ELSEVIER

Speech Communication 27 (1999) 187–207

SPEECH COMMUNICATION

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 [1]

Hideki Kawahara [*,2], Ikuyo Masuda-Katsuse [3], Alain de Cheveigné [4]

ATR Human Information Processing Research Laboratories, 2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan

Received 12 February 1998; received in revised form 24 September 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. The method also consists of a fundamental frequency (F0) extraction 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 reproductive quality.

Keywords: Speech analysis; Pitch-synchronous; Spline smoothing; Instantaneous frequency; F0 extraction; Speech synthesis; Speech modification

1. Introduction

The need for flexible speech modification methods is increasing in both commercial and scientific fields. Various sophisticated methods have been proposed (McAulay and Quatieri, 1986; Stylianou et al., 1995; Narendranath et al., 1995; Veldhuis and He, 1996) but their flexibility and resultant speech quality have been limited. This suggests that an appropriate representation of speech has yet to be found, or it may simply indicate that we have not fully explored the potential of existing representations. If we revisit old concepts with 'Auditory Scene Analysis' (Bregman,

* Corresponding author. E-mail: kawahara@sys.wakayama-u.ac.jp

[1] Speech files available. See http://www.elsevier.nl/locate/specom

[2] Present address: Department of Design Information Science, Faculty of Systems Engineering, Wakayama University, 930 Sakaedani, Wakayama, Wakayama 640-8510, Japan. Also a project head under CREST.

[3] Present address: Institute of System and Information Technologies/KYUSHU, 2-1-22-7F, Momochihama, Sawara-ku, Fukuoka 814-0001, Japan.

[4] Present address: Laboratoire de Linguistique Formelle, CNRS/Université Paris 7, case 7003, 2 place Jussieu, 75251 Paris CEDEX 05, France.

0167-6393/99/$ – see front matter
PII: S0167-6393(98)00085-5

1990) in mind, perhaps we can develop this potential. Indeed, special emphasis on ecological point of views found in the book, promoted our investigations toward seeking robust procedures to extract stable structures for periodic sounds which are vitally important. Speech sound for human is such an example. It is our belief when underlying principle of procedures are conceptually simple and shows robustness under ecologically valid conditions, it is very likely that human Auditory Scene Analysis functions already take advantage of the principle at least at the computational level in Marr's terminology (Marr, 1982).

The channel VOCODER (Dudley, 1939) which separates spectral and source information to manipulate and transmit speech sounds, is a good example of a simple and appealing idea made powerful by introducing such views. The channel VOCODER and its modern variant, Linear Predictive Coding (LPC) (Itakura and Saito, 1970; Atal and Hanauer, 1971) are potentially very flexible in parameter manipulations, because there are few inter-related constraints between spectral and source parameters. VOCODER-based representations are also attractive, because they are conceptually easy to understand and have direct correspondence to the speech production mechanism and the auditory periphery. However, the resynthesized speech quality of VOCODERs suffers from buzziness induced by pulsive excitation even without parameter manipulations. This quality is also degraded when parameter manipulations are large. On the other hand, sophisticated methods based on iterative procedures that approximate desired manipulated short term Fourier transformations (STFTs) (Veldhuis and He, 1996) have superior reproduction quality with a small amount of manipulations. However, they deteriorate rapidly when parameter manipulations increase. Intricate relations between the manipulated spectral parameters and waveforms also make it difficult to get insight from these representations. Simple concepts like the channel VOCODER may seem outdated, but if the VOCODER allows for a high quality reproduction, the simpler the better.

There have been many effective proposals (Abrantes et al., 1991; Stylianou et al., 1995; Caspers and Atal, 1987; Griffin and Lim, 1988) for reducing the buzziness of VOCODER-type methods by manipulating synthetic source signals. We also proposed a unique method for this problem based on a group delay manipulation of all-pass filters in our previous reports (Kawahara, 1997; Abe et al., 1996). The method is already built into our speech modification system. This topic leads to an interesting discussion on temporal aspects of human auditory perception. The detailed discussions will be given in the other paper, because it makes this paper too complicated.

Then, the major remaining problem is eliminating errors in spectral estimation. It is necessary to remove any periodicity interferences from the time–frequency representation for the representation to be usable in reproducing a spoken sound in a different fundamental frequency (F0) or in a different vocal tract length. Parametric models like LPC are also susceptible to signal periodicity (El-Jaroudi and Makhoul, 1991) even though they can alleviate constraints posed by the uncertainty principle. These interferences are observed as an apparent increase in random variations in spectral representations. In a sense, it is 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 must 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 that does not have any trace of periodicity.

A flexible speech manipulation also introduces a requirement for F0 trajectories. Conventional F0 extraction methods based on interval measurements usually provide stepwise trajectories, especially for low-pitched voices. This stepwise structure is a trace of the source periodicity and is harmful to F0 modifications. Therefore, it is desirable to have a F0 extraction method that provides a smooth trajectory.

The goal of this paper is to introduce the implementation of a very high quality speech analysis-modification-synthesis method as a channel VOCODER based on the aforementioned conditions. Information reduction is not intended in this paper since our primary interests are quality and

flexibility of manipulations. The paper consists of four sections. First, a pitch-adaptive spectral smoothing to eliminate periodicity interference is discussed. Then, an instantaneous-frequency-based F0 extraction method is introduced to provide reliable and smooth F0 trajectories. Next, a system utilizing the proposed representations to manipulate speech parameters is introduced. Finally, we present examples of real speech analysis and manipulations.

2. Elimination of periodicity interference

In this section, a method for eliminating the spectral interference structure caused by signal periodicity is systematically introduced. First, the basic principle of the proposed method is introduced as an adaptive smoothing of a time–frequency representation. Then, a compensatory time window design is shown that reduces this time–frequency smoothing problem to a smoothing problem in a frequency domain. Finally, a procedure that eliminates and compensate for 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 periodic variation in the frequency domain. If the signal is purely periodic and the period is an integer multiple of the sampling period, a pitch-synchronous analysis can perfectly eliminate temporal variations by using a rectangular window, whose 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 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 such signals change all the time, and each repetition is not the same as the previous period due to natural source related fluctuations. The sharp discontinuities at both ends of the rectangular window makes the analysis highly sensitive to such minor fluctuations. Spectral smearing, which is caused by this discontinuity and fluctuations, is detectable due to the wide spectral dynamic range of natural speech signals. It is also not practical to extract a portion representing an impulse response, because responses corresponding to formant peaks do not die out within a pitch period, unless the pitch is extremely low.

The other approach to eliminate periodicity-related interferences introduces a spectrum model that embodies periodicity effects. This approach was proposed for LPC parameter estimation (El-Jaroudi and Makhoul, 1991). The results that used synthetic speech signals indicated that interferences due to the signal periodicity are well compensated. However, this approach does not provide reliable estimates for natural speech, because this method assumes that the auto-correlation 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 a natural speech spectral envelope, which are not modeled in auto-regressive part, result in an unpredictable smearing of the auto-correlation function. In addition, 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 factors suggest that the elimination process of periodicity interferences should neither rely on 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 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; these axes represent the global source characteristics and the shapes and movements of articulation organs. In this interpretation, a periodic signal $s(t)$ 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. Furthermore, 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 non-stationary repetitive aspect of speech waveforms is as follows:

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

where $\alpha_k(t)$ represents the time varying amplitude of kth harmonic component, $\omega(\tau)$ represents a common time varying fundamental angular frequency, $\omega_k(\tau)$ represents a time varying fluctuation angular frequency of kth 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 $\omega(\tau)$ and $\omega_k(\tau)$) 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 a sinusoidal representation (McAulay and Quatieri, 1986; Boashash, 1992a) but the procedure in using this formulation differs substantially.

A short-term Fourier analysis of this signal yields a time–frequency representation of the signal $F(\omega, t)$, known as a spectrogram (Cohen, 1989). 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. Therefore, it is desirable to use a time windowing function, $w(t)$, which has equivalent relative resolution in both the time and frequency domains to take full advantage of the available partial information.

