A sustained vowel replacing algorithm based on iterative formant filtering

Chun-Tang Hsu; Yi-Wen Liu
Department of Electrical Engineering, National Tsing Hua University, Taiwan

ABSTRACT
In electronic music, formant filtering refers to transforming a source signal to sound like vowels. However, existing plugins, such as iZotope VocalSynth 2, and The Orb do not focus on preserving the timbre of the source audio after filtering. In this paper, an algorithm is proposed for replacing a sustained vowel of one singer with another vowel while preserving the timbral characteristics. Here we assume that a timbre source signal (TSS) and a vowel source signal (VSS) are available by recording from singer 1 and 2, respectively. Spectral envelopes below 5kHz of both signals are extracted first. Then, the ratio of the two envelopes is used for fine-tuning the TSS power spectrum and this process will be repeated iteratively. Upon completion, the output file is meant to preserve the timbre of singer 1 while its vowel identity, including the accent, has been changed to that of singer 2. To evaluate the synthesis quality, sustained vowels sung by 11 singers were recorded, and the proposed algorithm was applied for cross-synthesis. 11 subjects are invited to tell if synthesized timbre resembles singer 1 or singer 2. Results show that 7 subjects successfully chose singer 1 with accuracy > 70%.

Keywords: Vowel, Formant filtering

1. INTRODUCTION
Human brains could extrapolate the timbre features of a specific person upon hearing few or one instance of his sustained vowel voice (1). For instance, people could imagine what the sustained vowels /a/, /i/, /u/, /e/ of an individual might sound like by listening to his sustained vowel /o/ only. This transformation process only modifies phoneme category rather than timbre characteristic, so people still know the voice is pronounced by the same person.

In speech synthesis, the spectral envelope of the power spectrogram determines the timbre of the vowel. Speech is fully intelligible through the telephone bandwidth (nominally only 200-3200 Hz) (2). Plus, the vowel category is determined by first few formant positions. Therefore, the low-frequency band mostly contributes to which vowel the given audio sounds like. Human’s vowel perception can be generalized by the first three formant frequencies (3). For example, a formant filtering plugin called The Orb (4) provides a function that transforms any sound into a vowel-like formant-filtered sound with a set of three band-pass filters. However, filtering signal with few band-pass filters would discard voice details such as high frequency component, leading to some degradation of audio quality. Other plugin like iZotope VocalSynth 2 (5) uses scientific modeling of the human vocal tract to adjust human vocal characteristics like nasality, vowel shapes, and formants for a smooth, textural vocal treatment to any audio signal. Although the plugin provides flexible controlling parameters for voice manipulation, it does not yet support vowel transformation like synthesizing transformed vowel according to the timbre of the input audio.

In this work, an iterative formant filtering algorithm is proposed to simulate the extrapolation ability mentioned above. The method is applied to transform one of Japanese 5 vowels, namely /a/, /i/, /u/, /e/, /o/, recorded from a subject to any other 4 vowels without changing one’s timbre characteristic. In previous work, Bohn (6) has already proposed a formant filtering method to achieve this kind of effect. However, in that work the only resynthesis experiment was transforming a Hungarian word “fésü” (fE:SU) to “fásü” (fA:SU). Whether other transformation combinations might work properly as well remains unknown.
calculate the ratios of formants of the sounds and derive a multiplication factors to modify spectral envelope. This might not be a problem if the source spectrogram has clear harmonic series at 0–5 kHz. As shown in Figure 1, the harmonic series of /e/ is prominent at 0 to 5 kHz and the pronunciation of /é/ is similar to Hungarian “é” in “fésű”. However, other vowel such as /u/ does not show prominent harmonic series above 1 kHz. Whenever the vowel /u/ is transformed to the other vowel with clear harmonic series above 1 kHz, such as /e/, the high-frequency part of /e/ cannot be fully recovered by multiplication in the spectral domain.

2. Data preparation

2.1 Dataset

A glissando dataset was built for synthesis and evaluation. 26 subjects (18 males, 8 females) between the age of 22 and 25 were recruited, including 1 female professional singer, 5 amateur male singers and 20 singers who did not receive any singing training. Each subject was told to sing vowels /a, i, u, e, o/ respectively in glissando form. The start and end of the pitch were chosen by the subject, and they could use their most comfortable pitch range to sing. The recording was done in a recording studio with Surface Pro 4 laptop microphone. The digital sampling rate are 44100 Hz. A subject was asked to sing the glissando vowel /a/ and the DIO algorithm (7,8) in WORLD vocoder (9) was used to analyze the pitch contour of the recorded audio data. The pitch of the other four vowels, which are /i, u, e, o/, sung by the subject should cover the same range of /a/.

