SPEECH REPRESENTATION AND TRANSFORMATION USING ADAPTIVE INTERPOLATION OF WEIGHTED SPECTRUM: VOCODER REVISITED

Hideki KAWAHARA

ATR Human Information Processing Research Laboratories, Kyoto 619-02, Japan
kawahara@hip.atr.co.jp

ABSTRACT

A simple new procedure called STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) has been developed. STRAIGHT uses pitch-adaptive spectral analysis combined with a surface reconstruction method in the time-frequency region, and an excitation source design based on phase manipulation. It preserves the bilinear surface in the time-frequency region and allows for over 600% manipulation of such speech parameters as pitch, vocal tract length, and speaking rate, without further degradation due to the parameter manipulation.

1. INTRODUCTION

The need for flexible speech modification methods is increasing in both commercial and scientific fields. Various sophisticated methods have been proposed [1], but their flexibility and resultant speech quality has been limited. If we consider old concepts in the context of the enormous progress in computational power in recent years, then a simple and appealing idea like the VOCODER [2], which separates spectral and source information in order to manipulate and transmit speech sounds, is potentially very powerful. The major problem in such separation has been interactions between source characteristics and temporal fine structures like periodicity. A simple change in how one views this problem may provide the answer.

This work was motivated by the need for a flexible and high-quality analysis-synthesis method to use in our ongoing experiments on auditory feedback[3]. To attain this goal, we addressed in turn three sub-problems: 1) elimination of periodicity effects in spectral analysis 2) reliable F0 estimation, 3) fine control of phase information in resynthesis. These were solved using a combination of Fourier and wavelet analysis techniques. The major difference with prior methods is that ours involves information expansion, rather than reduction. This allows great flexibility of manipulation.

2. METHOD

2.1. Elimination of periodicity effects

The central idea of the proposed method is that it considers the periodic excitation of voiced speech to be a sampling operation of a surface $S(\omega, t)$ in a three-dimensional space defined by the axes of time, frequency, and amplitude, which represent the global source characteristics and the shapes and movements of articulation organs. In this interpretation, a periodic signal $s(t) = s(t + n\tau_0)$ with a fundamental period τ_0, is thought to provide information about the surface for every τ_0 in the time domain and every $f_0 = 1/\tau_0$ in the frequency domain.

A more dependable representation of this repetitive aspect of speech waveforms is as follows.

$$s(t) = \sum_{k \in N} \alpha_k(t) \sin\left(\int_{t_0}^{t} k(\omega(\tau) + \omega_k(\tau))d\tau + \phi_k\right) \quad (1)$$

This equation implies that a speech waveform is a nearly harmonic sum of FM (frequency modulated) sinusoids modulated by AM (amplitude modulated) parameters. We assume that $\alpha_k(t)$ represents a sampled point of the surface $S(\omega, t)$. This equation also suggests that a fundamental frequency derived from a different frequency range may have a slightly different value. This equation is very close to the sinusoidal representation's[4], but the procedure to make use of this formulation is quite different.

Short-term Fourier analysis of this signal yields a time-frequency representation of the signal $F(\omega, t)$, known as a spectrogram [5]. The spectrogram exhibits an almost regular structure from the signal periodicity in both the time and frequency domains. It is desirable to use a time windowing function $w(t)$ which has isometric resolution in both the time and frequency domains and has the minimum uncertainty.

$$w(t) = \frac{1}{\tau_0} e^{-\pi(t/\tau_0)^2} \quad (2)$$

Since the fundamental period ($\tau_0(t) = 2\pi/\omega(t)$) varies with time, the analysis window size also adaptively follows the change.

Our goal is to reconstruct a smoothed time-frequency representation $S(\omega, t)$, which has no trace of interference caused by the periodicity of the signal based on the partial information given by the adaptive window analysis. The simplest form which requires four nearest grid points to determine the surface is the bilinear equation given below.

$$S_p(\omega, t) = (a_p\omega - b_p)(c_p t - d_p) \quad (3)$$

where subscript p represents the p-th patch element and $\{a_p, b_p, c_p, d_p\}$ are constants to define the shape. This piecewise bilinear approximation of the surface provides a good approximation of the original surface, if the original surface is reasonably smooth.

It is possible to calculate this piecewise bilinear representation based on the data of sampling points by a first-order spline function using sampling points as knot points. However, algorithms based only on knot points are numerically fragile for real speech signals, because real speech signals are not precisely periodic and consist of natural fluctuations and noises. Instead, we propose using an interpolation function which provides equivalent piecewise bilinear representations when the sampled data in the time-frequency representation are given only on the grid points.

