Perception of Vocal Tremor

Vocal tremors characterize many pathological voices, but acoustic-perceptual aspects of tremor are poorly understood. To investigate this relationship, 2 tremor models were implemented in a custom voice synthesizer. The first modulated fundamental frequency (F0) with a sine wave. The second provided irregular modulation. Control parameters in both models were the frequency and amplitude of the F0 modulating waveform. Thirty-two 1-s samples of /a/, produced by speakers with vocal pathology, were modeled in the synthesizer. Synthetic copies of each vowel were created by using tremor parameters derived from different features of F0 versus time plots of the natural stimuli or by using parameters chosen to match the original stimuli perceptually. Listeners compared synthetic and original stimuli in 3 experiments. Sine wave and irregular tremor models both provided excellent matches to subsets of the voices. The perceptual importance of the shape of the modulating waveform depended on the severity of the tremor, with the choice of tremor model increasing in importance as the tremor increased in severity. The average frequency deviation from the mean F0 proved a good predictor of the perceived amplitude of a tremor. Differences in tremor rates were easiest to hear when the tremor was sinusoidal and of small amplitude. Differences in tremor rate were difficult to judge for tremors of large amplitude or in the context of irregularities in the pattern of frequency modulation. These results suggest that difference limens are larger for modulation rates and amplitudes when the tremor pattern is complex. Further, tremor rate, regularity, and amplitude interact, so that the perceptual importance of any one dimension depends on values of the others.

KEY WORDS: vocal tremor, vocal quality, speech synthesis, analysis by synthesis, vocal pathology

Steady state phonation is never perfectly steady. Phonation is characterized by slow modulations of fundamental frequency, usually in the range of 2–12 Hz, that reflect normal instabilities in human neurological control (e.g., Aronson, Ramig, Winholtz, & Silber, 1992; Titze, 1994). Although such modulations in normal voices are usually not perceptually prominent, these slight tremors contribute to the natural quality of the voice. More extreme and perceptually salient patterns of variability may also occur and are important aspects of overall vocal quality. Perceptually salient frequency modulations that are enhanced and exploited for artistic purposes in singing are usually called vibrato, and prominent, involuntary modulations are termed tremor (Titze, 1995a).

Both tremor and vibrato are commonly (and somewhat vaguely) defined as quasi-periodic, quasi-sinusoidal modulations of the fundamental frequency of phonation (e.g., Hibi & Hirano, 1995; Hori, 1989b; Morsomme, Orban, Remacle, & Jamart, 1997). Sundberg (1995) proposed four parameters to describe frequency modulations in vibrato: the rate of fundamental frequency (F0) modulation, its amplitude or extent...
about the mean F0 (usually 0.5–2 semitones; see Horii, 1989a, for review), the shape of the modulating waveform (generally more or less sinusoidal; Horii, 1989b), and the consistency of the frequency, amplitude, and shape of the modulating waveform. Tremor has been studied much less than vibrato, and description of vocal tremors in terms of these (or other) parameters is lacking. Existing studies of tremor have focused almost exclusively on acoustic measures of tremor rates and co-occurring short-term F0 variations like jitter and shimmer (e.g., Ackermann & Ziegler, 1994; Brown & Simonson, 1963; Ramig & Shipp, 1987). The pattern and regularity of long-term F0 modulations have not been studied in tremulous pathologic voices, but informal observations have revealed large departures from sinusoidality and regularity (e.g., Ackermann & Ziegler, 1994; Aronson et al., 1992). The perceptual importance of these different aspects of frequency modulation in pathologic voice has also never been investigated. Thus, the literature does not provide a clear description of vocal tremor or a way of predicting whether a voice will sound tremulous or tremor-free. In fact, authors have seemed uncertain about whether tremor is a single phenomenon or several different phenomena reflecting different underlying pathologies. Terms like flutter and wow have been proposed to designate different modulation rates (Aronson et al., 1992; Titze, 1994), although pathologic voices do not always have a single clear rate of frequency modulation (Aronson et al., 1992; Winholtz & Ramig, 1992).

