VOQUAL'03, Geneva, August 27-29, 2003

Exemplar-based Voice Quality Analysis and Control using a High Quality Auditory Morphing Procedure based on STRAIGHT

Hideki Kawahara

Faculty of Systems Engineering
Wakayama University / ATR, Japan
kawahara@sys.wakayama-u.ac.jp

Abstract

This paper tries to introduce a new strategy and tools for voice quality research that complements conventional approaches. A very high-quality speech analysis, modification and synthesis procedure STRAIGHT, which is basically a channel VOCODER based on a pitch-synchronous analysis synthesis framework, was extended to implement auditory morphing in terms of spectral, pitch and voice quality parameters. This extension enables voice quality modification by parametric transformation using STRAIGHT. It also enables an exemplar-based research strategy for perceptual aspects of voice quality analysis and control. In other words, manipulated synthetic voice having virtually equivalent naturalness to *natural* voice introduces a mean to perform a unique research strategy called *systematic downgrading,* that is suitable especially for para and non-linguistic aspects of human vocalization. In addition to morphing procedure, a set of visualization techniques were introduced based on fixed-point analyses in the time and the frequency domain for assisting exploratory data analysis that is indispensable in voice quality research.

1. Introduction

Voice quality is an important attribute that makes speech far richer than a mere acoustic instantiation of its corresponding textual transcription. There has been a large body of investigations on voice quality in terms of speech production and perception. Research strategies employed in these investigations can be grouped into two categories. One is the analytical approach and the other is the synthetic approach. However, it is still a hard problem to control voice quality of synthetic and/or enhanced speech in terms of perceptually relevant control parameters and to provide natural and vivid voice quality to produced speech. This article tries to propose the other alternative, an exemplar-based approach, for voice quality research and applications to provide an complementary methodology to attack this hard problem.

The exemplar-based approach requires a certain set of parameters to be able to precisely reproduce a given example and a mean to traverse between examples in the parameter space. In other words, it requires a high-quality auditory morphing, which was recently implemented based on STRAIGHT [1].

This article explores necessary extensions to the original STRAIGHT for making it applicable to morphing voice quality related parameters and demonstrates its use. It also introduces an event based analysis and visualization technique [2] to facilitate exploratory data analysis in voice quality research.

2. Brief introduction to STRAIGHT

It is worthwhile to briefly introduce STRAIGHT (Speech Transformation and Representation based on Adaptive Interpolation of weiGHTed spectrogram) [3, 4], that was originally designed for speech perception research using a source filter architecture. STRAIGHT is basically a channel VOCODER based on F_0 adaptive procedures. The procedures are grouped into three subsystems; a source information extractor, a smoothed time-frequency representation extractor, and a synthesis engine consisting of an excitation source and a time varying filter. Underlying principles and implementation of the second and the third component are given in the following paragraphs. Those for the first component, source information extractor, which are the central issues in this paper, are described in the following sections.

Separating speech information into mutually independent filter parameters and source parameters is important for flexible speech manipulations. A F_0 adaptive complimentary time window pair and F_0 adaptive spectral smoothing based on a cardinal B-spline basis function effectively remove interferences due to signal periodicity from the time-frequency representation of the signal. The filtering component is implemented as the minimum phase impulse response calculated from the smoothed time-frequency representation through several stages of FFTs. This FFT-based implementation enables source F_0 control with a finer frequency resolution than that is determined by the sampling interval of the speech signal. This implementation also enables suppression of "buzz-like" timbre, which is common in conventional pulse excitation, by introducing group delay randomization in the higher frequency region. However, in previous studies, there was no dependable methodology to extract control parameters of this group delay randomization from the speech signal under study. This paper introduces new procedures to extend the source information extractor and the excitation source of STRAIGHT to solve these problems. The procedures also provide useful means to visualize quality related source parameters [4, 5, 2].

STRAIGHT is capable of resynthesizing various voice quality from a set of parametric representations. The current implementation, that is introduced in this paper, consists of a smoothed time-frequency representation, F0 and a time-frequency aperiodicity index. It indicates that voice quality is embedded in combinations of these parameters. It is possible to correlate these parameters with conventional voice quality related indices. For example, spectral tilt can be calculated from the time-frequency representation and HNR also can be calculated from the aperiodicity map. However, they are summarized indices and are not sufficient to reproduce vivid quality. It is important to find a way to control "vivid"-ness of voice quality. That is our main interest and the reason why an exemplar based approach is proposed.

