JOINT ESTIMATION OF VOICE SOURCE AND VOCAL TRACT PARAMETERS AS APPLIED TO THE STUDY OF VOICE SOURCE DYNAMICS

Hideki Kasuya *, Kikuo Maekawa †, and Shigeru Kiritani ‡
**Faculty of Engineering, Utsunomiya University, Utsunomiya, Japan, † The National Language Research Institute, Tokyo, Japan, ‡Graduate School of Medicine, University of Tokyo, Tokyo, Japan*

ABSTRACT

A novel method is presented for the joint estimation of voice source and vocal tract (formant/anti-formant) parameters from the acoustic speech signal. The method is based on the ARX (auto-regressive with exogenous input) model with a glottal flow waveform as an input to the system. A nonlinear optimization strategy is employed to estimate glottal flow parameters, whereas an extended Kalman filter algorithm is used to extract vocal tract parameters. The amount of turbulence noise components included in the voiced speech segment is also measured by a bank of band pass filters. The method is applied to the study of voice source dynamics as a conveyer of paralinguistic and vocal quality features that include perceived naturalness of an utterance. It is shown that this method is useful to gain profound understanding of voice source dynamics when used under certain limitations.

1. INTRODUCTION

Voice source dynamics convey not only linguistic but paralinguistic and vocal quality features. Perceived talker individuality is signaled by voice source dynamics as well. In order to lead to profound understanding of cognitive aspects of paralinguistic and vocal quality features, much effort has to be made to represent voice source dynamics quantitatively at articulatory, aerodynamic and acoustic levels. Paralinguistic and vocal quality features are conveyed primarily by quasi-periodic glottal flow dynamics and turbulence noise components. Many descriptive glottal flow models have been proposed [1-10]. The Liljencrants-Fant (LF)[3] and Rosenberg-Klatt (RK) [6]models, among other things, have gained much attention. The LF model at present seems to be a standard for glottal flow analysis. Meanwhile, the role of turbulence noise has been investigated primarily in relation to perceived vocal qualities of normal and dysphonic voices. Perceived breathy and strained vocal qualities are to a great extent associated with the amount and spectral characteristics of the noise in the speech signal [6,11-19]. Several methods have been proposed to estimate the statistical characteristics of the noise included in the speech signal [12,20-23], which have been successfully applied to sustained phonations of a vowel in most cases.

A linear source-filter model of speech production has played an important role in basic areas of speech production and perception [24-26]. However, deep insights into the acoustic and aerodynamic properties of voice source dynamics have given rise to the introduction of an interactive voice source model with sub-glottal and supra-glottal systems [3,27-31]: *e.g.* acoustic coupling with the sub-glottal system in breathy voice, flow ripples observed during the glottal opening phase, carry-over of a response to the glottal pulse to the next cycle in high-pitch voice, effects of the narrow vocal tract constriction, and so on. It is very difficult, however, to incorporate all of these interactive phenomena into a well-formed mathematical model of the glottal flow. In this paper, we restrict ourselves to a rather classic linear source-filter model, since we believe that this simple approximation can still provide many fruitful observations of voice source dynamics and could be extended to more complicated, realistic situations in the next step.

Automatic and semi-automatic inverse filtering methods have been extensively used to estimate the glottal flow waveform from a speech signal that has been carefully recorded on a digital audio tape-recorder to avoid phase distortion [32-42]. In this method, one has to have full knowledge of the acoustic characteristics of the vocal tract before performing inverse filtering, while one needs to know the voice source properties to estimate the vocal tract characteristics, thus giving rise to a contradiction. Vocal tract characteristics are usually estimated in an assumed closed interval of the signal using an LPC-based method. This method brings about another problem, *i.e.* difficulty in finding a closed interval and in applying it to speech of a short closed interval. The idea of joint estimation of both glottal flow and vocal tract parameters for vowel-type speech sounds was proposed by Fujisaki and Ljungqvist [4] and by Milenkovic [43]. But they suffer to some extent from instability of the convergence of the parameters to be estimated and from difficulty in applying to the speech utterance including anti-resonance.

In this paper, we propose a novel method for the joint estimation of voice source and vocal tract parameters from a speech signal [44,45] and attempt to separate turbulence noise components included in the utterance, which often play an important role as an acoustic correlate of perceived vocal qualities. Examples of its application to the study of voice source dynamics as a conveyer of paralinguistic and vocal quality features are then presented. An integrated software system based on the method is also briefly described which consists of analysis, editing and synthesis modules [45].

2. ARX SPEECH PRODUCTION MODEL

2.1 Voicing Source Model

Many descriptive models have been proposed to quantitatively represent glottal flow [1-10]. The Liljencrants-Fant (LF) model [3] has gained much attention and has become a reference for glottal flow analysis. However, it is somewhat complicated for speech analysis-synthesis applications. The Rosenberg-Klatt model [1,6], on the other hand, is mathematically simple but to a certain extent has limitations in representing a large variety of glottal flow waveforms. It has recently been revised to overcome this drawback [9]. In this paper, we use the RK model to represent a differentiated glottal flow including radiation characteristics, because of its mathematical simplicity. The joint estimation method bellow can easily be extended to more sophisticated flow models. The RK model generates a rudimentary waveform defined as

$$g(n) = 2an - 3bn^2, \quad 0 \le n \le T * OQ$$

$$= 0, \qquad T * OQ < n < T,$$

$$a = (27 * AV)/(4 * OQ^2 * T)$$

$$b = (27 * AV)/(4 * OQ^3 * T^2),$$

where *T* is the pitch period, *AV* the amplitude parameter and *OQ* the open quotient of the glottal open phase for the duration of a complete glottal cycle. $g(n)$ is zero in the closed phase. The differentiated glottal flow waveform $u(n)$ is generated by smoothing $g(n)$ with a low pass filter where tilt of the spectral envelope is adjusted by a spectral tilt parameter *TL* (Figure 1).

