VocaRefiner: An Interactive Singing Recording System with Integration of Multiple Singing Recordings

Tomoyasu Nakano **Masataka Goto**
National Institute of Advanced Industrial Science and Technology (AIST), Japan
{t.nakano, m.goto}[at]aist.go.jp

ABSTRACT

This paper presents a singing recording system, *VocaRefiner*, that enables a singer to make a better singing recording by integrating multiple recordings of a song he or she has sung repeatedly. It features a function called *clickable lyrics*, with which the singer can click a word in the displayed lyrics to start recording from that word. Clickable lyrics facilitate efficient multiple recordings because the singer can easily and quickly repeat recordings of a phrase until satisfied. Each of the recordings is automatically aligned to the music-synchronized lyrics for comparison by using a *phonetic alignment* technique. Our system also features a function, called *three-element decomposition*, that analyzes each recording to decompose it into three essential elements: F_0, power, and spectral envelope. This enables the singer to select good elements from different recordings and use them to synthesize a better recording by taking full advantage of the singer's ability. Pitch correction and time stretching are also supported so that singers can overcome limitations in their singing skills. VocaRefiner was implemented by combining existing signal processing methods with new estimation methods for achieving high-accuracy robust F_0 and group delay, which we propose to improve the synthesized quality.

1. INTRODUCTION

When singers perform live in front of an audience they only have one chance. If they forget the lyrics or sing out of time with the accompaniment then these mistakes cannot be corrected, though singing out-of-tune could be fixed by using real-time pitch correction (e.g., Auto-tune or [1]). However, when vocals are recorded in a studio setting, the situation is quite different. Many attempts, or "takes", at singing the entire song, or sections within it, can be recorded. Indeed, if time and cost are not an issue, this process can continue until either the singer or someone else (*e.g.,* a producer or recording engineer) is completely satisfied with the performance. The vocal track which eventually appears on the final recording is often reconstituted from different sections of various takes and, to a greater and greater degree, subjected to automatic pitch correction (*e.g.,* Auto-tune) to "fix" any notes which are sung out of tune. What is left over at the end of this process is simply discarded as it is of no further use. This "standard" process of recording singing voice is summarized in the left side of Figure 1.

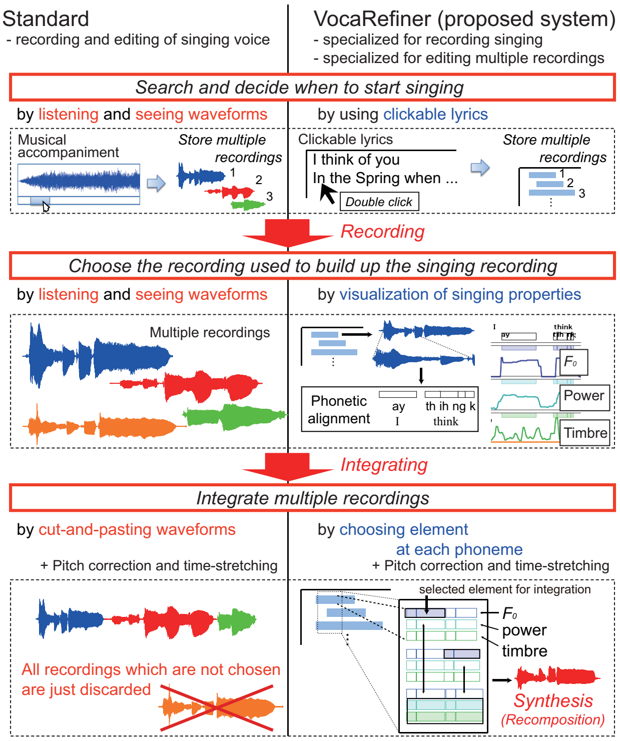

Standard
- recording and editing of singing voice
VocaRefiner (proposed system)
- specialized for recording singing
- specialized for editing multiple recordings
Search and decide when to start singing
by listening and seeing waveforms
by using clickable lyrics
Musical accompaniment
Store multiple recordings
1
2
3
Clickable lyrics
I think of you
In the Spring when ...
Double click
Store multiple recordings
Recording
Choose the recording used to build up the singing recording
by listening and seeing waveforms
by visualization of singing properties
Multiple recordings
Phonetic alignment
ay
I
th ih ng k
think
F0
Power
Timbre
Integrating
Integrate multiple recordings
by cut-and-pasting waveforms
by choosing element at each phoneme
+ Pitch correction and time-stretching
+ Pitch correction and time-stretching
All recordings which are not chosen are just discarded
selected element for integration
F0
power
timbre
Synthesis
(Recomposition)

Figure 1. A comparison of VocaRefiner with the standard recording and editing procedure.

Although this procedure for recording and editing vocals is widespread, it has some drawbacks. First, it is extremely time-consuming to manually listen through multiple takes and subjectively determine the "best" parts to be saved for the final version. Second, the manipulation of multi-track waveforms through "cut-and-paste" and the use of pitch correction software requires specialist technical knowledge which may be too complex for the amateur singer recording music in their home.

To address these shortcomings, we have developed an interactive singing recording system called *VocaRefiner*, which lets a singer make multiple recordings interactively and edit them while visualizing analysis of the recordings. VocaRefiner has three functions (shown in the right side of Figure 1) which are specialized for recording, editing and processing singing recordings.

1. *Interactive recording with clickable lyrics*: This allows a singer to immediately navigate to the part of the song he or she wants to sing without the need to visually inspect the audio waveform.

