Restructuring speech representations using a pitch adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds

Hideki Kawahara[*], Ikuyo Masuda-Katsuse [†] and Alain de Cheveigné[‡]
ATR Human Information Processing Research Laboratories,
2-2 Hikari, Seika-cho, Soraku-gun, Kyoto 619-02, Japan
kawahara@sys.wakayama-u.ac.jp

September 22, 1998

Abstract

A set of simple new procedures has been developed to enable the real-time manipulation of speech parameters. The proposed method uses pitch-adaptive spectral analysis combined with a surface reconstruction method in the time-frequency region, and an excitation source design based on group delay manipulation. It also consists of a fundamental frequency (F0) extraction method using instantaneous frequency calculation based on a new concept called 'fundamentalness'. The proposed procedures preserve the details of time-frequency surfaces while almost perfectly removing fine structures due to signal periodicity. This close-to-perfect elimination of interferences and smooth F0 trajectory allow for over 600% manipulation of such speech parameters as pitch, vocal tract length, and speaking rate, while maintaining high reproduction quality.

1 Introduction

The need for flexible speech modification methods is increasing in both the commercial and scientific fields. Various sophisticated methods have been proposed [18, 24, 19, 25], but their flexibility and resultant speech quality have been limited. This may suggest that the appropriate representation of speech has not been found yet or it may simply indicate that we have not yet explored the full potential of existing representations. If we revisit old concepts having views underlying "Auditory Scene Analysis" in mind, we will find the latter is the case.

The channel VOCODER [10], which separates spectral and source information in order to manipulate and transmit speech sounds, is a good example of a simple and appealing idea made powerful by introducing such point of views. The channel VOCODER and its modern variant, LPC (Linear Predictive Coding) [14, 4] are potentially very flexible in parameter manipulations, because there are little inter-related constraints between spectral and source parameters. The VOCODER-based representations are also attractive, because they are conceptually easy to understand and have direct correspondence to both the speech production mechanism and the auditory periphery. However, the re-synthesized speech quality by VOCODERs suffers from buzziness induced by the pulsive excitation even without parameter manipulations. It also degrades when parameter manipulations were large. On the other hand, sophisticated methods based on iterative procedures to approximate desired manipulated STFT (short term Fourier transformation) have superior reproduction quality when the amount of manipulations were small. However, they deteriolate rapidly when parameter manipulations increase. Intricate relations between the manipulated spectral paremeters and waveforms also make it difficult to get insight from these representations. Simple concepts like the channel VOCODER may seem outdated. However, if it allows for a high quality reproduction, it is true the simpler the better.

There are numbers of proposals to reduce buzziness of VOCODER-type methods and found effective. Then, the major remaining problem is errors in spectral estimation. It is necessary to remove any periodicity interferences from the time-frequency representation for the representation to be usable for reproducing a speech sound in a different fundamental frequency (F0) or in a different vocal tract length. Parametric models like LPC are also susceptible to signal periodicity[12], even

[*] Currently, he is also a professor of Wakayama University and a project head under CREST

[†] Currently, Matsushita Electric Industrial co., LTD.

[‡] Paris 7th University / CNRS. He participated in this project while he was at ATR as a visiting researcher.

though they can alleviate constraints posed by the uncertainty principle. These interferences are observed as apparent increase in random variations in spectral representations. It is, in a sense, contradictory and frustrating that periodicity induces such difficulty in speech analysis and manipulation because voiced sounds are perceived to be smoother and richer than unvoiced sounds at least for human listeners. Speech representations have to take advantage of the periodic nature of voiced sounds instead of treating it as a problem. In other words, we need a stable spectral representation which does not have any trace of periodicity.

The flexible speech manipulation also introduces a requirement on F0 trajectories. Conventional F0 extraction methods based on interval measurements usually provide stepwise trajectories, especially for low-pitched voices. This stepwise structure itself is a trace of the source periodicity and is harlmful for F0 modifications. It is therefore desireble to have a F0 extraction method which prvides a smooth trajectory.

The goal of this paper is to introduce how a very high quality speech analysis-modification-synthesis method is implemented as a channel VOCODER based on the answers mentioned above. Information reduction is not intended in this paper. Quality and flexibility of manipulations are the primary interests. The paper consists of four sections. Firstly, a pitch-adaptive spectral smoothing to eliminate priodicity interference is discussed. Secondly, an instantaneous-frequency-based F0 extraction method is introduced to provide reliable and smooth F0 trajectoryies. Thirdly, a system to utilize the proposed representations to manipulate speech parameters is introduced. Finally, examples of real speech analysis and manipulations are presented.

2 Elimination of periodicity interference

In this section a method to eliminate spectral interference structure caused by signal periodicity is systematically introduced. At first, the basic principle of the proposed method is introduced as adaptive smoothing of time-frequency representation. Then, a compensatory time window design is introduced to reduce this time-frequency smoothing problem to a smoothing problem in a frequency domain. Finally, a procedure to compensate and eliminate the major implementational problem of this formulation, over-smoothing, is introduced.

2.1 Background

When the length of a time window for spectral analysis is comparable to the fundamental period of the signal repetition, the resultant power spectrum shows periodic variation in the time domain. When the length of a time window spans several repetitions, the resultant power spectrum shows pariodic variation in the frequency domain. If the signal is purely periodic and the pariod is an integer multiple of the sampling pariod, a pitch-synchronous analysis can perfectly eliminate temporal variations by using a rectangular window which length is an integer multiple of the fundamental period in samples. If the size of the rectangular window is set equal to the fundamental period, variations both in the time domain and the frequency domain can be eliminated.

However, this approach is not practical for analyzing natural speech signals, because the fundamental frequencies of those signals are changing all the time and each repetition is not the same as the pervious period due to natural source related fluctuations. The sharp discontinuitis at the both ends of the rectangular window makes the analysis highly sensitive to such minor fluctuations. Spectral smearing caused by this discontinuity and fluctuations is detectable because of the wide spectral dynamic range of natual speech signals. It is also not practical to extract a portion which represents an impulse response, because responses corresponding formant peaks do not die out within a pitch period unless the pitch is extremely low.

The other approach to eliminate periodicity related interferences is to introduce a spectrum model which embodies periodicity effects. This approach was proposed for LPC parameter estimation[12]. The results using synthetic speech signals indicated that interferences due to the signal periodicity are compensated well. However this approach does not provide reliable estimates for natural speech, because this method assumes that the autocorrelation function of the periodic source is a regular pulse train. This assumption does not hold for natural speech when the spectrum model only represents the auto-regressive components of natural speech. In general, the moving average components of natural speech spectral envelope which are not modeled in auto-regressive part, result in a unpredictable smearing of the autocorrelation function. And the smearing introduces a significant bias in the periodicity compensation process. In other words, this model based approach is fragile for natural speech signals.

These suggest that the elimination process of the periodicity interferences should not rely on neither strong spectrum models nor perfect periodicity. The desired method has to be robust for natural fluctuations and F0 estimation errors.

2.2 Pitch-adaptive smoothing

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, fre-

quency, 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. In other words, voiced sounds are assumed to provide partial information about the surface. The goal of spectral analysis that enables flexible manipulation is to recover the surface $S(\omega, t)$ using this partial information.

However, speech is neither purely periodic nor stable. Also the estimation process of the fundamental frequency inevitably introduces estimation errors. The desired algorithm has to take these factors into account. A more dependable representation of this nonstationary repetitive aspect of speech waveforms is as follows.

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

where $\alpha_k(t)$ represents the time varying amplitude of k-th harmonic component, $\omega_k(\tau)$ represents a time varying fundamental angular frequency of k-th component, and ϕ_k represents the initial phase at time t_0. This equation implies that a speech waveform is a nearly harmonic sum of FM (frequency modulation: represented by $\omega_k(\tau)$) sinusoids modulated by AM (amplitude modulation: represented by $\alpha_k(t)$) 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. The form of this equation is very close to that of the sinusoidal representation [18, 5], but the procedure in using this formulation differs quite a bit.

Short-term Fourier analysis of this signal yields a time-frequency representation of the signal $F(\omega, t)$, known as a spectrogram [8]. The spectrogram exhibits an almost regular structure from the signal periodicity in both the time and frequency domains. This representation is a result of smearing due to the time-frequency representation of the time windowing function. The uncertainty principle introduces a trade-off relation between frequency resolution and temporal resolution of the windowing function. It is therefore desirable to use a time windowing function $w(t)$ which has equivallent relative resolution in both the time and frequency domains to take the full advantage of the available partial information.

