IMPROVING MELODY EXTRACTION USING PROBABILISTIC LATENT COMPONENT ANALYSIS

Jinyu Han[⋆] [*]   *Ching-Wei Chen*[†]

[†] Gracenote, Inc.
2000 Powell Street Suite 1500, Emeryville, CA 94608, USA
[⋆]Northwestern University
2133 Sheridan Road, Evanston, IL 60208, USA

ABSTRACT

We propose a new approach for automatic melody extraction from polyphonic audio, based on Probabilistic Latent Component Analysis (PLCA). An audio signal is first divided into vocal and non-vocal segments using a trained Gaussian Mixture Model (GMM) classifier. A statistical model of the non-vocal segments of the signal is then learned adaptively from this particular input music by PLCA. This model is then employed to remove the accompaniment from the mixture, leaving mainly the vocal components. The melody line is extracted from the vocal components using an auto-correlation algorithm. Quantitative evaluation shows that the new system performs significantly better than two existing melody extraction algorithms for polyphonic single-channel mixtures.

Index Terms— Melody Extraction, Probabilistic Latent Component Analysis, Singing Voice Detection and Extraction

1. INTRODUCTION

Melody is one of the most basic and easily recognizable traits of musical signals. The main melody of a song is usually defined as the pitch sequence that a human listener is most likely to perceive and associate with that piece of music. Knowing the melody of a song is useful in numerous applications, including music recognition, analysis of musical structure, and genre classification. Although humans have a natural ability to identify and isolate the main melody from polyphonic music, automatic extraction of melody by a machine remains a challenging task.

In polyphonic music, there are multiple instruments and sound sources playing simultaneously. Determining the main melody from such an audio recording involves extracting a single dominant pitch contour out of a mixture of concurrent spectral events. In this paper, melody is defined as the pitch contour of the lead vocal in a song. This is a reasonable assumption since when music contains a singing voice, many people remember and recognize that piece of music by the melody line of the lead vocal part.

Many melody extraction algorithms have been proposed over the last decade. Generally speaking, they can be classified into two categories. Systems inspired by multi-pitch estimation employ different probabilistic models for pitch candidate selection, followed by a pitch tracker that finds the most probable melodic line [1, 2]. These systems consider the probabilistic relationships between the main melodic source and the polyphonic audio mixture. Previous algorithms [3, 4, 5, 6] have used GMM, Hidden Markov Model (HMM) and Particle Filtering to model this relationship.

Systems based on source separation use statistical methods to model the lead singing components and the background accompaniment separately. The main melody line is then extracted from the singing components, based on the assumption that the main melody is usually the vocal melody. GMM, Gaussian Scaled Mixture Model (GSMM), Instantaneous Mixture Model (IMM) and Non-negative Matrix Factorization (NMF) are popular generative models for each individual component [7, 8]. The above-mentioned methods usually introduce generative models for the signal. Another possibility is the use of classification schemes such as Support Vector Machine in [9].

Our proposed system belongs to the second category. In contrast to previous methods requiring a source/filter model, our algorithm is based on adaptively learning a statistical model for each component of the music from the mixture itself. In this paper, we are concerned with polyphonic music containing singing voice and accompaniment. Based on the assumption that the sound produced by the accompaniment is similar during both the non-vocal and vocal parts of the song, a probabilistic model for the accompaniment is learned from the non-vocal segments of the mixture and then used to remove the accompaniment from the polyphonic mixture. After the accompaniment is suppressed in the mixture, the melody line of the music can be more easily extracted from the remaining singing components of the signal.

Although our system is only tested to extract melody from mixtures containing singing voice, it is noted that it could be easily applied for melody extraction from music with a lead instrument.

The paper is organized as follows. Section 2 introduces the probabilistic model used in our system. Section 3 presents the overall melody extraction algorithm. Experimental Results are provided in Section 4. Section 5 concludes this paper.

