Speech Segregation Using an Auditory Vocoder With Event-Synchronous Enhancements

Toshio Irino, *Senior Member, IEEE*, Roy D. Patterson, and Hideki Kawahara, *Senior Member, IEEE*

***Abstract*—We propose a new method to segregate concurrent speech sounds using an auditory version of a channel vocoder. The auditory representation of sound, referred to as an "auditory image," preserves fine temporal information, unlike conventional window-based processing systems. This makes it possible to segregate speech sources with an event synchronous procedure. Fundamental frequency information is used to estimate the sequence of glottal pulse times for a target speaker, and to repress the glottal events of other speakers. The procedure leads to robust extraction of the target speech and effective segregation even when the signal-to-noise ratio is as low as 0 dB. Moreover, the segregation performance remains high when the speech contains jitter, or when the estimate of the fundamental frequency F0 is inaccurate. This contrasts with conventional comb-filter methods where errors in F0 estimation produce a marked reduction in performance. We compared the new method to a comb-filter method using a cross-correlation measure and perceptual recognition experiments. The results suggest that the new method has the potential to supplant comb-filter and harmonic-selection methods for speech enhancement.**

***Index Terms*—Auditory image, auditory scene analysis, channel vocoder, comb filter, pitch/F0 extraction.**

I. Introduction

HUMAN listeners can segregate speakers in a multispeaker environment, and the mechanisms whereby humans perform speech segregation are important for automatic speech recognition. It is a major topic in the field of computational auditory scene analysis (CASA) (e.g., [1]). Typically, CASA models proposed for solving segregation problems are directly or indirectly based on models of auditory processing [2]–[5]. Nevertheless, most speech segregation systems still use nonauditory forms of spectral analysis like the short-time Fourier transform (STFT) and its relatives. There are several reasons for this: the STFT involves far less computation than an auditory model; performance is not necessarily improved by the use of an auditory model when it is restricted to frame-wise processing with window averaging; the STFT and sinusoid models have advantages when using harmonic enhancement methods (including comb filtering and harmonic selection [6], [7]), since the harmonics appear regularly spaced on a linear frequency axis. However, systems that focus on extracting harmonics at strict integer multiples of a fundamental frequency, F0, have serious limitations, because F0 estimation is imperfect when the signal-to-noise ration (SNR) between the target speech and background noise is low, *and* the error increases in proportion to harmonic number.

From the production point of view, a voiced speech sound is a stream of glottal pulses, each of which is followed by a complex resonance with information about the shape of the vocal tract at the time of the pulse. It has been argued that the auditory system extracts the pitch of the sound from neural spike times using a process analogous to auto-correlation) [8]–[10], and the "correlogram" has been proposed as a basis for speaker segregation, and for the reconstruction of the target speech [3], [11]. However, auto-correlation involves frame-wise processing which smears temporal fine structure, including both the pulse-resonance structure *and* aspects of temporal asymmetry that we hear [12]–[16].

We propose a new method for speaker segregation and target-speaker resynthesis based on the auditory image model (AIM) [17], [18]. AIM was developed to provide a representation of the "auditory images" we perceive in response to everyday sounds [19], [20]. The auditory representation preserves fine temporal information, unlike conventional window-based processing, and this makes it possible to do synchronous speech segregation. We describe a procedure for extracting the glottal events of a target speaker using F0 information, which are then used, first to enhance the auditory image of the target speaker, and then to resynthesize the target speech. The system is a form of channel vocoder [21], [22] that will be referred to as an "auditory vocoder." We also illustrate the robustness of the system in situations where competing sources cause errors in the fundamental frequency estimation.

II. Auditory Vocoder

The system has three components (Fig. 1): the AIM [17], [18], a robust F0 estimator [e.g., [22], [23]], and a synthesis module based on the channel vocoder concept. In Section II-A, we describe an event-synchronous version of AIM and show how it can be used for speaker separation. In Section II-B, we describe

Manuscript received January 31, 2005; revised August 27, 2005. This work was supported in part by a project grant from the Faculty of Systems Engineering, Wakayama University, in part by the Japan Society for the Promotion of Science, under Grant-in-Aid for Scientific Research (B), 15300061, and in part by the U.K. Medical Research Council under Grant G9900369. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Judith Brown.

T. Irino and H. Kawahara are with the Faculty of Systems Engineering, Wakayama University, Wakayama 640-8510, Japan (e-mail: irino@sys.wakayama-u.ac.jp; kawahara@sys.wakayama-u.ac.jp).

R. D. Patterson is with the Centre for Neural Basis of Hearing, Department of Physiology, University of Cambridge, Cambridge CB2 3EG, U.K. (e-mail: roy.patterson@mrc-cbu.cam.ac.uk).

Digital Object Identifier 10.1109/TASL.2006.872611

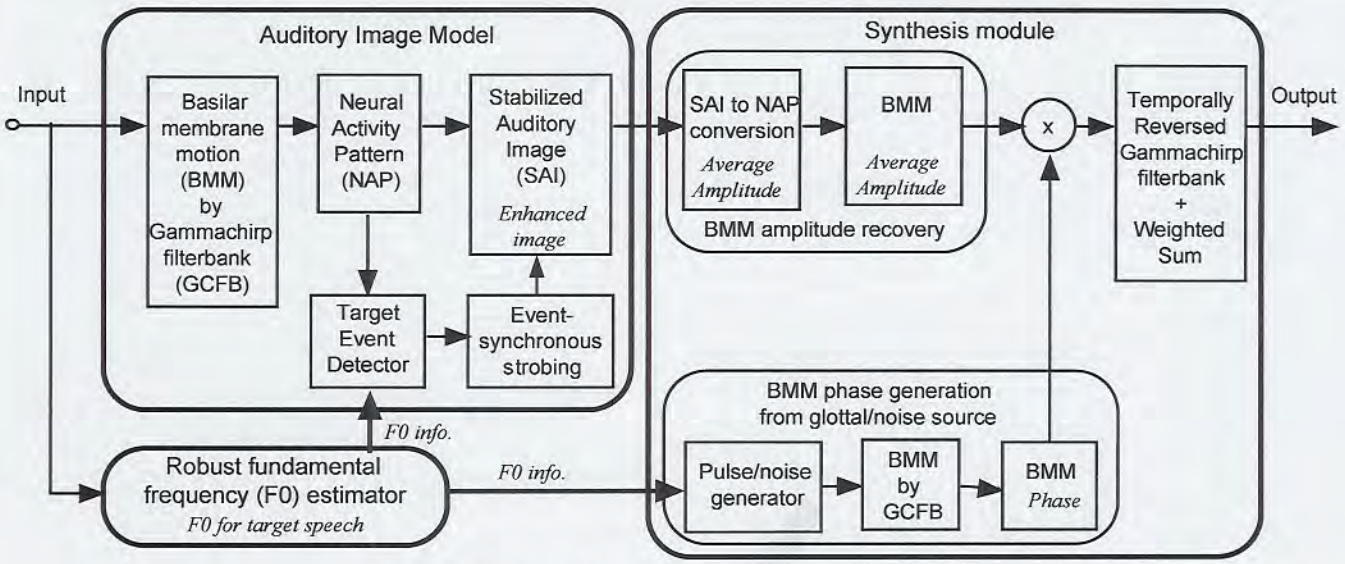

Fig. 1. Configuration of the auditory vocoder for speech segregation.

a method for signal resynthesis from the auditory image (AI) using a channel vocoder.

