Proceedings of 20th International Congress on Acoustics, ICA 2010

23–27 August 2010, Sydney, Australia

Analysis and synthesis of singing with hoarse vocal expressions

Hideki Kawahara (1), Hanae Itagaki (1), Yoshika Wada (1), Masanori Morise (2)
Ryuichi Nisimura (1) and Toshio Irino (1)

(1) Department of Design Information Sciences, Wakayama University, Wakayama, Japan
(2) Collage of Information Science and Engineering, Ritsumeikan University, Kusatsu, Japan

PACS: 43.75.Rs, 43.75.Zz, 43.70-h, 43.72.Ar, 43.72.Ja, 43.66.Jh

ABSTRACT

Strong vocal expressions in singing use hoarse voice effectively in various manners. However, analysis and synthesis of such voice quality have been a challenging topic with virtually little success. An excitation structure extraction framework called XSX was introduced to represent such complex structured vocal excitation with various types of aperiodicity as an integral component of TANDEM-STRAIGHT, a widely used speech analysis, modification and resynthesis framework. TANDEM-STRAIGHT is basically a source-filter model extended by introducing temporally stable power spectral representation for periodic signals and F0 adaptive spectral envelope estimation based on the consistent sampling theory. The excitation source signal used in TANDEM-STRAIGHT is a mixture of pulses and colored random signals. The source signal parameters are extracted by XSX and an aperiodicity extraction procedure. XSX is based on spectral division and inverse Fourier transform of power spectra by their spectral envelopes, which were calculated for a set of periodicity candidates. Combining salience scores for each candidate yields an integrated measure to detect locally periodic components. The aperiodicity extraction procedure is based on long-range linear prediction of band-pass signals by a set of Quadrature Mirror filters applied to the original and the time-warped signals. This data-driven approach enables to extract and represent complex excitation structures such as diplophonia. The analysis results are used to design voice excitation source, which is capable of adding/modifying hoarse vocal expressions and enables morphing between two or more expressive performance examples.

INTRODUCTION

Non-periodic voices play indispensable roles in expressive speech, traditional theatrical performance, various types of singing and other vocal activities. This article introduces applications of a new method for analysis and representation of such complex voices. The method is called XSX (eXcitation Structure eXtractor) (Itagaki et al. 2009). XSX is an integral part of a speech analysis, modification and synthesis system called TANDEM-STRAIGHT (Kawahara et al. 2008) and was successfully applied to analyze Noh voice (Fujimura et al. 2009), which is a Japanese traditional theatrical performance.

“Hoarse” voice covers a wide range of vocal expressions. Numbers of objective measures were used to quantify “hoarseness,” such as jitter, shimmer, spectral tilt fluctuations and broadband noisy components. But, they are primarily for descriptions. This article, on the contrary, does not try to apply such existing measures to describe “hoarse” vocal expressions. Instead, it tries to generate, replicate and manipulate such expressions. It introduces a rich structured signal representation, XSX, to perceptually precisely replicate “hoarseness.” This precise replication is necessary to apply it for an exemplar based approach by using speech morphing as a tool for exploratory investigations.

The following sections introduce XSX with background explanation on TANDEM-STRIAGHT. Then various singing voices having complex excitation structure are analyzed and visualized using XSX. Finally, discussions and demonstrations of manipulation are presented.

BACKGROUND: TANDEM-STRAIGHT

TANDEM-STRAIGHT and its precursor STRAIGHT (Kawahara et al. 1999) are both based on a simple concept that periodic excitation of voiced sounds is a built-in sampler of underlying smooth time-frequency representations. In this section, only TANDEM-STRAIGHT is outlined, because it supersedes STRAIGHT practically as well as theoretically.

TANDEM-STRAIGHT decomposes input speech into three sets of representations, source information, spectral information and aperiodicity information. These representations are, if necessary, modified and used to synthesize processed speech sounds. Focus of this article is on the source information representation by XSX. It uses two key ideas, which are the bases of spectral information extraction in TANDEM-STRAIGHT. One is temporally stable power spectral representation (Morise et al. 2007) and the other is F0 adaptive spectral smoothing based on consistent sampling (Unser 2000). Refer appendix for details.

