AN AUDITORY VOCODER RESYNTHESIS OF SPEECH FROM AN AUDITORY MELLIN REPRESENTATION

PACS : 43.72.Ar; 43.71.An; 43.66.Ba

Toshio Irino [*]; Roy D. Patterson [**]; and Hideki Kawahara [+]

[*] NTT Communication Science Laboratories, NTT Corporation / CREST-JST
2-4, Hikaridai Seika-cho Soraku-gun Kyoto, 619-0237, JAPAN
Tel: +81-774-93-5333; Fax: +81-774-93-5155
E-mail: irino@cslab.kecl.ntt.co.jp
[**] Centre for the Neural Basis of Hearing, Physiology Department, Cambridge University
Downing Street, Cambridge, CB2 3EG, UK.
Tel: +44-1223-333819; Fax: +44-1223-333840
E-mail: roy.patterson@mrc-cbu.cam.ac.uk
[+] Wakayama University / ATR / CREST-JST
930 Sakaedani, Wakayama, Wakayama 640-8510, JAPAN
Tel: +81-734-57-8461; Fax: +81-734-57-8112
E-mail: kawahara@sys.wakayama-u.ac.jp

ABSTRACT

An auditory Mellin transform has been proposed to segregate information about the size and shape of the vocal tract automatically; the process is also independent of glottal pitch. In this paper, we describe a method for resynthesizing speech from the Mellin representation using a high quality vocoder (STRAIGHT), and a nonlinear function to map between the two representations of speech. This enables us to replay the coded speech to evaluate the glottal event detector that precedes the Mellin transform, and the spectral information recovered from the Mellin representation.

1. INTRODUCTION

Speech analysis/synthesis schemes based on vocoders [1] are an essential tool in speech research and signal processing. Traditional vocoders include linear predictive coding (LPC) which is essentially spectral analysis in a linear frequency domain; they are common in mobile phone systems. Recently, STRAIGHT [2] was used improve the sound quality of a research vocoder by adaptively manipulating the Short-Time Fourier Transform (STFT). In addition, a mel, log-spectral approximation filter (MLSA) [4] was developed to resynthesize speech from mel-frequency cepstral coefficients (MFCC) within the vocoder framework. Another vocoder method based on MFCC was also proposed for speech morphing [19].

The success of MFCC in automatic speech recognition (ASR) [3] is often attributed to its auditory origins, but the cepstral calculation following the mel-spectral analysis is cannot be justified in terms of auditory processing. The windowing in the mel-spectral calculation eliminates fine temporal detail that human listeners hear [5]. The purpose of this project was to develop a method of resynthesizing sounds from a better auditory representation in an effort to improve speech morphing, noise suppression, and speech segregation.

Auditory models are analysis systems by nature; they do not normally include resynthesis since perception does not require resynthesis. Simple analysis/resynthesis systems have been produced with the wavelet transform and linear filterbanks on Mel, Bark, or ERB scales. The sounds can be resynthesized from the output of the filterbank when all of the magnitude and phase information is preserved. Sound resynthesis from a nonlinear, level-dependent filterbank has also been achieved using the gammachirp auditory filterbank [6,7]. An iterative method has also been developed to resynthesize sound from an auto-correlation representation computed after auditory spectral analysis [8], although local minima prevent derivation of a unique mapping from the autocorrelation representation. Unlike the vocoder, however, these resynthesis schemes do not support modification of the coded speech, such as fundamental frequency (F0) conversion or spectral morphing.

We have previously proposed an "Auditory Vocoder" [16,17] to resynthesize speech from an auditory representation referred to as the Mellin Image (MI) [9,10] that segregates the size and shape information of incoming sounds. The idea was to link the two F0-independent representations produced by the MI and STRAIGHT [2]. In this paper, we introduce event-synchronous processing into the Auditory Vocoder to improve the sound quality, and demonstrate the importance of glottal event detection in these systems.

