ACCOMPANIMENT SEPARATION AND KARAOKE APPLICATION BASED ON AUTOMATIC MELODY TRANSCRIPTION

Matti Ryynänen, Tuomas Virtanen, Jouni Paulus, and Anssi Klapuri

Tampere University of Technology
Institute of Signal Processing
P.O.Box 553, FI-33101 Tampere, Finland
{matti.ryynanen, tuomas.virtanen, jouni.paulus, anssi.klapuri}@tut.fi

ABSTRACT

We propose a method for separating accompaniment from polyphonic music and its karaoke application, both based on automatic melody transcription. First, the method transcribes the lead-vocal melody of an existing polyphonic music piece, where the transcription consists of a MIDI note sequence and a detailed fundamental frequency (F0) trajectory for each note. Based on the note F0 trajectories, the method uses sinusoidal modeling to estimate, synthesize, and remove the lead vocals in the piece, thus producing separated accompaniment of the piece. User sings along with the separated accompaniment similar to karaoke while the user singing can be tuned to the transcribed melody. This will help non-professional singers to produce more appealing karaoke performances. The quality of separated accompaniments was quantitatively evaluated with approximately one hour of polyphonic music, including material from a commercial karaoke DVD.

1. INTRODUCTION

Karaoke is a popular form of entertainment where a non-professional singer sings along with a song accompaniment where lead vocals are not present. Usually, the lyrics of a song are synchronously displayed on a screen to aid the singer. In general, music is distributed in a form where all the instruments and the lead vocals are mixed together in a monophonic or stereophonic audio stream, thus making the material unsuitable for karaoke usage. Therefore, song accompaniments for karaoke are conventionally produced in recording studios by professional musicians, which is both time-consuming and costly. In addition, it is impossible to provide karaoke-song collections including all the consumers' favorite songs. Therefore, it would be useful to have a method for producing the song accompaniment directly from user-selected audio. There exist methods for separating vocals and music, such as [1, 2]. In particular, the method presented in [2] does the separation based on melody transcription.

Karaoke performances are usually given by non-professional singers who may sing out-of-tune, i.e., the singing pitch differs noticeably from the original melody notes. However, there exist commercial products for real-time tuning of singing given reference notes or a pitch track. The reference can be embodied in karaoke media, and it also enables scoring of karaoke performances, such as in SingStar[1] application, for example. The reference can be simultaneously visualized with the user's singing pitch to give feedback.
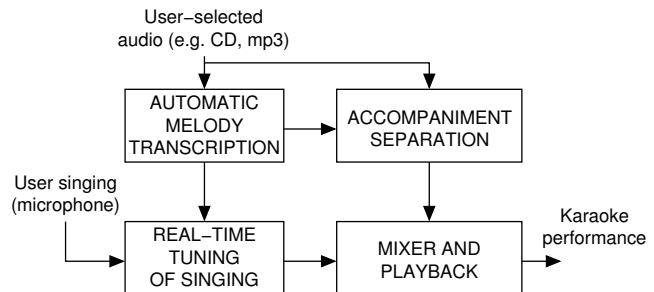
Fig. 1. A block diagram of the proposed method.

A method for producing the reference would be desirable, since it is not directly available for most music pieces.

The proposed method solves the two above-mentioned problems based on automatic melody transcription and accompaniment separation. Figure 1 shows a block diagram of the method. First, the lead-vocal melody of a user-selected music piece is automatically transcribed. Based on the transcription, the vocal signal is estimated and synthesized from the music piece using sinusoidal modeling and subtracted from the original audio to produce a song accompaniment. During a karaoke performance, the user listens to the separated accompaniment and sings along. If desired, the user singing can be tuned to the transcribed lead-vocal melody in real time. The melody transcription and the accompaniment separation run faster than real time and allow causal processing, thus enabling immediate start of a karaoke performance of any music piece. Currently, the method does not take into account the lyrics. However, both the transcribed melody and the synthesized vocals facilitate lyrics alignment for karaoke applications.

The paper is organized as follows. Section 2 describes the melody transcription method, followed by an explanation of the accompaniment separation in Section 3. Section 4 introduces a method for tuning the user singing. Section 5 reports quantitative evaluation results for the accompaniment separation, and Section 6 concludes the paper.

