influence in the results. To apply the algorithm to our recordings, we have defined the maximum range of the notes as 100 cents, the minimum length as 150 ms, and we have considered four steps per semitone.

Table 3. Best result of each semitone step division level.

Steps per Semitone	Minimum Duration (s)	Maximum Range (Cents)	F-Score
1	0.15	100	0.882
2	0.2	100	0.896
3	0.15	150	0.902
4	0.15	100	0.902
5	0.15	100	0.900
6	0.15	150	0.905
Continuous	0.2	100	**0.908**

3.2. Comparison with a Standard Note Detection Algorithm

We have compared our note detection system with Tony, a state-of-the-art algorithm that uses HMMs with note onset, stable and offset states to detect notes in multiple pitch tracks calculated emphasising different frequency ranges. As our speech/singing discrimination algorithm only uses the 'note/no note' decision, we have compared precisely this aspect in the two algorithms. With this purpose, we have used the Bertsolaritza database explained in Section 2.1. We have labelled 10 ms spaced frames with a binary label (note/no note) with both algorithms and calculated the agreement between systems using the kappa score [35] in each audio file. The histogram of these scores is shown in Figure 8.

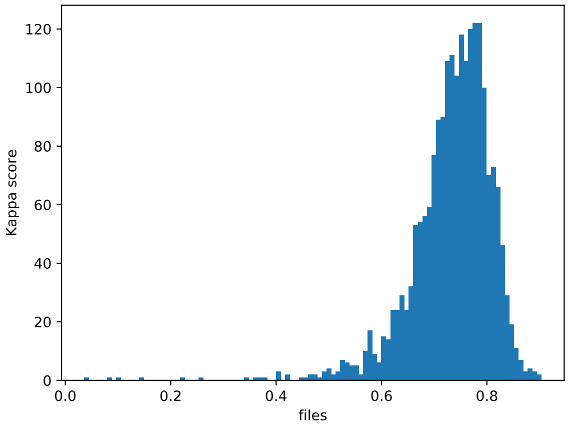


Figure 8. Histogram of kappa score between notes detected by Tony and our algorithm.

We can observe that the agreement between both note detection systems is strong for most files in the database, with a mean kappa of 0.73 ± 0.08.

3.3. Speech and Singing Discrimination

We have compared our algorithm with other methods training them in our Bertsolaritza database and using the NUS database for test purposes. We have selected methods that are suitable to work with segments of different duration as it is the case of Bertsolaritza database. We have trained three different GMM classifiers: one with the first difference of f_0 as suggested in [1] (Δf_0), another with the distribution of the Discrete Fourier Transform (DFT) of f_0 [12] (DFT-f_0) and the last one with MFCC parameters (MFCC). The details of the calculation of these methods are explained in [33]. In addition, we have also considered using note labels from Tony instead of our note detection algorithm for the calculation of PN parameter (Tony). In the proposed system, the statistical classifiers used as last stage have been the SVM [36,37] and the Extreme Learning Machine (ELM) [38,39]. None of these methods consider parameters related with the duration of the segment to classify, even in an indirect way. They use normalised histograms, percentage values and majority votes, which are parametrizations that do not depend on the duration of the segment and allow the classification of segments with different length.

The performance of the classifiers has been measured using unweighted Precision, Recall and F-score, defined as the macro-averaged measure parameter for each of the classes. We used unweighted mean of the score because our Bertsolaritza database is heavily imbalanced, with more singing segments than speech segments as seen in Section 2. Macro-averaging considers all classes equally and, in the case of imbalanced datasets, is more convenient [40].

We have used the 5x2cv paired t test [41] with the structure shown in Figure 9 to compare the systems. To achieve this, we split the Bertsolaritza database in 10 sections with the only condition that all the segments of each speaker had to be in a single section. After this, we made five iterations of splitting the database in two blocks of five random sections with no block repetition. In each of the iterations, we created a first score by using one block as train and the other as test. A second score is calculated by rotating test and train, training set becomes test and vice versa. This gives us 10 scores in total (two per iteration) that are used to calculate the means and variance values presented in Tables 4 and 5 and the p scores in Table 6. The procedure ensures that no speaker is present both in training and test block in any iteration.