Assume that the time window function $w(t)$ to have the following form.

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

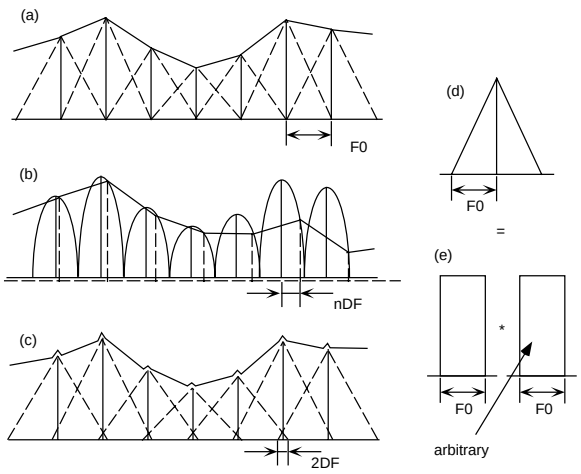

Figure 1: Interpolation and smoothing in a one-dimensional case. (a) illustrates a piecewise linear interpolation and equivalent smoothing operation using the smoothing function (d) constructed by a convolution operation of two first-order cardinal B-spline functions (e). Actually (d) is a second-order cardinal B-spline function. (b) illustrates the effects of F0 errors on an interpolation operation that only relies on knot points. Note that the distortion increases proportionally with the harmonic number. Here, 'DF' denotes the error in F0 estimation and 'n' represents a harmonic number. (c) illustrates the effects of F0 errors on the smoothing operation. Note that the distortion is localized and independent of the harmonic number.

$$W(\omega) \quad = \quad \frac{\tau_0}{\sqrt{2\pi}} e^{-\pi(\omega/\omega_0)^2} \tag{3}$$

where $W(\omega)$ represents the Fourier transform of $w(t)$, and $\omega_0 = 2\pi f_0$. Since the fundamental period ($\tau_0(t) = 2\pi/\omega_0(t)$) 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. This is considered to be a surface reconstruction problem that is based on partial information. It is therefore necessary to provide constraints for the problem so that a unique solution can be obtained. One reasonable constraint is to have only local information used.

Consider the one dimensional case. The simplest reconstruction method using discretized partial information like amplitudes of harmonic components of voiced sounds, is to connect harmonic peaks by straight lines[]. In other word, it is to represent the reconstructed surface as a piecewise first order polynomial.

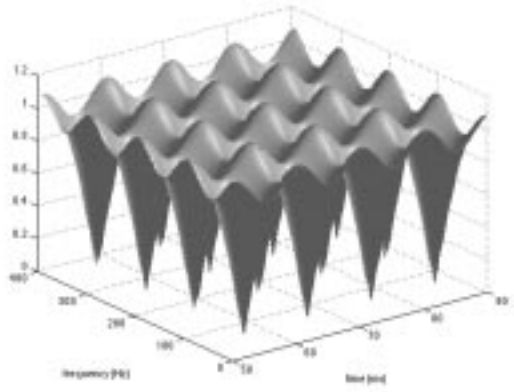

Figure 2: Isometric 3D spectrogram of a regular pulse train (100 Hz).

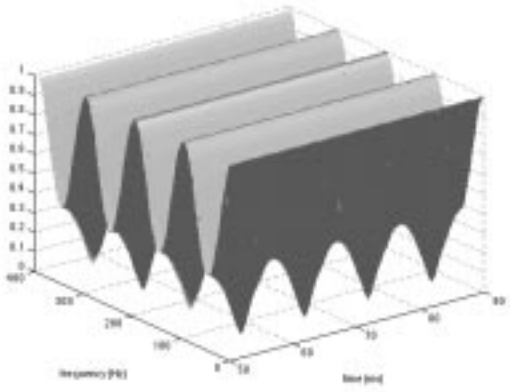

Figure 3: 3D spectrogram of a regular pulse train (100 Hz) using a pitch synchronous modification.

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. They are also sensitive to small errors in fundamental frequency estimation. Instead, we propose using a smoothing function that provides an equivalent piecewise linear representation. It is a convolution (smoothing) using a 2nd-order cardinal B-spline function. The procedure is illustrated in Figure 1. Smoothing operation is preferable to interpolation operation for real speech, where noise and error in F0 estimation are inevitable, because the smoothing operation is less sensitive to those problems and resultant errors are localized, as shown in the Figure.

In our previous papers [16, 15], a two-dimensional smoothing procedure was prpoosed to implement the basic idea. In this report, a set of pitch adaptive time windows are introduced to calculate phase insensitive power spectra which reduces this two-dimensionla smoothing operation into one dimensional smoothing operation mentioned above.

2.3 Power spectrum with reduced phasic interference

In this section, a compensatory set of time windows, which provides an effectively phase insensitive spectrogram is introduced. It means that it is not necessary to apply the 2nd-order cardinal B-spline smoothing function to remove the periodic interference from such spectrogram. Instead, it is sufficient to perform the spline-based smoothing only in the frequency domain, once the temporal interference is effectively eliminated.

First of all, an examplar periodic interference using

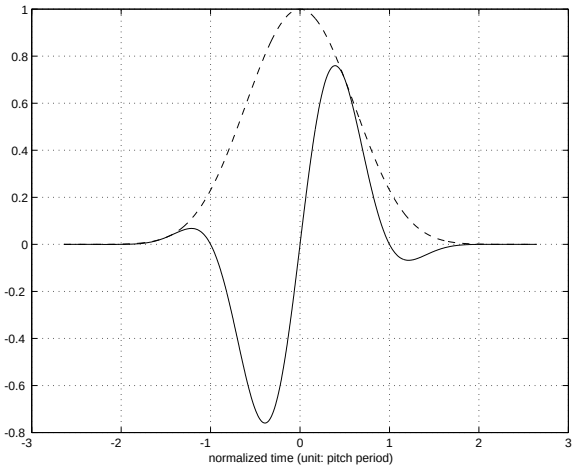

Figure 4: Compensatory time window (solid line) and its original time window (dashed line).

isometric real valued time window is illustrated. Figure 2 shows interference for a test signal consisting of regular pulses with a constant fundamental period (10 ms). Regular 'holes' exist both in the time and frequency domains where neighboring frequency components become out of phase.

It is possible to remove temporal interferences around peaks by employing a pitch synchronous analysis by constructing a new time windowing function $w_p(t)$ based on a cardinal B-spline basis function which size is adaptive to the fundamental period. The second-order cardianl B-spline function is selected because it places the second-order zeros on the other harmonic frequencies. This makes the resultant spectrum less sensitive to F0 estimation errors. Figure3 shows a 3D

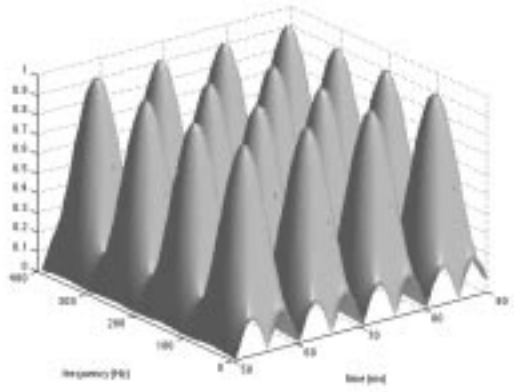

Figure 5: 3D spectrogram of a regular pulse train using a compensatory time window.

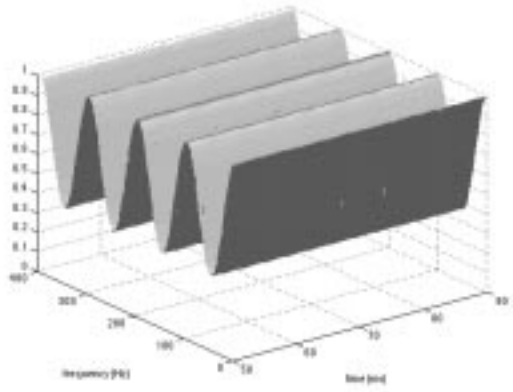

Figure 6: 3D spectrogram with reduced phase effects.

spectrogram of the same pulse train using the time window given in the following equation.

$$w_p(t) = e^{-\pi\left(\frac{t}{\tau_0}\right)^2} \odot h(t/t0) \quad (4)$$

$$h(t) = \begin{cases} 1-|t| & |t|<1 \\ 0 & \text{otherwise} \end{cases}$$

where $\odot$ represents convolution. The spectrogram illustrates that periodic interference is still exists in spectral vallay areas.