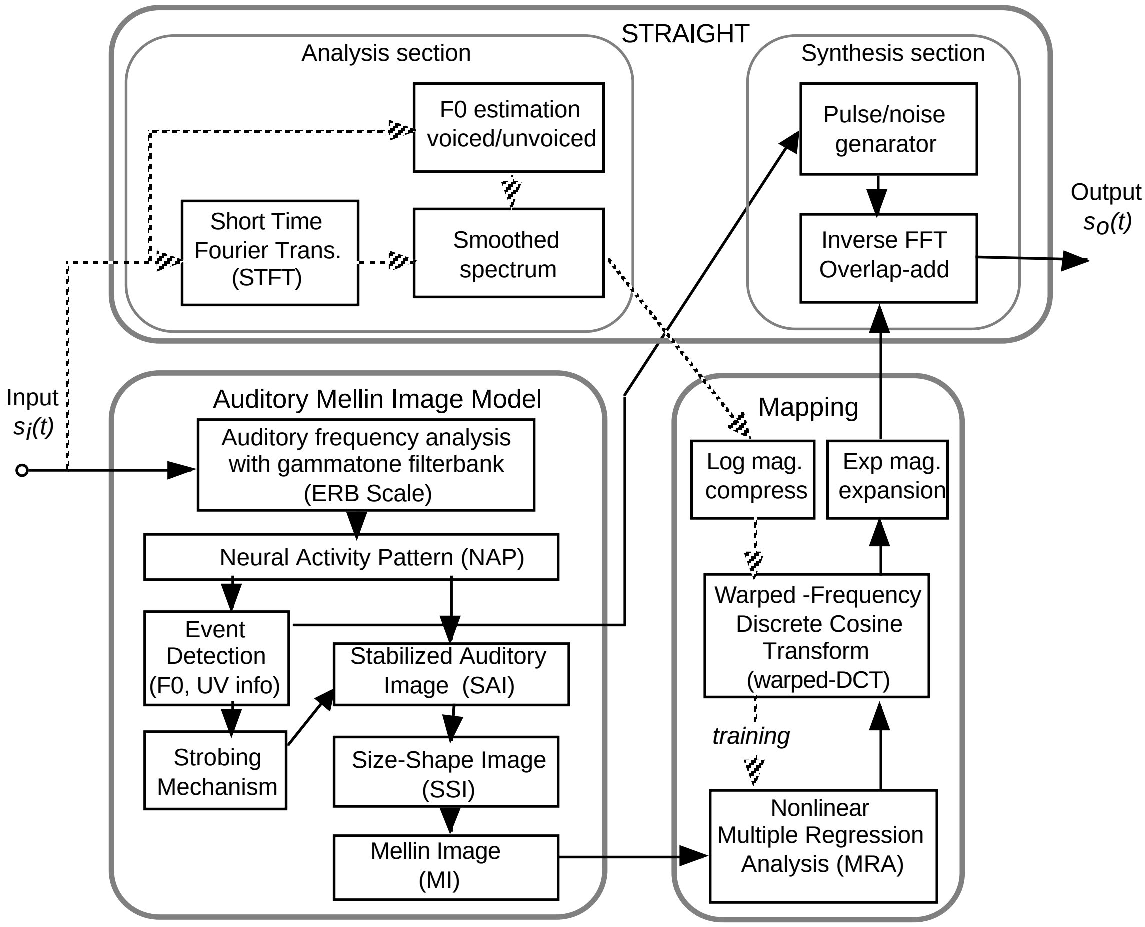
STRAIGHT
Analysis section
F0 estimation voiced/unvoiced
Short Time Fourier Trans. (STFT)
Smoothed spectrum
Synthesis section
Pulse/noise genarator
Inverse FFT Overlap-add
Output $s_o(t)$
Input $s_i(t)$
Auditory Mellin Image Model
Auditory frequency analysis with gammatone filterbank (ERB Scale)
Neural Activity Pattern (NAP)
Event Detection (F0, UV info)
Stabilized Auditory Image (SAI)
Strobing Mechanism
Size-Shape Image (SSI)
Mellin Image (MI)
Mapping
Log mag. compress
Exp mag. expansion
Warped -Frequency Discrete Cosine Transform (warped-DCT)
training
Nonlinear Multiple Regression Analysis (MRA)

Figure1. Block diagram for the sound resynthesis system

Section 2 explains the system architecture and the signal processing applied by each module. Section 3 describes the effect of the system in terms of mapping error and sound quality. Section 4 describes some potential applications of the system.

2. SYSTEM ARCHITECTURE AND PROCESSING

The system (Fig. 1) has three components: STRAIGHT, the Auditory Mellin Image, and a mapping block to link them together.