Further, studies of the acoustic characteristics of frequency modulation provide only a partial insight into vocal tremor, because they do not describe how the acoustic features of tremors determine their perceptual salience. Judgments about voice disorders by patients and clinicians are heavily influenced by their perception of vocal deviation, so knowledge about vocal tremor perception is important. Unless the perceptual relevance of acoustic characteristics is known, there is no way to determine what acoustic features are important to listeners or how to measure them to best represent the perceived tremor. Speech synthesis allows experimental investigation of hypotheses about acoustic-perceptual relations and provides a method to manipulate candidate acoustic variables, whose perceptual significance is then assessed by presentation to listeners. In this way, synthesis offers a technique for identifying and quantifying the perceptually important acoustic characteristics of vocal tremors, and, thus, for finding evidence to help resolve many issues surrounding the description of frequency modulation in pathological voices.

This study describes such an investigation. We measured F0 variation over time for a variety of pathological voices, synthetically copied the voices using different models of this variation, and then used listener judgments to determine how well these acoustic models captured the perceptually salient characteristics of the frequency modulations. In Experiment 1 we examined the perceptual importance of the shape and regularity of the modulating waveform; in Experiment 2 we examined the perceptual significance of the amplitude of frequency modulation; and in Experiment 3 we examined the importance of the rate of frequency modulation.

**Analysis and Synthesis Techniques**

**Overview**

Because it is unclear what acoustic parameters should be used to describe frequency modulation in pathological voices, Sundberg’s (1995) four-parameter characterization of vibrato was adopted as a framework for analysis and synthesis, with one modification. Acoustic analyses and pilot perceptual studies indicated that frequency modulation in many pathological voices is both nonsinusoidal in shape and irregular in rate. Therefore, we assumed that irregularities in the pattern of frequency modulation do not occur without simultaneous irregularities in the frequency of modulation. Given this assumption, we combined the shape and regularity of the modulating waveform into a single binary “wave shape” parameter, which allowed the experimenter to select a sine wave or irregular tremor model. Figure 1 illustrates these models. Figure 1A shows the F0 track for a synthetic voice synthesized with a sine wave tremor, which sinusoidally modulates F0 above and below its specified mean value, and a typical output of the irregular tremor model is shown in Figure 1B. In this model, the pattern of irregular F0 modulation is established by passing white noise through an 8-pole Butterworth low-pass filter with cutoff frequency equal to the maximum modulation rate (which is described next). This produces an irregular pattern of frequency modulation. Note that the pattern of modulation remains independent of the rate of modulation in this framework. For example, tremors can be created for which F0 changes slowly but irregularly, quickly but irregularly, and so on.

The rate and amplitude of frequency modulation (i.e., how fast F0 varied and how far it varied above and below the mean F0) can also be manipulated independently in both the sine wave and irregular tremor models. In the sine wave model, the rate of frequency modulation represents the frequency of the modulating sine wave. In the irregular tremor model, the frequency modulation rate represents the maximum rate of change in F0. When this irregular model is applied, the synthesizer output in tremor cycles/second is approximately half the nominal value of this parameter. Therefore, parameter values were doubled when the irregular model was applied.

Amplitude modulations may occur along with frequency modulations in speech waveforms, resulting in
cycles that increase and decrease in amplitude in a regular pattern. However, at least in normal voices, these are largely a secondary effect of interactions of the changing harmonics of the voice with the (relatively) fixed vocal tract resonances (Horii, 1989a; Horii & Hata, 1988). Additional amplitude modulation may occur as a result of simultaneous modulation of glottal resistance or vibration of portions of the supraglottic vocal tract, but these factors account for relatively little variance in amplitude (Hibi & Hirano, 1995). Thus, in theory, patterns of frequency variation are of primary concern in the description of both tremor and vibrato, and the magnitude of amplitude modulation and phase relationships between “frequency tremor” and “amplitude tremor” appear to be largely artifactual (see Sundberg, 1995, for review). For this reason, neither tremor model included parameters to vary amplitude, although amplitude modulations did emerge from these models, presumably as a result of movement of harmonics toward and away from resonance peaks as F0 varied. These models also did not specify tremor phase, so that the initial and final points of the modeled tremor did not necessarily match those of the original voice samples.

**Algorithms**

Frequency for each cycle of phonation was calculated in the sine wave tremor model as

\[ F_0(t) = F_0_{\text{nom}} + D_{\text{Hz}} \cdot \sin(2\pi \cdot T_{\text{Hz}} \cdot t), \]

where \( t \) is time, \( F_0_{\text{nom}} \) is the mean F0 specified in the synthesizer, \( D_{\text{Hz}} \) is the peak amplitude of the modulating sinusoid (the amplitude of the tremor, in Hz), and \( T_{\text{Hz}} \) is the repetition rate (the modulation frequency, also in Hz) of the tremor.

