Journal of Voice
Vol. 16, No. 4, pp. 443–459
© 2002 The Voice Foundation

Perceived Pitch of Synthesized Voice with Alternate Cycles

Xuejing Sun and Yi Xu

Department of Communication Sciences and Disorders, Northwestern University, Evanston, Illinois

Summary: Both in normal speech voice and in some types of pathological voice, adjacent vocal cycles may alternate in amplitude or period, or both. When this occurs, the determination of voice fundamental frequency (defined as number of vocal cycles per second) becomes difficult. The present study attempts to address this issue by investigating how human listeners perceive the pitch of alternate cycles. As stimuli, vowels /a/ and /i/ were synthesized with fundamental frequencies at 140 Hz and 220 Hz, and the effect of alternate cycles was simulated with both amplitude- and frequency-modulation of the glottal volume velocity waveform. Subjects were asked to judge the pitch of the modulated vowels in reference to vowels without modulation. The results showed that (a) perceived pitch became lower as the amount of modulation increased, and the effect seems to be more dramatic than would be predicted by existing hypotheses, (b) perceived pitch differed across vowels, fundamental frequencies, and modulation types, that is, amplitude versus frequency modulation, and (c) the prediction of perceived pitch was best made in the frequency domain in terms of subharmonic-to-harmonic ratio. These findings provide useful information on how we should assess the pitch of alternate cycles. They may also be helpful in developing more robust pitch determination algorithms. **Key Words:** Pitch—Alternate cycles—Subharmonics—Modulation—Subharmonic-to-harmonic ratio.

INTRODUCTION

Because of the complex nature of speech production, the voicing part of human speech is not purely periodic. Rather, it often contains a variety of irregularities.[1] For example, jitter (small random variation in period) and shimmer (small random variation in amplitude) can often be seen in normal speech. Sometimes, the variations are more substantial and systematic, turning the signal into alternate amplitude cycles or alternate period cycles. As described by Klatt and Klatt,[2] “normal voicing suddenly changes to a vibration model where the first of a pair of periods is delayed and reduced in amplitude and the first pulse may disappear entirely” (p. 840). This type of voice can occur both in normal voice and in pathological voice.[2–4] When such voice patterns occur, the determination of voice fundamental frequency (F_0) becomes difficult because it is uncertain whether each individual cycle or every two alternate cycles should be considered as one pitch period.

This type of voicing patterns have been observed and studied by many researchers.[1–6] While better understanding of the production of voice with alternate pulse cycles has been achieved through these studies, determining F_0 for this type of voice still remains a

Accepted for publication April 16, 2002.

Portions of this work were presented at the 139th meeting of the Acoustical Society of America, Atlanta, GA, June 2000

Address correspondence and reprint requests to Xuejing Sun, Department of Communication Sciences and Disorders, Northwestern University, 2299 N. Campus Dr., Evanston, IL 60208, USA.

e-mail: sunxj@northwestern.edu

problem. To our knowledge, to date, no satisfactory solutions have been found. The determination of F_0 is important as F_0 carries important speech information, such as voice quality, intonation, and emotion, and so forth. For F_0 to function in speech, presumably, it should be perceivable as pitch. It is therefore important to understand how the pitch of alternate cycles is perceived by human listeners. This understanding will be valuable for describing voice quality, studying tone and intonation in speech, and developing effective pitch determination algorithms. Note that another significant perceptual property of alternate cycles is the roughness sensation.[2,7] However, we only focus on the perceived pitch in the present study and leave the investigation of roughness to future studies. Also note that the term "pitch" usually refers to a perceptual quality, whereas "fundamental frequency" (F_0) is a physical property. Therefore, the unit hertz for fundamental frequency may not be the best choice for describing pitch. However, it is known that the perceived pitch of a tone with appropriate intensity level can be measured linearly in hertz when the F_0 is below 1000 Hz. In our work, we are dealing with speech signals, the pitch of which is well below this limit. Hence, we describe perceived pitch on the hertz scale in this study, which, we believe, does not limit the applicability of current results in general.

According to Titze,[1,8] alternate cycles in speech waveform primarily reflect the vibratory patterns of the vocal folds. The glottal pulse signals generated by the vibration of the vocal folds can be classified into three types, as shown in Figure 1.[1,8] Type 1 is a nearly-periodic signal, which presumably occurs most commonly in normal speech. This pattern of signal remains stable in the long term, although there are small random variations from cycle to cycle both in period and in amplitude, known as jitter and shimmer. A type 2 signal is characterized by *conspicuously* alternating high and low amplitude pulses or

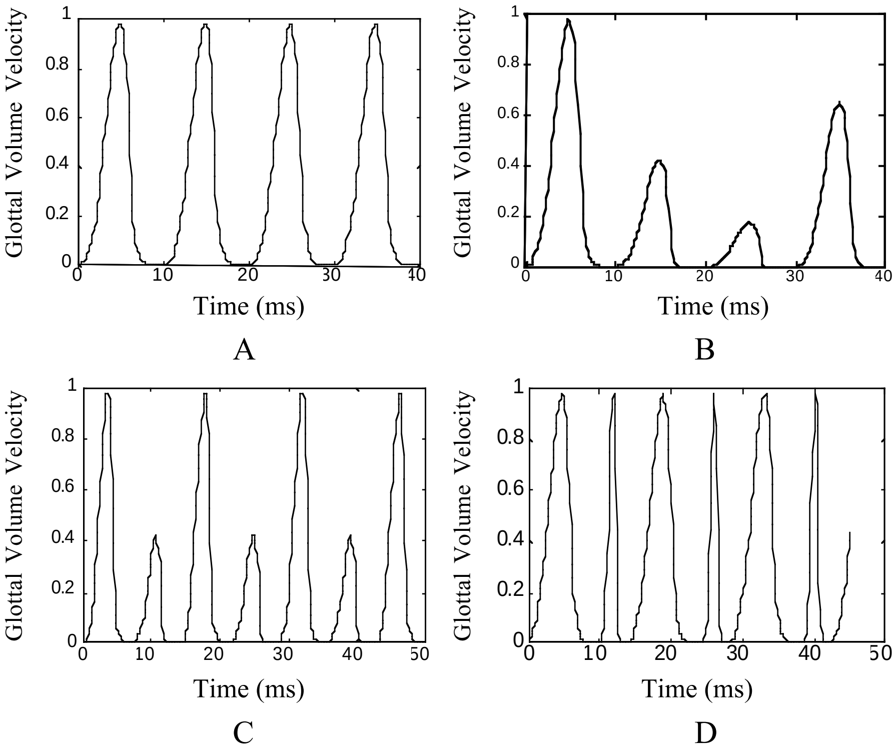

FIGURE 1. Schematic representations of three types of glottal signal in the time domain. The classification scheme follows Titze (1994, 1995): (A) type 1 signal—nearly-periodic vibration pattern; (B) type 3 signal—vibrations without any apparent periodic structures; (C) type 2 signal with amplitude alternation; (D) type 2 signal with period alternation.

alternating long and short periods. A type 3 signal does not have any apparent periodic structures. Assuming the basic shape of the voiced sounds in speech is determined by the glottal source signal, we regard alternate cycles in speech as the result of type 2 glottal source signal.

In Figures 1C and 1D, the signals can be viewed as the result of amplitude modulation (AM) and frequency modulation (FM), respectively. That is, the slowly varying component modulates the faster component. In this case, the ratio between the two components is one-half. The low frequency component is often called subharmonic. The subharmonic can be any integer fraction of the fundamental frequency (e.g., 1/2, 1/3, 1/4, …, 1/n). According to Titze,[8] subharmonic generation can occur when there is left-right asymmetry in the mechanical or geometric properties of the vocal folds. Svec et al[4] offered an alternative explanation that the subharmonic vibratory pattern of vocal folds could result from a combination of two vibrational modes whose frequency ratio is 3:2. The effect of subharmonics on speech can be either amplitude modulation or frequency modulation. As a result, adjacent cycles in speech can have alternate amplitudes and/or alternate periods.

