DIGITAL PATTERN PLAYBACK FOR EDUCATION IN DIGITAL SIGNAL PROCESSING AND SPEECH SCIENCE

Takayuki Arai

Dept. of Information and Communication Sciences
Sophia University, Tokyo, Japan

ABSTRACT

We developed a digital version of Pattern Playback to convert a spectrographic representation of speech back into a speech signal. Pattern Playback was originally developed by Cooper and his colleagues from Haskins Laboratories in the late 1940s. We used our Digital Pattern Playback (DPP) for instruction in digital signal processing and speech science. The original DPP used two different algorithms: amplitude modulation and fast Fourier transform. The new DPP uses additive synthesis of sinusoidal harmonics, which is easier for undergraduate college students to understand. We also designed a scientific exhibition with DPP at a science museum for children and adults. DPP is educational for a wide variety of people, from children to technical students.

Index Terms— digital pattern playback, education, speech science, sound spectrogram, voice-print puzzle

1. INTRODUCTION

In this study, we developed Digital Pattern Playback (DPP) [1,2], a digital version of Pattern Playback. Pattern Playback (PP) is a device that converts a spectrographic representation of speech back into a speech signal. PP was developed by Cooper and his colleagues from Haskins Laboratories in the late 1940s.

PP has contributed greatly to the development of research in speech science [3-5]. By converting a spectrogram into sound, it is possible to test which acoustic cue projected on the sound spectrogram is important for speech perception. One can simplify the acoustic cue and/or systematically change an aspect of the acoustic cue, redraw a spectrographic representation, and synthesize stimulus sounds. The locus theory is an example of a study which shows the importance of PP, as it revealed the importance of the second formant trajectory of a following vowel for the perception of a preceding stop consonant [6].

The DPP we developed [1,2] was based on two different algorithms: amplitude modulation (AM) and fast Fourier transform (FFT). In the AM method, the amplitudes of the harmonics were modulated by the darkness pattern of a spectrogram. This concept is similar to the original PP, which was based on the source filter theory of speech production. In the original PP, a light source and tone wheel generated an optical set of harmonics, and the amplitudes of the harmonics were optically modulated by a given spectrogram. Our algorithm in [1,2] simulated a similar process in digital form.

In the FFT method, a time slice of a given spectrogram is treated as a logarithmic spectrum of that time frame, and the spectrum is converted back into the time domain by the inverse FFT. In this method, we set the phase components to zero. This is because our goal is not to perfectly reconstruct the original speech signal, but to implement the PP digitally for pedagogical purposes.

Both the AM and FFT methods conceptually achieve the same result; the main difference between them is how we view the process, from either the time or frequency domain.

Our first goal was to re-design the DPP algorithm for lab experiments targeting second-year undergraduate students in the Faculty of Science and Technology at Sophia University. Given our target audience, we avoided the concepts of amplitude modulation and fast Fourier transform. Instead, we based our algorithm on additive synthesis of sinusoidal harmonics, which is more easily understood at this level. The resulting algorithm is based on Fourier synthesis and has the same characteristics as the AM and FFT methods.

Our second goal was to apply this DPP technique to a hands-on activity for children and adults in order to show that DPP is useful for everyone.

In this paper, we first use our DPP algorithm in a lab experiment. We then describe an exhibition of DPP at a science museum.

2. DPP ALGORITHM BASED ON ADDITIVE SYNTHESIS OF SINUSOIDAL HARMONICS

Our newly proposed DPP algorithm is based on additive synthesis of sinusoidal harmonics as shown in Fig. 1. In this algorithm, a time slice of a given spectrogram, or a short-time spectrum, is treated as a Fourier series of that time frame, and the Fourier coefficients are converted to the time-domain waveform by Fourier synthesis. As in the FFT method, we do not reconstruct the original phase; we simply set the phase components to zero (the Fourier cosine synthesis). For lab experiments at the undergraduate level, we also kept the simplicity of the original PP, including no pitch change during playback.

2.1. New algorithm for DPP

In this algorithm, we first reduce the frequency resolution of each short-time spectrum to obtain only the spectral envelope of a given time frame (especially for a spectrogram obtained by a narrow-band analysis). This reflects the vocal-tract filter of a given instance. Then, each harmonic cosine waveform whose fundamental frequency is the same is multiplied by the corresponding Fourier cosine coefficient obtained from the short-time spectrum ("*a*" coefficients in Fig. 1).