A new methodology, exemplar based approach and a new research tool, auditory morphing based on STRAIGHT have been introduced. It is important to provide a research strategy to take advantage of these elements. "Systematic downgrading"

is one such strategy we have been adopting.

3. Systematic downgrading

A strategy called systematic downgrading was originally proposed in the context of research on scat singing [6, 7, 8], where non-linguistic and pare-linguistic information plays indispensable roles. The central idea of "systematic downgrading" is to keep test stimuli as ecologically relevant (in other words, highly natural) as possible. It is important to use ecologically relevant stimuli, because human perceptual systems can be highly nonlinear, meaning it is generally difficult to draw dependable conclusions for human responses to highly complex signals (for example speech) only based on responses to elementary stimuli such as tone, tone bursts, clicks, noise and synthetic speech. [1] It is also important to have means to manipulate physical parameters of the stimuli in a well-defined manner. The STRAIGHT-based morphing fulfils requirements on ecological relevance (high quality resynthesized speech) and precise control of physical parameters simultaneously.

The following steps outline "systematic downgrading" in case of investigating regularities in voice quality.
(1) Prepare the reference speech and the target speech having typical voice quality.
(2) Morph the reference speech to the target speech by careful manual transformation of parameters.
(3) Extract regularities in the manual transformation and design series of approximation functions of the transformation.
(4) Morph the reference speech by the approximation functions AND refine it with additional manual modifications.
(5) Repeat step (3) and (4) until satisfactory approximation function is designed.

The procedure is a generalized version of the "null point procedure", which is a common practice to minimize disturbances to the system under study. It keeps the critical subjective evaluation to be performed only for high-quality (ecologically relevant) stimuli. This is especially important in voice quality research, because quality is the most fragile attribute against various kinds of distortions. It also should be noted that the step (3) inevitably is exploratory, even though multivariate analysis may help acquire insights. The first step of such exploratory investigation would be to implement selective morphing. Similar examples can be found in our Eurospeech'30 paper [9] and a literature on emotional speech [10].

4. Auditory morphing

Morphing is a procedure to regenerate a signal from a representation on a shortest trajectory between anchor points in an abstract distance space with a distance metric d_{fx}. When there is no ambiguity, it is also possible to extrapolate the shortest trajectory outside the anchor points.

It is necessary to introduce an approximation that yields practical implementation of this general morphing procedure. One such approximation for speech morphing is to define the new distance d_{cp} [2] as a composite operations of a coordinate transformation $\mathcal{T}$ and a localized distance metric d_{pp}.

$$d_{fx} \simeq d_{cp} = d_{pp}(s_{ref}(\lambda,\tau)), s_{tgt}(\mathcal{T}(\lambda,\tau))) \qquad (1)$$

where ref and tgt represents "reference" and "target" respectively. If the transformation $\mathcal{T}$ does not have any penalty due to the transformation, and if the localized distance metric is Euclidean, the morphing procedure is reduced to a linear interpolation on representations represented on the reference coordinate. The proposed procedure described below is based on this approximation. This procedure is analogous to visual morphing when the time-frequency coordinate and the attributes on the coordinate system are replaced by the shape and color (including intensity and texture).

There are several technical issues to implement the procedure. Specifically, the coordinate system and the localized distance metric must reflect auditory perceptual characteristics, and the transformation must be as simple as possible. In this article, the time-frequency plane is used as the coordinate system. The transformation is represented as a simple piecewise bilinear transformation, because, unlike the image morphing, the time-frequency coordinate is not isotropic. In our preliminary experiences, it was found that for morphing emotional speech samples digitized at 44.1 kHz 16bit, only up to 5 anchor points on a frequency axis at one temporal location and up to 4 temporal anchor points for one CV syllable were sufficient. [3] For the fundamental frequency, it is relevant to morph the parameter in the log-frequency domain, because the F0 dynamics is represented in terms of a linear dynamical equation in the log-frequency domain [11]. For the spectral density, morphing is calculated on dB representation, because it is one of relevant approximations of intensity perception. The time-frequency periodicity index [4, 2] is also transformed by the same mapping function. That is one of important cues for voice quality representation and control that will be introduced in the following section.

