Manuscript for

J. Logopedics Phoniatrics Vocology,
Special Issue of Jan Gauffin Memorial Symposium
(16 October 2008, Stockholm)

Noh Voice Quality[i]

By

Osamu Fujimura,

Kiyoshi Honda,

Hideki Kawahara,

Yasuyuki Konparu,

Masanori Morise,

J. C. Williams[ii]

January 18, 2009 Version: noh-jlpv090118

There are figures, figure captions, and TOC as separate files.

[[The separation of unit names from numbers is not consistent.]]

1. Noh: The Traditional Japanese Play

Noh is a traditional theatrical art in Japan with a documented history of over a millennium[iii]. The tradition is currently maintained by multiple schools of discipline in different regions of Japan, conducting public performances regularly, each school independently or jointly. The formal performance centers on the site [ʃite], the principal actor, who represents his school of Noh[iv], wearing an evocative mask and gorgeous silk costume specific to the dancing and singing of a particular play. The play is normally performed on a wooden stage of a traditionally designed Noh theatre, but on special occasions during the summer (takigi-noh), performances are held outdoors. The theatre hall specifically for Noh performance has a particular acoustic design with resonating pots hidden under the floor for acoustic effects of stepping of the site. The integrated singing and dancing of the site is often assisted by one (or more) subsidiary performer, who introduces (or intervenes) the site performance providing some verbal context of the story or forming a dialogue. The site's dance with simultaneous singing is accompanied by drums, a wood wind instrument (hu-e) and a male vocal chorus, sitting by the edges of the square stage. The subtle, suppressed expression of deep emotion that is conveyed by the performance of Noh strongly reflects the influence of Zen Buddhism.

In Noh, the artistically refined form of highly emotional message is delivered largely by the special voice quality and rhythmic patterns that the site uses in singing. The text of the song is written in old Japanese inheriting from the time of the playwright[v]. The pronunciation of each phrase is highly stylized with liberally artistic prosodic manipulation; native Japanese audience usually cannot follow the text from the singing of the site, unless the published text is studied beforehand, which is the case for some of the audience. In fact, it seems typically not the linguistic expressions that play the principal role of conveying the emotional message: many of the audience attending the performance are deeply moved emotionally by sitting in the audience and listening to the site's voice. Foreigners without understanding the Japanese language often attend Noh performances and become emotionally affected in tears. The facial expression of the site is totally hidden behind the mask throughout the performance. While the expressive stepping on the floor in dancing, which also helps coordinating instrumental and vocal accompaniment, must play a strong role in the proceeding of the play, it is beyond doubt that the site's voice quality and manipulation of prosody in singing principally transmits the emotional message.

With the site of the Konparu Noh School in Nara, one of the oldest conservatories of the art of Noh, as one of the collaborators of this study, we attempt to approach toward understanding some of the mysterious properties of the voice of Noh. By studying this extreme form of vocal expression, we hope

that not only will we understand the acoustical nature of this classical art better, but also will be able to uncover inherently human characteristics of daily conversational voice. Expressive speech in our daily conversation is likely related to the highly refined and extreme form of expressivity in Noh in some respects. We note also that the controlled voice in Noh may share some aspects of pathologically affected voice, and exploring signal properties of the voice production in Noh singing may provide another approach toward exploring into pathological conditions of phonation as well as idiosyncratic normal voice characteristics of individuals. For all these purposes, from a basic research point of view, we need a comprehensive descriptive framework for voice characteristics.

The present paper does not intend to arrive at a quantitative description of the voice qualities in Noh performances in general. It is questionable whether one can identify critical voice characteristics of singing in Noh by collecting and analyzing a wide range of Noh voice signals by signal processing techniques that are traditionally available. In fact, as we show here, the descriptive framework of voice signal properties that are currently available in speech science is not sufficiently effective to describe Noh voice, and probably, in fact, any truly lively conversational voice. We also have noted, in previous preliminary studies, there were other types of artistic use of special human voice characteristics that seemed unusual in more or less similar ways to what we find in Noh [Mazo et al. 1994].

We suspected that the extended notion of voice quality should be particularly related to the deviation from the traditional assumption of periodicity about voiced signals [Kawahara, Fujimura & Konparu 2006]. In other words, we must be able to describe special (in some sense controlled) fluctuations of signal parameters, such as F0 (see infra), amplitude or intensity, spectral envelope or waveform within a glottal period. In fact, we have identified here some unusual or hitherto unnoticed or undiscussed properties of voice signals in highly expressive and emotionally moving parts of the site's singing. Given a newly developing signal processing methods (see infra), by collaboration of their inventors as coauthors, we are now in a position to start exploring quantitative measures for describing some of aperiodic voice signals that are produced by a strongly nonlinear physiological system of human voice generation. At least, we can point to some physical properties of voice signals that appear to be worth pursuing for quantification.

2. Aperiodicity in voice signals

In speech science grounded on the source-filter theory [Fant 1960], voiced signals in quasi-stationary portions of vowel-like acoustic segments are assumed

to be basically periodic. The time function representing the volume velocity of air passing through the glottis, viz. the gap between the vibrating vocal folds, is treated as a periodic signal repeating the same waveform for a fundamental period, reflecting the stationary vibratory movement of the vocal folds in phonation. Any deviation from this assumption of signal periodicity has to be treated as one or more of the voice signal parameters that slowly (in comparison with the fundamental period) change in time as a signal modulation. The random or chaotic components of the voice source signal due to air turbulence is assumed to be additive in the linear acoustic system, and the nonlinearity of the vocal fold tissue caused by inherent tissue properties including hysteresis must be assumed to result in no significant deviations from quasi-periodicity of the voice source signal. Accordingly, the tissue vibration and the resultant airflow modulations are, in our standard treatment, assumed to be produced by an oscillatory system that is slowly changed through time-dependent system parameters under physiological control.