The amount of amplitude modulation can be defined as a percentage in the form of the following:[1]

$$M = \frac{A_i - A_{i+1}}{A_i + A_{i+1}} 100 \qquad (1)$$

where A_i and A_{i+1} are the amplitudes of consecutive pulses (see Figure 1C), and M is the modulation index which can vary from 0 to 100%. Similarly, in frequency modulation, we may have

$$M = \frac{T_i - T_{i+1}}{T_i + T_{i+1}} 100 \qquad (2)$$

where T_i and T_{i+1} are the periods of consecutive pulses (see Figure 1D), and M is the amount of frequency modulation in percentage.

The modulation index indicates the amount of difference between two adjacent cycles in terms of amplitude or period. We shall call this modulation index “glottal modulation index” (GMI) since it describes the glottal volume velocity waveform. Modulated glottal volume velocity waveform presumably results in alternate cycles in radiated speech waveform. Figure 2 shows three waveforms of synthetic vowel /a/ with amplitude modulation with different glottal modulation indices.

The alternate cycles observed in the output speech waveform are the combined effects of vocal tract filtering and vibration pattern of the source signal, which can be measured by a modulation index on the radiated speech waveform directly. To ease our description, we term it as “signal modulation index” (SMI). The variation of signal modulation index could be caused by different glottal modulation indices or different formant structures. Since two cycles that are different in period receive different resonant contributions from the vocal tract, most likely their amplitudes in speech waveform would be different. This has been observed in the present study using a Klatt-style formant synthesizer.[2] Similar phenomena have been found in jitter and shimmer by Murphy,[9] who stated that “it is interesting to note that jitter cannot exist independently of shimmer for the radiated speech waveform” (p. 2870). The complicated interaction between the source and the vocal tract often makes it difficult to judge from the speech waveform whether we have amplitude or frequency modulation at the source, and what is the level of modulation, although we could infer this by employing the glottal inverse filtering technique.

The modulation index, either GMI or SMI, describes the behavior of alternate cycles in the time domain. In the frequency domain, the manifestation of amplitude modulation or frequency modulation is the presence of subharmonics, as can be seen in Figure 3. In Figure 3A, for a signal without modulation, the distance between the harmonics is 140 Hz. For a signal with modulation (Figures 3B and 3C), where the subharmonic frequency is 70 Hz, some spectral components with lower amplitude appear on the spectra, reducing the distance between adjacent spectral lines to 70 Hz. The amplitude of the subharmonics reflects the level of modulation: the greater the amplitude, the higher the level of modulation.

As mentioned earlier, voice with alternate cycles often occurs in normal speech and the determination of their fundamental frequency is critical for voice and speech research. With the current lack of solid acoustic and physiological theory to guide us in this regard, studying the perceived pitch of such voice can provide us important information necessary for analyzing its fundamental frequency. To our knowl-

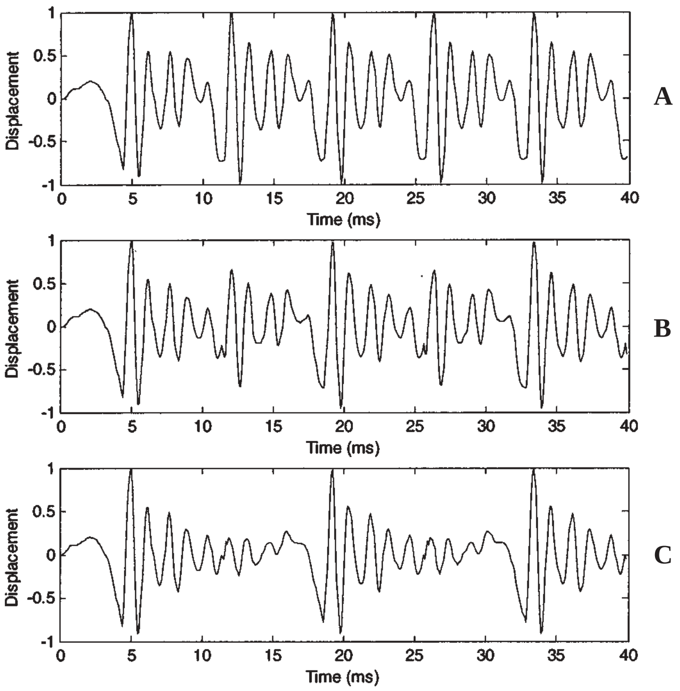

FIGURE 2. Synthetic vowel /a/ showing different amount of alternate cycles at three different modulation levels: (A) without modulation; (B) amplitude modulation with glottal modulation index = 50%, and (C) amplitude modulation with glottal modulation index = 90%.

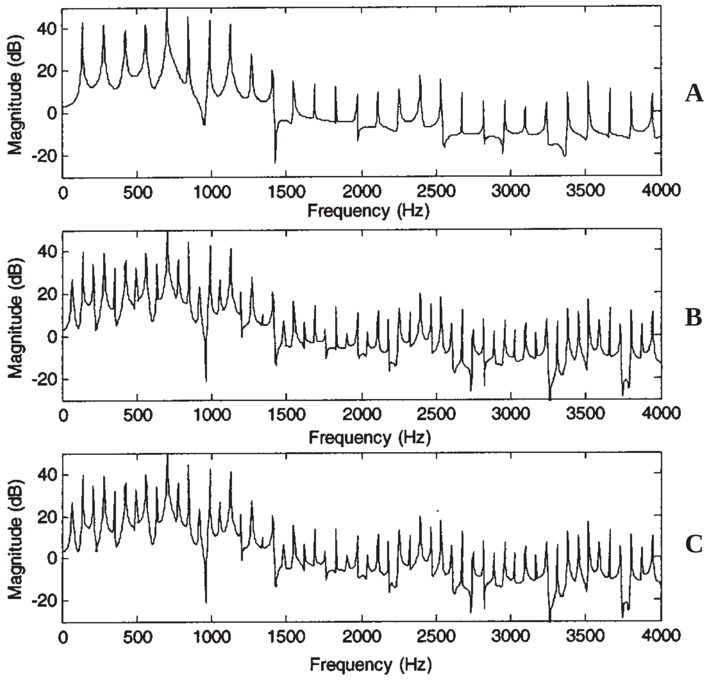

FIGURE 3. Spectra of synthetic vowel /a/ showing the effect of different amount of amplitude modulation on the magnitude of subharmonics: (A) without modulation; (B) amplitude modulation with glottal modulation index = 50%; and (C) amplitude modulation with glottal modulation index = 90%.

edge, there has been only one systematic investigation of perceived pitch of voice with alternate cycles, which was done very recently by Bergan and Titze.[7] Their study tried to find the relationship of subharmonics and the modulation index with perceived pitch and roughness of vocal signals. In the study, vowel /a/ was synthesized with various degrees of amplitude modulation and frequency modulation. The subharmonic being considered was at $F_0/2$. Three conditions of fundamental frequencies, that is, 100 Hz, 200 Hz, and 300 Hz, were used. It was found that the crossover point to the lower pitch (i.e., that associated with subharmonic) occurred between 10% and 30% of modulation, which varied according to the modulation type and F_0. The study also found that the crossover point for FM usually came earlier than that for AM.

From what we could tell, the modulation thresholds obtained by Bergan and Titze[7] were in terms of glottal modulation index rather than signal modulation index. It is known that during speech production the vocal tract exerts nonlinear filter effects on glottal source signals.[2] Hence, the radiated speech waveform may deviate from the glottal signal even though the primary pattern is retained. Since it is the radiated waveform that we hear to perceive pitch in speech, it is important to also understand the relationship between signal modulation index and perceived pitch. Titze[8] suggests that, for alternate amplitude cycles, at 10% modulation, pitch may not be different from that of unmodulated signals, whereas in the vicinity of 50% modulation, a significant pitch change should occur in pitch perception. This hypothesis appears to refer to signal modulation index rather than glottal modulation index.