A. Event-Based Auditory Image Model

We briefly summarize the conventional stabilized auditory image [17]–[20], [24] and describe the new concept of "event-synchronous strobing" and the resultant enhancement.

1) Stabilized Auditory Image: AIM performs its spectral analysis with a gammatone, or gammachirp, filterbank [17], [25]–[27] on a quasilogarithmic (ERB) frequency scale [28]. The impulse response of the gammachirp auditory filter is

$$g_c(t) = at^{n-1}\exp(-2\pi b\mathrm{ERB}(f_r)t) \times \exp(j2\pi f_r t + jc\ln t + j\phi) \quad (1)$$

where a, b, c, and n are parameter values, ϕ is phase, f_r is the asymptotic frequency of the filter, and $\mathrm{ERB}(f_r)$ denotes the equivalent rectangular bandwidth at frequency, f_r. The gammachirp filter with peak (or center) frequency, f_c, is denoted $g_c(f_c, t)$ because f_c is different from the asymptotic frequency, f_r when $c \neq 0$. When $c = 0$, (1) is a gammatone filter. In the following experiment, we used a linear gammatone filter. The output of the filterbank, $S_B(f_c, t)$, is AIM's simulation of basilar membrane motion (BMM). If the signal is $s(t)$, then $S_B(f_c, t)$ is

$$S_B(f_c, t) = \int_0^\infty g_c(f_c, \tau_1)s(t-\tau_1)d\tau_1. \quad (2)$$

In the cochlea, the mechanical vibration is dynamically compressed and the inner hair cells convert the motion into neural pulse trains. In the model, the output of (2) is half-wave rectified and compressed to convert it into a simulated neural activity pattern (NAP), $S_N(f_c, t)$. So, the NAP has detailed temporal information about the decay of formant resonances within glottal periods. Note that we do not apply compression in the current auditory vocoder because informal listening indicated that the sound quality was better without it.

We assume that the brainstem performs a time-interval analysis of the temporal structure; in AIM, the process is simulated by a form of "strobed temporal integration" (STI). It is this processing that converts the neural pattern into an AI. Briefly, the activity in each neural channel is monitored to identify local maxima in the neural activity, and these local maxima are used to control temporal integration. The process is controlled by the derivative of the envelope of the activity. The derivative of the envelope is decomposed into the convolution of the input signal and the derivative of the gamma function in the gammachirp. So, it is referred to as "delta gamma" [15], [16]. The local maxima occur regularly when the signal is periodic or quasiperiodic, as in the voiced parts of speech. Temporal integration is strobed on each of the local maxima and consists of taking the current segment of the NAP (about 35 ms in duration) and adding it into the corresponding channel of the AI with whatever is currently in that channel, point by point. Mathematically, if the signal $s(t - kt_p)$ is strictly periodic with period t_p, and strobing is ideal, then the SAI has the form

$$A_I(f_c, \tau) = \sum_{k=0}^{\infty} S_N(f_c, \tau + kt_p)e^{-\xi\tau}e^{-\eta kt_p} \quad (3)$$

where f_c is the peak-frequency of one auditory filter, τ is the time-interval axis of the SAI, and k is an integer. The exponential scalars, ξ and η, delimit the maximum time-interval and the persistence of the AI. So, each channel of the SAI is the decaying sum of a sequence of segments of neural activity that have been shifted by one pitch period. The process is similar to measuring time intervals from the largest NAP pulses to smaller NAP pulses that are nearby in time, and making a dynamic interval-histogram of the results, that is, an interval histogram with a continuous decay appropriate to the decay of auditory perception. The half life of the auditory image is estimated to be about 30 ms [17]–[20].

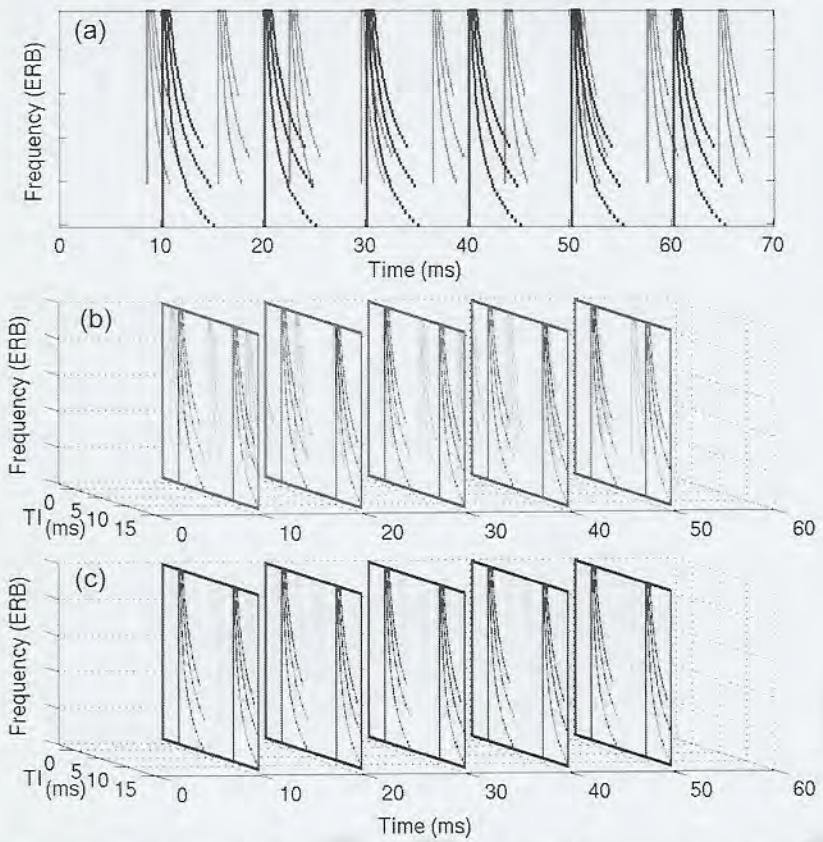

Fig. 2. Schematic plot for event-synchronous strobing and the enhancement of a target speaker in concurrent speech. (a) Compensated NAP for a segment of target speech (black lines; 10-ms pitch period) with concurrent vowels (gray lines; 7-ms pitch period). (b) Time sequence of two-dimensional auditory images (AIs) generated from the NAP by event-synchronous strobing every 10 ms. (c) SAI after temporal integration across successive frames.

STI converts the time dimension of the NAP [e.g., Fig. 2(a)] into the time-interval dimension of the AI [e.g., Fig. 2(b)]. At the same time, it removes the propagation lag associated with auditory filtering; it aligns the peaks of the responses to the glottal pulse across frequency and assigns peaks from the most recent glottal pulse to the origin of the time-interval dimension. STI is applied separately to each channel of the filterbank output, and the SAI is the complete array of stabilized neural patterns. The combination of strobing and a 30-ms half-life stabilizes the vowel pattern for as long as the sound is stationary. In this way, AIM constructs an auditory image of the signal automatically based on its internal structure.

