SPECTRAL SEQUENCE COMPENSATION BASED ON CONTINUITY OF SPECTRAL SEQUENCE

Masato Akagi, Mamoru Iwaki and Noriyoshi Sakaguchi

Japan Advanced Institute of Science and Technology
1-1 Asahidai, Tatsunokuchi, Nomi, Ishikawa, 923-1292 Japan

akagi@jaist.ac.jp

ABSTRACT

Humans have an excellent ability to select a particular sound source from a noisy environment, called the "Cocktail-Party Effect" and to compensate for physically missing sound, called the "Illusion of Continuity." This paper proposes a spectral peak tracker as a model of the illusion of continuity (or phonemic restoration) and a spectral sequence prediction method using a spectral peak tracker. Although some models have already been proposed, they treat only spectral peak frequencies and often generate wrong predicted spectra. We introduce a peak representation of log-spectrum with four parameters: amplitude, frequency, bandwidth, and asymmetry, using the spectral shape analysis method described by the wavelet transformation. And we devise a time-varying second-order system for formulating the trajectories of the parameters. We demonstrate that the model can estimate and track the parameters for connected vowels whose transition section has been partially replaced by white noise.

1. INTRODUCTION

Although recent speech recognition systems give high recognition rates for clean speech, their speech recognition accuracy is reduced in adverse environments. Humans, on the other hand, can communicate with each other even in the presence of many speakers or a lot of surrounding noise. This is because they can have an excellent ability to select a particular sound source from a noisy environment, called the "Cocktail-Party Effect" and to compensate for physically missing sound, called the "Illusion of Continuity"[1][2]. If we could model these abilities, the models would be able to extract clean speech from noisy speech or predict inaudible speech. This clean speech could then be accurately recognized by recent systems, so such models could improve the performance of automatic speech recognition.

This paper proposes a spectral peak tracker as a model of the illusion of continuity (or phonemic restoration) and a spectral sequence prediction method using the spectral peak tracker.

Although some models have already been proposed, such as spectral peak frequency trajectory extrapolation using second-order systems[3] and spectrum sequence estimation using the IFIS [4], those models treat only spectral peak frequencies or often generate wrong predicted spectra.

To overcome these drawbacks, we introduce a peak representation of the log-spectrum with four parameters: amplitude, frequency, bandwidth, and asymmetry, using the spectral shape analysis method described by the wavelet transformation[5] as a model of the primary auditory cortex. And a time-varying second-order system is devised to formulate the trajectories of the parameters.

To evaluate the model, synthesized connected vowels with the transition section partially replaced by white noise, which causes the illusion of continuity, were provided and processed by the model. The results show that the model can estimate and track the four parameters even in noisy sections and can compensate for the connected vowel spectrum sequences. Additionally, since this model uses a second-order system, it can overshoot reduced-spectrum sequences caused by coarticulation, by determining appropriate second-order system features.

2. SPECTRAL SEQUENCE PREDICTION MODEL

The spectral sequence prediction model proposed in this paper is illustrated in Fig. 1. This model consists of spectral analysis, spectrum representation like that in the primary auditory cortex(A1), spectrum peak extraction, spectrum peak prediction, and spectrum reconstruction.

2.1. Spectral Analysis

Input sound waves are transformed into log-cepstrum sequences. To compensate for the bias in the estimated cepstrum, an unbiased cepstrum estimation[6] is used.

Figure 2 shows an example of the estimated log-power spectrum sequence of the Japanese vowel /a/ with sampling frequency of 20 kHz and cepstrum order of 60. The estimated spectrum sequence has little turbulence and stable spectral peaks. The frequency axis shows ERB-rate[7], which is said to have a good correspondence in the physiology of auditory peripherals and psychology.