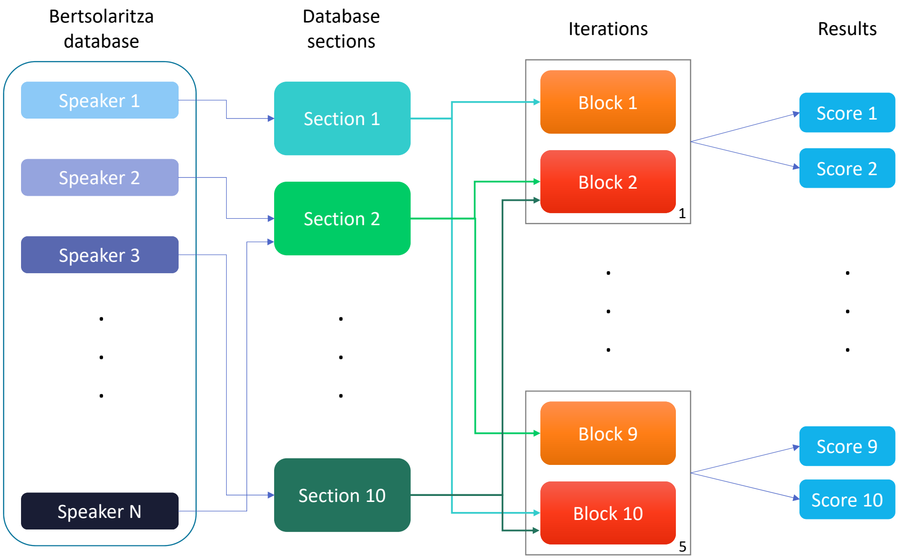


Figure 9. 5x2cv test structure.

To test the generalisation capability of the algorithm, the NUS database has been classified using the algorithms trained with all the Bertsolaritza database and without any special adaptation. The results are shown in Tables 4 and 5 for the classification of singing and speech, respectively. Results using a SVM classifier are presented for our proposed method because it performed better than the EML classifier. GMM trained with MFCCs is the best method in the Bertsolaritza database, although differences in results with our proposed method are not statistically significant (see Table 6). When we apply this model to NUS database, it gets very good results for the singing class, comparable to our method. In the Bertsolaritza database, singing and speech are mixed in the same file, while, in the NUS database, each file contains only one type of voice. In the MFCC method, file-wise mean and variance normalisation is applied for the calculation of the MFCCs. If there are enough data with high speaker variability in the training database, the MFCC method can learn the characteristics of both classes and generalise to other databases. However, if data are not enough, this generalisation is not good [33]. The pitch-based methods Δf_0 and DFT of f_0 do not get good results compared with our system. We think that this is due to the presence of short voice segments in the database. The methods are suitable when long voice segments are present. The proposed algorithm and the one that uses notes detected by Tony have similar scores to the MFCC-GMM method in the Bertsolaritza database and all of them generalise well enough and get similar results in the NUS database. Although F-score results are similar, Precision and Recall values are not. MFCC and Tony methods get good recall for singing and therefore good precision for speech, but not so good values for singing precision and speech recall. On the contrary, our algorithm gets more balanced precision and recall values. The algorithm of Tony is designed to analyse singing voice and not speech. Therefore, it is likely that it has a strong bias towards a singing class when identifying notes. In the case of the MFCC method, the bias is due to the unbalance in favour of the singing class in the training data.

Table 4. Results of singing classification.

Method	**Precision**		**Recall**		**F-Score**	
	Bertsolaritza	**NUS**	**Bertsolaritza**	**NUS**	**Bertsolaritza**	**NUS**
Δf_0	0.95 ± 0.00	0.79	0.85 ± 0.03	0.73	0.90 ± 0.01	0.76
DFT-f_0	0.93 ± 0.00	0.76	0.83 ± 0.01	0.77	0.88 ± 0.01	0.76
MFCC	**0.99** ± **0.00**	0.88	**0.97** ± **0.00**	**0.97**	**0.98** ± **0.00**	**0.92**
Tony	0.97 ± 0.00	0.83	0.92 ± 0.01	0.95	0.94 ± 0.01	0.88
Proposed	0.98 ± 0.00	**0.92**	0.96 ± 0.00	0.92	0.97 ± 0.00	**0.92**

Table 5. Results of speech classification.

Method	**Precision**		**Recall**		**F-Score**	
	Bertsolaritza	**NUS**	**Bertsolaritza**	**NUS**	**Bertsolaritza**	**NUS**
Δf_0	0.58 ± 0.04	0.72	0.81 ± 0.01	0.78	0.68 ± 0.03	0.75
DFT-f_0	0.53 ± 0.01	0.73	0.76 ± 0.01	0.72	0.63 ± 0.01	0.73
MFCC	**0.90** ± **0.01**	**0.96**	**0.97** ± **0.00**	0.85	**0.93** ± **0.01**	0.90
Tony	0.73 ± 0.04	0.93	0.88 ± 0.00	0.78	0.80 ± 0.02	0.85
Proposed	0.87 ± 0.02	0.91	0.94 ± 0.00	**0.91**	0.90 ± 0.01	**0.91**

To assess the statistical significance of the results, we calculated the p-value for the results of all the alternative systems when compared with our proposed system. Table 6 shows the p-values for speech and singing detection. Considering these values and a significance level of $\alpha = 0.05$, the differences in performance of the proposed system is statistically not significant comparing to the MFCC system and it is significant comparing it with the rest of the systems.