2. *Visualization of singing analysis*: This enables the singer to see an analysis of the recorded singing which captures three essential elements of singing voice: F_0 (pitch), power (loudness), and spectral envelope (voice timbre).

3. *Integration by recomposition and manipulation*: This allows the singer to select elements among multiple recordings at the phoneme level and recombine them to synthesize an integrated result. In addition to the direct recombination of phonemes, VocaRefiner also has pitch-correction and time-stretching functionality to give the user even more control over their performance.

The use of these three functions draws out the latent potential of existing singing recordings to the greatest degree possible, and enables the amateur singer to use advanced technologies in a manner which is both intuitive to use and enhances creative possibilities of music creation through singing.

The remainder of this paper is structured as follows. In Section 2 we present an overview of the main motivation, the target users for VocaRefiner, and the originality of this study. In Section 3 we describe VocaRefiner's functionality and usage. The signal processing back-end which drives VocaRefiner is described in Section 4 along with results on the performance of the F_0 detection method. In Section 5 we discuss the role and potential impact of VocaRefiner in the wider context of singing, and finally, in Section 6 we summarize the key outcomes from the paper.

We provide a website with video demonstrations of VocaRefiner at http://staff.aist.go.jp/t.nakano/VocaRefiner/.

2. VOCAREFINER: AN INTERACTIVE SINGING-RECORDING SYSTEM

This section describes the goal of our system and shortcomings of standard approaches. To achieve the goal and to overcome the shortcomings, we then propose our original solutions of VocaRefiner.

2.1 Goal of VocaRefiner

The aim of this study is to enable amateur singers recording music in their home to create high-quality singing recordings efficiently and effectively. Many amateur singers have recently started making personal recordings of songs and have uploaded them to video and audio sharing services on the web. For example, over 600,000 music video clips including singing recordings by amateur singers have been uploaded to the most popular Japanese video-sharing service *Nico Nico Douga* (http://www.nicovideo.jp). There are many listeners who enjoy such amateur singing which is illustrated by the fact that, as of April 2013 on Nico Nico Douga, over 4250 video clips by amateur singers received over one hundred thousand page views, over 190 video clips had more than one million page views, and the top five video clips had more than five million page views.

In Japanese culture, it is common for the singers not to show their faces in video clips. In this way, their recordings can be appreciated purely on the quality of the singing. In fact, amateur singers have become very well-known just by their voices and released commercially-available compact discs from recording companies. This is a kind of the new culture for music creation and appreciation driven by massive influx of user-generated content (UGC) on web services like Nico Nico Douga.

This creates a need and demand for making personal singing recordings at home. Most amateur singers record their singing voice at home without help from other people (*e.g.,* studio engineers). To fully produce the recordings, they must complete the entire process shown in the left side of Figure 1 by themselves. To create high-quality singing recordings, singers typically use traditional recording software or a digital audio workstation on a personal computer to recording multiple takes of their singing, again and again until they are satisfied. They then cut-and-paste multi-track waveforms and sometimes use pitch correction software (*e.g.,* Auto-tune). This traditional approach is inefficient and time-consuming, and requires specialist technical knowledge which may be a barrier for some would-be singers. We therefore study a novel recording system specialized for personal singing recording. Our eventual goal with this work is to facilitate and encourage even greater numbers of singers to create vocal recordings with better control and to actively participate in UGC music culture.

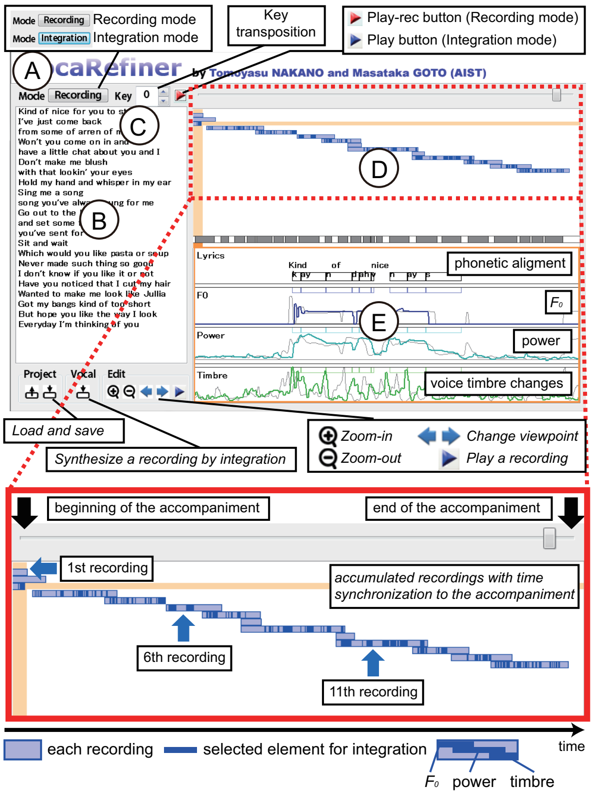
2.2 Originality of VocaRefiner

In this paper we present an alternative to the standard approach of recording singing voice by providing a novel interactive singing recording system *VocaRefiner*. It has an original efficient and effective interface based on visualizing analysis of singing voice and driven by signal processing technologies. We propose a novel use of the lyrics to specify when to start the singing recording and also propose an interactive visualization and integration of multiple recordings.

Although lyrics have already been clickable on some music players [2], they only allowed users to change the playback position for listening. VocaRefiner presents a novel use of lyrics alignment for recording purposes.