Let $h_t(\lambda,\tau)$ be an interpolation function. Then, the operation to calculate the smoothed representation is given by the following equation.

$$S(\omega,t)=\sqrt{g^{-1}\left(\iint_D h_t(\lambda,\tau)g(|F(\omega-\lambda,t-\tau)|^2)d\lambda d\tau\right)} \quad (4)$$

where D represents the support of the interpolation function $h_t(\lambda,\tau)$. The interpolation function $h_t(\lambda,\tau)$ which fulfills the requirement to preserve the bilinear surface, is the product of two crossing triangular ridges as defined below.

$$h_t(\lambda,\tau)=\frac{1}{4}(1-|\lambda/\omega_0(t)|)(1-|\tau/\tau_0(t)|) \quad (5)$$

where $\omega_0(t)=2\pi f_0(t)$ and $[-\omega_0(t)\le\lambda\le\omega_0(t),\ -\tau_0(t)\le\tau\le\tau_0(t)]$. Note that this operation is local both in the time domain and in the frequency domain, and this makes the procedure less sensitive to f_0 errors.

In Equation 4, $g(\ )$ defines what quantity to be preserved through the interpolation. For example, the identity mapping, $g(x)=x$, preserves the energy of the signal and the 1/3 power law, $g(x)=x^{1/3}$, preserves the perceived loudness, approximately. This nonlinearity is also used to control any over-smoothing caused by multiple smoothing by $h_t(\lambda,\tau)$ and $w(t)$. It is used in conjunction with the inverse filtering of the time frequency representation of $w(t)$ in terms of $h_t(\lambda,\tau)$.

2.2. Reliable F0 extraction

In the proposed STRAIGHT procedure, it is crucially important to extract reliable fundamental frequencies at a fine time and frequency resolution, especially in the re-synthesis stage. This is made possible to extract the instantaneous frequency of the fundamental component of the signal. This may sound strange to some readers, because in order to extract the fundamental frequency, it is necessary to know the fundamental frequency in advance.

This apparent contradiction is solved by introducing a measure to represent the 'fundamental-ness' without using *a-priori* knowledge about the fundamental frequency. A fairly wide class of analyzing wavelets makes the fundamental component have the smallest FM and AM. Then, one can introduce a factor-of-merit to represent this 'fundamental-ness' using the magnitudes of FM and AM.

Using an analyzing wavelet $g_{AG_1}(t)$ made from a complex Gabor filter having a slightly finer resolution in frequency (*i.e.* $\eta>1$), the input signal can be divided into a set of filtered complex signals $D(t,\tau_0)$.

$$D(t,\tau_0) = |\tau_0|^{-\frac{1}{2}}\int_{-\infty}^{\infty} s(t)\overline{g_{AG_1}\left(\frac{t-u}{\tau_0}\right)}du \quad (6)$$

$$g_{AG\tau}(t) = g_\tau(t-\tau/4)-g_\tau(t+\tau/4) \quad (7)$$

$$g_\tau(t) = e^{-\pi\left(\frac{t}{\eta\tau}\right)^2}e^{-j\frac{2\pi t}{\tau}} \quad (8)$$

The 'fundamental-ness' index $M(t,\tau_0)$ is calculated for each channel (τ_0) based on this output. The definition of the index is given as follows.

$$M = -\log\left[\int_\Omega\left(\frac{d|D|}{du}\right)^2du\right]+\log\left[\int_\Omega|D|^2du\right] -\log\left[\int_\Omega\left(\frac{d\arg(D)}{du}\right)^2du\right]+2\log\tau_0 \quad (9)$$

where integration interval Ω is set proportional to the size of the corresponding analyzing wavelet. This index M is normalized and scalable without any adjustments. Extracting F0 means simply to find the maximum index of M in terms of τ_0 and to calculate the instantaneous frequency using the outputs of neighboring channels around τ_0.