Both glottal modulation index and signal modulation index are based on observations in the time domain. Recall that the manifestation of both amplitude modulation and frequency modulation in the frequency domain is the presence of subharmonics. This means that we may be able to use one parameter to describe both amplitude modulation and frequency modulation. Inspecting Figure 3 again, we can see clearly that along with the increase of glottal modulation index, the magnitude of the subharmonic components increases with respect to harmonic components. When the amplitude of the subharmonics is low, or more exactly, when the amplitude ratio between the subharmonics and harmonics is low, the subharmonics probably have no effect on pitch perception. When the amplitude ratio is sufficiently high, pitch may be perceived as one octave lower. The effect of amplitude ratio on pitch may also be explored from the perspective of masking, which has been studied extensively in psychoacoustics,[10] where the amplitude ratio between signal and masker signal (signal-to-masker ratio) determines whether the signal is detectable. In our case, if viewing harmonic components as the masker and subharmonic components as the signal, we may say that as the signal-to-masker ratio increases, the harmonics have fewer and fewer masking effects and the subharmonics become more and more audible. When the ratio becomes sufficiently large, listeners perceive a new signal whose pitch corresponds to the subharmonic frequency. Thus, to summarize, it seems possible that we could use the amplitude ratio between subharmonics and harmonics as our frequency domain parameter for alternate cycles, which we refer to as subharmonic-to-harmonic ratio (SHR). Murphy[11] has also postulated the possible use of amplitude variation of harmonics and subharmonics on predicting perceived pitch.

RESEARCH QUESTIONS

The goal of the present study is, therefore, to assess the perceived pitch of voice with alternate cycles. In particular, we would like to examine the relationship between perceived pitch, modulation index, and subharmonic-to-harmonic ratio. The design of the study was based on three basic assumptions: (a) the observed alternate cycles in speech waveform are primarily the result of type 2 signals; (b) type 2 signals can be modeled by either amplitude modulation or frequency modulation as defined in Equations (1) and (2); and (c) in the frequency domain, type 2 signals are manifested by the appearance of subharmonics, the frequency of which is at an integer ratio of the fundamental frequency.

Five specific questions are asked in the present study:

1. What is the relationship between glottal modulation index and perceived pitch?
2. What is the relationship between signal modulation index and perceived pitch, and between glottal modulation index and signal modulation index?
3. Given the amount of amplitude or frequency modulation, will perceived pitch differ across

different fundamental frequencies and different vowels?

4. Do amplitude modulation and frequency modulation have different effects on perceived pitch?
5. Finally, how does subharmonic-to-harmonic ratio change with glottal modulation index, and how well can it be used to predict perceived pitch of alternate cycles?

Questions 1 and 4 and part of question 3 above were already investigated by Bergan and Titze.[7] Thus, it will be interesting to see how closely we can replicate their findings in the present study. The rest of the questions have not been asked before. Their answers may help us further understand the perceived pitch of voice with alternate cycles.

METHODS

Subjects

Thirteen native speakers of American English (6 males and 7 females) between the ages of 18 and 36 participated in the experiment. All reported having normal hearing, vision, and language ability, and none reported any formal musical training. Prior to the experiment, subjects were asked to sign an informed consent form. Subjects were paid for their participation.

Stimuli/apparatus

Synthetic vowels were used as stimuli in the present study. An in-house formant synthesizer based on the framework of KLSYN88 synthesizer[2] and the LF (Liljencrants–Fant) voice source model were employed to generate signals.[12] Two synthetic vowels, /i/ and /a/, with alternate cycles were generated. The procedure was as follows: (a) using the LF voice source model to produce glottal pulse waveform; (b) modulating the glottal wave by varying the amplitude or period of every other glottal pulse based on Equations (1) and (2); (c) synthesizing different vowels by varying the formant frequencies; and (d) saving the output to individual files.

Similar to Bergan and Titze,[7] in the present study, we examined only subharmonic at ½ of F_0, which resulted from the simplest pattern of modulation. More complex cases, that is, ⅓, ¼, can be investigated by extending the current work. The specific steps of producing type 2 signals are as follows:

1. For amplitude modulation, we first assume the amplitude of the first cycle A_i as 1, then for a given modulation index, the amplitude of the second cycle A_{i+1} can be derived from Eq. (1) as:

$$A_{i+1} = \frac{100 - M}{100 + M} A_i \tag{3}$$

2. Similarly, for frequency modulation, from Eq. (2) we can have:

$$T_{i+1} = \frac{100 - M}{100 + M} T_i \tag{4}$$

To simulate a real-life case, we further put a constraint on T_{i+1} and T_i. Supposing the fundamental period is T_0, we require: $T_{i+1} + T_i = 2T_0$. Then we have:

$$T_{i+1} = \frac{100 - M}{100} T_0 \tag{5}$$

$$T_i = \frac{100 + M}{100} T_0 \tag{6}$$

The modulated synthetic vowels were generated at two fundamental frequencies, 140 and 220 Hz. For each fundamental frequency, 10 amplitude modulated and 10 frequency modulated vowels were synthesized by increasing the value of glottal modulation index from 0 to 90% at steps of 10, that is, 0, 10, . . . , 90. Each modulated signal was to be presented three times during the experiment. In total, there were 240 (2 fundamental frequencies × 2 vowels × 2 modulation types × 10 modulation levels × 3 repetitions) stimuli used in the experiment.

Subjects were asked to determine the pitch of each stimulus by matching it to a series of reference signals. Synthetic vowels without modulation were used as the reference signals. Two series of reference vowel signals (/i/ and /a/) were synthesized with fundamental frequencies ranging from 70 Hz to 140 Hz and from 110 Hz to 220 Hz, respectively. The underlying assumption of this range is, according to previous discussion, perceived pitch of alternate cycles should be in the range from the original F_0 to one-half. The resolution of the fundamental frequencies of the reference signals, that is, the smallest frequency differences between any two test signals, was 1

Hz. The sampling rate of the signals was 8 kHz and the duration of each signal was 400 ms.

In Bergan and Titze[7] triangular waves were used as the reference signals. They admit that ideally the same synthesizer that generates the modulated signals should be used, but it was not adopted because real-time control over the F_0 was not available. In the current study, we presynthesized the reference signals. The disadvantage is that the accuracy is fixed at 1 Hz. Nevertheless, we think that this resolution is sufficient as psychoacoustic studies have shown that the frequency difference limen (DL) of human ear is approximate in the range of (0.5, 1) Hz for frequency from 125 to 250 Hz.[13]

The sound level was in the range of (60, 65) dB SPL for all the signals. The duration of the whole experiment was about 90 minutes on average. The signals were presented to the subjects via a set of binaural headphones. Calibration of the equipment was performed to ensure accurate sound level. A program written in Java on a Macintosh computer controlled the entire experiment procedure.

Design

A repeated measure design was employed. The independent variables were vowel (/i/, /a/), fundamental frequency (140 Hz, 220 Hz), modulation type (frequency and amplitude), and glottal modulation index (from 0% to 90% with step size 10). The dependent variable was perceived pitch. For each subject, there were 80 (2 vowels × 2 fundamental frequencies × 2 modulation types × 10 modulation indices) experimental conditions, and each stimulus had three repetitions.

Procedure

All tests were conducted in a sound-treated booth in the Speech Perception Laboratory at Northwestern University. The subject was seated comfortably in the booth facing a computer monitor with headphones on. In each trial, the subject was asked to select a reference vowel that had pitch most similar to the modulated vowel. The pitch of the reference vowel could be changed by moving a scrolling bar on the screen. There was also a number indicating the current pitch of the reference vowel in hertz. Before the decision was made, the subject could listen to the stimulus and the reference vowel as many times as necessary.

Before the real trials, the subject was asked to go through several practice trials until he/she became familiar with the task. There were three sessions, and within each the 80 modulated synthetic vowels were presented in a random order. Subjects could take a break after each session at will. When the experiment was completed, all pitch values determined by the subject were written into a file for later analysis.

Analysis

For statistical analysis and calculation of glottal modulation index, signal modulation index, and subharmonic-to-harmonic ratio, the following procedures were taken. First, the three repetitions for each subject within each condition were averaged. In order to compare the results of different fundamental frequencies, normalization was applied to the data. For F_0 at 140 Hz, 140 Hz would be 1 while 70 Hz would be 0.5; for F_0 at 220 Hz, 220 Hz would be 1 while 110 Hz would be 0.5. Then a four-way repeated measure analysis of variance (ANOVA) was performed. The four factors were fundamental frequency, vowel, modulation type, and glottal modulation index. The alpha level was set to 0.05. For glottal modulation index, we were interested in its main effect, that is, whether perceived pitch was affected significantly by varying glottal modulation index. For modulation type, vowel, and fundamental frequency, we wanted to examine their interaction with glottal modulation index. A Sheffe's post hoc test was followed to determine between which two modulation indices there was a significant pitch change.

