ROLES OF VOICE SOURCE DYNAMICS AS A CONVEYER OF PARALINGUISTIC FEATURES

Hideki KASUYA*, Masanori YOSHIZAWA*, and Kikuo MAEKAWA**

* Faculty of Engineering, Utsunomiya University ** National Language Research Institute

ABSTRACT

This paper focuses on the role of frequency spectral tilt (TL) of the voicing source waveform, the glottal turbulent noise and fluctuations of the source spectrum on period-to-period basis in the cognition of paralinguistic features as in suspicious and disappointed renditions. The TL is closely related to duration of the return phase of the differentiated glottal waveform. Perceptual experiments have shown that the larger the TL, the lower the degree of suspicion and existence of the source spectral fluctuations increases that degree over the entire range of the TL. In the case of disappointment, the larger the TL, the higher the degree of disappointment and inclusion of the glottal turbulent noise emphasizes disappointment when the TL is small.

1. INTRODUCTION

Acoustic correlates of paralinguistic features include (1) F0 contour, (2) F0 range, (3) intensity, (4) glottal flow waveform, (5) amount of the glottal turbulent noise, (6) fluctuations in the glottal flow, e.g. jitter, shimmer and flow waveform variations, (7) syllable duration, and (8) formant characteristics. The correlates from (1) through (6) are all associated with voice source dynamics. Among these source characteristics, the F0 pattern and F0 range have been primarily studied by many researchers and shown to be salient acoustic cues to paralinguistic features [1,2]. The syllable duration and formant characteristics [3] have also been reported to be an essential conveyer of paralinguistic features. Less attention has been paid, however, to roles of the glottal flow waveform shape, glottal turbulent noise and fluctuations. This may be partly because an adequate quantitative model for acoustic speech production has not been established which allows us to manipulate systematically these acoustic source parameters for cognitive experiments of paralinguistic features.

We have developed a novel joint estimation method of voice source and vocal tract (formant/antiformant) parameters from the speech signal based on the autoregressive with exogenous input (ARX) model [4]. This model incorporates a mathematical voice source model as well as fluctuations of the source waveform. The voice source parameters estimated from the utterance using the method could be systematically manipulated and used to re-synthesize speech of various voice qualities for perceptual experiments. The entire process from the acoustic analysis to re-synthesis has been implemented in an analysis-synthesis-editing system [4].

In this paper, we focus on the role of spectral tilt parameter (TL) of the source waveform, which is closely related to the duration of the return phase of the differentiated glottal waveform, the glottal turbulent noise and fluctuations of the source frequency spectrum on period-to-period basis in the cognition of paralinguistic features. A male adult uttered a Japanese word /eki/ ("a railroad station" in English) in three different paralinguistic renditions, neutral (declarative), suspicious, and disappointed. The ARX joint estimation method analyzed the speech utterances and yielded formant and voice source parameters.

In the disappointment rendition, amount of the glottal turbulent noise and the TL value were systematically changed to produce synthetic speech of various voice qualities, keeping the F0 contour and duration of the original utterance unchanged. In the suspicious rendition, on the other hand, the TL value and source frequency spectral fluctuations were orderly manipulated. The synthetic speech materials were all subjected to the perceptual experiments, in which ten male and female subjects judged the degree of disappointment or suspicion, for each of the randomly presented synthetic speech materials.

It has been found that the larger the TL, the higher the degree of disappointment and inclusion of the glottal turbulent noise emphasizes disappointment when the TL is small. In the case of suspicion the larger the TL, the lower the degree of suspicion and existence of the source spectral fluctuations increases that degree over the entire range of the TL values.