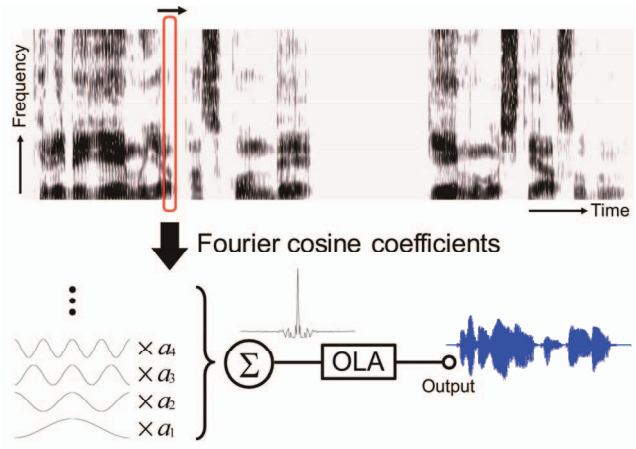

Fig. 1 Schematic representation of DPP algorithm based on the additive synthesis of sinusoidal harmonics. The overlap-add (OLA) technique was applied depending on the desired f0 contour.

```
Fs = 16000;          % sampling frequency (in Hz)
f0 = 100;            % fundamental frequency (in Hz)
T = Fs/f0;           % fundamental period (in sample)
f_pixel = 80;        % # pixel of the vertical axis
t_pixel = 240;       % # pixel of the horizontal axis
S = imread('samples.jpg');    % read image file
output = zeros(T*t_pixel,1);  % initialization
for frame=1:t_pixel           % loop for time frame
   ns = T*(frame-1)+1;        % start sample
   ne = ns+T-1;               % end sample
   for k = 1:f_pixel          % loop for Fourier syn.
      A = 255-double(S(f_pixel-k+1,frame));
                              % Fourier coefficient
      output(ns:ne) =         % adding cos func.
          output(ns:ne)+A*cos(2*pi*k*[-T/2:T/2-1]/T)';
   end
end
soundsc(output, 16000)        % play output
```

Fig. 2 Matlab script for additive synthesis of sinusoidal harmonics for DPP.

Finally, by adding the harmonic cosine waveforms, each of which has the Fourier cosine coefficient as its amplitude, we get a periodic temporal waveform. This periodic waveform can be viewed as the response of the vocal-tract filter of that time frame. The input signal is the impulse train with a fundamental period which is the reciprocal of the fundamental frequency. We truncate one cycle of the periodic waveform and treat it as an impulse response of the vocal-tract filter for that time frame. Finally, we place the impulse response along the time axis frame-by-frame. If we place the impulse responses based on the desired instantaneous fundamental frequency (f0), the resulting waveform can have a time-varying f0 contour. When the impulse responses overlap each other, the overlap-add (OLA) technique was applied (Fig. 1). Because of the simplicity of the lab experiments at the second-year undergraduate level, we use a constant f0 contour. We carefully select a common time duration for the fundamental period for Fourier synthesis and the desired fundamental period of the constant f0 contour, so that the OLA is a simple concatenation.

Thus, Fourier synthesis adding the harmonic cosine waveforms with the same fundamental frequency yields an impulse response of the filter at a given time frame. In general, the impulse response can have a longer duration than the fundamental period of the Fourier synthesis, and the amplitudes of the edges of one cycle can be greater than zero. However, if we use a longer fundamental period for the Fourier synthesis, we can reduce the edge effect. To increase the fundamental period, we need more samples in the frequency domain to increase the frequency resolution of the short-time spectra. Fourier synthesis with cosine waveforms (zero phase) yields even smaller amplitudes at the edges than synthesis with other phase functions, such as, the sine transform. Therefore, we used the cosine transform and truncated the periodic waveform with the fundamental period.

In theory, we can use a variety of sets of values for each parameter. Figure 2 shows a Matlab script that we use for the Lab experiments of the undergraduate course at Sophia University. In this case, we used a sampling frequency of 16 kHz and a fundamental frequency of 100 Hz, so the fundamental period was 160 samples. An input image file is 'samples.jpg', and the input image *S* with the size of 240 x 80 pixels is converted into a speech signal 'output' by adding the harmonic cosine waveforms. Figure 3 shows an example of an English sentence. The parameters here are the same as those used in Fig. 2.

2.2. Evaluation

In the fall semester of the academic years from 2009 through 2011, we used the DPP algorithm in lab experiments for second-year college students of the Department of Information and Communication Sciences at Sophia University. First, we gave the students a simple introduction to the Matlab language and signal processing, including Fourier synthesis. We then let them analyze an audio signal by using a function to compute a sound spectrogram. After that we used the Matlab script in Fig. 2 to see how a spectrogram can be converted back to a speech sound. With this script, students can learn/review the concept of Fourier synthesis and the mechanism of speech production.