2.2 ARX Speech Production Model

The speech production process is modeled as an IIR filter with an equation error as follows:

$$s(n) + \sum_{i=1}^{p} a_i s(n-i) = \sum_{j=1}^{q} b_j u(n-j) + e(n),$$

where $s(n)$ and $u(n)$ denote a speech signal and a differentiated glottal flow waveform at time n, respectively. In the equation, a_i and b_j are vocal tract filter coefficients, p and q are model orders, and $e(n)$ the equation error. When $e(n)$ is assumed to be white, the equation is referred to as the ARX (auto-regressive with exogenous input) model. By performing the *z*-transform on the equation, one gets the following:

$$S(z) = \frac{B(z)}{A(z)} U(z) + \frac{1}{A(z)} E(z),$$

where $S(z), U(z)$ and $E(z)$ are the *z*-transform of the speech signal $s(n)$, voicing source $u(n)$, and equation error $e(n)$, respectively. $B(z)/A(z)$ indicates the vocal tract transfer characteristic for the voiced sound, $1/A(z)$ represents the one for the unvoiced sound. When the zeros of the acoustic transfer characteristics for the unvoiced segment are considered, it results in the ARMAX (auto-regressive moving average with exogenous input) model, which is usually a more sophisticated approximation to the speech production. Since the ARMAX model brings about another problem, *i.e.* increase in the number of model parameters to be estimated, we use the ARX model at present as an approximation of the transfer characteristics of the turbulence noise produced in the vocal tract.

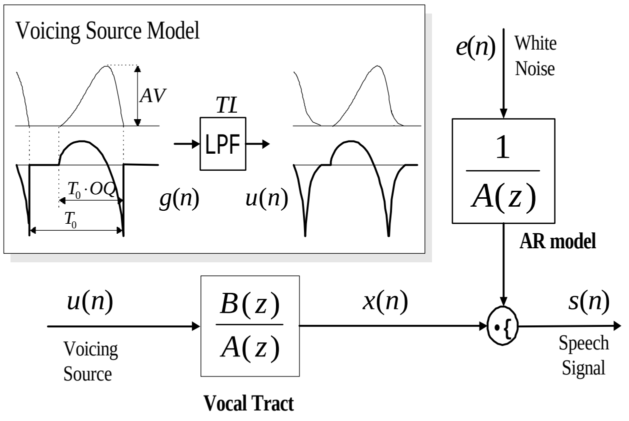
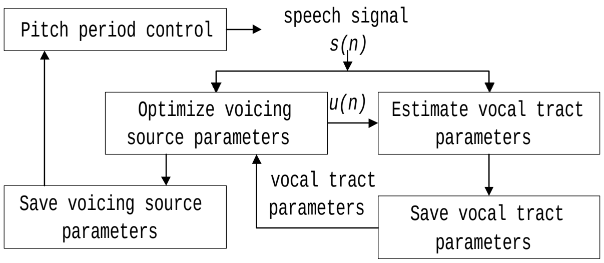


Figure 1. ARX (auto-regressive with exogenous input) model and Rosenberg-Klatt voicing source waveform.

3. ANALYSIS ALGORITHM

A speech signal is composed of voiced, unvoiced and silent segments. The voiced segment includes not only quasi-periodic components but sometimes turbulence noise, as in breathy and strained phonations. The ARX speech production model described above implements in nature noise generation process, although an acoustic transfer characteristic of the noise is slightly different from that of an interactive system [6,19].

The analysis frame is synchronized with an estimated pitch period in the quasi-periodic segment, whereas it is shifted every 5 ms in the aperiodic segment. Voiced segments are specified beforehand by the user with an interactive labeling tool. For the voiced segment, the key point of the analysis is how to divide the speech signal into the glottal waveform approximated by the RK model and the vocal tract characteristic represented by the system transfer function $B(z)/A(z)$. An adaptive pitch-synchronous algorithm has been developed to estimate the voicing source and vocal tract parameters directly from the speech signal [44,45]. Figure 2 represents the block diagram of the analysis procedure for a voiced segment. The voicing source parameters in the current analysis period are optimized so as to minimize the mean square equation error (MSEE). In order to estimate the vocal tract (IIR) filter coefficients, the filter is ex-



Pitch period control
speech signal
s(n)
Optimize voicing source parameters
u(n)
Estimate vocal tract parameters
Save voicing source parameters
vocal tract parameters
Save vocal tract parameters

Figure 2. Block diagram of the pitch-synchronous ARX analysis of voiced segments.

panded into a time-variant system so that a Kalman filter algorithm can be used. In the algorithm, the input signal is a differentiated glottal flow waveform $u(n)$ approximated by optimized values of the RK parameters and the output is a speech signal $s(n)$. In the estimation of the RK parameters, the optimization was made by the simulated annealing algorithm which usually required much computation time [44]. Taking sensitivity of each of the RK parameters to the MSEE value into account, an improved algorithm is presented where only OQ is optimized and the other parameters are estimated in different ways described below.

3.1. Estimation of Voicing Source Parameters

Optimization of the RK voicing source parameters is performed so as to minimize the MSEE value as mentioned above. The MSEE is a function of the RK parameters, *AV, TL* and *OQ*. *AV* is estimated as the coefficient b_0 of $B(z)$ using the Kalman filter algorithm. *TL*, on the other hand, is estimated from the inverse filtered signal using the method proposed by Lu, *et al* [39]. These simplifications save computation time significantly without sacrificing accuracy. *OQ*, which is important and highly sensitive to the MSEE, is estimated by a simpler optimization method [45]. The estimation is carried out on a cycle-to-cycle basis. In addition to the glottal waveform, a glottal noise component defined by the noise amplitude parameter (*NA*) is estimated from the prediction error within the glottal open phase.

