Internal Use Only
非公開

TR - H - 200

0063

Speech transformation using adaptive interpolation of time-frequency representation and all-pass filters

Hideki Kawahara

1996. 8. 30

ATR人間情報通信研究所

〒619-02　京都府相楽郡精華町光台2-2　☎ 0774-95-1011

ATR Human Information Processing Research Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan
Telephone: +81-774-95-1011
Facsimile: +81-774-95-1008

Speech transformation using adaptive interpolation of time-frequency representation and all-pass filters

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

August 15, 1996, 12:04 A.M.

Abstract

A simple new procedure called STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) has been developed, using pitch-adaptive spectral analysis combined with a surface reconstruction method in the time-frequency region, and an excitation source design based on phase manipulation of all-pass filters. The proposed interpolation preserves the bilinear surface in the time-frequency region and allows for over 600% manipulation of such speech parameters as pitch, vocal tract length, and speaking rate, without introducing further degradation due to parameter manipulation. A new design procedure of all-pass filters also reduces the characteristic degradation caused by the usual pulse excitation which can be annoying, especially under headphone listening conditions.

PACS No. 43.72Ar, 43.72Lc, 43.60Lg, 43.70Jt

I. Introduction

The need for flexible speech modification methods is increasing in both commercial and scientific fields. Various sophisticated methods have been proposed (Veldhuis and He, 1996), but flexibility and the resultant speech quality are still limited, especially when large modification is necessary. If we consider old concepts in the context of the enormous progress in computational power in recent years, then a simple and appealing idea like the VOCODER (Dudley, 1939), which separates spectral and source information in order to manipulate and transmit speech sounds, is potentially very powerful.

The major problem in such separation was interactions between source characteristics and temporal fine structures like periodicity. If it were possible to eliminate source effects from spectral representations completely, it would enable independent manipulation of various speech parameters without degradations. A simple change in point of view on this problem seems to provide the answer.

II. Background

Voiced sounds are perceived as smoother than unvoiced sounds, unless the speech sounds are pathological. However, conventional speech analysis methods failed to represent this smoothness in an objective manner. For example, interactions between an analysis time window for short-term Fourier transform and the relative phase of the source excitation introduce magnitude variations in the resultant spectrum. Longer time windows, which span several cycles of the speech fundamental period, effectively circumvents this problem but introduces another interference in the form of harmonic structure due to periodicity.

Various sophisticated algorithms were proposed to estimate spectral envelope and to separate the source information and the system information (Itakura and Saito, 1970; Atal and Hanauer, 1971; El-Jaroudi and Makhoul, 1991). The conventional methods used to estimate spectral envelope for voiced speech rely on modeling the target spectrum. Linear prediction analysis provides a straightforward algorithm to estimate model parameters using normal equation, however the non-Gaussian nature of periodic sounds introduces non-negligible bias in estimated parameters. Other methods to circumvent these problems require iterative procedures to optimize specific cost functions which are not suitable for real-time applications.

Some algorithms tried to incorporate periodicity to make the estimation more precise (El-Jaroudi and Makhoul, 1991). Others tried to separate components due to periodicity. They demonstrated good performance in simulations using synthetic materials as well as representative

examples of real speech. However, in our follow-up experiments, applying these sophisticated methods to real speech did not always produce satisfactory results. These difficulties may be attributable to differences between the underlying model and the mechanism of real speech production processes as well as practical implementational issues. For example, homomorphic filtering provides an elegant means to deconvolve the source information and the transfer function, but special care must be taken with the time windowing function which can have serious effects on a frequency region where the signal energy is weak.

The procedure proposed here is designed to be applicable to real-time conversion of speech sounds which we need for ongoing research on interactions between speech perception and production using modified auditory feedback information (Kawahara and Williams, 1995). The proposed method consists of only feedforward procedures which heavily use FFT in implementation. This feedforward computation makes the proposed method suitable for real-time applications, when enough computational power is available.

A. Problem to be solved

To implement this speech manipulation procedure, it is necessary to solve the following three problems for the method to be flexible and have good speech quality.

- The first problem is to eliminate source effects from spectral representations.
- The second problem is to control the pitch of synthetic speech sounds with finer frequency resolution than that determined by the sampling frequency.
- The third problem is to design a set of excitation source waveforms which have completely flat spectra while having the desired temporal structures using multi pulses.

III. Principle of the proposed method

All these problems were solved using a conventional Fourier analysis. One major difference in our method is that it uses information expansion rather than reduction to achieve desired flexibility and high reproduction quality. In the current implementation, the amount of information to represent an exemplar speech sound is 200 times greater than that for storing the original waveform. The other important difference is that a systematic procedure to design the source signal using all-pass filters is proposed, which enables finer control on pitch and delicate timbre.

A. Elimination of periodicity effect by adaptive interpolation

The central idea of the proposed method is to consider the periodic excitation of voiced speech to be a sampling operation of a surface $S(\omega, t)$ in a three dimensional space defined by time, frequency, and amplitude axes. In this interpretation, a periodic signal $f(t) = f(t + n\tau_0)$ with the 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 only provide partial information about the surface.

Short-term Fourier analysis of this signal yields a time-frequency representation of the signal $F(\omega, t)$, know as spectrogram (Cohen, 1989). The spectrogram exhibits regular structure due to signal periodicity in both time and frequency. The uncertainty relation between frequency resolution and temporal resolution of the spectrogram representation requires a reasonable selection of the time window to extract the surface information, with isometric resolution in both frequency and time domains. The time window function $w(t)$, which fulfills this requirement and has the minimum uncertainty, has the following form.

$$w(t) = \sqrt[4]{2}e^{-\pi(t/\tau_0)^2} \tag{1}$$

In real speech, since the fundamental frequency varies with time, $\tau_0(t)$ is used in the following equations. This also indicates that the analysis window has to be modified adaptively to the speech fundamental frequency.

Our goal is to calculate a smoothed time-frequency representation $S(\omega, t)$, which has no effects caused by the periodicity of the signal based on the partial information given by the adaptive window analysis. It is a surface reconstruction problem based on partial information. It is neccessary to provide constraints for the problem to have a solution. One reasonable constraint is to use local information only. A constraint, which requires to use nearest three points to determine a surface, has a trivial answer, a triangular planer surface. This answer is not interesting, because it does not provide a smooth surface. The next possible constraint, which requires to use four nearest points has an interesting solution. The most simple form which fulfills this requirement is the bilinear equation given below. In other words, the surface is represented as a patchwork which consists of piecewise bilinear surfaces.

$$S_p(\omega, t) = (a_p\omega - b_p)(c_p t - d_p) \tag{2}$$

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

It is possible to calculate this piecewise bilinear representation using the data on sampling points. For example, using scaling and translation in both time and frequency, it is possible to map a time-frequency region onto a region defined by $(0,0),(1,0),(0,1),(1,1)$. Then the bilinear surface is represented explicitly using the sampled value on each corner. Let $s_{00}, s_{10}, s_{01}, s_{11}$ be the sampled values on the corresponding vertices. Then the value on the bilinear surface $S_p(\lambda,\tau)$ is represented by the following equation.

$$S_p(\lambda,\tau) = s_{00} + (s_{01} - s_{00})\tau + (s_{10} - s_{00})\lambda + (s_{00} - s_{01} - s_{10} + s_{11})\lambda\tau \tag{3}$$

However, this procedure is numerically fragile for real speech signals because real speech signals are not precisely periodic and consist of natural fluctuations. Instead, we propose to use an interpolation function which provides equivalent piecewise bilinear representation when the sampled data in the time-frequency representation are given only on the grid points spaced τ_0 in time domain and f_0 in frequency domain.