It is possible to design a compensatory time window that produces maxima where the original spectrogram has 'holes'. The compensatory time window for the pitch synchoronous time window $w_p(t)$ has the following form. The sinusoidal modulation of Equation 5 is designed to perform frequency conversion and phase shifting at the same time to fulfil the requirement. Consider a set of neighboring harmonic components. The sinusoidal modulation upconverts the lower harmonic component to the amount of F0/2 and downconverts the higher harmonic component to the amount of F0/2. It also shifts their phases towards the opposite directions to the amount of $\pi/2$ each. This phase shift makes out-of-phase in-phase and produces maxima where the original spectrogram has 'holes'. The window function is illustrated in Figure 4.

$$w_c(t) = w_p(t)\sin\left(\pi\frac{t}{t_0}\right) \quad (5)$$

The power spectrum using this compensatory window is shown in Figure 5. A power spectrum with reduced phasic interference $P_r(\omega,t)$ is represented as a weighted squared sum of the power spectra $P_c(\omega,t)$ and $P_o(\omega,t)$ using this compensatory window and the original time window, respectively.

$$P_r(\omega,t) = \sqrt{P_o^2(\omega,t) + \xi P_c^2(\omega,t)} \quad (6)$$

where ξ is selected to minimize temporal variation of the resultant spectrogram. A numerical search procedure provided 0.13655 for the optimum blending factor ξ. Figure 6 illustrates the reduced phasic interference spectrogram obtained using this ξ.

Using this optimum blending factor, it is possible to eliminate needs for temporal smoothing using the 2nd-order cardianl B-spline smoothing function. It also enables slower frame rate for calculating spectrogram. These changes from the previous implementation are very effective to reduce the demand on the computational power.

2.4 Compensation of over-smoothing in the frequency domain

One problem with the algorithm described in the previous section is over-smoothing, which occurs due to smearing caused by the time windowing function. For example, in the frequency domain, a power spectrum which is calculated by short term Fourier transformation, does not have a line spectral structure. Each harmonic component is smoothed by the frequency domain representation of the time windowing function. Then the smoothing function $h(\lambda)$ in the frequency domain has to operate on this already smoothed spectral representation to eliminate interferences caused by the signal periodicity. This operation successfully removes the interference, however at the same time it also smoothes the underlying spectral structure. This over-smoothing is illustrated in Figure 7. The over-smoothing $v(\omega)$ is also represented as a function of the frequency representation of the time window $W(\omega)$ and the 2nd-order cardinal B-spline smoothing function in

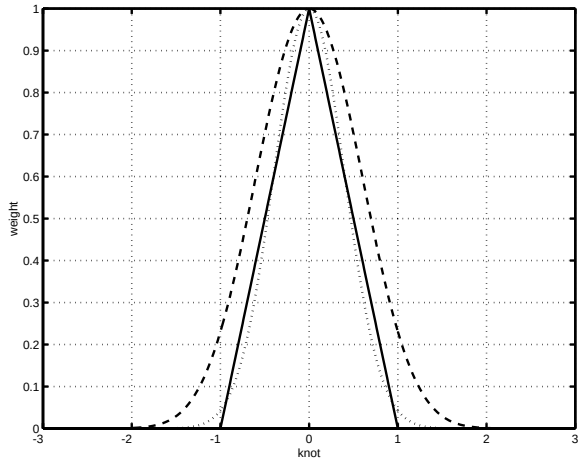

Figure 7: Over smoothing example. The desired basis function of the 2nd-order cardinal B-spline (solid line) is smoothed out (dashed line) by the smoothing effect of the time window (dotted line).

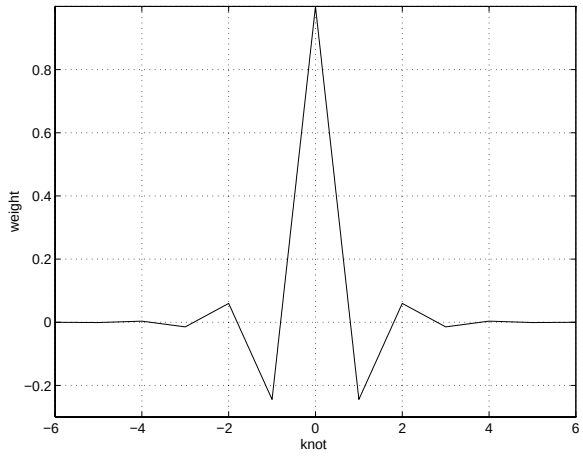

Figure 8: The optimal 2nd-order smoothing function.

the frequency domain $h(\omega)$.

$$v(\omega) = \int_{-\infty}^{\infty} W(\omega - \lambda)h(\lambda)d\lambda \tag{7}$$

Assume that the over-smoothed spectrum is reasonably approximated by the convlution of $v(\omega)$ and the line spectrum sampled from the underlying continuous spectrum at each harmonic frequency. Then it is possible to recover the original spectral values at each harmonic frequency by applying a compensating operation which transforms $v(\omega)$ to have only one non-zero value at each harmonic frequency. In other words, the problem is reduced to an inverse filtering problem. The requirement is written as follows. The goal is to find a set of coefficients $c_k(k = -N...N)$.

$$u_l = \sum_{k=-N}^{k=N} c_k v_{l-k} \tag{8}$$

where

$$u_l = \begin{cases} 1 & (\ l = 0) \\ 0 & (\text{otherwise}) \end{cases} \tag{9}$$

where N is a reasonably large integer which effectively make v_N negliglble and v_k represents $v(k\omega)$. It is necessary to use more than $2N$ + 1 set of this relation to find a unique solution to this problem. It yields the following simultaneous linear equations.

$$\mathbf{u} = \mathrm{H}\mathbf{c} \tag{10}$$

$$\mathbf{u} = [u_{-M}, u_{-M+1}, \ldots, u_0, \ldots, u_{M-1}, u_M]'$$

$$\mathbf{c} = [c_{-N}, c_{-N+1}, \ldots, c_0, \ldots, c_{N-1}, c_N]'$$

$$\mathrm{H} = \begin{array}{c} \\ -M \\ \vdots \\ k \\ \vdots \\ M \end{array} \begin{array}{c} \begin{array}{ccccc} -N & \cdots & l & \cdots & N \end{array} \\ \begin{pmatrix} v_{-N-M} & \cdots & v_{l-M} & \cdots & v_{N-M} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ v_{k-N} & \vdots & v_{k+l} & \vdots & v_{k+N} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ v_{-N+M} & \cdots & v_{l+M} & \cdots & v_{N+M} \end{pmatrix} \end{array}$$

where [ ]′ represents transpose of a matrix. The solution is represented as follows.

$$\mathbf{c} = (\mathrm{H}'\mathrm{H})^{-1}\mathrm{H}'\mathbf{u} \tag{11}$$

Figure 8 shows the optimal smoothing function for a previously introduced isometric Gaussian time window. Figure 9 also shows the shape of the compensated over-smoothing function. Note that only one values at each node point is non-zero.

However, this optimal smoothing function introduced a new problem. We required the algorithm to be localized. The size of support of the optimal smoothing function is effectively 3 times larger than the original smoothing function. It is desirable to calculate a quasi optimal smoothing function with smaller support. The desired quasi optimal smoothing function can be calculated by making N small, 1 for example. Figure 10 shows the shape of the recovered impulse using a quasi optimal smoothing function which consists of three second-order cardinal B-spline functions (N = 1).

3 Extraction of smooth and reliable F0 trajectories

For flexible and high-quality modification of speech parameters, it is also important to extract F0 trajectories which do not have any trace of interferences caused by

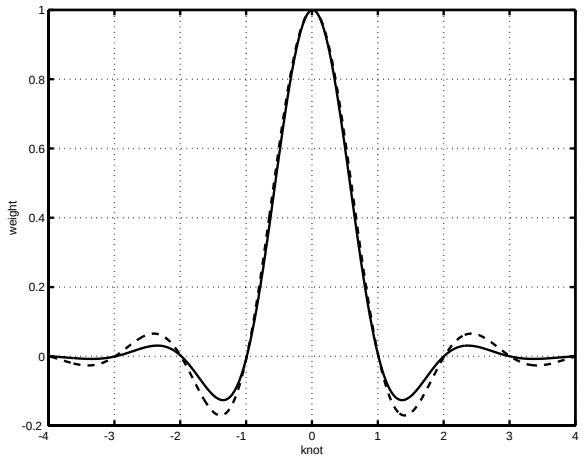

Figure 9: Recovered impulse using the optimal smoothing functions (solid line: 2nd-order, and dashed line: 4th-order).