Table 6. p-value of the results of the proposed optimized algorithm compared with the rest of the systems.

Method	p **(Singing)**	p **(Speech)**
Δf_0	8.099×10^{-5}	2.858×10^{-5}
DFT-f_0	0.010	0.001
MFCC	0.084	0.063
Tony	0.004	0.001

3.4. Analysis of Computation Times

We measured the time needed by each speech/singing discrimination method to train and classify the Bertsolaritza database using 5x2cv cross-validation. The processes have been run in an Intel Xeon CPU E5-2660 v2. The results obtained are shown in Table 7.

We can see that all processing times are comparable, except for the one of the GMM built with MFCC. Considering that the proposed system has the best classification results overall, we can see that the better results achieved by the MFCC system are produced at the expense of bigger dimensionality and computation time.

Table 7. Computation times for training and classifying in the Bertsolaritza database.

Method	Train Duration	Classification Duration
Δf_0	**0:04:03**	0:03:59
DFT-f_0	0:04:15	0:03:51
MFCC	7:12:33	0:17:04
Tony	0:04:31	0:04:13
Proposed	0:04:08	**0:03:46**

4. Discussion

In this article, we present a novel method of discriminating speech and singing segments using only two parameters derived from f_0. The system has been tested and compared with two systems based also on f_0 analysis and another one based on spectral information in two databases with different characteristics. The proposed method gets better results than the other systems using pitch parameters and equivalent results compared to the spectrum analysis system. It has been also proven that the proposed one generalises the classification in other databases without bias to the singing class and has a lower computation time.

We observed that the two parameters used for the classification in the proposed system, PV and PN, are very discriminative in databases with diverse characteristics: different languages, singing proficiency levels, recording conditions and quality, etc. These parameters are also capable of classifying voice segments of short length and this provides a high flexibility to the classification system.

As a by-product, an algorithm to detect notes in a singing voice has been created. The algorithm is easy and intuitive to tune. It only uses three configuration parameters that are directly related with the characteristics of a specific music style, in this case, the Western music. The test has shown that the algorithm is able to cope with audio files containing music and speech segments. It is capable of obtaining musical information from the singing voice without trying to classify speech as singing. Other existing note detection algorithms are tuned to deal with music signals and therefore they try to find notes in all the segments they process. On the contrary, our algorithm has been designed expressly for detecting notes in signals with speech and music and it is not biased to any of the classes. We compared our note detection algorithm with the state-of-the-art algorithm of Tony and we have seen that the agreement between the two methods regarding note detection is strong.

The parameter sensitivity analysis of the note detection algorithm showed that changes in the minimum duration have more effect on the speech/singing classification results than the maximum pitch range. If the maximum pitch range considered is high, all the pitch variations in the signal would be included in the range and then the only parameter with influence on the results has a minimum length. When the minimum length considered is very small, the pitch variation inside the segment usually is not very large and then all segments are treated as musical notes, which gives poor speech/singing classification results regardless of the maximum pitch range applied. In general, the singing voice has a greater pitch range than speech, but it has smaller local pitch changes once the vibrato has been smoothed in the pitch curve.

For future work, we consider it interesting to apply the note labelling algorithm to different music styles and datasets and find the optimal values for the configuration parameters valid for other singing styles. It would be also very interesting to develop new objective measures for the evaluation of note detection algorithms. In addition, we are considering the possibility to test the potential of LSTMs for performing the note detection.

Author Contributions: Conceptualization, X.S., D.T. and E.N.; Methodology, X.S. and L.S.; Software, X.S. and D.T.; Validation, X.S. and E.N.; Formal analysis, X.S., D.T. and L.S.; Investigation, X.S.; Resources, X.S., L.S. and I.S.; Data curation, X.S.; Writing—original draft preparation, X.S. and E.N.; Writing—review and editing, X.S., E.N., I.S. and I.H.; Visualization, X.S. and E.N.; Supervision, E.N., I.S. and I.H.; Project administration, E.N., I.S. and I.H.; Funding acquisition, E.N., I.S. and I.H.