2. MELODY TRANSCRIPTION

Melody transcription refers to the automatic extraction of a parametric representation (e.g., a MIDI file) of the lead-vocal melody from a polyphonic music piece. A melody is a sequence of consecutive
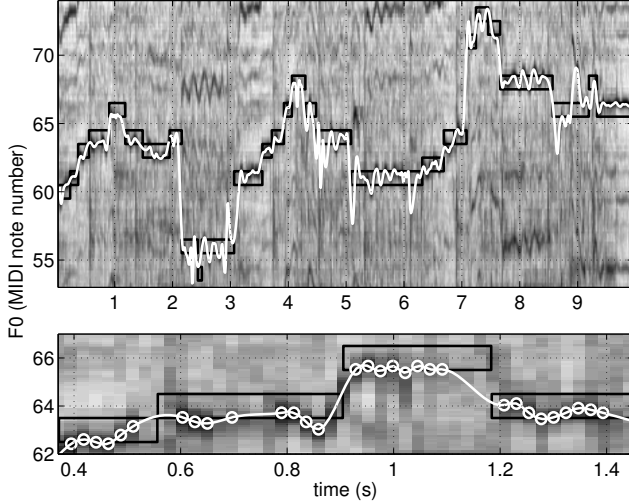
Fig. 2. Automatically transcribed melody and its detailed F0 trajectory for the beginning of a verse in "Sick Sad Little World" by Incubus. See text for details.

notes and rests, where the i:th note in the sequence is defined by its fundamental frequency n_i, onset time, and offset time. Here, fundamental frequency (F0) values are expressed on a scale of unrounded MIDI note numbers $f_{\text{MIDI}} = 69 + 12 \log_2(f_{\text{Hz}}/440)$.

The motivation for using melody transcription is twofold. First, melody transcription produces a useful mid-level representation which facilitates robust accompaniment separation. The separation is performed only during the transcribed notes, thus preserving original audio quality during rests. The transcribed notes also allow robust estimation of a detailed F0 trajectory for the melody which is utilized in the separation. Secondly, the transcribed melody is used as a reference to which the user singing is tuned during a karaoke performance.

We use a melody transcription method [3] which is an improved version of an earlier method [4]. Briefly, the method uses a frame-wise pitch salience estimator to measure the strength of different fundamental frequencies in 92.9 ms analysis frames with 23.2 ms interval between successive frames (i.e., 4096 and 1024 samples at 44.1 kHz sampling frequency f_s). In the following, $s_t(f)$ denotes the salience of fundamental frequency f in frame t. Fundamental frequency f ranges from MIDI note 44 to 84 with approximately 800 values distributed between these limits. The estimated saliences, their time differentials, and a measure of incoming spectral energy are used as features for computing observation likelihoods for melody notes, other instrument notes, and silence or noise segments, each of which is modeled using a hidden Markov model. The parameters of the models have been estimated from several hours of music. A musicological model also uses the saliences to estimate the musical key of the piece and employs the corresponding between-note transition probabilities. These probabilities are modeled with a note bigram trained with thousands of melodies. The Viterbi algorithm is used to find the optimal path through the melody note models in order to produce a transcribed melody note sequence.

After transcribing the melody into notes, a detailed F0 trajectory is estimated for each note n_i as follows. For each frame t in the time region between the note onset and offset, the maximum-salience F0

is obtained by

$$\widehat{f}_t = \operatorname*{argmax}_f s_t(f), \text{ where } |f - n_i| < 3. \quad (1)$$

The condition $|f - n_i| < 3$ in Eq. (1) limits the possible search range in frequency to ± 3 semitones from the transcribed note F0 n_i. The mean $\bar{s}_i$ of the detected salience maxima for the note is then calculated. Only the F0 values $\widehat{f}_t$, for which $s_t(\widehat{f}_t) > \alpha \bar{s}_i$, are preserved for further processing, where α was empirically set to 0.8. This is done to avoid less reliable frames. The preserved F0 values are finally used to interpolate a detailed F0 trajectory at 10 ms time grid for the note time regions, i.e., the trajectory is not defined during rests. The interpolation is performed using piecewise cubic splines. In the following, the interpolated F0 trajectory is denoted with $\widetilde{f}_u$ where u denotes the frame index in the 10 ms grid.

Figure 2 shows an example transcription with melody notes n_i (the black boxes) and the detailed F0 trajectory $\widetilde{f}_u$ (the white line). The gray-level intensity on the background indicates the salience values $s_t(f)$, where darker color shows greater salience. The lower panel shows a close-up of a few notes, where the white circles indicate the preserved F0 values. Piecewise cubic splines are fitted to these points and then used to interpolate the detailed F0 trajectory.