Temporally stable power spectral representation

Provided that the spectral representation of a time windowing function only (effectively) covers two harmonic components of a periodic signal, the power spectrum of the signal consists of sinusoidally varying components. Because the fundamental period of this variation is equal to reciprocal of the fundamental frequency of the periodic signal, averaging two (short term) power spectra calculated half fundamental period apart cancels out the sinusoidal temporal variations. It is the underlying idea of calculating temporally stable power spectrum. This algorithm is named TANDEM (Morise et al. 2007) and the resultant

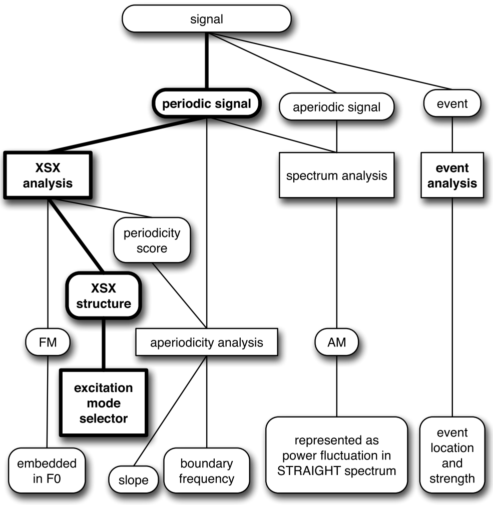

Figure 1: Schematic diagram of excitation information

spectrum is called TANDEM spectrum.

F0 adaptive spectral smoothing

TANDEM reduces the original two-dimensional time-frequency smoothing problem to a one-dimensional discrete to analog conversion problem (Kawahara et al. 2008). By using a rectangular spectral smoother having F0 (fundamental frequency) for its width, periodic spectral variations due to the temporally periodic excitation are completely removed. Excessive spectral smearing caused by combination of this spectral smoothing and spectral representation of the time window function is compensated using the (spectral) compensating digital filter, which is designed based on the consistent sampling theory. The resultant spectrum is called STRAIGHT spectrum.

F0 AND EXCITATION STRUCTURE EXTRACTION

Since the STRAIGHT spectrum mentioned above only consists of spectral envelope information and the TANDEM spectrum that is used to calculate the STRAIGHT spectrum consists of both envelope and periodicity information in a multiplicative manner, dividing the TANDEM spectrum by the STRAIGHT spectrum yields periodicity information and a constant bias. This resultant spectrum after removing the constant bias is called periodicity spectrum. This forms the basis of XSX.

Specialized periodicity detector

Fourier transform of the periodicity spectrum has a dominant peak at the fundamental period. However, designing this detector requires F0 information in advance, but it is not always possible. This apparent contradiction is resolved by introducing a weighting function that covers lower harmonic components prier to calculate Fourier transform. This preprocessing makes the peak height represent (approximate) salience of the input periodicity around assumed F0 for calculating TANDEM and STRAIGHT spectra. In other words, this is a periodicity detector specialized to the assumed F0.

Integration of specialized detectors

F0 extractors have to detect periodicity spanning from 40 Hz to 800z Hz typically (de Cheveigné and Kawahara 2002). Integration of specialized detectors is necessary to fulfill this requirement. This integration is implemented in two steps. The first step is to calculate individual salience functions using specialized detectors by assuming F0 candidates equidistance on the logarithmic frequency axis. The second step is to combine individual salience functions shaped by using a unimodal weighting function.

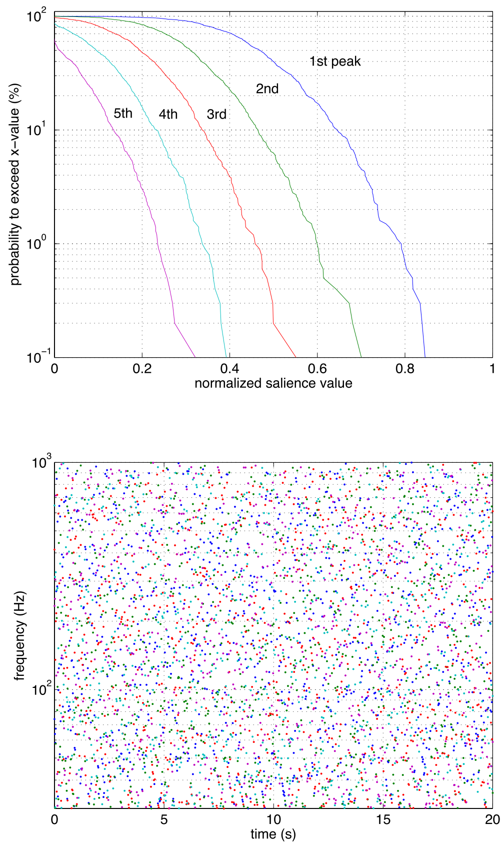

Figure 2: White noise input. Upper plot represents probability of each salience peak to exceed values indicated by the horizontal axis. Lower plot shows extracted peak frequencies