The primary consideration of such slow parametric control, in phonetics and related areas, has been the fundamental frequency of the voice source signal, often called the pitch signal in speech technology (along with signal intensity or amplitude). The pitch signal involves discontinuous switching between voiced and unvoiced acoustic segments, separately from the description of the smoothly and one-dimensionally changing numeric parameter called F0 (fundamental frequency) during each voiced interval of an utterance, setting aside transient phenomena of voice onset and offset, which is usually approximated by an abrupt onset and offset of the periodic signal amplitude. There have been discussions expressing concerns about the insufficiency of this assumption of scalar pitch signal (see, for example, Fujimura [1968]), but experimental quantification of this issue has been difficult due to the lack of effective approximations that could treat computationally the deviations from the standard assumptions of the quasi-periodicity.

In this paper, in conjunction with the development of a new computational method by Kawahara and Morise, called TANDEM-STRAIGHT, which represents a substantial improvement of the widely used speech signal processing algorithms (see, e. g., Flanagan [1972]), we will present our current progress [Kawahara, et al. 2008a,b]. The new method of description still assumes basic quasi-periodicity for arbitrarily selected segments of aperiodic signals, but it allows substantially more flexibility in representing an approach based on multiple hypotheses of approximate local periodicity. In particular, it allows, for the first time, displaying coexisting and locally changing frequencies that we have to describe in cases such as gradual transitions between harmonic structure with a fundamenatl frequency and its subharmonics (see infra for deeper discussion of this issue).

We consider F0 as one of many physical attributes of voice quality [Laver 1980]. Its variation within an utterance characterizes conditions of voicing in human speech, in part under conscious control and in part not, and in part with slow and smooth phonetic changes from syllable to syllable [Fujimura 2000, 2007] and in part apparently accompanied by random fluctuation in the process of voice source signal production. It is such temporal variability of voice signals that we are particularly concerned with in this study. The variability pertains not only to the fundamental frequency (as observed as phase (or informally frequency) modulation) or amplitude of the waveform (amplitude modulation), but also to the waveform within each glottal cycle or its corresponding spectral envelope as a function of frequency. Instability of such physical properties of the quasi-periodic signal is inherent to human voice, and in fact, some deviation from a perfect stationary repetition of the same signal segment is a critically necessary property for a sound signal to be accepted as human voice.

It is often suspected that aperiodicity of voice signal reflects emotions in expressive speech. The nature of aperiodicity, its exact physical properties in voice, in relation to specific percepts, is largely not yet understood, except that some well-defined situations such as tremolo in music have been quantitatively studied. Noh singing provides us with challenging research opportunities.

3. Expressive voice in Noh: Yorobôsi example

One typical aperiodicity that is often observed in normal conversational utterance is either a sudden or gradual introduction of subharmonics, particularly, but not necessarily, toward the end of a phonetic phrase often in conjunction with the phenomenon termed vocal fry. In highly emotional speech such as Russian lament utterances, subharmonics appear in many portions of phrases with abrupt appearances and disappearances, in relatively low frequencies not necessarily with respect to the fundamental component [Mazo et al. 1994]. In Noh, we have identified similar occurrences of subharmonics, but the variability of harmonic frequencies as observed in narrowband spectrograms is more complex and includes peculiar phenomena hitherto not reported to the authors' knowledge.

We chose a part of recorded acoustic signals from Noh singing, which seemed typical for a highly expressive singing in Noh. The text, with gloss in parentheses, giving some context of this excerpt, is given in Appendix. Fig. 1 illustrates an example in a compressed waveform temporally aligned with a narrowband spectrogram. The signal was recorded in the specialized Noh theatre of the Prefectural Public Hall in Nara City, Japan, without audience, without a Noh mask, and in plain clothes. The site, one of the coauthors, sung an excerpt from a Noh play called Yorobôsi, depicting an overjoyed (in deep and strong emotion, not

apparently sounding joyful) scene of a middle-aged man. It is a scene of an utmost delight of a blind man (Yorobôsi), who was abandoned by his father by someone's malicious accusation when he was a small boy, became blind, and now lives in a Buddhist temple. The father had discovered the truth and was in deep regret; he happened to visit the temple (guided by Buddha's mercy), standing by his son. In the play, Yorobôsi, without realizing his father was there, is in a fantasy of seeing the moonlight. His joyful emotion of seeing the light, after being blind for years, is then again overwhelmed by the reality of being a blind man [Konparu 2006].