Funding: This research has been partially supported by UPV/EHU (Ayudas para la Formación de Personal Investigador), the Spanish Ministry of Economy and Competitiveness with FEDER support (MINECO/FEDER, UE) (RESTORE project, TEC2015-67163-C2-1-R) and by the Basque Government under grant KK-2018/00014 363 (BerbaOla) and IT355-19.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BLSTM	Bidirectional Long Short-Term Memory
DFT	Discrete Fourier Transform
GMM	Gaussian Mixture Models
HMM	Hidden Markov Models
LSTM	Long Short-Term Memory
MFCC	Mel Frequency Cepstral Coefficients
PCM	Pulse Coded Modulation
SVM	Support Vector Machine
VAD	Voice Activity Detector

References

1. Ohishi, Y.; Goto, M.; Itou, K.; Takeda, K. Discrimination between Singing and Speaking Voices. In Proceedings of the 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, 4–8 September 2005; pp. 1141–1144.
2. Deutsch, D.; Lapidis, R.; Henthorn, T. The speech-to-song illusion. *Acoust. Soc. Am. J.* **2008**, *124*, 2471. [CrossRef]
3. Falk, S.; Rathcke, T. On the speech-to-song illusion: Evidence from German. In Proceedings of the Speech Prosody 2010—Fifth International Conference, Chicago, IL, USA, 10–14 May 2010.
4. Mesaros, A.; Virtanen, T. Automatic recognition of lyrics in singing. *EURASIP J. Audio Speech Music Process.* **2010**, *2010*, 1–11. [CrossRef]
5. Loscos, A.; Cano, P.; Bonada, J. Low-Delay Singing Voice Alignment to Text. In Proceedings of the 1999 International Computer Music Conference (ICMC), Beijing, China, 22–27 October 1999.
6. Saunders, J. Real-time discrimination of broadcast speech/music. In Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, Atlanta, GA, USA, 9 May 1996; Volume 2, pp. 993–996.
7. Carey, M.J.; Parris, E.S.; Lloyd-Thomas, H. A comparison of features for speech, music discrimination. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP99) (Cat. No.99CH36258), Phoenix, AZ, USA, 15–19 March 1999; Volume 1, pp. 149–152.
8. El-Maleh, K.; Klein, M.; Petrucci, G.; Kabal, P. Speech/music discrimination for multimedia applications. In Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100), Istanbul, Turkey, 5–9 June 2000; Volume 4, pp. 2445–2448.
9. Khonglah, B.K.; Prasanna, S.M. Speech/music classification using speech-specific features. *Digit. Signal Process.* **2016**, *48*, 71–83. [CrossRef]
10. Gerhard, D. Pitch-based acoustic feature analysis for the discrimination of speech and monophonic singing. *Can. Acoust.* **2002**, *30*, 152–153.
11. Schuller, B.; Rigoll, G.; Lang, M. Discrimination of speech and monophonic singing in continuous audio streams applying multi-layer support vector machines. In Proceedings of the 2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763), Taipei, Taiwan, 27–30 June 2004. [CrossRef]
12. Thompson, B. Discrimination between singing and speech in real-world audio. In Proceedings of the 2014 IEEE Spoken Language Technology Workshop (SLT), South Lake Tahoe, NV, USA, 7–10 December 2014; pp. 407–412. [CrossRef]
13. Tsai, W.H.; Ma, C.H. Speech and singing discrimination for audio data indexing. In Proceedings of the 2014 IEEE International Congress on Big Data, Anchorage, AK, USA, 27 June–2 July 2014; pp. 276–280. [CrossRef]