As mentioned in the Introduction, Titze's hypothesis[8] about perceived pitch of alternate cycles considers the equivalent of our signal modulation index rather than glottal modulation index. On the other hand, in Bergan and Titze,[7] the crossover point was measured in terms of glottal modulation index. It is, therefore, interesting to calculate the signal modulation index and examine its relation with perceived pitch for comparison. This calculation was done by locating the major peak or valley of two adjacent cycles using the corresponding stimulus without modulation as reference and computing the values following Equations (1) and (2). However, for quite a few stimuli we could not reliably locate the major peaks or valleys, especially for frequency modulated signals. As a result, only signal modulation index for amplitude modulated signals was reported. As only the threshold for amplitude modulation was mentioned in Titze,[8] we were still able to compare the relation between signal modulation index and perceived pitch obtained in the present study with Titze's hypothesis, as will be discussed later.

For subharmonic-to-harmonic ratio, we first calculated its value for all stimuli (see the description below), and compared subharmonic-to-harmonic ratio with pitch change at each modulation level. A regression analysis was also performed on pitch change and subharmonic-to-harmonic ratio to see how the two are related. Here we only briefly describe the major steps for calculating subharmonic-to-harmonic ratio. A more detailed description can be found in Sun.[14]

A speech signal is first split into 40 ms short frames, on which a fast Fourier transform (FFT) is applied. A logarithmic transformation is then taken on the linear frequency scale, and the results are interpolated by the cubic-spline method.[15] The log frequency scaled spectrum is shifted leftward at odd orders, that is, $\log_2(1)$, $\log_2(3)$, $\log_2(5)$, . . . These shifted versions are added together, which is equivalent to compressing the spectrum at odd orders. That is, harmonics at f, 3f, 5f, … are added together. Similarly, the spectra shifted at even orders $\log_2(2)$, $\log_2(4)$, . . . are also added together. The amount of shifting is determined by the ratio between the upper cutoff frequency and half the fundamental frequency. In the present study, the two fundamental frequencies are 140 and 220 Hz. The cutoff frequency is 4000 Hz. Then, the local maximum value is found within a half-octave range centered at 70 and 110 Hz, respectively, on the spectrum, which is the sum of the shifted spectral at even orders. After locating the position of the local maximum, we identify the value of this particular position on the spectrum, which is the sum of shifted spectra at odd orders. The assumption is that by shifting the spectrum at even orders, we obtain the sum of all harmonics below the cutoff frequency of 70 Hz (or 110 Hz), which ideally should be the maximum value. In practice, however, because of the resolution of FFT, numerical interpolation, and rounding, we usually can only get a local maximum value around 70 Hz (or 110 Hz). Similarly, we can get the sum of subharmonics at 70 Hz by locating a local maximum on the spectrum, which is the summation of shifted spectra at odd orders. Finally, by dividing the two summation values, we obtain the subharmonic-to-harmonic ratio.

RESULTS

The three factors (fundamental frequency, vowel, modulation type) result in a total of $2 \times 2 \times 2 = 8$ conditions. For each condition, the mean values of perceived pitch is plotted against glottal modulation index (see Figure 4). The title of each figure indicates the combination of the three factors. We can see that, in general, (1) perceived pitch becomes lower as the modulation level increases; and (2) pitch drops more quickly with frequency modulation.

Figure 4 shows that perceived pitch is affected significantly by varying the level of modulation, and that the relationship is nonlinear. In Figure 4, the shape of the function curve is nearly flat at the vicinity of 0 and 90% of modulation, with a sharp transition in the middle. It should be noted that in Figure 4, we can see that the average highest perceived pitch corresponding to 0% of modulation is always lower than rather than equal to 140 Hz (or 220 Hz). This is possibly because: (1) pitch matching is a difficult task; (2) as the highest value provided was 140 Hz (or 220 Hz), whenever the subject made an error, it would make the perceived pitch lower; (3) subjects tend to be conservative, that is, they do not want to choose the highest value all the time.

The ANOVA results are presented in Table 1, which show the effects of fundamental frequency, vowel, modulation type, and glottal modulation index. As pointed out earlier, we are only interested in some of the ANOVA results. Thus, in Table 1, we list the results of main effect of glottal modulation index, the interaction between glottal modulation index, and three other factors, namely, fundamental frequency (140 Hz and 220 Hz), modulation type (amplitude modulation and frequency modulation), and vowel (/i/ and /a/). The main effect of glottal modulation index is significant, namely, a significant interaction between modulation type and glottal modulation index. The interaction between fundamental frequency and glottal modulation index is also significant, which indicates that the effect of glottal modulation index on pitch perception is different at different frequencies. The interaction between vowel and glottal modulation index is significant, but to a lesser extent. This means that although the vocal tract has an effect

