# SINGING VOICE DETECTION IN POP SONGS USING CO-TRAINING ALGORITHM

*Swe Zin Kalayar Khine, Tin Lay Nwe, Haizhou Li*

Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613
{zkkswe, tlnma, hli}@i2r.a-star.edu.sg

## ABSTRACT

We propose a co-training algorithm to detect the singing voice segments from the pop songs. Co-training algorithm leverages *compatible* and partially *uncorrelated* information across different features to effectively boost the model from unlabeled data. We adopt this technique to take advantage of abundant unlabeled songs and explore the use of different acoustic features including vibrato, harmonic, attack-decay and MFCC (Mel Frequency Cepstral Coefficients). The proposed algorithm substantially reduces the amount of manual labeling work and computational cost. The experiments are conducted on the database of 94 pop solo songs. We achieve an average error rate of 17% in segment level singing voice detection.

***Index Terms*****—** Co-training algorithm, Singing voice detection, Hidden Markov Model, Timbre.

## 1. INTRODUCTION

Detection of singing voice is needed in applications such as singer identification, singing voice separation, music information retrieval, music transcription and summarization. In Karaoke applications, singing voice segments have to be detected to carry out lyric alignment. One of the most important characteristics of music is the presence of singing voice [1]. We define vocal as singing voice with or without instrumental accompaniment. Pure instrumental music, or nonvocal, refers to the segments that only have music without singing. In singing voice modeling, we train vocal and nonvocal models from labeled song database. In singing voice detection, a common approach is in two steps: 1) extracting features from the song, 2) labeling the segments into vocal and nonvocal classes.

The unlabeled songs are easier to obtain than the labeled ones. In this paper, we propose a novel co-training algorithm for training a vocal/nonvocal classifier. The algorithm starts with a small set of labeled songs to bootstrap the model; it uses knowledge learnt from one feature to probabilistically label a database. The resulting labels are in turn used to train classifier based on another features. This repeats across different features extracted from the same database. In this way, we leverage the *compatible* (features giving consistent prediction for labeling) and partially *uncorrelated* (features providing different views to the labeling problem) information across different features to effectively boost the model from unlabeled data.

Blum and Mitchell [2] introduces the co-training algorithm on web page classification consisting of two redundantly sufficient sets of features are trained separately using a small set of labeled web pages on each view. Each algorithm's prediction on new unlabeled web pages are used to augment the training set for the other feature. The learning scheme of co-training is Naive Bayes. As stated by Chan, Koprinska and Poon [3], the performance of co-training also depends on the learning algorithm it uses. They use Support Vector Machines that outperforms Naive Bayes on email classification. Muller, Rapp and Strube [4] employ decision tree classifier for co-training approach with independent features using the small sets of training labeled German texts in a loop to label the unlabeled German texts. Lee, Kan and Lai [5] employed co-training with PARCELS classifier that uses based on separate stylistic and lexical views of the web block. Their co-training process results outperform single-view result. There are redundantly sufficient features to extract from data so that an email having only one set of features or another can be classified [15]. In summary, the co-training algorithm works when an initial classifier of reasonable performance and the redundantly sufficient features are available. In this paper, we employ co-training algorithm using HMM (hidden Markov model) [9] classifier. The system co-trained on three features outperforms that using single feature.

The rest of this paper is organized as follows. We study perceptually motivated acoustic features and their characteristics in section 2. In section 3, Detail of the co-training algorithm and HMM classifier, in section 4, describe the pop song database, experiment setup and results. Finally, we conclude our study in section 5.

## 2. ACOUSTIC FEATURES

Several perceptually motivated features, namely harmonic, vibrato and timber features, characterize song segments. We use subband filters on octave frequency scale in formulating these acoustic perceptual features.

### 2.1. Vibrato

Vibrato is a useful cue for vocal/nonvocal discrimination [7]. It is a periodic, rather sinusoidal, modulation of pitch and timbre of a musical tone [8]. The style of the individual singer can develop a vocal vibrato function [9].

Not all instruments such as percussion instruments can produce vibrato because some have fixed pitches which cannot be varied by sufficiently small degrees [10]. The contemporary flutist use vibrato lavishly. However, the tone most recommended for eighteen and early nineteen century flutist was probably produced without vibrato [17]. String player for instance guitarist choose to vary pitch from below, only up to the nominal note and not above it. The performers or singers regard vibrato as an ornament. And, the style of the vibrato is decided by themselves [9].