3.2. Estimation of Vocal Tract Parameters

In order to estimate the coefficients of the vocal tract transfer function, the same algorithm as the one proposed by Ding, *et al.* [44] is used, except that the system gain parameter $b_0(n)$, being related to the amplitude of the glottal flow waveform, is newly introduced in the equation. Although the time-varying filter coefficients are computed at every sampling point in the iteration, the formant-antiformant values are obtained by solving for the roots of the polynomials at only three points within the pitch period: the last of the period, the middle of the open phase and the middle of the closed phase. In theory, a set of formant values estimated at every point during the iteration of the analysis should be used, but in practice it fails once in a long while. Therefore, a set of formant-antiformant values, which has the maximum number of formants, is saved as the representative of the vocal tract parameters within the period. This is because 1) calculation of formant values at every sampling point takes immense computation time and a large capacity for the storage of the parameters is needed, 2) an estimated vocal tract filter is not always stable around the glottal closure instant, and finally 3) it is not unreasonable to assume time-invariant nature of the vocal tract movement within one pitch period. From preliminary experiments, we have also found that the filters estimated at the three points within the period mentioned above are stable.

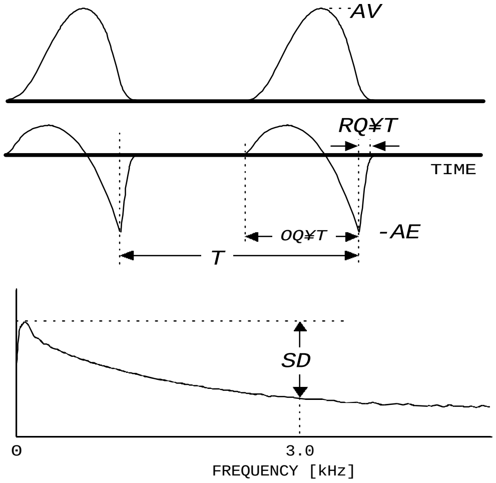


Figure 3. Voicing source parameters derived from the RK glottal flow waveform model.

3.3. Selection of Model Order

For the selection of an optimum order of the model, the method by Ding, *et al*. is used [44]. If this does not work well, the user can easily change the order.

3.4. Voicing Source Parameters

As in the case of the LF model [8], various voicing source parameters are derived from the RK waveform (Figure 3). Pitch frequency (*F0*), amplitude of the voicing waveform (*AV*) and open quotient (*OQ*) parameters are primary voicing source parameters that are already included in the RK waveform model. In addition to these three parameters, we introduce three more parameters: the negative peak value of $u(n)$ (*AE*), which is nearly proportional to AV/OQ in the RK model and corresponds to E_e in the LF model, spectral decay (*SD*) in dB at 3 kHz from the maximum value of the spectrum of the RK waveform and a return phase quotient (*RQ*) which is defined as the ratio of the interval needed for a unit step response of a low pass filter (*TL*) to reach a value of 0.9 to the pitch period *T*. *AE* is an amplitude of a differentiated glottal flow at the excitation point and a direct determinant of the amplitude of the output speech signal. Difference between the fundamental frequency component level (*H1*) and the second harmonic level (*H2*) is also employed as an indicator (*H1-H2*) of loose contact of the vocal folds during phonation [46].

To summarize, we have *T0* (=1/*F0*), *AV*, *AE*, *OQ* and *RQ* as the time-domain voicing source parameters, *F0*, *SD* and *H1-H2* as the frequency-domain parameter, and the turbulence noise parameter *NNP* described in the next section.

4. ESTIMATION OF TURBULENCE NOISE

Turbulence noise sometimes exists in the voiced speech segment, typically in breathy phonation. Importance of the noise is well known for the female voice, breathy voice and strained voice. Several methods have been proposed to estimate the noise component included in the voiced speech segment [20-23]. They have proved to work well for sustained phonation of a vowel but are a little short of accuracy when applied to continuous speech.

In this paper, we present a simple but robust method to measure the amount of the noise. A speech signal is fed to a bank of band-pass filters. On each of the outputs, a modified autocorrelation function is computed. If the function value regarded as periodic, otherwise as aperiodic. The sum of the powers of the aperiodic outputs of the band-pass filters are obtained every 5 ms and normalized noise power (= 10 log[noise power/total signal power]) is computed. The normalized noise power (*NNP*) is a quantitative measure of the amount of the noise component.

5. ANALYSIS-EDITING-SYNTHESIS SYSTEM

5.1. Synthesis Strategy

A cascade formant synthesizer is used to synthesize both voiced and unvoiced speech. The simplicity is the advantage of our system over the one used by Klatt [6], in which a complicated cascade-parallel structure is employed. The RK model is used to synthesize voiced speech, whereas the M-sequence white noise is used to synthesize unvoiced speech [45]. In the formant synthesizer, gain of each of the second order resonators is always 0 dB at 0 Hz. But the amplitude of the vocal tract transfer function $B(z)/A(z)$ that is estimated by the Kalman filter may not always be 0 dB at 0 Hz. As a result, an estimated amplitude parameter AV ($= kb_0$, k : constant), can not be used directly in such a formant synthesizer. Adjustment of the amplitude parameter *AV* is made so as to minimize the root mean square error between the original and synthesized speech signals. Since we define the pitch period *T* as an interval between the two successive *GCI*'s in the synthesis, the RK glottal waveform of one pitch period begins with the close phase of the previous cycle, followed by the RK waveform of the open phase of the current cycle. In such a way, the *GCI* keeps the same value, even when the *OQ* is changed. In order to simulate the turbulence noise in the voiced component, a value of the *NNP* is converted to an amplitude parameter *NA* that drives a vocal tract filter $1/A(z)$. The output of the filter $1/A(z)$ is added to the output of the filter $B(z)/A(z)$ driven by a glottal flow $u(n)$, resulting in the synthetic speech signal (Figure 1). For the unvoiced segment, an exogenous input $x(n)$ is regarded as zero.