The frequency allocation in the first step is based on the fact that the frequency response of this salience function is proportional to the assumed F0. In other words, salience functions of detectors designed by using different assumed F0 have an identical shape when represented on the logarithmic frequency axis. This logarithmic equidistance allocation of salience functions makes them overlap equally.

The weighting in the second step is to refine approximation of salience function, because the (approximate) salience function calculated by Fourier transform of the weighted periodicity spectrum has spurious peaks. The weighting on the logarithmic frequency axis is to suppress those spurious peaks. Integrating these refined salience function yields the integrated salience function which covers desired possible F0 (and other periodic component) range. Refer appendix for details.

XSX: excitation structure extractor

Each specialized periodicity detector covers a focused region spanning less than one octave. The integrated salience values outside this overlapping region are mutually independent. Therefore, detecting multiple peaks of the integrated salience function enables detection of local periodicity other than F0. Analysis results of XSX in each analysis frame are a set of these peak frequencies of local periodicity peaks of the salience function and the corresponding salience values.

Natural speech sounds deviate from the mathematical definition of periodic signals in various aspects. XSX extracts specific aspects of such deviations. Figure 1 summarizes those excitation related variations and processing components in TANDEM-STRAIGHT system. Components represented using thick lines in Fig. 1 are focus of this article.

ANALYSIS OF KNOWN SIGNALS

To illustrate behavior of XSX, known signals were analyzed. The first test signal is white noise. This provides “background noise level” of the extracted salience value.

White noise

Figure 2 shows results extracted from a white noise signal. The upper plot shows probability of each peak value to exceed the normalized salience value indicated on the horizontal axis. Probabilities corresponding to the first five prominent peaks in each frame are plotted using different lines. The lines are color corded according to the order of salience. For example, the primary peak of salience (blue line) exceeds 0.8 with less than 1% probability. Note that it is virtually impossible for random component to have secondary (dark green line) or weaker peaks larger than 0.7 in terms of normalized salience (less than 0.1% even for the secondary peak).

The lower plot shows corresponding frequencies of salience peaks. The frequency is calculated by parabolic interpolation using three salience values including the peak and its neighboring detector outputs. The dot color represents the order of salience. The color-coding scheme is the same as the upper plot. However, color is not important in this case. Important point is that peak frequency distribution is uniform in terms of logarithmic frequency.

Pulse location modulation

The second test signal models one aspect of diplophonia. This signal has three repetition rates; the first one is usual fundamental frequency corresponds to each interval franked by pulses of both sides. The second one is the repetition rate of paired intervals, coupling short and long adjacent intervals forms a larger unit. The last one is displaced each interval.

Every other pulse of a 200 Hz pulse train is gradually increased its displacement in this test signal. Figures 3 and 4 show XSX analysis results. The upper plot shows salience values of extracted first five large peaks as multiple time series. The lower plot shows frequencies with the same color scheme as the upper plot. Annotations are added to make plots legible when printed in black and white format. By comparing blue dots in both plots, it is indicated that the original F0 (200 Hz) dominates in the initial 0.3 s. Afterwards, two neighboring pulse intervals are combined to form the larger unit, which corresponds to the subharmonic frequency (100 Hz). The upper plot shows turn over of dominance in salience. The initial green line changes its color when it crosses with the initial blue line around 0.3 s.

Figure 4 shows magnified view of the same results. The frequency plot indicates that there are two distinctly different interval lengths, which correspond to the long and the short intervals resulted by the pulse displacement. By comparing the upper and the lower plot, it is indicated that contiguously aligned green dots corresponds to local peak region of salience green dots. Green dots in upper plot have dips around frequency transition in the lower plot. It also should be noted that those peak salience values of green dots are close to 0.7, indicating that they are very unlikely caused by random fluctuation of the input signal.