Three different types of vibrato are shown in Figure 1. Vibrato is characterized by two parameters: the extent and the rate. The vibrato extent describes how far the frequency of partial fluctuates left or right from a note within a vibrato cycle. The vibrato rate specifies the number of fluctuations per second. Although vibrato excursions to the left and right are not balanced at Type-A and Type-B, vibrato excursions to the up and down from the note are balanced at Type-C. Type-C vibrato has narrower pitch fluctuation and faster rate.

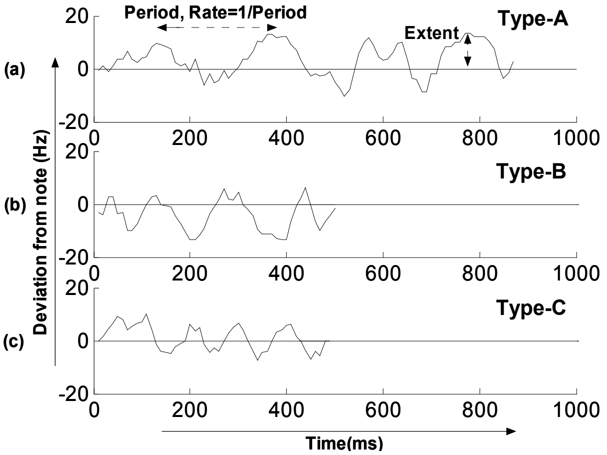


**Figure 1. Three types of vibrato waveforms observed at the note of D6, 1174.6Hz being normalized to 0 at Y-axis**

Vibrato filter has two cascaded layers of subband as demonstrated in Figure 2. The first layer consists of the overlapped 96 trapezoidal bandpass filters which are tapered between $\pm 0.5$ semitone and $\pm 1.5$ semitone. The singing voice contains high frequency harmonics [11], so our subband filters span up to 8 octaves (16 kHz). The tapered and overlapped trapezoidal filters allow vibrato fluctuations of adjacent notes. The second layer has 5 non-overlapped rectangular filters of equal bandwidth for each trapezoidal subband [9].

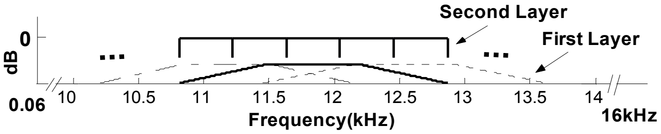


**Figure 2. A bank of two cascaded subband filters**

In Figure 3, the upper panel illustrates the spectrum partial. The middle panel shows the frequency response of the vibrato filter and the lower panel demonstrates the instantaneous amplitude output of the vibrato filter which can track the local maxima to derive the vibrato extent [9]. The different types of vibrato undulations are captured by the vibrato filters.

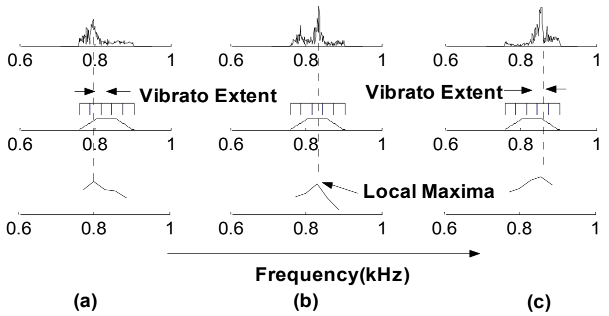
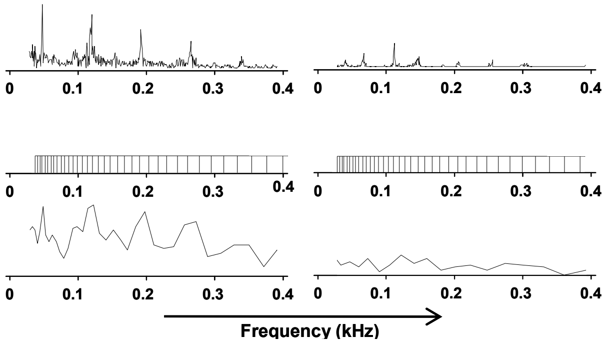


**Figure 3. Vibrato fluctuations and vibrato filtering observed at the note G#5, 830.6Hz. (a) Vibrato fluctuates left (b) no fluctuation (c) Vibrato fluctuates right.**

## 2.2. Harmonic

Rocamore [12] mentioned that the harmonic of the singing voice is high because the partials of the singing voice are located at multiples of the fundamental frequency and some of them are overlapped with the harmonic of the musical instrument from the accompaniment. The start of the singing voice makes a rapid increase in the energy level of the music signal [11]. To capture the difference of the harmonic spectral intensity between vocal and nonvocal segments, we implement the bandwidths of the harmonic filters with $\pm 0.5$ semitone from each note and the filters span up to 16 kHz as shown in the middle panel of Figure 4. In Figure 4, the upper panel illustrates the harmonic spectrum of the vocal and nonvocal signals and the lower panel demonstrates the output of the harmonic filters.