5.2. Analysis-Editing-Synthesis System

Each of the estimated voice source and formant/anti-formant parameters is re-sampled every 5 ms and displayed on the graphical window. An editing tool as a part of the system enables the user to manipulate the parameters on the graphical window. A number of editing functions are provided: modification, addition and deletion of the value, simultaneous modification and scaling of the multiple values, saving and printing the parameter values, and so on [45]. Using the parameter values so modified, one can immediately synthesize speech.

6. EVALUATION OF THE ARX METHOD

Since the formant estimation algorithm, which is essentially the same as the one developed in our previous paper [44], was already evaluated using synthetic speech signals and proved to be much superior to some other well-accepted methods [44], we focus in this paper on the accuracy of the voice source parameters estimated by the ARX method.

6.1. Experiment

A male adult pronounced in an anechoic room a Japanese word /e'ki/ ("railroad station" in English) in three paralinguistic renditions, *i.e.* the neutral (declarative), suspicion and disappointment. The apostrophe indicates the lexical accent. A speech signal was recorded together with a high-speed digital image of the vocal cord vibration obtained through a fiberscope and an electro-glottographic (EGG) signal, as shown in Figure 4 [47]. The digital images were recorded at a rate of 4,500 frames per second. The vowel segment of /e/ of the neutral utterance was subjected to the analysis for the evaluation.

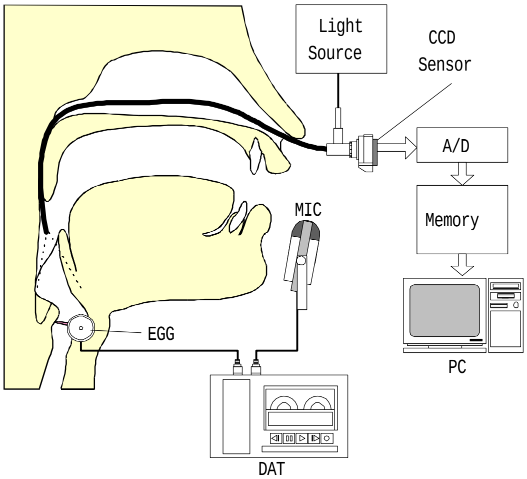


Figure 4. Experimental setup to record the speech signal, EGG signal and high-speed image through a fiberscope.

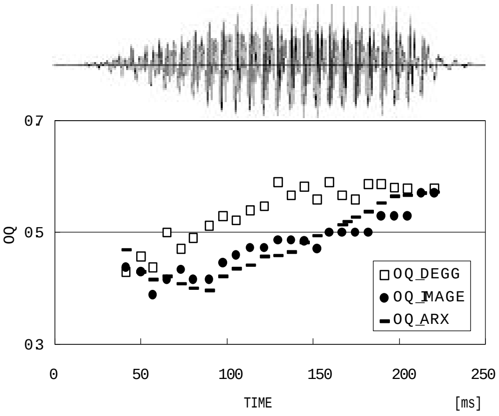


Figure 5. The speech waveform and OQ trajectories measured from the DEGG signal, high-speed digital image and ARX method for the vowel segment /e/ of a Japanese word /e'ki/.

6.2. Results and Discussion

A comparison was made on the glottal closure instant (*GCI*) and open quotient (*OQ*) values between the measurements from EGG signals and the ARX method. The *GCI* was defined on the EGG signal as a time point indicating a negative peak of a differentiated EGG (DEGG) signal. Standard deviation of the difference between the two measurements was 0.077 ms, ensuring agreement between the two measurements. Figure 5 illustrates the *OQ* values estimated by the three methods, *i.e*. the digital image, EGG and ARX, for the same vowel segment. On the DEGG signal, an open phase was defined as an interval between the positive and the negative peaks within the pitch period. Although a tendency for the *OQ* value to increase gradually toward the end of the vocalic segment of this particular sample is similar among the three measurements, the *OQ* values measured by the ARX method agree well with the ones obtained from the high-speed digital images. This comparison certainly ensures reliability of the voice source measurements by the ARX method, but it should be noted that the measured values depend upon an initial value in the optimization. When a reasonable initial value is provided for a vocalic segment, our joint estimation algorithm seems to keep track of the voice source dynamics quite well.

7. VOICE SOURCE DYNAMICS ASSOCIATED WITH PARALINGUISTIC ASPECTS OF THE UTTERANCE

Paralinguistic information of the utterance is largely conveyed by voice source dynamics [48-50]. Significant acoustic features of the dynamics, among others, include pitch frequency (*F0*) pattern, F0-Range, glottal flow intensity, glottal flow waveform (or its spectral characteristic as a counterpart) and its temporal property, turbulence noise and syllable duration. Our analysis-editing-synthesis system described earlier is quite useful for investigating systematically roles of these acoustic features in terms of cognition of paralinguistic and vocal quality features. Hypotheses can be tested using synthetic speech that is generated by the system by manipulating the acoustic parameters obtained from the utterance. This section presents an application of the ARX system to the study of voice source dynamics associated with paralinguistic and vocal quality features.