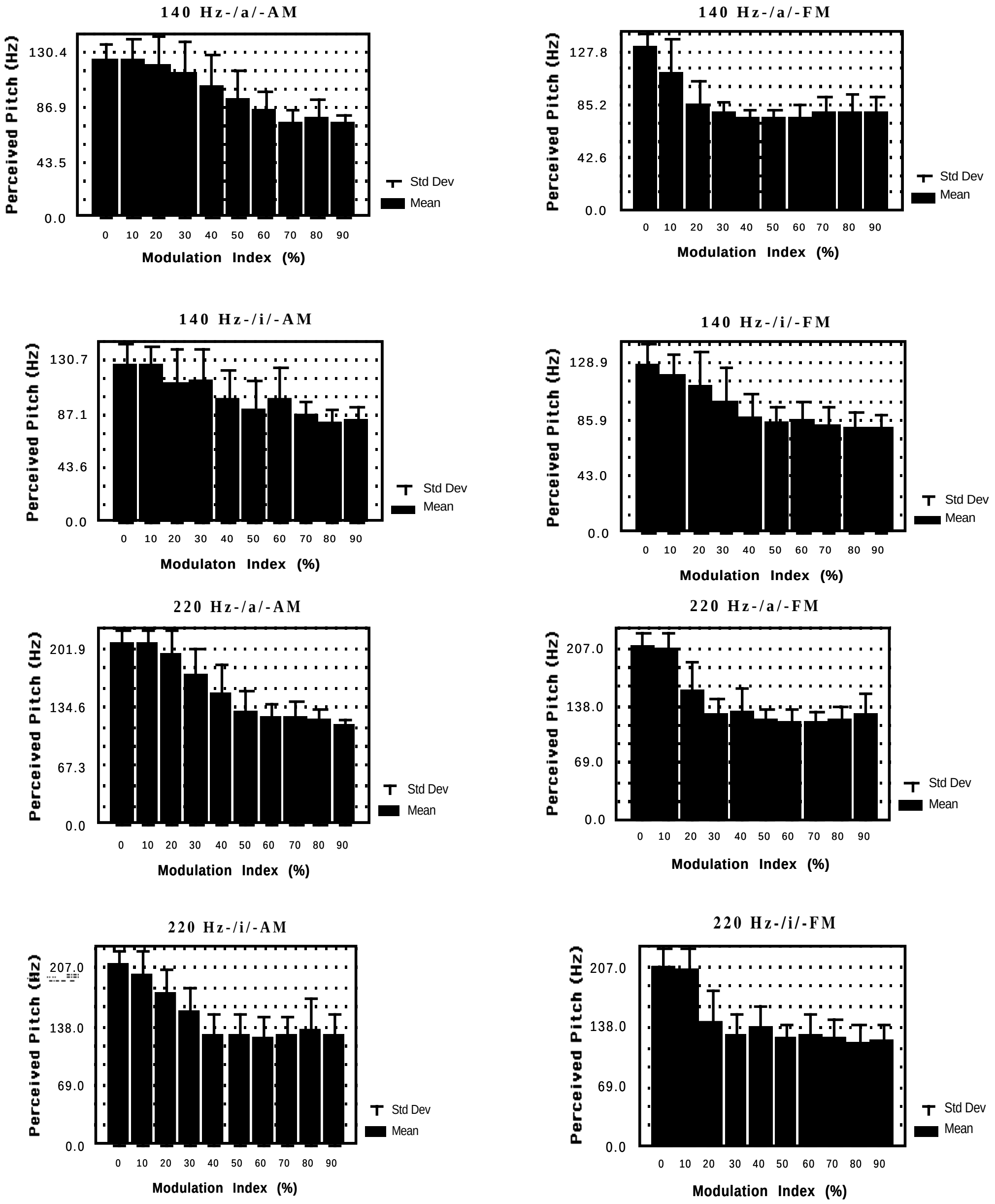

FIGURE 4. Variation of perceived pitch with glottal modulation index. The x-axis is glottal modulation index from 0 to 90%, whereas the y-axis is the frequency corresponding to perceived pitch from 0 to 140 Hz or 220 Hz. The eight graphs correspond to eight experimental conditions, which are combinations of fundamental frequency (140 Hz and 220 Hz), vowel (/a/ and /i/), and modulation type (amplitude modulation and frequency modulation).

on pitch perception, it is not as significant as other factors. In terms of the difference between two adjacent modulation indices, Sheffe's post hoc tests show that there is a significant difference between 20 and 30% of modulation, but not elsewhere. This indicates that around 20 to 30% of modulation, there is a substantial change in pitch perception. This is similar to the results obtained by Bergan and Titze,[7] where the crossover points usually occurred between 10% and 30% modulation.

Table 2 shows signal modulation indices for all eight conditions at different modulation levels. From

TABLE 1. *ANOVA Results for the Effects of Glottal Modulation Index, Modulation Type, Fundamental Frequency, and Vowel on Perceived Pitch*

	***F*-value**	***P*-value**
Glottal modulation index	87.725	< 0.0001
F_0 × glottal modulation index	5.230	< 0.0001
Modulation type × glottal modulation index	9.713	< 0.0001
Vowel × glottal modulation index	2.055	0.0399

TABLE 2. *Signal Modulation Indices (SMI) and Corresponding Pitch at Different Levels of Glottal Modulation (Amplitude Modulation)*

	140 Hz				**220 Hz**			
Glottal modulation index (%)	**/a/**		**/i/**		**/a/**		**/i/**	
	SMI	**Pitch**	**SMI**	**Pitch**	**SMI**	**Pitch**	**SMI**	**Pitch**
0	0.000	0.889	0.000	0.912	0.000	0.947	0.000	0.957
10	0.033	0.897	0.069	0.907	0.018	0.947	0.057	0.903
20	0.068	0.863	0.121	0.809	0.037	0.888	0.121	0.806
30	0.106	0.822	0.198	0.823	0.056	0.772	0.279	0.703
40	0.146	0.747	0.264	0.716	0.097	0.684	0.369	0.578
50	0.197	0.661	0.311	0.657	0.141	0.593	0.369	0.582
60	0.300	0.598	0.363	0.709	0.214	0.566	0.377	0.569
70	0.381	0.537	0.411	0.610	0.305	0.566	0.306	0.577
80	0.519	0.559	0.487	0.574	0.419	0.546	0.689	0.610
90	0.626	0.524	0.607	0.585	0.534	0.516	0.793	0.586

Table 2, we can see that signal modulation index is usually smaller than the corresponding glottal modulation index. Subjects perceived a significant pitch change with a much smaller modulation index than 50% as suggested by Titze.[8] Note that except for the 220 Hz-/i/-AM group, all other groups show very consistent patterns, that is, as glottal modulation index varied from 0 to 90%, signal modulation index monotonically increased from 0 to 0.5 or 0.6, and perceived pitch decreased monotonically. For 220 Hz-/i/-AM, we were unable to compute signal modulation index reliably.

Table 3 shows subharmonic-to-harmonic ratios for all eight conditions at different modulation levels. It can be seen clearly that subharmonic-to-harmonic ratio increases as glottal modulation index (also see Figure 5) increases. When glottal modulation index equals zero, subharmonic-to-harmonic ratio is the lowest across all conditions, while at 90% of modulation, subharmonic-to-harmonic ratio approaches 1. Moreover, frequency modulation generally has higher subharmonic-to-harmonic ratio than amplitude modulation, which may explain why frequency modulation has more dramatic effect on pitch perception. It should be note that with 0% of modulation, theoretically subharmonic-to-harmonic ratio should be zero as there are no subharmonics in our synthetic speech. However, because we process the signals digitally, roundoff errors or the like are inevitable. Thus, we usually can only obtain a small value rather than zero. Also, note that in some cases SHR can be greater than 1. Besides the aforementioned reason

TABLE 3. *Subharmonic-to-Harmonic Ratios (SHR) at Different Levels of Glottal Modulation (Amplitude Modulation [AM] and Frequency Modulation [FM])*

	140 Hz				220 Hz			
	/a/		/i/		/a/		/i/	
Glottal modulation index (%)	AM	FM	AM	FM	AM	FM	AM	FM
0	0.093	0.082	0.160	0.150	0.088	0.077	0.053	0.075
10	0.113	0.352	0.178	0.195	0.134	0.367	0.191	0.422
20	0.154	0.719	0.245	0.255	0.192	0.614	0.306	0.453
30	0.194	1.022	0.326	0.303	0.254	0.906	0.302	0.568
40	0.235	1.101	0.404	0.424	0.316	1.026	0.298	0.672
50	0.283	0.997	0.474	0.502	0.383	1.041	0.294	0.751
60	0.347	0.918	0.539	0.624	0.462	1.009	0.385	0.881
70	0.438	0.932	0.593	0.783	0.552	1.020	0.519	0.913
80	0.564	0.951	0.646	0.904	0.659	1.048	0.685	0.948
90	0.734	1.002	0.695	0.989	0.792	1.076	0.849	0.940

from calculation, it could also be caused by the nonlinear filtering effect of the vocal tract which makes some spectral components more prominent than others. When the modulation level is deep enough, the subharmonic is no longer "subharmonic," instead, it becomes the real harmonic. As a result, pitch becomes one octave lower and is no longer ambiguous. Thus, in this case, computing subharmonic-to-harmonic ratio is equivalent to computing the ratio between the sum of the harmonics at odd orders and the sum of the harmonics at even orders. This ratio can be a bit smaller or greater than 1 depending on the particular spectral structure.

In order to relate pitch changes to variations in subharmonic-to-harmonic ratio, we further performed the following procedures: (1) for normalized pitch values, subtracting them from 1 to obtain the amount of pitch change (see Figure 6); (2) performing regression analyses for subharmonic-to-harmonic ratio versus pitch change values, and glottal modulation index versus pitch changes values (see Tables 4 and 5). Figures 5 and 6 show that the general trends of pitch change and subharmonic-to-harmonic ratio are quite similar to each other. Table 4 and Figure 7 further show that across all conditions subharmonic-to-harmonic ratios are highly correlated with pitch changes with minimum $r^2 = 0.7329$. On the other hand, for glottal modulation index and pitch change, the r^2 values are much lower in general, with minimum $r^2 = 0.462$ (Table 5).