2.1. STRAIGHT

STRAIGHT [2] is fundamentally a vocoder with analysis and synthesis sections. During the analysis, the F0 is accurately extracted to smooth out the periodic bouncing inherent in short-term spectral analysis. The resulting STRAIGHT spectrum is essentially F0 independent. During the synthesis, pulses or noise with a flat spectrum are generated in accordance with voicing information and the F0. The speech is then resynthesized from the smoothed spectrum with the pulse, or noise, component using an inverse FFT and the overlap-add technique.

2.2. Auditory Mellin Image Model

The auditory model used to produce the Mellin Image [9,10] is based on the Auditory Image Model of perception [12,13]. The model performs its spectral analysis with a gammatone filterbank on the ERB scale. The output is half-wave rectified and logarithmically compressed. Then, adaptive thresholding is applied in each channel to produce a simple form of Neural Activity Pattern (NAP). The NAP is then converted into a Stabilized Auditory Image (SAI) using a very simple strobe mechanism [11]. It is like calculating the times between neural pulses in the auditory nerve and constructing and array of time-interval histograms, one for each channel of the filterbank.

In this project, an event detector was introduced to locate glottal pulses accurately for use as strobe signals, and to estimate F0 accurately. The glottal pulse is extracted from a temporal profile of the NAP after compensation for the group delay in the gammatone filterbank. The local peaks of the summary NAP are extracted as event locations using an algorithm similar to adaptive thresholding [12]. The performance of the event detector is described in section 3.

The vertical axis of the SAI is ERB frequency; the horizontal axis is 'time-interval from the strobe point'. A glottal segment with width 1/F0 is extracted from the SAI to render the representation independent of F0. The segment is, then, converted into a Size-Shape Image (SSI) whose abscissa, *h*, is the product of Time-Interval and Filter-Frequency; the ordinate remains ERB frequency. The Mellin Image (MI) is derived from the SSI using spatial-frequency decomposition (or high-resolution, cepstral decomposition); complex sinusoids are applied along each line of constant *h* in the SSI. The vertical axis of the MI corresponds to cepstral order; the horizontal axis remains the time-interval/peak-frequency product, *h* [9,10].

The vertical profile of the MI, averaged across *h*, is similar to the mel-frequency cepstral coefficients (MFCC) derived from an F0-independent spectral representation. The MFCC derived from the STFT is essentially F0-dependent because of periodic spectral ripples.

We developed a new version of the auditory image model to evaluate the role of glottal event detection in SAI production, and the quality of the spectral information in the MI. The timing of glottal pulses was explicitly extracted from the NAP, as input for the mapping function and to enable event-synchronous construction of the Mellin Image.

2.3. Mapping block

We developed a function to map between the Mellin Image and the STRAIGHT spectrum; both are essentially F0-independent representations.

2.3.1. Strategy

The STRAIGHT spectrum was converted into a form of MFCC representation that corresponds to the vertical profile of the MI. Conventional mel-cepstral analysis does not include an analytic method for perfect inversion to a spectral representation. We substituted an orthogonal function, namely warped-frequency DCT, for both mel-cepstral analysis and synthesis as described in section 2.3.2. Then, a nonlinear Multiple-Regression Analysis (MRA) was used to map between the two MFCC-like representations as described in sections 2.3.3 and 2.3.4.

2.3.2. Warped frequency DCT

The logarithmic magnitude of the STRAIGHT spectrum was converted into a cepstral representation using a warped-frequency version of the Discrete Cosine Transform (DCT); specifically,

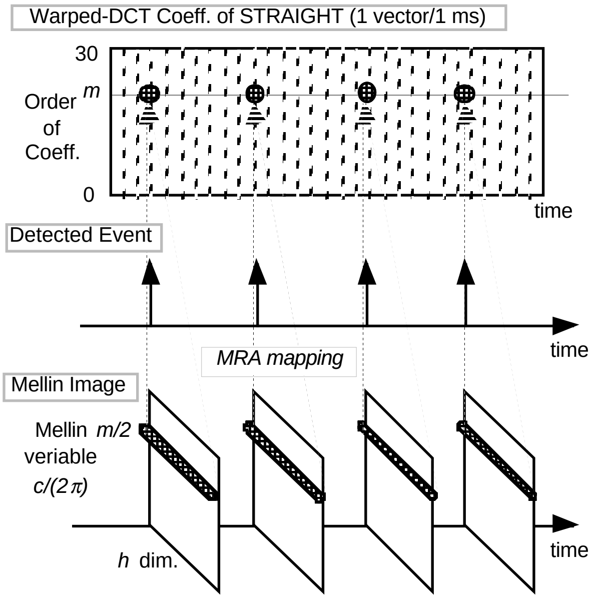

Figure2. Mapping from the MI to the warped-DCT coefficients.

