Acoust. Sci. & Tech. **27**, 6 (2006)

ACOUSTICAL LETTER

Digital pattern playback: Converting spectrograms to sound for educational purposes

Takayuki Arai*, Keiichi Yasu and Takahito Goto

Department of Electrical and Electronics Engineering, Sophia University, 7–1 Kioi-cho, Chiyoda-ku, Tokyo, 102–8554 Japan

(*Received 27 June 2006, Accepted for publication 28 July 2006*)

Keywords: Education in acoustics, Speech science, Acoustic phonetics
PACS number: 43.70.Bk, 43.10.Sv [doi:10.1250/ast.27.393]

1. Introduction

Pattern playback, a device that converts a spectrographic representation back to a speech signal, was developed by Cooper and his colleagues from Haskins Laboratories in the late 1940s [1] and has contributed tremendously to the rapid development of research in speech science [2–4]. By converting a spectrogram into a sound, we can test which acoustic cue projected on the spectrogram is important for speech perception. Furthermore, we can simplify the acoustic cue and/or systematically change an aspect of the acoustic cue, redraw a spectrographic representation, and synthesize stimulus sounds. By doing this, many studies have been conducted, such as the study of the locus theory, which accounts for the importance of the second formant trajectory of a following vowel for the perception of a preceding stop consonant [5].

Today, we can easily implement a modern pattern playback with digital technology, and this is valuable for pedagogical applications. Thus, in this study, we implement a digital pattern playback and explore its usefulness for education [6].

2. Principle

In the original "pattern playback" [1], the light source and tone wheel generate an optical set of harmonics at 120 Hz, and the amplitudes of the harmonics are modulated by a given spectrogram. The spectrogram is placed on the top of a belt moving at a constant speed, and an amplitude-modulated signal is output from the loudspeaker.

This analog version of pattern playback can easily be implemented with modern digital technology. In fact, Nye *et al.* reported a digital version of the pattern playback from Haskins Laboratories using a PDP-11 computer system [7]. In this study, we propose two simple but versatile algorithms for digital pattern playback.

The first algorithm, or the AM method, is based on the concept of amplitude modulation (AM). In this algorithm, the amplitudes of harmonics are modulated by the darkness pattern of a spectrogram as shown in Fig. 1. This is somewhat similar to the original pattern playback based on the source filter theory of speech production. Changing the fundamental frequency of the harmonics yields a variation in pitch, and it eventually allows us to put intonation onto the output sounds. As an alternative option, we can also use a noise source, instead of the harmonic source, to produce unvoiced sounds.

Many studies discuss how to reconstruct the original phase components from a spectrographic representation (e.g., [8]). However, the original pattern playback, even without the reconstruction of phase components, is still extremely powerful for educational purposes because it shows the importance of formant transitions, et cetera. Furthermore, we want to implement a simple, digital system that everybody can use. For this reason, our system does not reconstruct the phase components or change the fundamental frequency during playing back.

The second algorithm, or the FFT method, is based on the fast Fourier transform (FFT). In this algorithm, 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 as shown in Fig. 2. Because we are not reconstructing the original phase, we simply set the phase components to zero.

Because our aim is a simple algorithm with no pitch change during playback, we have carefully chosen a frame shift dependent on the fundamental period. In other words, we used the frame shift that exactly matches the desired fundamental period. To do this, we first reduce the frequency resolution of the spectrum to obtain only the spectral envelope (especially for a spectrogram obtained by a narrow-band analysis), which reflects the vocal-tract filter. Then, by taking the inverse FFT, we get an impulse response of the filter for that time frame. Finally, we place the impulse response along the time axis frame-by-frame with the time interval of the frame shift, which is also equivalent to the fundamental period. We are technically able to change the time intervals to place the impulse responses depending on the instantaneous pitch contour, although we maintain a constant fundamental period.

In theory, we can use a variety of sets of values for each parameter. In practice, we use the following values. For the sampling frequency, 8 to 16 kHz is preferable. For the frame length, 256 or 512 points is optimal. We can use a frame shift of 3–13 ms. This range is suitable for producing a speech sound uttered by an adult male or female, because the fundamental period is set to the frame shift. We often use the frame shift of 10 ms, as when the fundamental frequency is 100 Hz. We can reconstruct an intelligible speech sound as long as the spectrum within a frame is represented at about 40 points or more up to 8 kHz. A non-linear transformation of input contrast values is also optional.