2. PROBABILISTIC LATENT COMPONENT ANALYSIS

Probabilistic Latent Component Analysis (PLCA) decomposes a multi-dimensional distribution as a mixture of latent components where each component is given by the product of one-dimensional marginal distributions. Recently, it has been shown that PLCA is numerically identical to NMF for two-dimensional input, and

[*] The first author performed the work while at the Gracenote Media Technology Lab

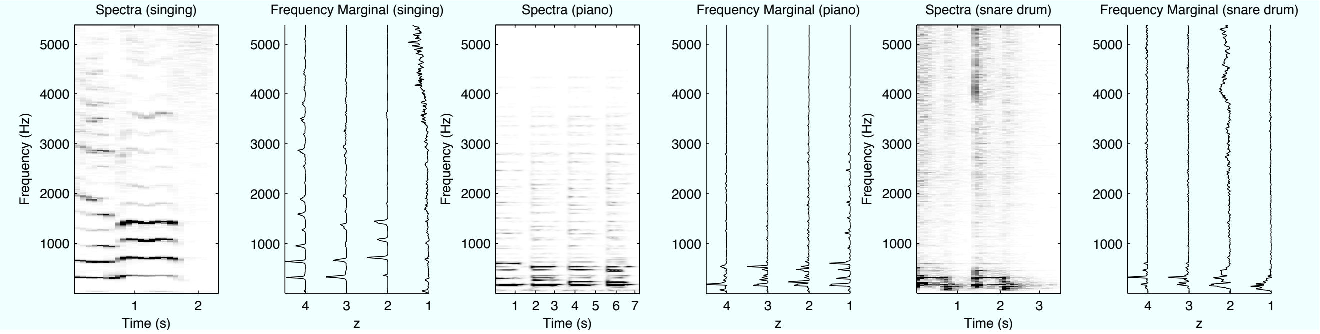

Fig. 1. Example of PLCA models of three different sounds. The two left plots display a spectrogram of singing voice and a set of derived frequency marginals. Likewise the middle two and right two display the same information for a piano sound and snare drum sound. A set of four latent variables is introduced for conditional independence. Note how the derived marginals in different cases extract representative spectra for each sound

non-negative tensors for arbitrary dimensions [10], however, PLCA presents a much more straightforward way to make easily extensible models.

The basic model of PLCA is defined as:

$$P(\mathbf{x}) = \sum_{z \in \mathbf{Z}} P(z) \Pi_{j=1}^{N} P(x_j|z) \tag{1}$$

where $P(\mathbf{x})$ is an N-dimensional distribution of the random variable $\mathbf{x} = x_1, x_2, ..., x_N$. $\mathbf{Z}$ is a set of latent variables introduced to achieve the conditional independence of $\mathbf{x}$, and $P(x_j|z)$ are one dimensional distributions. This model effectively represents a mixture of marginal distribution products to approximate an N-dimensional distribution based on conditional independence. The objective is to discover the most appropriate marginal distributions.

The estimation of the marginal $P(x_j|z)$ can be performed using the Expectation Maximization (EM) algorithm [10]. In the expectation step, the posterior of the latent variable z is estimated:

$$P(z|\mathbf{x}) = \frac{P(z)\Pi_{j=1}^{N} P(x_j|z)}{\sum_{z'} P(z')\Pi_{j=1}^{N} P(x_j|z')} \tag{2}$$

and in the maximization step, the marginals are re-estimated as follows:

$$P(z) = \int P(\mathbf{x}) P(z|\mathbf{x}) d\mathbf{x} \tag{3}$$

$$P^*(x_j|z) = \int \ldots \int P(\mathbf{x}) P(z|\mathbf{x}) dx_k, \forall k \neq j \tag{4}$$

$$P(x_j|z) = \frac{P^*(x_j|z)}{P(z)} \tag{5}$$

Given the spectrogram of a piece of polyphonic music, PLCA can be used to explicitly model the spectrogram as a two-dimensional distribution in time and frequency

$$P(f,t) = \sum_{z} P(z) P(f|z) P(t|z) \tag{6}$$

where $P(f,t)$ represents the distribution of the spectrogram, and $P(f|z)$ and $P(t|z)$ are conditional distributions along the frequency and time dimensions. While the time axis marginals are not particularly informative, the frequency axis marginals contain a dictionary of the spectrogram which best describes the sound represented by the input.

These frequency marginals can be used as a model for certain kinds of sounds such as singing voice, speech or particular instruments. Fig. 1 shows three sets of frequency marginals learned from three different kinds of sounds: singing voice, piano and snare drum. It is shown that the extracted frequency marginals capture a unique energy distribution along the frequency dimension for the sound. For example, the frequency marginals extracted from singing voice display clear harmonic structures for the vowel sounds and high frequency distribution for the fricative at the end, while the marginals from the snare drum have a flatter and more uniform distribution.