$$\text{ERB-rate} = 21.4 \log_{10}(4.37 f[\text{kHz}] + 1) \tag{1}$$

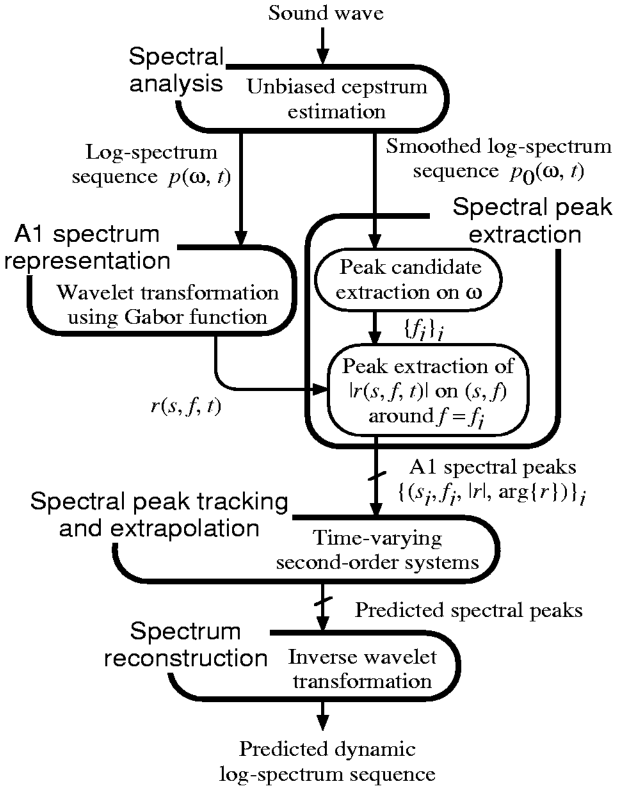
Sound wave
Spectral analysis
Unbiased cepstrum estimation
Log-spectrum sequence p(ω, t)
Smoothed log-spectrum sequence p0(ω, t)
A1 spectrum representation
Wavelet transformation using Gabor function
Spectral peak extraction
Peak candidate extraction on ω
{fi}i
Peak extraction of |r(s, f, t)| on (s, f) around f = fi
r(s, f, t)
A1 spectral peaks {(si, fi, |r|, arg{r})}i
Spectral peak tracking and extrapolation
Time-varying second-order systems
Predicted spectral peaks
Spectrum reconstruction
Inverse wavelet transformation
Predicted dynamic log-spectrum sequence

Figure 1: Spectral sequence prediction model.

2.2. Representation Model of Spectrum in Primary Auditory Cortex

As a function of the primary auditory cortex(A1), Wang and Shamma[5] focused on the spectral pattern analysis along three independent dimensions: a logarithmic frequency axis, a local symmetry axis, and a local spectral bandwidth axis, and they described them using a wavelet transformation. In this paper, following their approach, we adopt the Gabor function as a wavelet, which is given by

$$\psi(\omega) = e^{-\left(3\frac{10^{-1/8}-1}{1+10^{-1/8}}\phi_c\omega\right)^2} e^{i\phi_c\omega}. \qquad (2)$$

Here, ϕ_c is the center angular frequency, and the logarithmic frequency is represented by ERB-rate. Let $r(s, f, t)$ be the A1 spectrum representation at time t; it is calculated through wavelet transformation of the input log-spectrum sequence $p(\omega, t)$ as follows:

$$r(s, f, t) = \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} \psi^* \left(\frac{\omega - f}{s}\right) p(\omega, t)\, d\omega, \qquad (3)$$

where f, s, and $\arg\{r(s, f, t)\}$ represent logarithmic angular frequency, local spectral bandwidth, and local symmetry, respectively, and ψ^* means the complex conjugate of ψ. The absolute value $|r(s, f, t)|$ means the amplitude of the A1 spectrum representation on the above three axes.

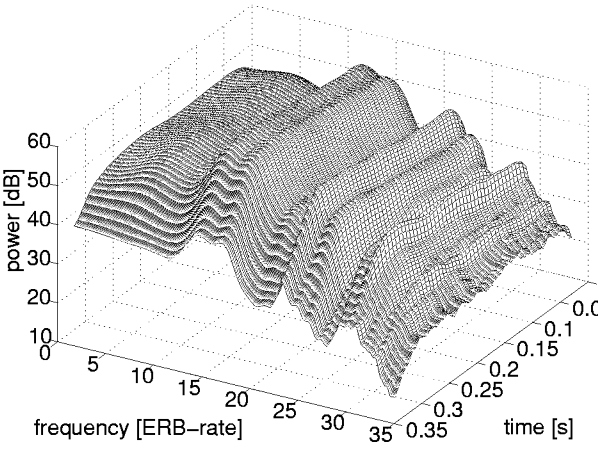

Figure 2: Log-spectrum sequence of vowel /a/ by unbiased estimation.