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

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

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

$$\begin{aligned} h_t(\lambda,\tau) &= \frac{1}{4}(1-|\lambda/\omega_0(t)|)(1-|\tau/\tau_0(t)|) \\ &\quad \text{where } \omega_0(t) = 2\pi f_0(t), \text{ and} \\ &\quad [-\omega_0(t) \le \lambda \le \omega_0(t),\ -\tau_0(t) \le \tau \le \tau_0(t)] \end{aligned} \tag{5}$$

In Equation 4 $g(\ )$ defines what measure to preserve through the interpolation. For example, identity mapping, $g(x) = x$, preserves the energy of the signal and the $1/3$ power low, $g(x) = x^{1/3}$, preserves the perceived loudness, approximately.

If the original surface has a bilinear portion on it, the proposed interpolation operation preserves the surface intact. If the smoothing function, which is calculated from the time window function to delive the spectrogram, is a localized even function, and if the bilinear portion is reasonably wide, then the most of bilinear surface is also preserved. In other words, the effects of source periodicity are perfectly removed.

B. Minimum phase impulse response and fine pitch control

The transformed speech $s(t)$ is generated using the following equation.

$$\begin{aligned} s(t) &= \sum_{t_i \in Q} \frac{1}{\sqrt{G(f_0(t_i))}} v_{ti}(t - T(t_i)) \\ &= \frac{1}{\sqrt{2\pi}} \sum_{t_i \in Q} \frac{1}{\sqrt{G(f_0(t_i))}} \int_{-\infty}^{\infty} V(\omega, t_i)\Phi(\omega) e^{j\omega(t-T(t_i))} d\omega \qquad (6) \\ &\text{where } T(t_i) = \sum_{t_k \in Q, k<i} \frac{1}{G(f_0(t_k))} \end{aligned}$$

where Q represents a set of positions of excitation for synthesis, $G(\,)$ represents pitch modification and $j = \sqrt{-1}$. The all-pass filter function $\Phi(\omega)$ is used to control fine pitch and the temporal structure of the source signal and is described in the next section. In the discrete-time system, range of integration in the equation to derive $v_t(\tau)$ becomes $[-\pi, \pi]$ using normalized angular frequency $\omega = 2\pi f / f_s$, where f_s represents 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 $M(S(u(\omega), r(t)), u(\omega), r(t))$, where $M(\,)$, $u(\,)$ and $r(\,)$ represent manipulations in amplitude, frequency, and time axes respectively.

$$\begin{aligned} V(\omega, t) &= \exp\left(\frac{1}{\sqrt{2\pi}} \int_0^{\infty} h_t(q) e^{j\omega q} dq\right) \qquad (7) \\ h_t(q) &= \begin{cases} 0 & (q < 0) \\ c_t(0) & (q = 0) \\ 2c_t(q) & (q > 0) \end{cases} \\ \text{and} \quad c_t(q) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-j\omega q} \log M(S(u(\omega), r(t)), u(\omega), r(t)) d\omega \end{aligned}$$

where q represents quefrency.

1. Fine pitch control by phase manipulation

In resynthesizing speech from its modified amplitude spectrogram, a minimum phase impulse response was calculated using the complex cepstrum. This method inevitably consists of an inverse Fourier transformation stage and makes it sensible to use phase manipulation to add further control of source characteristics. One such function is a finer control of fundamental frequency. Since linear rotation of phase with frequency corresponds to a shift in time domain, it can be

used to control fractional pitch. Equation 6 embodies this fine control. The discrete-time system only allows sampling points to be located on discrete positions, the event position calculated by $T(t_i)$ is represented as the sum of the discrete location n/f_s and the fractional part T_f. Fine pitch control is implemented as a phase shift $\Phi_1(\omega)$ defined below.

$$\Phi_1(\omega) = e^{-j\omega f_s T_f} \tag{8}$$

2. Consideration of unvoiced sounds

Care must be taken for non-periodic sounds like fricatives and plosives. It is necessary to use a fixed-window analysis to represent non-periodic signals. It is reasonable to refer to human performance in a gap-detection task of wide-band noise (Formby and Muir, 1988) to decide this fixed window length. In the current implementation, a fixed time window corresponding to 400Hz fundamental frequency was selected. A Gaussian random noise was used as the excitation source for the non-periodic sounds.

C. Excitation source design by phase manipulation

Phase manipulation also provides the means to control detailed timbre of the synthetic sound. Even though there are no differences between different all-pass filters in amplitude response, which is considered to be the major contributing factor of timbre, a different all-pass filter adds a different timbre to the resynthesized speech, especially under headphone listening conditions. This observation is not contradictory to reports on phase effects on timbre found in the literature (Plomp and Steeneken, 1969; Blauert and Laws, 1978; Patterson, 1987). Simultaneous masking experiments using synthetic speech revealed that there is a phase-dependent threshold variation which sometimes approaches 20 dB in peak-to-peak differences (Kawahara, 1986). This effect may be responsible for phase effects on timbre because this effect provides a possible mechanism for the faint signals to contribute to total timbre. Faint signals, which are normally inaudible under simultaneous masking by a continuous speech-like noise, can be audible if they are located in low threshold regions assosiated with a periodic signal like a synthetic speech. However, the phase effects on timbre are not well understood and need systematic investigation.

1. Requirements

All-pass filters which meets the following requirements are suitable for speech synthesis and musical instruments.

(1) Energy should be localized in time.

(2) Temporal spread of energy in time and in frequency should be controlled to meet a specific purpose.

(3) Temporal asymmetry in energy distribution should be implemented if necessary.

A design method using random numbers to generate all-pass filters is also desirable.

2. All-pass filter design

The following representation of all-pass filters fulfills the first and the second requirements. However, there still exists some coupling between temporal parameters and frequency parameters which makes the control difficult.

$$\Phi_2(\omega) = \exp\left(-j\rho(\omega)\sum_{k\in P}\alpha_k \sin(k\omega)\right) \tag{9}$$

where P represents a set of integer indices. This equation can be considered an implementation of all-pass filters using multi-pulses. The temporal spread is determined by $\{\alpha_k\}$, a temporal weighting function. Relative temporal spread in different frequency region is determined by $\rho(\omega)$, a frequency weighting function. For simplicity, it is better to use the following normalization constraints.

$$\begin{aligned} 1 &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\rho(\omega)d\omega \\ \rho(\omega) &\geq 0 \end{aligned} \tag{10}$$

Because the Equation 9 is a non-linear summation of sinusoids, the corresponding impulse response, inverse Fourier transform of the equation, has many cross terms. Then it is necessary to investigate relations between parameters in Equation 9 and the spread in the time domain. Let $P = k, \alpha_k = \alpha, \rho(\omega) = 1$ for the simplest case. Then its time domain representation $\phi_2(t)$, which corresponds to the frequency representation $\Phi_2(\omega)$, yields the following equation.

$$\begin{aligned} \phi_2(t) &= \frac{1}{\sqrt{2\pi}}\int_{-\pi}^{\pi} e^{-j\alpha \sin k\omega} e^{j\omega t} d\omega \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\pi}^{\pi}(1 + (-j\alpha \sin k\omega) + \frac{1}{2!}(-j\alpha \sin k\omega)^2 + \frac{1}{3!}(-j\alpha \sin k\omega)^3 \\ &\quad + \ldots + \frac{1}{m!}(-j\alpha \sin k\omega)^m + \ldots)e^{j\omega t} d\omega \end{aligned} \tag{11}$$

This expansion shows that even though the number of elements of P is one, there are an infinite numbers of pulses spaced every k sampling points. But the signal $\phi_2(t)$ is effectively localized in time because the height of each pulse decreases very rapidly.