Many students became interested in the output speech because it is relatively intelligible, enabling them to guess the contents of the utterance. It was successful enough to pique the students' interest and teach them the basics of speech signal processing and speech science. Some students, for example, showed their enthusiasm by trying to change the constant pitch frequency to another frequency or change the speaking rate of the output sound.

After the lab experiment, each student wrote a report. Responses varied, but in general feedback was positive. Following are some comments made by students:

- I did not understand the concept of Fourier synthesis when I learned it in math class, but I now have better understanding of this topic. Now I know even a complex speech signal consists of a set of simple sinusoidal waveforms.
- It was fun to learn this topic, because I was able to have a better understanding of what type of information is important in a speech signal.

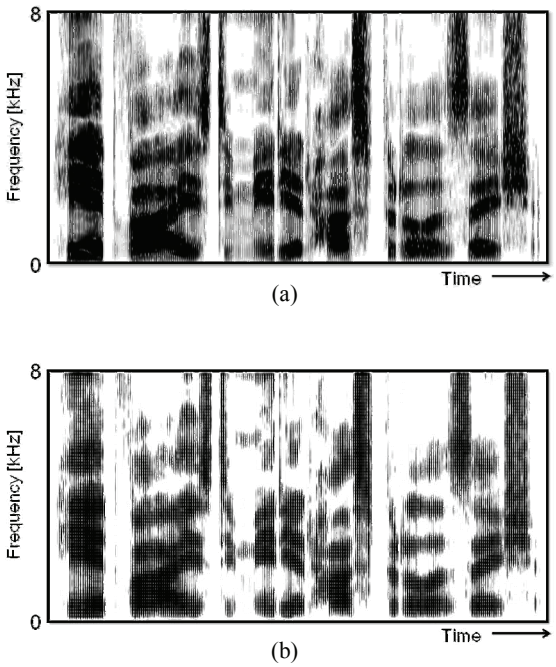

Fig. 3 Sample result for an English sentence. (a) original and (b) synthesized signals.

Table 1 Number of students for the answers of two questions of the questionnaire for the lab experiment.
Q1) Did you understand the algorithm of the DPP?
Q2) Did you get interested in the DPP?

		Q2		
		Not at all	Moderately	Very much
Q1	Not at all	4	7	3
	Moderately	2	35	43
	Very much		1	4

- I was surprised because I did not know a speech signal is replicable from a sound spectrogram.

In 2011, we asked students to answer to the questionnaire after the lab experiment, and we collected answers from 112 students. Table 1 shows number of students for the answers based on a part of the questionnaire. From this table, approximately 86% of students understood the algorithm of the DPP and approximately 94% of students got interested in the DPP. We found that not only the students who understood the algorithm but also the students who did not fully understand it got interested in the DPP.

3. DPP AT A SCIENCE MUSEUM

DPP is accessible not only for college and graduate students, but also for younger children. We found that a DPP demonstration is even effective when we capture a printed spectrogram on a sheet of paper with a camera [2]. This is true not only for graduate or undergraduate students, but also younger generation including young children. Furthermore, since hands-on activities are often more suitable for children [7], we designed an exhibition called "Voice-Print Puzzle," originally designed for a science museum, Sony ExploraScience, in Tokyo [8].

Figure 4 shows the setup used at the Sony ExploraScience exhibition. In this exhibition, we converted speech sounds into spectrographic form, and printed each spectrogram on a plastic plate with the corresponding label in Japanese orthography. The speech sounds were a subset of Japanese mora (i.e. syllables) composed of either a single vowel or a consonant followed by a vowel. Some of the plates are shown in the box at the lower righthand corner of Fig. 4.

Then, we set up a DPP environment with a video camera as shown on the top of Fig. 4. The input image was converted into a speech sound by the DPP algorithm running on a hidden computer. In this exhibition, we used a digital video camera, but a small web camera could also be used in a smaller setting, such as a classroom.

Fig. 4 Voice-Print Puzzle at Sony ExploraScience.

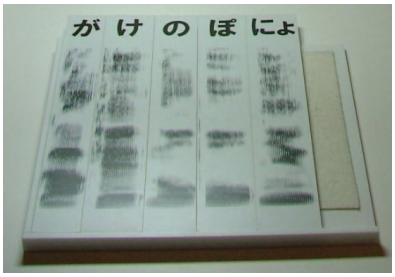

Fig. 5 Five morae (syllables) are lined up next each other under the camera.

Fig. 6 A spectrographic representation of your own speech can be displayed on the monitor screen at a booth called "Voice Analyzer" next to the booth for the "Voice-Print Puzzle."