Frequency modulation in the irregular tremor model followed the following equation:

\[ F_0(t) = F_0_{\text{nom}} + \frac{D_{\text{Hz}}}{D_{\text{max}}} \left[ r(t) \ast H(T_{\text{Hz}}, t) - \frac{1}{2} \right], \]

where \( * \) denotes time domain convolution, \( H \) is the impulse response of an 8-pole Butterworth low pass filter with cutoff frequency \( T_{\text{Hz}} \), \( r(t) \) is white noise uniformly distributed on \([0,1]\), and \( D_{\text{max}} \) is the maximum excursion of \( r \ast H \) from 0.5.

**Voice Samples**

The voices of 32 speakers (15 male and 17 female) with vocal pathology were selected at random from a library of samples recorded under identical conditions. Speakers represented a variety of primary diagnoses, including essential vocal tremor, vocal fold mass lesions, vocal fold paralysis, adductory spasmodic dysphonia, reflux laryngitis, glottal incompetence, and laryngeal web. They ranged from mildly to severely dysphonic. The frequency, amplitude, regularity, or perceptual prominence of any tremor were not criteria for voice selection, because no perceptual evidence exists to support distinctions between flutter, wow, and
tremor. Further, by including voices that varied widely in tremor prominence, we hoped to learn which acoustic characteristics affect the perceptual salience of a tremor. Five experienced listeners (including the first and third authors) assessed tremor severity for each voice on a 3-point scale. All ratings for each voice agreed exactly or within one scale value, which was considered adequate for the rather coarse level of measurement required here. Accordingly, these values were averaged and used to place voices into one of three categories according to tremor severity. Seven voices had mild tremors (mean rating for each voice < 1.5; SD across raters and voices = 0.28), 20 voices had moderate-to-prominent tremors (1.5 < mean rating < 2.2; SD = 1.62), and 5 voices had severe tremors (mean rating > 2.2; SD = 0.65).

Speakers were recorded as part of a clinical phonatory function analysis. They were asked to sustain the vowel /a/ for as long as possible, at comfortable levels of pitch and loudness. Voice signals were transduced with a 1” Bruel and Kjaer condenser microphone held a constant 5 cm off axis. Voice samples were low-pass filtered at 8 kHz and digitized at 20 kHz. A 1-s segment was excerpted from the middle of these productions, antialias-filtered, and downsampled to 10 kHz for further analysis.

Analyses of Fundamental Frequency

Frequency analyses were undertaken to provide estimates of the parameters needed to synthesize tremors. For each original voice sample, a negative peak, positive peak, or zero crossing that could be reliably identified for each cycle throughout the voice time series was selected. This event was marked throughout the sample by an automatic algorithm. Event marking was verified by the first author. For highly aperiodic stimuli, event locations cannot be considered precise by the standards of perturbation analysis (e.g., Titze, 1995b), but repeat analyses of the most severely aperiodic voices indicated that locations were replicable within ±2 samples (a range of 0.4 ms). Because these values are considerably less than the just-noticeable differences for F0 in this range (which are greater than 2 Hz; e.g., Rossing, 1990), this relatively coarse resolution was considered sufficient for tremor modeling. The frequency of phonation was calculated for each marked cycle of phonation and rounded to the nearest 0.1 Hz for subsequent analyses.

Estimating Tremor Parameters

Figure 2 shows how the rate and amplitude of frequency modulation were estimated for each voice sample. First, the frequency of each cycle of phonation (calculated as the reciprocal of its period) in the original voice sample was plotted against time. Plots were smoothed with a 2-point moving average, and the rate of frequency modulation was estimated visually by counting cycles. (Experiment 3 assesses the adequacy of this estimation procedure.) Estimation based on demodulation techniques (e.g., Winholtz & Ramig, 1992) was attempted, but was abandoned because multiple peaks often emerged from these relatively short, highly irregular stimuli. Estimated rates of frequency modulation ranged from 1 to 12 Hz (mean = 4.5 Hz). This range exceeds that typically found for patients with prominent vocal tremors. For example, Ackermann and Ziegler (1994) reported vocal tremor rates of 20%–30% of the mean F0, and Brown and Simonson (1963) reported tremors ranging in frequency from 4 to 8 Hz.
However, those studies included only patients whose primary presenting complaint was tremor, and patients with milder symptoms were excluded from study (but were included here).