DISCUSSION

The relationship between perceived pitch and modulation index in all conditions shows similar trends. That is, when modulation index increases, perceived pitch becomes lower, eventually changing into approximately one-half the original value. This indicates that the presence of subharmonics in speech has a pitch-lowering effect, and perceived pitch is determined by the energy contained in the subharmonic.[8]

It is interesting to note that the relationship between glottal modulation index and perceived pitch is not linear (Figures 4 and 6). For example, for frequency modulated signals, the general pattern is that, when glottal modulation index is less than 20%, there are no significant pitch changes, whereas when glottal modulation index is greater than 50%, pitch becomes one-half the original value. For amplitude modulated signals, the threshold for perceiving one half the original pitch is greater than 50%. In general, frequency modulation seems to have greater effects on perceived pitch than amplitude modulation, which is consistent with Bergan and Titze.[7]

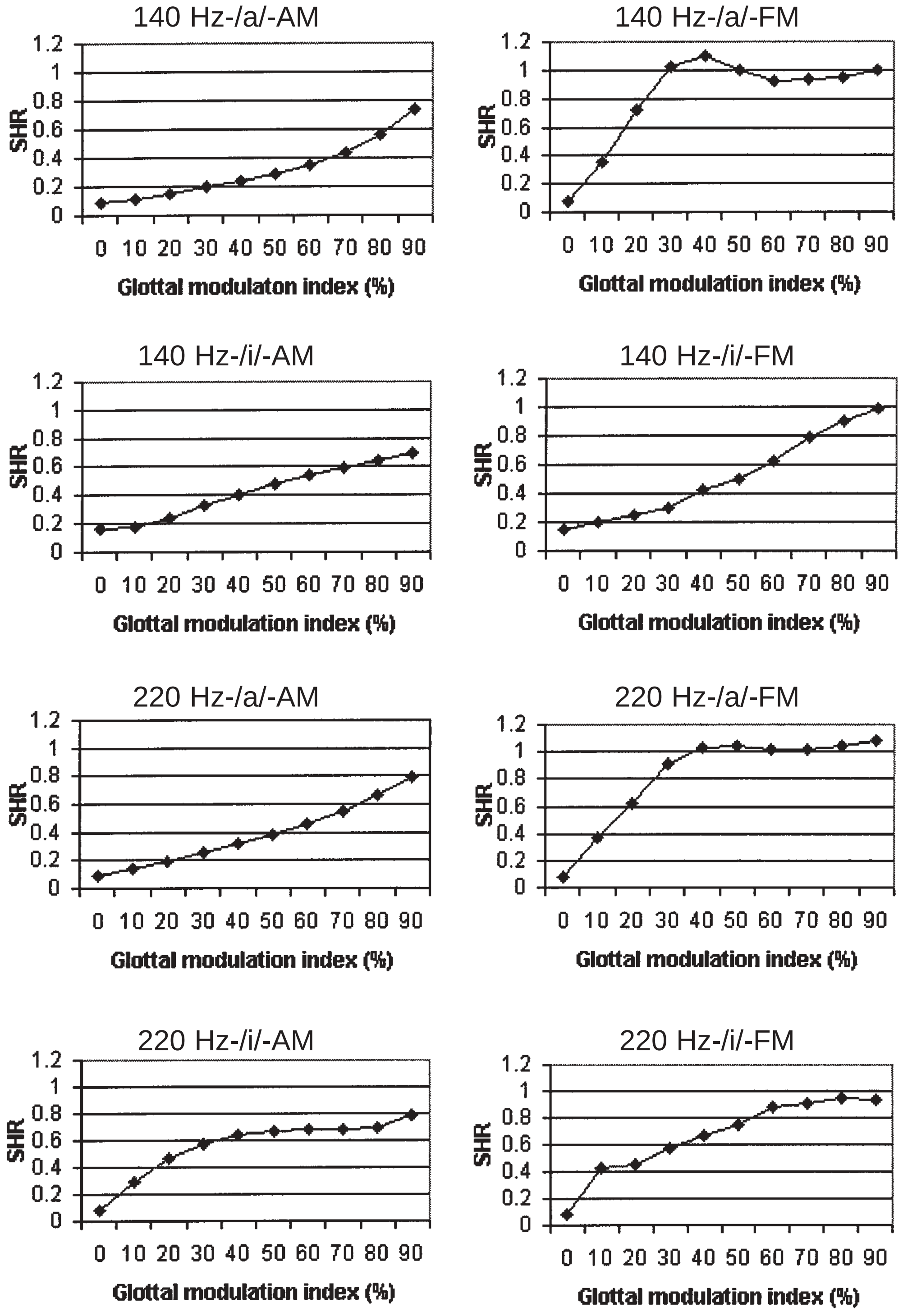

FIGURE 5. Subharmonic-to-harmonic ratio (SHR) versus glottal modulation index. The x-axis is glottal modulation index from 0 to 90%, and the y-axis is subharmonic-to-harmonic ratio. The eight graphs correspond to eight experimental conditions, which are combinations of fundamental frequency (140 Hz and 220 Hz), vowel (/a/ and /i/), and modulation type (amplitude modulation and frequency modulation).

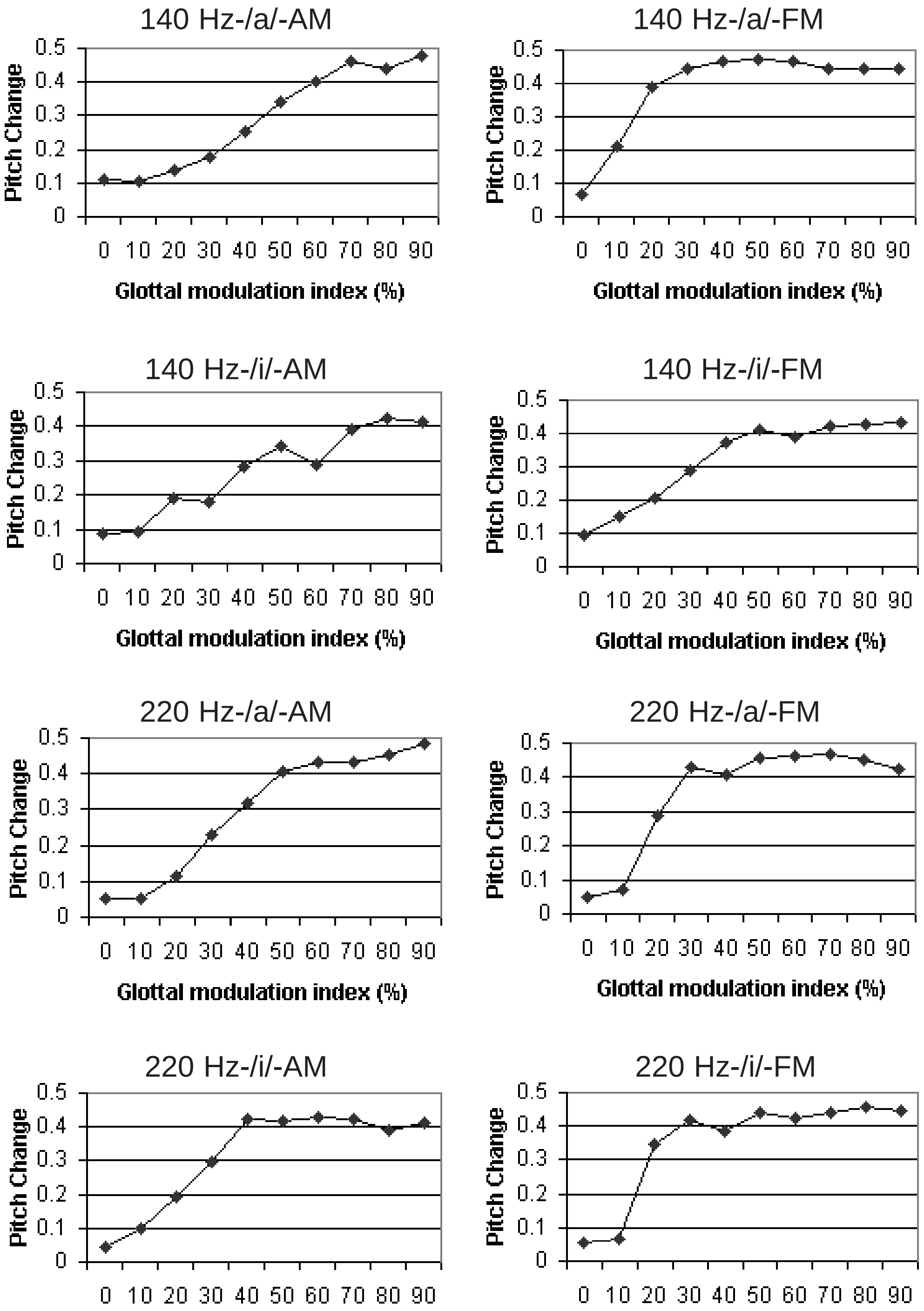

FIGURE 6. Pitch change with glottal modulation index. The x-axis is glottal modulation index from 0 to 90%, and the y-axis is the frequency corresponding to the amount of pitch change from 0 to 0.5. The amount of pitch change is obtained by subtracting the normalized pitch values from 1. The eight graphs correspond to eight experimental conditions, which are combinations of fundamental frequency (140 Hz and 220 Hz), vowel (/a/ and /i/), and modulation type (amplitude modulation and frequency modulation).