Assume ε to represent the value of effective zero, then the limit for the coefficient α is represented in terms of k, ε and specification for the spread in time Δt.

$$\begin{aligned} \alpha &\leq \left(\varepsilon\Gamma\left(\frac{\Delta t}{k}\right)\right)^{\frac{k}{\Delta t}} \\ \Delta t &\geq k \end{aligned} \tag{12}$$

where $\Gamma()$ represents the gamma function and the spread Δt is represented in terms of number of samples. One reasonable selection of ε is to use the value of LSB (least significant bit) of A/D and D/A conversion.

The next step is to check the relation between the weighting function $\rho(\omega)$ and the spread in time and frequency. The group delay of Equation 9 yields:

$$\begin{aligned} d(\omega) &\triangleq j\frac{d\log\Phi(\omega)}{d\omega} \\ &= \rho(\omega)\sum_k \alpha_k k\cos(k\omega) + \frac{d\rho(\omega)}{d\omega}\sum_k \alpha_k \sin(k\omega) \\ &= \sum_k \alpha_k A_k(\omega)\cos(k\omega - \varphi_k(\omega)) \end{aligned} \tag{13}$$

where $A_k(\omega)$ and $\varphi_k(\omega)$ are given by the following equations.

$$A_k(\omega) = \sqrt{k^2\rho^2(\omega) + \left(\frac{d\rho(\omega)}{d\omega}\right)^2} \tag{14}$$

$$\varphi_k(\omega) = \arctan\left(\frac{\frac{d\rho(\omega)}{d\omega}}{k\rho(\omega)}\right) \tag{15}$$

These equations indicate that $\rho(\omega)$ represents the shape of group delay closely, if $\rho(\omega)$ is reasonably smooth.

To implement the third requirement, the following equation introduces asymmetry in time.

$$\Phi_3(\omega) = \exp\left(-j\int_{-\pi}^{\pi}\beta\left(j\frac{d\log\Phi_2(\lambda)}{d\lambda}\right)d\lambda\right) \tag{16}$$

where $\beta(\ )$ represents a smooth even function monotonic in both positive and negative directions. One such example is given here.

$$\beta_1(\dot{x}) = \exp(-|x|) + |x| - 1 \tag{17}$$

This is a variation inspired by ROEX (ROunded EXponential) function proposed by Patterson (Patterson et al., 1982).

Random numbers can be used to generate a set of α_k, taking the boundary given by Equation 12 into account. Let us call this all-pass filter $\Phi_4(\omega)$.

3. Use of all-pass filters

All-pass filters $\Phi_1(\omega),\ldots,\Phi_4(\omega)$, designed by the procedure mentioned above can be used together because the product of the all-pass filters is an all-pass filter. Each all-pass filter has relevant role: $\Phi_1(\omega)$ is used for fine control of fundamental frequency, $\Phi_2(\omega)$ can be used to reduce annoying pulsive timbre caused by impulse excitation, $\Phi_3(\omega)$ can simulate room acoustics by diffuse phase. Generating $\Phi_4(\omega)$ for every event allows us to simulate degradation of periodicity in a specific frequency region. The final feature is effective in implementing voiced fricative sounds and gradual transitions from voiced to unvoiced sounds.

IV. Implementation

This section describes implementation issues which were not covered in the previous sections. The proposed method is currently implemented as a set of functions on the MATLAB system. Built-in user interface functions and advanced scientific visualization and sonification functions of MATLAB obviated the need for the development of dedicated interface functions for the proposed method. The implementation allows users to have full access to the internal variables to take advantage of these MATLAB functions. It also made it possible to use the system on different machines without modification of the source codes. Procedures are designed to be compatible with various sampling frequencies (8 kHz to 48 kHz) used in our laboratories.

A. Extraction of source information

The proposed procedure relies heavily on pitch extraction and voiced and unvoiced distinction. However, precision for fundamental frequency extraction is not necessarily high for analyzing data because the interpolation-based formulation has made the procedure insensitive to small errors in window size. Higher precision is necessary in the speech production phase.

In the current implementation, the lag window method proposed by Sagayama (Sagayama and Furui, 1978) is modified to meet the requirements of the proposed method. The pitch extraction method is based on smoothing and normalization of autocorrelation function and is suitable for our FFT based procedure. Low pass filtering and parabolic curve fitting around the τ_0 peak is introduced to implement finer frequency resolution. A voiced and unvoiced distinction procedure is implemented using the maximum normalized autocorrelation at τ_0, segmental power, and zero-crossing.

It is also possible to use commercial pitch extraction systems for the pitch extraction stage of this method because users have full access to MATLAB functions and internal variables.

B. Example

A series of speech transformation experiments were conducted using utterances spoken by two female and one male native speakers of American English. Th words ''right'' and ''light'' were recorded using an omnidirectional microphone (SONY ECM-955) and a DAT recorder (SONY DTC-10) at 48 kHz sampling rate and 16 bit resolution. These recorded data were played back and A/D converted at 22050 Hz 16 bit using an A/D-D/A interface (PAVEC MD-8000).

Figure 1 shows the waveform and F0 of ''light'' spoken by a female speaker. F0 contour is displayed only for the voiced portion. The following analyses used this F0 data to adjust the time window size adaptively.

Figure 2 illustrates the spectrogram calculated using the isometric window. The periodicity of the speech signal is shown as regularly distributed white dots on the spectrogram. This regular structure interferes with visual inspection of formant structure.

Figure 3 shows the interpolated spectrogram based on the proposed method. The regular white dot interference caused by signal periodicity is completely removed while the essential details of the spectrogram are retained. Formant structure is clearly isolated visually.

Four different three-dimensional (3D) plots have been prepared to illustrate the effects of the time window and interpolation. A region defined by $220 < t < 320$ (ms) and $0 < f < 2000$ (Hz) is selected for this purpose. This region corresponds to the transition from /a/ to /i/ and consists of the first and the second formants (F_1 and F_2 respectively).

Figure 4 shows a wide-band spectrogram, where the length of the time window is $1/\sqrt{2}$ of the isometric window length. The periodicity of the speech signal introduces periodic temporal variation of the surface.

Figure 5 shows a narrow-band spectrogram, where the frequency resolution associated with the time window is $1/\sqrt{2}$ of the isometric window. The periodicity of the speech signal introduces periodic variation of the surface along the frequency axis.

Figure 6 shows the spectrogram calculated using the isometric window. The periodicity of the speech signal is shown as the grid-like structure of the surface.

Figure 7 shows the interpolated spectrogram based on the proposed method. The grid-like interference caused by signal periodicity is completely removed, while details of the surface shape are retained.

This interpolated spectrogram is used to resynthesize speech. In this example, all-pass filters shown in Figure 8 were used. The all-pass filter shown in plot (a) is generated using $P = 1, \alpha_1 = 2\text{ms}, \rho(\omega) = 1$ using Equation 9. The all-pass filter shown in plot (b) is generated using $P = 13, \alpha_1 = 4\text{ms}, \rho(\omega) = 1$. The all-pass filter shown in plot (c) is generated in the following way. A set $\{\alpha_k\}$ is weighted by $k\exp(-k^2/k_0^2)$ and generated using random numbers.

The number of elements in P was 40 and $\rho(\omega) = c_0\omega^2$, where c_0 represents the normalization factor. The all-pass filter shown in plot (c) is an example, since this filter is generated using random numbers, different each time. The all-pass filter shown in the last plot is the convolution of all these all-pass filters and is still an all-pass filter.

Figure 9 illustrates the original speech waveform and the resynthesized speech waveform. They preserve some similarity, but the fine structures are different because the phase information in the original speech is abandoned.