4.1. Illustration of the method

Figures illustrate how to define and to use anchor points for auditory morphing based on STRAIGHT. Figure 1 shows anchor points that were manually defined for morphing two speech samples of a word spoken under different emotional context.

The information given by Figure 1 is used to deform time-frequency coordinate system of the target speech to make it aligned with the reference speech. Figure 2 shows how the target uniform time-frequency grid is deformed. By using this deformation, all spectral and source parameters in two end points (two speech examples) are made aligned in the target's time-frequency coordinate system. This alignment makes morphing simple interpolation at each time-frequency location.

5. Source information extraction and control

This section briefly introduces tools for source information extraction using instantaneous frequency and group delay as key concepts[12]. Source information extracted in this stage consists of the F_0 and aperiodicity measures both in the frequency and in the time domain.

The frequency domain parameters, F0 and aperiodicity index are used to extend STRAIGHT to be able to morph voice quality related parameters. However, the time domain parameters are not implemented as a component of STRAIGHT to control voice quality. It is introduced here to provide a tool to visualize source information that is possibly related to voice quality.

[1] Synthetic speech made from formant synthesizers and LPC or low order cepstral vocoders excited by pulse and noise source are too simple to represent variabilities of speech quality.

[2] Strictly speaking, this approximation does not define distance metric. The approximation does not satisfy the requirement for distance metric; $d_{cp}(a,b) = d_{cp}(b,a)$. However, until it is inevitable, this simplified definition is used in this article.

[3] However, it was found by morphing examples provided by VOQUAL'03 that the number of anchor points on the frequency axis sometimes exceeded 8 and for morphing vibrato, there may be a chance that it requires temporal anchor points for each cycle of vibrato.

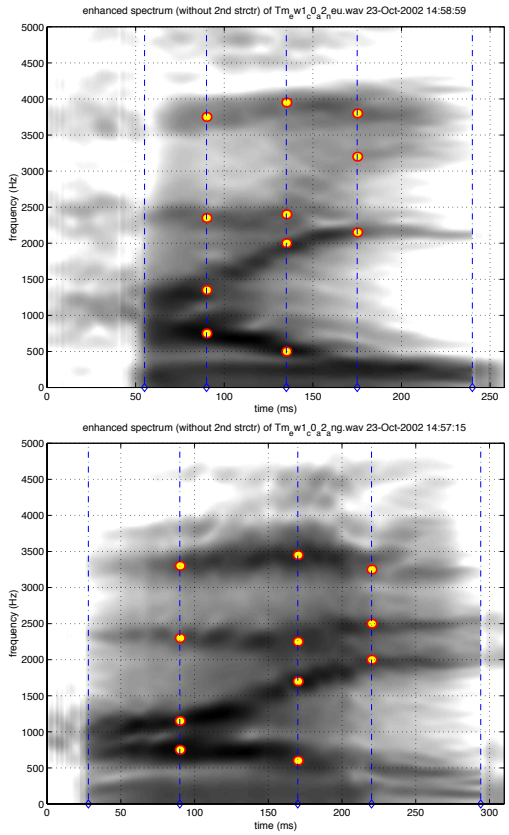

Figure 1: *Smooth spectrographic representations of words played by a male actor under neutral (upper) and angry (lower) emotional conditions. Anchor points in the time-frequency domain are plotted as open circles and temporal anchors are plotted as vertical dash-dot lines. (Lower frequency portion ($\leq$ 5000 Hz) of time-frequency representations are shown to clarify contrasts.) (This figure is the excerpt from our ICASSP paper [1].)*

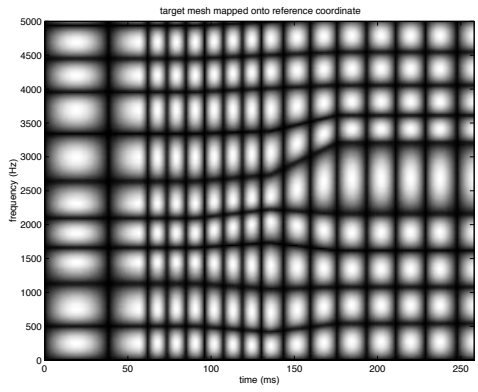

Figure 2: *Regular time-frequency grid in the target coordinate transformed into the reference coordinate. (This figure is the excerpt from our ICASSP paper [1].)*

5.1. Frequency domain analysis

Speech signals are not exactly periodic. F_0s and waveforms are always changing and fluctuating. The instantaneous frequency based F_0 extraction method used in this paper was proposed[4] to represent these nonstationary speech behavior and was designed to produce continuous and high-resolution F_0 trajectories suitable for high-quality speech modifications. The estimation of the aperiodicity measures in the frequency domain is dependent on this initial F_0 estimate, which is based on a fixed point analysis of a mapping from filter center frequencies to their output instantaneous frequencies.

