Spectrographic representation of speech based on the short-time Fourier transform

Takayuki Arai (Sophia University)

1. Introduction

Spectrographic representations of speech, that is, *sound spectrograms* or simply *spectrograms*, are widely used in speech-related fields [1]. A speech signal consists of multiple frequency components dynamically changing in time. Therefore, a speech signal is often represented in the time-frequency plane. In the 1940's, the spectrograph was invented to print spectrograms [1]. Today, we can easily compute spectrograms using the short-time Fourier transform (STFT) [2]. In discrete form, the STFT is defined as follows:

$$X[n,k] = \sum_{n=0}^{M-1} x[n+m]w[m]e^{-j(2\pi/N)km}, \qquad 0 \le k \le N-1$$

where $w[m]$ is the window of length M with samples beginning at $m = 0$.

With the STFT, we analyze a speech signal frame-by-frame with a frame length of a couple of tens of milliseconds. Within each frame, a window function is multiplied with the original signal, and the windowed signal is transformed into the frequency domain by the Fourier transform.

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. Pattern playback has contributed tremendously to the rapid development of research in speech science [3-5]. By converting a spectrogram into 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. Today, we can easily implement a modern pattern playback with digital technology [6]. We implemented a digital pattern playback and confirmed its usefulness for pedagogical applications [7,8].

2. Digital Pattern Playback

In the original "pattern playback", 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.

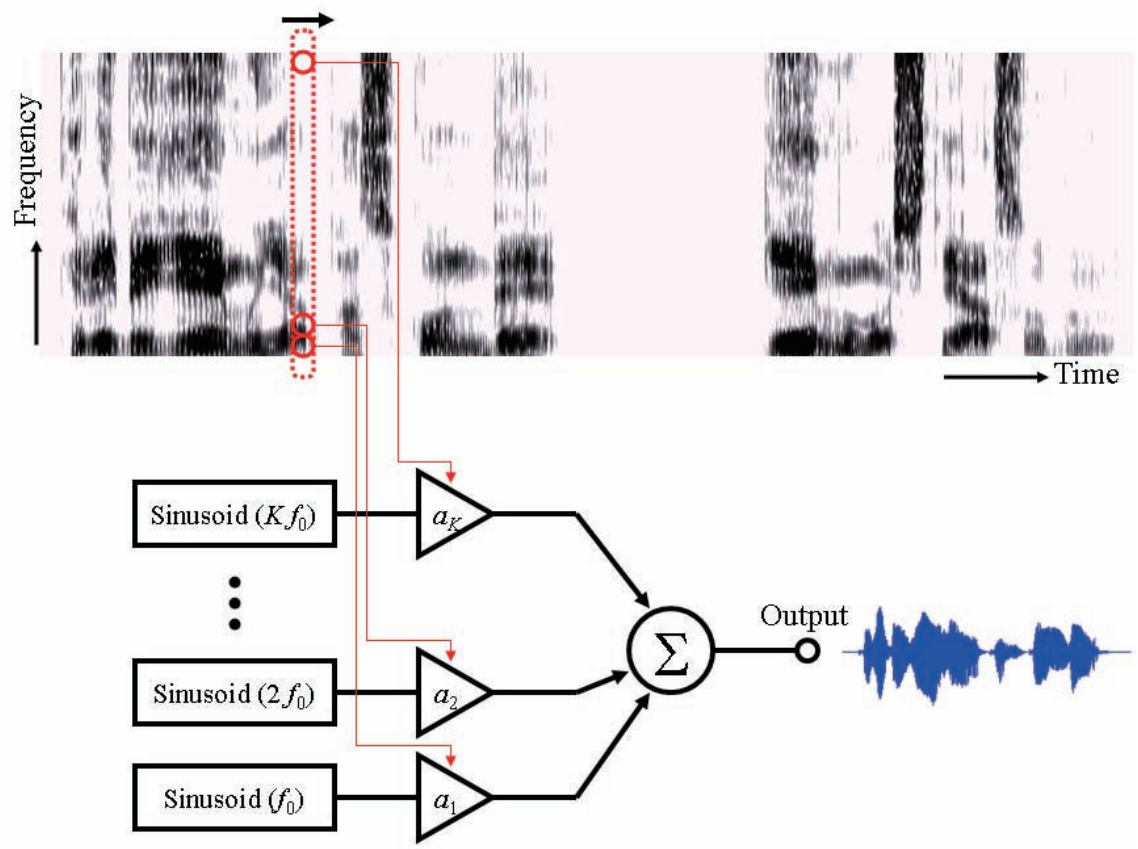

Fig. 1 Block diagram for the Digital Pattern Playback (AM-based algorithm) [7,8].

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 [6]. 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.