2) Event Detection: To effect segregation of concurrent speakers, we identify the stream of strobe pulses from the target speaker and use them to enhance the AI for the target speech. To this end, we introduce an event-detection mechanism in the NAP processing to locate the glottal pulses of the target speaker. These events are used as strobe signals to convert the NAP into a SAI. Fig. 3(a) shows about 30 ms of the NAP of a male vowel. The abscissa is time in milliseconds; the ordinate is the center frequency of the gammatone auditory filter in kilohertz. The mechanism identifies the time interval associated with the repetition of the neural pattern, and the interval is the fundamental period of the speech at that point. Since the NAP is produced by convolution of the speech signal with the auditory filter, and the group delay of the auditory filter varies with center frequency, it is necessary to compensate for group delay across channels when estimating event timing. Fig. 3(b) shows the NAP after group delay compensation; the operation aligns the responses of the filters in time to each glottal pulse. The solid line in Fig. 3(c) shows the temporal profile derived by summing across channels in the compensated NAP in the region below 1.5 kHz. The peaks corresponding to glottal pulses are readily apparent in this representation.

We extracted peaks locally to reduce spurious responses due to noise in the NAP. The algorithm is similar to pitch period estimation using a threshold function with an exponential decay [29]. If the amplitude is $x(t)$, then the threshold value $x_T(t)$ is

$$x_T(t) = \max\left\{x(t-\Delta t)\cdot e^{-\ln 2\cdot\left(\frac{\Delta t}{T_h}\right)}, x(t)\right\} \qquad (4)$$

where Δt is the sampling time, and T_h is the half-life of the exponential decay. The dashed line is the threshold used to identify peaks, which are indicated by circles on vertical lines in Fig. 3(c). After a peak is found, the threshold decreases gradually to locate the next peak. We also introduced a form of prediction to make the peak detection robust. The threshold is reduced by a certain ratio when the detector does not find activity at the expected period, defined as the median of recent periods. This is indicated by the sudden drop in threshold toward the end of each cycle. The time of the event is defined as the time of the local peak of $x_T(t)$. Tests with synthetic sounds confirmed that this algorithm works sufficiently well when the input is clean

Please make this figure a little bit smaller,

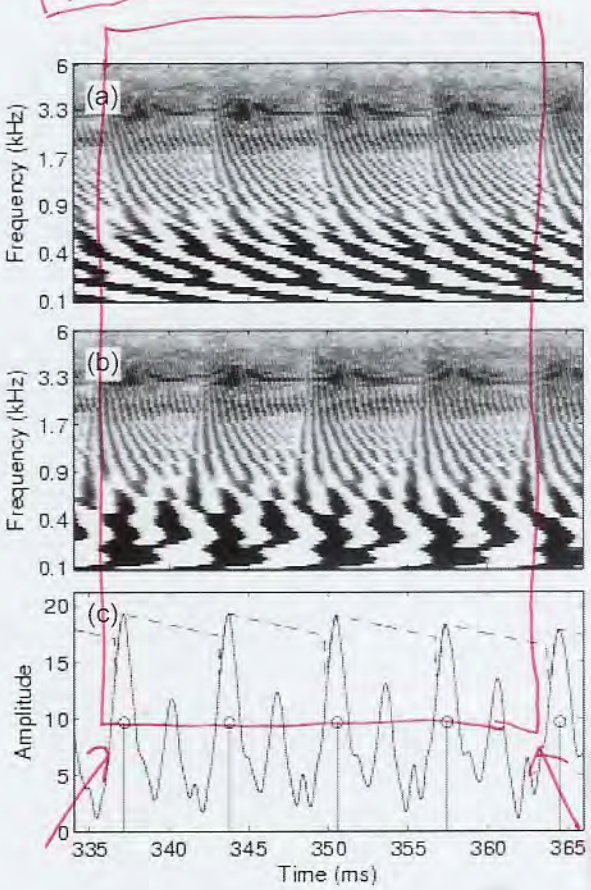

Fig. 3. Auditory event detection for clean speech. (a) Neural activity pattern (NAP) of a male vowel. (b) NAP after group delay compensation. (c) Temporal profile of the compensated NAP in the region below 1.5 kHz (solid line). The dashed line is a threshold used to identify peaks; the circles and vertical lines shows the estimated event times.

speech. It is, however, difficult to apply this method under noisy conditions particularly when the SNR is low. This is why we enlisted F0 information to improve event–time estimation.

3) Robust F0 Estimator for Event Detection: It is easier to estimate fundamental frequency F0 than event times for a given SNR. The latest methods for F0 estimation are robust in low SNR conditions and can estimate F0 to within 5% of the true value in 80% of voiced speech segments of babble, at SNRs as low as 5 dB [23]. So, we developed a method of enhancing event detection with F0 information.

Fig. 4 shows a block diagram of the procedure. First, candidate event times are calculated using the event detection mechanism described in the previous section. The half-life of the adaptive threshold is reduced to avoid missing events in the target speech. The procedure extracts events for both the target and background speech when both are voiced. Then we produce a temporal sequence of the pulses located at the candidate times. For every frame step (e.g., 3 ms), the value of F0 for the target speech in that frame is converted into a temporal function consisting of a triplet of gaussian functions spaced at 1/F0; the standard deviation of the gaussians in this example was 0.3 ms. The triplet function is, then, cross-correlated with the event–pulse sequence to find the best matching lag time. At this best point,

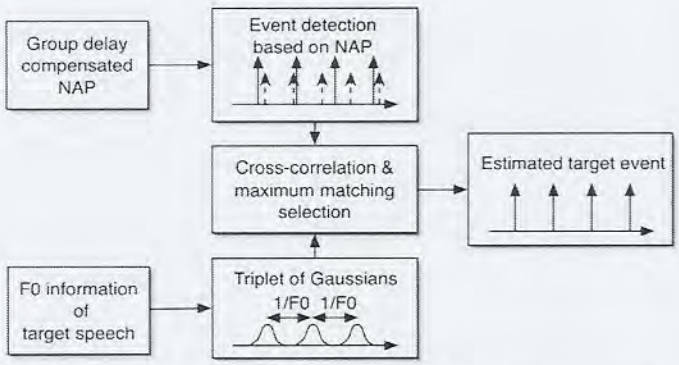

Fig. 4. Schematic representation of the estimation of glottal events for the target speaker. Events for the target sound (solid arrows) and background speech (dashed arrows) were first estimated from the compensated NAP. The target event (rightmost box) was estimated from the compensated NAP by cross-correlating with the triplet of Gaussians.

the temporal sequence and the triplet of gaussians are multiplied together so as to derive a value similar to the likelihood for each event. The likelihood–like value becomes very large when the periods of the two sequences match. However, it is not necessarily the case when the estimate of F0 contains an error, or when the candidate event is derived from regular glottal pulses. Nevertheless, we must have event times for strobing, and so, the likelihood–like value is accumulated until it exceeds a criterion threshold value. Then, the candidate event time is taken to be the event time for the target speech signal. The number of the event times increases as the threshold value decreases and vice versa. The threshold value was set manually in the following experiment, but it should be possible to introduce a statistical model to set the threshold optimally.