5.1.1. F_0 estimation

The F_0 estimation method of STRAIGHT assumes that the signal has the following nearly harmonic structure.

$$x(t) = \sum_{k=1}^{N} a_k(t) \cos\Big[\int_0^t (k\omega_0(\tau) + \omega_k(\tau))d\tau + \phi_k(0)\Big], \quad (2)$$

where $a_k(t)$ represents a slowly changing instantaneous amplitude. $\omega_k(\tau)$ also represents slowly changing perturbation of the k-th harmonic component. In this representation, F_0 is the instantaneous frequency of the fundamental component where $k = 1$. The F_0 extraction procedure also uses instantaneous frequencies of other harmonic components to refine F_0 estimates.

By using band-pass filters with complex number impulse responses, filter center frequencies and instantaneous frequencies of filter outputs provide an interesting means for the sinusoidal component extraction. Let $\lambda(\omega_c, t)$ be the mapping from the filter center angular frequency ω_c to the instantaneous frequency of filter output. Then, angular frequencies of sinusoidal components are extracted as a set of fixed points Ψ based on the following definition.

$$\Psi(t) = \{\psi \mid \lambda(\psi, t) = \psi, \; -1 < \frac{\partial}{\partial\psi}(\lambda(\psi, t) - \psi) < 0\}. \quad (3)$$

This relation between filter center frequencies and harmonic components were reported by number of authors[13, 14]. Similar relation to resonant frequencies was also described in modeling auditory perception[15]. In addition to these findings, a geometrical properties of the mapping around fixed points was found very useful in source information analysis[4].

The signal to noise ratio of the sinusoidal component and the background noise (represented as C/N: carrier to noise ratio hereafter) is approximately represented using $\frac{\partial\lambda}{\partial\psi}$ and $\frac{\partial\lambda}{\partial\psi t}$. Please refer to [4] for details. Combined with this C/N estimation method, the following nearly isotropic filter impulse response is designed.

$$w_s(t, \omega_c) = (w(t, \omega_c) \odot h(t, \omega_c))\, e^{j\omega_c t}, \quad (4)$$

$$w(t, \omega_c) = \exp(-\omega_c^2 t^2 / 4\pi\eta^2),$$

$$h(t, \omega_c) = \max\left\{0, 1 - \left|\frac{\omega_c t}{2\pi\eta}\right|\right\}, \quad (5)$$

where $\odot$ represents convolution and η represents a time stretching factor, that is slightly larger than 1 to refine frequency resolution (1.2 is used in the current implementation). With a log-linear arrangement of filters (6 filters in one octave), fundamental harmonic component can be selected as the fixed point having the highest C/N. Finally, the initial F_0 estimate is used to select several (in our case, lower three) harmonic components for refining F_0 estimate using C/N and the instantaneous frequency for each harmonic component.

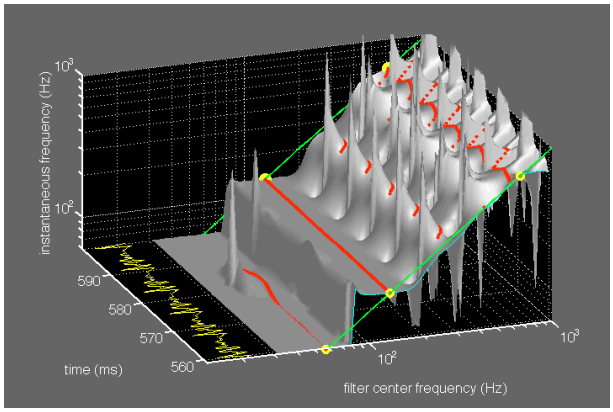

Figure 3: *Three dimensional representation of the filter center frequency to the output instantaneous frequency map. The surface represents the mapping. Red dots represent the fixed point of the mapping. F0 can be found as fixed points on the unique flat surface. Note that other fixed points are not stable in time. Speech material is the sustained Japanese vowel /a/ spoken by a male speaker. Temporal stretch parameter $\eta = 1.1$ was used.*

Figure 3 shows an example to illustrate how the log-linear filter arrangement makes the fundamental component related fixed point salient. It is clearly seen that the mappings stay flat only around the fundamental component.