To ensure that the sound quality of recording data is reliable, each audio files needs to pass a judgment program with the assistance from the WORLD vocoder. The judging criteria are elaborated below:

- The pitch analyzed by DIO algorithm should not contain any pitch doubling.
- The pitch difference between two adjacent frames should be less than 0.35 note. The frame length is 10 ms with 0 overlap.
- Human-based pronunciation check

If the recorded audio files failed to satisfy the criteria listed above, they would be either discarded or fixed by some tricks described below:

- DIO algorithm returns 0 if the pitch of the specific segment is indeterminable. These frames were automatically discarded.
- The frames were sorted in pitch order before running pitch resolution test
- If there was a slight defect in the audio file, the subject is allowed to record two audio files and merge them by crossfading. However, this approach might lead to some discontinuous pattern when WORLD vocoder resynthesize the glissando audio, so subjects would be asked to try their best to complete recording at one go.
2.2 Sustained voice resynthesis

As shown in Figure 2, for each glissando audio files, WORLD vocoder is used to extract audio features, which include smoothed spectral envelope, aperiodic parameter, and f0. When resynthesizing sustained voice, a pitch value is selected. An audio frame with the pitch value closest to the selected pitch value would be used for synthesis. Finally, a sustained voice can be synthesized by duplicating the frame by N times and feeding the features back into WORLD vocoder.

3. The Vowel Transformation Algorithm

The transformation process consists of three steps: Spectral envelope extraction, VSS iterative envelope fitting, and synthesis. Since the magnitude perception of human ears is logarithmic, each frame’s power spectrum is transformed to log scale,

\[ \Omega = \log_{10}(|F(w(t))|) \]  

where w(t) is window function and F stands for Fourier transform. In our case, the Hann window is applied and the window length and hop size are 8192 and 441 respectively. Because the most vowel characteristic presents in the low frequency part, for male, only the first 500 of 8192 FFT points (approximately 0–2700 Hz) will be subsequently subject to filtering during the iteration process. Female sound input requires 800–1000 points to represent the vowel features because female generally has higher formant frequency (10).

3.1 Spectral envelope extraction

Spectrums of the same vowels at different pitch share almost the same spectral envelope (11). Once the filter’s envelope is derived, it can be applied on input source with any arbitrary pitch. Several techniques have been proposed to extracted spectral envelope and the most commonly used approach are linear prediction (LP) (12) and the cepstrum (13) method. Nonetheless, both methods have some limitations: The LP envelope tends to overestimate peaks (14), and the cepstrum envelope contains too much ripples, leading to pitch shifting after performing formant filtering (see Figure 4).

Given a power spectrum with logarithm magnitude, the spectral envelope is extracted by performing cubic spline interpolation alone peaks. To extract peak locations, an algorithm is devised to search along the frequency axis for local maximums. The extraction target is the peak of each harmonic series. Since the harmonic peaks are almost equally spaced in linear frequency scale, a minimum horizontal distance between

Figure 3 – “x” marks stand for peaks founded by the proposed algorithm.
two peak is set to avoid finding extra peaks between two harmonic component. Figure 3 shows the result of peaks extraction algorithm performed on the power spectrum of vowel /a/. Finally, each peaks are interpolated with a cubic spline line, forming a spectral envelope. The result is shown in Figure 4. This method is quite similar with the Empirical Envelope Estimation Algorithm (EEEA) proposed by Meng (15). The different part is that the proposed algorithm uses all harmonic peaks instead of prominent peaks when performing spline line interpolation.

![Figure 4 – Spectral envelope /a/ extraction result.](image)

### 3.2 VSS iterative envelope fitting

In this part, the power spectrum of TSS is filtered iteratively until it fits that of VSS. To achieve this, both the envelopes of TSS and VSS are extracted first. A transformation filter is derived by calculating the ratio of VSS envelope to TSS envelope. After applying the filter to the TSS power spectrum, a new spectral envelope will be calculated again. The iterative filtering process would make the spectral envelope of TSS approximate to VSS’s spectral envelope.

Since the modified 500 points power spectrum will be combined with the rest of frequency bins at the synthesis stage, the connection part might have large discontinuity. To address this issue, the last few points of 500 points filter are multiplied by a fade out kernel when performing formant filtering. Note that this process is also done on log magnitude spectrum. The kernel is defined as $0.7^n$ where $n$ is the kernel size. Empirically, the length of the fade out kernel is chosen to be 20.