Because it is not known which aspects of frequency variation are perceptually important in tremor, two procedures were used to estimate the extent of frequency variation above and below the mean F0 (the tremor amplitude). In the first, tremor amplitude was estimated by calculating the absolute difference between the frequency of each phonatory cycle and the mean frequency over the entire voice sample. The mean of these absolute differences was used as the initial estimate of tremor amplitude. In the second procedure, tremor amplitude was estimated on the basis of the maximum deviations above and below the mean F0. In this case, the absolute differences between the mean frequency and the maximum and minimum F0 values in each sample were calculated, and the average of these two values was used as the estimate of tremor amplitude. Estimates of tremor amplitude based on average deviations from the mean F0 and based on the maximum and minimum frequencies ranged from 0.6 Hz to 10.3 Hz (mean difference = 4.8 Hz), matched-pairs \( t(31) = -9.74, p < .01 \).

**Voice Synthesis**

Every voice was modeled with both the sine wave and irregular tremor models, the relative merits of which were assessed in Experiment 1. A formant synthesizer implemented in MATLAB (MathWorks, 2001) allowed users to specify F0, the shape of the estimated volume velocity derivative, the spectrum of the inharmonic component of the voice (the noise spectrum), signal-to-noise ratio, formant frequencies and bandwidths, and the tremor parameters described above.\(^2\) Initial parameter estimates for synthesis were derived from acoustic analyses of the voices as follows. Formant frequencies and bandwidths were estimated using autocorrelation linear predictive coding (LPC) analysis (e.g., Markel & Gray, 1976) with a window of 25.6 ms (increased to 51.2 ms when stimulus F0 was near or below 100 Hz). A preliminary estimate of the volume velocity derivative was derived by inverse filtering a single glottal pulse from the microphone recordings. The resulting waveform was fit with a Liljencrants-Fant (LF) source model (Fant, Liljencrants, & Lin, 1985), the parameters of which then specified the harmonic part of the source (see Gerratt & Kreiman, 2001, for further details). The frequency of this cycle (i.e., the reciprocal of its period, as above) served as the initial value of F0. The noise spectrum was estimated by a cepstral-domain comb filter similar to that described by de Krom (1993), which removed the harmonic part of the signal. The residual was then inverse filtered to remove the vocal tract parameters, leaving the inharmonic part of the source. This was fitted with a 25 segment piece-wise linear approximation, which served to specify the noise spectrum.

The synthesis procedure is described in detail elsewhere (Gerratt & Kreiman, 2001). Briefly, the synthesizer sampling rate was fixed at 10 kHz. To overcome quantization limits on modeling F0, the source time series was synthesized pulse by pulse using an interpolation algorithm that tracked the precise beginning of each source pulse relative to sample times. The overall effect is equivalent to digitizing an analog pulse train with pulses of the exact desired frequencies at the fixed 10 kHz sample rate. A 100 tap finite impulse response filter was synthesized for the noise spectrum, and a spectrally shaped time series was created by passing white noise through this filter. The LF pulse train was added to this noise time series to create a complete glottal source time series. The ratio of noise to LF energy was adjusted so that the noise-to-periodic energy ratio approximated the value calculated from the original voice sample. Finally, the complete synthesized source was filtered through the vocal tract model (estimated through LPC analysis, as described above) to generate a preliminary version of the synthetic voice.

Within the synthesizer, the operator adjusted the above parameters from their preliminary estimated values as necessary to achieve the optimal perceptual match to the original voice. In particular, the output of the inverse filter was satisfactory as a starting point for fitting the LF model for all the present stimuli. Parameters of the LF model were always adjusted until the resulting synthetic stimuli provided good perceptual and spectral matches to the original voices. Therefore, any errors in the inverse filtering were not fatal to the final synthetic stimuli. After these adjustments were made, all synthesizer parameters were held constant across experimental conditions. Only tremor-related parameters were varied experimentally.

---

\(^2\)The analysis and synthesis software described in this article is available at http://www.medsch.ucla.edu/glottalaffairs/software_of_the_boga.htm

\(^3\)Jitter and shimmer were not modeled separately from the noise component.
Method

Listeners

Five expert listeners (4 speech-language pathologists and 1 phonetician, including the third author) participated in this experiment. Listeners ranged in age from 25 to 55 years (M = 39). All had daily clinical or research exposure to disordered voices, and all reported normal hearing.