The trajectory split around 200 Hz found in Fig. 3 also suggests that there are two different intervals which form larger unit

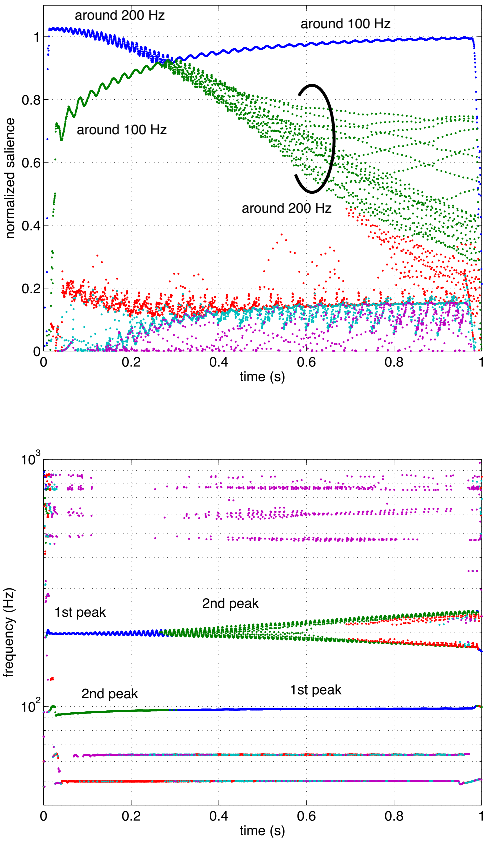

Figure 3: Location modulated pulse train. (upper) salience of extracted peaks and (lower) their peak frequencies

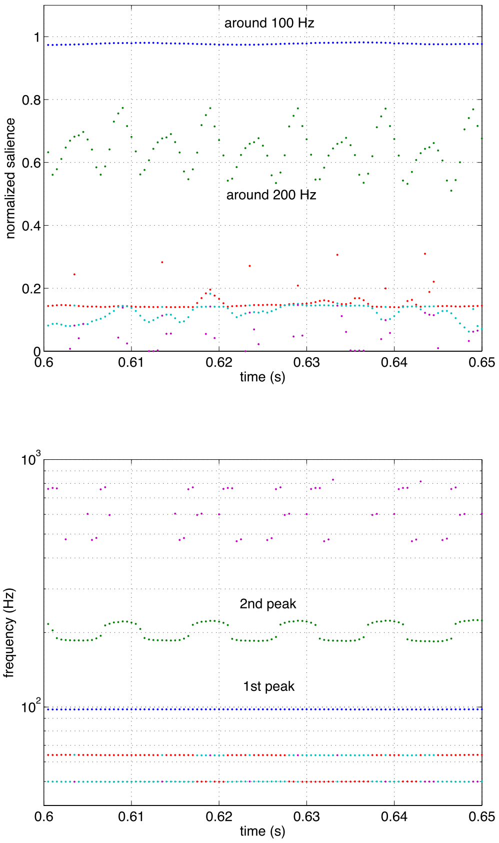

Figure 4: Magnified views of pulse displaced signal results

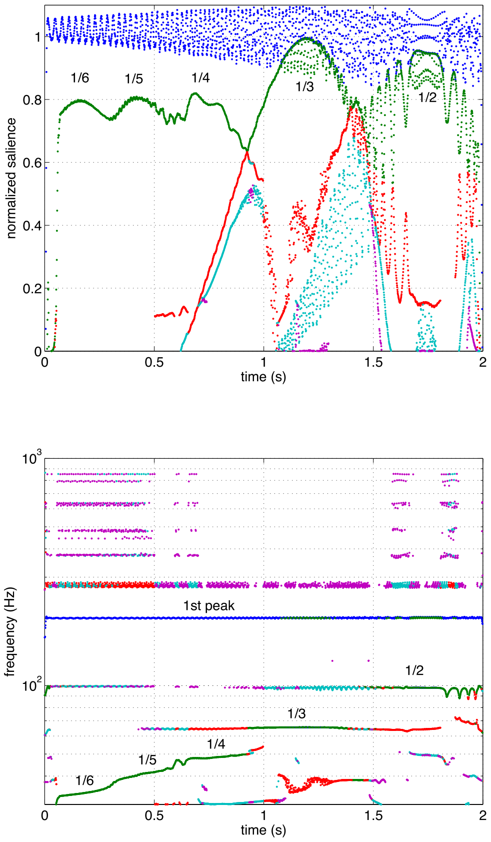

Figure 5: AM complex signal with nonlinearity. Note that salience peaks of second place correspond to where the original F0 is integer multiple of candidate frequencies. Annotation indicates ratio $1/N$ to the original F0.

intervals.