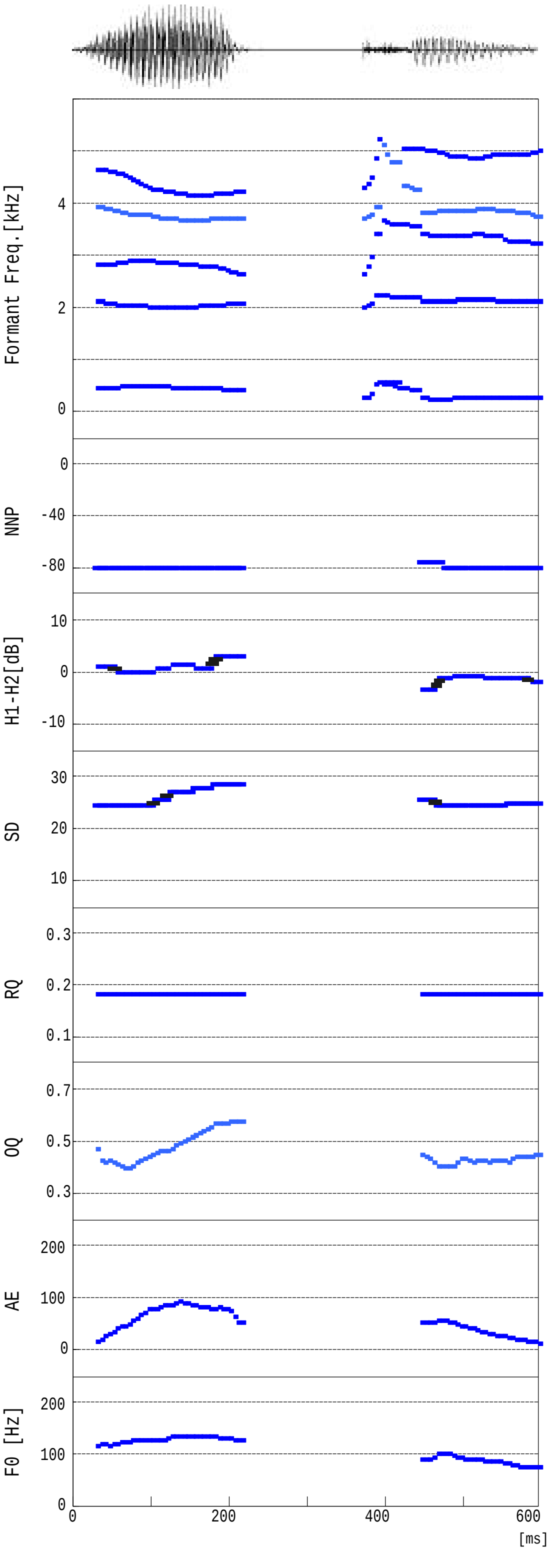
7.1. Experiment

The voicing source and noise parameters were analyzed for the speech and EGG signals and high-speed digital images of the utterances in the three paralinguistic renditions described in the previous section. Synthetic speech generated by the system was also used for informal perceptual tests.

7.2. Results and Discussion

Figures 6 (a), (b) and (c) illustrate panels of the speech waveform and the voice source parameters analyzed in the neutral, suspicion, and disappointment renditions, respectively. From the top, are shown the speech waveform, the first five formant frequency trajectories, *NNP*, *H1-H2*, *SD*, *RQ*, *OQ*, *AE*, and *F0*.

Let us first consider voice source dynamics in the suspicious utterance in relation to the "neutral." Effects of



TIME

Figure 6(a). A panel of the speech waveform, first five formant frequency trajectories and various voicing source parameters for the neutral utterance /e'ki/.

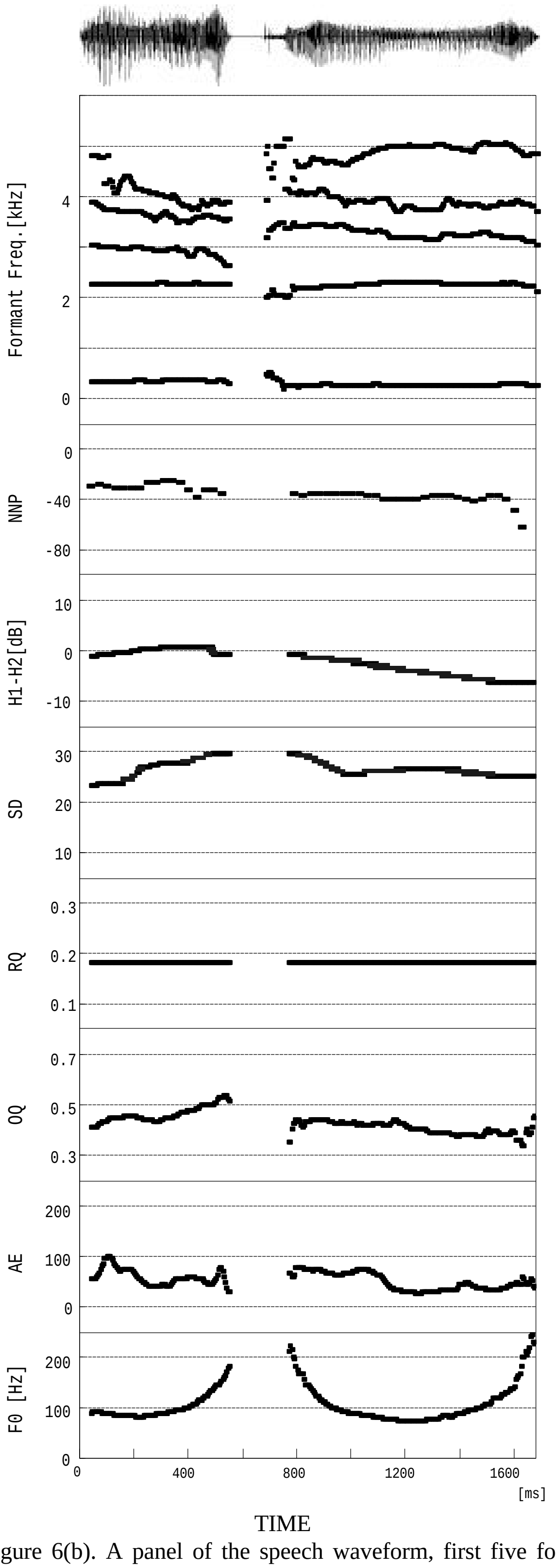


Figure 6(b). A panel of the speech waveform, first five formant frequency trajectories and various voicing source parameters for the suspicious utterance /e'ki/.