Stimuli

The original voice sample and two synthetic versions of each voice (one created with each tremor model) were used in this experiment. Synthetic stimuli differed only in the tremor model applied (sine wave vs. irregular tremor), with all other synthesizer parameters held constant. The amplitude of frequency modulation equaled the average deviation from the mean F0 for all stimuli. The specified rate of modulation for irregular tremors was twice that used for sine wave tremors, as described above, so that output rates from the two models were roughly equivalent. The two sets of synthetic stimuli did not differ significantly in mean F0, F(1, 62) = 0.06, p > .05, or in standard deviation of F0, F(1, 62) = 4.65, p > .01, indicating that only the pattern of F0 modulation, and not the amount of variation in F0, distinguished the two tremor models.

All stimuli were 1 s in duration. They were equalized for peak amplitude, and onsets and offsets were multiplied by 50-ms ramps to eliminate click artifacts prior to presentation.

Procedure

Listeners heard the two synthetic versions of each voice, each paired with the corresponding original sample. They were asked to rate the similarity of the synthetic to the original stimulus on a 100-mm visual analog scale ranging from exact same (0 mm) to very different (100 mm). They were asked to focus their attention primarily on the tremor component of the voice and to try to consider the overall modulation pattern, instead of making their judgments solely on the basis of the beginning and end points of the frequency contours. An additional 12 voice pairs (20%, selected at random) were repeated, for a total of 76 trials per listener.

Stimuli within a pair were separated by 500 ms. Which stimulus (synthetic or natural) occurred first in a pair varied at random, with the constraint that each occurred first an equal number of times. Pairs of voices were randomized separately for each listener.

Testing took place in a double-walled sound booth. Stimuli were presented in free field over good quality speakers at a constant comfortable listening level. Listeners controlled the rate of stimulus presentation and were able to replay voice pairs as desired before making their responses. Test time totaled approximately 15 min.

Results and Discussion

Across models, listeners judged the match of the synthetic stimuli to the natural targets to be excellent. The average rating for stimuli synthesized with the sine wave tremor model was 25.6 on a 100-point scale (0 indicating that the two stimuli were identical; SD = 26.3). The mean rating for stimuli synthesized with the irregular tremor model was 24.1 (SD = 25.9). Analysis of variance (ANOVA) showed a significant interaction between voice and tremor model, F(31, 256) = 1.57, p < .05, indicating that one tremor model did not consistently perform better than the other. Instead, which tremor model provided the better match to the original voice depended on the pattern of F0 variability. Listeners differed significantly in the level of their ratings, with some using much more of the rating scale than others, F(4, 299) = 8.27, p < .01. However, no interaction was observed between raters and stimulus versions, F(1, 299) = 0.90, p > .01, indicating the pattern of preferences was consistent across subjects.

The limited number of listeners and significant differences among listeners in their use of the rating scale made it difficult to formally evaluate differences between the two tremor models for individual voices. However, across voices, the difference between ratings for the two tremor models depended in part on the severity of the vocal tremor (simple linear regression; F[1, 158] = 3.76, p < .05), with selection of the appropriate model increasingly affecting acceptability of the synthesized stimuli as tremor severity increased. In particular, when tremors are mild, listeners appear relatively insensitive to the precise details of the F0 contour. For example, Figure 3 shows F0 tracks for a voice with a relatively small amount of tremor. The sinusoidal F0 contour shown in panel B follows the original tremor fairly closely, while the irregular tremor in panel C forms a highly smoothed version of the contour. Both tremor models provided excellent perceptual matches to the original voice (sine wave tremor model: mean rating = 10.0 on the 100-point scale; irregular tremor model: mean rating = 3.8). Note, however, that this voice also contains substantial amounts of high-frequency noise, as evidenced by the short-term variations in the F0 contour in panel A. We
speculate that listeners equally preferred the highly smoothed and less smoothed F0 contours because they had difficulty distinguishing the two patterns of moderate long-term variation in the context of significant short-term variation in F0. We return to this hypothesis in the General Discussion section.