Let us assume that the time window function $w(t)$ has the following form:

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

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

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 this change.

Our goal is to reconstruct a smooth 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 based on partial information. Thus, it is necessary to provide constraints for the problem so that a unique solution can be obtained. One reasonable constraint is to use only local information.

Let us consider a one-dimensional operation. The simplest reconstruction, which uses discretized partial information like amplitudes of harmonic components of voiced sounds, connects harmonic peaks with straight lines. In other words, it represents the reconstructed surface as a piecewise first order polynomial. 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 represen-

tation. It is a convolution (smoothing) using a 2nd-order cardinal B-spline function. The procedure is illustrated in Fig. 1. This smoothing operation is preferable to interpolation for real speech, where noise and error in F0 estimation are inevitable because the smoothing operation is less sensitive to such problems and resultant errors are localized, as shown in the figure.

In previous papers (Kawahara and Masuda, 1996; Kawahara, 1997) we proposed a two-dimensional smoothing procedure to implement the basic idea. In this report, we introduce a set of pitch-adaptive time windows to calculate the phase insensitive power spectra that reduce this two-dimensional smoothing operation into the one-dimensional smoothing operation mentioned above.

2.3. Power spectrum with reduced phasic interference

In this section, a compensatory set of time windows providing an effectively phase insensitive spectrogram is introduced. As a result, it is unnecessary to apply the 2nd-order cardinal B-spline smoothing function to remove the periodic interference from such a spectrogram. Instead, it is sufficient to perform the spline-based smoothing only in the frequency domain, once the temporal interference is effectively eliminated.

First, an exemplar periodic interference using an isometric real valued time window is illustrated. Fig. 2 shows the interference for a test signal consisting of regular pulses with a constant fundamental period (10 ms). Regular 'holes' exist 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 whose size is adapted to the fundamental period. The 2nd-order cardinal B-spline function is selected because it places the 2nd-order zeros on the other harmonic frequencies. This makes the resultant spectrum less sensitive to F0 estimation errors. Fig. 3 shows a three-dimensional spectrogram of the same pulse train using the time window given in the following equation:

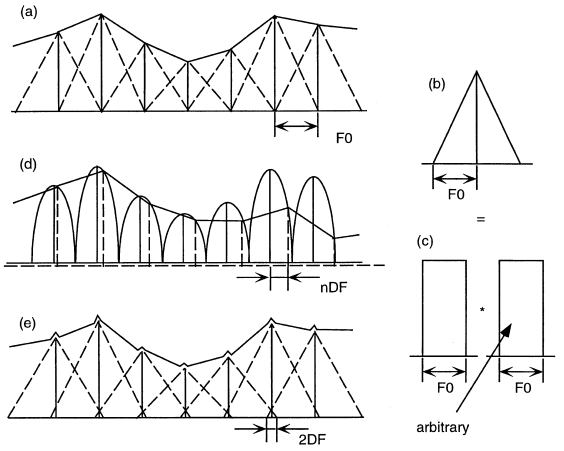

Fig. 1. Interpolation and smoothing in one-dimension: (a) a piecewise linear interpolation and equivalent smoothing using smoothing function (b) constructed by convolution of two first-order cardinal B-spline functions (c). (b) Second-order cardinal B-spline function. (d) Effects of F0 errors on interpolation only relying on knot points. Note distortion increases proportionally with harmonic number. 'DF' denotes error in F0 estimation and 'n' represents a harmonic number. (e) Effects of F0 errors on smoothing. Note localized distortion independent of harmonic number.

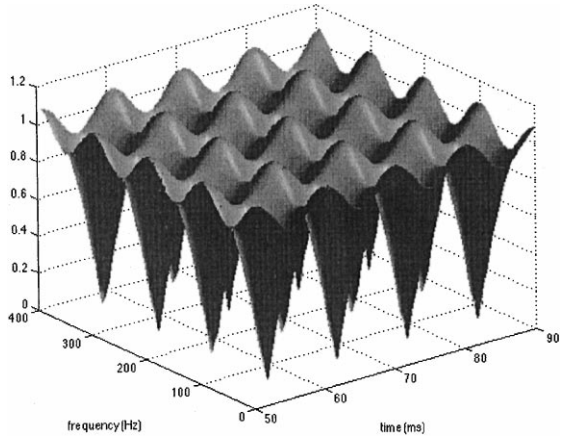

Fig. 2. Three-dimensional spectrogram of regular pulse train (100 Hz). The isometric Gaussian time window is used to calculate this spectrogram. Vertical scale is linear and represents absolute value of spectrogram.

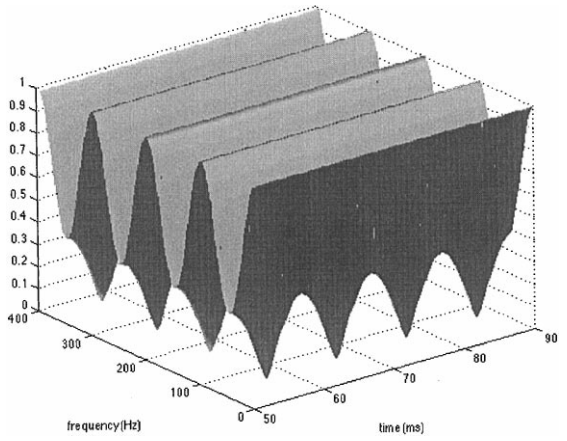

Fig. 3. Three-dimensional spectrogram of regular pulse train (100 Hz) with pitch synchronous modification to the original Gaussian window.

$$w_{\mathrm{p}}(t) = \mathrm{e}^{-\pi(t/t_0)^2} \odot h(t/0),$$

$$h(t) = \begin{cases} 1 - |t|, & |t| < 1, \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

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

It is also possible to design a compensatory time window that produces maxima where the original spectrogram has 'holes'. The compensatory time window for the pitch-synchronous time window, $w_{\mathrm{p}}(t)$, has the following form. The sinusoidal modulation of Eq. (5) is designed to perform frequency conversion and phase shifting for fulfilling the requirement. Consider a set of neighboring harmonic components. The sinusoidal modulation converts the lower harmonic component up to the amount of F0/2 and converts the higher harmonic component down 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 an out-of-phase in-phase and produces maxima where the original spectrogram has 'holes'. The window function is illustrated in Fig. 4.

$$w_{\mathrm{c}}(t) = w_{\mathrm{p}}(t) \sin\left(\pi \frac{t}{t_0}\right). \tag{5}$$

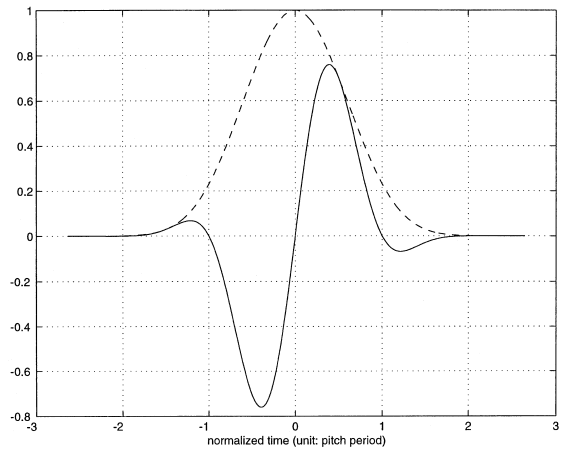

Fig. 4. Compensatory time window (solid line) and its original time window (dashed line).

