IberSPEECH 2018
21-23 November 2018, Barcelona, Spain

Speech and monophonic singing segmentation using pitch parameters

Xabier Sarasola, Eva Navas, David Tavarez, Luis Serrano, Ibon Saratxaga

AHOLAB University of the Basque Country (UPV/EHU), Bilbao, Spain

{xsarasola,eva,david,lserrano,ibon}@aholab.ehu.eus

Abstract

In this paper we present a novel method for automatic segmentation of speech and monophonic singing voice based only on two parameters derived from pitch: proportion of voiced segments and percentage of pitch labelled as a musical note. First, voice is located in audio files using a GMM-HMM based VAD and pitch is calculated. Using the pitch curve, automatic musical note labelling is made applying stable value sequence search. Then pitch features extracted from each voice island are classified with Support Vector Machines. Our corpus consists in recordings of live sung poetry sessions where audio files contain both singing and speech voices. The proposed system has been compared with other speech/singing discrimination systems with good results.

Index Terms: audio segmentation, voice discrimination, singing voice, pitch

1. Introduction

There have been many studies dealing with the analysis and development of technologies for speech. Lately there is also an interest in studying singing speech. Both areas are intimately related, but it is not possible to apply directly the same techniques to both problems. For instance, phone alignment systems with good performance for speech do not get good results for singing voice [1]. Also automatic recognition systems degrade their performance when dealing with singing speech [2]. Despite this different behaviour with speech technologies both types of voice share the same nature and therefore are very close. In fact, singing voice is considered an extension of common speech and sometimes it is not easy to distinguish them. Even for humans it is sometimes difficult to correctly discriminate speech from singing voices [3], [4]. Nevertheless, there is an agreement about some characteristics that behave differently between speech and singing [5], [6]:

- Voiced/unvoiced ratio is higher in singing voice, as vowels tend to be longer to accommodate to the duration of notes.
- Dynamic ranges of energy and F0 in singing voice are higher than in speech.
- Vibrato areas can appear in long sustained notes in singing voice. This phenomenon never occurs in speech.
- The pitch continuity is different in speech and singing voice. Pitch in singing voice is more similar to a series of discrete values and on the contrary in speech it is a continuously varying curve.

Two types of automatic speech/singing classification systems have been proposed. The first scheme consists on using short-time features of the signal, like for instance spectral information in the form of MFCCs, to classify it at frame level [7]. In [8], a wide range of short-time features are analysed for frame level classification of polyphonic music, singing voice and speech. The second approach uses long-term features for classification. In previous works, distribution of pitch values [9], pitch parameters [10] and note grammars [11], [12] are used. These long term features can be calculated for the whole file or for a window that slides through the audio file. The former case corresponds to speech/singing classification and the later to segmentation. For the segmentation problem, the results improve using long windows up to 1000 ms [13].

Bertsolaritza is a live improvisation poetry art from Basque Country. In these live sessions a host introduces the singer and proposes the topic for the improvised verses that will be sung a capella. The singers are professional verse creators, but not professional singers. Many live sessions of bertsolaritza are recorded and saved together with the corresponding transcriptions by Bertso associations. All these recordings are very useful for analysis of the singing style and as training database for bertso synthesis. Each recording includes speech from the host, sung verses and overlapped applauses. Thus, a good segmentation system is required to isolate the segments of interest.

Our goal is to create a tool to automatically segment the bertsolaritza audio files to get the singing voice.

The rest of the paper is organized as follows. Section 2 explains the proposed segmentation system using only two parameters derived from pitch. Section 3 details the algorithm developed to assign a musical note label to each audio frame. Section 4 describes the results of the segmentation system and compares it to other speech/singing voice classification systems. Finally, Section 5 presents the conclusions of the work.

2. Proposed segmentation system

The main scheme of the proposed system can be seen in Figure 1. First voice is detected in the audio files using a voice activity detection (VAD) algorithm based on Hidden Markov Models and Gaussian Mixture Models (GMM-HMM). Pitch contour is calculated and for each voiced segment pitch parameters are extracted. These parameters are classified as speech or singing by a previously trained Support Vector Machine (SVM).