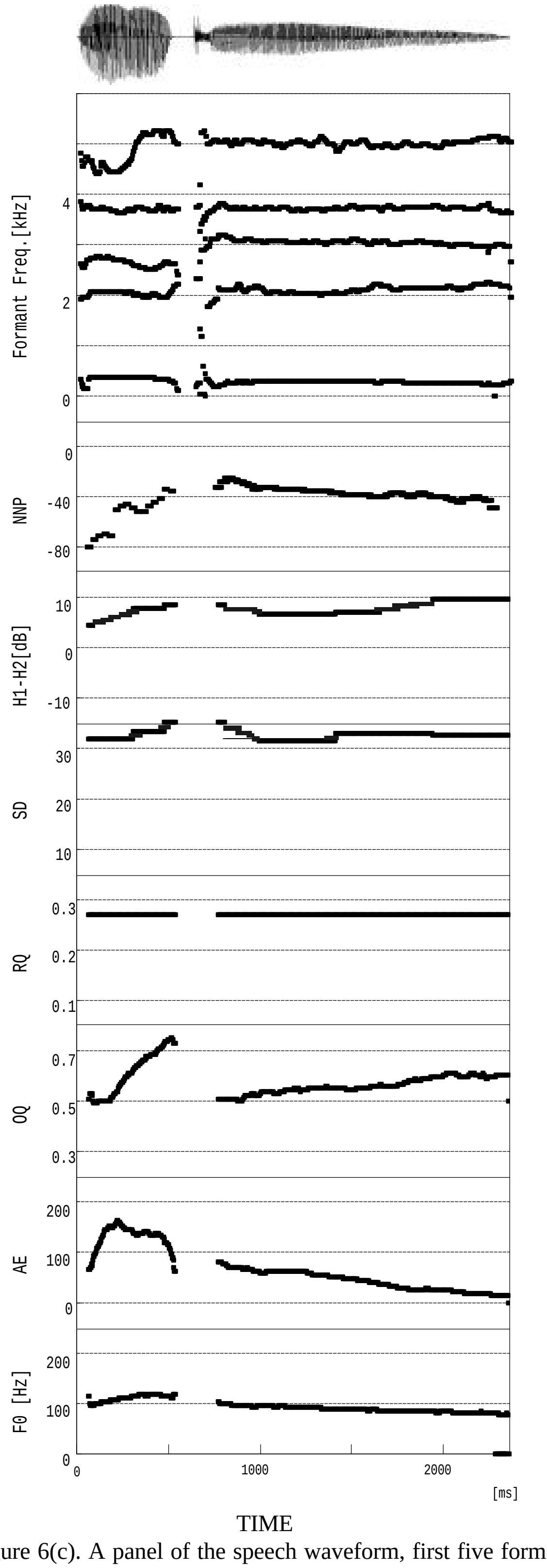


Figure 6(c). A panel of the speech waveform, first five formant frequency trajectories and various voicing source parameters for the disappointed utterance /e'ki/.

paralinguistic aspects of the utterance on the *F0* pattern have extensively been investigated [50]. Increase in *F0* values in the final syllable and thereby in the *F0* dynamic range has been reported to be a significant acoustic correlate of “suspicion” as can be seen in Figure 6 (b). *F0* increases gradually toward the final part of the first syllable /e/ and then rapidly falls down in the initial part of the second syllable /ki/ to realize accentuation placed on the first syllable /e/. *F0* finally moves up to signal “suspicion” in the final part of the second syllable. This *F0* pattern as well as longer duration of the utterance could be a primary acoustic correlate of “suspicion.” In addition to the *F0* control as mentioned above, this particular speaker maneuvers the larynx in a more complicated manner to emphasize “suspicion.” The *OQ* contour keeps very small values even in the last part of /e/, whereas the neutral utterance in Figure 6 (a) reveals higher values of *OQ* in the corresponding portion, due to transition to the glottal opening for the articulation of a consonant /k/. The excitation amplitude (*AE*) reveals a very complex pattern in the first syllable as compared with the one of “neutral”, which suggests slightly unstable vocal fold vibrations. This unstable vocal fold vibration was observed in the high-speed images of the vocal fold vibration as well. The high-speed images also show that the laryngeal framework is constricted during the vowel and even in the voiceless portion of the consonant /k/ and the left and right arytenoid cartilages are in contact. These observations may suggest high activities of the internal laryngeal muscles in the first syllable to emphasize the degree of suspicion. The perceived vocal quality of the first syllable was clearly pressed and harsh [51], which is reflected in the increased *NNP*. This large noise measure is thought to be caused by aperiodicity of the vocal fold vibration rather than generation of turbulence noise around the glottis. In the vocalic segment of the second syllable, /i/, the vocal quality is still pressed and harsh; the noise measure again keeps a large value and the open quotient becomes further smaller toward the end of the segment. The small OQ value and shorter pitch period in the time domain are interpreted in the frequency domain as the decrease in the values of the spectral parameters *SD* and *H1-H2*.