The power spectrum using this compensatory window is shown in Fig. 5. A power spectrum with reduced phasic interference, $P_{\mathrm{r}}(\omega, t)$, is represented as a weighted squared sum of the power spectra, $P_{\mathrm{c}}(\omega, t)$ and $P_{\mathrm{o}}(\omega, t)$, using this compensatory window and the original time window, respectively.

$$P_{\mathrm{r}}(\omega, t) = \sqrt{P_{\mathrm{o}}^2(\omega, t) + \xi P_{\mathrm{c}}^2(\omega, t)}, \tag{6}$$

where ξ minimizes the temporal variation of the resultant spectrogram. A numerical search proce-

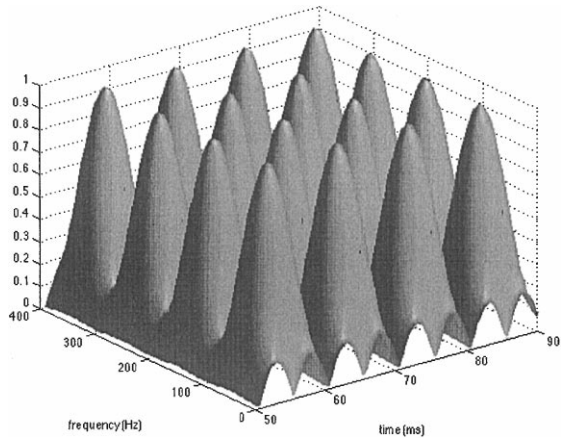

Fig. 5. Three-dimensional spectrogram of regular pulse train using compensatory time window. Note that peaks are located at the positions of 'holes' in the isometric spectrogram.

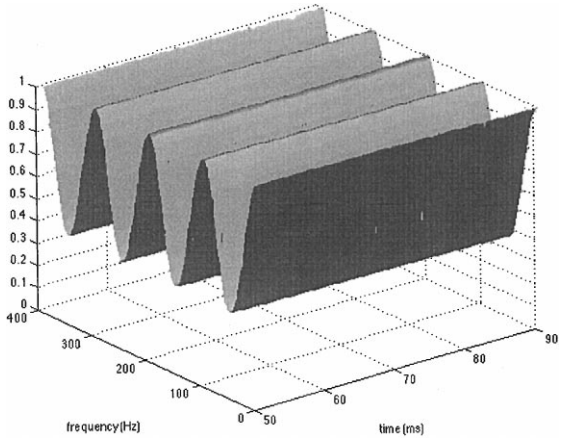

Fig. 6. Three-dimensional spectrogram with reduced phase effects.

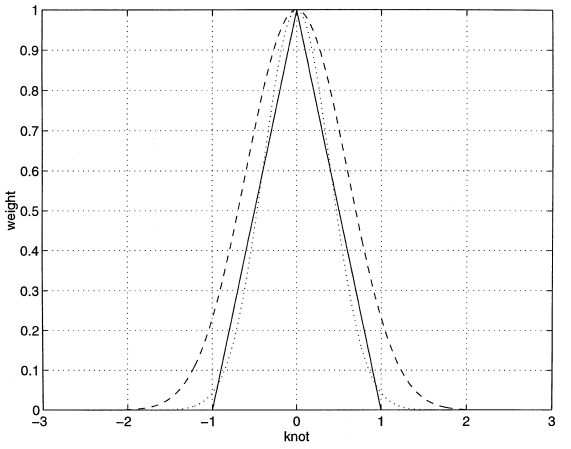

Fig. 7. Over smoothing example. Desired basis function of 2nd-order cardinal B-spline (solid line) is smoothed out (dashed line) by smoothing effect of time window (dotted line).

dure provided 0.13655 for the optimum blending factor ξ. Fig. 6 illustrates the reduced phasic interference spectrogram obtained using this ξ.

Using this optimum blending factor, it is possible to eliminate the need for temporal smoothing using the 2nd-order cardinal B-spline smoothing function. It also enables a slower frame rate for calculating the spectrogram. Such changes from the previous implementation are very effective in reducing the computational demand.

2.4. Compensation of over-smoothing in the frequency domain

One problem with the algorithm described in the previous section is over-smoothing, which is a result of smearing caused by the time windowing function. For example, in the frequency domain, a power spectrum calculated by a 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 the already smoothed spectral representation to eliminate interferences caused by the signal periodicity. This operation successfully removes the interference, but it also simultaneously smoothens the underlying spectral structure. This over-smoothing is illustrated in Fig. 7. The over-smoothing $v(\omega)$ is also represented as a convolution of the frequency representation of the time window $W(\omega)$ and the 2nd-order cardinal B-spline smoothing function in the frequency domain $h(\omega)$.

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

If the over-smoothed spectrum is reasonably approximated by the convolution 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 $(k = -N, \ldots, N)$.

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

where

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

N is a reasonably large integer that effectively makes v_N negligible and v_k represents $v(k\omega)$. It is

necessary to use more than the $2N+1$ set of this relation to find a unique solution. This yields the simultaneous linear equations:

$$\begin{aligned}
&\boldsymbol{u} = H\boldsymbol{c},\\
&\boldsymbol{u} = [u_{-M}, u_{-M+1}, \ldots, u_0, \ldots, u_{M-1}, u_M]',\\
&\boldsymbol{c} = [c_{-N}, c_{-N+1}, \ldots, c_0, \ldots, c_{N-1}, c_N]',
\end{aligned} \tag{10}$$

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

where $[\,]'$ represents the transpose of a matrix. The solution is represented as follows:

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

Fig. 8 shows the optimal smoothing function for a previously introduced isometric Gaussian time window.

Fig. 9 also shows the shape of the compensated over-smoothing function. Note that only one value at each node point is non-zero.

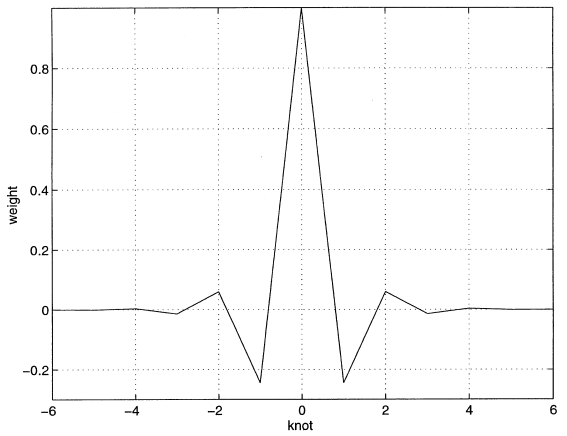

Fig. 8. Optimal 2nd-order smoothing function.

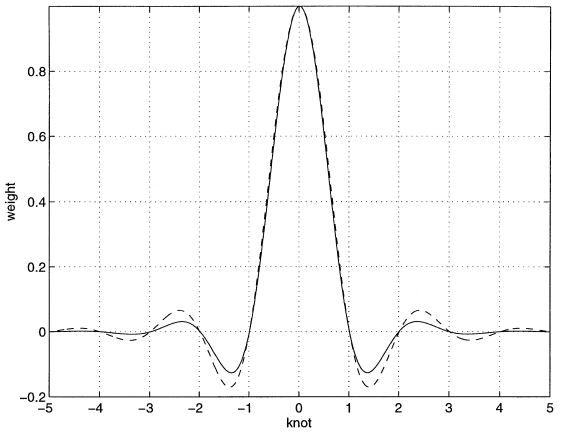

Fig. 9. Recovered impulse using the smoothing functions (solid line: 2nd-order; dashed line: 4th-order).

However, this optimal smoothing function presented a new problem. We required the algorithm to be localized. The support size of the optimal smoothing function is effectively three times larger than the original smoothing function. It is desirable to calculate a quasi-optimal smoothing function with less support. The desired quasi-optimal smoothing function can be calculated by making N small, 1 for example. Fig. 10 shows the shape of the recovered impulse using a quasi-optimal smoothing function that consists of three 2nd-order cardinal B-spline functions ($N=1$).