4) Event Synchronous Strobed Temporal Integration: Using the target event times, we can enhance the AI for the target speech. Fig. 2(a) shows a schematic of a NAP after group delay compensation, for a segment of speech where there are concurrent vowels from two speakers. The target speech with a 10-ms glottal period is converted into the black activity pattern, while the background speech with a 7-ms period is converted into the gray pattern. For every target event time (every 10 ms), the NAP is converted into a two-dimensional AI as shown in Fig. 2(b). The horizontal axis of the AI is time-interval (TI) from the event time; the vertical axis is the same as the NAP (i.e., the tonotopic dimension of the cochlea). As shown in Fig. 2(b), the activity pattern for the target speech always occurs at the same time–interval, whereas that for the background speech changes position. So, we get a "stabilized" version of the target speech pattern in this "instantaneous" AI. Thus, (3) becomes

$$A_{Ii}(f_c, \tau, t_e) = S_N(f_c, \tau + t_e)e^{-\xi\tau} \tag{5}$$

where t_e is the time of the event, and τ is the time–interval axis of the SAI. When the event detection contains errors, the image deteriorates. So, reliable event-detection is essential for this process.

Temporal integration is performed by weighted averaging of successive frames of the instantaneous AI as shown in Fig. 2(c),

and this enhances the pattern for the target speech relative to that for the background speech. Specifically, the weighting is

$$A_I(f_c, \tau, t_e) = \sum_{t_e=0}^{\infty} W_A(t_e) A_{Ii}(f_c, \tau, t_e). \quad (6)$$

In this way, the fine structure of the target pattern is preserved and stabilized when the event detector captures the glottal events of the target speech correctly. The weighted averaging is essentially equivalent to conventional STI where the weighting function $W_A(t)$ is a ramped exponential with a fixed half-life. We initially used a hanning window spanning five successive SAI frames; tests indicated that window shape does not have a large effect as long as the window length is within a reasonable range. Simple weighted averaging is not necessarily the best method to reduce the noise in the two-dimensional image, particularly when the response to concurrent speech appears at essentially random time intervals in the instantaneous SAI [e.g., Fig. 2(b)]. In the following evaluation, we applied simple median filtering over five successive event-frames since we found it reduces the noise level. It should be possible to improve the sound quality more by applying advanced image processing techniques for noise reduction to the fine structure in the SAI.

B. Synthesis Procedure

Having obtained the SAI for the target speech, we then need to resynthesize the sound from the SAI. The synthesis module was developed using the concept of the channel vocoder [21], [22].

1) Auditory Channel Vocoder: A channel vocoder consists of a bank of band-limited filters and pulse/noise generators controlled by previously extracted information about fundamental frequency and voicing. In the current case, there was a pair of gammachirp/gammatone filterbanks for analysis and synthesis (the leftmost and rightmost boxes in Fig. 1) because of the temporal asymmetry in the filter response. But the synthesis filterbank is not essential when we compensate for the group delay of the analysis filter. What is essential for effective resynthesis is a good estimate of BMM which has to be recovered from the amplitude information in the SAI and the phase information generated from the F0 information.

2) Mapping Function From SAI to NAP: The synthesis module of Fig. 1 shows that the mapping from SAI to NAP is essential for resynthesizing the enhanced speech sounds. Previously [30], [31], we used the fine temporal structure in both the SAI and NAP to design a mapping function using nonlinear, multivariate, regression analysis, MRA. Statistical training was required to formulate the mapping between multiple inputs and a single output. It would be ideal when there is sufficient data for the training and sufficient time for the training, but oftentimes there is not. In this paper, we made a simple, fixed circuit for the mapping to provide the baseline performance of the auditory vocoder. We assumed that the frequency profile of the SAI would provide a reasonable representation of the main features of the target speech, if the profile mainly represents the activity of the target speech in the SAI. The first two-dimensional image in Fig. 2(c) shows that the target feature is concentrated in the time-interval dimension when the filter center frequency f_c is high, and it overlaps the next cycle when the center frequency is low. The activity of the interfering speech appears at a low level in the rest of the SAI. So, when we cumulate the activity within the area of the target feature, it would produce a reasonable profile. The formula for converting the activity from a SAI to a NAP is

$$S_S(f_c, t_e) = \frac{1}{T_{\max} - T_{\min}} \int_{T_{\min}}^{T_{\max}} W_S(f_c, \tau) A_I(f_c, \tau, t_e) d\tau \quad (7)$$

where $W_S(f_c, \tau)$ is a weighting function that produces a flat NAP from a regular SAI. The range of accumulation $[T_{\min}, T_{\max}]$ on the time-interval axis was

$$T_{\min} = -1; \quad T_{\max} = \min\left\{5, \max\left(1.5, \frac{2}{f_c}\right)\right\} \text{(ms)}. \quad (8)$$

$T_{\max}$ is 5 ms when f_c is low and 1.5 ms when f_c is high. The frequency-dependent range is intended to contain the ridges of activity shown in the first image of Fig. 2(c). At this point, the constant values, -1, 1.5, and 5 ms, were selected manually in accordance with preliminary simulations. When the values increase, the activity includes more distracter components, When the values decrease, the resonance information of the target reduces gradually.

The amplitude of the NAP $S_S(f_c, t_e)$ is recovered by (7) for every event time. The amplitude between successive events was recovered by linear interpolation which is sufficient for good quality resynthesized sound. Then, the amplitude of the BMM is recovered using the inverse function.

3) Phase Retrieval and Resynthesis: Since the phase information is completely lost in this process, it is necessary to generate phase information for resynthesizing the sounds. We use F0 information to generate the glottal pulse and noise in accordance with the target speech sound, as shown in the bottom box of the synthesis module in Fig. 1. We employed the generator in STRAIGHT [22], since it is known to produce pulses and noise with generally flat spectra, and it does not affect the amplitude information. The waveform of the pulse and noise sequence is analyzed by the complex version of the analysis filterbank in AIM. Then the phase information is calculated and multiplied with the amplitude of the recovered BMM. Finally, a temporally-reversed auditory filterbank is applied to reintroduce the group delay of the original analysis filterbank. The resynthesized sounds are then derived by simply summing up the channel outputs.

The resulting resynthesized sounds were reasonably good even though a simple, fixed circuit was used for the mapping function between the SAI and the NAP.

III. Evaluation

Segregation methods based on harmonic selection require precise F0 information. However, this is unrealistic; natural vocal vibration contains jitter and there is commonly background noise. So F0 estimation is never error-free in practice,

and it is important to evaluate how robust the system is to jitter and errors in F0 estimation. In this section, we compare the new method to a comb-filter method using sound mixtures in three ways:

1) with visual inspection of spectra (Section III-A);
2) with an objective measure involving cross-correlation (Section III-B);
3) with perceptual experiments (Section III-C).

First, we estimated F0 and the degree of voicing in advance from the clean target speech using an algorithm in STRAIGHT [22]. The F0 errors were a proportion of the estimated values, and the proportion was varied. Then a mix of target speech and background noise was produced with either babble or concurrent speech to evaluate the auditory vocoder relative to the comb-filter method. The target and concurrent speech sounds were selected randomly from a database of four-mora Japanese words, using the same male speaker to make the task difficult. The babble was produced by adding speech from ten speakers; a segment of a few minutes was produced, and samples with the appropriate duration were extracted with random start times. The unvoiced parts of these speech sounds were a problem for both methods, and so they were passed through without processing in Sections III-A and III-C, and they were removed when evaluating the cross-correlation in Section III-B.