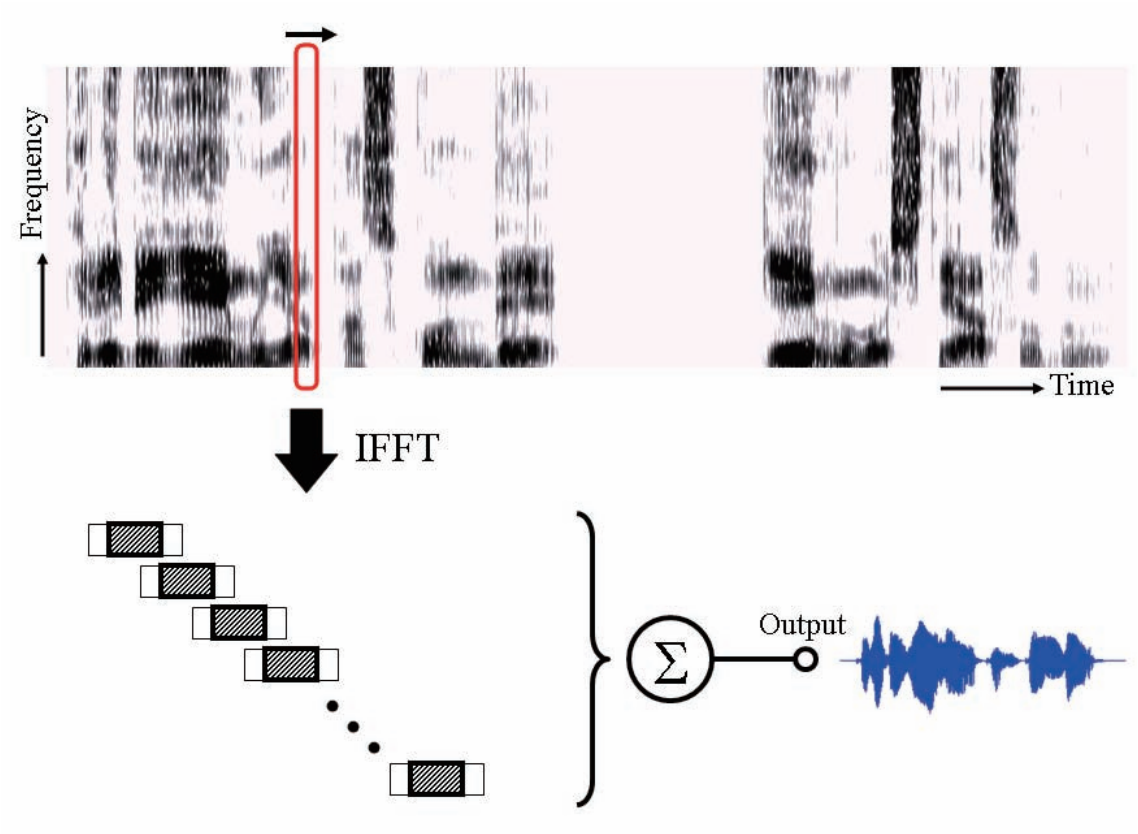

Fig. 2 Block diagram for the Digital Pattern Playback (FFT-based algorithm) [7,8].

Many studies discuss how to reconstruct the original phase components from a spectrographic representation (e.g., [9]). 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.

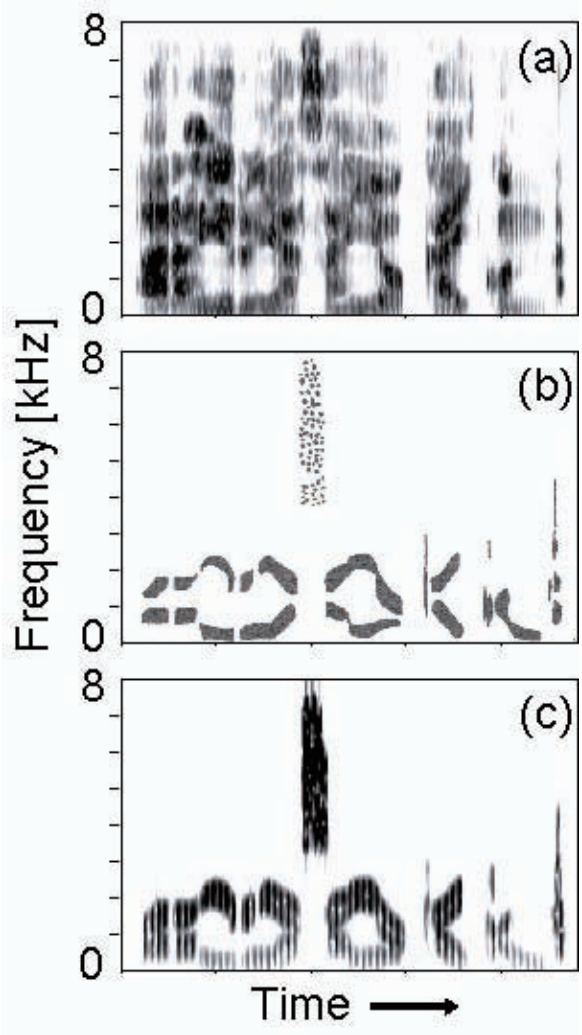

Fig. 3 Spectrograms of an utterance “arayuru yasai-o kaikonda”: (a) original signal, (b) simplified version of (a), and (c) reconstructed signal using the FFT method from (b) [8].

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. Conclusions

We reviewed the STFT and the Pattern Playback, and furthermore, a modern pattern playback was implemented with digital technology. 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] R. D. Kent and C. Read, *Acoustic Analysis of Speech*, 2nd ed., (Singular, San Diego, CA, 2001).

[2] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, (Prentice Hall, Englewood Cliffs, NJ, 1989).

[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*, 37, 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.*, 24(6), 597-606, 1952.

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

[6] 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 Report on Speech Research*, SR-44, 95-107, 1975.

[7] T. Arai, K. Yasu and T. Goto, “Digital pattern playback,” *Proc. Autumn Meet. Acoust. Soc. Jpn.*, pp. 429-430 (2005).

[8] T. Arai, K. Yasu and T. Goto, “Digital pattern playback: Converting spectrograms to sound for educational purposes,” Acoust. Sci. Tech., 27(6), 393-395 (2006).

[9] M. Slaney, “Pattern playback from 1950 to 1995,” *Proc. IEEE Int’l Conf. Systems, Man and Cybernetics Conf.*, 4, 3519-3524, 1995.