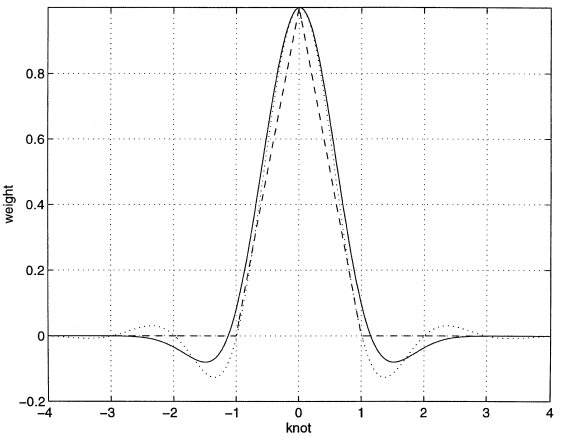

Fig. 10. Recovered impulse using optimal smoothing function and the quasi-optimal smoothing function (solid line: quasi-optimal; dotted line: optimal; dashed line: basis function).

3. Extraction of smooth and reliable F0 trajectories

For flexible and high-quality modification of speech parameters, it is important to extract F0 trajectories without any trace of interferences caused by phasic interaction between the 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 Eq. (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 (Boashash, 1992a, b; Cooke, 1993; Abe et al., 1995, 1996) of the fundamental component in Eq. (1). Therefore, extracting the instantaneous frequency of the fundamental component directly is better if we use the signal model represented by Eq. (1).

3.1. Basic principle

The fundamental frequency is extracted as the instantaneous frequency of the signal's fundamental component in the proposed method. This may sound strange to some readers because selecting the fundamental component requires prior knowledge of the fundamental frequency.

This apparent contradiction is solved by introducing a measure to represent the fundamentalness without using a priori knowledge of the fundamental frequency. A fairly wide class of analyzing wavelets makes the output corresponding to the fundamental component have smaller FM and AM than 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 the FM and AM modulation magnitudes are minimum and to have a monotonic relation with the modulation magnitudes.

Fig. 11 illustrates how this definition works when analyzing a complex sound with several harmonic components. When no harmonic component is within the response area of the analyzing wavelet ((a) in the figure), fundamentalness 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 ((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 ((c) in the figure), the highest signal to noise ratio causes the fundamentalness to be maximized. When the frequency of a harmonic component other than the fundamental component agrees with the best (or characteristic) frequency of the analyzing wavelet ((d) in the figure), even though the signal to noise ratio in terms of the specific harmonic component provides the highest value, the fundamentalness is not high. This is because two or more harmonic components are located within the response area as a result of the intended

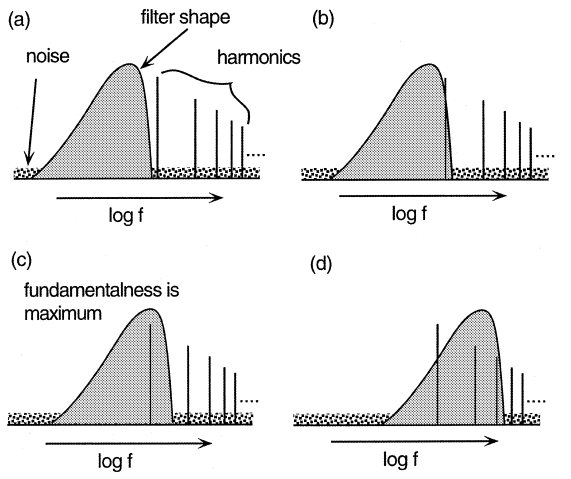

Fig. 11. Schematic explanation of 'fundamentalness'. (a) An analyzing wavelet (filter shape) does not cover harmonic component. (b) Only fundamental component inside receptive area. (c) Fundamental component located at best frequency of analyzing wavelet. (d) Higher harmonic component located at best frequency of analyzing wavelet.

filter shape design and these components produce beating that consists of AM and FM. The other cases also provide lower fundamentalness. Thus, maximum fundamentalness assures that the filter actually corresponds to the fundamental component in question.

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 with a slightly finer resolution in frequency (i.e. $\eta > 1$, η represents the temporal stretching factor), the input signal can be divided into a set of filtered complex signals $D(t, \tau_c)$.

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

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

$$g(t) = e^{-\pi(t/\eta)^2} e^{-j2\pi t}. \tag{15}$$

The characteristic period of the analyzing wavelet τ_c is used to represent the corresponding filter channel. The shift and subtraction in Eq. (14) puts zeros at zero-frequency (DC component) and at harmonic frequencies of even order components.

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:

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

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 represents 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 factors 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 and amplitude envelope normally consist of rapid movements that carry prosodic information. Removing the contribution of the monotonic F0 movement and amplitude change reduces artifacts on the fundamentalness evaluation caused by prosodic components.

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

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

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

where μ_{FM} and μ_{AM} represent monotonic FM change and AM change, respectively. Extracting F0 simply means finding the maximum index of M_c in terms of τ_c and calculating the average (or more specifically, interpolated) instantaneous frequency using the outputs of the channels neighboring τ_c.

The instantaneous frequency $\omega(t)$ of one filter output signal $D(t, \tau_c)$ is calculated using the following equation. This equation is derived to remove unwrapping in the usual calculation of instantaneous frequency.

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

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

3.2. Evaluation of F0 extraction performance

Details regarding the performance and properties of the proposed F0 extraction method are important because they 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.

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, 40 to 0 dB in 10 dB steps. Only a 100 Hz pulse train was tested because the proposed procedure is scalable and independent of the sampling frequency and F0. Table 1 lists the results. The F0 search range was from 40 to 800 Hz, and no post-processing was involved. The ratio between center frequencies of adjacent channels was $2^{1/12}$. The last line shows the result obtained when an envelope signal was used as the input instead of using the 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.

Table 1
Relation between S/N and rms (root mean square) error in F0 extraction for a pulse train with white noise (last row is result when envelope of original signal calculated using Hilbert transform used as input to proposed F0 extraction procedure)

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

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

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

3.2.2. Speech and EGG database

A speech database with simultaneous EGG (Electro Glott Graph) recordings, provided by Nick Campbell (ATR Interpreting Telephony Research Laboratories), was used to evaluate the practical behavior of the proposed F0 extraction method. A subset of 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 s for the male data and 266 s for the female data. An improved AMDF method (de Cheveigné, 1996) and the 'get_f0' procedure in the ESPS (Secrest and Doddington, 1983) were also tested.

F0s were extracted every 1 ms. This exceedingly fine temporal resolution was meaningful for performance evaluation, because F0 extraction errors are dependent on noise and fluctuations in an analysis segment. A comparative performance was evaluated based on the EGG recordings as shown in Table 2. F0 differences greater than 20% were counted as errors.

Note that the error rate may be further reduced by introducing a heuristic weighting on M_c. These results indicate that the method is competitive or supersedes existing F0 extraction methods.

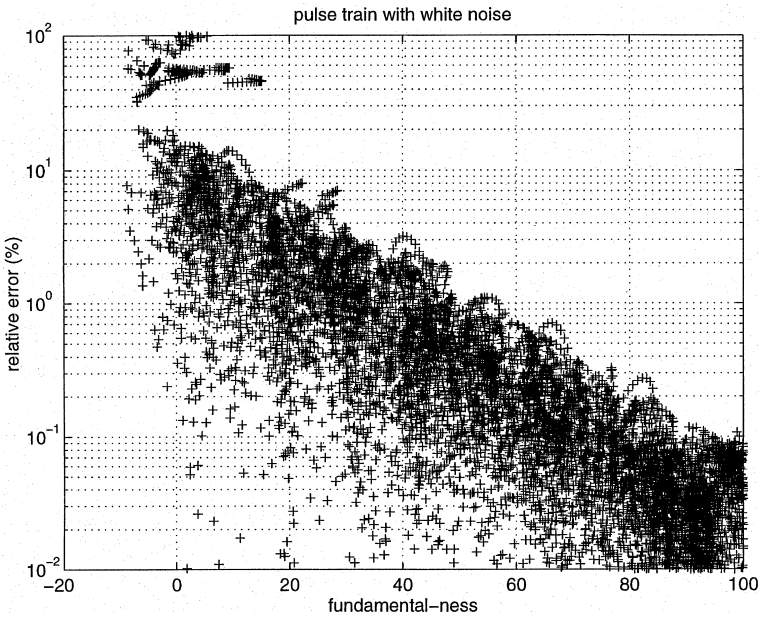

Fig. 12. Scatter plot of 'fundamentalness' and F0 errors.