14. Schuller, B.; Schmitt, B.J.B.; Arsić, D.; Reiter, S.; Lang, M.; Rigoll, G. Feature selection and stacking for robust discrimination of speech, monophonic singing, and polyphonic music. In Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands, 6 July 2005; pp. 840–843. [CrossRef]
15. Kum, S.; Nam, J. Joint Detection and Classification of Singing Voice Melody Using Convolutional Recurrent Neural Networks. *Appl. Sci.* **2019**, *9*, 1324. [CrossRef]
16. Hsu, C.; Wang, D.; Jang, J.R.; Hu, K. A Tandem Algorithm for Singing Pitch Extraction and Voice Separation From Music Accompaniment. *IEEE Trans. Audio Speech Lang. Process.* **2012**, *20*, 1482–1491. [CrossRef]
17. Chan, T.; Yeh, T.; Fan, Z.; Chen, H.; Su, L.; Yang, Y.; Jang, R. Vocal activity informed singing voice separation with the iKala dataset. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Queensland, Australia, 19–24 April 2015; pp. 718–722. [CrossRef]
18. Lehner, B.; Widmer, G.; Bock, S. A low-latency, real-time-capable singing voice detection method with LSTM recurrent neural networks. In Proceedings of the 23rd European Signal Processing Conference (EUSIPCO), Nice, France, 31 August–4 September 2015; pp. 21–25. [CrossRef]
19. Leglaive, S.; Hennequin, R.; Badeau, R. Singing voice detection with deep recurrent neural networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, Australia, 19–24 April 2015; pp. 121–125. [CrossRef]
20. Ono, N.; Miyamoto, K.; Le Roux, J.; Kameoka, H.; Sagayama, S. Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram. In Proceedings of the IEEE 16th European Signal Processing Conference, Lausanne, Switzerland, 25–29 August 2008; pp. 1–4.
21. Hsu, C.L.; Jang, J.S.R. On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset. *IEEE Trans. Audio Speech Lang. Process.* **2009**, *18*, 310–319.
22. Black, D.A.; Li, M.; Tian, M. Automatic identification of emotional cues in Chinese opera singing. In Proceedings of the International Conference on Music Perception and Cognition and Conference for the Asian-Pacific Society for Cognitive Sciences of Music, Seoul, Korea, 4–8 August 2014.
23. Duan, Z.; Fang, H.; Li, B.; Sim, K.C.; Wang, Y. The NUS sung and spoken lyrics corpus: A quantitative comparison of singing and speech. In Proceedings of the 2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, Kaohsiung, Taiwan, 29 October–1 November 2013; pp. 1–9. [CrossRef]
24. Sarasola, X.; Navas, E.; Tavarez, D.; Erro, D.; Saratxaga, I.; Hernaez, I. A Singing Voice Database in Basque for Statistical Singing Synthesis of *bertsolaritza*. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, 23–28 May 2016.
25. Hain, T.; Woodland, P.C. Segmentation and classification of broadcast news audio. In Proceedings of the Fifth International Conference on Spoken Language Processing, Sydney, Australia, 30 November–4 December 1998.
26. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. (Methodol.)* **1977**, *39*, 1–22. [CrossRef]
27. Mauch, M.; Cannam, C.; Bittner, R.; Fazekas, G.; Salamon, J.; Dai, J.; Bello, J.; Dixon, S. Computer-aided melody note transcription using the Tony software: Accuracy and efficiency. In Proceedings of the First, International Conference on Technologies for Music Notation and Representation, Paris, France, 28–30 May 2015; pp. 23–30.
28. Ellis, A.J. On the musical scales of various nations. *J. Soc. Arts* **1885**, *1688*, 485–532.
29. Benson, D.J. Music: A mathematical offering. *Math. Intell.* **2008**, *30*, 166.
30. Moon, Y.S.; Kem, J. Fast normalization-transformed subsequence matching in time-series databases. *Trans. Inf. Syst.* **2007**, *E90-D*, 2007–2018 . [CrossRef]
31. Kostakis, O.K.; Gionis, A.G. Subsequence Search in Event-Interval Sequences. In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15), Santiago, Chile, 9–13 August 2015; pp. 851–854. [CrossRef]
32. Boersma, P. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In Proceedings of the Institute of Phonetic Sciences, Amsterdam, The Netherlands, 14–17 August 2017; Volume 17, pp. 97–110.

33. Sarasola, X.; Navas, E.; Tavarez, D.; Serrano, L.; Saratxaga, I. Speech and monophonic singing segmentation using pitch parameters. In Proceedings of the IberSPEECH 2018, Barcelona, Spain, 21–23 November 2018; pp. 147–151.
34. Bregman, A.S. *Auditory Scene Analysis: The Perceptual Organization of Sound*; MIT Press: Cambridge, MA, USA, 1994.
35. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
36. Suykens, J.A.; Vandewalle, J. Least squares support vector machine classifiers. *Neural Process. Lett.* **1999**, *9*, 293–300. [CrossRef]
37. Scholkopf, B.; Smola, A.J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*; MIT Press: Cambridge, MA, USA, 2001.
38. Huang, G.; Huang, G.B.; Song, S.; You, K. Trends in extreme learning machines: A review. *Neural Netw.* **2015**, *61*, 32–48. [CrossRef] [PubMed]
39. Salerno, V.; Rabbeni, G. An extreme learning machine approach to effective energy disaggregation. *Electronics* **2018**, *7*, 235. [CrossRef]
40. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [CrossRef]
41. Dietterich, T.G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.* **1998**, *10*, 1895–1923. [CrossRef] [PubMed]