3. ACCOMPANIMENT SEPARATION

The accompaniment separation is based on a signal model

$$x(k) = v(k) + b(k), \quad (2)$$

where the mixture audio signal $x(k)$ is considered as a linear sum of the vocals $v(k)$ and the accompaniment $b(k)$. Index k denotes time in audio samples. The vocal signal $v(k)$ is further modeled as a sum of sinusoids

$$v(k) = \sum_{d=1}^{D} a_d(k) \sin(\theta_d(k)), \quad (3)$$

where $a_d(k)$ and $\theta_d(k)$ are the amplitude and phase of the d:th harmonic partial at time k. D is the number of partials, which was set to 40 in our simulations. In the model, each partial is represented with an individual sinusoid. A good perceptual separation quality was obtained using the quadratic polynomial-phase model [5] and the following analysis procedure.

The procedure first estimates the amplitude, frequency, and phase of each partial in 40 ms frames for each frame index u and then interpolates them between adjacent frames to obtain the samplewise amplitudes and phases. The following processing takes place in an individual frame and the frame indices are omitted to simplify the notation. The frequency f_d (in Hz) of each partial $d = 1, \ldots, D$ is set equal to $f_d = 440 \cdot d \cdot [2^{(\widetilde{f}_u - 69)/12}]$, i.e., the partial frequencies are constrained to integer multiples of the fundamental frequency f_1. Normalized cross-correlation c_d between the analysis frame and a complex exponential having the partial frequency f_d is calculated as

$$c_d = \frac{\sum_k x(k) \exp(i2\pi k f_d/f_s) w(k)^2}{Z}, \quad (4)$$

where $w(k)$ is the Hamming window centered to the temporal position of frame u, f_s is the sampling frequency, and $Z = \sum_k w(k)^2/2$ is a normalization constant. The amplitude of the partial is then obtained as the magnitude of the correlation
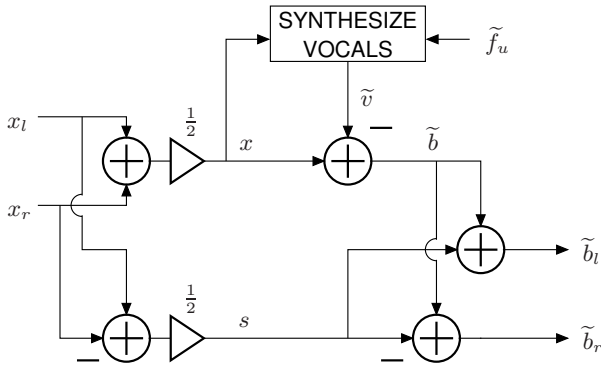
$$a_d = |c_d| \quad (5)$$

Fig. 3. Overview of the accompaniment separation for stereophonic signals.

Fig. 4. An example of user singing F0 illustrated with black dashed line. The gray boxes and the solid line show the transcribed melody and its F0 trajectory, respectively.

and phase as its angle

$$\theta_d = \angle c_d \, . \tag{6}$$

Since the partial frequencies are harmonically related and the exponentials therefore uncorrelated with each other, the above procedure produces a least-squares [6, Chap. 8] estimate of stationary windowed sinusoids even when applied individually for each partial.

Samplewise amplitudes $a_d(k)$ in Eq. (3) are obtained by linear interpolation between adjacent frames. Samplewise phases $\theta_d(k)$ are obtained by quadratic phase interpolation method [5] with parameter value $\lambda = 4/5$ so that phases and frequencies of only the adjacent frames are required in the interpolation. The synthesized partials are finally summed to obtain the estimated vocal signal $\widetilde{v}(k)$. In particular, the signal contains zeros where there is no transcribed melody notes. The separated accompaniment signal $\widetilde{b}(k)$ is then obtained by $x(k) - \widetilde{v}(k)$.

The above-described processes of melody transcription and accompaniment separation are carried out on monophonic mixture signal $x(k)$. In a usual case of stereophonic inputs, processing is applied to the middle signal $x(k) = \frac{1}{2}(x_l(k) + x_r(k))$, where $x_l(k)$ and $x_r(k)$ are the left and the right channels, respectively. In our experiments, we noticed that this approach works well in practice, since commercial music recordings are usually produced so that the lead vocals are mixed at center panning to prevent canceling the vocals out if the recording is played in mono. However, leaving the separated accompaniment as mono signal may sound dull compared to original recording. Therefore, the method adds the side information, $s(k) = \frac{1}{2}(x_l(k) - x_r(k))$, to the separated accompaniment. This is illustrated in Fig. 3 where the left and the right channel of the separated accompaniment are simply obtained by $\widetilde{b}_l(k) = \widetilde{b}(k) + s(k)$ and $\widetilde{b}_r(k) = \widetilde{b}(k) - s(k)$, respectively. This approach was found to produce good results. The accompaniment separation could be performed individually for both the left and right channels. However, this doubles the computational load of separation and produced no audible improvement in our experiments.