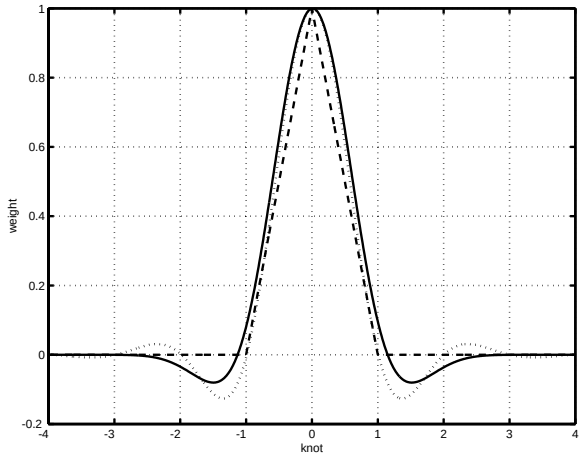

Figure 10: Recovered impulse using the optimal smoothing function and the quasi-optimal smoothing function (solid line: quasi-optimal, dotted line: optimal, and dashed line: basis function).

phasic interaction between analysis time window and the waveform of a signal. Pitch extraction algorithms based on the usual definition of periodicity do not behave well for this purpose, because a natural speech signal is neither purely periodic nor stable. For example, nearly periodic signals represented by Equation 1 do not satisfy the usual definition of periodicity shown below.

$$s(t + T_0) = s(t) \tag{12}$$

Where, T_0 is the period. Pitch extraction algorithms based on the usual definition of periodicity try to find T_0 to minimize some distance measure between $s(t + T_0)$ and $s(t)$ for a certain duration. However, there is no reason for the $1/T_0$ extracted in that manner to agree with the instantaneous frequency[5, 6, 1, 2] of the fundamental component in Equation 1. Therefore, it is better to extract the instantaneous frequency of the fundamental component directly, if we use the signal model represented by Equation 1.

3.1 Basic principle

In the proposed method, the fundamental frequency is extracted as the instantaneous frequency of the fundamental component of the signal. This may sound strange to some readers, because selecting the fundamental component requires knowledge of the fundamental frequency in advance.

This apparent contradiction is solved by introducing a measure to represent the 'fundamentalness' without using *a-priori* knowledge about the fundamental frequency. A fairly wide class of analyzing wavelets makes the output, which corresponds to the fundamental component, have smaller FM and AM than the other outputs. These analyzing wavelets correspond to a frequency response that is designed to have a steeper cut-off at the higher end and a slower cut-off at the lower end. Let us define the 'fundamentalness' to have the maximum value when FM and AM modulation magnitudes are minimum and to have a monotonic relation with the modulation magnitudes.

Figure 11 illustrates how this definition works in the case of analyzing a complex sound with several harmonic components. When no harmonic component is within the response area of the analyzing wavelet (case (a) in the figure) , 'fundamentalness' in this case provides the background noise level. When the fundamental component is inside the response area but not at the best (or characteristic) frequency of the analyzing wavelet (case (b) in the figure), 'fundamentalness' is not very high because of the low signal to noise ratio. When the frequency of the fundamental component agrees with the best (or characteristic) frequency of the analyzing wavelet (case (c) in the figure), the highest signal to noise ratio causes 'fundamentalness' to be maximum. When the frequency of a harmonic component other than the fundamental component agrees with the best (or characteristic) frequency of the analyzing wavelet (case (d) in the figure), even though the signal to noise ratio in terms of the specific harmonic component provides the highest value, 'fundamentalness' is not high, because two or more harmonic components are located within the response area due to the intended filter shape design. The other cases also provide lower 'fundamentalness'. Thus, maximum 'fundamentalness' assures that the filter actually corresponds to the fundamental component in question.

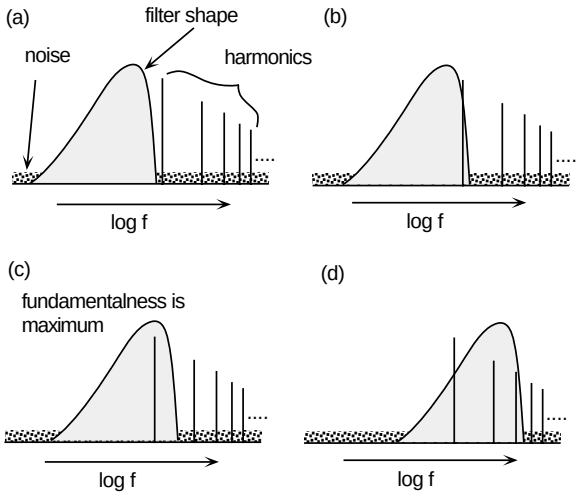

Figure 11: Schematic explanation of how 'fundamentalness' works. (a) an analyzing wavelet (filter shape) covers no harmonic component. (b) only the fundamental component is inside the receptive area. (c) the fundamental component is located at the best frequency of the analyzing wavelet. (d) a higher harmonic component is located at the best frequency of the analyzing wavelet

There are many functions that have the required filter shape. A Gabor function is one example. A factor-of-merit is introduced to represent this 'fundamentalness' using the FM and AM magnitudes as follows.

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

$$D(t,\tau_c) = |\tau_0|^{-\frac{1}{2}} \int_{-\infty}^{\infty} s(t)\overline{g_{AG}\left(\frac{t-u}{\tau_c}\right)}du \quad (13)$$

$$g_{AG}(t) = g(t-1/4) - g(t+1/4) \quad (14)$$

$$g(t) = e^{-\pi\left(\frac{t}{\eta}\right)^2} e^{-j2\pi t} \quad (15)$$

The characteristic period of the analyzing wavelet τ_c is used to represent the corresponding filter channel.

The 'fundamentalness' index $M(t, \tau_c)$ is calculated for each channel (τ_c) based on the output. The definition of the index is given as follows.

$$M = -\log\left[\int_\Omega \left(\frac{d|D|}{du}\right)^2 du\right] + \log\left[\int_\Omega |D|^2 du\right] - \log\left[\int_\Omega \left(\frac{d^2\arg(D)}{du^2}\right)^2 du\right] + \log\Omega(\tau_c) + 2\log\tau_c \quad (16)$$

where the integration interval $\Omega(\tau_c)$ is set proportional to the size of the corresponding analyzing wavelet and is a function of τ_c. The first term represents the magnitude of AM component. The magnitude of AM is normalized by the second term which represens the total energy. The third term represents the magnitude of FM component. The magnitude of FM is normalized by the fifth term which represents squared frequency. The fourth term is the normalization factor of the temporal integration interval. These normalization make the index M dimension-less number, meaning it is scalable to any F0s and sampling frequencies.

In practice, it is better to use a slightly modified definition, because the F0 trajectory normally consists of rapid movements that carry prosodic information. Removing the contribution of the monotonic F0 movement reduces artifacts on the 'fundamentalness' evaluation caused by prosodic components.

$$M_c = -\log\left[\int_\Omega \left(\frac{d|D|}{du} - \mu_{AM}\right)^2 du\right] - \log\left[\int_\Omega \left(\frac{d^2\arg(D)}{du^2} - \mu_{FM}\right)^2 du\right] + \log\left[\int_\Omega |D|^2 du\right] + \log\Omega(\tau_0) + 2\log\tau_0 \quad (17)$$

$$\mu_{AM} = \frac{1}{\Omega}\int_\Omega \left(\frac{d|D|}{du}\right) \quad (18)$$

$$\mu_{FM} = \frac{1}{\Omega}\int_\Omega \left(\frac{d^2\arg(D)}{du^2}\right) \quad (19)$$

Extracting F0 simply means finding the maximum index of M_c in terms of τ_0 and calculating the average (or more specifically, interpolated) instantaneous frequency using the outputs of the channels neighboring τ_0.

The instantaneous frequency $\omega(t)$ of one filer output signal $D(t, \tau_0)$ is calculated using the following equation.

$$\omega(t) = 2f_s \arcsin\frac{|y_d(t)|}{2} \quad (20)$$

$$y_d(t) = \frac{D(t+\Delta t/2,\tau_0)}{|D(t+\Delta t/2,\tau_0)|} - \frac{D(t-\Delta t/2,\tau_0)}{|D(t-\Delta t/2,\tau_0)|}$$

3.2 Evaluation of F0 extraction performance

It is important to mention details about the performance and properties of the proposed F0 extraction method because it provides an interesting insight into how to integrate the source characteristics and the spectral characteristics.