Figure 4 shows an example of the source information display of STRAIGHT. It illustrates how C/N information is used for finding the fundamental component. C/N information is shown on the top panel and the bottom panel. Please refer to the caption for explanation.

As mentioned in the previous paragraph, this F_0 estimation procedure consists of the C/N estimation for each filter output as its integral part. It is potentially applicable to aperiodicity evaluation. However, application of this procedure to higher harmonic components is computationally excessively expensive. A simple procedure given in the next paragraph is proposed to extract the virtually equivalent information.

5.1.2. Aperiodicity measure

Time domain warping of a speech signal using the inverse function of the phase of the fundamental component makes the speech signal on the new time axis have a constant F_0 and regular harmonic structure[4]. Deviations from periodicity introduce additional components on inharmonic frequencies. In other words, energy on inharmonic frequencies normalized by the total energy provides a measure of aperiodicity.

Similar to Eq. 4, a slightly time stretched Gaussian function, convoluted with the 2nd order cardinal B-spline basis function that is tuned to the fixed F_0 on the new time axis, is designed to have zeros between harmonic components. A power spectrum calculated using this window provides the energy sum of periodic and aperiodic components at each harmonic frequency and provides the energy of the aperiodic component at each in-between harmonic frequency. This enables aperiodicity evaluation to be a simple peak picking of the power spectrum calculated on the new time axis. A cepstral liftering to suppress components having quefrencies greater than F_0 is introduced to enhance robustness of the procedure.

Let $|S_S(\omega)|^2$ represent the smoothed power spectrum on the new time axis. Then, let $|S_U(\omega)|^2$ and $|S_L(\omega)|^2$ represent the upper and the lower spectral envelopes respectively. The upper envelope is calculated by connecting spectral peaks and the lower envelope (bottom line) is calculated by connecting spectral valleys. The aperiodicity measure is defined as the lower envelope normalized by the upper envelope. The bias due to the liftering in the proposed procedure is calibrated by a table-look-up based on the simulation results using known aperiodic signals. The actual aperiodicity measure $P_{AP}(\omega)$ in the frequency domain is calculated as a weighted average using the original power spectrum $|S(\omega)|^2$ as the weight.

$$P_{AP}(\omega) = \frac{\int w_{ERB}(\lambda;\omega)|S(\lambda)|^2 \mathcal{T}\left(\frac{|S_L(\lambda)|^2}{|S_U(\lambda)|^2}\right) d\lambda}{\int w_{ERB}(\lambda;\omega)|S(\lambda)|^2 d\lambda} \quad (6)$$

where $w_{ERB}(\lambda;\omega)$ represents simplified auditory filter shape for smoothing the power spectrum at the center frequency ω. $\mathcal{T}(\ )$ represents the table-look-up operation.