1. Sound files

These original sound files and the manipulated files are located under the following URL.
httpd://www.hip.atr.co.jp/~kawahara/aiff/

Informal listening tests demonstrated that the resynthesized speech was sometimes indistinguishable from the original when heard over speakers. However, there was still perceptible degradation while listening carefully under headphones, but further degradation caused by parameter manipulation was negligible, even with modifications of up to 600%. Also, the manipulated speech still sounded ''natural'' after 200% to 600% modification of speech parameters, even though it is necessary to re-define ''naturalness'' for these large manipulations. For example, 500% manipulation simultaneously of pitch and frequency converted male speech into a naturally sounding cat's mew.

V. Possible applications

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

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

The interpolated spectrogram is ideal for visual inspection of speech features like formant structures because interference caused by the source periodicity is removed perfectly. It may also be usable for pre-processing of format extraction algorithms.

VI. Further problems

The proposed method is categorized as an analysis-and-synthesis method which was believed to have poorer speech quality while having greater flexibility than waveform-based methods. The reproduced sound by the proposed method seems to provide a counter example to this understanding. It suggests that the original concept of VOCODER still holds, and that the speech quality based on the analysis-and-synthesis scheme can be improved further. This implies that the precise reproduction of source signals is not necessary for high quality speech reproduction. Rather, there can be equivalent classes in which corresponding source signals have the identical timbre while having different waveforms. It is practically as well as theoretically important to characterize these equivalent classes in terms of some statistical measures.

VII. Conclusion

A new method to represent and manipulate speech signals based on an adaptive time window and spectrogram interpolation is introduced. The method yields very high flexibility in parameter manipulation without further degradation while keeping high reproduction quality. This may help promote research to investigate relations between physical parameters and perceptual correlates.

Acknowledgement

The author would like to express sincere appreciation to colleagues at ATR, Dr. Roy Patterson of MRC Cambridge, and Dr. Alain de Cheveigné of CNRS for discussions. Also he wishes to express special thanks to his collaborator, J. C. Williams, for discussions and encouragement.

References

Atal, B. S. and Hanauer, S. L. (1971). "Speech analysis and synthesis by linear prediction of speech wave," J. Acoust. Soc. Am. **50**, 637--655.

Blauert, J. and Laws, P. (1978). "Group delay distortion in electroacoustical systems," J. Acoust. Soc. Am. **63**, 1478--1483.

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

Dudley, H. (1939). "Remaking speech," J. Acoust. Soc. Am. **11**, 169--177.

El-Jaroudi, A. and Makhoul, J. (1991). "Discrete all-pole modeling," IEEE Trans. **SP-39**, 411--423.

Formby, C. and Muir, K. (1988). "Modulation and gap detection for broadband filtered noise signals," J. Acoust. Soc. Am. **84**, 545--550.

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

Kawahara, H. (1986), "On detectability of noise bursts in speech," Technical Report EA86-39, (IEICE Japan, Tokyo). [in Japanese].

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

Oppenheim, A. and 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. Am. **82**, 1560--1586.

Patterson, R. D., Nimmo-Smith, I., Weber, D. L., and Milroy, R. (1982). "The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram and speech threshold," J. Acoust. Soc. Am. **72**, 1788--1803.

Plomp, R. and Steeneken, H. J. (1969). "Effects of phase on the timbre of complex tones," J. Acoust. Soc. Am. **46**, 409--421.

Sagayama, S. and Furui, S. (1978). "Pitch extraction using the lag window method," Proc. IECE Japan **1235-5**, 263. [in Japanese].

Veldhuis, R. and 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.

Figures

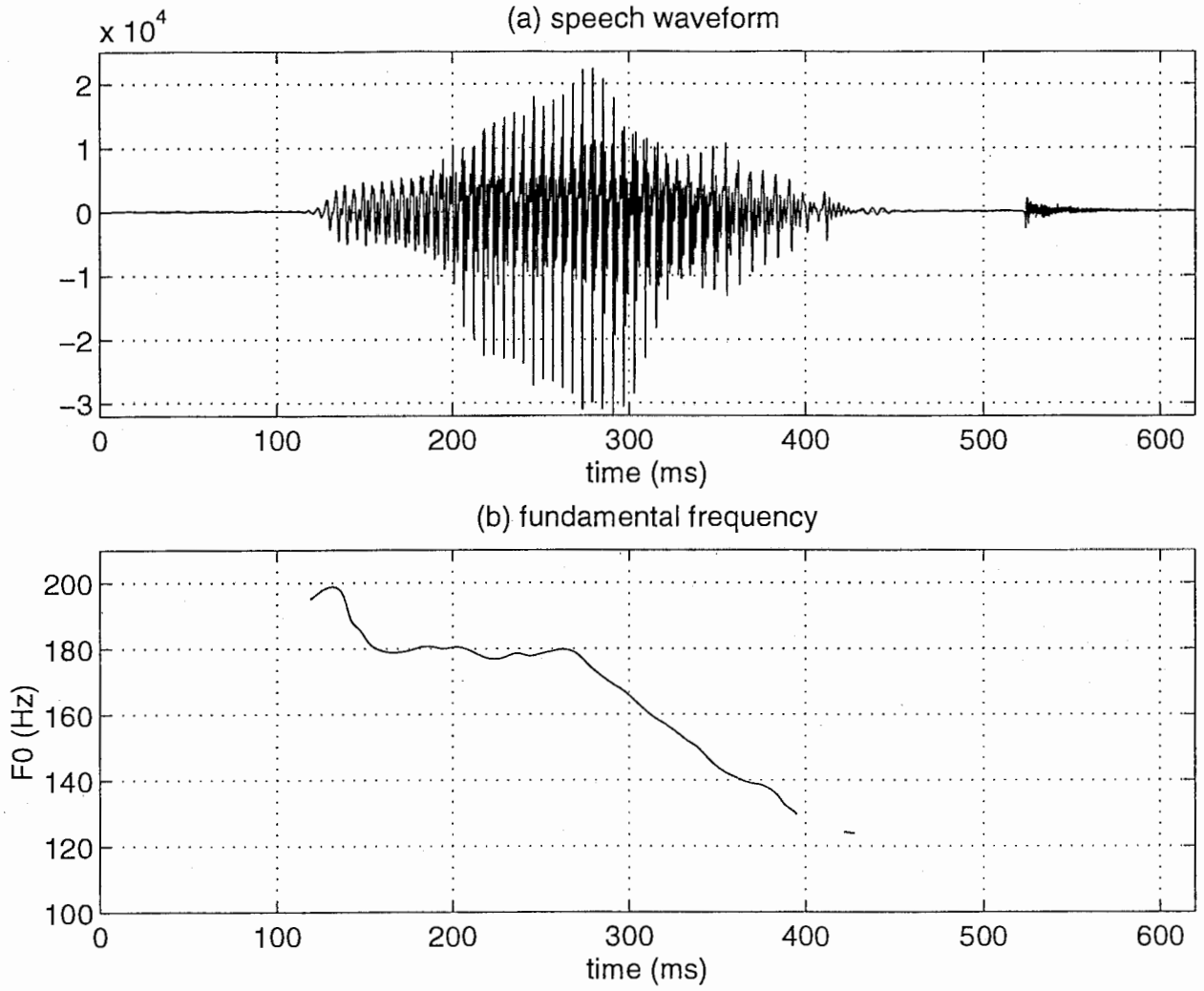

FIG. 1. Example data. (a) original speech waveform and (b) extracted fundamental frequency F0.