A. Comparison Based on the Spectrogram

Fig. 5(a) and (b) shows the original isolated target-speech wave for the Japanese word "imafu" and its spectrogram, respectively. The target sound was mixed with babble at an SNR of 0 dB to produce the wave in Fig. 5(c). At this level, a listener needs to concentrate to hear the target speech in the distracter. The spectrogram in Fig. 5(d) shows that the features of the target speech are no longer distinctive. When the F0 error is 5%, the *target* speech extracted by the auditory vocoder [Fig. 5(e)] was distorted but it was still intelligible; the *distracter* speech was disrupted to the point where it no longer sounded like speech, and this alteration of the perception greatly reduced the interference that it would otherwise have produced. The spectrogram in Fig. 5(f) shows that vocoder resynthesis restores the harmonic structure of the target speech but not that of the distracter speech. It is easier to listen to the target sound than the mix of target speech and distracter speech, or the mix of target speech and babble. This is partially because the pitch pattern is regenerated using F0 information.

For comparison, we applied a comb-filter method based on the STFT to the same sound, again with an F0 error of 5%. The target speech extracted in this way has a hollow character and the babble is relatively strong. The target speech wave in Fig. 5(g) is noisier than that in Fig. 5(e). The spectrogram in Fig. 5(f) shows that much of the harmonic structure is lost by the comb-filter method, particularly for the higher harmonics. This follows directly from the fact that the error in this method increases in proportion to harmonic number. As F0 error increased, the quality of the target speech sounds degraded and the background noise increased. This is a fundamental defect of comb-filter and harmonic-selection methods.

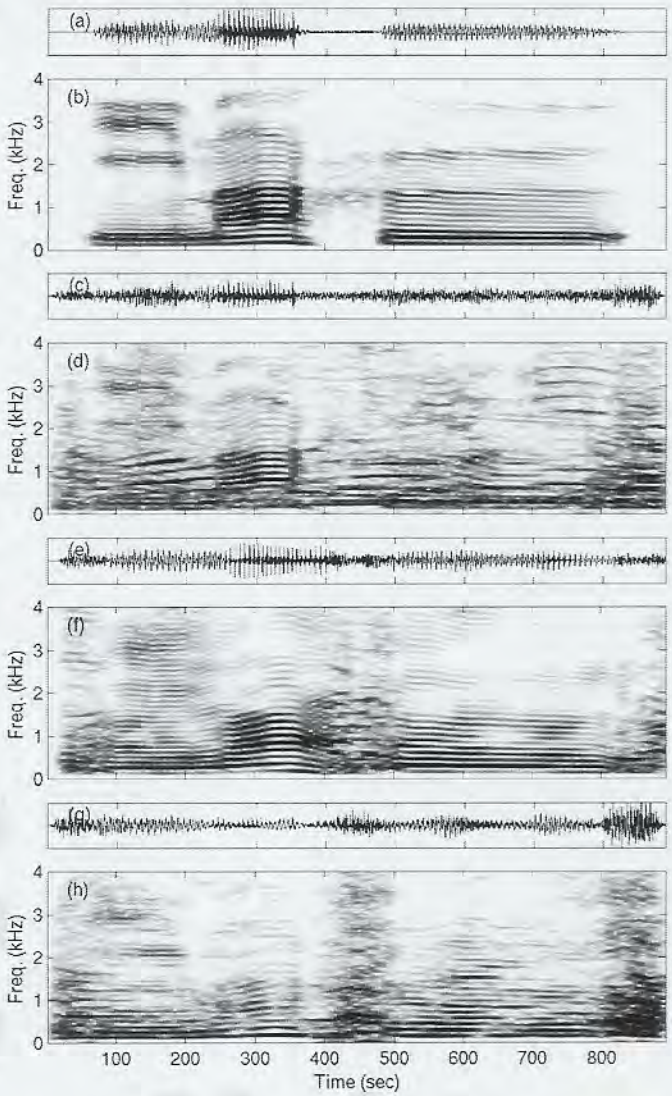

Fig. 5. Waveforms and spectra when the error rate of the F0 estimation is 5%. (a) and (b) Original waves for the Japanese word "imafu." (c) and (d) Mix of sound and babble (SNR = 0 dB). (e) and (f) Segregated target speech from the auditory vocoder. (g) and (h) Segregated target speech from the comb-filter.

B. Objective Evaluation

This section describes an objective measure designed to quantify the observations in Section III-A and the difference in performance between the auditory vocoder and the comb-filter method.

1) Cross-Correlation Measure: It is difficult to define the SNR at the output of the auditory vocoder because it is difficult to separate the components of the target speech and background noise due to the nonlinear processing. Moreover, it is difficult to define an ideal binary mask or "oracle" mask [1] to quantify the degree of segregation. However, we can calculate a cross-correlation function between the output sound and the original input (either the target speech or the background noise). When they are highly correlated, a large peak is observed at the appropriate delay. Since the auditory vocoder introduces nonuniform time

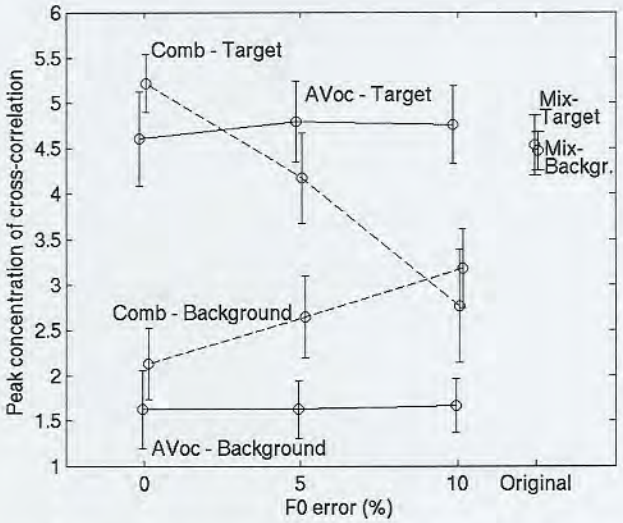

Fig. 6. Peak concentration of the cross-correlation function when the background is babble and the SNR is 0 dB. The abbreviations "AVoc," "Comb," and "Mix" indicate the outputs of the auditory vocoder, the comb-filter, and the original mix, respectively. "Target" and "Background" are the original target speech and background sounds. So, for example, "AVoc-Target" indicates the output of the auditory vocoder was cross-correlated with the clean target speech.