Multiple recordings were also not fully utilized for integration into the final high-quality recording, with most recordings being simply discarded if they are not explicitly selected. For example, recordings with good lyrics but incorrect pitch and recordings with correct-pitch singing but a mistake in the lyrics generally cannot be used in the final recording. However, VocaRefiner can make full use of bad recordings that would otherwise be discarded in the standard approach.

Although there has not been much research into the assistance of singing recording, some studies exist for visu-



Kind of nice for you to st
I've just come back
from some of arren of m
Won't you come on in and
have a little chat about you and I
Don't make me blush
with that lookin' your eyes
Hold my hand and whisper in my ear
Sing me a song
song you've alw ung for me
Go out to the
and set some
you've sent for
Sit and wait
Which would you like pasta or soup
Never made such thing so good
I don't know if you like it or not
Have you noticed that I cut my hair
Wanted to make me look like Jullia
Got my bangs kind of too short
But hope you like the way I look
Everyday I'm thinking of you

Mode Recording Recording mode
Mode Integration Integration mode
Key transposition
Play-rec button (Recording mode)
Play button (Integration mode)
A
VocaRefiner
by Tomoyasu NAKANO and Masataka GOTO (AIST)
Mode Recording
Key 0
C
D
B
E
Lyrics
Kind of nice
phonetic aligment
F0
F_0
Power
power
Timbre
voice timbre changes
Project
Vocal
Edit
Load and save
Synthesize a recording by integration
Zoom-in
Zoom-out
Change viewpoint
Play a recording
beginning of the accompaniment
end of the accompaniment
1st recording
accumulated recordings with time synchronization to the accompaniment
6th recording
11th recording
each recording
selected element for integration
time
F_0
power
timbre

Figure 2. An example VocaRefiner screen. The recordings are displayed as rectangles.

alizing analysis of singing voices to improve singing skills [3, 4]. Singing analysis has also been used for other purposes, such as time-stretching based on phase vocoder [5], voice-conversion [6], and voice-morphing [7]. However, we believe that no other research currently exists which deals with both the analysis and integration of multiple singing recordings as in VocaRefiner.

3. INTERFACE OF VOCAREFINER

The VocaRefiner system, shown in Figure 2, is built around components which encapsulate the following three functions:

1. Interactive recording by clickable lyrics
2. Visualization by automatic singing analysis
3. Integration by recomposition and manipulation

These functions can be used within the two main modes of VocaRefiner, "recording mode" and "integration mode", which are selected using button Ⓐ in Figure 2.

In recording mode the user first selects the target lyrics of the song they wish to sing (which can currently be in English or Japanese, marked Ⓑ) and loads the musical accompaniment.

To facilitate the alignment of lyrics with music and clickable lyric functionality, the representation of the lyrics must be richer than a simple text file containing the words of the song. It must also contain timing information - where each word has an associated onset time and the lyrics must also include the pronunciation of each word. It is possible to estimate this information automatically, however this process can produce some errors which require manual correction. Given the normal text file of lyrics, we therefore automatically convert it into the VocaRefiner format and then manually correct errors if any.

The accompaniment can include a synthesized guide melody or vocal (e.g. prepared by a singing synthesis system) to make it easier for the user to sing along with the lyrics. In the case where the user is recording a cover version of an original song they can include the original vocal of the song for this purpose.

If the user is unable to sing the song in original key, they can make use of a transposition function (marked Ⓒ), to shift the accompaniment to a more comfortable range.

3.1 Interactive Recording with Clickable Lyrics

The clickable lyrics function, which is built around the time-synchronization of lyrics to audio (described in Section 4.1), enables a singer who makes a mistake in the pitch or lyrics to start singing that part again immediately. Such seamless re-recording can offer a new avenue for recording singing, in particular for the amateur singer recording at home. One case where this could be particularly useful is when attempting to sing the first note of a song, where it can be hard to hit the right note straight away. Using clickable lyrics, the singer can repeat the phrase they will to sing recording each version until they are happy they have it right. By recording vocals in this way, a singer could also easily try different styles of singing the same phrase (storing each one aligned to the accompaniment), which could help them to experiment more in their singing style.

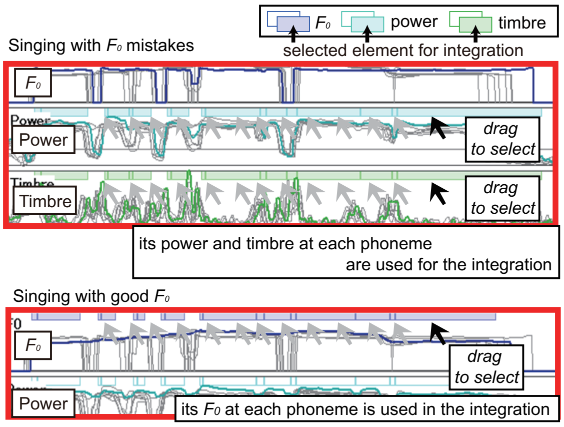
Because the lyrics and music are synchronized in time, when the singer clicks the lyrics, the accompaniment is played back on headphones (to prevent recording the accompaniment as well as the vocal) from the specified time and the voice sung by the user is recorded in time with the accompaniment. In addition, if the singer only wants to sing a particular section of the song, this section can be selected using the mouse.

The recording process can also be started by clicking the "play-rec" button indicated by the red triangle (close to Ⓒ) or by using the mouse to drag the slider located to the right of the button.

With this type of functionality, the clickable lyrics component can facilitate the efficient recording of multiple takes, where a singer can repeat an individual phrase over and over until they are satisfied. In this way, our work extends existing work into lyrics and audio synchronization [2], which has, up until now, only been applied to playback systems which cannot record and align singing input.