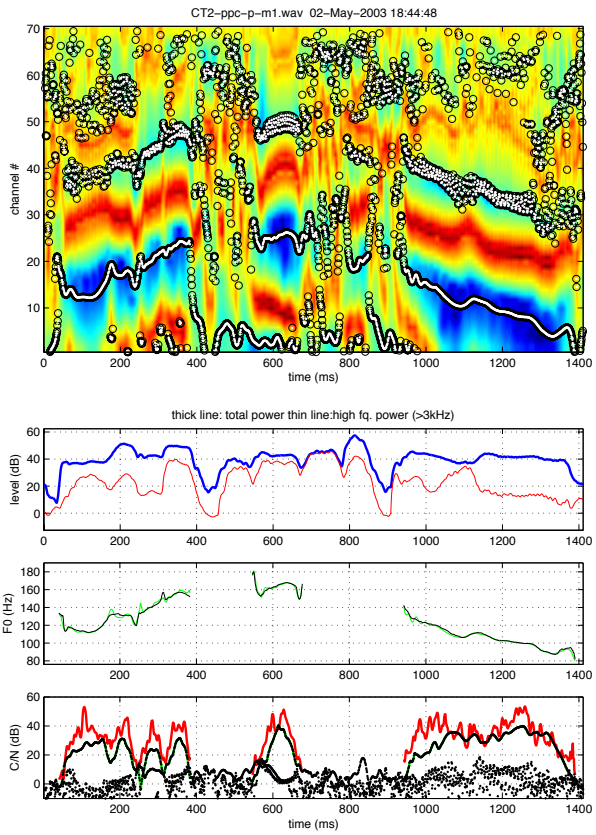

Figure 4: *Extracted source information from the French singing file* `CT2-ppc-p-m1.wav`. *The top panel represents fixed points extracted using a circle symbol with a white center dot. The overlaid image represents the C/N ratio for each filter channel (24 channels/octave center frequency allocation covering from 80 Hz to 600 Hz in this example). The lighter the color the higher the C/N. The middle panel shows the total energy (thick line) and the higher frequency (> 3 kHz) energy (thin line). The next panel illustrates an extracted F_0. The bottom panel shows the C/N ratio for each fixed point. Note that one C/N trajectory is outstanding. It corresponds to the fundamental component.*

5.2. Time domain concentration measure

Signals having the same aperiodicity measure may have perceptually different quality. This difference is associated with the

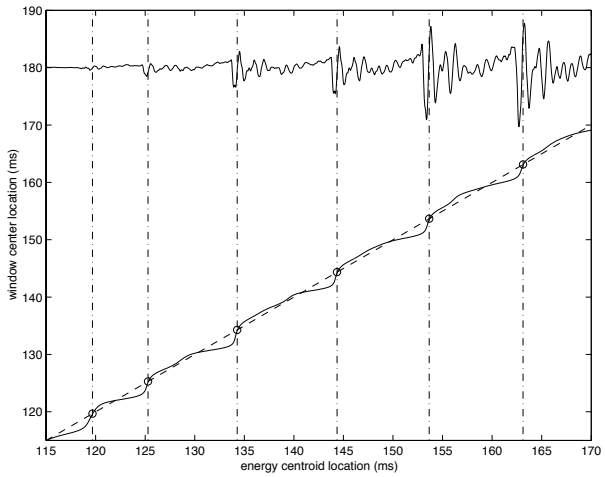

Figure 5: *Time domain event extraction. The original speech waveform is plotted at the top of the figure. The figure shows the onset of a Japanese vowel sequence /aiueo/ spoken by a male speaker. The solid line, which is close to the diagonal dashed line, represents the mapping from the energy centroid to the window center location. Small circles represent the extracted fixed points.*

temporal structure of the aperiodic component and can be extracted using the acoustic event detection and characterization method based on a fixed point analysis of a mapping from time window positions to windowed energy centroids[5].

5.3. Group delay based event extraction

Speech can be interpreted as a collection of acoustic events. The response to vocal fold closure characterizes voiced sounds, and a sudden explosion of the vocal tract characterizes stop consonants. Fricatives can also be characterized as a collection of temporarily spread noise bursts.

Similar to the F_0 extraction based on fixed points, acoustic events are extracted as a set of fixed points $T(b)$ based on the following definition.

$$T(b) = \{\tau \mid \tau(b,t) - t = 0, \quad -1 < \frac{\partial}{\partial t}(\tau(b,t) - t) < 0\}, \tag{7}$$

where $\tau(b,t)$ represents mapping from the center location t of the time window to its output energy centroid, and "b" represents the parameter to define the size of the window. For the sake of mathematical simplicity, Gaussian time window is used in our analysis.

Figure 5 illustrates how the energy based event detection works. The energy centroid trajectory crosses the identity mapping upward at several locations; they are fixed points[4].

A group delay based compensation of event location was introduced, because the event location defined by Eq. 7 is inevitably consists of a delay due to impulse response of the system under study. Usually, the interesting location is not the energy centroid; instead, it is the origin of the response. The proposed method[5] uses the minimum phase impulse response

[4]To make representation intuitive, the horizontal axis of the figure represents the energy centroid instead of window center. This illustrates how energy centroid is attracted by local energy concentration.