$$\Psi_m(z) = \begin{cases} \dfrac{\sqrt{1-\alpha^2}\,z^{-1}}{1-\alpha z^{-1}}\left(\dfrac{z^{-1}-\alpha}{1-\alpha z^{-1}}\right)^{m-1} & (m>0) \\ 1 & (m=0) \end{cases} \qquad (1)$$

The real part of the frequency response of this filter, $\mathrm{Re}[\Psi_m(\omega)]$, is a normalized, orthogonal function when $\{\omega \mid 0 \le \omega \le \pi\}$. α is a coefficient which determines the degree of frequency warping [12] given by

$$\tilde{\omega} = \omega + 2\arctan\{\alpha\sin\omega/(1-\alpha\cos\omega)\}. \qquad (2)$$

When α is zero, $\mathrm{Re}[\Psi_m(\omega)] = \cos(m\omega)$, which is a discrete cosine. When α is between 0 and 1, $\mathrm{Re}[\Psi_m(\omega)]$ corresponds to a cosine component defined on the warped frequency scale, $\tilde{\omega}$, with a weighting function to maintain the orthogonality. So, the real function, $\mathrm{Re}[\Psi_m(\omega)]$, is used as a kernel for a warped-frequency version of the DCT, and hence the name, warped-DCT . It is also real. When the sampling frequency is 12 kHz and $\alpha = 0.56$, the warped frequency scale, $\tilde{\omega}$, is close to the ERB scale.

The warped-DCT coefficients were calculated from the smoothed, log-magnitude spectrum of STRAIGHT. A simple warped-DCT analysis and synthesis of the STRAIGHT spectrum does not affect the sound quality appreciably when the maximum order, m, is 30.

2.3.3. Arrangement of the mapping function

We performed event-synchronous mapping (Fig. 2) based on the output of the new glottal event detector in the MI module of the analysis/systhesis system (Fig. 1). The new function simplified the mapping as compared with that described previously [16,17], and enabled us to improve the quality of the synthetic speech at the same time.

The Mellin Image (MI) is two-dimensional, and an image is produced for each event detected; whereas, the warped-DCT coefficients have the same frame rate as the STRAIGHT spectrum (1 ms in this case). As noted above, the vertical axis of the MI corresponds to cepstral order. The spatial frequency, $c/(2\pi)$, is defined as cycles within the range of the ERB scale (100-6000 Hz) [9,10], and so a $c/(2\pi)$ value of $m/2$ corresponds roughly to the mth order of the warped-DCT coefficients. Figure 2 shows the arrangement of the mapping function between the MI ($c/(2\pi)$ value of $m/2$) and the warped-DCT coefficients (mth order). The mapping procedure is described for an arbitrary value of m in the *2.3.4*. The procedure is repeated for all m values between 0 and 30. Note that only the real part of the MI was mapped into real warped-DCT coefficients. That is, the size normalization of the MI was ignored in this implementation.

2.3.4. Choice of the mapping function

The MI is produced by a highly non-linear process: log-compression and adaptive thresholding in the NAP, followed by non-linear temporal integration. The STRAIGHT spectrum is also nonlinear albeit to a lesser degree. So, nonlinear, Multivariate Regression Analysis (MRA) was introduced to accommodate the difference in the number of coefficients and the nonlinearities.

We have adopted a form of nonlinear MRA that avoids iterative calculation for computational efficiency. It also enables us to avoid the problems associated with local-minima and over-learning inherent in iterative learning. A method developed for nonlinear Auto-Regressive (AR) analysis [15] was modified to produce the nonlinear MRA since the mathematical formulation is quite similar. This nonlinear MRA also includes the linear case in the formulation.