2.2.1. *Minimum phase impulse response and fine pitch control*

The extracted f_0 (in fine resolution) is used to re-synthesize speech signal $y(t)$ using the following equation.

$$y(t) = \sum_{t_i\in Q}\frac{1}{\sqrt{G(f_0(t_i))}}\ v_{ti}(t-T(t_i))$$

$$v_{ti}(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}V(\omega,t_i)\Phi(\omega)e^{j\omega(t)}d\omega \quad (10)$$

$$\text{where } T(t_i)=\sum_{t_k\in Q,k<i}\frac{1}{G(f_0(t_k))}$$

where Q represents a set of positions of the excitation for synthesis, $G(\ )$ represents the pitch modification.The all-pass filter function $\Phi(\omega)$ is used to control the fine pitch and the temporal structure of the source signal and is described in the next section.

$V(\omega,t_i)$ represents the Fourier transform of the minimum phase impulse response [6] which is calculated from the modified amplitude spectrum $A(S(u(\omega),r(t)),u(\omega),r(t))$, where $A(\ )$, $u(\ )$ and $r(\ )$ represent manipulations in the amplitude, frequency, and time axes, respectively.

$$V(\omega,t) = \exp\left(\frac{1}{\sqrt{2\pi}}\int_0^\infty h_t(q)e^{j\omega q}dq\right) \quad (11)$$

$$h_t(q)=\begin{cases}0 & (q<0)\\ c_t(0) & (q=0)\\ 2c_t(q) & (q>0)\end{cases}$$

$$\text{and}\qquad c_t(q)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-j\omega q}\log A d\omega$$

where q represents the quefrency.

2.3. Excitation source design

The phase manipulation mentioned in the previous section allows the control of the excitation source waveform by phase manipulation. This is necessary, because while there is no degradation in speech quality caused by parameter manipulation using the STRAIGHT procedure, there is some initial degradation in speech quality under headphone listening when no temporal fine structure control is employed. Literature on phase effects in timbre perception [7] has suggested that temporal fine-structure control, based on an all-pass filter design, provides the answer.

2.3.1. *Design goals*

An all-pass component which meets the following requirements is suitable for speech synthesis and applications for musical instruments.
(1) Temporal spread of energy in time and in frequency should be controlled to meet a specific purpose.
(2) Temporal asymmetry in energy distribution should be implemented if necessary.
(3) A systematic method using random numbers to generate all-pass filters is also desirable.

2.3.2. How to generate the signal

The temporal fine structure is designed using the following equation.

$$\Phi_2(\omega) = \exp\left(-j\rho(\omega)\sum_{k\in P}\alpha_k \sin(k\omega)\right) \qquad (12)$$

where P represents a set of indices and $\rho(\omega)$ represents the frequency-weighting function used to control the temporal energy spread in each frequency region. The spread of multipulses in the simplest case, where only one sinusoidal phase component exists and $\rho(\omega) = 1$, can be controlled using the following relation. Let ε be a small number which is effectively equivalent to zero for a specific purpose.

$$\begin{aligned} |\alpha| &\leq (\varepsilon\Gamma(\beta+1))^{\frac{1}{\beta}}\, 2^{\frac{\beta-1}{\beta}} \qquad (13)\\ \beta &= \frac{\Delta t}{k} \end{aligned}$$

where $\Gamma(\ )$ represents the Γ function and Δ represents the desired spread in time, which is represented in the number of samples. The spread for multiple components is derived as the product of each component spread.

The all-pass filter design using random numbers is based on the group delay design, because it allows a more intuitive control of the temporal structure than by using the phase characteristics directly. Let $n(t)$ be Gaussian white noise and $W_s(\tau)$ be a weighting function in the spatial frequency domain. The desired spread d_g of the target group delay function $d_4(\omega)$ is calculated using the following set of equations.

$$\begin{aligned} d_4(\omega) &= \frac{d_g x(\omega)}{\sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}|x(\omega)|^2\, d\omega}} \qquad (14)\\ x(\omega) &= \rho(\omega)F^{-1}\left(W_s(\tau)N(\tau)\right)\\ W_s(\tau) &= |\tau|\exp\left(-\pi(\tau/\tau_{bw})^2\right) \qquad (15) \end{aligned}$$

where $N(\tau)$ is the initial random group delay function (even function) made from $n(\omega)$. $F^{-1}()$ denotes the inverse Fourier transform. The phase characteristic $\Phi(\omega)$ is calculated by integrating $d_4(\omega)$.

In order to introduce asymmetry into the group delay, we can use the following equation.

$$\Phi_3(\omega) = \exp\left(-j\int_{-\pi}^{\omega} r\left(j\frac{d\log\Phi_2(\lambda)}{d\lambda}\right)d\lambda\right) + c_0 \qquad (16)$$

where $r()$ represents an arbitrary and smooth even function. The following function is one example.

$$r_1(x) = \exp(-|x|) + |x| - 1 \qquad (17)$$

This asymmetry is found to introduce an interesting timbre.