2.3. Extraction of Spectral Peaks

The amplitude of the A1 spectrum representation has maximal values $|r(s_0, f_0, t)|$ at any time t. This means that there exists a strong resonance of frequency f_0 with bandwidth s_0. Here, we call them A1 spectral peaks and extract them.

Since many spectral peaks generally appear at once, it is not easy to extract them accurately without any unexpected loss. So firstly for the smoothed spectrum $p_0(\omega, t)$ from cepstrum order of 50, maximal points larger than its autoregression line are searched for as initial candidates of maximal points in A1, *i.e.* $\{f_i\}_{i=1,2,\ldots}$. Finally, maximal values of $|r(s, f, t)|$ are searched for in A1 around f_i to extract spectral peak.

The extracted spectral peaks at every time t are represented by four parameters: peak frequency $f_0(t)$, bandwidth $s_0(t)$, symmetry $\arg\{r(s_0, f_0, t)\}$, and amplitude $|r(s_0, f_0, t)|$. They are considered to be spectral peaks considering the A1 characteristics.

2.4. Tracking and Extrapolation of Spectral Peaks

If spectral peak predictors are introduced for time sequences of the extracted spectral peaks, they should be able to simulate illusions of temporal continuity. The predictors are required to track spectral peaks in non-noise sections and to extrapolate spectral peaks while keeping the velocity in noise sections.

In this paper, second-order systems are introduced for predicting and tracking four parameters of spectral peaks.

$$a_i(1-w)y''_{i,j}(t) + \{b_i(1-w) + c_i w\}y'_{i,j}(t) + y_{i,j}(t) = (1-w)x_{i,j}(t) \quad (i = 1, 2, 3, 4) \qquad (4)$$

The system outputs a predicted value $y_{i,j}(t)$ of $x_{i,j}(t)$, which is the j-th spectral peak of the i parameter. Here, $i = 1$ for frequency, $i = 2$ for phase, $i = 3$ for amplitude, and $i = 4$ for bandwidth. The value of w is set to $w = 0$ in noise-free sections and $w = 1$ in noise sections, so the system varies its characteristics depending on whether or not there is noise. The presence of noise is automatically detected by checking whether the frame-wise effective power becomes much bigger than that in the previous frame. Conversely, if the effective power is sufficiently smaller than that in the previous frame, the noise is recognized as having disappeared.

2.4.1. Tracking of Peaks

In a noise-free section, the prediction and tracking system behaves as a second-order system. Assuming that, for a constant value input, the system should output the same constant value, we make a second-order discrete time system for the continuous time system as follows:

$$y_{i,j}[n] = G_i x_{i,j}[n-1] - \alpha_i y_{i,j}[n-1] - \beta_i y_{i,j}[n-2], \quad (5)$$

where $G_i = \frac{2}{2a_i+b_i}$, $\alpha_i = \frac{2(1-2a_i)}{2a_i+b_i}$, $\beta_i = \frac{2a_i-b_i}{2a_i+b_i}$.

2.4.2. Extrapolation of Peaks

It is known that the last tone of an upward/downward sweep tone with successive white noise is perceived higher/lower than its true frequency. This suggests that the spectral peaks in a noise section should be extrapolated holding the aspect that existed just before the noise was entered. Thus we used the following system:

$$y_{i,j}[n] = y_{i,j}[n-1] + (y_{i,j}[n-1] - y_{i,j}[n-2]), \qquad (i = 1, 2) \qquad (6)$$

for the frequency f and phase $\arg\{r(s, f, t)\}$ to be extrapolated while keeping the velocity.

Since spectral peaks tend to blur as the duration of the noise section becomes longer, it seems that the amplitude should decrease and the bandwidth should become wider. Thus we used the following system:

$$y_{3,j}[n] = \frac{c_3}{c_3+1} y_{3,j}[n-1], \qquad (7)$$

for amplitude extrapolation. Additionally, interpreting the above relationship between bandwidth and amplitude as meaning that the product of them should be constant, we extrapolate the bandwidth as

$$y_{4,j}[n] = \frac{c_4+1}{c_4} y_{4,j}[n-1], \; c_4 = c_3. \qquad (8)$$

2.4.3. Tracking and Extrapolation of Peaks

For the spectral peak prediction system, the choice of its input is an important issue because many spectral peaks exist at once. In this paper, the input parameter $x_{i,j}[n-1]$ is taken from the nearest peak to the previous prediction in the frequency axis.