The explanation variable of the MRA is the $m/2$th vector of the MI extracted at the event time. The vector was set to $\mathbf{x}_k = \{x_{k1}, x_{k2}, ..., x_{kp}\}$ for the kth MI (i.e., the kth event). The dependent, or response, variable was the mth warped-DCT coefficient derived with STRAIGHT at the event time. The response variable was set to y_k.

Once the mapping parameter is estimated from training data, the analysis part of STRAIGHT is no longer required. The speech is analyzed by the auditory Mellin model. For every event, the warped-DCT coefficients are recovered by MRA. Coefficients for intervening 1-ms bins were interpolated linearly. Then the STRAIGHT spectrum was recovered by inverse warped-DCT and exponentiation (The previous study [16] employed approximate inversion of the logarithmic function; the current algorithm employs precise inversion.) The speech was resynthesized using the pulse/noise generator and the F0 and voicing information extracted with the auditory model in Fig. 1.

Table I, RMS error in dB for closed and open data for variation of the speaker and MRA.

	Male (MHT)		Female (FTK)	
	Linear MRA	Nonlinear MRA	Linear MRA	Nonlinear MRA
closed	-15.9	-16.8	-15.4	-16.7
open	-15.0	-15.7	-14.4	-15.6

2.3.5. Nonlinear MRA

We used the following nonlinear MRA model for y_j,

$$y_k = \sum_{i=1}^{p} \{\phi_i + \pi_i \exp(-\gamma \bar{x}_k^{\,2})\} x_{ki} + \varepsilon_k \qquad (3)$$

where ϕ_i, π_i, and γ are model parameters, x_{ki} is the ith component in the vector for the kth MI, $\bar{x}_k$ is the average value of x_{ki} for all i, and ε_k is an error term. This formulation reduces to linear MRA when $\pi_i = 0$.

In the original paper on nonlinear auto-regressive analysis[15], the maximum likelihood (ML) estimate is shown to be approximated by the least squared error (LSE). So, we used the LSE for estimating parameters ϕ_i and π_i when γ is a constant. In this case, matrix algebra can be used to solve the problem without iteration. The equation for all data is

$$Y = X\beta + \varepsilon \qquad (4)$$

$$\beta = (\phi_1, \pi_1, \phi_2, \pi_2, \ldots, \phi_p, \pi_p)^T \qquad (5)$$

$$\mathbf{x}_k = \left(x_{k1}, x_{k1}\exp(-\gamma\,\bar{x}_k^{\,2}), \ldots, x_{kp}, x_{kp}\exp(-\gamma\,\bar{x}_k^{\,2})\right)^T \qquad (6)$$

$$X = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N)^T \qquad (7)$$

$$Y = (y_1, y_2, ..., y_N)^T. \qquad (8)$$

The parameters are estimated using LSE as

$$\hat{\beta} = (X'\,X)^{-1} X'\,Y \qquad (9)$$

which is the same formalization as linear MRA. This is an important advantage of this model.

It is, however, necessary to determine the constant, γ, in advance. Following the method used in the nonlinear AR model [13], we determined γ using $\varepsilon_\gamma = 0.00001$, the maximum of the average value $\bar{x}_k$, and a factor A_γ.

$$\gamma = -A_\gamma \cdot \ln \varepsilon_\gamma / \max_{1 \le k \le N}(\bar{x}_k{}^2) \qquad (10)$$

The degree of the fit depends on the value of A_γ. So, we varied A_γ and re-estimated the parameters to find the model that minimized the error.

3. EXPERIMENTS

3.1. Data and conditions

We used male (MHT) and female (FTK) speech from an ATR database of 503 sentences to estimate and evaluate the mapping function. The sampling rate for STRAIGHT and the warped-DCT was set to 12 kHz to match the frequency range (100 Hz - 6kHz) of the auditory filterbank of the MI where the sampling rate was 20 kHz. The number of the data, N, used to estimate the mapping parameters was 10000 which corresponds to about 16 sentences (8 male and 8 female). The vector length was 21 for *h* values of between 0 and 5 in the MI. So, the length of the explanation vector $\mathbf{x}_k$ for one MI in Eq. 6 was 42 for the nonlinear case and 21 in the linear case where the response variable, y_k, is a scalar.