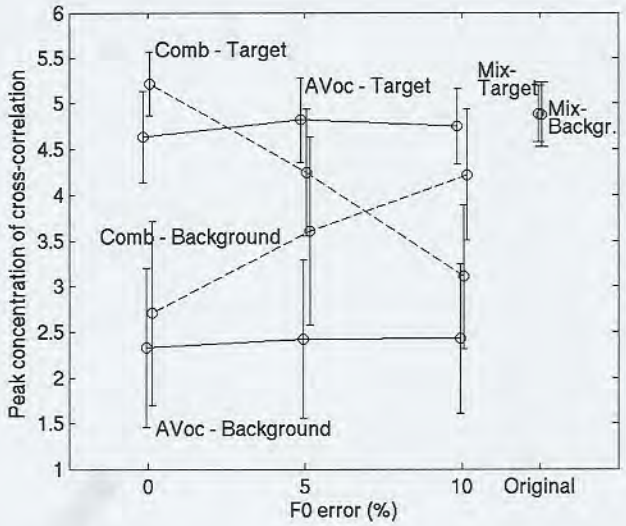

Fig. 7. Peak concentration of the cross-correlation function when the background is concurrent speech and the SNR is 0 dB.

delays, we defined a measure referred to as the "peak concentration" of the cross-correlation function. It is

$$X_P = \sqrt{\frac{\frac{1}{2T_P}\int_{-T_p}^{T_p}\{X_C(\tau)\}^2\,d\tau}{\frac{1}{2T_s}\int_{-T_s}^{T_s}\{X_C(\tau)\}^2\,d\tau}} \tag{9}$$

where $X_C(\tau) = \int_{-T_S}^{T_S} x(t)y(t+\tau)dt$, $x(t)$ is the input signal, $y(t)$ is the output signal, T_s is the signal duration, and T_p is a window that defines acceptable time delays. The width of T_p was 20 ms in this case; its value did not have much effect on the results. The peak concentration value is calculated only during the voiced parts of speech. When there is F0 error, the F0 trajectories are different in the target speech and the output of the auditory vocoder, and this mismatch leads to inaccurate estimation of the peak concentration. So, the target sound was given the same F0 trajectory using STRAIGHT [22].

2) Results: We calculated the peak concentration of the cross-correlation function between the system output and the original target speech (designated "Target"), or between the system output and the original background sound (designated "Background"), for a set of 50 words drawn from a database [32], [33]. The background was babble or concurrent speech at an SNR of 0 dB. The system outputs were the output of the auditory vocoder with F0 error levels of 0%, 5%, and 10% ("AVoc"); the output of the comb filter with F0 error levels of 0%, 5%, and 10% ("Comb"); and a simple mix of the original target speech and background noise ("Mix") to establish the baseline.

Fig. 6 shows peak concentration as a function of F0 error in percent, when the background was babble. The peak concentration values for "Mix-Target" and "Mix-Background" are almost the same because the SNR is 0 dB. Both of the original sounds are highly correlated with the mixed sounds. As SNR increases, the peak concentration value for "Mix-Target" increases and the value for "Mix-Background" decreases. So, the ratio of the peak concentration values is a form of SNR.

When the F0 error is 0%, the peak concentration value for "Comb-Target" is high and the value for "Comb-Background" is low. This implies that the target is enhanced and the background noise is suppressed at the output of the comb-filter. The difference is smaller when the F0 error is 5%, and the peak concentration values are reversed when the F0 error is 10%. This implies that the SNR is very low (sometimes negative) at the output of the comb filter. The components of the background noise leak through the harmonic filters when they are misaligned by F0 error, and the components of the target are reduced by these same filters.

In contrast, the peak concentration values for "AVoc-Target" are consistently high (about 4.5) and the values for "AVoc-Background" are constantly low, independent of the level of F0 error. This indicates that the component corresponding to the background noise at the output of the auditory vocoder does not sound similar to the original background noise, as observed in Section III-A. The auditory vocoder prevents the background noise leaking through in its original form giving the vocoder an advantage over the comb-filter method.

Fig. 7 shows peak concentration when the background was a single, concurrent speaker with an SNR of 0 dB. Although the variability is greater, and some of the values are larger, the general pattern of results is the same as when the background was babble (Fig. 6). The peak concentration of the cross-correlation function provides a rough estimate of the SNR in the output speech, and as such, it provides an objective measure for nonlinear speech processors. The cross-correlation function corresponds to the cross-power spectrum between two extended sounds. However, it does not necessarily reflect local distortion due to nonlinear processing. Accordingly, we also performed an experiment to evaluate the perceptual effect of the distortion.

C. Perceptual Experiments

Speech intelligibility experiments were performed with Japanese four-mora words randomly selected from a database, which is controlled for word familiarity and phonetic balance [32], [33].

1) Conditions, Subjects, and Task: There were five conditions: (1) a simple mix of the original target and background (designated "Mix"), (2,3) the output of the auditory vocoder with F0 errors of 0% ("AVoc (0)") and 5% ["AVoc (5)"], and (4,5) the output of the comb filter with F0 errors of 0% ["Comb (0)"] and 5% ["Comb (5)"]. For each condition, 25 words were selected from one male speaker. There was no overlap of words in these experiments. The background noise was either babble or concurrent speech at an SNR of 0 dB. The SNR was defined using the root mean square (rms) value of the whole sound. The SNR of 0 dB was chosen to make the task moderately difficult. It was very difficult to identify the words even for a simple mix with babble when the SNR was -5 dB; it was easy to identify the words for this mix when the SNR 5 dB or more.

The signal level was digitally equalized in terms of the equal loudness level, or L_{Aeq}, which is the rms level of the entire speech sound after A-weight filtering. The sounds were played back using M-audio FireWire 410 at 48 kHz and presented to the subjects in a sound-proof room through headphones (Sennheiser HD-580) at a level of 70 dBA. The words were presented singly in random order and followed by a 5-s response interval.

Four young adults with normal hearing thresholds between 250 and 8000 Hz participated in the experiments. The subjects were instructed to write down the words in Japanese "kana" characters, which uniquely identify the CV and V syllables of Japanese words. The response sequence was compared with the sequence presented to produce recognition scores, separately, for words, vowels, and consonants.

2) Results: The speech recognition scores for the five conditions when the background noise was babble at an SNR of 0 dB are presented in Fig. 8. The word recognition score for the original mix is 64% which is better than the scores for processed sounds. The degradation due to signal processing has been repeatedly reported [7], [34]. The word recognition scores for the comb-filter are 55% and 14% when the F0 error is 0% and 5%, respectively. The F0 error greatly affects the quality of the output sounds. The word recognition scores for the auditory vocoder are 11% and 31% when the F0 errors are 0% and 5%, respectively. We do not understand the reason why the score is worse when there is no F0 error. It results from poor consonant recognition; the vowel recognition scores are as high as 59% and 67%, although the vowels were always processed. In contrast, the word and vowel recognition scores for the comb-filter are 14% and 47% when the F0 error is 5%. The results show the auditory vocoder has the potential to outperform the comb-filter when there is error in F0 estimation.

The speech recognition scores for the five conditions when the background noise is concurrent speech at an SNR of 0 dB are presented in Fig. 9. For the original mix, the subjects were required to answer with one of the two words which were presented simultaneously and were almost equally loud. Although there is no correct target, the score was calculated assuming

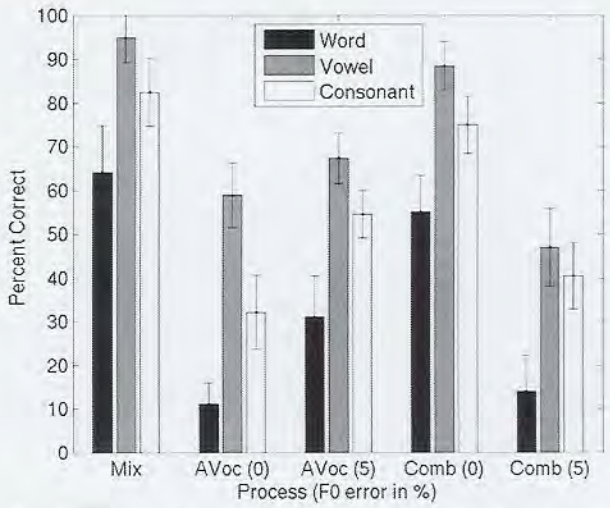

Fig. 8. Intelligibility of words, vowels, and consonants when the background is babble and the SNR is 0 dB. "Mix" indicates the original mix of target speech and babble. "AVoc (0)" and "AVoc (5)" indicate the output of the auditory vocoder when the F0 error was 0 or 5%, respectively. "Comb (0)" and "Comb (5)" indicate the output of the comb-filter when the F0 errors was 0 or 5%, respectively.