Experiment 2

This experiment examined the perceptual impact of the amplitude of frequency modulation—that is, how far a tremor deviates above and below the mean F0. As Figures 1–3 indicate, pathological voices do not necessarily have a single, well-defined modulation amplitude. Some tremor cycles in a voice depart much more (or much less) from the mean F0 than others do, but it is not known how listeners respond perceptually to variations in modulation amplitude. For this reason, we evaluated three different approaches to modeling the amplitude of F0 modulation—one based on the average deviation from the mean F0 in the original voice, one based on the maximum deviations from the mean F0, and one that was selected perceptually.

Method

Stimuli

The 32 voice samples from Experiment 1 were used in this study, along with three synthetic versions of each original voice. All versions of a given voice used whichever tremor model was judged the best match in Experiment 1 (10 sine wave tremors, 22 irregular tremors), but versions differed in the manner in which tremor amplitudes and rates were estimated. For the first version of a given voice, modulation amplitude was estimated based on the average deviation from the mean F0, as described in the Method section. The second version was created with deviations based on the maximum and minimum values of F0 in a sample. Both of these versions used the same estimated modulation rate. The third synthetic version of each voice used perceptually rather than acoustically derived estimates for tremor parameters. This condition was included to test the adequacy of estimation procedures for all the parameters used to model tremors and was created using whatever modulation rate and amplitude provided the best perceptual result, in the opinion of the first author (who created all the stimuli). In these stimuli, adjustments were made to modulation rates in 9 of 32 stimuli, to correct apparent errors in estimating rates for the more irregular stimuli. Values of the amplitude of frequency modulation were also adjusted in 18 of 32 stimuli. Most of these adjustments
resulted in values between the two estimates used in the other stimuli; on average, modulation amplitudes for the perceptually modeled stimuli were slightly but significantly larger than those based on average deviations from the mean F0 (average difference = 0.9 Hz), matched-pairs $t(31) = 3.83$, $p < .01$. Perceptually derived values for the rate of frequency modulation did not differ consistently from the original estimates (average difference = 1.06 Hz), matched-pairs $t(31) = 0.78$, $p > .01$. A repeated-measures ANOVA showed that these three sets of synthetic stimuli did not differ significantly from the original voice samples or from each other in mean F0, $F(3, 93) = 3.92$, $p > .01$, but did differ significantly in the amount of variability in F0 (measured as the standard error of the mean; see Table 1), $F(3, 93) = 58.06$, $p < .01$.

**Listeners**

Ten expert listeners (6 speech-language pathologists, 3 otolaryngologists, and 1 phonetician, including the third author) participated in this experiment. Listeners ranged in age from 25 to 55 years ($M = 38.4; SD = 10.6$). Each had daily clinical or laboratory exposure to pathological voice stimuli, and all reported normal hearing.

**Procedure**

Listeners heard the three synthetic versions of each voice, each paired with the original sample. An additional 19 voice pairs (20%, selected at random) were repeated, for a total of 115 trials per listener. As in Experiment 1, listeners were asked to judge the similarity of each synthetic token to the original voice, on a 100 mm visual analog scale ranging from exact same (0 mm) to very different (100 mm). Other procedures were identical to those used in Experiment 1. Test time totaled approximately 20 min.

**Results and Discussion**

Results of Experiment 2 confirmed that the two tremor models provided excellent copies of naturally occurring frequency modulations. Of the 96 stimuli (comprising all three synthetic versions), 51 received mean ratings of 25 or less on a 100-point scale, and 93 of 96 had mean ratings of 50 or less. Listeners again differed significantly in their levels of rating, $F(9, 929) = 31.38$, $p < .01$, but these differences did not interact with stimulus version, $F(18, 929) = 0.56$, $p < .01$, indicating that all listeners shared the same general pattern of preferences.

An ANOVA showed a significant effect of stimulus version on the acceptability of the stimuli, $F(2, 929) = 11.01$, $p < .01$. Scheffé post hoc comparisons indicated that stimuli based on maximum frequency excursions (mean rating = 29.5) were less acceptable overall than those modeled using the average deviation from the mean F0 (mean rating = 22.4; $p < .01$). Stimuli based on maximum frequency excursions produced stimuli whose tremors sounded too exaggerated. The effect was independent of tremor severity; that is, more severe tremors did not benefit from emphasis on the extremes of frequency deviation, $F(1, 929) = 6.12$, $p > .01$. Modeling based on the average deviations from the mean F0 did produce stimuli with consistently less frequency variability than the original voices, one-sample $t(31) = 4.89$, $p < .01$ (see Table 1), but these small differences in frequency variability apparently were considered perceptually acceptable.