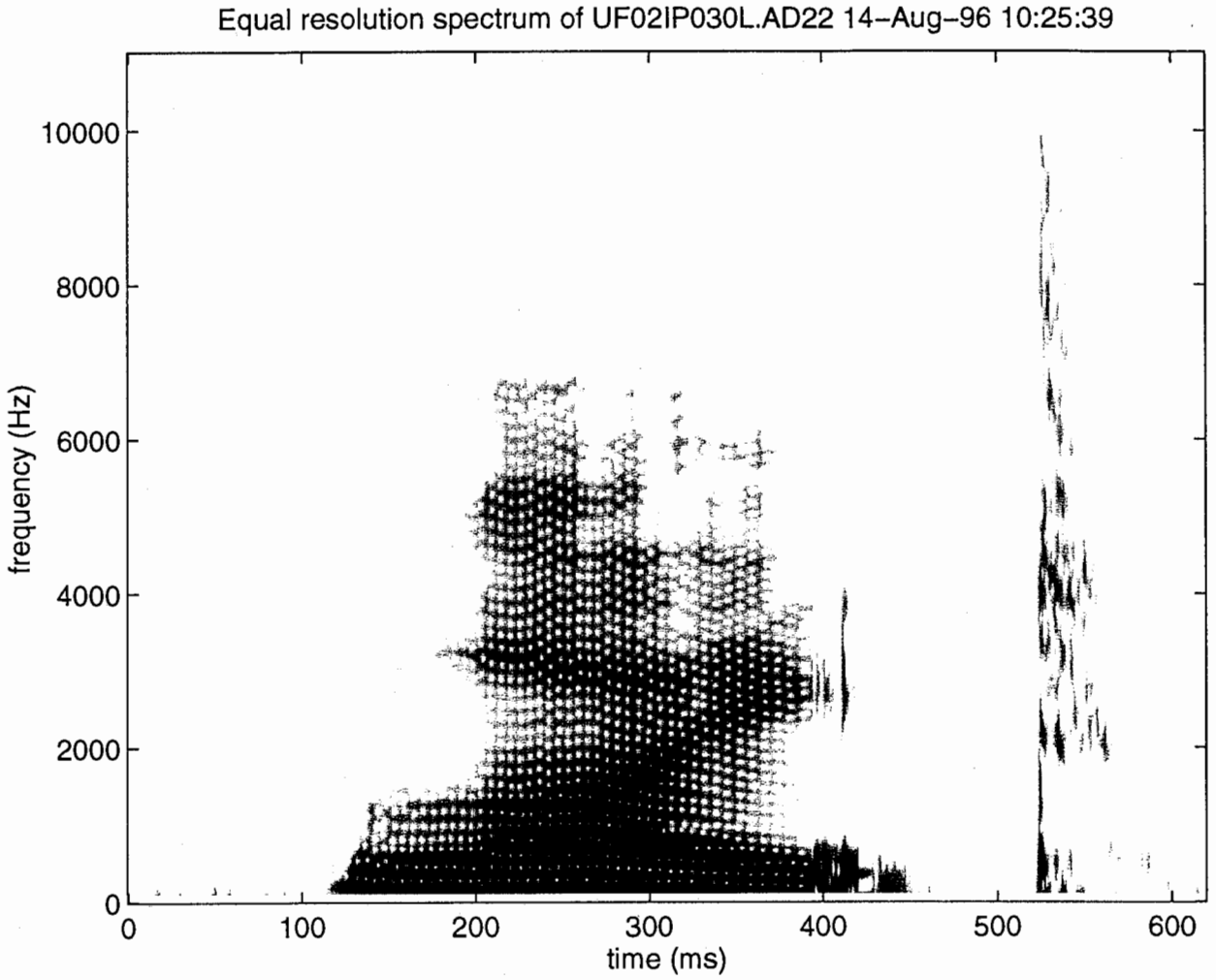

FIG. 2. Original equal resolution spectrogram.

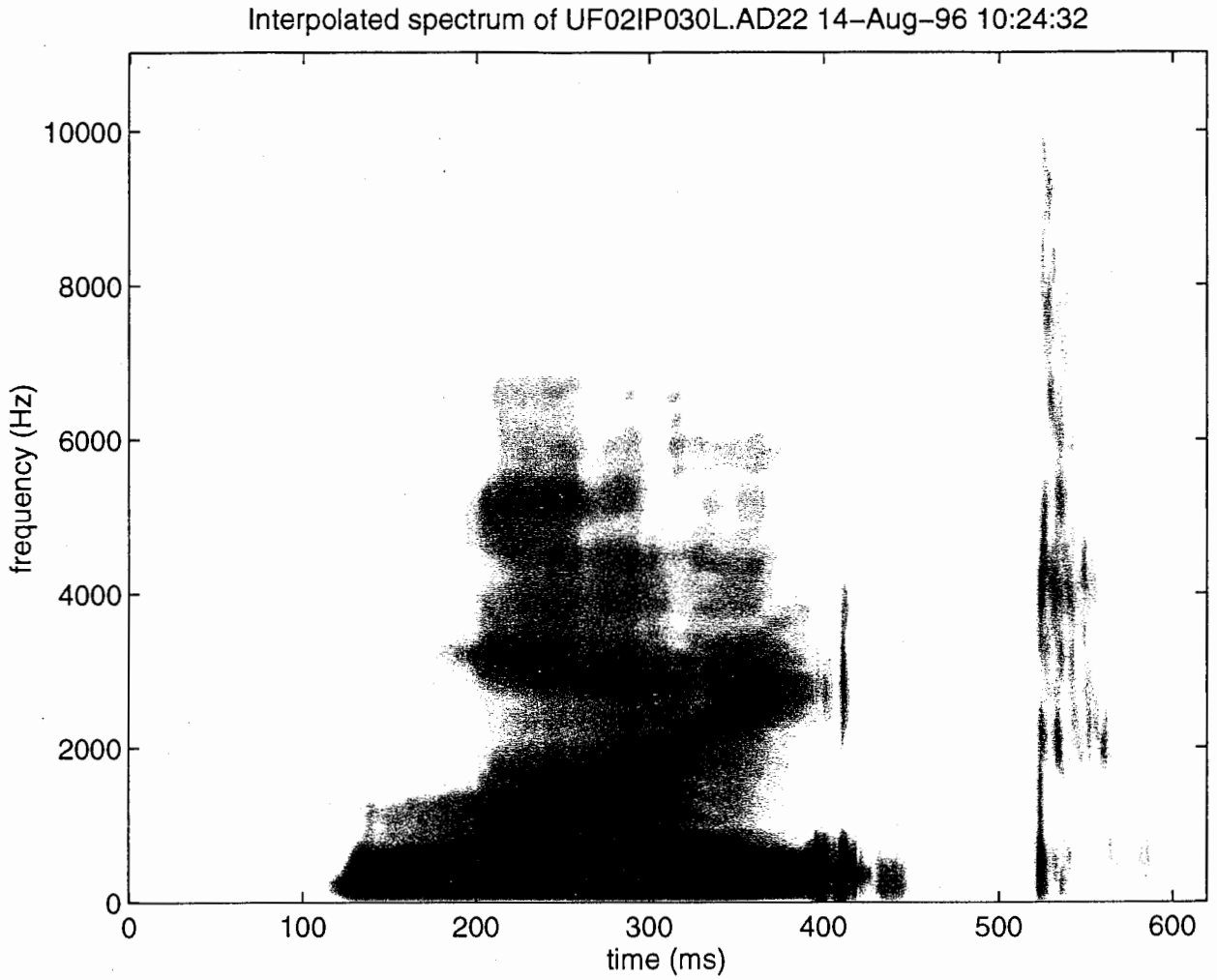

FIG. 3. Interpolated spectrogram.