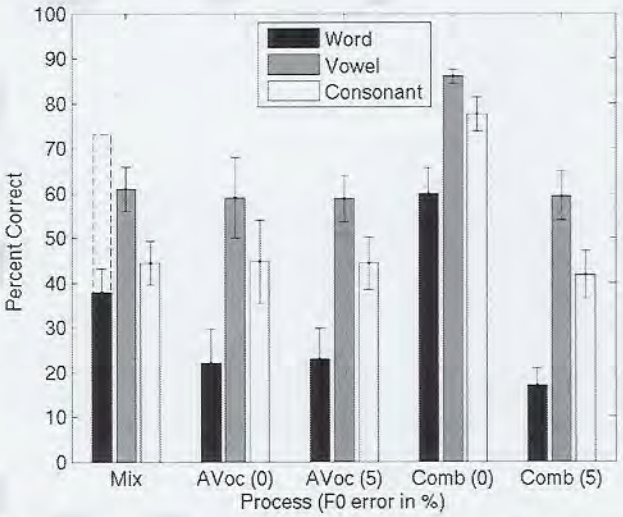

Fig. 9. Intelligibility of words, vowels, and consonants when the background was concurrent speech and the SNR was 0 dB.

one of them, chosen arbitrarily, was the target speech. When we allow either of the words was the target speech, the score almost doubles as indicated by the dashed bar in the left most column.

The comb-filter with no F0 error successfully draws the attention of the subjects to the target sound. The other conditions, "AVoc (0)," "AVoc (5)," and "Comb (5)," produce roughly the same results. The performance of the comb-filter is affected by the F0 error while the performance of the auditory vocoder is almost independent of F0 error.

IV. Conclusion

We have proposed an auditory vocoder for speech segregation. Unlike many conventional signal processing methods, the auditory vocoder does not apply a windowing function directly to the sound wave. Instead, it produces a temporal sequence of

two-dimensional frames similar to video signals by an event-synchronous procedure. The representation is referred to as the auditory image. The procedure enables us to preserve fine structure in the speech signals and to apply image processing techniques to enhance important features.

We compared the performance of the auditory vocoder with a popular comb-filter method, using visual inspection of spectrograms, an objective cross-correlation measure, and perceptual recognition experiments. The cross-correlation measure indicated that the performance of the auditory vocoder was independent of the degree of F0 error up to 10%, whereas the performance of the comb-filter method degrades progressively as F0 error increases. The perceptual experiments showed that the auditory vocoder was better than the comb-filter method whenever there was F0 error. The comb-filter method was better when there was no F0 error, but the normal situation always involves some degree of F0 error. The processing introduces perceptual distortion that limits the recognition score to values less than those of the original mix.

The event-synchronous enhancement was only applied to the voiced parts of the target speech in the evaluation. If it is applied to the unvoiced parts of speech, it introduces a little more degradation of speech quality because the event-detection mechanism would operate even if there were no glottal pulses. It should be noted, however, that the comb filter disrupts sound quality badly in the unvoiced parts of the speech. So the auditory vocoder is more effective in the unvoiced parts of speech, and particularly so around the voicing boundaries of the target speech, which are generally difficult to detect in background noise.

Although the performance of the auditory vocoder was better than that of the comb-filter, the advantage was not as great as might have been expected. The analysis of speaker segregation presented in this paper suggests there are several aspects of the processing that could be improved, and which would lead directly to better performance. First, the strobing that segregates the speakers by their glottal events is a crucial part of the processing, but the current version is clearly not as robust to background noise as human speech recognition. Auditory scene analysis indicates that a more sophisticated form of cross-channel integration would improve the robustness and thus the sound quality. Second, it should be possible to improve the enhancement of target features in the sequence of auditory images; specifically, advanced image processing techniques could be used to enhance the important features and determine the optimal range for the accumulation of information [(8)] in the synthesis of voiced features. For the unvoiced parts of speech, we need to find a way of introducing knowledge about speech features as they appear in the auditory image, as a means of constraining recognition decisions. Finally, it is essential to develop a robust method of F0 estimation that can accurately estimate the degree of voicing in the target speech in the presence of background noise.

In summary, the auditory vocoder provides a good framework for the incorporation of knowledge about auditory signal processing, in a form that can support incremental progress on CASA problems like speaker segregation.

REFERENCES

[1] P. Divenyi, Ed., *Speech Separation by Humans and Machines*. Norwell, USA: Kluwer, 2004.

[2] G. J. Brown and M. P. Cooke, "Computational auditory scene analysis," *Comput. Speech, Lang.*, vol. 8, pp. 297–336, 1994.

[3] M. Slaney, D. Naar, and R. F. Lyon, "Auditory model inversion for sound separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP-94)*, 1994, vol. II, pp. 77–80.

[4] D. P. W. Ellis, "Prediction-driven computational auditory scene analysis," Ph.D. dissertation, Dept. Elec. Eng Comp. Sci., Mass. Inst. Technol., Cambridge, MA, 1996.

[5] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 684–697, May 1999.

[6] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *J. Acoust. Soc. Amer.*, vol. 60, pp. 911–918, 1976.

[7] J. S. Lim, A. V. Oppenheim, and L. D. Braida, "Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. ASSP-26, no. 4, pp. 354–358, Aug. 1978.

[8] J. C. R. Licklider, "A duplex theory of pitch perception," *Experientia*, vol. 7, pp. 128–133, 1951.

[9] M. Slaney and R. F. Lyon, "A perceptual pitch detector," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP-90)*, 1990, pp. 357–360.

[10] R. Meddis and M. J. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification," *J. Acoust. Soc. Amer.*, vol. 89, no. 6, pp. 2866–2882, 1991.

[11] M. Weintraub, "A computational model for separating two simultaneous talkers," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP-86)*, 1986, pp. 81–84.

[12] R. D. Patterson, "The sound of a sinusoid: Spectral models," *J. Acoust. Soc. Amer.*, vol. 96, pp. 1409–1418, 1994.

[13] ——, "The sound of a sinusoid: Time-interval models," *J. Acoust. Soc. Amer.*, vol. 96, pp. 1419–1428, 1994.

[14] M. A. Akeroyd and R. D. Patterson, "A comparison of detection and discrimination of temporally asymmetry in amplitude modulation," *J. Acoust. Soc. Amer.*, vol. 101, pp. 430–439, 1997.

[15] T. Irino and R. D. Patterson, "Temporal asymmetry in the auditory system," *J. Acoust. Soc. Amer.*, vol. 99, no. 4, pp. 2316–2331, 1996.

[16] R. D. Patterson and T. Irino, "Modeling temporal asymmetry in the auditory system," *J. Acoust. Soc. Amer.*, vol. 104, no. 5, pp. 2967–2979, 1998.