2. ARX SPEECH ANALYSIS AND SYNTHESIS METHOD

2.1 Voicing Source Model

Many descriptive models have been proposed to quantitatively represent glottal flow. The Liljencrants-Fant (LF) model [5] 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 [6], on the other hand, is mathematically simple but to a certain extent has limitations in representing a large variety of glottal flow waveforms. 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

\[ g(n) = 2an - 3bn^2, \quad 0 \leq n \leq T \]
\[ a = (27 \cdot AV) / (4 \cdot OQ^2 \cdot T) \]
\[ b = (27 \cdot AV) / (4 \cdot OQ^2 \cdot T^2) \]

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

\[ s(n) + \sum_{i=1}^{P} a_i \cdot s(n-i) = \sum_{j=1}^{Q} b_j \cdot u(n-j) + e(n), \]

an equation error as follows:

where \( s(n) \) and \( u(n) \) denote a speech signal and a differentiated glottal flow waveform at time \( n \), respectively. In the equation, \( a \) and \( b \) 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,

\[ S(z) = \frac{B(z)}{A(z)} \cdot U(z) + \frac{1}{A(z)} \cdot E(z) \]

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 taken into consideration, it results in the ARMAX (autoregressive 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.

### 2.3 Analysis and Synthesis

An improved ARX analysis-synthesis method has been successful in overcoming most of the deficiencies that the previous algorithm has suffered [4], making it possible to apply it to the speech utterance of various voice quality variations including high-pitch voice. The detailed description of the new method is described elsewhere [7].

### 3. EXPERIMENTS

#### 3.1 Speech Materials

A male adult pronounced 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. Speech signal were recorded on a digital tape recorder and then down-sampled at a sampling frequency of 11.025 kHz in a computer for the analysis and synthesis. In the ARX analysis-synthesis, orders of the ARX equation were set at \( p=12 \) and \( q=0 \).

#### 3.2 Experiment I

Experiment I was designed to decide the range of the acoustic voice source parameters in which resulting synthetic speech sounded natural. Utterances in the suspicion rendition represented pressed and rough vocal quality, implying possible contribution of the \( TL \) parameter and fluctuation of the glottal flow. In our ARX model, the flow fluctuation is approximated by rapid variation of the \( TL \) value from period to period of the vocal fold vibration. The \( TL \) contour of the utterance was thus disintegrated into two components, \( i.e. \) the mean and the deviation from the mean component. The mean component is called simply \( TL \), whereas the deviation component is called fluctuation, bellow.

The disappointment utterance, on the other hand, sounded breathy, suggesting inclusion of the glottal noise as well as strong spectral tilt of the glottal flow.

#### 3.2.1 Synthetic Speech Materials

Synthetic speech stimuli for the cognitive experiment of suspicion were composed of combination of the \( TL \) values varied in four steps from 1 to 15 dB and on/off of the original fluctuation, resulting in eight stimuli (=4\times2). Stimuli for the disappointment, on the other hand, were synthesized in combination of the \( TL \) values in four steps from 1 to 15 dB and the glottal noise amplitude (\( NA \)) at six different levels from 0 (no noise included) to amplification of the original noise by a
factor of 2.5. Twenty-four stimuli (=4x6) were prepared for the experiment of the disappointment.

3.2.2 Method

Five male adults participated in the experiment. They were asked to make judgments on naturalness of synthetic stimuli on three scales: 3 for acceptable, 2 for almost acceptable, and 1 for not acceptable. All the subjects judged each of the stimuli ten times.

3.2.3 Results and Discussion

Results are shown in Table 1 (a) for the suspicion and (b) for the disappointment. In the tables, the synthetic speech stimuli with too large TL values and/or with excessive glottal noise generates unacceptable quality of naturalness. By setting a threshold at 2.0, we have six stimuli for the suspicion and 12 synthetic samples for the disappointment for the next experiment.

Table 1(a): Naturalness scores for the synthetic speech of the suspicion.

<table>
<thead>
<tr>
<th>Fluctuation</th>
<th>TL [dB]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 5 10 15</td>
</tr>
<tr>
<td>Yes</td>
<td>2.7 2.8 2.3 1.8</td>
</tr>
<tr>
<td>No</td>
<td>2.9 2.7 2.1 1.6</td>
</tr>
</tbody>
</table>

Table 1(b): Naturalness scores for the synthetic speech of the disappointment.