3.2. Error in the warped-DCT domain

The resulting mapping function was evaluated in the warped-DCT domain. Table I shows the rms error between the original and mapped warped-DCT coefficients for the sentences used in the parameter estimation (closed, sentences: MHT_A01 for male and FTK_A01 for female) and in the test (open, MHT_A50 and FTK_A50). The error unit is dB, that is, the relative value of the rms amplitude for all of the warped-DCT coefficients. Larger negative values indicate a better fit. The nonlinear parameter, A_γ, was not particularly sensitive and so it was set to 1.

The nonlinear MRA is always effective; the improvement is between -0.7 and -1.3 dB. The rms errors for the closed data are better than the errors for the open data, but the difference is a maximum of 1.1 dB. The differences between male and female sentences are less than 0.1 dB for nonlinear MRA and 0.6 dB for linear MRA. The results show that the mapping function is sufficiently general to accommodate the variation between sentences and speakers in this data set.

The error values are much smaller than in the previous report where they were between -7.6 and -13.8 dB [16,17]. Moreover the differences between conditions are smaller than in the previous results. Both of these improvements are due to the introduction of the event-synchronous processing for the MI.

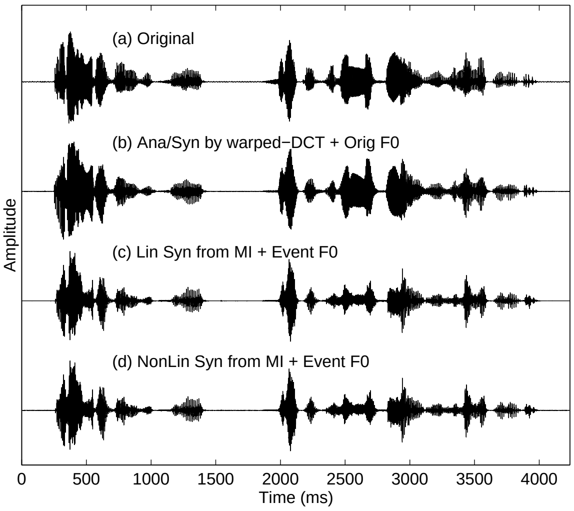

Figure 3. (a) Waveform of the sound MHT_A01. (b) Simple analysis/synthesis sound using the warped-DCT coefficients of 0th-30th order derived from a STRAIGHT smoothed spectrum. (c) Resynthesized sound from the mapped warped-DCT coefficients from the Mellin Image (MI) using linear MRA, and (d) using nonlinear MRA.

3.3. Waveform and sound quality

Figures 3(a) and 3(b) show the waveforms of the original sound, MHT_A01, and the sound resynthesized from the STRAIGHT spectrum with the warped-DCT decomposition (0th - 30-th order). The waveforms are very similar. The sound quality is virtually the same as that produced by STRAIGHT without warped-DCT decomposition; the difference would probably be inaudible over loud speakers.

Figures 3(c) and 3(d) show the waveforms of the sounds resynthesized from the MI using linear and nonlinear MRA. The waveforms are quite similar to those in Fig. 3(b) and the sound quality is good but it is not as close to the original as the STRAIGHT sound with warped-DCT decomposition. There is almost no difference between the versions of MI sound produce with linear and nonlinear MRA. The event-synchronous method successfully eliminates the intrusive clicks reported previously [16,17]. So, the new resynthesis framework works reasonably well.

The difference between the STRAIGHT and MI resynthesis is caused partly by errors in the mapping function and partly by inaccuracies in the event detection, both of which would appear to be amenable to improvement by standard techniques.

4. APPLICATIONS

The discussion to this point has concentrated on the processing and evaluation of the Auditory Vocoder. This section considers potential applications for perceptual experiments and signal processing.

4.1. Perceptual experiments

The current system segregates glottal event information and spectral information, and then combines them for playback. It is possible to resynthesize with a click train composed of impulses located at the event times. Although the click train contains no spectral information, the sentence is recognizable after having heard the original speech once. Shannon et al. [18] have performed psychophysical experiments on temporal cues in speech which relate recognition performance to the degree of spectral information preserved in resynthesis. They used bandpass filters to shape the spectra. Since the warped-DCT is an orthogonal transform, it is possible to manipulate spectral information more systematically than with simple bandpass filters. It would be interesting to compare human perception with the auditory representation in this type of psychophysical experiment.