TABLE 4. *r^2 and Probability Values of Regression of Subharmonic-to-Harmonic Ratio over Perceived Pitch Change at Ten Glottal Modulation Levels for All Eight Experimental Conditions*

Experimental conditions (F_0 × vowel × modulation type)	r^2	Probability
140 Hz-/a/-AM	0.8189	0.0003
140 Hz-/a/-FM	0.9629	< 0.0001
140 Hz-/i/-AM	0.9379	< 0.0001
140 Hz-/i/-FM	0.758	0.001
220 Hz-/a/-AM	0.8359	0.0002
220 Hz-/a/-FM	0.9401	< 0.0001
220 Hz-/i/-AM	0.913	< 0.0001
220 Hz-/i/-FM	0.7329	0.0016

TABLE 5. *r^2 and Probability Values of Regression of Glottal Modulation Index over Perceived Pitch Change at Ten Glottal Modulation Levels for All Eight Experimental Conditions*

Experimental conditions (F_0 × vowel × modulation type)	r^2	Probability
140 Hz-/a/-AM	0.9413	< 0.0001
140 Hz-/a/-FM	0.462	0.0183
140 Hz-/i/-AM	0.9203	< 0.0001
140 Hz-/i/-FM	0.8513	< 0.0001
220 Hz-/a/-AM	0.9074	< 0.0001
220 Hz-/a/-FM	0.5979	0.0053
220 Hz-/i/-AM	0.6981	0.0016
220 Hz-/i/-FM	0.606	0.0049

From Table 2, similar patterns for signal modulation index can be observed, although they are not as consistent as those of glottal modulation index. This shows that the nonlinear relationships between signal modulation index and perceived pitch exist as implied in Titze,[8] except for the threshold values for pitch change. Instead of 50% of amplitude modulation as suggested by Titze,[8] starting from 20%, subjects begin to perceive a significant pitch change. At 50% of signal modulation index for amplitude modulation, the stimuli are perceived as one octave lower than the reference signals. Also, from Figure 4, we see that the critical value at which there is a significant pitch change varies across conditions. In other words, there may not be a fixed percentage because other factors, such as fundamental frequency and vowel, could have influences on perceived pitch as well.

This suggests that time domain parameters, that is, glottal modulation index and signal modulation index, may not be ideal indicators of pitch change. This is because (1) they are only surface measures of alternate cycles and do not give us an in-depth explanation of the perceived pitch; (2) they behave quite differently across modulation type, vowel, and fundamental frequency; and (3) in practice, obtaining modulation index either manually or automatically from real voice or speech is a difficult task.

In contrast to modulation index, subharmonic-to-harmonic ratio seems to provide us a more direct indication of perceived pitch. First, by comparing Figures 5 and 6, we see that with the increase of glottal modulation index, both the amount of pitch change and subharmonic-to-harmonic ratio increase in a similar manner. Table 3 reveals that subharmonic-to-harmonic ratio usually increases faster for frequency modulated signals than for amplitude modulated signals, which could explain why frequency modulation has a more dramatic effect on perceived pitch than amplitude modulation. In Table 4, the r^2 values are fairly high at all conditions between pitch change and subharmonic-to-harmonic ratio, with only two below 0.8. As can be observed in Figure 7, the relationship between subharmonic-to-harmonic ratio and pitch change seems more linear, and the general trend is quite similar in all eight graphs. The above results are encouraging because (1) subharmonic-to-harmonic ratio provides us a unified yet direct way to describe both alternate amplitude cycles and alternate period cycles; (2) the relationship between subharmonic-to-harmonic ratio seems to be robust under various conditions, which means subharmnic-to-harmonic ratio could potentially predict perceived pitch well; and (3) subharmonic-to-harmonic ratio can be obtained automatically.[14]

Subharmonic-to-harmonic ratio and its calculation method with some modifications have been applied to pitch determination tasks.[14] In this algorithm, subharmonic-to-harmonic ratio is computed and evaluated to determine whether the subharmonic is strong

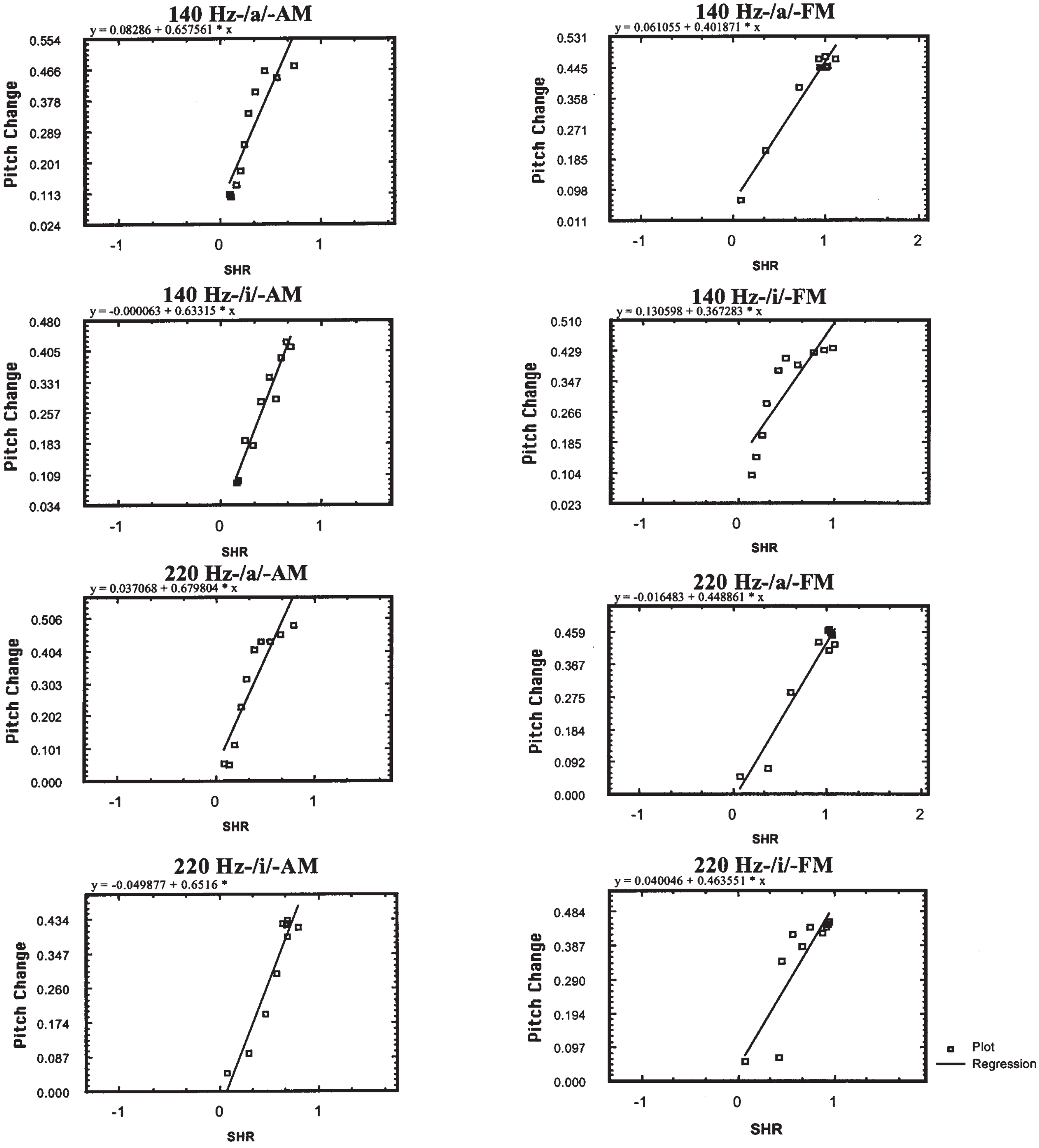

FIGURE 7. Regression analysis on pitch change and subharmonic-to-harmonic ratio. The x-axis is the subharmonic-to-harmonic ratio and y-axis is the frequency corresponding to the amount of pitch change. The eight graphs correspond to eight experimental conditions, which are combinations of fundamental frequency (140 Hz and 220 Hz), vowel (/a/ and /i/), and modulation type (amplitude modulation and frequency modulation).

enough to be an F_0 candidate. The evaluation results have shown that it substantially outperforms other state-of-the-art pitch determination algorithms being compared.