A preliminary test on the baseline performance of the proposed fundamental frequency extraction method was conducted.

S/N	% 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 (root mean square) error in F0 extraction for a pulse train with white noise. The last row represents the result when the envelope of the original signal that was calculated using Hilbert transform is used as the input to the proposed F0 extraction procedure.

3.2.1 Pulse train and white noise

A preliminary test was conducted using white noise and a pulse train. The average signal to noise power ratio was manipulated from infinity, 40dB to 0 dB in 10dB steps. Only a 100Hz pulse train was tested, because the proposed procedure is 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 involved. The ratio between center frequencies of adjucent channels is $2^{1/12}$. The last line shows the result obtained when an envelope signal was used as the input instead of using signal waveform itself.

Note that the rms (root mean square) deviation from the true fundamental frequency of the pulse train is proportional to the relative noise amplitude. When the S/N is 40dB in the F0 frequency region, a 0.13% rms deviation is possible. The method is relatively robust. Even under a 0dB signal to noise ratio, a 40% success rate for F0 extraction and deviation of less than 6% rms are obtained. The tolerance to noise is increased 6dB by using the envelope signal (that is calculated using Hilbert transform of the original signal) as the input signal. In this case, an 86% success rate for F0 extraction and deviation of less than 5% rms are obtained.

Figure 12 shows a scatter plot of relative F0 extraction errors versus the 'fundamentalness' index. Figure 13 also shows the average relation between the 'fundamentalness' and rms (root mean square) error of F0 extraction. The linear relationship is due to the simple definition of the instantaneous frequency. This relation enables one to estimate the F0 extraction error based on the 'fundamentalness' index at the same time as the F0 extraction. This is a very useful attribute of our method and the 'fundamentalness' can be used to implement a coding scheme without a U/V (unvoiced-voiced) decision process.

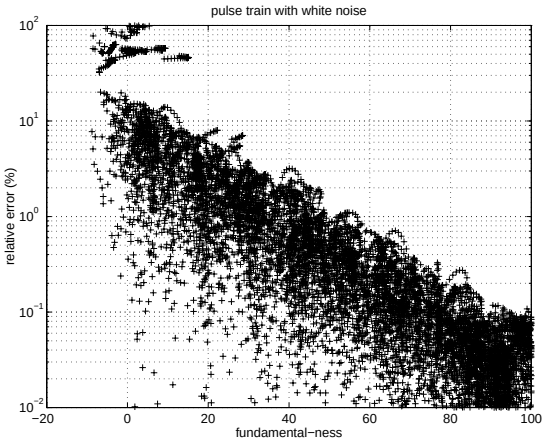

Figure 12: Scatter plot of 'fundamentalness' and F0 errors.

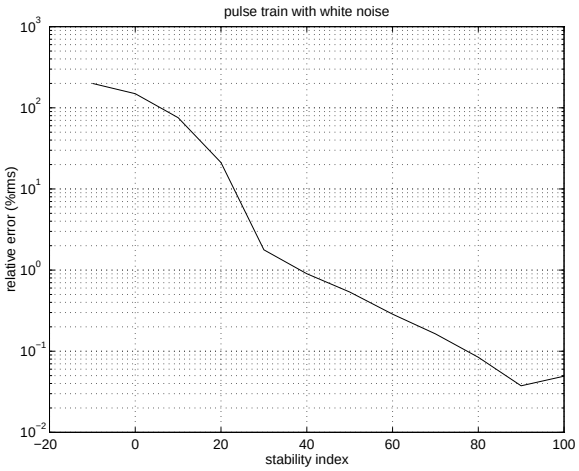

Figure 13: Relation between 'fundamentalness' and rms F0 errors.

3.2.2 Speech and EGG database

A speech database with simultaneous EGG (Electro Glott Graph) recordings, made available by Nick Campbell (ATR Interpreting Telephony Research Laboratories), was used to evaluate the practical behavior 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 [9] and the 'get_f0' procedure in the ESPS [22]

F0s were extracted every 1ms. This exceedingly fine temporal resolution is meaningful for performance evaluation, because F0 extruction errors are dependent

errors with the proposed method			
	ordinary	subharmonic	total
NC:	2.86%	0.06%	2.92%
FHS:	0.96%	0.27%	1.23%
errors with improved AMDF			
	ordinary	subharmonic	total
NC:	1.90%	0.70%	2.60%
FHS:	0.87%	1.48%	2.35%
errors with get_f0 (ESPS)			
	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.

on noise and fluctuations in a analysis segment. A comparative performance evaluated based on the EGG recordings is given in Table 2. F0 differences greater than 20% were counted as errors. Note that by introducing a heuristic weighting on M, the error rate may be further reduced. These results indicate that the method is competitive or supersedes existing F0 extraction methods.

It is interesting to apply the proposed F0 extraction procedure both to EGG data and to corresponding speech, because the procedure only concentrates on the fundamental component while the conventional EGG measurement is based on measuring the interval between glottal closures which are inevitably affected by components other than the fundamental component.

Figure 14 and Figure 15 show histograms of error counts between the EGG results and speech results for the two subjects. The figures represent results with a heuristic weighting. In the gross error case, more than 20% of the F0 discrepancies are less than 0.8% in total. Moreover, more than 50% of the female data are within 0.3% of the EGG F0. This performance is one of the best ever, even with current technical standards.

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.)

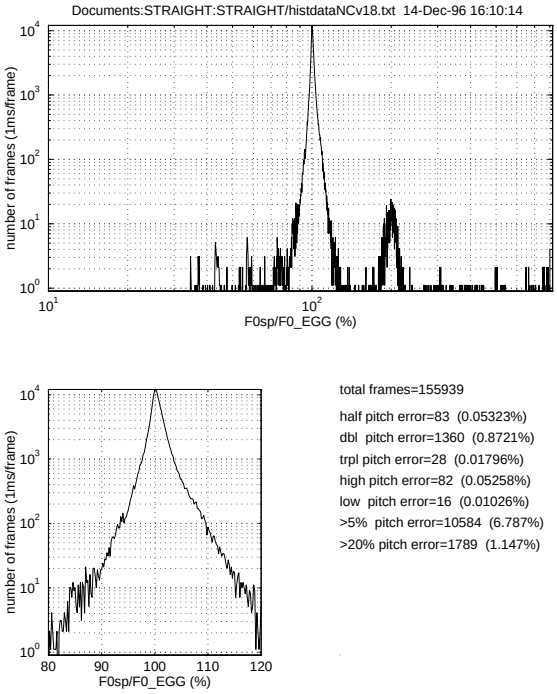

Figure 14: Histogram of extracted F0 in relation to F0 extracted from the EGG data. (Speaker: NC (male), with heuristics)

4 Integration to a speech manipulation system

These component procedures are integrated into a speech manipulation system as a channel VOCODER. Figure 16 illustrates the structure of the proposed system. It should be mentioned that there is an alternative implementation based on the sinusoidal model given in Equation 1. We will focus on the latter implementation here, because it provides simpler control on the perceptual attributes of the resynthesized sounds, especially for speech sounds.

The variable filter part of Figure 16 is implemented using a minimum phase impulse response and overlap and add procedure. The source generation part also employs all-pass filters to reduce the buzzy timbre due to a conventional pulse excitation. A brief description about the all-pass filter design can be found in our previous paper and detailed discussions will be given in the other paper.