With this setup, children can line up different plates next each other (Fig. 5) to make arbitrary word(s), phrase(s), or short sentences under the camera. An example phrase is shown at the lower left corner of Fig. 4. The image captured by the camera is also displayed on the monitor screen in the middle of this figure. When the children press a button (the red button in Fig. 4), the system automatically computes the output sound and plays it immediately. This exhibition was very popular with children, because they naturally love puzzles, and they could play with and listen to the sounds at the same time.

In the Sony ExploraScience exhibition, there were three exhibition booths before the Voice-Print Puzzle: Mechanism of Voice, Voice Visualizer and Voice Analyzer. Mechanism of Voice is a hands-on exhibition where visitors experimented with vowel production by placing different physical models of the human vocal tract on top of a loudspeaker. The loudspeaker produced different sound sources simulating human glottal sounds. There were ten different vocal-tract models including a long set of five Japanese vowels for male speech and another short set for female speech. We provided three sound sources, a low-f0 voice source, a noise-like source for male voiced and unvoiced speech, and a high-f0 voice source for female speech. Visitors could test any one of 30 (10 x 3) combinations to produce different vowel sounds.

In the Voice Visualizer booth, visitors could speak into a microphone and see an artistic representation of their speech displayed on a big screen. The display was based on frequency analysis, and the size and direction of the image on the screen changed depending on the frequency and power of the input signal in real time.

In the Voice Analyzer booth, visitors could speak into a microphone and see a spectrographic representation of their speech (Fig. 6). A sample was provided, with a simple explanation of how we can read information from the spectrogram, especially consonants and vowels.

Having these booths next to each other had a synergetic effect on visitors, giving them a greater understanding of speech overall. Although it would have been difficult for the children to fully understand the concepts on their own, we frequently observed that parents first understood and then explained it to their children. Thus, the exhibitions were successful in educating not only children but adults as well.

4. CONCLUSIONS

DPP is suitable for speech science education at the college and graduate level, because it enables us to test which acoustic cue projected on the spectrogram is important for speech perception. In this study, we developed a new DPP algorithm based on additive synthesis of sinusoidal harmonics. This algorithm is much simpler and is more suitable than the original ones for an undergraduate-level lab experiment, because it is based on Fourier synthesis and students are not required to know the concepts of amplitude modulation or fast Fourier transforms in advance. This simplicity is, we believe, important for educational purposes.

We also explored how to use DPP in a scientific exhibition. It turned out that visitors, both children and adults, enjoyed the booths and gave positive feedback. In the future we plan to make available a DPP interface that runs in real time. This would be especially suited for children, who primarily learn from such hands-on experimentation.

5. ACKNOWLEDGMENTS

I would like to thank the members of our lab and the staff at Sony ExploraScience for their assistance. This work was partially supported by a Grant-in-Aid for Scientific Research (21500841) from the Japan Society for the Promotion of Science and a grant of the Sophia University Open Research Center from MEXT.

6. REFERENCES

[1] T. Arai, K. Yasu and T. Goto, "Digital pattern playback," *Proc. Autumn Meet. Acoust. Soc. Jpn.*, pp. 429-430, 2005.

[2] T. Arai, K. Yasu and T. Goto, "Digital pattern playback: Converting spectrograms to sound for educational purposes," *Acoustical Science and Technology*, Vol. 27, No. 6, pp. 393-395, 2006.

[3] F. S. Cooper, A. M. Liberman and J. M. Borst, "The interconversion of audible and visible patterns as a basis for research in the perception of speech," *PNAS*, Vol. 37, pp. 318-325, 1951.

[4] F. S. Cooper, P. C. Delattre, A. M. Liberman, J. M. Borst and L. J. Gerstman, "Some experiments on the perception of synthetic speech sounds," *J. Acoust. Soc. Am.*, Vol. 24, No. 6, pp. 597-606, 1952.

[5] J. M. Borst, "The use of spectrograms for speech analysis and synthesis," *J. Audio Eng. Soc.*, Vol. 4, pp. 14-23, 1956.

[6] R. D. Kent and C. Read, *Acoustic Analysis of Speech*, 2nd ed., Singular, San Diego, CA, 2001.

[7] T. Arai, "Education system in acoustics of speech production using physical models of the human vocal tract," *Acoustical Science and Technology*, Vol. 28, No. 3, pp. 190-201, 2007.

[8] T. Arai, "Speech exhibition at a science museum," *Proc. Spring Meet. Acoust. Soc. Jpn.*, pp. 1387-1390, 2009 (in Japanese).