Meanwhile, “disappointment” reveals a low and flat *F0* contour and very long duration of the utterance as shown in Figure 6 (c). Both *OQ* and *RQ* show larger average values than “suspicion” and the “neutral.” In particular, *OQ* becomes considerably larger toward the final parts of the first syllable /e/ and the second syllable /ki/. Large values of *SD* and *H1-H2* in the spectral domain reflect well these tendencies observed in the voicing source waveform. In the final part of the first syllable /e/, *NNP* increases, which implies existence of turbulence noise or strong aperiodicity as in the case of “suspicion.” However, the high-speed digital images indicate that the left and right arytenoid cartilages are always separated at the back in the same vowel segment. In fact, vocal quality of this segment is clearly breathy, suggesting existence of turbulence noise a little downstream at the glottis. This is a good example of how the source of the noise figure *NNP* can be interpreted well by taking other voicing source parameters into account. Large OQ and RQ suggest low activities of the internal laryngeal muscles.

The ARX analysis-editing-synthesis system is a useful tool to investigate systematically perceptual significance of each of the voicing source parameters, since it provides with a convenient speech synthesis environment in manipulating the parameter values. Using this system, informal perceptual experiments were performed to ascertain the contribution of each of the voicing source and noise parameters in the perception of paralinguistic information. *F0* and *AE* trajectories were the most important parameters as the acoustic correlates of the three types of speaking styles. *NNP* and *OQ* trajectories also proved to enhance each speaking style, particularly in “suspicion.”

8. CONCLUSION

An automatic method based on the ARX (auto-regressive with exogenous input) model is presented for the joint estimation of voice source and vocal tract (formant/anti-formant) parameters from the speech signal. A nonlinear optimization algorithm employed in the method does not always guarantee the best estimate of the parameters, but reasonable initial values for the optimization help the algorithm operate successfully. The algorithm has been developed on the MS-Windows platform, providing with a useful environment for the analysis-synthesis experiment.

The system was applied to the study of voice source dynamics associated with paralinguistic and vocal quality features. Preliminary perceptual experiments showed that not only the pitch and intensity patterns but the dynamics of other voicing source parameters as the amount of turbulence noise present in the voiced segment, fluctuations of the source parameters and the glottal flow waveform were significantly important as a conveyer of paralinguistic and vocal quality features. We believe that the concept of the joint estimation proposed in this paper could develop further by incorporating more sophisticated glottal flow models which typically take the interaction with the sub- and supra-glottal systems into account.

ACKNOWLEDGMENTS

This work has been supported in part by the Grant-in-Aid for Scientific Research (10878051 and 10410110) from the Ministry of Education, Science and Culture. Special thanks are due to Dr. W.-Z. Zhu who developed an interactive software system of the ARX method, Dr. E. Murano, Mr. M. Matsuda and Mr. H. Imagawa for their valuable comments and assistance made during the work.

REFERENCES

[1] Rosenberg, A. 1971. Effect of glottal pulse shape on the quality of natural vowels. J. Acoust. Soc. Amer., 49, 583-590.
[2] Fant., G., Liljencrants, J. and Lin, Q. 1985. A four-parameter model of glottal flow. *STL-QPSR*, 4/85, 1-3.
[3] Fant, G. 1986. Glottal flow: models and interaction. *Journal of Phonetics*, 14, 393-399.
[4] Fujisaki, H. and Ljungqvist, M. 1986. Proposal and evaluation of models for the glottal source waveform. *Proc. Int. Conf. Acoust. Speech and Signal Process.*, ICASSP-86, 1605-1608.
[5] Fant, G. and Lin, Q. 1988. Frequency domain interpretation and derivation of glottal flow parameters. *STL-QPSR*, 2-3/88, 1-21.
[6] Klatt, D. and Klatt, L. 1990. Analysis, synthesis and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Amer.*

87, 820-857.

[7] Qi, Y. and Bi,. 1994. A simplified approximation of the four-parameter LF model of voice source. *J. Acoust. Soc. Amer*. 49,583-590.

[8] Fant, G. 1995. The LF-model revisited. Transformations and frequency domain analysis. *STL-QPSR* 2/95, 1-83

[9] Veldhuis, R. 1998. A computationally efficient alternative for Liljencrants-Fant model and its perceptual evaluation. *J. Acoust. Soc. Amer*. 103, 566-571.

[10] Alku, P., Vilkman, E., and Laukkanen, A.-M. 1998. Parameterization of the voice source by combining spectral decay and amplitude features of the glottal flow. *J. Speech, Language, and Hearing Research*, 41, 990-1002.

[11] Fritzell, B., Hammarberg, B., Gauffin, J., Karlson, I., and Sundberg, J. 1986. Breathiness and insufficient vocal fold closure. *Journal of Phonetics*, 14, 549-553.

[12] Kasuya, H., Ogawa, and H., Kikuchi, Y. 1986. An adaptive comb filtering method as applied to acoustic analysis of pathological voice. *Proc. ICASSP*, 1, 669-672.

[13] Holmberg, E.B., Hillman, R.E. and Perkell, J.S. 1988. Glottal airflow and pressure measurements for soft, normal and loud voice by male and female speakers. *J. Acoust. Soc. Amer*. 84, 511-529.

[14]Price, P.J. 1989. Male and female voice source characteristics: Inverse filtering results. *Speech Communication*, 8, 261-277.

[15] Kasuya, H. and Ando, Y. 1991. Acoustic analysis, synthesis, and perception of brathy voice. In: Gauffin, J. and Hammarberg, B. (Eds.) *Vocal fold physiology*, 251-258.

[16]Childers, D.G. and Lee, C.K. 1991. Voice quality factors: Analysis, synthesis and perception. *J. Acoust. Soc. Amer*. 90, 2394-2410.

[17] Karlson, I. 1992. Modelling voice variations in female speech synthesis. *Speech Communication*, 11, 491-495.

[18] Sodersten, M., Hertegard, S., Hammarberg, B. 1995. Glottal closure, transglottal air flow, and voice quality in healthy middle-aged women. *J. Voice*, 9, 182-197.

[19] Hanson, M. 1997. Glottal characteristics of female speakers: Acoustic correlates. *J. Acoust. Soc. Amer*. 101, 466-481.