3. EXAMPLES

A set of experiments involving analysis, modification and re-synthesis using real speech data, was conducted under the following conditions:
(1) Use of a sampling frequency of 22,050Hz with 16-bit linear A/D converted speech.
(2) Analysis of isolated words spoken by one male and two female subjects.
(3) Setting of the FFT length at 1024 to avoid moiré patterns with harmonic structures and frequency sampling.
(4) Analysis every 1 ms to produce 513 x 1000 data points per second.

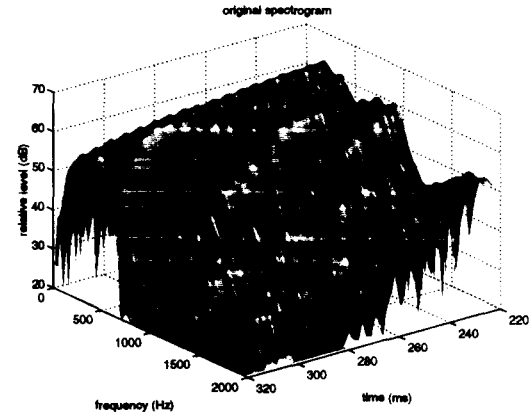

Figure 1. Spectrogram of female speech 'light' using a pitch-adaptive window.

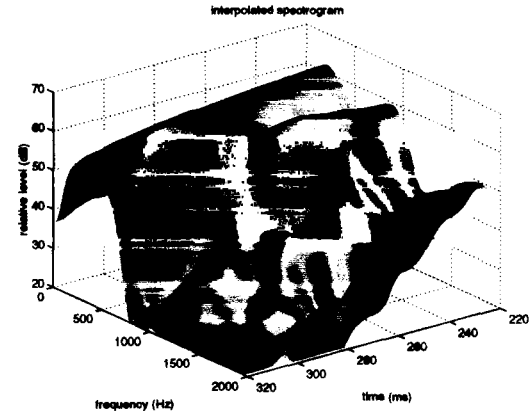

Figure 2. Smoothed spectrogram of female speech 'light' using a pitch-adaptive window.

(5)Extraction of fundamental frequencies every 1ms in a search range of from 40 Hz to 800 Hz without any post-processing.
(6) Introduction of 70 Hz spatial frequency smoothing by $W_s(\tau)$ and 2400Hz high pass weighting by $\rho(\omega)$ when generating random all-pass filters. The target group delay d_g is set to 2ms. Note that no asymmetry was introduced in the current experiments.

3.1. Resynthesized speech quality

The original sound files and the manipulated files are located at the following URL:
http://www.hip.atr.co.jp/~kawahara/aiff/
Informal listening tests demonstrated that re-synthesized speech signals are sometimes indistinguishable even under careful headphone listening conditions. Also, further degradations caused by parameter manipulation have been found to be negligible, even with modifications of 600%. Figure 1 shows the original equal-resolution spectrogram and Figure 2 shows the smoothed spectrogram. These are also accessible by visiting our STRAIGHT web page given above.

3.2. Evaluation of F0 extraction performance

3.2.1. Pulse train and white noise

A systematic test using white noise and a pulse train was conducted. The signal to noise ratio was manipulated from

S/N ratio	% success	standard deviation
∞	100%	0.004 Hz
40 dB	100%	0.13 Hz
30 dB	100%	0.28 Hz
20 dB	100%	0.86 Hz
10 dB	95.7%	2.77 Hz
0 dB	43.0%	6.34 Hz
0 dB (envelope)	86.5%	5.22 Hz

Table 1. Relation between S/N and rms error in F0 extraction for a pulse train with white noise.

proposed method errors			
	ordinary	subharmonic	total
NC:	2.86%	0.06%	2.92%
FHS:	0.96%	0.27%	1.23%
improved AMDF errors			
	ordinary	subharmonic	total
NC:	1.90%	0.70%	2.60%
FHS:	0.87%	1.48%	2.35%
get_f0 (ESPS) errors			
	ordinary	subharmonic	total
NC:	0.31%	2.65%	2.96%
FHS:	3.28%	0.93%	4.21%

Table 2. Comparative performance of the proposed method, an improved AMDF method and a commercial method.

infinity, 40dB to 0 dB in 10dB steps. Only a 100Hz pulse train was tested, because the proposed procedure is completely scalable and independent of the sampling frequency and F0. Table 1 illustrates the results. The F0 search range was from 40 Hz to 800 Hz, and no post-processing was made. The last line shows the result when an envelope signal is used as the input.