Note that once the frequency marginals are known for a certain sound in a mixture, they can be used to extract this kind of sound from the mixture in a supervised way [11] . In the next section, we describe how this model can be used in an un-supervised way for melody extraction.

3. METHOD DESCRIPTION

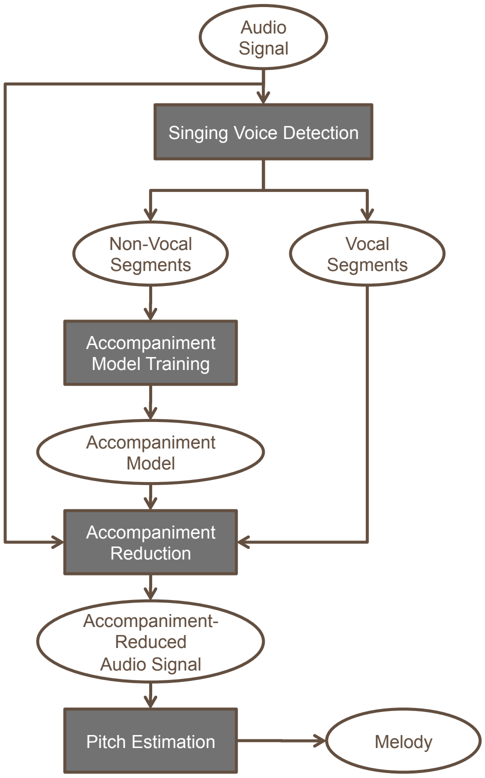

Fig. 2. System overview

Assume that we have a polyphonic audio signal featuring a singing voice and multiple instruments. Previous work [12] used a set of training instruments to learn a model space which fits each individual instrument based on Latent Component Analysis. In contrast to [12], our method does not use pre-trained models, instead, the models for the accompaniment and singing voice are learned adaptively from the mixture itself.

Our system illustrated in Fig. 2 follows a generic procedure for adapted source separation such as presented in [7] and [13]. We deal with the melody extraction problem in four stages.

In the first stage, *Singing Voice Detection*, the mixture is divided into vocal and non-vocal partitions using a trained Gaussian Mixture Model (GMM) classifier similar to [4]. Let $\mathbf{X_v}$ be the spectrogram of the vocal partition containing the singing voice and $\mathbf{X_{nv}}$ be the non-vocal partition with only accompaniment. The frequency marginals distribution $P_{nv}(f|z)$ for the accompaniment are then learned from $\mathbf{X_{nv}}$ in the *Accompaniment Model Training* stage using the PLCA model described in Section 2. $z \in \mathbf{Z_{nv}}$ is the set of latent variables extracted from $\mathbf{X_{nv}}$.

In the third stage, *Accompaniment Reduction*, the singing voice is extracted from the mixture as follows. Assuming that the accompaniment stays stable during both the non-vocal and vocal segments of the music, $\mathbf{X_v}(\mathbf{f},\mathbf{t})$ can be decomposed into two sets of frequency marginals:

$$X_v(f,t) = \sum_{z\in\mathbf{Z_{nv}}} P(z)P_{nv}(f|z)P(t|z) + \sum_{z\in\mathbf{Z_v}} P(z)P_v(f|z)P(t|z) \tag{7}$$

where $\mathbf{Z_{nv}}$ is the same set of latent variables extracted from $\mathbf{X_{nv}}$, and $\mathbf{Z_v}$ is the set of additional latent variables we added to explain the singing voice in $\mathbf{X_v}(\mathbf{f},\mathbf{t})$. We perform PLCA on $\mathbf{X_v}$ as we usually do when learning both the frequency and the time marginals, but we make sure that the frequency marginals corresponding to $\mathbf{Z_{nv}}$ are fixed to $P_{nv}(f|z)$ as we update only the remaining ones using the same training procedure as before.