![Figure 5 – A demonstration of the iterative process of formant filtering. In this case, the vowel /i/ is transform to /a/.](image)
3.3 Synthesis

Finally, the power spectrum of each frame is reconstructed by juxtaposing the 500 points with the rest
high frequency bins. Combining the modified power spectrum and the original phase spectrum of the
timbre source, the audio can be reconstructed by calculating the inverse Fourier transform of each frames
and performing overlap and add.

4. Evaluation

A questionnaire is devised in order to justify whether the synthesized audio actually preserves the
timbre of TSS. The questionnaire consists of ten questions and each question has seven scales (0-6) that the
evaluators can choose. For each question, subjects would be asked to listen to three audio files. The first
two files are WORLD vocoder resynthesized voices, which are the same vowel sung by two different
singers (singer 1, singer 2) at the same pitch. The third file is synthesized audio which takes singer 1 as
TSS and singer 2 as VSS. Subjects should tell to what extent the synthesized audio’s timbre sounds like
singer 1 or singer 2. If the audio sounds 80% more like singer 1, subjects should choose “0”. Similarly,
subjects should choose “6” if the audio sounds 80% like singer 2. The scale size is 10 %, and the center
scale “3” indicates that the subject can not tell who the synthesized audio sounds like. For example,
subjects would listen to the sustained vowel /i/ of singer 1 and 2 at the begin. Then, they have to listen to a
sustained vowel /a/ audio synthesized by singer 1’s /i/ (TSS) and singer 2’s /a/ (VSS) sustained vowel. The
goal is to verify the synthesized voice preserves the timbre characteristic of singer 1 and the vowel
characteristic of singer 2.

Among 10 questions, first 5 questions synthesized /a/ by singer 1’s /i/ and singer 2’s /a/ and another 5
synthesized /i/ by singer 1’s /a/ and singer 2’s /i/. The reason for examining /a/ and /i/ was these two
vowels’ power spectrums had less similarity as shown in Figure 1. Also, boosting the low amplitude of /i/
vowel at 500 Hz to 2000 Hz might be challenging. Therefore, testing the performance of converting /i/ to
/a/ would be a good indicator for the algorithm. The audio files used in the questionnaire were generated by
randomly picking several singers from the glissando dataset. This experiment only includes male singing
voice. Using different gender’s TSS and VSS could change the timbre of TSS significantly, unless the pitch
of both sources were close. Accordingly, 11 male singers were chosen for generating the 10 questions
mentioned above.

Figure 6 – The result of the hearing test questionnaire. The first row (Q0-Q4, from left to right)
examines the performance of synthesized voice /a/ by taking /i/ as input and the second row (Q5-Q9)
examines /i/ synthesized by /a/.

Figure 6 shows the response of each question. Both TSS and VSS would be randomly assigned a
number 0 or 6 and the subjects have no idea which source signal belong to which number. According to the
result of figure 6, the second case, e.g., /i/ synthesized by /a/ is more likely for subjects to believe that the
synthesized voice is similar to TSS. Also, the synthesis quality is case sensitive: while the response of Q8 is
unanimously correct, Q6’s timbre similarity to TSS is not as recognizable. The reason for this could be that
too many iterations are run or too many interpolation points are used for interpolation so the spectral
envelope of TSS overly approximates that of VSS.
5. DISCUSSIONS

Our algorithm successfully transfers one’s vowel into any other four Japanese vowels while the timbre characteristic remains unchanged. However, it has some limitations. First, the pitch of both sources should not be too much different. The difference might introduce some noise in the low frequency part of the synthesized vowel. Consequently, it is highly recommended to tune the pitch of VSS and TSS to the same level before filtering. Second, the performance of the algorithm highly depend on the accuracy of spectral envelope, so the incorrect extraction result such as finding some wrong peaks locations would yield a degraded or distorted result. Last but not least, our algorithm has better performance when TSS and VSS come from the same singer. If different gender’s TSS and VSS are used, the pitch of both sustained singing signals need to be close. When same pitch signals are not available, tuning the pitch of input signal by a formant preservation method (13) might be required.

6. FUTURE WORK

When different singer sources are used for synthesis, the timbre characteristic of filtered results might lay between two singers and the degree depends on how many iterations are run and how wide the filter is. This shows that spectral envelope is one of the key point that makes a voice sound like a specific person. Currently, to achieve the best synthesis quality, parameters such as kernel length, number of points used for interpolation, iteration times are chosen empirically. In the future, an automatic parameters selecting strategy need to be proposed to ensure the stability and generality of the algorithm.

ACKNOWLEDGEMENTS

This research is supported by the Ministry of Science and Technology of Taiwan (MOST 108-2634-F-007-003).

REFERENCES