[20] Yang, C.-S. and Kasuya, H. 1995. Least square estimation of laryngeal noise in speech signals. *J. Acoust. Soc. Jpn*., 16, 123-126.

[21] Stylianou, Y., Laroche, J., and Moulines, E. 1995. High-quality speech modification based on a harmonic+noise model. *Proc. EUROSPEECH*, 451-454.

[22] Yegnanarayana, B., d'Alessandro, C., and Darsinos, V. 1998. An interactive algorithm for decomposition of speech signals into periodic and aperiodic components. *IEEE Trans. Speech and Audio Processing*, 6, 1-11.

[23] d'Alessandro, C., Darsinos, V. and Yegnanarayana, B. 1998. Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources. *IEEE Trans. Speech and Audio Processing*, 6, 12-23.

[24] Stevens, K.N. and House,A.S. 1955. Development of a quantitative description of vowel articulation. *J. Acoust. Soc. Amer.*, 27, 484-493.

[25] Fant, G. 1960. *The acoustic theory of speech production*. The Hague, Mouton (2nd edition 1970).

[26] Flanagan, J. L. 1972. *Speech analysis, synthesis, and perception* (2nd edition,1972). New York, Springer Verlag.

[27] Rosenberg, M. 1983. An interactive model for the voice source. In:Bless, D.M. and Abbs, J.H. (Eds.) *Vocal Fold Physiology*, College-Hill: San Diego,CA, 155-165.

[28] Bickley, C.C. and Stevens, K.N. 1986. Effects of a vocal tract constriction on the glottal source: Experimental and Modeling Studies. *Journal of Phnetics*, 14, 373-382.

[29] Childers, D.G. and Wong, C.-F. 1994. Measuring and modeling vocal source-tract interaction. IEEE Trans. Biomedical Engr., 41, 663-671.

[30] Fant, G. 1997. The voice source in connected speech. Speech Communication, 22, 125-139.

[31] Titze, I.R. and Story, B.H. 1997. Acoustic interactions of the voice source with the lower vocal tract. J. Acoust. Sco. Amer., 101, 2234-2243.

[32] Gobl, C. 1988. Voice source dynamics in connected speech. *STL-QPSR*, 1/88, 123-159.

[33] Gobl, C. and Chasaide, A.N. 1988. The effects of adjacent voiced/voiceless consonants on the vowel voice source: A cross Language study. *STL-QPSR*, 2-3/88, 23-59.

[34] Gobl, C. 1989. A preliminary study of acoustic voice quality correlates. *STL-QPS*R, 4/89, 9-22.

[35] Gauffin, J. and Sunberg, J. 1989. Spectral correlates of glottal voice source waveform characteristics. *J. Speech and Hearing Research*, 32,556-565.

[36] Strik, H. and Boves, L. 1992a. Control of fundamental frequency, intensity and voice quality in speech. *Journal of Phonetics*, 20, 15-25.

[37] Strik, H. and Boves, L. 1992b. On the relation between voice source parameters and prosodic features in connected speech. *Speech Communication*, 11, 167-174.

[38] Alku, P. 1992. Glottal wave analysis with pitch-synchronous iterative adaptive inverse filtering. *Speech Communication*, 11, 109-118.

[39] Lu, J.-L., Ando, Y., and Kasuya, H. 1992. Study of voicing source characteristics based on a semiautomatic glottal inverse filtering method. *J. Acoust. Soc. Jpn*., 48, 642-648 (in Japanese).

[40] Hertegard, J. , Gauffin, J. and Karlson, I. 1992. Physiological correlates of the inverse filtered flow waveforms. *J. Voice*, 4,52-63.

[41] Fant, G. and Kruckenberg, A. 1996. Voice source properties of the speech code. *TMH-QPSR*, 4/96, 45-56.

[42] Alku, P. and Vilkman, E. 1996. Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering. *Speech Communication*, 18, 131-138.

[43] Milenkovic, P. 1986. Glottal inverse filtering by joint estimation of an AR system with a linear input model. *IEEE Trans. ASSP*, 34, 28-42.

[44] Ding, W., Kasuya, H., and Adachi, S. 1995. Simultaneous estimation of vocal tract and voice source parameters based on an ARX model. *Trans. on Information and Systems, Institute of Electronics, Information and Communication Engineers of Japan*, E78-D, 738-743.

[45] Zhu, W.-Z and Kasuya, H. 1998. A speech analysis-synthesis-edting system based on the ARX speech production model. *J. Acoust. Soc. Jpn*., 19, 223-230.

[46] Stevens, K.N. and Hanson, M. 1994. Classification of glottal vibration from acoustic measurements. In :Fujimura, O. and Hirano, M. (Eds.) *Vocal Fold Physiology*. Singular Publishing Group: San Diego,CA, 147-170.

[47] Kiritani, S., Imagawa, H., and Hirose, H. 1996. Vocal cord vibration in the production of consonants- Observation by means of high-speed digitaln imaging using a fiberscope. *J. Acoust. Soc. Jpn*., 17, 1-8.

[48] Fant, G. and Kruckenberg, A. 199. The voice source in prosody. Proc. ICPhS95, 2, 622-625.

[49] Fant, G., Hertegard, S. and Kruckenberg, A. 1996. Focal accent and subglottal pressure. *TMH-QPSR*, 2/96, 29-32.

[50]Murray, I.R. and Arnot, J.L. 1993. Towards the simulation of emotion in synthetic speech: A review of the literature of human vocal emotion. *J. Acoust. Soc. Amer*., 93, 1097-1108.

[51] Kasuya, H. and Yang, C.-S. 1995. Voice quality associated with voice source. *J. Acoust. Soc. Jpn*., 51, 869-875 (in Japanese).