*e-mail: arai@sophia.ac.jp

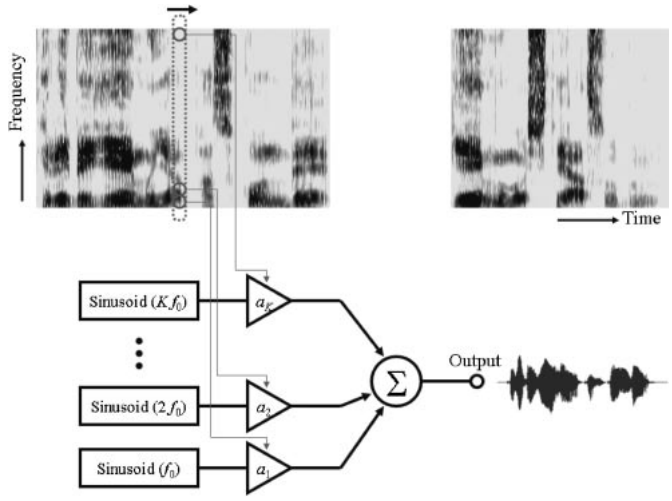

Fig. 1 Block diagram for the Digital Pattern Playback (AM-based algorithm).

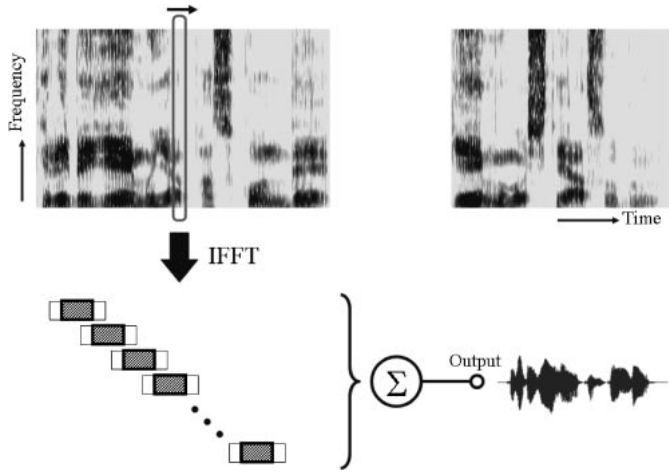

Fig. 2 Block diagram for the Digital Pattern Playback (FFT-based algorithm).

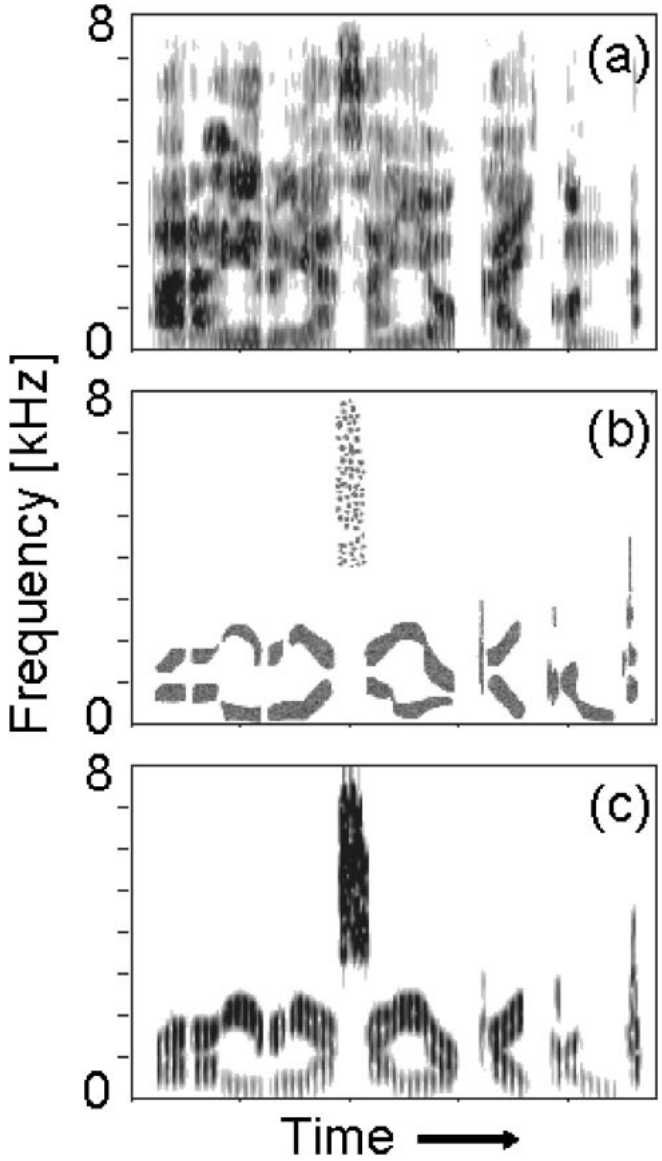

Fig. 3 Spectrograms of an utterance "arayuru yasai-o kaikonda": (a) original signal, (b) simplified version of (a), and (c) reconstructed signal using the FFT method from (b).