4. TUNING THE USER SINGING

The user singing can be automatically tuned to the transcribed melody during a karaoke performance. There exist commercial software and hardware implementations of pitch-shifting tuners,
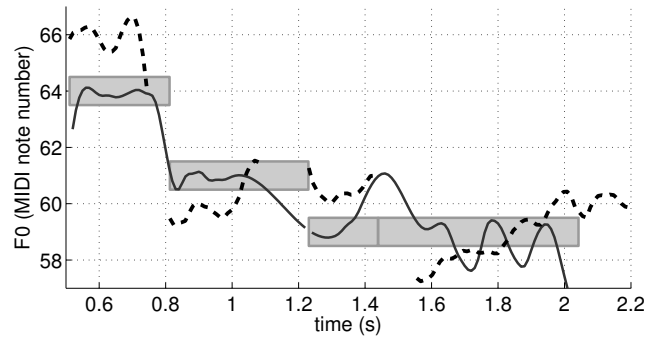
such as Antares Auto-Tune[2], which take the melody notes in MIDI format as input. Alternatively, singing can be tuned to the detailed F0 trajectory $\widetilde{f}_u$. In addition, user singing can be used to control the separation of the accompaniment so that the vocals in the original music piece are separated only when the user is singing. This minimizes the audibility of possible melody transcription errors.

The pitch shifting should preserve the formants of the user's singing voice to produce natural sounding output. For our purposes, we implemented a real-time frame-wise tuning algorithm. First, a singing frame of 23.2 ms is analyzed with the YIN algorithm [7] to give the fundamental frequency within the frame. The tuning method then measures the difference between the user singing F0 and the transcribed melody at the time. Then the Pitch Synchronous Overlap Add (PSOLA) pitch-shifting algorithm is used to tune the user singing in the frame to the original melody. We used a PSOLA implementation from [8].

Figure 4 shows an example of user singing compared to the transcribed melody and its detailed F0 trajectory. The gray boxes and the solid black line show the transcribed melody and its detailed F0 trajectory, respectively. The black dashed line shows the F0 values estimated from user singing with the YIN algorithm. During the first note, for example, user sings two semitones too high compared to the original melody. However, the PSOLA copes well with less than half-octave shifts up or down so the first note can be tuned to the original melody note with good quality.

Apart from real-time processing, the user singing could also be processed offline, which allows the use of temporal shifts, time stretching, and pitch shifting operations. For example, the user singing can be transcribed into notes and matched with the transcribed melody at note level, thus enabling very flexible editing possibilities.

5. ACCOMPANIMENT SEPARATION RESULTS

We evaluated the quality of the separated accompaniments using signal-to-noise ratio criterion. Given mixture signal $x(k)$ as input, the method produces $\widetilde{b}(k)$ which is an estimate of the accompaniment signal $b(k)$. The signal-to-noise ratio in decibels (dB) is then

---
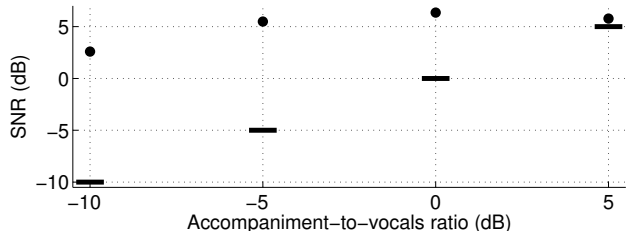
[2]See http://www.antarestech.com

1419

Fig. 5. Accompaniment separation results on the singing dataset with different mixing conditions, where the horizontal lines show the accompaniment-to-vocals ratios.