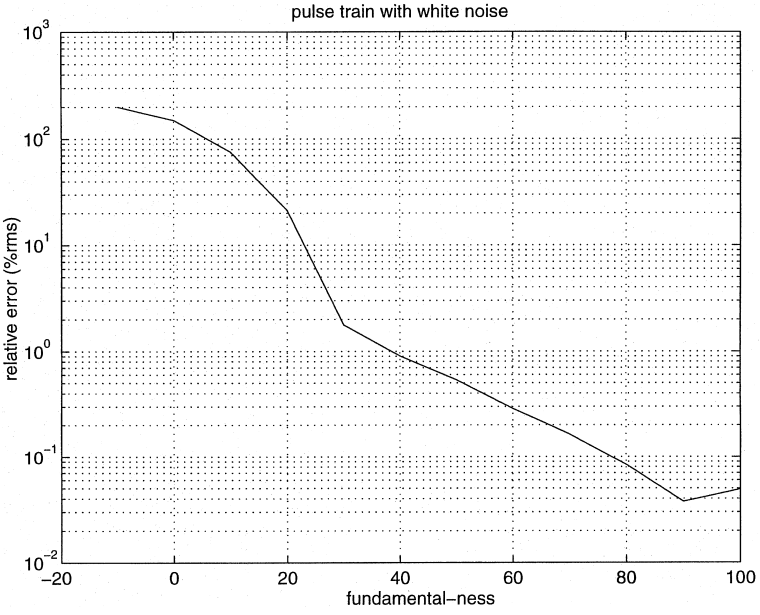

Fig. 13. Relation between 'fundamentalness' and rms F0 errors.

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

Table 2
Comparative performance of proposed method, improved AMDF method, and commercial method

	Ordinary (%)	Subharmonic (%)	Total (%)
Errors with the proposed method			
NC	2.06	0.06	2.92
FHS	0.96	0.27	1.23
Errors with improved AMDF			
NC	1.90	0.70	2.60
FHS	0.87	1.48	2.35
Errors with get_f0 (ESPS)			
NC	0.31	2.65	2.96
FHS	3.28	0.93	4.21

Figs. 14 and 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. The heuristic weighting was designed to suppress double pitch errors and half pitch errors by biasing fundamentalness. 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 obtained, even with current technical standards.

It should be pointed out that the F0 extraction procedure, which was developed as a part of the

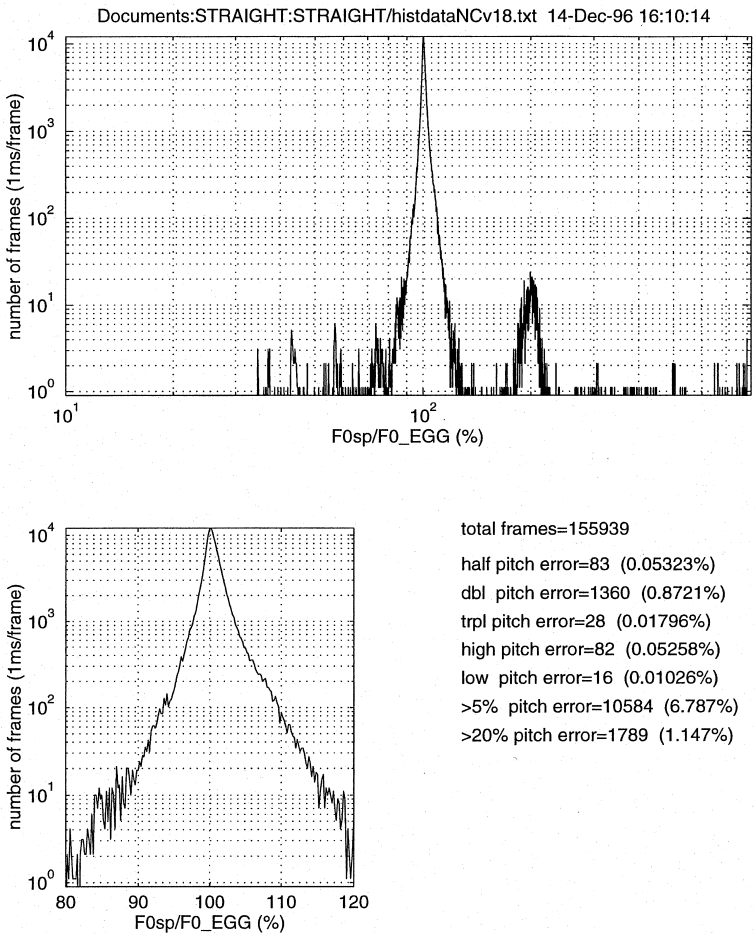

Fig. 14. Histogram of extracted F0 in relation to F0 extracted from EGG data. (Speaker: NC (male), with heuristics.)

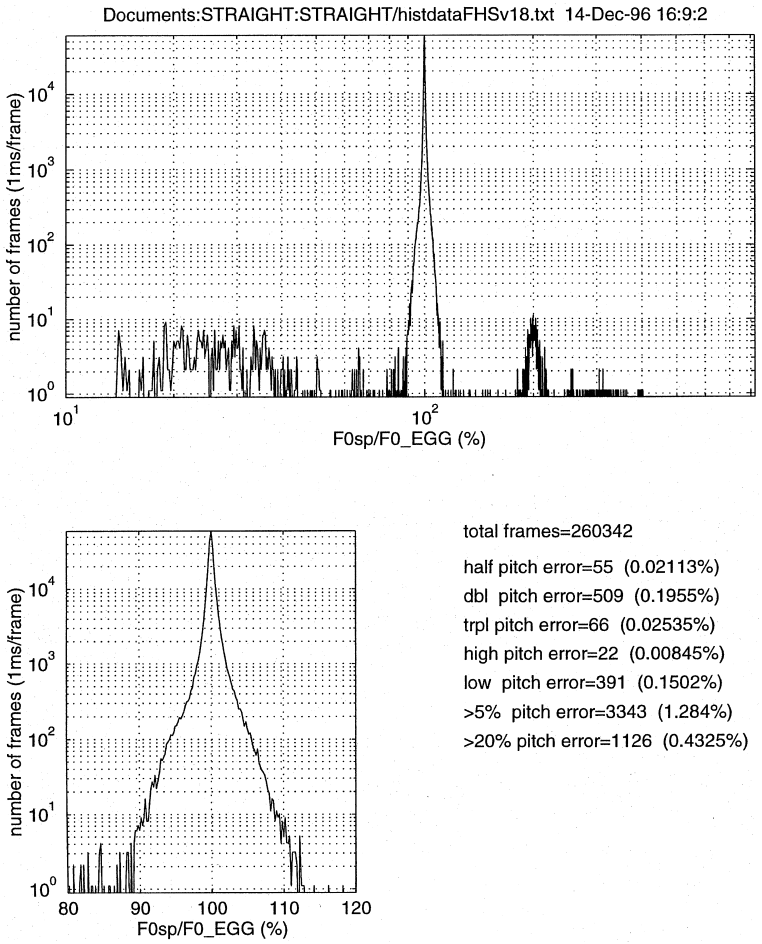

Fig. 15. Histogram of extracted F0 in relation to F0 extracted from EGG data. (Speaker: FHS (female), with heuristics.)

proposed speech modification method, demonstrated an extremely accurate and robust performance. It can be used as a general purpose procedure for extracting 'fundamental-like' components in arbitrary signals. We would like to suggest calling the procedure TEMPO (Time-domain Excitation extractor using Minimum Perturbation Operator).