The additional frequency marginals $P_v(f|z)$ we learned will best explain the lead singing voice in the mixture which is not present in the non-vocal partitions $\mathbf{X_{nv}}$. Once the marginals of the singing sources have been learned, we can reconstruct the spectrogram of the singing components $X_s(f,t)$ using only the distributions associated with $\mathbf{Z_v}$:

$$X_s(f,t) = \sum_{z\in\mathbf{Z_v}} P(z)P_v(f|z)P(t|z). \tag{8}$$

We assume the phase of the singing components is the same as the phase of the polyphonic audio, since the human ear is not sensitive to phase variations. Then X_s plus the original phase of the mixture can be converted to the time domain signal by a simple overlap-add technique. The time domain signal is passed to the fourth stage *Pitch Estimation* for final melody extraction. Given the singing voice extracted from the mixture, the main pitch sequence can be easily estimated by a simple auto-correlation technique similar to [14].

4. EXPERIMENT

To test the effectiveness of the PLCA model for accompaniment removal, we obtained a clip of rock music which contains a mix of four sources (singing voice, electric guitar, electric bass and drum kit), as well as a separate track of the singing voice. We manually divided the clip into a 15-second non-vocal segment and 14-second vocal segment. A statistical model for the accompaniment is learned from the non-vocal segment and then applied to reduce the accompaniment from the mixture as described in Section 3. Fig. 3 shows the result of this process on the vocal segment of the mixture. The spectrogram of the mixture is shown in Fig. 3(a). The spectrogram of the signal extracted from the mixture is plotted in (b) and the spectrogram of the separate vocal track is plotted in (c). The melody pitch estimation (red dots) extracted from the mixture is plotted against the ground-truth pitch (black solid line) extracted from the separated vocal track in (d). In this example, the detected pitch track matches well the ground-truth track with 80% overall accuracy. This example shows that our proposed system works well when we have a perfect *Singing Voice Detection* module.

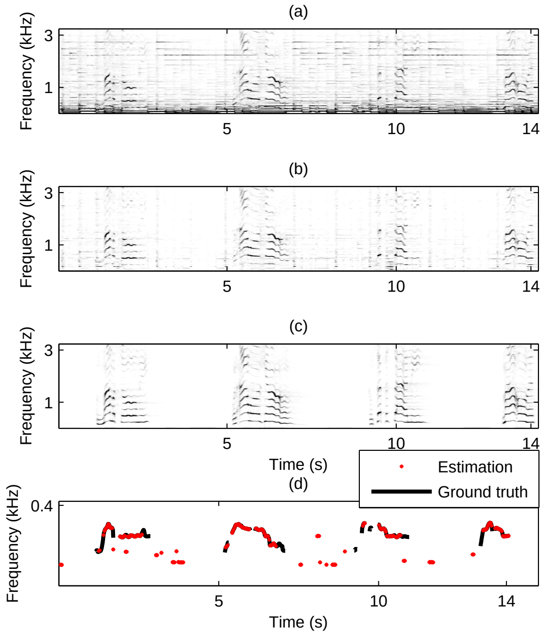

Fig. 3. Melody extraction on a clip of "Simple Man" by Lynyrd Skynyrd. (a) Spectrogram of the mixture. (b) Spectrogram of the extracted singing voice. (c) Spectrogram of the original singing voice before mixing. (d) Melody detection result.

Next we show quantitative evaluation of our system with an automatic singing voice detector.

The GMM-based singing voice detector is trained on a data set of 51 commercial songs across various genres. The ground-truth vocal/non-vocal segments are manually annotated by the authors. Mel-frequency cepstral coefficients (MFCCs) are used as the input feature for the classifier. We performed three-fold cross validation on this data set. The average precision of the classifier is 76% for vocal detection and 73% for non-vocal detection. The parameters for the best GMM classifier are used for the *Singing Voice Detection* module of the overall system.

The overall melody extraction system is tested on parts of the ISMIR 2005 training data set of 13 songs for audio melody extraction [9]. We only considered songs the database containing lead vocals, i.e., 9 songs, totaling about 270 seconds of audio, with two musical styles: jazz and pop. All test songs are single channel PCM data with 44.1 kHz sample rate and 16-bit quantization.

We compared our proposed system to two recent pitch/melody estimation systems: DHP [2] and LW [4]. DHP is a state-of-art multi-pitch estimation algorithm based on spectral peak and non-peak region selection. It outputs a likelihood score for each pitch hypothesis to indicate the confidence level of the estimate. The first pitch being detected is considered the predominant pitch in the sense that the score of this pitch hypothesis is the highest. LW is a predominant pitch detection algorithm based on channel/peak selection and HMM model. This algorithm is specially designed for extracting the singing voice melody from polyphonic audio. For both algorithms, we use the source code and recommended parameters provided by the authors. The pitch value is estimated every 10 milliseconds.