4.2. Speech processing applications

One long standing problem in speech-synthesis research is quantitative evaluation of the sound quality. The problem arises not only with text-to-speech synthesis, but also with audio coding, and noise suppression. The current method may enable us to evaluate the synthesis of all such applications within a unified framework.

5. CONCLUSIONS

An Auditory Vocoder is proposed to resynthesize sound from the auditory Mellin Image using STRAIGHT. The procedure circumvents the iterative process required in conventional auditory resynthesis. The sound quality is much improved by introducing event-synchronous processing. It may be possible to use the system to implement auditory forms of speech morphing, noise suppression and stream segregation.

Acknowledgments The authors wish to thank S. Katagiri for research support, and H. Kato for introducing us to nonlinear AR models. The project is partially supported by CREST of JST and the Medical Research Council of the U. K.

References

[1] Dudley, H., "Remaking speech," J. Acoust. Soc. Am., 11, pp.169-177, 1939.

[2] Kawahara, H., Masuda-Katsuse, I. and de Cheveigne, A. "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 Communication, 27, pp.187-207, 1999.

[3] Davis, S. B. and Mermelstein, P. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust., Speech, Signal Processing, ASSP-28, 357-366, 1980.

[4] Imai, S. "Cepstral analysis synthesis on the mel frequency scale," IEEE Int. Conf. Acoust., Speech Signal Processing (ICASSP-83), 93-96 , 1983.

[5] Patterson, R.D. "A pulse ribbon model of monaural phase perception," J. Acoust. Soc. Am., 82, 1560-1586, 1987.

[6] Irino, T. and Patterson, R.D. "A time-domain level-dependent auditory filter: the gammachirp," J. Acoust. Soc. Am., 101, pp.412-419, 1997.

[7] Irino, T. and Unoki, M.," An analysis/synthesis auditory filterbank based on an IIR implementation of the gammachirp," J. Acoust. Soc. Jpn., 20, 397-406, 1999.

[8] Slaney, M. "Pattern Playback from 1950 to 1995," IEEE Conf. Syst. Man, Cyben., Vancouver, Canada, 1995.

[9] Irino and Patterson, "Stabilised wavelet Mellin transform: An auditory strategy for normalising sound-source size," Eurospeech'99, Budapest, Hungary, 1999.

[10] Irino, T. and Patterson, R.D. "Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The Stabilised wavelet-Mellin transform," Speech Communication, 36 (3-4),pp.181-203, March 2002.

[11] Patterson, R.D., Allerhand, M. and Giguere, C. "Time-domain modelling of peripheral auditory processing: a modular architecture and a software platform", J. Acoust. Soc. Am., 98,1890-1894, 1995.

[12] http://www.mrc-cbu.cam.ac.uk/personal/roy.patterson/aim/

[13] Meddis, R., O'Mard, L. P., and Lopez-Poveda, E. A., "A computational algorithm for computing nonlinear auditory frequency selectivity,"J. Acoust. Soc. Am.,109, 2852-2861, 2001. http://www.essex.ac.uk/psychology/hearinglab/dsam/

[14] Strube, H. W.," Linear prediction on a warped frequency scale," J. Acoust. Soc. Am., 68, 1071-1076, 1980.

[15] Haggan, V. and Ozaki, T., "Modeling nonlinear random vibration using an amplitude-dependent autoregressive time series model", Biometrika, 68, 189 - 196, 1981.

[16] Irino, T., Patterson, R.D. , Kawahara, H., "Sound resynthesis from Auditory Mellin Image using STRAIGHT," CRAC workshop, Aalborg, Denmark, Sept. 2001.

[17] Irino, T., Patterson, R.D. , Kawahara, H., "Auditory VOCODER: Speech resynthesis from an auditory Mellin representation," IEEE ICASSP2002, May 2002.

[18] Shannon, R.V., Zeng, F-G, Kamath, V., Wygonski, J, and Ekelid, M., "Speech recognition with primarily temporal cues," Science, 270, pp.303-304, 1995.

[19] Slaney, M., Covell, M., and Lassiter, B.," Automatic audio morphing," IEEE Int. Conf. Acoust., Speech Signal Processing (ICASSP-96), Atlanta, GA.