[17] R. D. Patterson, M. Allerhand, and C. Giguère, "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform," *J. Acoust. Soc. Amer.*, vol. 98, pp. 1890–1894, 1995.

[18] S. Bleeck, T. Ives, and R. D. Patterson, "Aim-mat: The auditory image model in MATLAB," *Acta Acustica*, vol. 90, pp. 781–788, 2004.

[19] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory Physiology Perception, Proc. 9th Int. Symp. Hearing*, Y. Cazals, L. Demany, and K. Horner, Eds., 1992, pp. 429–446.

[20] R. D. Patterson, "Auditory images: how complex sounds are represented in the auditory system," *J. Acoust. Soc. Japan (E)*, vol. 21, pp. 183–190, 2000.

[21] B. Gold and C. M. Rader, "The channel vocoder," *IEEE Trans. Audio Electroacoust.*, vol. AU-15, no. 4, pp. 148–161, Dec. 1967.

[22] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "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," *Speech Commun.*, vol. 27, pp. 187–207, 1999.

[23] T. Nakatani and T. Irino, "Robust and accurate fundamental frequency estimation based on dominant harmonic components," *J. Acoust. Soc. Amer.*, vol. 116, no. 6, pp. 3690–3700, 2004.

[24] T. Irino and R. D. Patterson, "Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised wavelet mellin transform," *Speech Commun.*, vol. 36, no. 3–4, pp. 181–203, Mar. 2002.

[25] ——, "A time-domain, level-dependent auditory filter: The gammachirp," *J. Acoust. Soc. Amer.*, vol. 101, no. 1, pp. 412–419, Jan. 1997.

[26] ——, "A compressive gammachirp auditory filter for both physiological and psychophysical data," *J. Acoust. Soc. Amer.*, vol. 109, no. 5, pp. 2008–2022, May 2001.
[27] R. D. Patterson, M. Unoki, and T. Irino, "Extending the domain of center frequencies for the compressive gammachirp auditory filter," *J. Acoust. Soc. Amer.*, vol. 114, pp. 1529–1542, 2003.
[28] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, pp. 103–138, 2003.
[29] B. Gold and L. Rabiner, "Parallel processing techniques for estimating pitch periods of speech in the time domain," *J. Acoust. Soc. Amer.*, vol. 46, pp. 442–448, 1969.
[30] T. Irino, R. D. Patterson, and H. Kawahara, "Speech segregation using event synchronous auditory vocoder," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2003, vol. 5, pp. 525–528.
[31] ——, "Speech segregation based on fundamental event information using an auditory vocoder," in *Proc. 8th Eur. Conf. Speech Commun. Technol. (Eurospeech 2003, Interspeech 2003)*, Geneva, Switzerland, 2003, pp. 553–556.
[32] S. Sakamoto, Y. Suzuki, S. Amano, K. Ozawa, K. Kondo, and T. Sone, "New lists for word intelligibility test based on word familiarity and phonetic balance," (in Japanese) *J. Acoust. Soc. Jpn.*, vol. 54, no. 12, pp. 842–849, 1998.
[33] S. Sakamoto, N. Iwaoka, Y. Suzuki, S. Amano, and T. Kondo, "Complementary relationship between familiarity and SNR in word intelligibility test," *Acoust. Sci. Tech.*, vol. 25, no. 4, pp. 290–292, 2004.
[34] B. C. J. Moore, *Perceptual Consequences of Cochlear Damage*. Oxford, U.K.: Oxford Univ. Press, 1995.

Toshio Irino (SM'04) was born in Yokohama, Japan, in 1960. He received the B.S., M.S., and Ph.D. degrees in electrical and electronic engineering from Tokyo Institute of Technology, Tokyo, Japan, 1982, 1984, and 1987, respectively.

From 1987 to 1997, he was a Research Scientist at NTT Basic Research Laboratories, Tokyo Japan. From 1993 to 1994, he was a Visiting Researcher at the Medical Research Council—Applied Psychology Unit (MRC-APU, currently CBU), Cambridge, U.K. From 1997 to 2000, he was a Senior Researcher in ATR Human Information Processing Research Laboratories (ATR HIP). From 2000 to 2002, he was a Senior Research Scientist in NTT Communication Science Laboratories. Since 2002, he has been been a Professor of the Faculty of Systems Engineering, Wakayama University, Wakayama, Japan. He is also a Visiting Professor at the Institute of Statistical Mathematics. The focus of his current research is a computational theory of the auditory system.

Dr. Irino is a member of the Acoustical Society of America (ASA), the Acoustical Society of Japan (ASJ), and the Institute of Electronics, Information and Communication Engineers (IEICE), Japan.

Roy D. Patterson was born in Boston, MA, on May 24, 1944. He received the B.A. degree from the University of Toronto, Toronto, ON, Canada, in 1967 and the Ph.D. degree in residue pitch perception from the University of California, Los Angeles, in 1971.

From 1975 to 1995, he was a Research Scientist for the U.K. Medical Research Council, at their Applied Psychology Unit, Cambridge, England, focusing on the measurement of frequency resolution in the human auditory system, and computational models of the auditory perception. He also designed and helped implement auditory warning systems for civil and military aircraft, railway maintenance equipment, the operating theatres and intensive care wards of hospitals, and most recently, fire stations of the London Fire Brigade. Since 1996, he has been the Head of the Centre for the Neural Basis of Hearing, Department of Physiology, University of Cambridge, Cambridge, U.K. The focus of his current research is an "Auditory Image Model" of auditory perception and how it can be used to: 1) normalize communication sounds for glottal pulse rate and vocal tract length and 2) produce a size-invariant representation of the message in communication sounds at the syllable level. He has published over 100 articles in *JASA* and other international journals.

Dr. Patterson is a Fellow of the Acoustical Society of America.

Hideki Kawahara (SM'91) was born in Sapporo, Japan, on December 31 1949. He received the B.A. degree from Hokkaido University, Sapporo, in 1972 and the Ph.D. degree on the linear predictive analysis of vowels from ... in 1977 ***[AUTHOR: From Hokkaido University?]***.

From 1977 to 1992, he was a Research Scientist at NTT Basic Research Laboratories, Tokyo Japan, where he participated in designing telephone-transmission-quality assessment systems. He also participated in the development of broadband communication network systems. In 1989, he began research on signal processing models of hearing and neural networks. From 1992 to 1997, he was Head of the first department of ATR Human Information Processing Research Laboratories, Kyoto, Japan, where he promoted research on interaction between speech perception and production. Since 1997, he has been a Professor of the Faculty of Systems Engineering, Wakayama University, Wakayama, Japan. From 1997 to 2002, he has also been a Principal Investigator of the Auditory Brain Project sponsored by CREST JST, where he invented a high-quality speech analysis, modification, and resynthesis system referred to as STRAIGHT. His current research focus is the use of STRAIGHT inresearch on human speech perception with the focus on ecologically relevant, but well controlled, synthetic speech stimuli.

Dr. Kawahara is a board member of the Japanese Acoustical Society.

[28] T. Irino and R. D. Patterson, "A dynamic, compressive gammachirp auditory filterbank," IEEE Trans. Audio, Speech, Language Process., vol. 14, no. 6, The same issue (Nov. 2006)

T-SA-00809-2005 / SAP 1745