	Precision	Recall	F-measure	Accuracy
DHP	**0.52**	0.48	0.50	0.48
LW	0.09	0.086	0.09	0.19
Proposed	0.43	**0.80**	**0.55**	**0.61**

Table 1. Performance comparison of the proposed algorithm against DHP and LW, averaged across 9 songs

The metrics considered are *Precision*, *Recall*, *F-measure* and overall *Accuracy*. The results are summarized in Table 4. The LW algorithm performs poorly in all metrics. We believe the strong energy from accompaniment in the music causes the poor performance of LW, because the estimated pitches are found to match many pitches from the pitched accompaniment instruments. We also speculate that the parameters for LW may need to be specially tuned for a certain data set, even though we used the recommended values for all the parameters. DHP has the best precision measurement but it failed to output pitch estimates in the singing voice regions where there is strong interference from the percussion instruments, producing a much lower recall than our system. In the presence of other instrumental sounds, our proposed system achieves the best recall, F-measure and overall accuracy on this data set. The high recall of our system indicates that the proposed *Accompaniment Reduction* stage successfully suppresses the background instruments, leaving the singing voice as the predominant component in the extracted spectrogram. The relatively low precision is because that the background music is not completely removed from the mixture partly due to an imperfect *Singing Voice Detection* stage. Compared to the 80% overall accuracy achieved in the example presented in Fig.3, we believe the proposed system can perform significantly better with an improved singing voice detection technique.

5. CONCLUSION

We developed an unsupervised algorithm for melody extraction from single channel polyphonic music. Our system assumes no prior information on the type or the number of instruments in the mixture. We introduce Probabilistic Latent Component Analysis to model the accompaniment and lead vocal adaptively. Experimental results show that the PLCA model successfully suppressed the background music in the mixture audio. Quantitative evaluation showed our proposed algorithm is significantly better than two other melody extraction algorithms. The proposed system can be easily extended to extract the melody from a lead instrument or to a singing voice separation system. Although the proposed method does not require pretrained instrument models, its performance indeed depends on the performance of the singing voice detection. More advanced singing voice detection and pitch estimation techniques are currently under investigation.

6. REFERENCES

[1] A.P. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 6, pp. 804–816, 2003.

[2] Z. Duan, J. Han, and B. Pardo, "Harmonically informed pitch tracking," in *Proc. ISMIR*, 2009.

[3] M. Goto, "A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication*, vol. 43, no. 4, pp. 311 – 329, 2004.

[4] Y. Li and D. Wang, "Separation of singing voice from music accompaniment for monaural recordings," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, pp. 1475–1487, 2007.

[5] H. Fujihara, M. Goto, T. Kitahara, and H. Okuno, "A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 638 –648, 2010.

[6] S. Jo and C.-D. Yoo, "Melody extraction from polyphonic audio based on particle filter," in *Proc. ISMIR*, 2010.

[7] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of bayesian models for single-channel source separation and its application to voice/music separation in popular songs," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1564 –1578, 2007.

[8] J.-L. Durrieu, G. Richard, B. David, and C. Fevotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 564 –575, 2010.

[9] G. Poliner and D. Ellis, "A classification approach to melody transcription," in *Proc. ISMIR*, 2005.

[10] M. Shashanka, B. Raj, and P. Smaragdis, "Probabilistic latent variable models as nonnegative factorizationss," *Computational Intelligence and Neuroscience*, 2008.

[11] P. Smaragdis, B. Raj, and M. Shashanka, "Supervised and semi-supervised separation of sounds from single-channel mixtures," in *Proc. ICA*, 2007.

[12] G. Grindlay and D. Ellis, "A probabilistic model for multi-instrument polyphonic transcription," in *Proc. ISMIR*, 2010.

[13] S. Vembu and S. Baumann, "Separation of vocals from polyphonic audio recordings," in *Proc. ISMIR*, 2005, pp. 337–344.

[14] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise of a sampled sound," in *Proc. the Institute of Phonetic Sciences*, 1993, vol. 17, pp. 97–110.