Fig. 6. Accompaniment separation results for the karaoke DVD songs. The dots denote the results and the horizontal lines the estimated accompaniment-to-vocals ratios.

defined by

$$\mathrm{SNR}(b, \widetilde{b}) = 10 \log_{10} \left(\frac{\sum_k b(k)^2}{\sum_k (b(k) - \widetilde{b}(k))^2} \right). \qquad (7)$$

We used two different datasets for the evaluations. The first set includes 65 singing performances consisting of approximately 38 minutes of mono audio. For each performance, the vocal signal $v(k)$ was mixed with a synthesized MIDI accompaniment signal $b(k)$ to obtain the mixture signal $x(k)$. The mixing was adjusted to obtain musically meaningful accompaniment-to-vocals ratios -10, -5, 0, and 5 dB. We evaluated $\mathrm{SNR}(b, \widetilde{b})$ for each performance and accompaniment-to-vocals ratio, and report the average over the performances.

Figure 5 shows the results for the first dataset. When accompaniment and vocals were equally mixed with accompaniment-to-singing ratio of 0 dB, the vocal signal was suppressed on the average by 6.3 dB in the separated accompaniment. When the accompaniment was most loud (5 dB case), the melody transcription was harder resulting in only 0.8 dB suppression of vocal signal. On the other hand, when vocals were most loud (-10 dB case), also the suppression was greatest (12.6 dB) but the overall $\mathrm{SNR}(b, \widetilde{b})$ was 2.6 dB due to louder non-harmonic proportions of vocals remaining in the separated accompaniment.

The second dataset consists of approximately twenty minutes of stereophonic recordings from nine songs from a karaoke DVD. The DVD contains an accompaniment version of each song and also a version with lead vocals. The two versions are temporally synchronous at audio sample level so that the accompaniment version can be used as a reference with two channels $b_l(k)$, $b_r(k)$ and the lead-vocal version as the stereophonic input $x_l(k)$, $x_r(k)$.

Figure 6 shows the accompaniment separation results for each song on the dataset. The SNR values were individually evaluated for left and right channels and their average, $(\mathrm{SNR}(b_l, \widetilde{b}_l) + \mathrm{SNR}(b_r, \widetilde{b}_r))/2$, is reported. The difference in the results between the two channels was negligible. On the average over all the songs, the method suppressed vocals by 3.5 dB. The figure also shows estimated accompaniment-to-vocals ratio of each song with horizontal lines, where the vocals were obtained by subtracting the accompaniment from the lead-vocal version. It is possible to have SNR lower than the accompaniment-to-vocals ratio, for example, by separating other instruments instead of vocals. This is the reason for lower quality with song number 8.
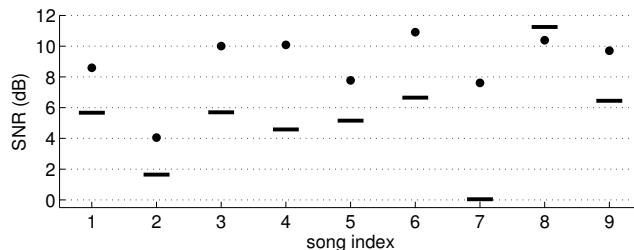
6. CONCLUSIONS

We have proposed a method for separating accompaniment from polyphonic music based on automatic melody transcription. This allows a user to obtain karaoke accompaniments from arbitrary commercial music recordings. The method provides a possibility for tuning the user singing to the original melody in real time. The melody transcription facilitates robust accompaniment separation based on the sinusoidal model. The quality of separated accompaniments was quantitatively evaluated using polyphonic music. The simulations show that the proposed method is able to suppress the original lead-vocal melody significantly. Audio examples of accompaniment separation can be found at http://www.cs.tut.fi/sgn/arg/matti/demos/karaoke.

7. REFERENCES

[1] 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 Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1564–1578, July 2007.

[2] Y. Li and D. L. Wang, "Separation of singing voice from music accompaniment for monaural recordings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1475–1487, May 2007.

[3] M. P. Ryynänen and A. P. Klapuri, "Automatic transcription of melody, bass line, and chords in polyphonic music," *Computer Music Journal*, vol. 32, no. 3, 2008, to appear.

[4] M. Ryynänen and A. Klapuri, "Transcription of the singing melody in polyphonic music," in *Proc. 7th International Conference on Music Information Retrieval*, 2006.

[5] Y. Ding and X. Qian, "Processing of musical tones using a combined quadratic polynomial-phase sinusoid and residual (QUASAR) signal model," *Journal of the Audio Engineering Society*, vol. 45, no. 7/8, 1997.

[6] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice Hall, 1993.

[7] A. de Cheveign and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

[8] P. Dutilleux, G. De Poli, and U. Zölzer, "Time-segment processing," in *DAFX - Digital Audio Effects*, U. Zölzer, Ed., chapter 7. John Wiley & Sons, 2002.