3. Auralization of the spectrogram

3.1. Batch version

A spectrogram saved as an image file was converted to speech signals by the proposed methods, that is, the AM and the FFT methods. We obtained intelligible speech sounds from both methods. In addition, we were able to directly convert a spectrogram printed on a sheet of paper to a speech signal immediately after capturing the image from a web camera, not via an image file. Figure 3(a) shows the spectrogram of an original speech signal. Figure 3(b) is a simplified version of the original spectrogram (a), and Fig. 3(c) is the spectrogram of a reconstructed signal from the simplified version using the FFT method. In this case, the sampling frequency was 16 kHz, the frame length was 16 ms, and the frame shift was 10 ms (therefore, the fundamental frequency was 100 Hz).

3.2. Real-time version

For real-time processing, we can implement a simple system with the FFT method by doing the following within each time frame: 1) capturing a time slice of the input spectrogram, 2) computing the waveform of the corresponding glottal cycle at that time frame by the inverse FFT, and 3) producing an acoustic signal based on the glottal waveform. This can be done with a digital signal processor (DSP), so that a user can slide his/her own spectrogram with a slow or fast speed and directly perceive the simultaneously changing output sound. This real-time aspect is very important in a pedagogical situation; the combination of the simultaneous sensations of tactility, somatosensory and auditory perception helps learners to understand the phenomenon more naturally, easily and intuitively.

3.3. Evaluation

Thirteen students participated in this evaluation experiment. They were studying Advanced Acoustics as a part of their two-year speech pathology course after graduating from

four-year colleges. First, we gave students an introductory lecture of speech science including spectrograms with a set of slides, which was followed by an initial written examination. After that, we gave an extra explanation and a demonstration on spectrograms using the proposed digital pattern playback. Afterwards, a second written examination was administered, with exactly the same problem set. Finally, we asked the students to answer to a questionnaire.

The problem set consisted of 16 multiple-choice questions; 4 questions (Problem Set A) are directly related to the demonstration, 8 questions (Problem Set B) are indirectly related to the demonstration, and 4 questions (Problem Set C) are not related to the demonstration but deal with the basics of speech science (all covered in the first lecture). In each of the problems, a part of the statement was missing and there were multiple fill-in-the-blank options, from which the students were asked to choose the best answer.

The overall correct rates of the first and second examinations were 78.4% and 82.2%, respectively. For Problem Set A, they were 75.0% and 92.3%; for Problem Set B, they were 75.0% and 74.0%; and for Problem Set C, they were 88.5% and 88.5%. Although all of the students received the same information twice for Problem Sets A and B in the lecture and the demonstration, we observed an improvement only for Problem Set A but not for Problem Set B. This implies that the demonstration with the proposed digital pattern playback is particularly effective for students to grasp the concepts directly related to the demonstration.

According to the questionnaire, 100% of the students were interested in digital pattern playback. In the questionnaire, we also asked the students to give a percentage for how useful they felt the lectures were, up until the computer demonstration. The average percentage was 26.9%. Because we showed them a demonstration using paper spectrograms converted back to sounds after the on-PC demonstration, we see that the students think it is more useful (73.1%) to see printed spectrograms being converted to speech sounds.

4. Conclusions

In this study, a modern pattern playback was implemented with digital technology. One might think it is impossible to produce a voiceless consonant, such as /s/, from a harmonic source with a constant fundamental frequency. However, it is interesting to note that scattered dots in high frequencies (see Fig. 3(b)) yield a close approximation to /s/ [4]. Because we confirmed that Digital Pattern Playback is effective in an educational demonstration, we will make this tool widely available. Audio and visual demonstrations of the education system are partly available at

http://www.splab.ee.sophia.ac.jp/Digital_Pattern_Playback/.

Acknowledgements

This research was partly supported by Grants-in-Aid for Scientific Research (A-2, 16203041 and C-2, 17500603) from the Japan Society for the Promotion of Science.

References

[1] http://www.haskins.yale.edu/featured/patplay.html
[2] 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*, **37**, 318–325 (1951).
[3] 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.*, **24**, 597–606 (1952).
[4] J. M. Borst, "The use of spectrograms for speech analysis and synthesis," *J. Audio Eng. Soc.*, **4**, 14–23 (1956).
[5] R. D. Kent and C. Read, *Acoustic Analysis of Speech*, 2nd ed. (Singular, San Diego, Calif., 2001).
[6] T. Arai, K. Yasu and T. Goto, "Digital pattern playback," *Proc. Autumn Meet. Acoust. Soc. Jpn.*, pp. 429–430 (2005).
[7] P. W. Nye, L. J. Reiss, F. S. Cooper, R. M. McGuire, P. Mermelstein and T. Montlick, "A digital pattern playback for the analysis and manipulation of speech signals," *Haskins Lab. Status Rep. Speech Res.*, SR-44, pp. 95–107 (1975).
[8] M. Slaney, "Pattern playback from 1950 to 1995," *Proc. IEEE Int. Conf. Systems, Man and Cybernetics Conf.*, Vol. 4, pp. 3519–3524 (1995).