2.1. GMM-HMM based VAD

The first stage of the proposed system divides the audio in three classes: voice, noise and silence. For this purpose we have used a GMM frame-level classifier smoothed with an HMM as used in [14]. We have considered 13 MFCC values and Δ and Δ^2 values with 25 ms window and 10 ms frame period. GMMs are trained using Expectation Maximization (EM) for speech, noise and silence. Frame-level classification usually creates noisy segmentation because in the transition segments fast label changes are produced. To avoid this, we have used an ergodic HMM of three states, one per class. This HMM has predefined transition probabilities with preference for remaining in the current state. In our case, the transition probability of the HMM

10.21437/IberSPEECH.2018-31

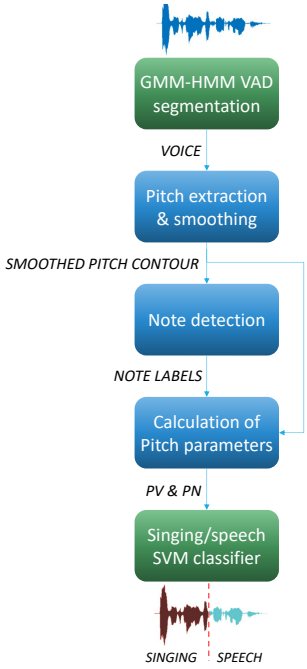

Figure 1: *Structure of the proposed speech/singing voice segmentation system*

outside the diagonal is 0.0001. This way, fast transitions and small discontinuities are removed. To classify the segments, the likelihood of observation provided by each model is calculated using expression (1).

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

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. Singing/speech classification

We have considered each audio segment labelled as voice by the GMM-HMM VAD as corresponding to a unique class: either speech or singing. We do not consider the option of having speech and singing in the same voice island because of the characteristics of our database. To evaluate a system that includes segmentation inside voice islands we would have to create and artificial audio files mixing our data. Therefore the problem is reduced to a binary classification of each voice island. We propose the use of two parameters derived from pitch to do the classification: 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 [15] with a frame period of 10 ms.

Voiced/unvoiced segments are obtained directly from the pitch curve and stable musical note segments are found using our algorithm explained in Section 3. The features for classification are calculated according to expressions (2) and (3).

$$PV = \frac{N_{VF}}{N_T} \quad (2)$$

$$PN = \frac{N_{NF}}{N_{VF}} \quad (3)$$

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. The vector containing these two parameters is classified using a SVM.

3. Note detection algorithm

Our algorithm to label the pitch curve with musical notes discretises the F0 curve in semitones expressed in cents and then searches for sequences of semitones that fulfil two conditions: to have enough duration and less variation range than a threshold. This algorithm is simpler than state-of-the-art algorithms [16] but our lack of labelled data made us create a method with minimum supervision. First we map the F0 value to cent scale with an offset to make all the possible values of f_{0c} positive according to expression (4).

$$f_{0c} = 1200 log_2(\frac{f_o}{f_{ref}}) + 5800 \quad (4)$$

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

To avoid possible instability due to vibrato, we apply a smoothing to the F0 curve. The smoothing consists on calculating the local maxima and minima envelopes and taking the average curve. The obtained smoothed pitch curve is rounded to the closest semitone value to discretise the sequence, as shown in Figure 2.

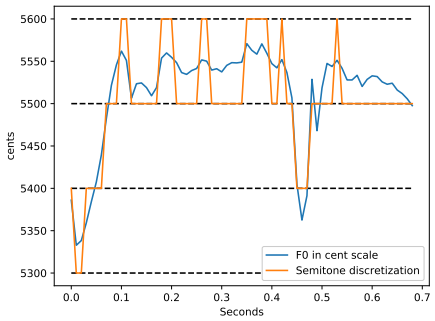

Figure 2: *Discretisation of F_0 curve*

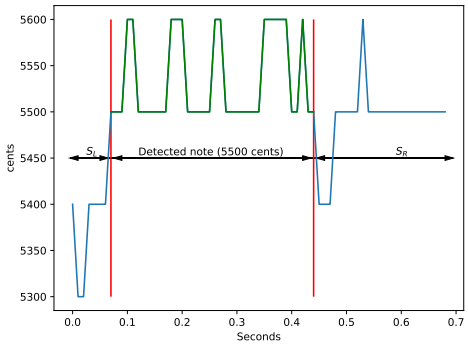

Figure 3: *Detected musical note and new split sequences*

Using the semitone discretised pitch curve we search for notes using subsequence search techniques [17], [18]. We

search the non-overlapping group of subsequences that fulfil the predefined conditions of minimum length and maximum range in amplitude expressed in equations 5 and 6.

$$Len(s) \geq L \tag{5}$$

$$max(s) - min(s) \leq R \tag{6}$$

where s is the semitone subsequence, R is the maximum amplitude range and L is the minimum length.