Further Scheffé comparisons indicated that stimuli created by perceptually adjusting synthesizer parameters (mean rating = 23.2) were also significantly preferred overall to stimuli based on maximum F0 deviations ($p < .01$), but did not differ in acceptability from stimuli based on average deviations in F0 ($p > .01$). Apparently, the reliable but small differences between these two sets of stimuli in frequency variability were not perceptually important in the context of the frequency irregularities that occur with pathological voices, although the larger differences in modulation amplitudes in stimuli based on maximum excursions in F0 are perceptually too extreme.

**Experiment 3**

This experiment investigated the perceptual effects of changes in the rate of F0 modulation. In this study, listeners heard synthetic stimuli that differed slightly in tremor rate, and were asked to determine which best matched the original voice sample. They also reported their confidence in their judgments. Patterns of listener preferences, combined with confidence ratings, provide more information about listeners’ ability to hear differences in tremor rates than would similarity ratings like those used in Experiments 1 and 2. Further, because the perceptibility of modulation rates was evaluated within

**Table 1. F0 characteristics of the stimuli in Experiment 2.**

<table>
<thead>
<tr>
<th>Stimulus</th>
<th>$M$ F0 (Hz)</th>
<th>SEM F0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original natural sample</td>
<td>169.1</td>
<td>0.16</td>
</tr>
<tr>
<td>Tremor amplitude based on average deviation from M F0</td>
<td>169.3</td>
<td>0.11</td>
</tr>
<tr>
<td>Tremor amplitude based on maximum deviation from M F0</td>
<td>169.6</td>
<td>0.27</td>
</tr>
<tr>
<td>Perceptually modeled</td>
<td>169.5</td>
<td>0.14</td>
</tr>
</tbody>
</table>

* Differences significantly from original voice sample ($p < .01$).
the context of differences among voices in tremor type and amplitude, this study also provided the opportunity to investigate potential perceptual interactions that may occur among different aspects of F0 modulation.

**Method**

**Stimuli**

Ten voices were selected from the original set of 32 to include a range of rates and extents of modulation. Five were modeled with sine wave tremors and five with irregular tremors (see Table 2). Nine synthetic versions of each original voice were created. The first of these (the central stimulus) was created using the tremor rate estimated from the pitch track, as described in the Analysis and Synthesis Techniques section above. Tremor rates were increased from this central value in steps of 0.25 Hz (0.5 Hz for the irregular tremor model) to create four stimuli; rates were decreased from this central value in steps of 0.25 Hz (0.5 Hz for the irregular tremor model) to create another four stimuli. Thus, tremor rates for each nine-member “family” of stimuli spanned a range of 2 Hz (or 4 Hz for the irregular tremor model), centered around the central stimulus value. All other synthesis parameters were held constant across stimulus versions.

Note that changes in tremor rates have different effects for the two tremor models. For the sine wave model, changes in tremor rate also produced changes in the ending point of the tremor. For example, decreasing the rate by 0.25 Hz means that the tremor’s final point will be 90° out of phase with respect to the basic stimulus version (because only three fourths of a cycle will be completed). Thus, stepwise modification of the tremor rate in the sine wave model assessed the perceptual importance of matching the end point of an F0 contour precisely, as well as the importance of the rate of frequency modulation. However, altering the rate for the irregular model had no consistent effect on tremor phase, because the model generated a different irregular pitch contour each time it was invoked.

**Listeners**

Ten expert listeners (1 speech-language pathologist, 1 otolaryngologist, and 8 phoneticians, including the first and third authors) participated in this experiment. Listeners ranged in age from 23 to 52 years ($M = 32.6$). Each had substantial experience evaluating voice quality, through daily clinical or laboratory encounters. All reported normal hearing.

**Procedure**

For each trial, listeners heard two pairs of voices (AB and AC). The first member of each pair (A) was always the original voice sample, and the second was one of the nine synthetic copies of that sample. In an additional 90 “same” trials, both voices in one pair were the original sample (for a total of 450 trials per listener). Listeners were asked to compare the pairs and decide whether B or C provided a better match to A. They were also asked to rate their confidence in each response on a 5-point scale ranging from wild guess (1) to positive (5). Listeners were able to play each pair as often as necessary before making a response. Stimuli were randomized for each listener and were presented in free field at a constant comfortable level in a double-walled sound booth. Voices within a pair were separated by 350 ms; the interpair interval was controlled by the listener, as was the rate at which trials were presented. To reduce listener fatigue, testing took place in two sessions, each lasting about 50 min.