AM complex signal with nonlinearity

The third test signal models hoarse voice by applying AM and asymmetric nonlinearity on a complex sound.

The instantaneous frequency $f_m(t)$ of the amplitude modulation signal $a_m(t)$ is exponentially increase starting from the initial value f_M.

$$f_m(t) = f_M \exp(g_e t), \tag{1}$$

where g_e is the growth rate of the instantaneous frequency.

The instantaneous amplitude $a_m(t)$ is calculated as follows and yields simpler form by substituting the definition of f_m.

$$\begin{aligned} a_m(t) &= 1 + \beta_m \sin\left(2\pi \int_0^t f_m(\lambda) d\lambda\right) \\ &= 1 + \beta_m \sin\left(\frac{2\pi f_M(\exp(g_e t))}{g_e}\right), \end{aligned} \tag{2}$$

where β_m represents the modulation depth of AM complex signal $x(t)$ defined below.

$$x(t) = a_m(t) \sum_{k=1}^{M} \alpha_k \sin(2\pi f_0 t), \tag{3}$$

where f_0 represents the fundamental frequency.

Finally, this complex signal is distorted using the following nonlinearity to yield the test signal $y(t)$.

$$y(t) = x(t) + cx^2(t), \tag{4}$$

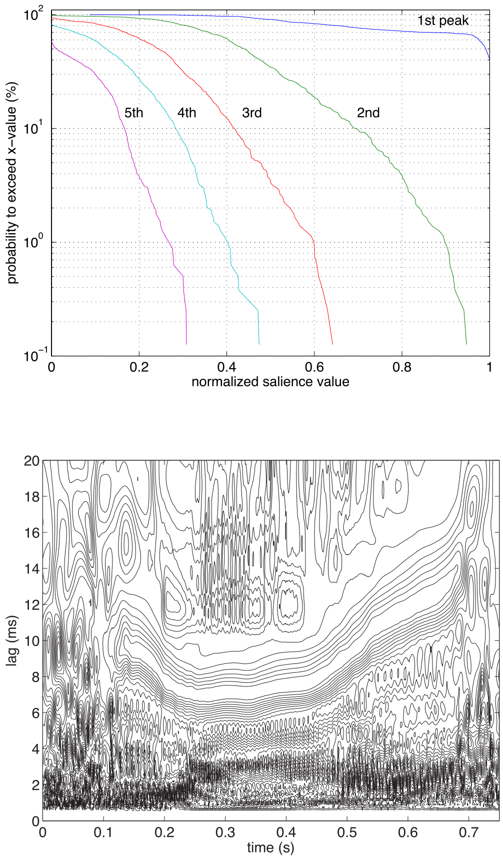

Figure 6: Natural Japanese vowel sequence /aiueo/ spoken by a male speaker. Upper plot represents probability of each salience peak to exceed values indicated by the horizontal axis. Lower plot shows contour map of salience value. The ridge shows where fundamental is

Figure 5 shows results extracted from the test signal $y(t)$ generated using $f_0 = 200$ Hz. The amplitude modulation frequency is exponentially increased. The modulation frequency is 33 Hz, 40 Hz, 50 Hz, 67 Hz and 100 Hz around 0.2 s, 0.4 s, 0.7 s, 1.2 s and 1.7 s respectively. The salience value has peaks around those locations indicating there exist larger temporal structure grouping several fundamental intervals.

ANALYSIS OF NATURAL SPEECH AND SINGING

This section presents two examples of natural speech analysis. Ordinary speaking voice and singing voice with “hoarse” expression.

Ordinary speaking voice

The first one is an ordinary voice speaking a Japanese vowel sequence /aiueo/ by a male speaker. Figure 6 shows the results. The upper plot illustrates salience peak value distribution. The distribution is clearly different from that of white noise shown in Fig. 2, indicating that speech is not random and has a (or several) periodic structure(s).