The algorithm is defined in the next steps:

- Consider the whole pitch curve as a collection of K voiced sequences of contiguous semitones expressed in cents each one with its own length $S = \{S_1, S_2, ..., S_K\}$.
- Define R as the maximum allowed variation range and L as the minimum length.
- Search the longest subsequence in each S_i $(1 \leq i \leq K)$ that fills the conditions.
- Label the longest subsequence found as a musical note. Between the possible semitones in the sequence, the most frequent one is selected as label.
- Split the remaining parts of the original sequence S_i in two new sequences: the subsequence in the left of the note found (S_{Li}) and the subsequence in the right of the note found (S_{Ri}) as shown in Figure 3.
- If any of the new generated subsequences fill the duration condition, S_i in S is substituted by them and the process begins again.
- When all sequences from S have been analysed the process finishes.

In this work we have established a maximum range of 100 cents and a minimum length of 150 ms, a standard minimum value in Western music [19].

4. Experiments and results

4.1. Datasets

As few publicly available database exist with speech and monophonic singing, we have used an excerpt of our Bertsolaritza database [20] to train the algorithms and the NUS Sung and Spoken Lyrics Corpus [21] to test them. In the Bertso database we manually labelled 20 audio files from 19 singers with a total duration of 60 minutes and 40 seconds. These audio files contain 32.8 minutes of singing voice and 2.87 minutes of speech. The 20 files were selected to cover the variability of the original Bertsolaritza database, considering recordings from different decades and gender. In the NUS database, each singer has recorded a singing and spoken version of 4 songs. The total number of different songs is 20 and 12 singers have made the recordings, 6 males and 6 females. As each recording contains either speech or singing voice, we used the VAD to obtain the voice segments and labelled them with the type of the recording.

Table 1 shows the distribution of singers and hosts by gender in the Bertso database. In most cases the speakers either sing or act as host, but some hosts give the topic for the improvised verses singing as well. These hosts appear in the recordings both singing and speaking.

In the Bertso database the audio files originally were in mp3. Both databases had 44100 Hz samplerate and have been downsampled to 16000 Hz and converted to Windows PCM files[1].

Table 1: *Number of speakers and singers in Bertso database*

	Singer	**Host**	**Singer and host**	**Total**
Female	7	6	2	15
Male	12	6	2	20
Total	19	12	4	

We analysed the Bertso database and the singing voice and speech segments never appear consecutively, i. e., there is always silence or noise between segments to classify. Therefore, considering the structure of both databases, each segment belongs only to a class. Additionally, we have studied the duration of the segments produced by speakers and singers: singing voice has longer durations than speech (mean duration of 3.69 and 1.51 seconds respectively in Bertso database and 3.87 and 1.92 seconds in NUS database).

The distribution of the proposed classifying features PV and PN in the databases can be seen in Figures 4 and 5. Speech is more scattered than singing voice, but taking into account both parameters a good discrimination of both classes can be achieved.

For the experiments, we split the Bertso database in 10 subsets for cross-validation tests. All the partitions considered include different singers in the train and test subsets. The NUS database is classified using the algorithms trained with Bertso database.

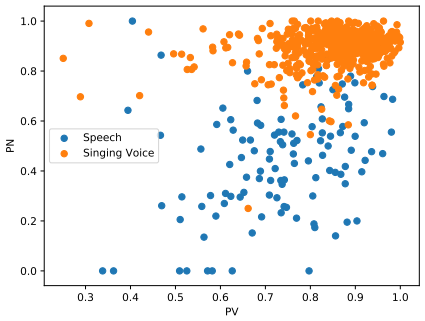

Figure 4: *Distribution of the classes in the Bertso database*

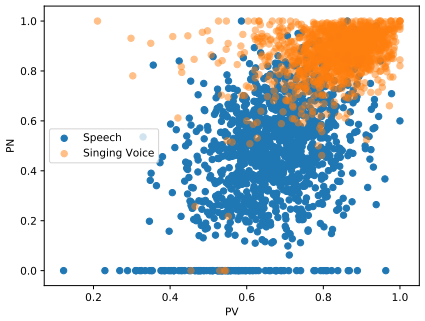

Figure 5: *Distribution of the classes in NUS database*

[1]Examples of signals contained in the Bertso database can be accessed at http://bdb.bertsozale.eus/en/web/bertsoa/view/1323t4.