4. Integration to a speech manipulation and re-synthesis system

The component procedures are integrated into a speech manipulation system for a channel VOCODER. Fig. 16 illustrates the structure of the proposed system. There is an alternative implementation based on the sinusoidal model given in Eq. (1). We will focus on the VOCODER implementation here, because it provides simpler con-

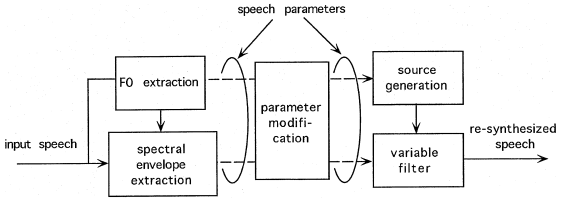

Fig. 16. Schematic diagram of proposed speech analysis-modification-synthesis system.

trol of the perceptual attributes of the re-synthesized sounds, especially for speech sounds.

The variable filter part of Fig. 16 is implemented using a minimum phase impulse response (Oppenheim and Schafer, 1989) and overlap and add procedure. The source generation part also employs all-pass filters to reduce the buzzy timbre resulting from a conventional pulse excitation. A brief description of the all-pass filter design is in our previous papers (Kawahara, 1997; Kawahara et al., 1996) and detailed discussions will be given in an another paper.

The minimum phase impulse response is preferred in the current implementation because our auditory system is sensitive to a specific type of phase characteristics, namely, temporal asymmetry (Patterson, 1987). A report on the detectability of group delay (Blauert and Laws, 1978) also suggests that the magnitudes of group delays, which are associated with spectral peaks and dips found in natural speech, are perceptually detectable. Therefore, it is 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 F0 (in fine resolution) is used to re-synthesize speech signal $y(t)$ using the following equation:

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

$$v_{ti}(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} V(\omega, t_i)\Phi(\omega)\mathrm{e}^{\mathrm{j}\omega(t)}\,\mathrm{d}\omega, \tag{21}$$

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 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, $G(x) = 2x$ doubles the original F0 and an implementation using a table and interpolation allows arbitrary modifications. 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 (Oppenheim and Schafer, 1989) 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.

$$V(\omega, t) = \exp\left(\frac{1}{\sqrt{2\pi}} \int_0^{\infty} h_t(q)\mathrm{e}^{\mathrm{j}\omega q}\,\mathrm{d}q\right), \tag{22}$$

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

and

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

where q represents the quefrency.

5. Analysis, modification and synthesis of real speech examples

A set of experiments involving analysis, modification, and re-synthesis of real speech data was conducted under the following conditions:

1. using sampling frequencies of 8, 12, 16 kHz, 22.050 Hz and 24 kHz with 16-bit linear A/D converted speech;
2. analyzing of isolated words and sentences spoken by male and female subjects in Japanese and English;
3. setting the FFT length at 1024;
4. analyzing every 1 ms to produce 513 × 1000 data points per second;

5. extracting fundamental frequencies every 1 ms in a search range from 40 to 800 Hz without any iterative post-processing.

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

5.1. F0 extraction

Fig. 17 shows the source information results using the proposed instantaneous-frequency-based procedure for an example utterance 'right' spoken by a female subject. The top plot shows the input waveform of the utterance. 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 the fundamentalness values associated with the extracted fundamental frequencies. The next plot illustrates the output power for channels consisting of the fundamental component. The bottom image illus-

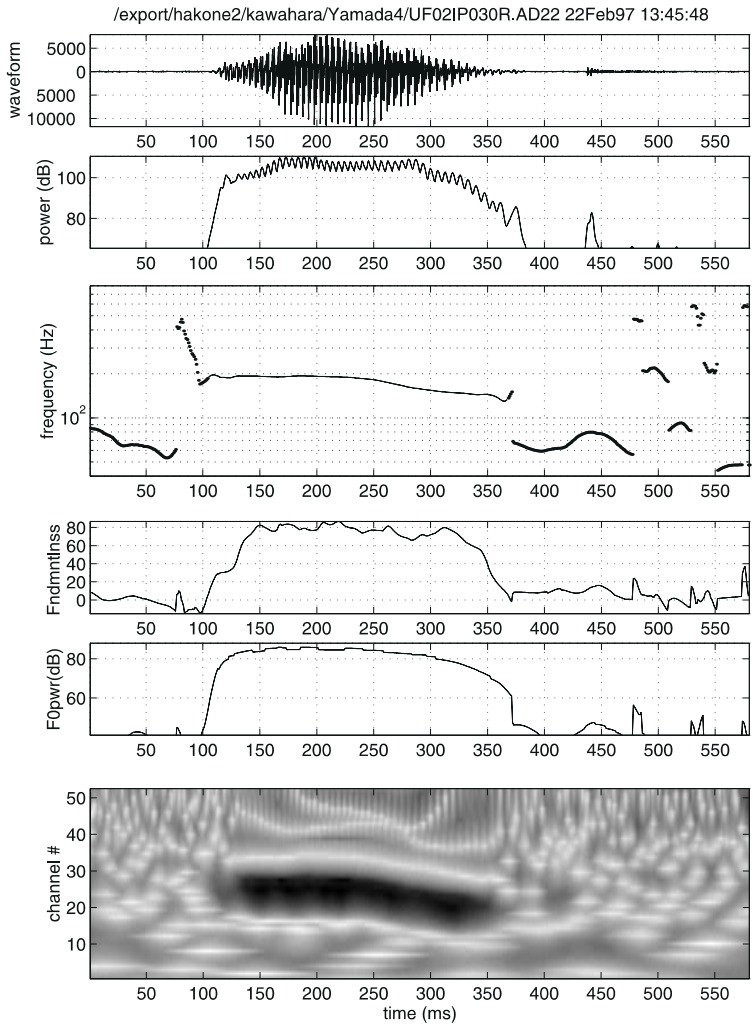

Fig. 17. Source information of female pronunciation of 'right' extracted by proposed instantaneous-based procedure.

trates 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 equally set to the fundamental period was also conducted for comparison. Fig. 18 shows a three-dimensional representation of the pitch-synchronous spectrogram. The figure illustrates the voiced portion. There are considerable spectral fluctuations in spectral valleys, while spectral peaks in the lower frequency region show smooth behavior.

The fluctuations in spectral valleys are artifacts due to sharp discontinuities at both edges of the rectangular time window. This can be concluded by comparison with Fig. 19. Fig. 19 shows a three-dimensional spectrogram using an isometric Gaussian time window, which is defined in Eq. (2) and is pitch-adaptive. Since 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. As such, the underlying spectral envelope is much smoother than the conventional pitch-synchronous analysis suggests.

Fig. 20 shows a spectrogram calculated by a reduced phasic interference procedure. The spectrogram also indicates that the strength of each harmonic component changes smoothly with time. Since the effective temporal resolution is about 1.4 times the isometric Gaussian window shown in Fig. 19, it is safe to conclude that the temporal change of the spectral envelope is actually smooth.

Fig. 21 shows the final spectral envelope after eliminating the temporal and frequency interfer-

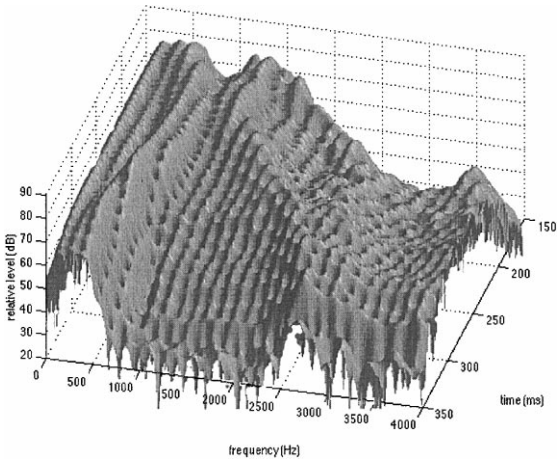

Fig. 19. Isometric Gaussian spectrogram.