<table>
<thead>
<tr>
<th>Relative Noise Level</th>
<th>TL [dB]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 5 10 15</td>
</tr>
<tr>
<td>x 2.5</td>
<td>1.4 1.3 1.3 1.3</td>
</tr>
<tr>
<td>x 2.0</td>
<td>1.8 1.7 1.6 1.5</td>
</tr>
<tr>
<td>x 1.5</td>
<td>2.2 2.1 2.1 1.9</td>
</tr>
<tr>
<td>x 1.0</td>
<td>2.8 2.5 2.5 2.3</td>
</tr>
<tr>
<td>x 0.5</td>
<td>2.6 2.8 2.6 2.4</td>
</tr>
<tr>
<td>None</td>
<td>2.1 2.0 1.6 1.4</td>
</tr>
</tbody>
</table>

3.3 Experiment II

Using synthetic speech materials, which were judged to be natural in Experiment I, a further perceptual experiment was performed. Experiment II was designed to understand quantitatively roles of the frequency spectral tilt parameter TL, amount of the glottal noise NA, and period-to-period fluctuation.

3.3.1 Synthetic Speech Materials

Six and 12 synthetic speech stimuli obtained in Experiment I were used in Experiment II as well, for the suspicious and disappointed renditions, respectively.

3.3.2 Method

Five male adults participated in the experiment. Figure 2 illustrates a schematic diagram for a paired comparison test, where the subjects were asked to judge which of the two stimuli presented in a soundproof room had a greater degree of a given rendition. In the renditions of suspicion and disappointment, the subjects listened to a total of 15 pairs (=6x5/2) and 66 (12x11/2) pairs, respectively. They judged five times for each of the pairs. Contribution of the acoustic parameters to the perceived degree of the rendition was evaluated by taking statistics given by (chosen times)/(all the presented times) [%] for each of the synthetic stimuli. This figure is called “relative degree of the rendition” below.

3.3.3 Results and Discussion

Figure 3 illustrates the relative degree of suspicion as a function of the frequency spectral tilt parameter TL in dB. Inclusion (YES) and exclusion (NO) of period-to-period fluctuations involved in the TL also affect the relative degree over the entire range of the TL.

Figure 3: Relative degree of the suspicious rendition as a function of the frequency spectral tilt parameter TL in dB. Inclusion (YES) and exclusion (NO) of period-to-period fluctuations in the TL also affect the relative degree over the entire range of the TL.

Figure 4: Relative degree of the disappointed rendition as a function of the frequency spectral tilt parameter TL in dB. Effects of the amount of the glottal noise on the relative degree are also shown in the cases of: no noise (NONE), a half of the original (x0.5), the original (x1.0), and 1.5 times as much as the original (x1.5).
TL, the lower the degree of suspicion and existence of the source spectral fluctuations increases that degree over the entire range of the TL values.

Figure 4 illustrates the relative degree of disappointment as a function of the TL [dB] value. Effects of the amount of the glottal noise on the relative degree are also shown in the figure. In this case, four levels of the glottal noise level are studied, i.e. no noise included, a half of the original noise level, the original and 1.5 times as much as the original. The figure indicates that the larger the TL, the higher the degree of disappointment and inclusion of the glottal turbulent noise emphasizes disappointment when the TL is small.

4. CONCLUSION

The role of such voice source dynamics as frequency spectral tilt (TL) of the voicing source waveform, the glottal turbulent noise and fluctuations of the source frequency spectrum on period-to-period basis in the cognition of paralinguistic features as in suspicious and disappointed renditions has been investigated. It has been found that the larger the TL, the lower the degree of suspicion and existence of the source spectral fluctuations increases that degree over the entire range of the TL. In the case of disappointment, the larger the TL, the higher the degree of disappointment and inclusion of the glottal turbulent noise emphasizes disappointment when the TL is small.

5. ACKNOWLEDGMENTS

The authors thank Mr. Takahiro Otsuka for his significant assistance in carrying out the experiments and Dr. Hiroki Mori for his valuable comments. The paper was supported in part by a Grant-in-Aid for Scientific Research from the Ministry of Education, Science, Sports and Culture, Japan (12680367).

6. REFERENCES