4.2. Comparison with other methods

To compare our algorithm with other methods we had tested them in our Bertso database and in NUS database. We have selected methods that are suitable to work with segments of different duration as it is the case of Bertso database. On the one hand, we have trained GMM classifiers with the parameters suggested in [12] ($\Delta F0$) and [13] (DFT of F0 distribution). On the other hand, we have also built a GMM classifier based on MFCC parameters. These methods are explained with more detail in the following subsections.

4.2.1. DFT of F0 distribution

As commented before, F0 provides very useful information to discriminate between speech and singing. The histogram of F0 gives information about the range and the distribution of F0 values that for singing voice will be concentrated around the frequencies corresponding to musical notes. In [13] GMMs are used to model the DFT of the F0 distribution and detect the deviation of the instantaneous pitch value from its mean. F0 values in Bertso database range from 75 to 500 Hz. 100 bin histogram is calculated where each segment corresponds to approximately 0.027 octaves. The histogram is normalized to have unit area and then modelled using 8 component GMMs for singing voice and speech.

4.2.2. Delta F0

The dynamics of F0 are another feature to consider in speech and singing voice discrimination. In [12] the $\Delta F0$ distribution of voiced segments is modelled with GMMs to discriminate speech and singing voice. We calculate the $\Delta F0$ using a Savitsky-Golay filter [22] with a window of 50 ms. An histogram of 100 bins is made from -50 to 50 Hz. The distribution of $\Delta F0$ in each voice segment is normalized to have unit area and then modelled with 16 component GMMs.

4.2.3. MFCC GMM

In [12] short-term spectrum features are used motivated by the presence of an additional resonance characteristic of singing speech addressed in [23]. We calculated 13 MFCC coefficients and their Δ with a frame period of 10 ms and a window of 25 ms applying a CMVN file-wise normalization. MFCC frames of speech and singing voice are modelled using 32 component GMMs. Each voice segment is assigned the class that gets the higher sum of log-likelihood for all the frames of the segment. We chose GMMs to model MFCC parameters because of the high dimensionality of the parameters.

4.3. Results of GMM-HMM VAD

The metric used to assess the VAD is the voice detection F-score defined as indicated in equation 7.

$$F = \frac{2TP}{2TP + FP + FN} \tag{7}$$

where TP is the duration speech classified as speech, FP is the duration of non-speech classified as speech and FN is the duration of speech classified as non-speech.

Table 2 shows the results for different number of Gaussian components. All of them get good results, over 0.96, and the number of components does not affect the performance significantly. We have selected the VAD with 32 components for the classification experiments.

Table 2: *Results of the GMM-HMM VAD for different number of Gaussian components*

Gaussians	F-Score
2	0.965 +/- 0.005
4	0.967 +/- 0.006
8	0.969 +/- 0.007
16	0.972 +/- 0.008
32	0.973 +/- 0.008
64	0.974 +/- 0.008

4.4. Results of speech/singing classification

The metric in each test will be the unweighted F-score, defined as the average of F-score for each of the classes. This metric is suitable when the classes are unbalanced as it is the case of Bertso database. The results of the cross-validation classification experiments are shown in Table 3. The proposed method gets the best results and can compete with spectrum based methods. The methods that use pitch derived parameters as $\Delta F0$ and DFT of F0 distribution get poorer results due to the short duration of the segments to classify. Although in other works they have proved useful, they are not suitable for the characteristics of Bertso database. The experiment in the NUS database shows similar results proving the validity of our method with professional singers and a different style. The GMM method results got worse for NUS database, probably because the audio files used to generate the models contain both speech and singing and the files in NUS database belong to one class. Therefore the normalization process affects both databases differently.

Table 3: *Results of speech/singing classification*

Method	Precision		Recall		F-score	
	Bert.	NUS	Bert.	NUS	Bert.	NUS
$\Delta F0$	0.78	0.76	0.83	0.74	0.80	0.74
DFT-F0	0.73	0.77	0.77	0.76	0.75	0.77
GMM	**0.95**	0.75	0.85	0.66	0.89	0.64
Proposed	0.91	**0.89**	**0.93**	**0.89**	**0.92**	**0.89**

5. Conclusions

An efficient singing/speech discrimination system that is suitable for classifying short segments has been developed. It uses only two parameters extracted from pitch to take the decision and classify each segment. As a by-product a flexible algorithm to label musical notes has also been produced. Currently we are testing LSTMs to segment singing and speech taking advantage of their capacity to classify sequences of different lengths.