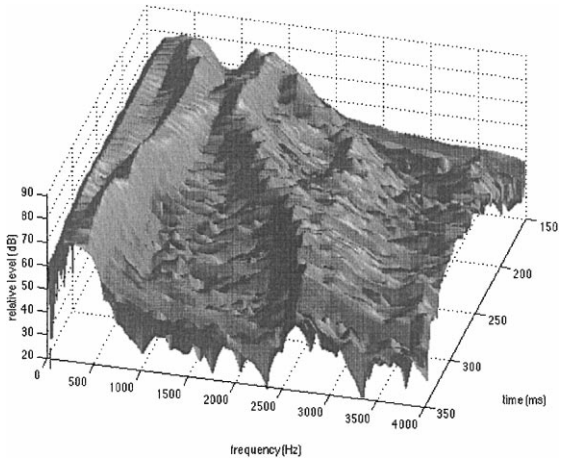

Fig. 18. Pitch-synchronous analysis of an utterance 'right' by a female subject. Only voiced portion is displayed.

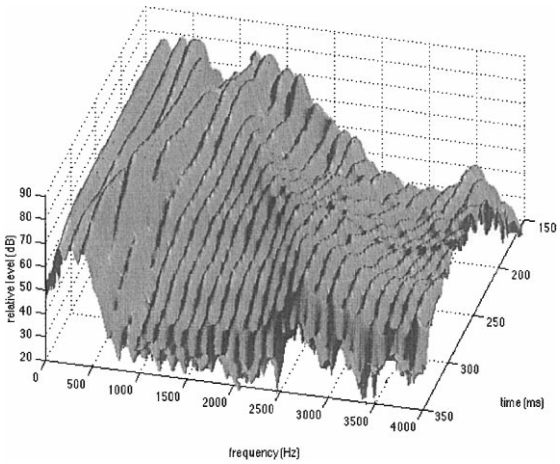

Fig. 20. Spectrogram with reduced phasic interference.

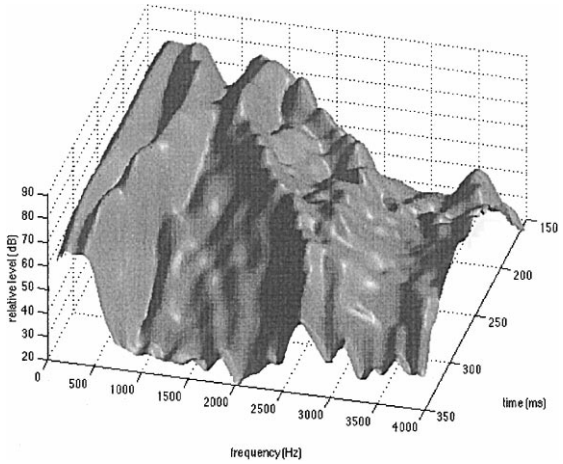

Fig. 21. Spectrogram without time-frequency interferences due to periodicity.

ences. Traces of harmonic components are effectively removed in the figure and the underlying spectral change, like formant transitions, is salient.

Other examples of pitch-synchronous analysis and pitch-adaptive smoothing for spectrum estimation are given in the Figs. 22–25. 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 two-dimensional spectral plots in Figs. 24 and 25. The displayed portion is the same as the three-dimensional plots.

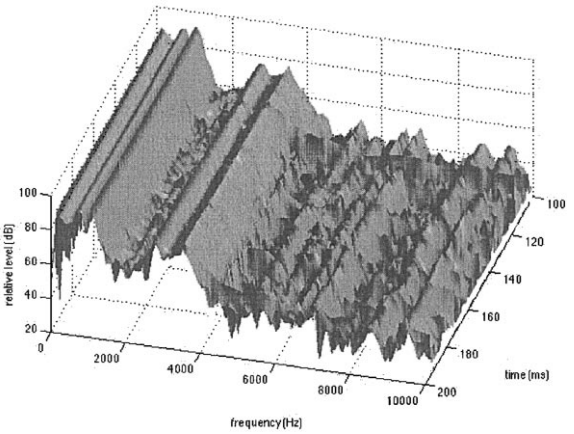

Fig. 22. Pitch-synchronous analysis.

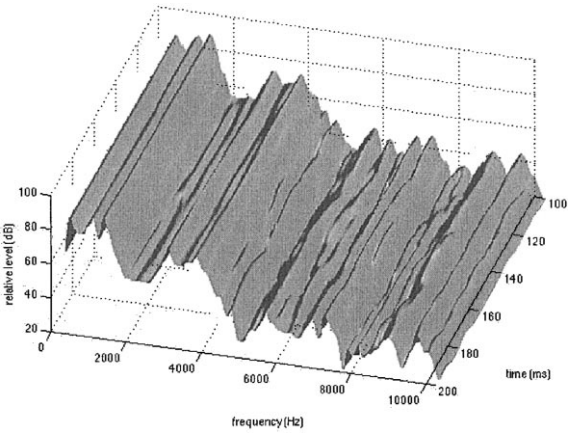

Fig. 23. Pitch-adaptive smoothing using cardinal B-spline.

5.3. Manipulations and re-synthesis

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, an inverse mapping function from the target temporal axis to the original temporal axis enables the transformation changing the speaking rate of the reproduced speech. A modification of the vocal tract length can be approximated by a linear conversion of frequency axis, because this is equivalent to modifying the wave propagation time from the glottis to the mouth. F0 modification is also trivial. These parameters are fed to the source generation part and the variable filter part in Fig. 16.

5.4. Re-synthesized speech quality

The original sound files and the manipulated files are located on the web site accompanying this journal. [5] Other manipulation 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 when listening carefully with headphones. With an unrealistic combination of ma-

[5] See http://www.elsevier.nl/locate/specom

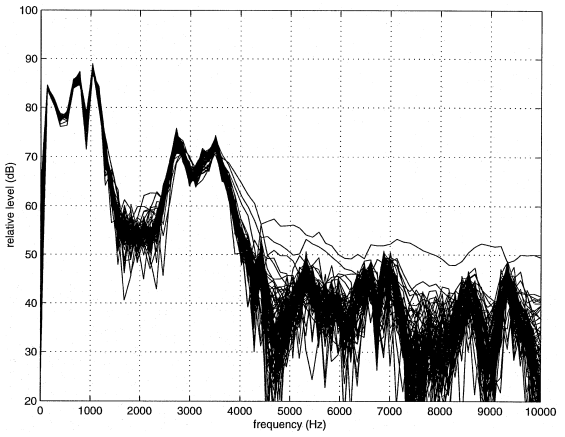

Fig. 24. Pitch-synchoronous spectra.

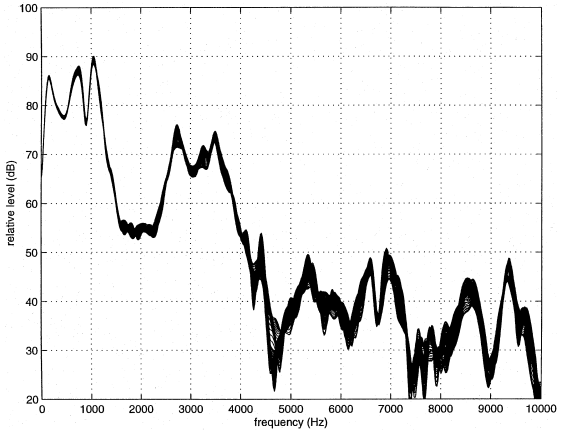

Fig. 25. Pitch-adaptive smoothed spectra using cardinal B-spline.

nipulation parameters such as a very short vocal tract and a low F0, the re-synthesized speech might sound strange, but is unlikely 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, the information rate is exceedingly high.