The excerpt in Fig. 1 corresponds to the text 'yorobôsi ga (subject grammatical particle) tune wa (under normal circumstance, before becoming blind) minaresi (familiar, the moonlight implied) kyôgai (surroundings) nareba (since it is)', sung in "joyful style". In the spectrogram, a marked frequency modulation of harmonic components (to be called F0 modulation, see infra) at a rate of about 5 Hz is observed in many acoustic segments of the utterance. Sometimes, the onset-offset of such a modulation is rather abrupt and the switching may be coupled with a rapid articulatory movement[vi] (e. g., in Fig. 1, [**i**][vii] to [**n**] at articulatory implosion, see the mark (a) at the bottom of the spectrogram). Even within the middle of the phonological syllable /gaJ/, for the [**a**] gesture, the effect of F0 modulation is seen clearly (only in higher harmonics, where the frequency change is amplified according to the multiplicity of the harmonic number), whereas for the succeeding gesture [**i**], the F0 modulation is stronger and is markedly visible only in the F1 region[viii]. The onset of this F0 modulation is rather abrupt at the end of the formant transition for the vowel change. In the succeeding syllable /na/, the F0 modulation is again reduced so that its wavy effect is seen at the high frequency region.

Fig. 1 about here

Subharmonic components appear briefly at the very beginning (see the mark (b)) of the phrase /kyoHgaJnareba/ (see endnote v about transcription), associated with a rapid rising transition of F0 with subharmonics. These subharmonic components are clearly identified between H1 and H2, H2 and H3, and H3 and H4. At the end of the palatalization gesture of the onset of /kyoH/, F0 becomes stable. Apparently, there are special laryngeal gestures employed toward the end of the phrase.[ix]

Fig. 2 is another excerpt from the same singing of "Yorobôsi". In this sample of the phrase, which corresponds to the text 'nani (whatsoever) utagai (doubt) mo (even) naniwa (place name, Ôsaka) e (bay) ni (at, grammatical particle for

location),' an unusual spectrographic pattern is observed particularly in the portions (a), (b), and (c). The marking (a) in the figure points to the appearance of subharmonic components between all adjacent pairs of harmonics, clearly indicating half of the F0 that is varied but remains persistent through this phrase; the subharmonics weaken and disappear gradually during the area (a), while the F0 modulation becomes very strong toward (b). The parts (a) and (b) correspond to a syllable /e/ of 'naniwa e', a phonologically short vowel without any consonant but phonetically extended and interrupted in the middle by a glottal stop, apparently, as indicated in the spectrogram by a blank about 240 ms. long. Before and after this voice interruption, which represents no phonological significance, there is a strong F0 modulation, which, particularly before the break toward the end of the area marked (b), is so strong that, as we saw above in Fig. 1, only in the lower harmonics the wavelike frequency modulation is clearly identified[x].

Fig. 2 about here

Closely adjacent to the break, both before and after the blank, particularly before, there is an identifiable time segment of the harmonic structure that shows strong subharmonics but maintains a relatively stable F0, suggesting a strong but transient amplitude perturbation of the signal associated with the glottal stop. At the beginning portion of the area marked by (a), there is another discontinuity of the spectrographic patterns, which unlike the break in (b) is not a silence period but enters a distinctly perturbed time segment in the vocalic transition from [a] to [e] (vowel hyatus). This discontinuous short segment preceding the first main portion of the vowel /e/ (between (a) and (b)) shows initially strong but gradually decaying subharmonics, continuing from the preceding brief voiced segment showing strong voicing perturbation (a). During the sustained articulation of the vowel [e], from (a) to (b), the voice becomes intensified (see waveform) and F0 becomes frequency modulated before the vibration becomes perturbed (subharmonics appear briefly) and finally interrupted (b).

The region marked (c) shows another particularly strongly perturbed voice. A careful listening of this portion indicates a perceived final pitch fall, which is consistent with what appears like a sharply falling and strong second harmonic. The frequency perturbation is so strong that toward the very end, the harmonic structure becomes obscure. Of a particular interest is the apparent convergence of the fundamental (going up) and the second harmonic (going down), which is contradictory to the concept of quasi-periodicity. A close look in this portion of the spectrogram by manipulation of the spectrograph suggests that a new subharmonic component arises near the end, which becomes stronger, relative to

the existing harmonics, but it disappears toward the very end (See Section 5 for further discussion).

The spectrographic appearance of changing harmonic structure may be deceptive. It is probably a manifestation of a strong physical interaction between the sinusoidal component in the movement of the vocal folds with the frequency of the second harmonic (which corresponds to the sinusoidal second harmonic oscillation component of the source signal emerging from the glottis), which is amplified by the first formant, the lowest mode of vocal tract oscillation. In other words, the air pressure oscillation exerts the maximum force variation onto the vocal fold tissue for the lowest vocal tract resonance at the glottal end, which represents the pressure node. As is well known in filter circuit design, when two oscillators (vocal folds in vibration and the vocal tract air colum in resonance) of similar resonant frequencies are weakly coupled by mutual interaction, the original oscillation frequencies are shifted according to the strength of coupling, and the vibration modes of the two independent systems become mutually pulled in and locked. If this is the case, it means that this special timbre of the voice in Noh cannot be handled by the independent source-filter approximation.

Fig. 3 shows a mysterious lack of correspondence between the spectrographic pitch behavior and perceived pitch change. This sample represents a phrase-final portion of the same Yorobôsi singing; the text is 'Sumiyosi (name of place in the Ôsaka area) no (possessive grammatical particle)'. Toward the end, a strong F0 modulation appears again, after the apparent spectrographic discontinuity from the time segment representing the vowel [o]. There is a stable and stationary portion with relatively little F0 modulation after the nasal murmur in syllable onset, which in turn succeeds the voiced but strongly perturbed high vowel [i] (with a preceding voiceless [s]). Note the phonetic interruption in the middle of the short syllable /yo/ is phonologically irrelevant.