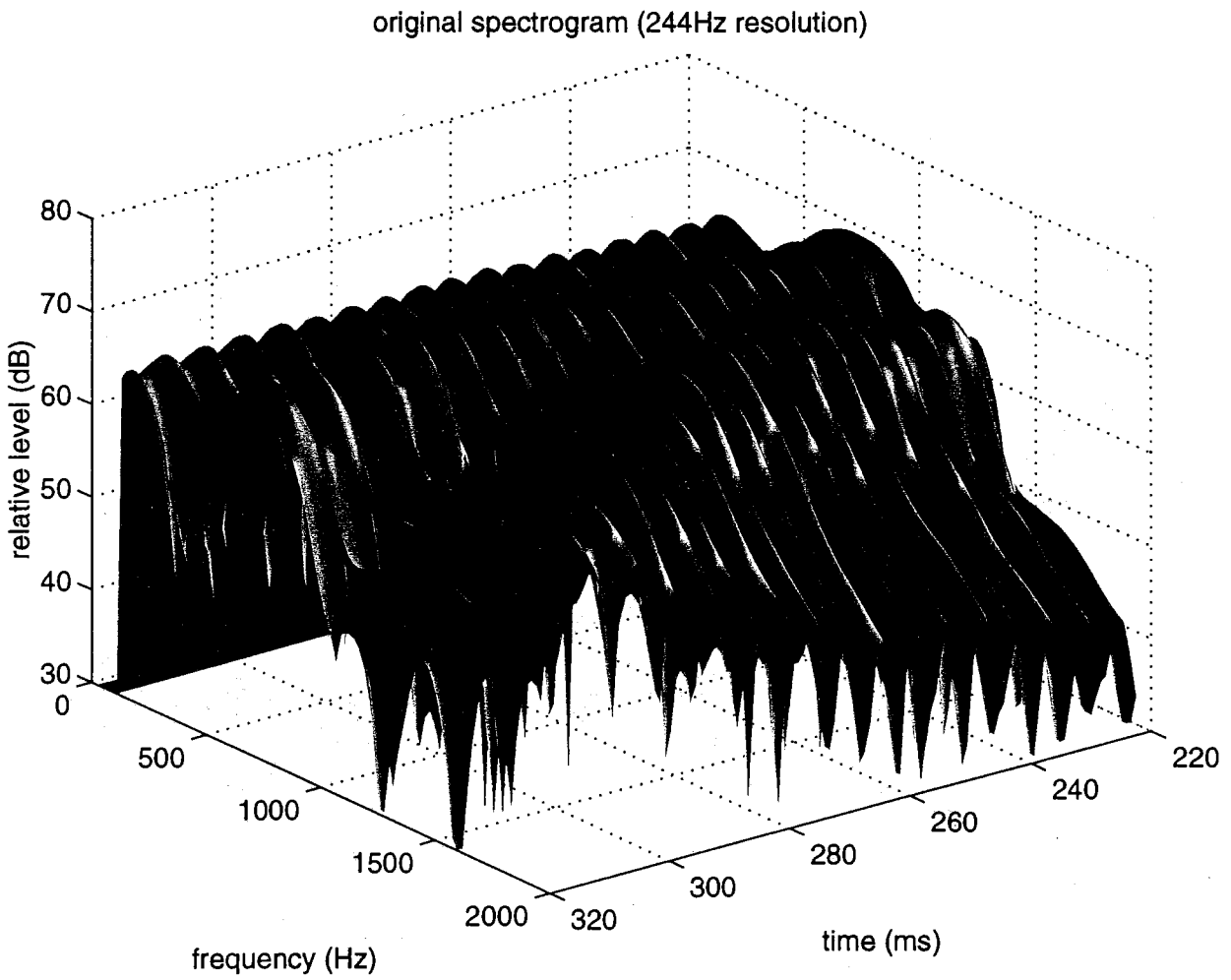

FIG. 4. 3D display of wide-band spectrogram.

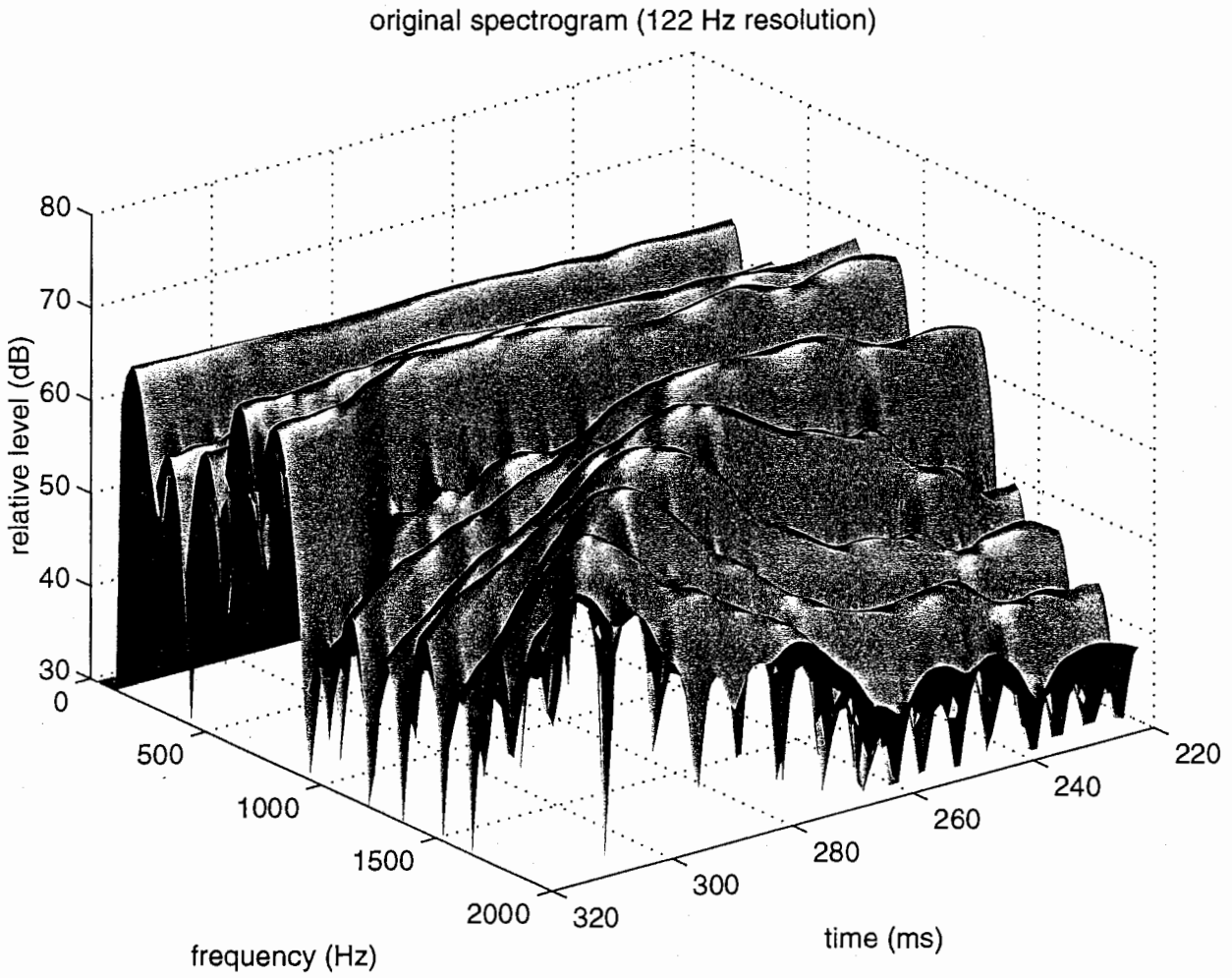

FIG. 5. 3D display of narrow-band spectrogram.