The lower plot shows salience value map represented using contour plot. The vertical axis represents lag and the horizontal axis represents time. The horizontally extended “U” shaped ridge in the middle corresponds to the fundamental component. This suggests that salience value is a robust clue to extract fundamental component. Peak picking (and following parabolic interpolation) along the frequency axis provides the salience peak plots in the previous figures.

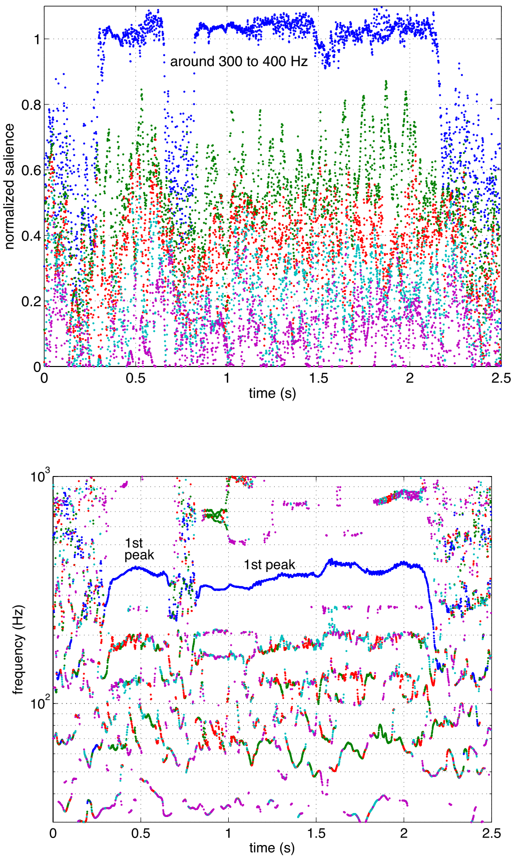

Figure 7: Professional singing voice with hoarse timbre. (Upper) normalized salience. (Lower) peak frequencies

Singing voice with hoarse expression

The second example is a singing voice recorded for CrestMuse project (CrestMuse last visited: 19 May 2010). Original J-POP songs were composed to clear copyright issues to use professional singing data in research. A song titled “Ride” for male singers is selected for the analysis in this article. The recording was taken place in one of professional recording studio with a professional Japanese pop singer. Recorded data was converted to WAVE format (44,100 Hz sampling frequency and 16 bit resolution) to be analyzed using Matlab. A fragment of recording that consists of typical hoarse expression was excerpted. The length of the excerpt was 2.5 s. The lyrics fragment of the excerpt is /kiesou na yume/ (... dreams to fade away ... in English). The final vowel /e/ has strong “hoarseness” in voice quality.

Figure 7 shows the analysis results. The upper plot shows salience values. Note that secondary peak values frequently exceed 0.7. This suggests that these secondary peaks are not caused by random fluctuations, indicating that there are other periodic components than the fundamental component. By inspecting the peak frequencies in the lower plot, no clear split of frequency trajectory is found. This may suggest that the hoarseness is due to AM rather than FM.

Figure 8 shows magnified view around 1.85 s, where strong hoarseness is perceived. The upper plot clearly indicates that the secondary peak represents the other strong periodic structure. (Because the values stay higher than 0.6 and sometimes exceed 0.8.) By comparing with the lower frequency plot, the secondary peaks (green dots) changes contiguously from 1.83 s to 1.89 s in the upper plot. This contiguous (green dots) trajectory is also found as a contiguous (green dots) trajectory around 70 Hz to 80 Hz in the peak frequency plot. This additional periodicity corresponds to 1/5 subharmonic of the primary F0. It also should be noted that salience variations behave similarly to the simulation results shown in Fig. 5. Please inspect around 0.4 s of Fig. 5 and 1.87 s of Fig. 8.