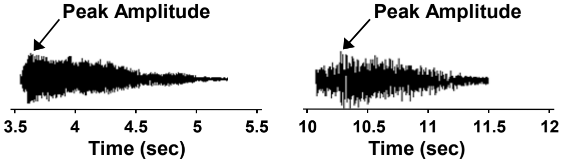
**(a) Vocal signal (b) Nonvocal signal**

**Figure 4. Harmonics and harmonic filtering.**

## 2.3. Timbre

For sounds that have the same pitch and loudness, timbre or sound quality is a general term for the distinguishable characteristics of a tone. Timbre is mainly determined by the harmonic content of a sound and the dynamic characteristics of the sound such as vibrato and attack-decay envelope of the sound [6]. The onset is very short about 6 ms for vowels independent of vowel colour. Musical instrument has longer onset time ranging from 20ms to 300ms. This is one of the factors that support singing voice to stand out of the musical accompaniments [13]. Attack-decay processes of vocal and nonvocal signals are shown in Figure 5 (a) and (b) respectively.

Figure 5 (a) demonstrates that the vocal signal arises a sudden attack to achieve its peak amplitude and the decay process is more gradual than the decay process of the nonvocal shown in Figure 5 (b). The nonvocal signal takes more time than the vocal signal to develop to its peak.



**(a) Vocal signal** **(b) Nonvocal signal**

**Figure 5. Attack-decay envelopes.**

## 2.4. Cepstral Coefficient Computation

A music signal is divided into frames of 20 ms with 13ms overlapping. Hamming window is applied to each frame to minimize signal discontinuities at the end of each frame and each audio frame is passed through vibrato filters. A total of 13 Octave Frequency Cepstral Coefficients ($OFCC_{vib}$) is computed from the log energies employing Discrete Cosine Transform. The feature coefficients with delta parameters from two neighboring frames are augmented to detain temporal information because delta parameters take care of vibrato rate and the attack-decay envelop in $OFCC_{har}$ and MFCC.

We replace the vibrato filters with the harmonic filters to calculate 13 Octave Frequency Cepstral Coefficients ($OFCC_{har}$) and MFCC filters [14] to compute 13 Mel Frequency Cepstral Coefficients (MFCC).

## 3. CO-TRAINING ALGORITHM

Suppose that we are able to extract two *compatible* and partially *uncorrelated* features $F_1$ and $F_2$ from a database. We now formulate the co-training algorithm as follows:

Given:

- F1 and F2 are redundantly sufficient sets of features
- L is a set of labeled training segments
- U is a set of unlabeled segments

Loop:

- Learn the classifier $C_1$ from L based on $F_1$
- Learn the classifier $C_2$ from L based on $F_2$
- Allow $C_1$ and $C_2$ to label the data in U
- Choose the labeled in U and add to L

In this paper, we propose the co-training algorithm for hidden Markov model (HMM) classifier with three features. First, we utilize the harmonic content and the dynamic characteristics of the sound such as vibrato and attack-decay to characterize the timbre feature effect. Then, we use MFCC feature to improve performance. Studies have shown [9] that the features are all effective in vocal/nonvocal detection. Intuitively, we expect that the vibrato, harmonic and MFCC spectral features have *compatible* views as far as vocal/nonvocal classification is concerned. However, it is arguable that the three views are *uncorrelated*. Studies [16] have also shown that the views do not have to be entirely *uncorrelated* for co-training to take effect. This motivates our attempt to explore multi-view co-training for vocal/nonvocal classifier.

The co-training is performed using three features from both annotated label training song segments and automatically labeled song segments as shown in Figure 6. Only the automatically labeled song segments are used for co-training.

In Figure 6, the first feature starts employing vibrato feature to extract the training labeled segments and unlabeled test songs. The HMM classifier learns and labels the unlabeled test songs. The automatically labeled song segments are added to the training labeled song segments to augment the training database as shown in the dotted line in Figure 6. We repeat the same process for harmonic feature and MFCC feature. We obtain the final vocal/nonvocal segments from the labeling using MFCC feature. We also implement a variation of the proposed co-training by removing the dotted lines. In other words, the second co-training approach starts with a small labeled database and continues the co-training only using automatically labeled data. The co-training can run in multiple iterations.