3.2 Visualization by Automatic Singing Analysis

Two types of visualization are implemented in VocaRefiner. The first of which addresses the timing information of multiple recordings. Each separate recording is indicated by a rectangle displayed at Ⓓ, as shown on Figure 2, whose length indicates its duration. The rectangles of multiple recordings, which appear stacked on top of one



F_0
power
timbre
Singing with F_0 mistakes
selected element for integration
F_0
Power
drag to select
Timbre
drag to select
its power and timbre at each phoneme are used for the integration
Singing with good F_0
F_0
drag to select
Power
its F_0 at each phoneme is used in the integration

Figure 3. Selecting voice elements to integrate.

another, can be used to see which parts of a song were sung many times, and can be useful for singers to find challenging parts requiring additional practice.

The second visualization shows the results of analyzing the singing recordings. This analysis takes place immediately after the each recording has taken place. First, the recording is automatically aligned to the lyrics via the pronunciation and timing using a *phonetic alignment* technique. VocaRefiner estimates and then displays *three elements* of each recording: F_0, power, and spectral envelope using techniques described in Section 4.2. These elements are used later for the recomposition of recordings from multiple takes.

An example of the analysis is shown at the point marked Ⓔ in Figure 2. The location of the rectangles in Figure 2 shows the onset and offset time of each phoneme. The blue line, the light green line, and the darker green line indicate trajectories of selected part used for integration of F_0, power, and voice timbre changes, respectively. The superimposed gray lines (which correspond to other recordings) are parts not selected for integration.

Such superimposed views are useful for seeing differences between the recordings without the need for repeated playback. In particular this can highlight recordings where the wrong note has been sung (without the need to listen back to the recording), and also show the singer the points where the timbre of their voice has changed.

3.3 Integration by Recomposition and Manipulation

The integration can be achieved by two main methods: "recomposition" and "manipulation" along with an additional technique for error repair. Their operation with VocaRefiner are described in the following subsections, and the technology behind them in Section 4.3.

3.3.1 Recomposition

The recomposition process involves direct interaction from the user where the elements they wish to use at each phoneme are selected with the mouse. These selected elements are used for synthesizing the recording.

In the situation where multiple recordings have been made for a particular section, VocaRefiner assumes that the most recently recorded take will be of good quality, and therefore selects this by default.

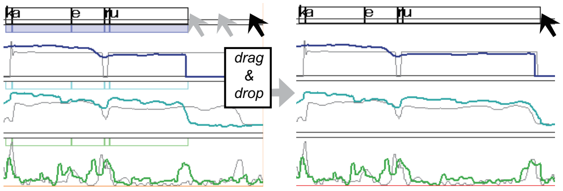


Figure 4. Time-stretching a phoneme. The length of the final phoneme /u/ is extended, and its F_0, power, and voice timbre are also stretched accordingly.

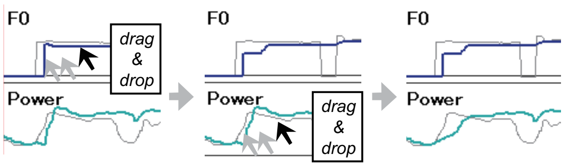


Figure 5. F_0 and power can be adjusted using the mouse.

3.3.2 Manipulation

Two modes of manipulation are available to the user, one which modifies the phoneme timing and the other which modifies the singing style. The modification of phoneme timing changes the phoneme onset and duration (via time-stretching), and the manipulation of singing style is achieved through changes to the F_0 and power.

A common situation requiring timing manipulation occurs when a phoneme is too short and needs to be lengthened. Figure 4 shows that when the length of the final phoneme /u/ is extended, the F_0, power, and spectral envelope of the phoneme are also stretched accordingly. Onset times can also be adjusted without the need for time-stretching.

Figure 5 shows that F_0 and power can be independently adjusted using the mouse. In addition to these local changes, the overall key of the recording can be also changed (Fig. 6) by global transposition.

3.3.3 Error Repair

Because occasional errors are unavoidable when recomposition and manipulation are based on the results of automatic analysis, it is important to recognize this possibility and provide the singer the means for correcting mistakes. The most critical errors that could require correction relate to the F_0 estimation and phonetic alignment. Such errors can be easily fixed through a simple interaction, as shown in Figure 7.

When an octave error occurs in F_0 estimation it can be repaired by dragging the mouse to specify the correct time-frequency range. In fact, octave errors can be eliminated by specifying the desired time-frequency range after recording. The more recordings of the same phrase there are, the easier it is to determine the correct time-frequency range, because the singer can make a judgement from many F_0 trajectories, where most have been correctly analysed.

Phonetic alignment errors are repaired by dragging the mouse to change the estimated phonetic boundaries. Fig-

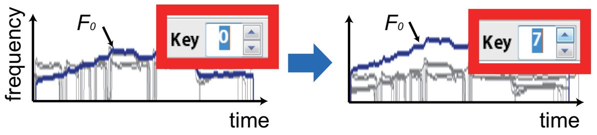


Figure 6. Example of a shift to a higher key.