Fig. 3 about here

While the harmonic pattern toward the end of this entire excerpt appears to indicate a distinct F0 fall, all the lower harmonics being consistent albeit wavy F0 modulation (viz. the higher the harmonic number, the larger the wavelike frequency modulation, as predicted by the harmonic structure of the signal), the perceived pitch change is clearly rising toward the end. At this point, before we saw what the new XSX display showed (Section 5), this phenomenon simply escaped our explanation. We noted, however, that the descent of the third harmonic seemed somewhat out of proportion compared to the fundamental and the second harmonic components. In fact, the fundamental component, though

relatively weak, is slightly rising unlike the second and the third harmonics, toward the end of the phrase. This phenomenon seems related to what we have just discussed above in relation to Fig. 2. We will revisit this issue and provide a tentative answer by our new computational pitch estimate (Section 5, in connection with Fig. 10).

It should be mentioned here, that throughout this excerpt, which contains different vowel qualities, the resonant enforcement of the very high frequency region near the top of the figure is very stable and strong, perhaps somewhat rising in frequency toward the end of each phonetic phrase. Even though a stability of very high modes of vocal tract resonance (high formants) can be explained by inevitably slow spatial change (i. e., constraint against large curvature of the surface) of the tongue surface shape compared to loops and nodes of the higher order modes vocal tract resonance, this may also reflect a stable configuration of the laryngeal cavity [Honda, et al. 2004; Takemoto, et al. 2006; Kitamura et al. 2006], which presumably corresponds to Sundberg's singing formant [Sundberg 1970].

4. TANDEM- STRAIGHT

STRAIGHT [Kawahara et al. 1999] is a modernized computational method for analyzing, synthesizing, and selectively modifying specific properties of speech signals. Conceptually, it follows the idea of VOCODER (see Dudley [1939], also Flanagan [1972]), which was invented more than 50 years ago. It evaluates short-term power spectra of the signals, temporally stabilizes them to represent phonetically relevant speech signal characteristics, removing perceptually unnecessary details through extracting effective parameter values of the time-varying spectra.

One key issue is the separation of the voice source characteristics that determine prosodic properties of phonetic signals from time-varying spectral timbre information [Fant 1960]. Perception of timbre of voice depends crucially on signal characteristics that can be regarded as independent from spectral properties reflecting speech articulations, which manifest sound contrasts for phonological functions such as distinctions among different vowels. Apart from such (so-called segmental) phonetic information related to spectral properties, we need time-varying signal properties that are represented as signal parameters in the waveforms of speech signals, called suprasegmental properties. Within this category of speech signal properties, there are those that are used characteristically in different languages for representing phonological distinctions, such as tones and accents for contrasting different words or phrases with inherently different meanings of such linguistic forms.

In speech technology, the aim of inventions, such as vocoders, was to automatically extract such phonologically pertinent segmental information from speech signals for machine processing. As we made advances, however, given the modern computational tools, we are now interested in properties of human communications that go beyond the scope of such processing capabilities concerning lexical properties of speech. In fact, we have been finding that the task of extracting linguistically relevant information was not completely separable from other tasks. In order to handle humanly relevant communicational information, we need to be able to process speech signal properties that traditionally have been considered irrelevant to linguistic messages. Such information is often identified as emotional; it pertains to highly expressive verbal messages.

The vocal signals representing such expressive messages often, or even invariably, are characterized by basic deviations from main assumptions about speech signal properties, viz. sequencing of quasi-stationary signal characteristics undergoing mechanical smoothing (coarticulation) and independence of source signal characteristics from vocal tract filter functions [Fant 2004, Fujimura 2007]. Apart from this deviation from the traditional basic assumption about speech signal characteristics, and also interacting with them as we will discuss in this paper, expressive voice must have fluctuations, which we call aperiodicity of voice signals. In order to address this issue correctly, which is well beyond the scope of this paper, we need to establish a non-traditional theory of voice signal production. At the same time, the conventional tools for representing signal characteristics such as the spectrograms and LPC (linear prediction coefficients) are not effective enough, as we will demonstrate below. Specifically, the traditional concept of the “pitch signal” in speech technology, in terms of separate voiced and unvoiced speech segments and F0 for voiced segments, is too gross an approximation for our present discussion. TANDEM-STRAIGHT as a new method for representing speech signal properties as a smoothed, cycle-by-cycle spectral representation in the time-frequency domain, appears to provide an alternative representation of more complex and necessary information for discussing voice quality.

STRAIGHT was proposed by Kawahara and his associates in 1999 as a new speech signal processing method [Kawahara et al. 1999]. Looking at the speech signal as a time function, the algorithm tries to identify approximate repetitions. By finding multiple candidate patterns for repetitions, it proposes multiple hypotheses of putative fundamental periods. Using each candidate time interval as the analysis window, it tries to compare many different fundamental

frequencies to evaluate their consistency for explaining locally acceptable periodicity as the best candidate. TANDEM improved this algorithm making it substantially more efficient [Kawahara et al. 2006a,b]. The TANDEM-STRAIGHT algorithm detects locally repeated appearance of any fragments of patterns in both time and frequency domains. The chosen best candidate pattern, if there is any one pattern that outstands in terms of consistency, using many different criteria in parallel (saliency test score), may be assumed as a candidate for the perceived pitch sensation.[xi].