6. Discussion

Preliminary examination of several utterances showed that the spectrogram analyzed by the new procedure was surprisingly smooth. This indicates that there is a lot of room for information reduction. The proposed representation is a good starting point for investigating information reduction because the re-synthesized sound from the 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 since 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 (Slaney et al., 1996), electric musical instrument synthesis, and efficient reuse of existing sound resources.

We would like to point out that this work was initially motivated by the need for a flexible and high-quality real-time analysis–synthesis method for our on-going experiments on auditory feedback (Kawahara and Williams, 1996). 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 analysis and wavelet transform. The major difference between our method and prior methods is that ours involves information expansion, rather than reduction. This information expansion allows greater manipulative flexibility and stimulates an interesting speculation about the role of information expansion along mammalian auditory pathways (Webster et al., 1992).

Analysis-and-synthesis methods such as ours were previously believed to give poorer speech quality than waveform-based methods. However, the reproduced sounds by the proposed method seem to counter this belief. Our results suggest that the original concept of the VOCODER still holds, and that speech quality based on analysis-and-synthesis schemes can be improved further. Perhaps, the 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 with some statistical measures.

It should be noted that the proposed set of procedures is not yet an ideal method of sound coding. The most elaborate part of our method is effective only for voiced speech and similar signals. Currently, our method uses simple STFTs to estimate magnitude spectrum for unvoiced speech. More sophisticated signal models like Multi Band Excitation (MBE) (Griffin and Lim, 1988; Dutoit and Leich, 1993), multi-pulse (Caspers and Atal, 1987) and others (Abrantes et al., 1991) must be incorporated to appropriately represent wider range of sounds. However, even with these shortcomings at this level of implementation, re-synthesized speech using current system is almost equivalent to natural speech in 'naturalness'. It also inherits conceptual simplicity and greater 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 allows us to test perceptual contributions of various spectral/temporal modifications in the vicinity of very natural reference signals. In other words, it provides us with a means to test human speech perception mechanisms with ecologically valid stimuli. Preliminary tests with the proposed method have suggested that human auditory perception is highly specialized for detecting changes that affect the interpretation of auditory scenes (Kawahara et al., 1996). The new concept fundamentalness also provides an interesting interpretation of the pitch perception of inharmonic partials produced by AM (Schouten et al., 1962). Furthermore, it is interesting to observe that robustness of our method to natural fluctuations and F0 estimation errors are resulted from a combination of the minimum uncertainty Gaussian time window and a cardinal B-spline smoothing that is a kind of harmonic cancellation (de Cheveigné, 1998). We believe experimental results using ecologically valid stimuli, a new interpretation suggested by fundamentalness, and underlying principles of the proposed method will provide interesting hints for developing a computational theory of Auditory Scene Analysis.

7. Conclusion

New procedures that 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 parameters. These procedures are integrated to implement a sophisticated channel VOCODER system. We would like to call this set of procedures STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum).

The proposed method offers greater flexibility for parameter manipulations without introducing the artificial timbre specific to synthetic speech signals while maintaining a high reproductive 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'.

Acknowledgements

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

References

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

Abe, T., Kobayashi, T., Imai, S., 1996. Robust pitch estimation with harmonics enhancement in noisy environments based on instantaneous frequency. In: Proc. ICSLP 96, Philadelphia, pp. 1277–1280.

Abrantes, A.J., Marques, J.S., Trancoso, I.M., 1991. Hybrid sinusoidal modeling of speech without voicing decision. In: Proc. Eurospeech 91, Paris, pp. 231–234.

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

Blauert, J., Laws, P., 1978. Group delay distortion in electroacoustical systems. J. Acoust. Soc. Amer. 63 (5), 1478–1483.

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

Boashash, B., 1992b. Estimating and interpreting the instantaneous frequency of a signal – part 2: Algorithms and applications. Proc. IEEE 80 (4), 550–568.

Bregman, A.S., 1990. Auditory Scene Analysis. MIT Press, Cambridge, MA.

Caspers, B., Atal, B., 1987. Role of multi-pulse excitation in synthesis of natural-sounding voiced speech In: Proc. IEEE Internat. Conf. Acoust. Speech and Signal Processing Vol. 4, pp. 2388–2391.

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

Cooke, M.P., 1993. Modelling Auditory Processing and Organisation. Cambridge University Press, London.

de Cheveigné, A., 1996. Speech fundamental frequency estimation. Technical Report TR-H-195, ATR-HIP.

de Cheveigné, A., 1998. Cancellation model of pitch perception. J. Acoust. Soc. Amer. 103 (3), 1261–1271.

Dudley, H., 1939. Remaking speech. J. Acoust. Soc. Amer. 11 (2), 169–177.

Dutoit, T., Leich, H., 1993. An analysis of the performance of the MBE model when used in the context of a text-to-speech system. In: Proc. Eurospeech 93, Berlin, pp. 531–534.

El-Jaroudi, A., Makhoul, J., 1991. Discrete all-pole modeling. IEEE Trans. SP 39, 411–423.

Griffin, D.W., Lim, J.S., 1988. Multiband excitation vocoder. IEEE Trans. on Acoustics Speech and Signal Processing 36 (8), 1223–1235.

Itakura, F., Saito, S., 1970. A statistical method for estimation of speech spectral density and formant frequencies. Trans. IECE Japan, 53-A, 36–43 (in Japanese).

Kawahara, H., 1997. Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited. In: Proc. IEEE Internat. Conf. Acoust. Speech and Signal Processing 2, Münich, 1303–1306.

Kawahara, H., Masuda, I., 1996. Speech representation and transformation based on adaptive time-frequency interpolation. Technical Report of IEICE, EA96-28, pp. 9–16 (in Japanese).

Kawahara, H., Williams, J.C., 1996. Effects of auditory feedback on voice pitch. In: Davis, P.J., Fletcher, N.H. (Eds.), Vocal Fold Physiology. Singular, Münich, Chapter 18, pp. 263–278.

Kawahara, H., Tsuzaki, M., Patterson, Roy D., 1996. A method to shape a class of all-pass filters and their perceptual correlates. Tech. Com. Psycho. Physio. the Acoust. Soc. Jpn., H-96-79, 1–8 (in Japanese).

Marr, D., 1982. Vision: A Computational Investigation into Human Representation and Processing of Visual Information. Freeman, New York.

McAulay, R.J., Quatieri, T.F., 1986. Speech analysis/synthesis based on a sinusoidal representation. IEEE Trans. ASSP 34, 744–754.

Narendranath, M., Murthy, H.A., Rajendran, S., Yenanarayana, B., 1995. Transformation of formants for voice conversion using artificial neural networks. Speech Communication 16, 207–216.

Oppenheim, A., Schafer, R., 1989. Discrete-Time Signal Processing. Prentice-Hall, Englewood Cliffs, NJ.

Patterson, R.D., 1987. A pulse ribbon model of monaural phase perception. J. Acoust. Soc. Amer. 82 (5), 1560–1586.

Schouten, J.F., Ritsma, R.J., Cardozo, B.L., 1962. Pitch of the residue. J. Acoust. Soc. Amer. 34, 1418–1424.

Secrest, B.G., Doddington, G.R., 1983. An integrated pitch tracking algorithm for speech systems. In: Proc. IEEE ICASSP83, pp. 1352–1355.

Slaney, M., Covell, M., Lassiter, B., 1996. Automatic audio morphing. In: Proc. IEEE Internat. Conf. Acoust. Speech and Signal Processing, Atlanta, pp. 1–4.

Stylianou, Y., Laroche, J., Moulines, E., 1995. High-quality speech modification based on a harmonic + noise model. In: Proc. Eurospeech 95, Madrid, pp. 451–454.

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

Webster, D.B., Popper, A.N., Fay, R.R., 1992. The Mammalian Auditory Pathway: Neuroanatomy. Springer, Berlin, 1992.