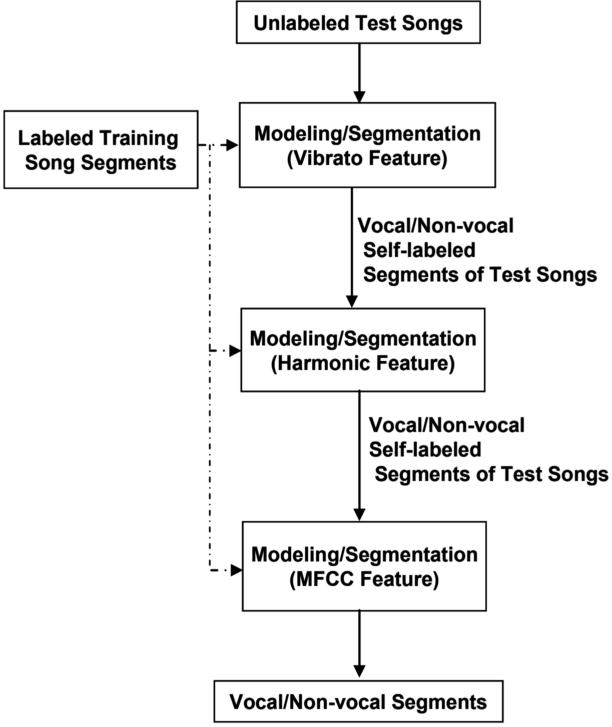

Unlabeled Test Songs
Labeled Training Song Segments
Modeling/Segmentation (Vibrato Feature)
Vocal/Non-vocal Self-labeled Segments of Test Songs
Modeling/Segmentation (Harmonic Feature)
Vocal/Non-vocal Self-labeled Segments of Test Songs
Modeling/Segmentation (MFCC Feature)
Vocal/Non-vocal Segments

**Figure 6. A co-training algorithm leveraging vibrato, harmonic and MFCC features in the training process**

## 4. EXPERIMENTS AND DISCUSSION

We study the effect of the co-training algorithm on the vocal detection system. We created a database of 94 popular pop solo songs consisting of English and Chinese songs. This song database is split into training database (TrainDB) containing 49 songs from 7 singers and test database including 45 songs from 14 singers (TestDB). The songs and the singers in TrainDB and TestDB are not overlapped. Every song is annotated manually to provide the ground truth labeled and each annotated vocal or nonvocal segment lengths range from 0.8 seconds to 12 seconds. The training songs are manually divided by vocal/nonvocal song segments and these vocal song segments are labeled by gender and tempo (High tempo, Medium Tempo and Low Tempo). Seven typed of labels are nonvocal, vocal male high tempo, vocal male medium tempo, vocal male low tempo, vocal female high tempo, vocal female medium tempo and vocal female low tempo. Window size is 20ms and frame shift is 7ms in all tests.

We train the continuous density HMM with four states and two Gaussian mixtures per state for all HMM models [9] using the labeled vocal/nonvocal song segments. Seven HMM models are trained using the TrainDB. We calculate approximately the likelihood score of the training song segments being generated by each of the 7 models. During testing, the test song is divided into 1s segments to extract the features and made the vocal/nonvocal detection decision. Each 1s segment is calculated with all the 7 models in the classifier and the model with the highest likelihood suggests the best match.

We conduct several singing voice detection experiments. We first conduct experiments using single feature such as MFCC, $OFCC_{vib}$ (Vibrato), $OFCC_{har}$ (Harmonic) and timbre feature which is determined by the combination of the harmonic content, vibrato and attack-decay (TBCC). We then conduct co-training experiments using automatically labeled data only (AL) and using automatically labeled data plus labeled data (ALL). Different combinations of features are also tried out in the experiments, such as, Vibrato + Harmonics and Vibrato + Harmonics + MFCC as shown in Table 1.

| Experiment Setup | MFCC | VIBRATO | HARMONIC | CO-TRAINING | AL | ALL | Error Rate (%) |
|---|---|---|---|---|---|---|---|
| MFCC | ✓ | | | | | | 20.36 |
| $OFCC_{vib}$ | | ✓ | | | | | 19.1 |
| $OFCC_{har}$ | | | ✓ | | | | 19.04 |
| TBCC | | ✓ | ✓ | | | | 21.36 |
| $CoAL_{vib,har}$ | | ✓ | ✓ | ✓ | ✓ | | 17.99 |
| $CoAL_{vib,har,MFCC}$ | ✓ | ✓ | ✓ | ✓ | ✓ | | 18.1 |
| $CoALL_{vib,har}$ | | ✓ | ✓ | ✓ | | ✓ | 17.94 |
| $CoALL_{vib,har,MFCC}$ | ✓ | ✓ | ✓ | ✓ | | ✓ | 17.05 |