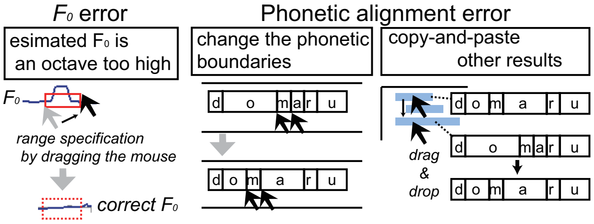

F_0 error
esimated F_0 is an octave too high
F_0
range specification by dragging the mouse
correct F_0
Phonetic alignment error
change the phonetic boundaries
copy-and-paste other results
d o m a r u
drag & drop

Figure 7. Error repair. An F_0 error is repaired by dragging the mouse to specify the correct time-frequency range (the red rectangle), and then a new F_0 trajectory is estimated from this range.

ure 7 shows the correction of the wrong duration of a phoneme /o/. Moreover, estimation results from other recordings can be used to correct errors by a simple copy-and-paste process. This function can be used to correct the situation where the alignment of a recording has many errors, for example, when a singer chose to hum the melody instead of sing the lyrics.

4. SIGNAL PROCESSING FOR THE IMPLEMENTATION

The functionality of VocaRefiner is built around advanced signal processing techniques for the estimation of F_0, power, spectral envelope and group delay in singing voice. While we make use of some standard techniques for this analysis, *e.g.*, F_0 [8, 9], spectral envelope [8], and group delay [10], and build upon our own previous work in this area [11,12] we also present novel contributions for F_0 and group delay estimation to meet the need for very high accuracy frequency and phase estimation in VocaRefiner. In evaluating the new F_0 detection method for singing voice (in Section 4.4), we demonstrate that our method exceeds the current state of the art.

Throughout this paper, singing samples are monaural solo vocal recordings digitised at 16 bit / 44.1 kHz. The discrete analysis time step (1 *frame-time*) is 1 ms. Time t in this paper is the time measured in frame-time units. All spectral envelopes and group delay are represented by 4097 frequency bins (8192 FFT length).

4.1 Signal Processing For Interactive Recording

Methods for estimating pronunciation and timing information and for transposing the key of the accompaniment are required for interactive recording. Phoneme-level pronunciation of English lyrics is determined using the CMU pronouncing dictionary[1] , and the pronunciation of Japanese

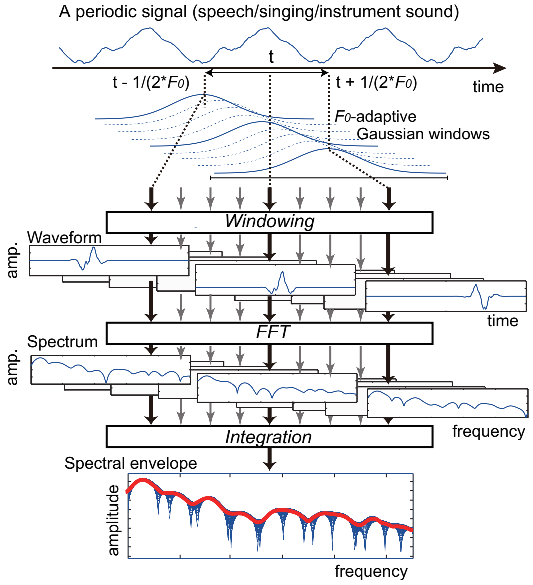


Figure 8. Overview of F_0-adaptive multi-frame integration analysis.

lyrics is estimated by using a Japanese language morphological analyzer MeCab[2] .

Timing information is estimated by first having the singer sing the target song once. The system then synchronizes the phoneme-level pronunciation of the lyrics with the recordings. This synchronization is called *phonetic alignment* and is estimated through Viterbi alignment with a monophone hidden Markov model (HMM). Two HMMs were trained with English and Japanese songs, respectively. The English songs came from the RWC Music Database (Popular Music [13], Music Genre [14], and Royalty-Free Music [13]) and the Japanese songs are in the RWC Music Database (Popular Music [13]).

When a singer wishes to transpose the key of the accompaniment in VocaRefiner, we use a well-known phase vocoder technique [5], which operates offline.

4.2 Signal Processing For Visualizing

A phonetic alignment method and three-element decomposition method are required for implementing this function. The phonetic alignment method is the same as that described above.

The system estimates the fundamental frequency (F_0), power, and spectral envelope of each recording.

$F_0(t)$ values are estimated using the method of Goto *et al.* [11]. $F_0(t)$ are linear-scale frequency values (Hz) estimated by applying a Hanning window whose length is 1024 samples (about 64 ms) and resampling at 16 kHz.

Spectral envelopes are estimated using F_0-adaptive multi-frame integration analysis [12]. This method can estimate spectral envelopes with appropriate shape and high temporal resolution. Figure 8 shows an overview of the analysis. First, F_0-adaptive Gaussian windows are used for spectrum analysis (F_0-*adaptive* analysis). Then neighborhood frames are integrated to estimate the target spectral envelope (*multi-frame integration* analysis).

[1] `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`

[2] `http://mecab.sourceforge.net/`

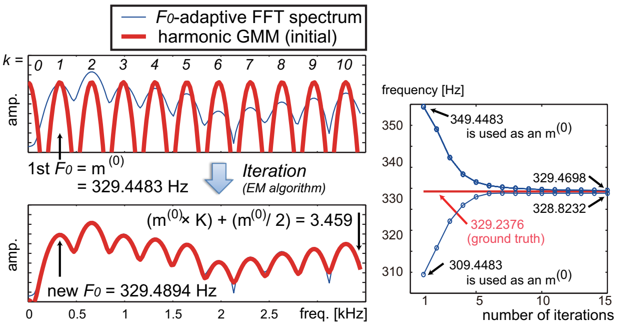


Figure 9. Iterative F_0 results estimated by the harmonic GMM.

The power is calculated from the spectral envelope by summation of the frequency axis at each time frame.