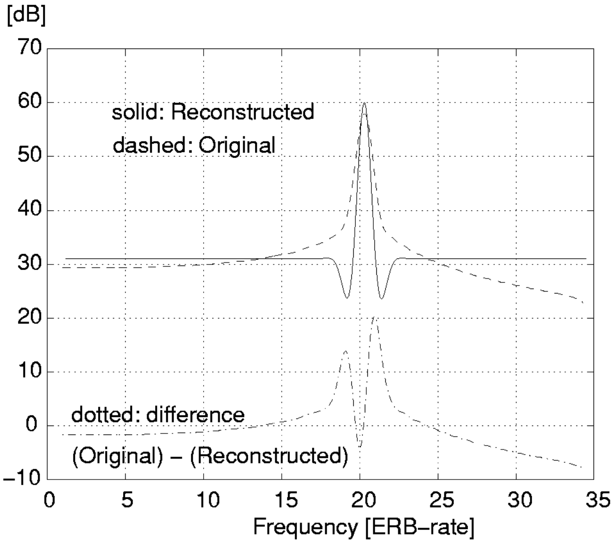

Figure 3: Reconstructed spectrum (solid line: reconstruction, dashed line: original).

It is noticeable that this choice provides a chance of interchange between two closer peaks.

2.5. Reconstruction of Spectrum

The log-power spectrum is reconstructed using only spectral peaks in A1 spectrum representation. It is calculated by the sum of the responses of predicted spectral peaks $\hat{r}(s_0, f_0, t)\phi(\frac{\omega - f_0}{s_0})$ and the mean of $p(\omega, t)$ over ω. It corresponds to the inverse wavelet transform. Figure 3 shows an example of the reconstruction of a single sweep tone at a moment.

3. SIMULATION OF TRACKING AND EXTRAPOLATION OF SPEECH SPECTRUM

A connected vowel (/a/ connected to /i/) was synthesized with a formant synthesizer. Its formants F1 to F3 were 800, 1200, and 2500 Hz in /a/ and 250, 2500, and 3000 Hz in /i/ with bandwidths 80, 120, and 150 Hz, respectively. The sampling frequency and F0 were 20 kHz and a constant 140 Hz, respectively. The transition section of /a/ and /i/ was 100 ms with linear transition in ERB-rate. The latter half of the transition section was replaced by enough large white noise, as shown in Fig. 4. The input log-power spectrum sequence was obtained by unbiased cepstrum estimation with order of 60, frame length of 25.6 ms and frame shift 6.4 ms, which is shown in Fig. 5. The prediction and tracking system was tuned using the parameters a_i, b_i, and c_i in eq.(4) so as to have the characteristic frequency of 20 Hz,

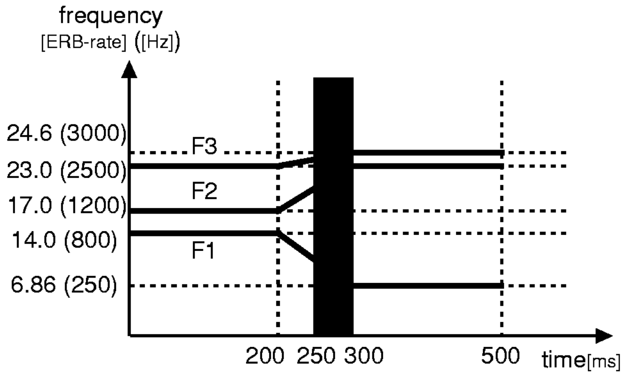

Figure 4: Synthesized connected vowel /a/-/i/ with transient replaced by noise.