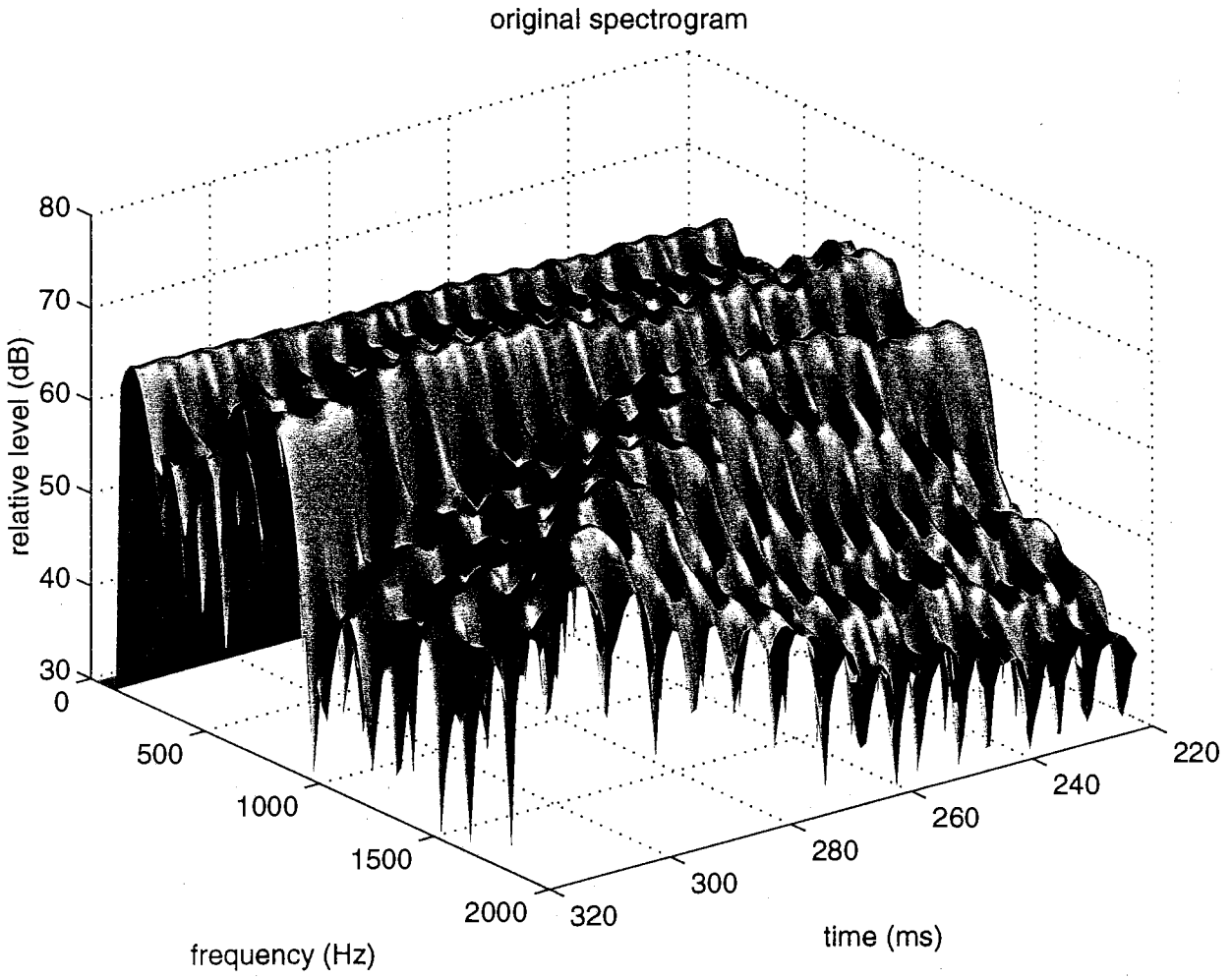

FIG. 6. 3D display of isometric resolution spectrogram.

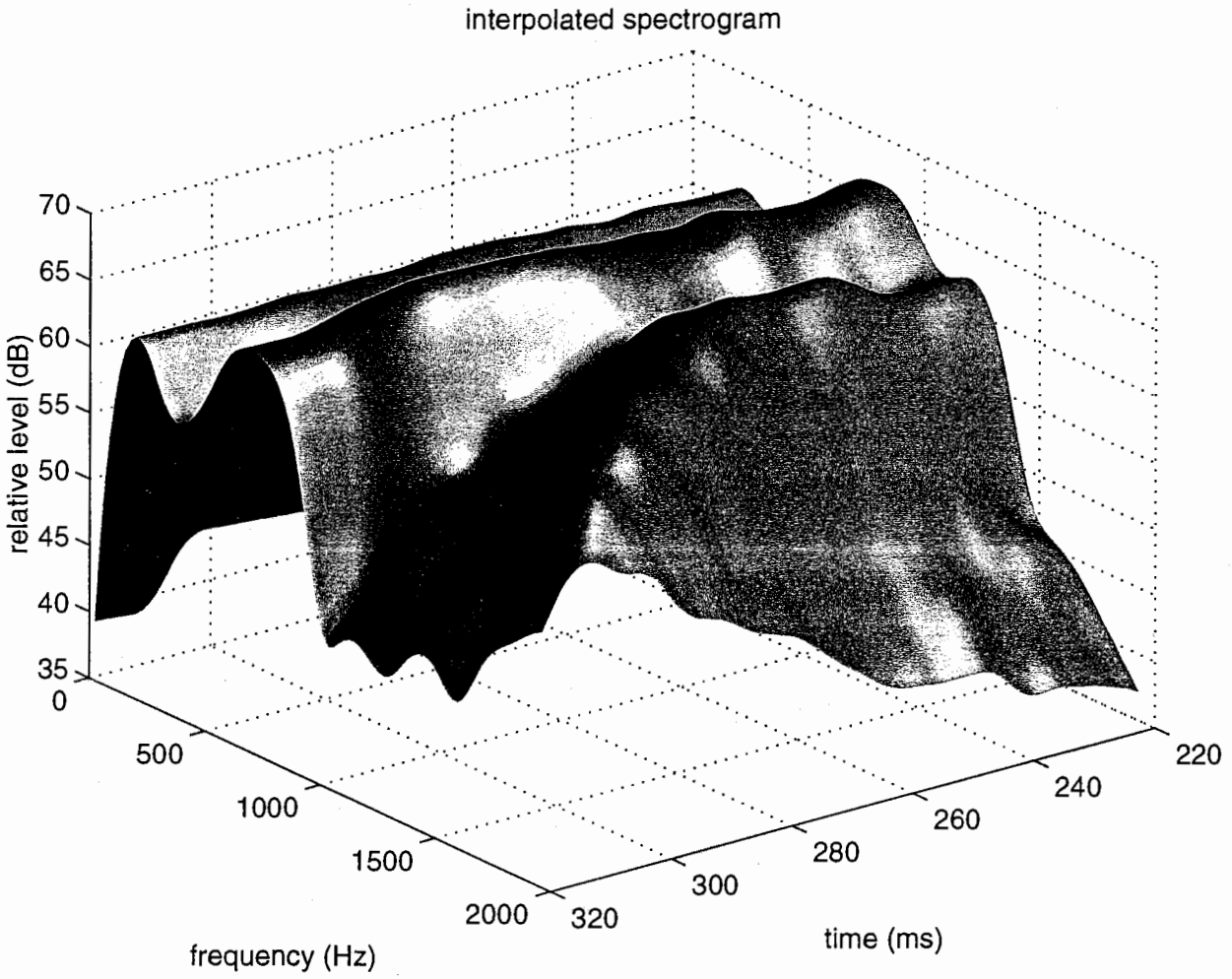

FIG. 7. 3D display of isometric resolution spectrogram.