3.2.2. Speech and EGG database

A speech database with simultaneous EGG recordings, made available by ATR Interpreting Telephony Research Laboratories, was used to evaluate practical behaviors of the proposed F0 extraction method. The data used in the test consisted of 208 sentences spoken by a male speaker and a female speaker. The total duration of voicing was 159 seconds for the male data and 266 seconds for the female data. An improved AMDF method[8] and the get_f0 procedure in the ESPS[1] system were used. No post-prosessing was mede in both procedures. F0s were extracted every 1ms. A comparative performance evaluated based on the EGG recordings is given in Table 2. F0 differences greater than 20% were counted as errors. By introducing a heuristic weighting on M, the error rate may be further reduced. It indicates that the method is competitive or supersedes existing F0 extraction methods.

4. POSSIBLE APPLICATIONS

Preliminary examination of several utterances showed that the interpolated spectrogram analyzed by the new procedure was surprisingly smooth. This indicates that there is a lot of room for information reduction; it is a good starting point for investigating information reduction because the resynthesized sound from this apparently smooth spectrogram preserved a considerable amount of fine details of the original speech quality.

[1] ESPS is a trdemark of Entropic Research Laboratory, Inc.

The magnitude spectrogram, which has no trace of the source periodicity, is a highly flexible representation for manipulation because any modification still directly corresponds to a feasible waveform through a complex cepstrum representation. This flexible representation and the nonparametric nature of the proposed method also open up various applications like voice morphing, electric musical instruments synthesis and efficient reuse of sound resources.

Analysis-and-synthesis methods such as ours were believed to give a poorer speech quality than waveform based methods. The reproduced sound by the proposed method seems to provide a counter example. It suggests that the original concept of the VOCODER still holds, and that the speech quality based on the analysis-and-synthesis scheme can be improved further.

Finally, it should be pointed out that the F0 extraction procedure developed as a part of the STRAIGHT procedure demonstrated an extremely accurate and robust performance. It can be used as a general purpose procedure to extract "fundamental-like" components in arbitrary signals. We would like to suggest calling the procedure TEMPO (Time-domain Excitation extractor using Minimum Perturbation Operator.)

5. CONCLUSION

A new method (STRAIGHT) to represent and manipulate speech signals, based on pitch adaptive spectrogram smoothing, is presented. STRAIGHT offers high flexibility in parameter manipulation with no further degradation, while maintaining a high reproduction quality. This may help promote research on the relation between physical parameters and perceptual correlates. The fundamental frequency extraction procedure (TEMPO) also provides a versatile method for investigating quasi-periodic structures in arbitrary signals.

ACKNOWLEDGMENT

The author would like to express his sincere appreciation to his colleagues at ATR, and to Dr. Roy Patterson of MRC Cambridge, and Dr. Toshio Irino of NTT for their discussions, and Dr. Alain de Cheveigné of CNRS for his discussions and evaluations of TEMPO and other methods. He also wishes to express special thanks to his collaborator, J. C. Williams, for discussions and encouragement.

REFERENCES

[1] R. Veldhuis and H. He. Time-scale and pitch modifications of speech signals and resynthesis from the discrete short-time fourier transform. *Speech Communication*, 18:257–279, 1996.

[2] H. Dudley. Remaking speech. *J. Acoust. Soc. Am.*, 11(2):169–177, 1939.

[3] H Kawahara and J. C. Williams. Effects of auditory feedback on voice pitch. In *The 9th Vocal Fold Physiology Symposium*, Sydney, 1995. [in print].

[4] Robert J. McAulay and Thomas F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. ASSP*, 34:744–754, 1986.

[5] L. Cohen. Time-frequency distributions - a review. *Proc. IEEE*, 77(7):941–981, 1989.

[6] A. Oppenheim and R. Schafer. *Discrete-Time Signal Processing.* Prentice Hall, Englewood Cliffs, NJ, 1989.

[7] R. D. Patterson. A pulse ribbon model of monaural phase perception. *J. Acoust. Soc. Am.*, 82(5):1560–1586, 1987.

[8] Alain de Cheveigné. Speech fundamental frequency estimation. Technical Report TR-H-195, ATR-HIP, 1996.