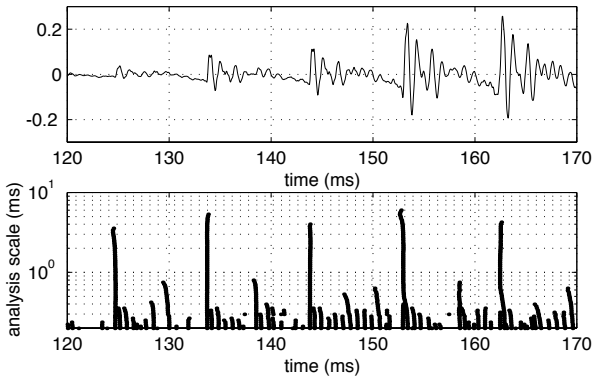

Figure 6: *Scale dependency of the detected event. The lower plot shows extracted event locations for different scale parameter σ_w. The upper plot shows the corresponding waveform.*

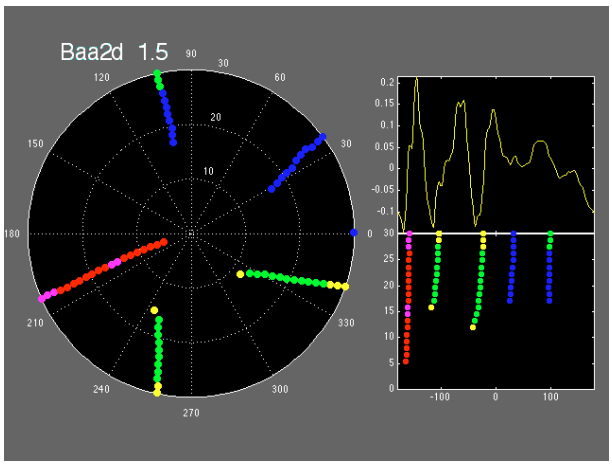

Figure 7: *Polar plot of event locations and its salience with multi resolution analysis. Angle in the left polar plot represents the phase of the fundamental component at event location. Right plot represents phase of the fundamental component as the horizontal axis. In both representations, salience is represented as the color (hot: strong, cool:weak) of symbols and radius represents the logarithmic scale (the outer the finer).*

calculated from the amplitude spectrum to compensate this inevitable delay. A test using a speech database with simultaneously recorded EGG(ElectroGlottoGram) signals[16] revealed that the proposed method provides estimates of vocal fold closure timings with the accuracy of 40 μs to 200 μs in terms of error standard deviation depending on the temporal spread of the events[5].

The analysis parameters of the event analysis method are an analysis window scale and a viewing frequency range. A systematic scale scanning in event analysis yields a hierarchical excitation structure of the signal[5].

Figure 6 shows an example of multi resolution event analysis. The same material was analyzed using scale parameters ranging from 0.1 ms to 10 ms. The vertical axis of the lower plot represents the scale parameter. Note that majority of fixed points are located at vocal fold closure instants.

Figure 7 shows the distribution of fixed points in terms of the phase of the fundamental component in two alternative

ways. The plots overlay fixed points extracted using 13 different window scales for one second of sustained vowel /a/ spoken by a male speaker. Radius of the right plot represents the scale parameter using logarithmic conversion $20\log(\sigma_w F_0) + 30$. A clear alignment of fixed points around 240 degree corresponds to closure of vocal fold and the other alignment around 0 degree seems to corresponds to its opening. By using these hierarchical representations and the frequency domain aperiodicity measure, a method to design excitation source can be derived.

6. Morphed sound files

Examples of auditory morphing using the French singing database provided by VOQUAL'30 can be found on the following URL.

`http://www.sys.wakayama-u.ac.jp/~kawahara/VQdemo/`

Please check which aspect of voice quality is morphed reasonably and which is not. At this point, vibrato was not morphed reasonably, partly because direct interpolation of two different FM modulation on F0 does not yield intermediate FM modulation and yields beating F0 modulation.

7. Discussion

There are many open questions to proceed this research direction further. First of all, redundancy in control parameters of voice quality suggests need for a unified framework to organize and translates findings done in different disciplines.

The aperiodicity index introduced in this article is closely related to HNR (Harmonic to Noise Ratio). It also is related to the decomposition method of speech signals into periodic and aperiodic components [17, 18]. It is important to investigate relations of these indices and to find the best combinations/selections of these indices for each applications such as diagnosis, control for synthesis and morphing. It also will be important to find a way to approximate complex time-frequency aperiodicity indices using a small number of parameters.