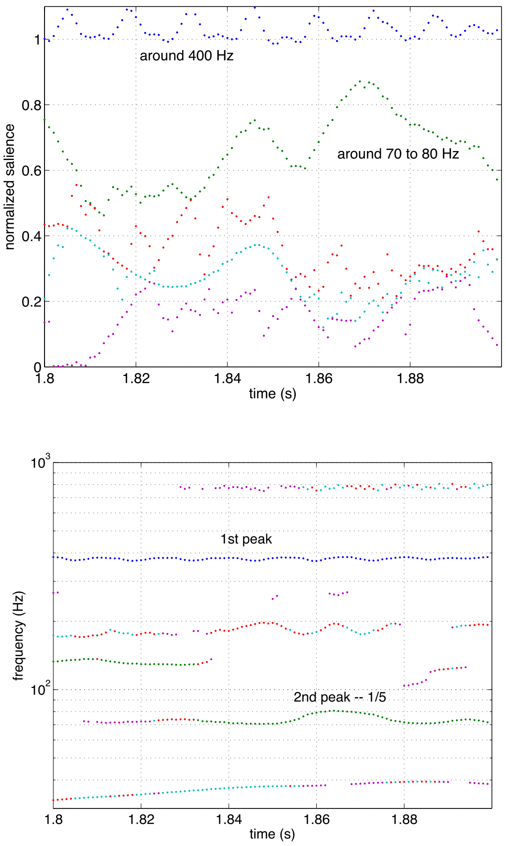

Figure 8: Magnified views of professional singing voice with hoarse timbre. (Upper) normalized salience. (Lower) peak frequencies

SYNTHESIS AND MANIPULATIONS

The last analysis results are used to resynthesize hoarse singing voice. Although it is successfully applied to resynthesis, further parameterization of modulation characteristics and types are necessary for flexible manipulation possible.

CONCLUSION

A framework for analysis and synthesis with flexible manipulation of hoarse, husky, creaky and other non-periodic voice quality is introduced based on a structural excitation extractor (XSX) designed for a sophisticated channel VOCODER TANDEM-STRAIGHT. Analysis results obtained using known synthesized test signals illustrated how excitation structures are extracted using XSX. Natural speech analysis results indicated that XSX provides rich information useful to replicate strong vocal expressions. Subjective sound quality tests and automation and parameterization of excitation structure extraction procedure are currently under study.

Acknowledgement Constructive comments by the reviewers are appreciated by the authors and were useful to improve readability. This study is partially supported by Grants-in-Aid for Scientific Research (A) 19200017 by JSPS and the Crest-

Muse project by JST.

REFERENCES

H. Akagiri, M. Morise, R. Nisimura, T Irino, and H. Kawahara. Evaluation and optimization of F0-adaptive spectral envelope estimation based on spectral smoothing with peak emphasis. In *Proc. ICA 2010*, 2010. (this proceeding).

CrestMuse. http://www.crestmuse.jp/index-e.html, last visited: 19 May 2010.

A. de Cheveigné and H. Kawahara. YIN, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.*, 111(4):1917–1930, 2002.

O. Fujimura, K. Honda, H. Kawahara, Y. Konparu, and M. Morise. Noh voice quality. *J. Logopedics Phoniatrics Vocology*, 34(4):157–170, 2009.

H. Itagaki, M. Morise, R. Nisimura, T Irino, and H. Kawahara. A bottom-up procedure to extract periodicity structure of voiced sounds and its application to represent and restoration of pathological voices. In *Proc. MAVEBA 2009*, pages 115–118, 2009.

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. *Speech Communication*, 27(3-4):187–207, 1999.

H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno. A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0 and aperiodicity estimation. In *Proc. ICASSP 2008*, pages 3933–3936. IEEE, 2008.

M. Morise, T. Takahashi, H. Kawahara, and T. Irino. Power spectrum estimation method for periodic signals virtually irrespective to time window position. *Trans. IEICE*, J90-D (12):3265–3267, 2007. [in Japanese].

M. Unser. Sampling – 50 years after Shannon. *Proceedings of the IEEE*, 88(4):569–587, 2000.

APPENDIX

TANDEM AND STRAIGHT SPECTRA

Let $x(t)$ represent a periodic signal having a fundamental period $T_0 = 1/f_0$ and $w(t)$ represent the windowing function. TANDEM spectrum $P_T(\omega,t)$ of the windowed signal centered around time t is calculated using the following equation.

$$P_T(\omega,t) = \frac{1}{2}\left(P(\omega,t-\frac{T_0}{4}) + P(\omega,t+\frac{T_0}{4})\right), \quad (5)$$
$$\text{where} \quad P(\omega,t) = \left|\frac{1}{\sqrt{2\pi}}\int w(\tau)x(t-\tau)e^{-j\omega\tau}d\tau\right|^2.$$

In the current implementation, a Blackmann window with the window length $2.5T_0$ is used.