The minimum phase impulse response is prefered in the current implementation because our auditory system is sensitive to a specific type of phase characteristics, in the other word, temporal asymmetry [21]. The

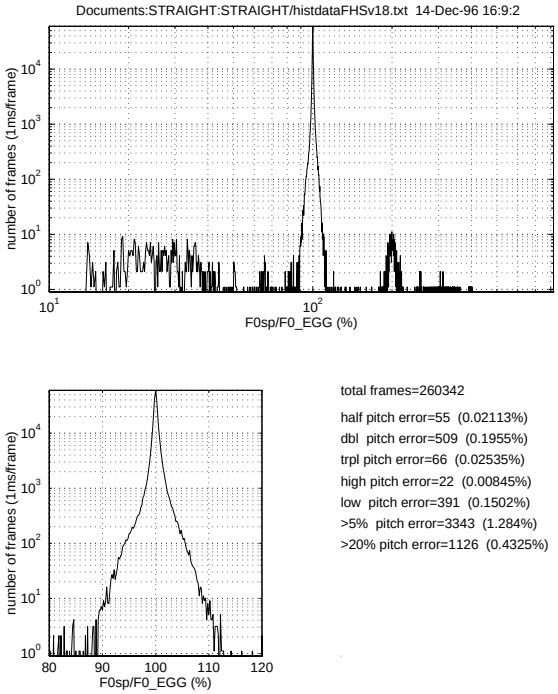

Figure 15: Histogram of extracted F0 in relation to F0 extracted from the EGG data. (Speaker: FHS (female), with heuristics)

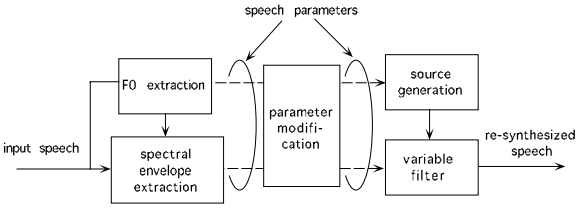

Figure 16: A schematic diagram of the proposed speech analysis-modification-synthesis system.

report on detectability of group delay also suggests that the magnitudes of group delays associated with spectral peaks and dips found in narural speech are perceptually detectable. It is therefore reasonable to adopt the physically feasible representation, the minimum phase impulse response, to implement the desired amplitude response.

4.1 Minimum phase impulse response and fine pitch control

A more formal description on the variable filter and the source generation parts are given here. In a source-filter model, the extracted f_0 (in fine resolution) is used to re-synthesize speech signal $y(t)$ using the following equation.

$$\begin{aligned} y(t) &= \sum_{l_i \in Q} \frac{1}{\sqrt{G(f_0(t_i))}} \; v_{li}(t - T(t_i)) \\ v_{li}(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} V(\omega, t_i)\Phi(\omega) e^{j\omega(l)} d\omega \quad (21) \\ &\text{where } T(t_i) = \sum_{l_k \in Q, k<i} \frac{1}{G(f_0(t_k))} \end{aligned}$$

where Q represents a set of positions in the excitation for the synthesis, and $G(\,)$ represents the pitch modification. $G(\,)$ can be an arbitrary mapping from the original F0 to the modified F0. For example, The all-pass filter function $\Phi(\omega)$ is used to control the fine pitch and the temporal structure of the source signal. For example, a linear phase shift in proportion to frequency is used to control F0 at a finer resolution than that determined by the sampling frequency.

$V(\omega, t_i)$ represents the Fourier transform of the minimum phase impulse response [20], 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.

$$\begin{aligned} V(\omega, t) &= \exp\left(\frac{1}{\sqrt{2\pi}} \int_0^{\infty} h_l(q) e^{j\omega q} dq\right) \quad (22) \\ h_l(q) &= \begin{cases} 0 & (q < 0) \\ c_l(0) & (q = 0) \\ 2c_l(q) & (q > 0) \end{cases} \end{aligned}$$

and

$$c_l(q) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-j\omega q} \log A(S(u(\omega), r(t)), u(\omega), r(t)) d\omega \quad (23)$$

where q represents the quefrency.

5 Analysis, modification and synthesis of real speech examples

A set of experiments involving analysis, modification and re-synthesis using real speech data, was conducted under the following conditions:
(1) Use of sampling frequencies of 8kHz, 12kHz, 16kHz, 22,050Hz, and 24kHz with 16-bit linear A/D converted speech.
(2) Analysis of isolated words and sentences spoken by male and female subjects in Japanese and English.
(3) Setting of the FFT length at 1024. (4) Analysis every 1 ms to produce 513 x 1000 data points per second.
(5)Extraction of fundamental frequencies every 1ms in a search range of from 40 Hz to 800 Hz without any iterative post-processing.

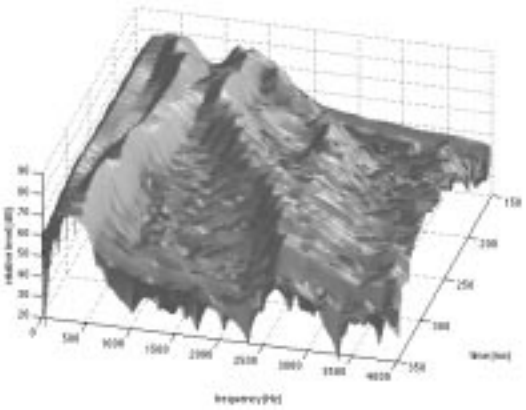

Figure 18: Pitch synchronous analysis: 3D representation.

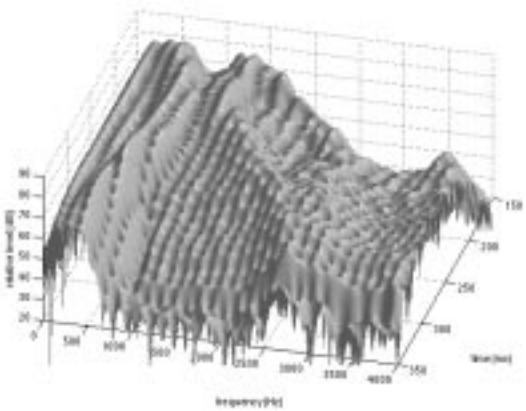

Figure 19: Isometric Gaussian spectrogram: 3D representation.

Examplese using real speech samples are shown to illustrate operations and various representations in the proposed method.

5.1 F0 extraction

Figure 17 shows results of the source information analysis using the proposed instantaneous-frequency-based procedure for a real speech example 'right' spoken by a female subject. The top plot shows the input waveform of an utterance of the word 'right' by a female speaker. The second panel shows the total power of wavelet outputs. The center plot is the extracted fundamental frequency. The thin line in the plot represents the voiced part, and the thick dots represent data points classified as unvoiced. The next plot shows 'fundamentalness' values associated with the extracted fundamental frequencies. The next plot illustrates output power that corresponds to channels consisting of the fundamental component. The bottom image illustrates a 'fundamentalness' map for all channels. Note that the voiced part corresponds to the salient dark blob in this map.

5.2 Spectral envelope extraction

The extracted F0 information is used to control the spectral envelope extraction procedure. The pitch synchronous analysis using a rectangular time window whose length is set equal to the fundamental period was also conducted for comparison. Figure 18 shows a 3D representation of the pitch synchronous spectrogram. The figure illustrates voiced portion. There are considerable spectral fluctuations in spectral valleys while spectral peaks in the lower frequency region shows smooth behavior.

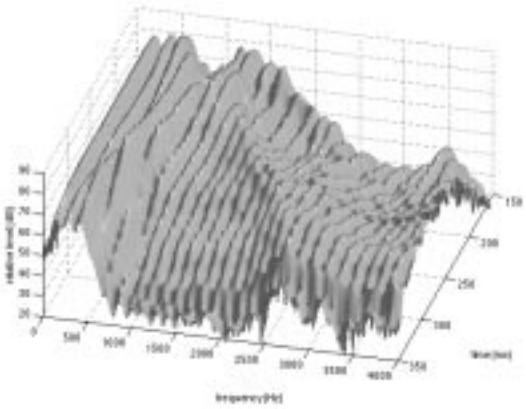

Figure 20: Spectrogram with reduced phasic interference: 3D representation.

Those fluctuations in spectral valleys are artifacts due to sharp discontinuities at the both edges of the rectangular time window. It can be concluded by comparison with Figure 19. Figure 19 shows a 3D spectrogram using an isometric Gaussian time window that is defined in Equation 2 and is pitch adaptive. As the Gaussian window has the minimum uncertainty, regular local peaks virtually represent the sampled values of the underlying spectral envelope. These peak values change more smoothly than the pitch synchronous spectrogram, both in the time domain and in the frequency domain, indicating that the underlying spectral envelope is much smoother than the conventional pitch synchronous analysis suggests.