4.3 Signal Processing For Integration

For high-quality resynthesis, the three elements should be estimated accurately and with high temporal resolution. For this purpose we propose a new F_0 re-estimation technique, called F_0*-adaptive* F_0 *estimation method*. It is highly accurate and has the requisite high temporal resolution. To generate the phase spectrum used in resynthesis we also propose a new method for estimating group delay [10].

4.3.1 F_0-adaptive F_0 estimation method

Using the technique in [11] we perform an initial estimate of the F_0 which we call the *1st* F_0 and use this as input to the F_0-adaptive F_0 estimation method. The basic idea behind our new method is that high temporal resolution can be obtained by shortening the analysis window length for F_0 estimation as much as possible. Moreover we exploit the knowledge that harmonic components at lower frequencies of the amplitude spectrum of FFT can be used to estimate F_0 accurately, as they contain relatively reliable information whereas aperiodic components often dominant at higher frequencies.

To obtain high accuracy and high temporal resolution, we propose a harmonic GMM (Gaussian mixture model). We fit the GMM to the FFT spectrum estimated by an F_0 *adaptive analysis* that uses F_0-adaptive Gaussian windows and uses the 1st F_0 used as an initial value. Hereafter, the 1st F_0 is described as $m^{(0)}$.

We designed an F_0-adaptation window by using a Gaussian function. Let $w(\tau)$ be a Gaussian window function of time τ defined as follows, where $\sigma(t)$ is the standard deviation of the Gaussian distribution and $F_0(t)$ is the fundamental frequency for analysis time t.

$$w(\tau) = \exp(-\frac{\tau^2}{2\sigma(t)^2}) \tag{1}$$

$$\sigma(t) = \frac{\alpha}{F_0(t)} \times \frac{1}{3} \tag{2}$$

To set the value of α, we follow the approach for high-accuracy spectral envelope estimation in [15] and assign α=2.5.

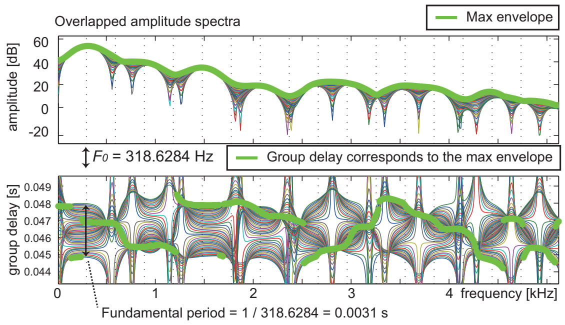


Figure 10. Overlapped STFT results showing the maximum envelope (top) and corresponding group delays (bottom).

A harmonic GMM $G(f; m, \omega_k, \sigma_k)$ for frequency f is designed as follows:

$$G(f; m, \omega_k, \sigma_k) = \sum_{k=0}^{K} \frac{\omega_k}{\sqrt{2\pi\sigma_k^2}} \exp(-\frac{(f-(m \times k))^2}{2\sigma_k^2}) \tag{3}$$

where K is the number of harmonics, for which K=10 was found to provide a high quality output. The Gaussian function parameters m, ω_k, and σ_k can be estimated using the well-known expectation and maximization (EM) algorithm, which is fitted to the F_0-adaptive FFT spectrum in the frequency range $[0, (K \times m^{(0)}) + m^{(0)}/2]$. In the iteration process of the EM algorithm, σ_k can be replaced with a range constraint, $[\epsilon, m]$, where $\epsilon = 2.2204 \times 10^{-16}$. The estimated m is used as the new estimated $F_0(t)$.

4.3.2 Normalized Group Delay Estimation Method Based on F_0-Adaptive Multi-Frame Integration Analysis

To enable the estimation of the phase spectrum for resynthesis, we propose a robust group delay estimation method. Although the previous method [12] relied upon pitch marks to estimate the group delay, the proposed method is more robust because it does not require them. The basic idea of this estimation is to use an F_0-adaptive multi-frame integration analysis based on the spectral envelope estimation approach in [12]. To estimate group delay, the F_0-adaptive analysis and a multi-frame integration analysis are conducted. In the integration, maximum envelopes are selected and their corresponding group delays are used as the target group delays. The group delay at each time can be estimated by using the method described in [10]. Figure 10 shows an example of extracting the maximum envelopes and corresponding group delays.

The estimated group delay has discontinuities along the frequency axis caused by the fundamental period. The group delay $\hat{g}(f,t)$ is therefore normalized with the range $(-\pi, \pi]$ and will be given by sin and cos functions as follows:

$$g(f,t) = \frac{\bmod\ (\hat{g}(f,t) - \hat{g}(\beta \times F_0(t), t), 1/F_0(t))}{F_0(t)} \tag{4}$$

$$g_\pi(f,t) = (g(f,t) \times 2\pi) - \pi \tag{5}$$

$$g_x(f,t) = \cos\left(g_\pi(f,t)\right) \tag{6}$$

$$g_y(f,t) = \sin\left(g_\pi(f,t)\right) \tag{7}$$

Here $\bmod\ (x,y)$ is a residual. The $\hat{g}(f,t) - \hat{g}(\beta \times F_0(t), t)$ component is used to eliminate an offset

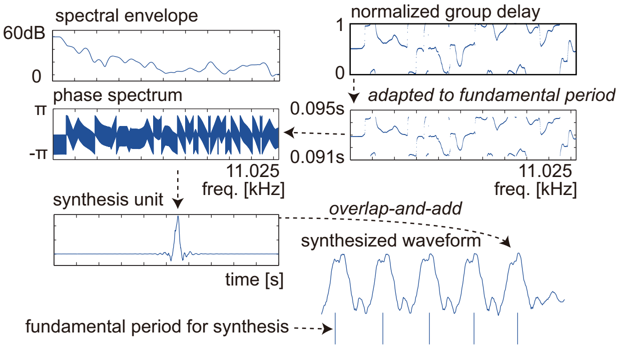


Figure 11. Singing synthesis by F_0-synchronous overlap-and-add method from spectral envelope and group delay.

of the analysis time, and β is set to 1.5 (an intermediate frequency between the first and second harmonics) as setting β=1.0 (the fundamental frequency) allowed undesirable fluctuations to remain.