STRAIGHT spectrum $P_{ST}(\omega,t)$ is calculated form this TANDEM spectrum by using the following set of equations.

$$P_{ST}(\omega) = \exp\left[L(\omega) + \tilde{q}(L(\omega - 2\pi f_0) + L(\omega + 2\pi f_0))\right], \quad (6)$$
$$\text{where} \quad L(\omega) = \log\left[P_S(\omega)\right],$$
$$P_S(\omega) = P_C(\omega + \pi f_0) - P_C(\omega - \pi f_0),$$
$$P_C(\omega) = \frac{1}{2\pi f_0}\int_{-3\pi f_0}^{\omega} P_T(\lambda)d\lambda,$$

where $\tilde{q}$ represents the adjusted first coefficient of the compensating digital filter, which is designed based on consistent sampling. Details of adjustment and truncation are presented in our other presentation of ICA2010 (Akagiri et al. 2010). Note that definitions of $L(\omega)$, $P_S(\omega)$ and $P_C(\omega)$ are slightly different in the detailed version given there. Please also note that the time coordinate t is dropped in these equations for readability.

F0 AND EXCITATION STRUCTURE EXTRACTION

Periodicity spectrum $P_P(\omega)$ is defined using the following equation.

$$P_P(\omega) = \frac{P_T(\omega)}{P_S(\omega)} - 1 \quad (7)$$

The (approximate) salience function $r_A(\tau)$ is a function of lag τ and calculated using the following equation.

$$r_A(\tau) = \int w_B(\omega)P_P(\omega)e^{j\omega\tau}d\omega, \quad (8)$$
$$\text{where} \quad w_B(\omega) = \begin{cases} 1+\cos(\frac{\pi\omega}{N\omega_0}) & |\omega| \le N\omega_0 \\ 0 & |\omega| > N\omega_0 \end{cases},$$

where parameter N determines range of harmonic components used to calculate periodicity. Since, this salience function is designed by assuming a specific fundamental frequency, it is better to explicitly represent the assumed frequency using f_c instead of f_0. Let $r_A(\tau; f_c)$ represent the approximate salience function designed using f_c.

The refined salience function $r(\tau; f_c)$ is defined by introducing a symmetric weighting function on the logarithmic frequency.

$$r(\tau; f_c) = w_L(\tau; f_c)r_A(\tau; f_c), \quad (9)$$
$$\text{where} \quad w_L(\tau; f_c) = \begin{cases} 1+\cos(\pi u(\tau)) & |u(\tau)| \le 1 \\ 0 & |u(\tau)| > 1 \end{cases},$$
$$u(\tau) = b_w \log_2(\tau f_c),$$

where b_w represents a parameter that determines sharpness of the salience function around the assumed periodicity f_c.

Integrated salience function

The salience functions defined above have an identical shape on the logarithmic frequency (as well as the logarithmic lag) axes. By placing assumed fundamental frequency f_c evenly on the logarithmic frequency axis, overlap of each salience function with neighboring functions is kept constant irrespective to f_c. This makes simple summation of logarithmically allocated salience functions $r(\tau; f_c)$ yield an integrated salience function $r_I(\tau)$ that covers wide frequency range.

$$r_I(\tau) = c_0 \sum_{f_c \in F_c} r(\tau; f_c), \quad (10)$$

where F_c represents the set of assumed frequencies for specialized detectors. The normalization constant c_0 is defined to make the salience value for periodic pulse train yield one. In our implementation, the assumed frequency $f_c(k)$ of the k-th detector is defined below.

$$f_c(k) = f_L 2^{\frac{k-1}{L}}, \quad (11)$$

where L represents the density of specialized detectors in terms of number of detectors in one octave and f_L represents the assumed frequency of the detector which covers the lowest end of the periodicity detection frequency range. The total number of detectors M is determined by the following equation.

$$M = \lceil L(\log_2(f_U) - \log_2(f_L))\rceil + 1, \quad (12)$$

where $\lceil x \rceil$ rounds x toward positive infinity and f_U represents the assumed frequency of the detector which covers the highest end of the periodicity detection frequency range.

The current examples are calculated using following set of parameters; $N = 4, b_w = 2.9, L = 3$. Optimization details of these parameters will be presented elsewhere.