**Table 1. Error rate (ER%) of singing voice detection on TestDB.**

We report the results in Table 1 to illustrate combinations of different settings in the experiments. $CoALL_{vib,har,MFCC}$, with an average error rate of 17.05%, outperforms all the other experiment setups. It is observed that $CoALL_{vib,har,MFCC}$ capture singing voice detection well by 5.8%, 20%, 10.5% and 10.7% relative error reduction over $CoAL_{vib,har,MFCC}$, TBCC, $OFCC_{har}$ and $OFCC_{vib}$. Experiments with co-training give better results than the single feature training method. Although $CoALL_{vib,har}$ and $CoAL_{vib,har,MFCC}$ take up more computations, they achieve clearly better results.

## 5. CONCLUSION

We have presented the co-training algorithm employing several features such as vibrato, harmonic, timbre and MFCC. We conducted experiment with different features and performed the co-training in two different settings. The first one combines the annotated labeled segments with automatically labeled segments in the co-training (with the dotted lines in Figure 6 taking effects); the second one only uses automatically labeled segments in the co-training. The results show that co-training algorithm is an effective tool that leverages different acoustic features from different views, that reduces the labeling work and improving the classifier performance.

## 6. REFERENCES

[1] G. Tzanetakis, "Song-specific Bootstrapping of Singing Voice Structure," *IEEE International Conference Multimedia and Expo*, vol. 3, pp. 2027-2030, 27-30, June,2004.

[2] A. Blum, and T. Mitchell, "Combining Labeled and Unlabeled Data with Co-Training," *IEEE International Conference Data Mining*, pp. 597-598, USA, 2001.

[3] J. Chan, I. Koprinska, and J. Poon, "Co-Training with a Single Natural Feature Set Applied to Email Classification," *ACM International Conference Web Intelligence*, pp. 586-589, 20-24, September, 2004.

[4] C. Muller, S. Rapp, and M. Strube, "Applying Co-Training to Reference Resolution," *40th Annual Meeting of the Association for Computational Linguistics*, pp. 352-359, Philadelphia, July, 2002.

[5] C.H. Lee, M.Y. Kan, and S. Lai, "Stylistic and Lexical Co-training for Web Block Classification," *ACM International Conference Web Information and Data Management*, pp. 136-143, Washington DC, USA, 2004.

[6] F. Winckell, *Music, Sound and Sensation*, Dover, NY, 1967.

[7] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata and H.G. Okuno, "F0 Estimation Method for Singing Voice in Polyphonic Audio Signal Based on Statistical Vocal Model and Viterbi Search," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5, pp. V-253-V-256, Toulouse, 14-19, May, 2006.

[8] R. Timmers, , and P. Desain, Vibrato: Questions and answers from musicians and science. *International Conference of Music Perception and Cognition*, England, 2000.

[9] T. L. Nwe, and H. Li, "Exploring Vibrato-Motivated Acoustic Features for Singer Identification," *IEEE Transactions on Audio, Speech and Language Processing,* vol. 150, pp. 519-530, February, 2007.

[10] "Vibrato", Word of the Day. Answers Corporation, 2006. *Ansers.com,* 13 Dec. 2006.
http://www.answers.com/topic/vibrato

[11] T. Zhang, "System and Method For Automatic Singer Identification," *IEEE International Conference Multimedia and Expo,* Baltimore, MD, 6-9, July, 2003.

[12] M. Rocamora, and P. Herrera, "Comparing Audio Descriptors for Singing Voice Detection in Music Audio Files," *Brazilian Symposium on Computer Music,* San Pablo, Brazil, September, 2007.

[13] W. Hackhaus, *Die Ausgleichsvorgange*. Zeitschrift fur Technische Physik, 1932.

[14] C. Becchetti, and L.P. Ricotti, *Speech Recognition Theory and C++ Implementation*, John Wiley & Sons, New York, 1998.

[15] S. Kiritchenko, and S. Matwin, "Email Classification with Co-Training," *2001 Conference of the Centre for Advanced Studies on Collaborative Research,* Toronto, Canada, 2001.

[16] I. Muslea, S. Minton and C. A. Knoblock. "Active + Semi-supervised learning = Robust Multi-View Learning," 2002, Proc. of the 9th Conference on Machine Learning, pp. 435-442.

[17] M. Robinson and V. Parrish, "Flute Vibrato," The Standing Stones. http://www.standingstones.com/flutevib.html