There are also discontinuities along time axis. These are smoothed along both the time and frequency directions using a 2-dimensional FIR low-pass filter. Since the estimated group delay of frequency bins under F_0 is known to be unreliable, we finally smooth the group delay of bins under F_0 so that it can take the same value of the group delay at F_0.

4.3.3 *Singing Synthesis Using Normalized Group Delay*

The singing-synthesis method used to make the final recording needs to reflect integrating and editing results. Our implementation of singing synthesis from spectral envelopes and group delays is based on the well-known F_0-synchronous overlap-and-add method (Fig. 11).

The normalized group delays $g_x(f,t)$ and $g_y(f,t)$ are adapted to the synthesized fundamental period $1/F_0(t)_{syn}$ as follows:

$$g(f,t) \;=\; \frac{1}{F_0(t)_{syn}} \times \frac{(g_\pi(f,t)+\pi)}{2\pi} \tag{8}$$

$$g_\pi(f,t) \;=\; \begin{cases} \tan^{-1}(\frac{g_y(f,t)}{g_x(f,t)}) & (g_x(f,t)>0) \\ \tan^{-1}(\frac{g_y(f,t)}{g_x(f,t)})+\pi & (g_x(f,t)<0) \\ (3\times\pi)/2 & (g_y(f,t)<0, g_x(f,t)=0) \\ \pi/2 & (g_y(f,t)>0, g_x(f,t)=0) \end{cases} \tag{9}$$

Then the phase spectrum used to generate the synthesized unit is computed from the adapted group delay. The phase spectrum can be obtained by integration of the group delay, as in [10].

4.4 Experiments and Results

To evaluate the effectiveness of the iterative F_0 estimation method we examine its use when applied as a secondary processing stage on three well-known existing F_0 methods: Goto [11][3] , SWIPE [9], and STRAIGHT [8]. In each case we provide our iterative F_0 estimation method with the initial output from these systems and derive a new F_0 result. The frequency range is used as [100, 700] Hz for all the methods.

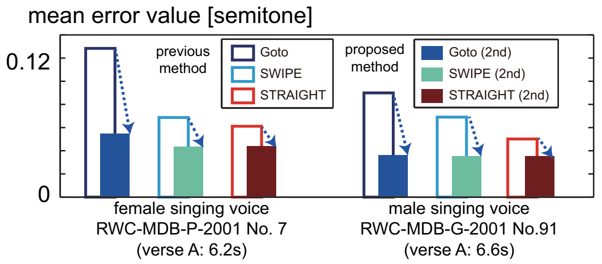


Figure 12. Estimation accuracies (mean error value) of the proposed re-estimation method (described as "2nd") compared with those of Goto [11], SWIPE [9], and STRAIGHT [8].

Estimation accuracy is determined by finding the mean error value, ϵ_f, defined by

$$\epsilon_f \;=\; \frac{1}{T_f}\sum_t |f_g(t)-f_n(t)| \tag{10}$$

$$f_n(t) \;=\; 12\times\log_2\frac{F_0(t)}{440}+69 \tag{11}$$

where T_f is the number of voiced frames, and $f_g(t)$ is the ground truth value. The $f_n(t)$ and $f_g(t)$ are log-scale frequency values relative to the MIDI note number.

To compare the performance of the algorithms, we use synthesized and resynthesized natural sound examples in the RWC Music Database (Popular Music [13] and Music Genre [14]). To prepare the ground truth, $f_g(t)$, we used singing voices resynthesized from natural singing examples using the STRAIGHT algorithm [8].

Results in Figure 12 show that the F_0 estimation across each of the methods is highly accurate, with very low, ϵ_f, both for male and female signing voice. Furthermore we can see that, for each of the three algorithms, the inclusion of our iterative estimation method improves performance. In this way, our iterative method could be applied to any F_0 estimation algorithm as an additional processing step to increase accuracy.

Regarding the estimation of spectral envelope and group delay, it is not feasible to perform a similar objective analysis. Therefore in Figure 13 we present a comparison between the estimated spectral envelope and group delay from a singing recording and a synthesized singing voice. By inspection it is clear that both the spectral envelope and group delay between the two signals are highly similar, which indicates the robustness of our method.

5. DISCUSSION

There are two ways to make high-quality singing content currently and in the future. One way is for singers to improve their voices by training with a professional teacher or using singing-training software. This can be considered the "traditional" way. The alternative is to improve one's singing "expression" skill by editing and integrating, *i.e.*, through practice and training with software tools. This paper presented a system for expanding the possibilities via this new emerging second way. We recognise that these two ways can be used for different purposes and have different qualities of pleasantness. We also believe that, in

[3] The 1st author reimplemented Goto's method for speech signals.