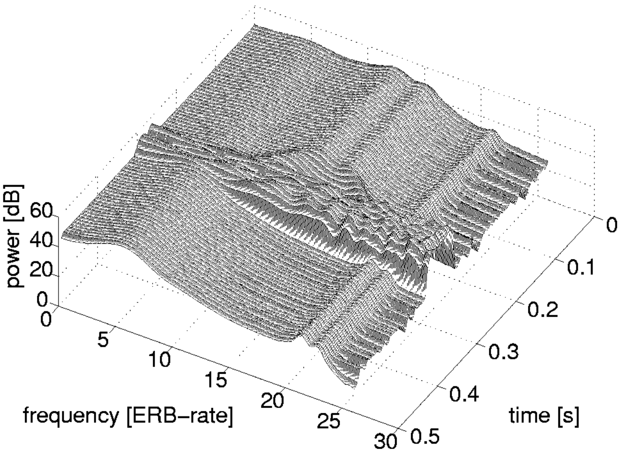

Figure 5: Input spectrum sequence of vowel in Fig. 4.

damping factor of 1, and $c_i = 0.0064$ in the noise-free section. Figure 6 demonstrates the prediction and tracking result. It is observed that the section replaced by noise was recovered well by the prediction and tracking.

Some auditory phenomenon[8] are simulated by the model, for example, the bouncing effect in crossing two sweep tones as an effect of spectral peak tracking (CD track No. 17) and the illusion of continuity as an effect of spectral peak extrapolation (CD track No. 29). In addition, reduced spectrum sequences caused by coarticulation may be recovered by overshooting the spectral peak.

4. CONCLUSIONS

Spectral peaks were represented by four parameters of the A1 spectral representation following Wang and Shamma[5]. A spectral peak prediction and tracking system having some perceptual knowledge was introduced. This system was able to predict and track the spectral peaks from sounds that included bursts of noise. Although it is hard

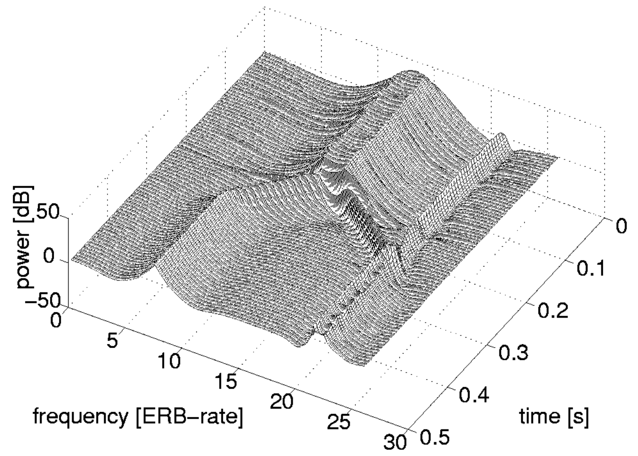

Figure 6: Reconstructed spectrum for input in Fig. 5.

to simply apply the results to speech recognition systems, the modeling of phonemic restoration will contribute to systems in the future.

5. ACKNOWLEDGMENTS

This work was supported by CREST and Grant-in-Aid for Scientific Research from the Ministry of Education (Nos. 07308026 and 10680374).

6. REFERENCES

1. Warren, R. M. "Perceptual restoration of missing speech sounds", *Science*, Vol. 73, 1970, pp. 1011–1012.
2. Bregman, A. S., *Auditory scene analysis: The perceptual organization of sound*, The MIT Press, London, 1990.
3. Aikawa, K., Kawahara, H., and Tsuzaki, M. "A Neural Matrix Model for Active Tracking of Frequency-Modulated Tones", *ICSLP'96*, Vol. 1, pp. 578–581.
4. Masuda-Katsuse, I., Kawahara, H., and Aikawa, K. "Speech Segregation Based on Continuity of Spectral Shapes", *Computational Auditory Scene Analysis CASA'97*, pp. 39–45.
5. Wang, K. and Shamma, S. A. "Spectral Shape Analysis in the Central Auditory System", *IEEE Trans. Speech Audio Processing*, Vol. 3, No. 5, 1995, pp. 382–395.
6. Imai, S. and Furuichi, C., "Unbiased Estimation of Log Spectrum", *IEICE Trans. A*, Vol. J70-A, No. 3, pp. 471–480.
7. Glasberg, B. R., and Moore, B. C. J. "Derivation of auditory filter shapes from notched-noise data", *Hear. Res.*, Vol. 47, 1990, pp. 103–138.
8. Bregman, A. S. and Ahad, P. A., *Demonstrations of Auditory Scene Analysis: The perceptual organization of sound*, MIT Press, 1996.