This idea is utilized in the new version of STRAIGHT (TANDEM-STRAIGHT) [Kawahara et al. 2008a] for F0 extraction. TANDEM is a new procedure for calculating temporally stable power spectrum for fluctuating signals, by using two time windows a half pitch period apart. TANDEM-STRAIGHT uses different F0 extractors. Outputs of extractors are integrated to yield final estimates [Kawahara et al. 2008b]. This multi-candidate structure enables us to process aperiodic signals that cannot be characterized by a single number of F0. In other words, our extractor extracts a representation of voice excitation structure rather than a scalar value. For this reason, this pitch evaluator is now tentatively called XSX, abbreviating excitation structure extractor. The best estimate of F0 candidate can be selected period by period.

This procedure can produce many different representations of the signal properties for different purposes. The one we use in this discussion is a display of candidates of F0 values for individual glottal vibration periods (resulting from different F0 assumptions) for the given speech signal. Each F0 candidate value is associated with an estimate of "salience", which may possibly be interpreted as the likelihood of the particular F0 value to represent the effective pitch of the voice signal in the given signal context.

4.1. Examples: Amplitude and frequency perturbation

The continuous transition between harmonic and subharmonic structures, as we have observed in complex voice signals in Noh, can be represented effectively by XSX. First let us see how the XSX procedure works by the use of simple mathematical examples.

Fig. 4 about here

Figure 4 shows extracted pitch information for an amplitude-modulated pulse train. The base fundamental frequency is 160 Hz, and a component at 80 Hz is its subharmonic at F0/2. As shown at the top panel of the figure, the signal starts as

a 160-Hz pulse train. The pulse amplitude of every other pulse is then gradually decreased until the amplitude becomes zero at 0.4 second from the beginning. Along the ordinate in the main panel, multiple candidates of pitch eliciting frequencies are shown. The dark points (or thick lines connecting them) suggest the best candidate at each frame.

In this figure, the transitional situation from the usual periodic excitation to subharmonic excitation is shown as a continuous process, rather than being forced, as in the traditional pitch detectors, between F0 and 1/2 F0 arbitrarily, or creating a false continuous frequency change between the two values. As shown by the gradually changing thickness of each line, which represents a computed salience of the frequency component presumably as the likelihood of eliciting pitch perception [Kawahara 2008b,], the F0 component disappears continuously and the F0/2 component (half frequency subharmonic) appears gradually showing two distinct frequencies in parallel.

Fig. 5 about here

Figure 5 shows a similar representation for a frequency-modulated pulse train. In this case, the pitch candidate at F0 starts to show a wavy frequency change, and as the waviness grows at some point, XSX splits the candidate into two continuously separating lines. Unlike in the case of amplitude modulation, however, XSX picks up the candidate at the subharmonic frequency abruptly. How the pitch sensation in such a situation is actually affected, compared to our predictions, remains to be perceptually studied[xii].

In cases of realistically complex voice signals, by observing the patterns of candidate frequency components using XSX, it may also be possible to infer the cause of subharmonic excitation in each case. For example, in Fig. 5, the amplitude modulation does not introduce any significant frequency perturbation in extracted candidate frequencies around 160 Hz; i. e., a straight horizontal line at this frequency gradually weakens and disappears at 0.4 s.[xiii]

In contrast, the frequency modulation clearly introduces a wavy perturbation pattern around 160 Hz, as discussed above (Fig. 5). What the XSX plotting actually shows is a cycle-by-cycle variation of the excitation time function interval and it should be interpreted in terms of a cycle-by-cycle fluctuation of vocal fold vibration, in the natural voice production; the situation such as Noh voice is not as simple as in the mathematical exercise above. Other candidates that appear in the display for natural situations may be in part obscured in human perception under the background salience level due to random fluctuation.

4.2. Synthesis and morphing

The TANDEM-STRAIGHT analysis and synthesis procedure produces quite naturally sounding signals even with some alteration of details, and the difference from the original signal is often perceptually very subtle. However, there is some noticeable difference apparently related to the appearance of the subharmonic components[xiv]. A carefully designed perceptual testing must be used to evaluate the details of signal properties in relation to individual signal characteristics that can be identified with XSX. On the other hand, the voice quality in the global situation of Noh performance no doubt produces very distinct expressive effects. It is a great challenge for us to understand this global effect, given that minute local properties now can be identified accurately in details.

TANDEM-STRAIGHT decomposes input speech signal characteristics into three partial presentations to portray different aspects of the signal properties: a smooth spectrographic representation (STRAIGHT-spectrogram), a spectrographic representation of aperiodicity, a fundamental frequency (F0) tracking. By using these partial representations, it is possible to generate "morphed" signals, which are produced by mixing two different signals in terms of their parameter representations. In this way we should eventually be able to, for example, continuously transform one of two exemplar Noh singing styles into the other (to the extent that the signal manipulation process succeeds) in order to test if the parameterized representation of voice quality captures pertinent physical properties for specific expressions.

5. Yorobôsi examples by XSX

Several samples excerpted from the materials described in Section 3 are analyzed here using the XSX computation.

Fig. 6 about here

Fig. 6 shows the subharmonic region, which we discussed in connection with Fig.1, the part marked (b). It is in the middle of the long vowel [**o...**] of /kyoHgaJ nareba/, where the spectrogram (Fig. 1) shows rapidly changing pitch and abrupt transition from subharmonics to harmonics without a significant change in articulation. Instability is indicated by highlighted marks in Fig. 6 but the component slightly above 100Hz is continuously the most salient pitch candidate from 36.95s to 37.16s.