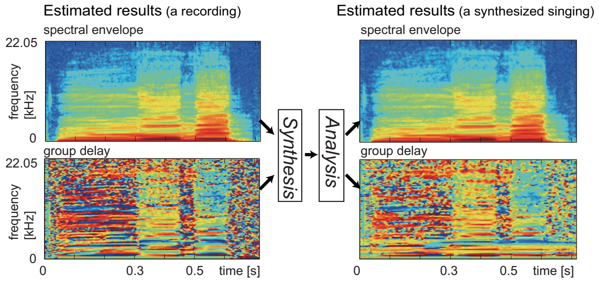


Figure 13. Examples of estimated spectral envelope and group delay and of analysis results for a synthesized singing voice.

the future, they could become equally important. A high-quality singing recording produced in the traditional way can create an emotional response in the listeners who appreciate the "physical control" of the singer. On the other hand, a high-quality singing recording improved using a tool like VocaRefiner can reach listeners in a different way, where they can appreciate the level of expression within a kind of "singing representation" created through skilled technical manipulation. In both cases, there is a shared common purpose of vocal expression and reaching listeners on a personal and emotional level.

The standard function of recording vocals has only focused on the acquisition of the vocal signal using microphones, pre-amps and digital audio workstations, etc. However, in this paper we explore a new paradigm for recording, where the process can become interactive. By allowing a singer to record their voice with a lyrics-based recording system opens new possibilities for interactive sound recording which could change how music is recorded in the future, *e.g.,* when applied to recording other instruments such as drums, guitars, and piano.

6. CONCLUSIONS

In this paper we present an interactive singing recording system called VocaRefiner to help amateur singers make high quality vocal recordings at home. VocaRefiner comes with a suite of powerful tools driven by advanced signal processing techniques for voice analysis (including robust F_0 and group delay estimation), which allow for easy recording, editing and manipulation of recordings. In addition, VocaRefiner has the unique ability to integrate the "best parts" from different takes, even down to the phoneme level. By selecting between takes and correcting errors in pitch and timing, an amateur singer can create recordings which capture the full potential of their voice, or even go beyond it. Furthermore, the ability to visually inspect objective information about their singing (*e.g.,* pitch, loudness and timbre) could help singers better understand their voices and encourage them to experiment more in their singing style. Hence VocaRefiner can also act as an educational tool.

In future work we intend to further improve the synthesis quality and to implement other music understanding functions including beat tracking and structure visualization [16], towards a more complete interactive recording environment.

Acknowledgments

We would like to thank Matthew Davies (CREST/AIST) for proofreading. This research utilized the RWC Music Database "RWC-MDB-P-2001" (Popular Music), "RWC-MDB-G-2001" (Music Genre), and "RWC-MDB-R-2001" (Royalty-Free Music). This research was supported in part by OngaCrest, CREST, JST.

7. REFERENCES

[1] K. Nakano, M. Morise, and T. Nishiura, "Vocal manipulation based on pitch transcription and its application to interactive entertainment for karaoke," in *LNCS: Haptic and Audio Interaction Design*, vol. 6851, 2011, pp. 52–60.

[2] H. Fujihara, M. Goto, J. Ogata, and H. G. Okuno, "Lyricsynchronizer: Automatic synchronization system between musical audio signals and lyrics," in *IEEE Journal of Selected Topics in Signal Processing*, 2011, pp. 311–316.

[3] D. Hoppe, M. Sadakata, and P. Desain, "Development of real-time visual feedback assistance in singing training: a review," *Journal of computer assisted learning*, vol. 22, pp. 308–316, 2006.

[4] T. Nakano, M. Goto, and Y. Hiraga, "MiruSinger: A singing skill visualization interface using real-time feedback and music cd recordings as referential data," in *Proc. ISMW 2007*, 2008, pp. 75–76.

[5] U. Zölzer and X. Amatriain, *DAFX - Digital Audio Effects*. Wiley, 2002.

[6] T. Toda, A. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Trans. ASLP*, vol. 15, no. 8, pp. 2222–2235, 2007.

[7] H. Kawahara, R. Nisimura, T. Irino, M. Morise, T. Takahashi, and H. Banno, "Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown," in *Proc. ICASSP 2009*, 2009, pp. 3905–3908.

[8] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch adaptive time-frequency smoothing and an instantaneous frequency based on F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.

[9] A. Camacho, *SWIPE: A Sawtooth Waveform Inspired Pitch Estimator for Speech And Music*. Ph.D. Thesis, University of Florida, 2007.

[10] H. Banno, L. Jinlin, S. Nakamura, K. Shikano, and H. Kawahara, "Efficient representation of short-time phase based on group delay," in *Proc. ICASSP1998*, 1998, pp. 861–864.

[11] M. Goto, K. Itou, and S. Hayamizu, "A real-time filled pause detection system for spontaneous speech recognition," in *Proc. Eurospeech '99*, 1999, pp. 227–230.

[12] T. Nakano and M. Goto, "A spectral envelope estimation method based on F0-adaptive multi-frame integration analysis," in *Proc. SAPA-SCALE Conference 2012*, 2012, pp. 11–16.

[13] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music databases," in *Proc. ISMIR 2002*, 2002, pp. 287–288.

[14] ——, "RWC music database: Music genre database and musical instrument sound database," in *Proc. ISMIR 2003*, 2003, pp. 229–230.

[15] H. Kawahara and M. Morise, "Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework," *Sadhana: Academy Proceedings in Engineering Sciences*, vol. 36, no. 5, pp. 713–727, 2011.

[16] M. Goto, K. Yoshii, H. Fujihara, M. Mauch, and T. Nakano, "Songle: A web service for active music listening improved by user contributions," in *Proc. ISMIR 2011*, 2011, pp. 311–316.