The time domain energy concentration parameter does actually modifies perceptual impression of noise. However, reliable extraction of the parameter and proper representation (for example time-frequency resolution) are not understood well. It is also a unsolved problem how to allocate perceptual aperiodicity into the frequency domain index and the time domain index.

8. Summary

An exemplar based approach on voice quality research was introduced using a high quality morphing procedure based on STRAIGHT. This approach will serve as a complementary tool to the conventional methods. Even with simple representations used in the original STRAIGHT still be able to provide authentic voice quality control. However, it was also pointed out that it is necessary to introduce a parametric representation of vibrato for morphing singing sounds realistically.

9. Acknowledgements

This work was supported by Grant-in-Aid for Scientific Research (B) 14380165 and e-society project (primary investigator: Prof. Kiyohiro Shikano of NAIST) of MEXT Japan. It was also supported by the advanced research grant of Wakayama University.

10. References

[1] Hideki Kawahara and Hisami Matsui, "Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation," in *Proc. ICASSP*, 2003, [in print].

[2] Hideki Kawahara, Jo Estill, and Osamu Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight," in *Proc. 2nd MAVEBA*, Firenze, Italy, 2001, [CD ROM].

[3] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[4] Hideki Kawahara, Haruhiro Katayose, Alain de Cheveigné, and Roy D. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," in *Proc. Eurospeech'99*, 1999, vol. 6, pp. 2781–2784.

[5] Hideki Kawahara, Yoshinori Atake, and Parham Zolfaghari, "Accurate vocal event detection method based on a fixed-point analysis of mapping from time to weighted average group delay," in *Proc. ICSLP'2000*, Beijin China, 2000, pp. 664–667.

[6] Hideki Kawahara and Haruhiro Katayose, "Scat singing generation using a versatile speech manipulation system, straight," in *141st meeting of the Acoust. Soc. Amer., Chicago*, 2001, vol. 109, pp. 2425–2426.

[7] H. Kawahara and H. Katayose, "Scat generation research program based on straight, a high-quality speech analysis, modification and synthesis system," *J. of IPSJ*, vol. 43, no. 2, pp. 208–218, 2002, [in Japanese].

[8] Hideki Kawahara, "Systematic downgrading for investigating "naturalness" in synthesized singing using straight: A high quality vocoder," in *143rd meeting of the Acoust. Soc. Amer., Pittsburgh*, 2002, p. 2334.

[9] Hisami Matsui and Hideki Kawahara, "Investigation of emotionally morphed speech perception and its structure using a high quality speech manipulation system," in *Eurospeech 2003*, Geneva, 2003, vol. 1, pp. –, [accepted for publication].

[10] Murtaza Bulut, Shrikanth S. Narayanan, and Ann K. Syrdal, "Expressive speech synthesis using a concatenative synthesizer," in *Proc. 7th Int. Conf. on Spoken Language Processing (ICSLP '02)*, Denver, 2002, vol. 1, pp. 1265–1268.

[11] H. Fujisaki, "A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour," in *Vocal Fold Physiology: Voice Production, Mechanisms and Functions*, O. Fujimura, Ed., New York, 1998, pp. 347–355, Raven Press.

[12] L. Cohen, *Time-frequency analysis*, Prentice Hall, Englewood Cliffs, NJ, 1995.

[13] F. J. Charpentier, "Pitch detection using the short-term phase spectrum," *Proceedings of ICASSP'86*, pp. 113–116, 1986.

[14] T. Abe, T. Kobayashi, and S. Imai, "Harmonics estimation based on instantaneous frequency and its application to pitch determination," *IEICE Trans. Information and Systems*, vol. E78-D, no. 9, pp. 1188–1194, 1995.

[15] Martin Cooke, *Modelling Auditory Processing and Organization*, Cambridge University Press, Cambridge, UK, 1993.

[16] Yoshinori Atake, Toshio Irino, Hideki Kawahara, J. Lu, S. Nakamura, and K. Shikano, "Robust fundamental frequency estimation using instantaneous frequencies of harmonic components," in *Proc. ICSLP'2000*, Beijing China, October 2000, PB(2)-26, pp. 907–910.

[17] B. Yegnanarayana, C. d'Alessandro, and V. Darsinos, "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 1, pp. 1–11, Jan. 1998.

[18] C. d'Alessandro, V. Darsinos, and B. Yegnanarayana, "Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 1, pp. 12–23, Jan. 1998.