6. Acknowledgements

This work 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 (BerbaOla).

7. References

[1] A. Loscos, P. Cano, and J. Bonada, "Low-delay singing voice alignment to text." in *ICMC*, 1999.

[2] A. Mesaros and T. Virtanen, "Automatic recognition of lyrics in singing," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, pp. 1–11, 2010.

[3] D. Deutsch, R. Lapidis, and T. Henthorn, "The speech-to-song illusion," *Acoustical Society of America Journal*, vol. 124, pp. 2471–2471, 2008.

[4] S. Falk and T. Rathcke, "On the speech-to-song illusion: Evidence from german," in *Speech Prosody 2010-Fifth International Conference*, 2010.

[5] S. R. Livingstone, K. Peck, and F. A. Russo, "Acoustic differences in the speaking and singing voice," *Proceedings of Meetings on Acoustics*, vol. 19, no. 35080, 2013.

[6] J. Merrill and P. Larrouy-Maestri, "Vocal features of song and speech: Insights from Schoenberg's Pierrot lunaire," *Frontiers in Psychology*, vol. 8:1108, 2017.

[7] W. Chou and L. Gu, "Robust singing detection in speech/music discriminator design," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 2. IEEE, 2001, pp. 865–868.

[8] B. Schuller, B. J. B. Schmitt, D. Arsić, S. Reiter, M. Lang, and G. Rigoll, "Feature selection and stacking for robust discrimination of speech, monophonic singing, and polyphonic music," *IEEE International Conference on Multimedia and Expo, ICME 2005*, vol. 2005, pp. 840–843, 2005.

[9] D. Gerhard, "Pitch-based acoustic feature analysis for the discrimination of speech and monophonic singing," *Canadian Acoustics*, vol. 30, no. 3, pp. 152–153, 2002.

[10] B. Schuller, G. Rigoll, and M. Lang, "Discrimination of speech and monophonic singing in continuous audio streams applying multi-layer support vector machines," *2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763)*, pp. 1655–1658, 2004. [Online]. Available: http://ieeexplore.ieee.org/document/1394569/

[11] W. H. Tsai and C. H. Ma, "Speech and singing discrimination for audio data indexing," *Proceedings - 2014 IEEE International Congress on Big Data, BigData Congress 2014*, pp. 276–280, 2014.

[12] Y. Ohishi, M. Goto, K. Itou, and K. Takeda, "Discrimination between Singing and Speaking Voices," *Interspeech*, pp. 1141–1144, 2005.

[13] B. Thompson, "Discrimination between singing and speech in real-world audio," *2014 IEEE Workshop on Spoken Language Technology, SLT 2014 - Proceedings*, pp. 407–412, 2014.

[14] T. Hain and P. C. Woodland, "Segmentation and classification of broadcast news audio," in *Fifth International Conference on Spoken Language Processing*, 1998.

[15] P. Boersma, "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*, vol. 17, no. 1193. Amsterdam, 1993, pp. 97–110.

[16] M. Mauch, C. Cannam, R. Bittner, G. Fazekas, J. Salamon, J. Dai, J. Bello, and S. Dixon, "Computer-aided melody note transcription using the tony software: Accuracy and efficiency," 2015.

[17] Y. S. Moon and J. Kem, "Fast normalization-transformed subsequence matching in time-series databases," vol. E90-D, no. 12, 2007, pp. 2007–2018.

[18] O. K. Kostakis and A. G. Gionis, "Subsequence Search in Event-Interval Sequences," 2015, pp. 851–854.

[19] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[20] X. Sarasola, E. Navas, D. Tavarez, D. Erro, I. Saratxaga, and I. Hernaez, "A singing voice database in Basque for statistical singing synthesis of bertsolaritza." in *LREC*, 2016.

[21] Z. Duan, H. Fang, B. Li, K. C. Sim, and Y. Wang, "The nus sung and spoken lyrics corpus: A quantitative comparison of singing and speech," in *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, Oct 2013, pp. 1–9.

[22] A. Savitzky and M. J. E. Golay, "Smoothing and differentiation of data by simplified least squares procedures." *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[23] J. Sundberg, "The acoustics of the singing voice," *Scientific American*, vol. 236, no. 3, pp. 82–91, 1977.