Figure 20 shows a spectrogram calculated by a reduced phasic interference procedure. It also indicates

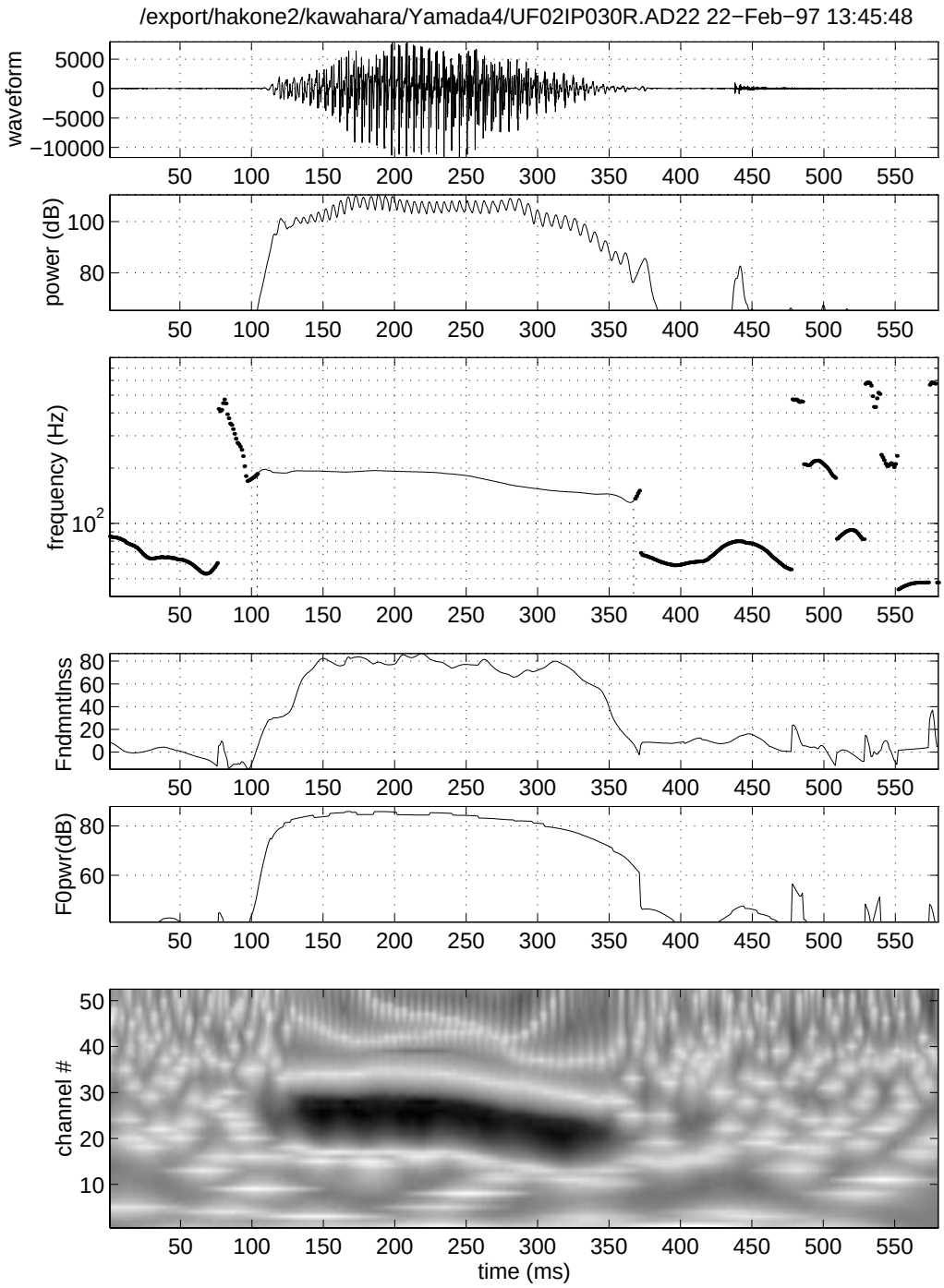

Figure 17: Source information of female pronunciation of 'right' extracted by the proposed instantaneous-based procedure.

that strength of each harmonic component changes smoothly with time. As the effective temporal resolution is about 1.4 times of the isometric Gaussian window shown in Figure 19, it is safe to conclude that the temporal change of the spectral envelope is actually smooth.

Figure 21 shows the final spectral envelope after eliminating both temporal and frequency interferences. Traces of harmonic components are effectively removed in the fugure and it makes undearlying spectral change like formant transitions salient.

Another examples of pitch synchronous analysis and

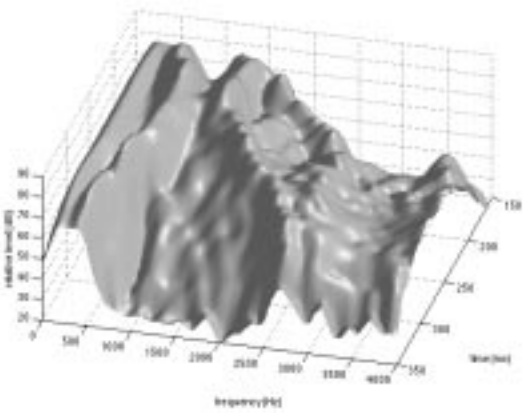

Figure 21: Spectrogram without the time-frequency interferences due to periodicity: 3D representation.

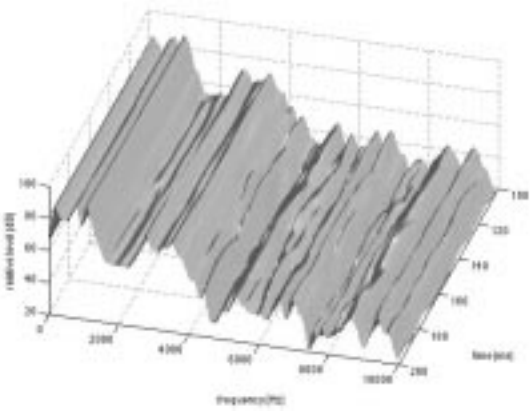

Figure 23: Pitch-adaptive smoothing using cardinal B-spline:3D representation.

,

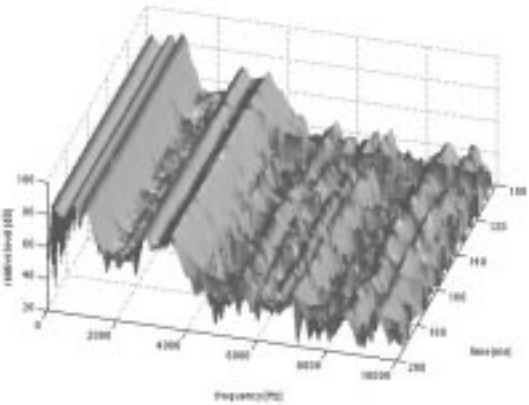

Figure 22: Pitch synchronous analysis:3D representation.

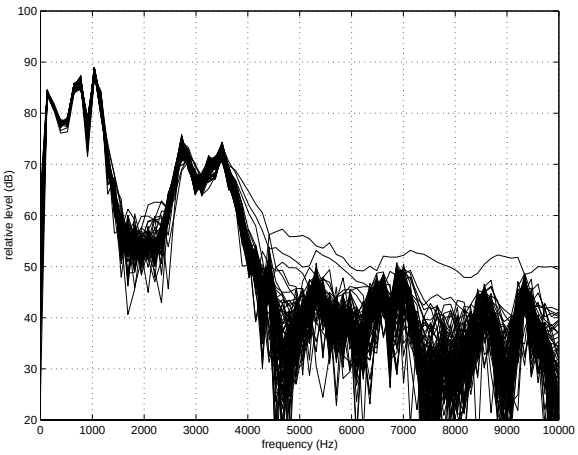

Figure 24: Pitch synchronous analysis:3D representation.

pitch adaptive smoothing for spectrum estimation are given in the following Figures 22, 24, 24 and 25. The sample is a sustained Japanese vowel /a/ spoken by a male subject.

Spectral valleys are also smeared in the pitch synchronous analysis. The smearing is clearly shown by comparing overlayed 2D spectral plots in Figures 24 and 25. The displayed portion is the same as the 3D plots.

5.3 Manipulations and resynthesis

The extracted F0 information and the spectral envelope are represented as a vector and a matrix of real numbers. These simple representations make parameter manipulation easy and direct. For example, to chage the speaking rate of the reproduced speech, an inverse mapping function from the target temporal axis to the original temporal axis enables the transformation. Modification of vocal tract length can be approximated by a linear conversion of frequency axis, because it is equivalent to modify the wave propagation time from glottis to mouth. F0 modification is also trivial. These parameters are fed to the source generation part and the variable filter part in Figure 16.