One way to interpret this is that the sum of the two glottal periods (9.0 ms.), corresponding to 180 Hz (5.56 ms.) and 290 Hz (3.45 ms.), if these two periods

alternate contiguously in pairs, is equivalent to a fundamental period corresponding to 110 Hz (9.1 ms.). This interpretation may invite some speculation that the two subsegments within each glottal cycle correspond to the open and closed glottis, the former reflecting the vocal fold movement modulating the airflow directly, while the latter, showing no airflow through the glottis (apart from the vertical movement of the vocal folds), reflecting the lowest vocal tract resonance affecting the vocal fold position by pressure variation, which is considerable because the glottis represents the pressure node of the air resonance. This latter speculation seems plausible, given that 290Hz is, in this vowel articulation with a strong bilabial rounding/protrusion, close to the first formant frequency. Note, however, that such an interpretation violates the usual assumption of source-filter independence, which can provide no explanation of any temporal substructure of the source signal within the glottal period. It should also be mentioned here that the traditional acoustic phonetics based on spectrography has no way to examine such details of voice signal properties.

Fig. 7 about here

This display by our XSX processing pertains to the mysterious region marked (c) of Fig. 2. The syllable is at the end of 'nani utagai mo naniwae ni'. Plots shown in Fig.7 illustrate very unpredictable variability of pitch sensation. Near the very end of this display, around 56.7 - 57.0s., a very low frequency component around 60-70Hz, though transient, becomes salient, interrupted by dominancy of a much higher component around 270Hz, which is near the first formant frequency of the vowel [i] (see discussion in Section 3 in connection with Fig. 2.)

Revisiting the mysterious pitch rise perceived toward the end of Fig. 3, pertaining to the vowel [o] at the end of the phrase 'sumiyosi no', we have analyzed the signal by XSX. The result as we see in Fig. 8 clearly indicates the most likely pitch we would hear to be rising, from about 110s. toward the end (in the upper portion of the figure continuing the steadily dominant next to top thick track in the left half of the display), though there are temporary interruptions by other candidates in the very low frequency range after about 108.95 s. Thus, despite no clear indication for a rising pitch that we detected when we examined the narrowband spectrogram in Fig. 3, we have computational prediction supporting this subjective perception of phrase final pitch rise.

Fig. 8 about here

This amazing result seems to suggest that our new TANDEM-STRAIGHT algorithm captures perceptually critical characteristics of the sound signal that cannot be revealed by standard sound analyses. We are in the process of examining the phenomenon in details.

6. Concluding Remarks

Prosody is not just an F0 contour. The concept of pitch in speech technology is not sufficient to describe speech prosody. Phonologically, tone and stress and rhythm are necessary to describe lexical and phrasal phonological patterns in the prosodic aspects of speech. Phonetically, voice quality and temporal organization of speech gestures manifest prosodic control as physical events, and one of principal aspects of voice quality, along with other aspects as we have discussed, is the pitch perception. F0, as a physical variable reflecting an abstract dimension of laryngeal control along with respiratory and mandibular controls, represents one manifestation of the tonal phonological variables [Fujimura 2008].

The prosodic control of voice quality, in part perceived as pitch change, is not performed completely based on any phonological or phonetic segmentation of speech units, whether phonemes or syllables, as accepted for descriptions of "normal" speech. When strong emotions are expressed, one has to violate general rules of speech organization. While syllables and their linear concatenation, through usual coarticulatory process and sometimes through special operations as specified in phonology (see Fujimura & Williams [1998, 2007]), serve as basic means for temporal organization of speech signals [Fujimura 1992, 2000], in an artistic performance, we know, we resort to occasional deviations from such normal processes. We observed, for example, rather conspicuous voicing interruptions (as well as rapid pitch changes, which we have found but not discussed here in details) in the middle of a syllable, whether phonologically long or short. Such somewhat discontinuous phonetic events are not necessarily linked to the mora, a phonetically important temporal unit for Japanese speakers.

When we discuss intricate control of voice quality in highly expressive voice, we need to go beyond the well-accepted assumption of source-filter independence in order to understand the physical characteristics of voice production. Not only voicing occasionally involves vocal fry and uses of subharmonics, but also inharmonic frequency components seem to characterize expressive voice quality indicating strong interaction between vocal fold vibration and vocal tract resonance. Aperiodicity of voice signals conveys crucial

communicative messages.

7. Acknowledgement

The current team of researchers as collaborators in this work was organized as a study group on Voice: Acoustic Characteristics and Emotional Expression of Speech (PI: Osamu Fujimura) constituting part of a larger research project on Wide Range Interdisciplinary Issues Accompanying Advanced Science and Technology, of the International Institute for Advanced Studies (IIAS), Kyôto, Japan, in part supported by the Japan Society for the Promotion of Science, during the fiscal years 2006 - 2008. We acknowledge the enthusiastic support of Dr. Jun'ichiro Kanamori, Director of IIAS, and his staff, as well as many external study group participants. We also acknowledge a preceding research project at IIAS on Art and Society, chaired by Professor Masako Sasaki, through which Yasuyuki Konparu became associated with IIAS, where Osamu Fujimura served as IIAS Fellow (2004 – 2005) participating in several study groups. Some speech scientists in Japan, including Sayoko Takano, Hirokazu Sato, and Ken'ichiro Sakakibara, contributed pertinent discussions in some of the study group meetings.