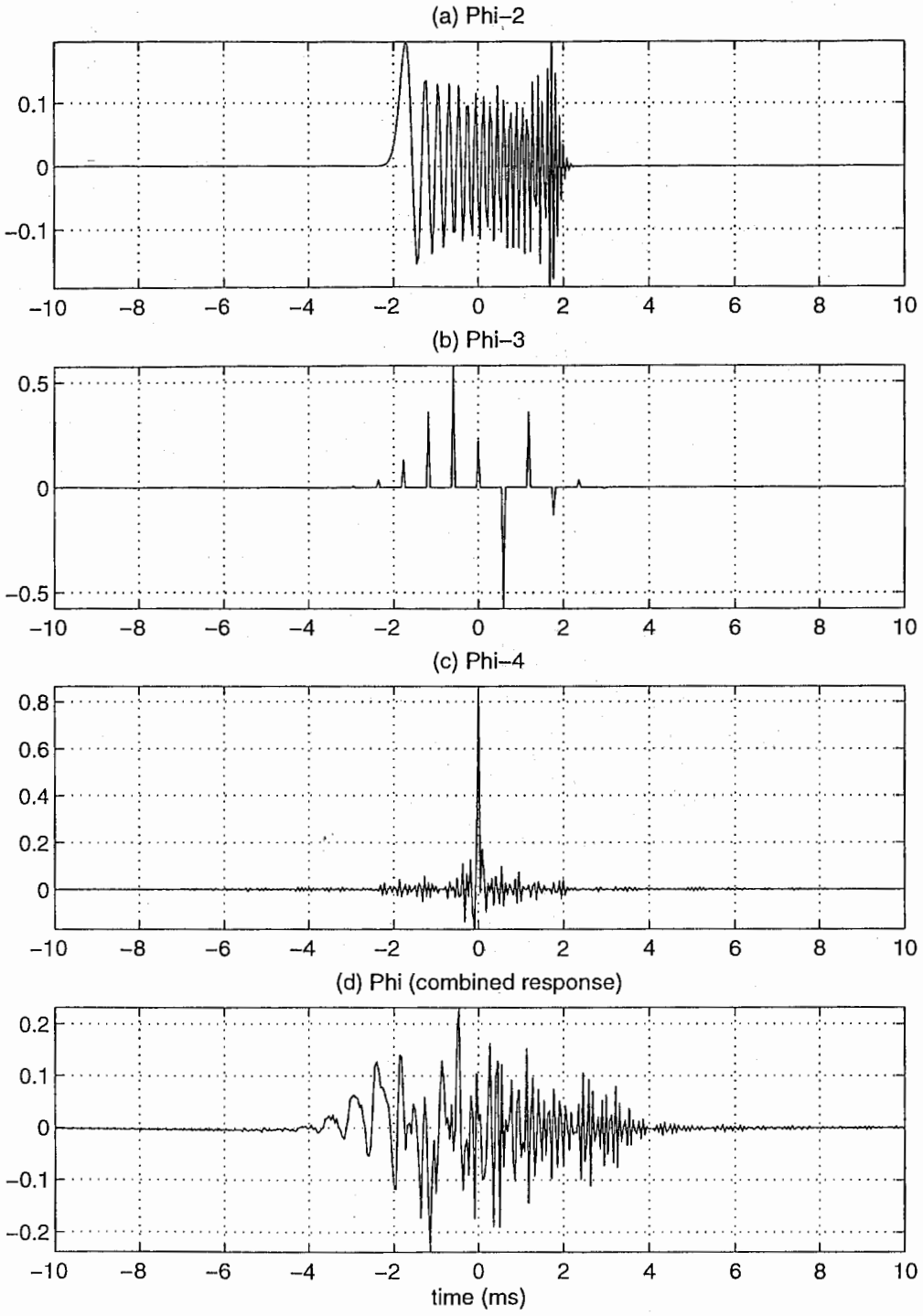

FIG. 8. All-pass filter impulse responses used for re-synthesizing speech.

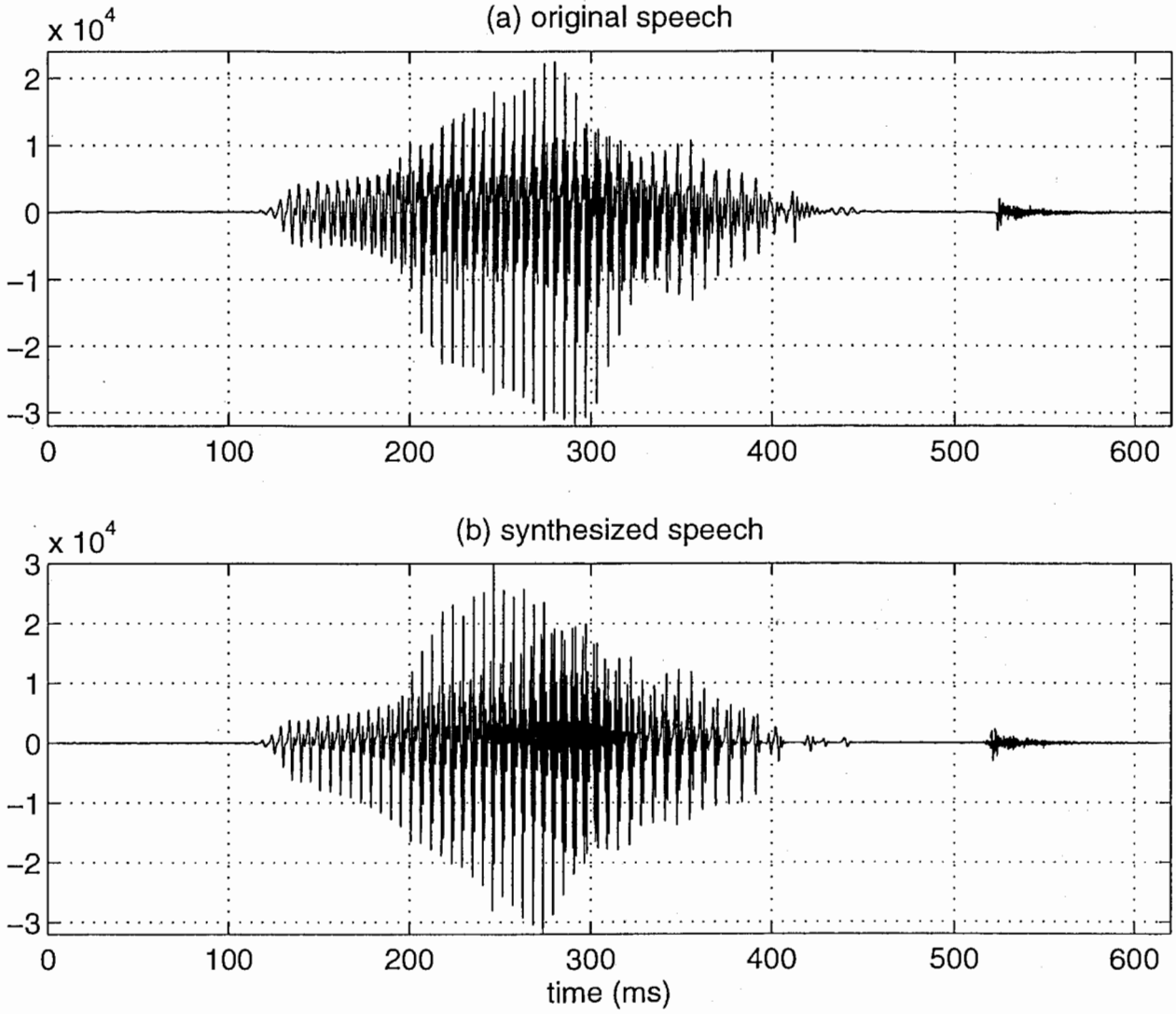

FIG. 9. The original speech waveform and the synthesized speech waveform.