Although not intended in this study, it would be interesting to relate subharmonic-to-harmonic ratio to pitch perception theories (e.g., Terhardt's virtual pitch concept[16]) and roughness phenomenon.[7,16] In Terhardt's pitch perception theory, each harmonic component produces a series of subharmonics which are potential pitch candidates, and the overall perceived pitch corresponds to the frequency that has the largest number of coincidences by counting all the candidates. In our case, when subharmonic-to-harmonic ratio is small, the subharmonic components have low probability to be resolved by the auditory system, thus contributing less to the counting process. On the other hand, a larger subharmonic-to-harmonic ratio implies that those subharmonic components are more likely to be resolved making the overall pitch one octave lower. Subharmonic-to-harmonic ratio could also potentially be used as a parameter to quantitatively describe voice qualities, such as roughness. For example, a rough voice may be characterized by a medium ratio value, whereas a ratio value close to either 0 or 1 may indicate a more "regular" voice.

Despite the advantages of SHR discussed above, some caveats need to be mentioned. First, when glottal modulation index ranges from 20 to 40%, the corresponding subharmonic-to-harmonic ratio is roughly in the range of 0.2 to 0.4. In this region, relatively large individual differences are observed, as indicated by the large standard deviations in Figure 4. Bergan and Titze[7] have also found extensive inter- and intrasubject variability. This uncertainty is probably because subjects can listen either holistically or analytically when presented with complex tones.[17] In our case, when subharmonic-to-harmonic ratio is within the medium range, subharmonic components are competing with harmonics, which could elicit different perceptions. Thus, the average pitch value in the figure may not represent the real perceived pitch in a strict sense. We would rather regard it as a region of less certainty. In order to not let the large individual differences in the ambiguous region smear the overall trend, we ran regression analyses on the mean values rather than on the raw data. In this way, the overall subharmonic-to-harmonic ratio can predict perceived pitch quite well.

Second, in our calculation of subharmonic-to-harmonic ratio, we treat all harmonics and subharmonics equally in the range up to half of the sampling rate, that is, 4 kHz in the present study. It has been shown that there are dominant harmonic regions for pitch perception.[10] For example, Hermes[15] uses frequencies lower than 1250 Hz in his pitch determination algorithm. In our experiment, we felt that harmonics higher than 1250 Hz might still contribute to pitch perception, although their contribution might be much less than that of the lower harmonics. We tried to compute subharmonic-to-harmonic ratio by multiplying the frequencies higher than 1500 Hz by an exponential decay coefficient to reduce the contribution of higher harmonics. However, the selection of the coefficient becomes a problem, for there is no theoretical foundation available. Besides, the design of the current study is not really appropriate for this purpose. Thus, we only report the results using our original method, which seems to be sufficient to illustrate the relationship between subharmonic-to-harmonic ratio and perceived pitch. Even with these caveats in mind, nonetheless, subharmonic-to-harmonic ratio still seems to be a better predictor of perceived pitch than glottal modulation index or signal modulation index.

Finally, 400-ms signals used in the present study may be an overly optimistic choice. In reality, alternate cycles in normal speech may not last that long. Thus, further studies are needed to examine the duration effect, if any. Note that in Bergan and Titze,[7] duration of the stimuli was not provided. Therefore, we could not compare our data with theirs in this respect.

CONCLUSIONS

In the present study, we modulated the glottal volume velocity signal in amplitude and in frequency, respectively, and used the modulated glottal signal to synthesize vowels /a/ and /i/ at 140 Hz and 220 Hz. We asked subjects to judge the pitch of these synthesized vowels. We found that as the modulation index increased, perceived pitch became lower, ranging from the original pitch to that one octave lower. We further found that with the same amount of glottal modulation index, the variation of the perceived pitch differed across vowels, fundamental frequencies, and modulation types. Specifically, there was a significant pitch change when glottal modulation index was increased from 20 to 30%. With the same glottal modulation index, frequency modulation had a greater pitch lowering effect than amplitude modulation. As glottal modulation index increased, signal modulation index also increased but with lower magnitude. Particularly for amplitude modulated signals, starting from 10% of signal modulation index, a significant pitch change was usually perceived. With signal modulation index at 50% or higher, most likely pitch was perceived as one octave lower. We also found that subharmonic-to-harmonic ratio, as a frequency domain parameter, seemed to be a better indicator of perceived pitch than modulation index. It correlated highly with pitch changes in all eight experimental conditions, and provided a unified way to describe both amplitude and frequency modulation in alternate cycles. The current results have important implications for the development of more effective pitch determination algorithms.

Acknowledgments The authors wish to thank Kimberly Fisher, Charles Larson, and two anonymous reviewers for helpful comments and suggestions on the manuscript. This work was partially supported by NIH grant DC03902.

REFERENCES

1. Titze IR. *Workshop on acoustic voice analysis—summary statement*. Denver, Colo: National Center for Voice and Speech; 1995.
2. Klatt DM, Klatt LC. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J Acoust Soc Am*. 1990;87(2):820–857.
3. Blomgren M, Chen Y, Ng ML, Gilbert HR. Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers. *J Acoust Soc Am*. 1998;103(5 Pt 1): 2649–2658.
4. Svec JG, Schutte HK, Miller DG. A subharmonic vibratory pattern in normal vocal folds. *J Speech Hear Res*. 1996; 39(1):135–143.
5. Herzel H, Berry D, Titze IR, Saleh M. Analysis of vocal disorders with methods from nonlinear dynamics. *J Speech Hear Res.* 1994;37:1008–1019.
6. Titze IR, Baken R, Herzel H. Evidence of chaos in vocal fold vibration. In: Titze IR, ed. *Vocal Fold Physiology: New Frontiers in Basic Science.* San Diego, Calif: Singular Publishing Group; 1993;143–188.
7. Bergan CC, Titze IR. Perception of pitch and roughness in vocal signals with subharmonics. *J Voice.* 2001;15(2):165–175.
8. Titze IR. *Principles of Voice Production.* Englewood Cliffs, NJ: Prentice–Hall, Inc.; 1994.
9. Murphy PJ. Perturbation-free measurement of the harmonics-to-noise ratio in voice signals using pitch synchronous harmonic analysis. *J Acoust Soc Am.* 1999;105:2866–2881.
10. Moore BCJ. *An Introduction to the Psychology of Hearing* 3rd ed. San Diego, Calif: Academic Press; 1989.
11. Murphy PJ. Spectral characterization of jitter, shimmer, and additive noise in synthetically generated voice signals. *J Acoust Soc Am.* 2000;107(2):978–988.
12. Fant G, Liljencrants J, Lin QG. A four-parameter model of glottal flow. *Speech Transmission Lab Quarterly Progress Status Report.* Vol 4. Stockholm: Royal Institute of Technology; 1985;1–13.
13. Wier CC, Jesteadt W, Green DM. Frequency discrimination as a function of frequency and sensation level. *J Acoust Soc Am.* 1977;61(1):178–184.
14. Sun, X. *Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio.* International Conference on Acoustics, Speech, and Signal Processing, Orlando, Fla; May 13–17, 2002.
15. Hermes DJ. Measurement of pitch by subharmonic summation. *J Acoust Soc Am.* 1988;83(1):257–264.
16. Terhardt E. Pitch, consonance, and harmony. *J Acoust Soc Am*. 1974;55:1061–1069.
17. Smoorenburg GF. Pitch perception of two-frequency stimuli. *J Acoust Soc Am*. 1970;48:924–942.