Professor Johan Sundberg organized the Jan Gauffin Memorial Symposium at the Royal Institute of Technology (KTH) in September 2008, and invited this paper as part of its program. Under the sponsorship of Professor Gunnar Fant, Osamu Fujimura spent one and a half year working daily with Jan Gauffin (Lindqvist at the time) from autumn 1963 through spring 1965 at KTH. This joint work resulted in a true friendship between Gauffin and Fujimura. They spent a few days together in Kamakura, toward the end of March 2007, when Gauffin made a trip to Japan knowing he was suffering from a terminal illness. They visited a Buddhist temple there and enjoyed the last time together, in the most quiet and peaceful surrounding.

With sincere and diligent collaboration of the current coauthors, this work is dedicated to the warm and fond memory of Jan Gauffin/Lindqvist of all his friends in the world who assembled in this symposium.

We are grateful for the hosts and sponsors of the most successful symposium.

8. References

Dudley, H. Remaking speech. J. Acoust. Soc. Am. **11**, 169-177. 1939.
Fant, G. Acoustic Theory of Speech Production. The Hague: Mouton Pub. 1960
Fant, G. Speech Acoustics and Phonetics. Dordrecht: Kluwer Pub. 2004
Fant, G., Liljencrants, J., & Lin, Q. G. A four-parameter model of glottal flow. STL-

QPSR, 4: 1-13, 1985.
Flanagan, Speech analysis synthesis and perception (second edition). Berlin: Springer Verlag. 1972.
Fujimura, O. An approximation to voice aperiodicity. IEEE Transaction of Audio & Electroacoustics. AU16, 68-72. 1968.
Fujimura, O. A note on voice fundamental frequency (pitch) in irregular voice. In Fujimura O. (ed.) Vocal Physiology: Voice Production, Mechanisms and Functions pp. 377-8. New York: Raven Press 1988.
Fujimura, O. The C/D model and prosodic control of articulatory behavior. Phonetica **57**, 128-38, 2000.
Fujimura, O. Onseikagaku Genron: Gengo no Honsitu o Kangaeru (Basic Speech Science: Exploration into the Nature of Language). Tokyo: Iwanami Publishers. 2007.
Fujimura, O. & Erickson, D. Acoustic Phonetics. In Hardcastle, W. J. & Laver, J. (eds.) Handbook of Phonetic Sciences, pp. 65-115. Oxford: Blackwell Pub. 1997.
Fujimura, O. & Williams, J. C. Syllable concatenators in English, Japanese, and Spanish. In Fujimura, O., Joseph, B. & Palek, B. (eds.) Proc. of LP'98, Prague: Charles University Press. pp. 461-98. 1999.
Fujimura, O. & Williams, J. C. Prosody and syllables. Phonological Studies **11**, 65-74. 2008
Fujisaki, H. Tanabe, Y. A time-domain technique for pitch extraction of speech. J. Acoust. Soc. Japan (E) **29**, 418–419, 1973..
Gauffin, J. & Sundberg, J. Spectral correlates of glottal souurce waveform characteristics, J. Speech & Hearing Research **32**, 556-65. 1989.
Honda, K., Takemoto, H., Kitamura, T., Fujita, S., & Takano, S. Exploring human speech production mechanisms by MRI. /IEICE Info. & Syst., E87-D/, 1050-1058. 2004
Kawahara, H., Masuda-Katsuse, I., 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**, 187-207, 1999.
Kawahara, H., Fujimura, O., Konparu, Y. Voice of Noh, 4th ASA/ASJ Joint Meeting, Honolulu, HI (paper 1pMU1), 2006.
Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T. TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. Proc. ICASP2008. Kawahara et al. 2008a.
Kawahara, H. Morise, M. Takahashi, Nisimura, T., R., Banno, H., Irino, T. A unified approach for F0 extraction and aperiodicity estimation based on a temporally stable power spectral representation, ISCA Tutorial and Research Workshop

(ITRW) on “Speech Analysis and Processing for Knowledge Discovery”, Aalborg, 4-6 June 2008. Kawahara et al. 2008b.
Kitamura, T., Takemoto, H., Adachi, S., Mokhtari, P., & Honda, K. [2006] Cyclicity of laryngeal cavity resonance due to vocal fold vibration. J. Acoust. Soc. Am., 120: 2239-2249.
Konparu, Y. Yorobousi to wa (About Yorobousi). Tôsin (Konparu Noh School Bul., Nara) No. 10. 2006.
Laver, J. The Phonetic Description of Voice Quality. Cambridge: Cambridge University Press 1980.
Liljencrants, J. Speech Signal Processing. In Hardcastle, W. J. & Laver, J. (eds.), The Handbook of Phonetic Sciences (Chapter 23). Oxford: Blackwell Pub. 1997.
Mazo, M., Erickson, D. & Harvey, T. Emotion and expression: Temporal data on voice quality in Russian lament. In O. Fujimura & M. Hirano (eds.), Vocal Fold Physiology: Voice Quality Control, pp. 173-87. San Diego: Singular Publ. 1994.
Sundberg, J. Formant structure and articulation of spoken and sung vowel. Folia Phoniatrica 22: 28-48, 1970.
Williams, J. C. Nô to koten girisya geki: Sankasya to site no kankyaku no kanten kara (From the audience-participant viewpoint: Noh and classical Greek theatre). Tôsin (Konparu Noh School Bulletin, Nara) No. **10**. 2006.