5.4 Resynthesized speech quality

The original sound files and the manipulated files are located on the web site which is accompanied with the Speech Communication Journal. Other ma-

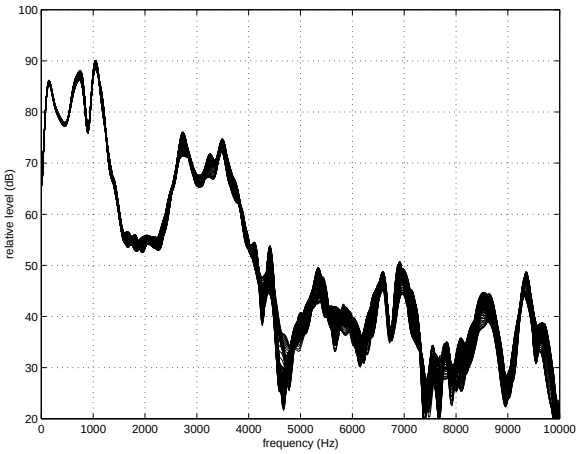

Figure 25: Pitch-adaptive smoothing using cardinal B-spline:3D representation.

,

nipulation examples are also presented on the same page. Informal listening tests have demonstrated that re-synthesized speech signals are sometimes indistinguishable from the natural ones even under careful headphone listening conditions. Witn an unrealistic combination of manipulation parameters, for example a very short vocal tract and a low F0, the re-synthesized speech sounds strange, but does not likely to have an artificial timbre. Although the proposed method is classified as an analysis-synthesis coding scheme, its quality as a coding system is comparable with waveform coding schemes like ADPCM. However, information rate is exceedingly high.

6 Discussion

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.

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 or a direct sinusoidal representation. This flexible representation and the non-parametric nature of the proposed method also open up various applications like voice morphing [23], electric musical instrument synthesis and efficient reuse of existing sound resources.

It is also have to be pointed out that this work was initially motivated by our need for a flexible and high-quality real-time analysis-synthesis method to use in our on-going experiments on auditory feedback [17]. As a result, the procedure consists only of feedforward algorithms and does not rely on iterations to optimize some criteria. Actually, the proposed method heavily uses a combination of Fourier and wavelet analysis techniques. The major difference between our method and prior methods is that ours involves information expansion, rather than reduction. This information expansion allows great flexibility in manipulation and stimulates an interesting speculation about the role of information expansion along with mammalian auditory pathways.

Analysis-and-synthesis methods such as ours were believed to give a poorer speech quality than waveform-based methods. The reproduced sounds by the proposed method, however, seem to provide a counter example. They suggest that the original concept of the VOCODER still holds, and that speech quality based on analysis-and-synthesis schemes can be improved further. This implies that precise reproduction of the source signal phase is not necessary for high-quality speech reproduction. Rather, there can be equivalent classes in which corresponding source signals have the same timbre while having different waveforms. It is practically as well as theoretically important to characterize these equivalent classes in terms of some statistical measures.

It has to be noted that the proposed set of procedures is not an ideal method in sound coding yet. The most elaborated part of our method is effective only for voiced speech and similar signals. Currently it uses simple STFT to estimate magnitude spectrum for unvoiced speech. It is necessary to incorporate more sophisticated signal models like MBE (MultiBand Excitation), [13, 11] multi-pulse [7] and others [3] to represent wider range of sounds appropriately. However, even with these shortcomings at this level of implementation, resynthesized speech using current system is almost equivalent to natural speech in terms of 'naturalness'. It also inherits conceptual simplicity and great flexibility in speech parameter control from the channel VOCODER. These characteristics make the proposed method a useful tool for speech perception and production research.

The proposed method allow us to test perceptual contributions of various spectral/temporal modifications in the vicinity of very natural reference signals. In the other words, it provides us a mean to test human speech perception mechanisms under ecologically valid situmuls conditions. Preliminary tests have suggested that human auditory perception is highly spe-

cialized for detecting changes which affect interpretation of auditory scene. The new concept 'fundamentalness' also provides an interesting interpretation of the pitch perception of inharmonic partials produced by AM. These, we believe, will provide interesting hints for developing a computational theory of auditory perception.

7 Conclusion

New procedures to represent and manipulate speech signals, based on pitch-adaptive spectral smoothing and instantaneous-frequency-based F0 extraction, have been presented. Elaborated procedures were designed for eliminating any traces of interferences caused by the signal periodicity to enable flexible manipulations of speech paremeters. These procedures are integrated to implement a sophisticated channel VOCODER system. We would like to call this set of procedures as STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum).

The proposed method offers high flexibility in parameter manipulations without introducing artificial timbre which is specific to synthetic speech signals, 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 also provides a versatile method for investigating quasi-periodic structures in arbitrary signals. These procedures may also provide an alternative approach for establishing 'the computational theory of Auditory Scene Analysis'.

Acknowledgment

The authors would like to express their sincere appreciation to their colleagues at ATR, to Dr. Roy Patterson of CNBH Cambridge and Dr. Toshio Irino of NTT (currently, of ATR) for their discussions. They also wish to express special thanks to their collaborator, J. C. Williams, for her discussions and encouragement. Finally, they would like to acknowledge that comments by the anonymous reviewers on the early version were very helpful to make this paper more readable.

References

[1] T. Abe, T. Kobayashi, and S. Imai. Harmonics estimation based on instantaneous frequency and its application to pitch determination. *IEICE Trans. Information and Systems*, E78-D(9):1188–1194, 1995.

[2] T. Abe, T. Kobayashi, and S. Imai. Robust pitch estimation with harmonics enhancement in noisy environments based on instantaneous frequency. In *Proc. ICSLP 96*, pages 1277–1280, Philadelphia, 1996.

[3] A. J. Abrantes, J. S. Marques, and I. M. Trancoso. Hybrid sinusoidal modeling of speech without voicing decision. In *Proceedings of Eurospeech 91*, pages 231–234, Paris, 1991.

[4] B. S. Atal and S. L. Hanauer. Speech analysis and synthesis by linear prediction of speech wave. *J. Acoust. Soc. Am.*, 50(2 pt.2):637–655, 1971.

[5] Boualem Boashash. Estimating and interpreting the instantaneous frequency of a signal – part 1: Fundamentals. *Proc. of IEEE*, 80(4):520–538, 1992.

[6] Boualem Boashash. Estimating and interpreting the instantaneous frequency of a signal – part 2: algorithms and applications. *Proc. of IEEE*, 80(4):550–568, 1992.

[7] Barbara Caspers and Bishnu Atal. Role of multipulse excitation in synthesis of natural-sounding voiced speech. In *Proceedings of IEEE int. Conf. Acoust., Speech and Signal Processing*, volume 4, pages 2388–2391, 1987.

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

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

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

[11] Thierry Dutoit and Henri Leich. An analysis of the performance of he mbe model when used in the context of a text-to-speech system. In *Proceedings of Eurospeech 93*, pages 531–534, Berlin, 1993.

[12] A. El-Jaroudi and J. Makhoul. Discrete all-pole modeling. *IEEE Trans.*, SP-39:411–423, 1991.

[13] Daniel W. Griffin and Jae S. Lim. Multiband excitation vocoder. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 36(8):1223–1235, 1988.

[14] F. Itakura and S. Saito. A statistical method for estimation of speech spectral density and formant frequencies. *Trans. IECE Japan*, 53-A:36–436, 1970. [in Japanese].

[15] Hideki Kawahara. Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited. In *Proceedings of IEEE int. Conf. Acoust., Speech and Signal Processing*, volume 2, pages 1303–1306, Muenich, 1997.

[16] Hideki Kawahara and Ikuyo Masuda. Speech representation and transformation based on adaptive time-frequency interpolation. *Technical Report of IEICE*, EA96-28:9–16, 1996. [in Japanese].

[17] Hideki Kawahara and J. C. Williams. Effects of auditory feedback on voice pitch. In Pamela J. Davis and Neville H. Fletcher, editors, *Vocal Fold Physiology*, chapter 18, pages 263–278. Singular Publishing Group, 1996.

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

[19] M. Narendranath, Hema A. Murthy, S. Rajendran, and B. Yenanarayana. Transformation of formants for voice conversion using artificial neural networks. *Speech Communication*, 16:207–216, 1995.

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

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

[22] B. G. Secrest and G. R. Doddington. An integrated pitch tracking algorithm for speech systems. *Proceedings of IEEE ICASSP83*, pages 1352–1355, 1983.

[23] Malcolm Slaney, Michele Covell, and Bud Lassiter. Automatic audio morphing. In *Proceedings of IEEE int. Conf. Acoust., Speech and Signal Processing*, pages 1–4, Atlanta, 1996.

[24] Yannis Stylianou, Jean Laroche, and Eric Moulines. High-quality speech modification based on a harmonic + noise model. In *Proceedings of Eurospeech 95*, pages 451–454, Madrid, 1995.

[25] 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.