Appendix

Here is a Romanized transcript (English gloss in parentheses) of a short part of Yorobôsi (a story of an abandoned boy, who became blind, meeting his father in a religious setting, a portion sung in extreme joy), selected from our recording at the Public Noh Theatre in Nara (no audience). We also made a comparable recording (also by Y. Konparu) in an anechoic chamber at ATR.

“Ara omosiroya (How nice) ware (I) môməku (blind) to narazarisi (not became) saki (prior to) wa, yorobôsi ga tunewa (always) minaresi (familiar seeing) Kyôgai (surroundings) nareba (since it is). Nani (whatever) utagai (doubt) mo naniwae ni (none whatsoever, pun with place name Naniwa), kôgetu (moon over water) terasi (shining) syôhû (wind through pine) huku (blowing), eiyano (through the night) syôsei (sound of temple gong) nani (what) no nasu (brings about) tokoro (result) zo ya (is it). Sumiyosi (place name) no, matu (pine) no koma (gap of trees) yori (through), nagamure ba (when I look), tuki (moon) oti kakaru (about to fall). Awadi (place name) sima (island) yama (mountatin) to (as) nagamesi (I saw) wa tuki (moon) kage (appearance) no.”

Endnotes

[i] This paper was presented by Fujimura at the Jan Gauffin Memorial Symposium, October 16, 2008, Royal Institute of Technology (KTH) in Stockholm.
[ii] In alphabetical order.
[iii] There is some similarity between the Noh in Japan and an ancient theatrical art in Greece [Williams 2006].
[iv] The term site is used as the name of the role as the principal player in a Noh performance and subsidiary roles are called waki and ture. The same term site is used designating the title of the person who represents the school of Noh and normally plays the role of site representing the school. A site of a school could play a role of 'waki' in a joint performance of multiple schools in special occasions.
[v] The most representative playwright of classical Noh numbers is Zeami (1363? ~1443?). The Noh score is available for each play number for learners of "utai" (the singing of Noh), it is a booklet of the traditional Japanese orthographic text with annotations suggesting special (traditional) pronunciations of some phrases and diacritic marks for particular pitch controls.
[vi] See, e.g., the switching across the syllable boundary between [**Nai**] and [**na**], from the vowel gesture
[vii] The glide gesture of the phonologically diphthongal syllable /gaJ/ in Japanese is strongly palatalized and tense with a distinctly fronted tongue body [**i**], and in fact it can be interpreted that, in this singing, the high front vowel gesture is phonetically separated from the nucleus as another syllable. Surrounded by slashes in this note and elsewhere in the paper, we are using the phonemoidal transcription [Fujimura & Erickson 1997] of Japanese. The ususal IPA transcription is given in bold, surrounded by square brackets.
[viii] The modulation amplitude is so large that, in higher frequency region above F1, the harmonic structure is obscured.
[ix] Presumably, the laryngeal cavity resonance (e.g., 3 kHz, see infra) is caused by a tight glottal adduction resulting in a small open quotient of the glottal area function, and a very small damping of vocal tract resonances.
[x] The frequency modulation of F0 is multiplied according to the harmonic number and higher harmonics do not visibly maintain the harmonic structure.
[xi] Such a detection mechanism of repetitive patterns may characterize human (or animal) cognition systems in gneral, reflecting a basic biiological mechanism related to evolution of species and their behavior.
[xii] XSX does not necessarily produce perceptually salient pitch related representation. XSX picks up any periodicity (repetitive events, including only one pair of events) in any part of time-frequency space.
[xiii] Each constituent periodicity detector of XSX tuned to a specific F0 is designed to have frequency resolution equal to the specific F0. Note that with the current setting, sidebands are not picked up in Fig. 5.
[xiv] If one of the subharmonic components, i. e., a component having a frequency that is any odd integer multiple of 1/2F0, e. g., 5/2F0, half way between the second and the third harmonics of an almost periodic sound, has significantly strong enough amplitude, then, theoretically, the waveform's fundamental period corresponds to 1/2F0, not F0. The inverse 1/F0 = T is the time interval, over which a repetition of the same waveform appears. For example, if there is a significantly strong enough component at 1/2 F0 in the spectrum of voiced sound, the waveform must have some difference between one glottal period corresponding to F0 and the next, and therefore, strictly speaking, the repetition of the source time function occurs for a sequence of

such alternating time intervals, which approximately look like the glottal period. A slightly reduced amplitute for every other glottal period of the vocal fold vibration (probably due to hysteresis of the vocal fold tissue), for example, would create such a situation. If such an amplitude reduction of every other glottal period starts apeearing gradually, then, at some point, the waveform repetition starts to have an obviously doubled fundamental period, which must be heard as having 1/2F0. Conversely, if a waveform has alternating amplitude for every other glottal cycle and if this alteration is significant enough, and if this aternating amplitude change becomes weaker and eventually disappears, then, at some point, one must start to hear a pitch that is higher than before by an octave. There could be a transitional portion of such slowly changing sound, where the perceived pitch is not clear as to which octave interval the sound belongs to, but there will never be any tone between the two appearing with any different tone name (if the ear is musically trained). TANDEM-STRAIGHT's F0 candidate display can represent such a situation correctly by showing two coexisting frequency components, an octave apart, with the relative amplitudes continuously changing (see Fig. 4 for amplitude modulation of sinusoidal signal). A traditional pitch extractor cannot show this situation because it is designed to show one F0 value at any point. Such a situation occurs fairly commonly for normal voice especially toward the end of a voiced phrase, where the voice becomes gradually weakens with descending F0, or gradually going into vocal fry.