**Results and Discussion**

Listeners usually judged that the original natural stimulus was the best match to itself, selecting this pair on 93% of “same” trials. Across the 10 voice families, error rates on these trials ranged from 1.1% to 13.3%. However, listeners were not especially confident of their choices. Mean confidence for comparing a synthetic stimulus to a natural voice was 3.77 ($SD = 1.38$) on the 5-point scale. This, combined with fact that some synthetic stimuli were confused with natural ones for every stimulus family, demonstrates the success of the synthesis at imitating the original voices.

The task of deciding which of two synthetic stimuli best matched the natural stimulus was difficult. Mean confidence for trials without “same” pairs was 2.10 ($SD = 1.38$).
To determine how well listeners were able to distinguish differences in tremor rates, we examined trials where one pair of voices included the central stimulus and the other pair included a second synthetic voice. Among synthetic stimuli, the central stimulus (created using the visually guided estimate of the tremor rate) was selected as the best match to the natural voice on about 62% of trials, which exceeded chance levels: sine wave tremors, $\chi^2(1) = 29.16, p < .01$; irregular tremors, $\chi^2(1) = 20.25, p < .01$. Responses on the remaining trials were distributed rather evenly across the other synthetic versions (see Table 3). For both sine wave and irregular tremors, no relationship was observed between listener preferences and the amount of difference between the second synthetic voice and the central stimulus in tremor rate: sine wave tremors, $F(1, 6) = 0.02, ns$; irregular tremors, $F(1, 6) = 3.03, ns$. For sine wave tremors, listeners' confidence did increase with the distance between the central stimulus and the second synthetic stimulus, $F(1, 398) = 5.19, p < .05$, but for irregular tremors, confidence did not vary with the amount of difference between stimuli, $F(1, 398) = 1.77, ns$. These results suggest that listeners are not especially sensitive overall to small differences in tremor rate and that the tremor phase (the actual frequency endpoint) of a stimulus had little effect on listeners' judgments. Given this relative insensitivity, our somewhat informal method of estimating tremor rates appears adequate.

For sine wave tremors, listeners' confidence was higher for synthetic stimuli paired with the natural stimuli having the fastest central tremor rates, indicating that differences in rate are easier to hear for faster rates, $F(1, 178) = 74.97, p < .05$. However, greater tremor amplitudes significantly reduced the ability to detect rate differences, $F(1, 178) = 28.16, p < .05$. Similar effects of rate manipulations were not observed for irregular tremors: tremor rate, $F(1, 178) = 3.15, ns$; tremor amplitude, $F(1, 178) = 0.74, ns$.

These results suggest that perceptual sensitivity to differences in tremor rate depends on the complexity of the total pattern of frequency variation. As the amplitude of a sine wave tremor increases, the complexity of the pattern increases, so listeners have more difficulty isolating and attending to rate alone. Similarly, as the pattern of frequency modulation departs from sinusoidality, listeners are increasingly unable to resolve the changes in tremor, apparently because the background tremor pattern is uncertain. When the overall pattern of frequency modulation is most complex, listeners apparently respond to overall levels of F0 variability (measured by the standard error of the mean, e.g., as in Experiment 2 above), rather than to precise patterns of frequency change.

### General Discussion

In this study, we examined the perception of frequency modulation in pathological voices, using a variety of voice synthesis strategies. Sine waves provided a good approximation to the perceived pattern of frequency modulation for some voices, but other voices were better modeled with an irregular modulating waveform. The perceptual importance of the shape of the modulating waveform appears to depend on the severity of the tremor, with the choice of model increasing in importance as the tremor increases in severity and salience. The average deviation from the mean F0 better approximated the perceived amplitude of a tremor than did the maximum deviations in F0, although listeners were not particularly sensitive to small changes in tremor amplitudes. Differences in tremor rates were easiest to hear when the tremor was sinusoidal and of small amplitude. Differences in rate were difficult to judge for tremors of large amplitude, or in the context of irregularities in the pattern of frequency modulation.

Amplitude modulation (as opposed to the amplitude of frequency modulation; see footnote 1) was not explicitly modeled in this study, although amplitude modulations occurred in all the natural and synthetic stimuli, presumably due to the movement of harmonics toward and away from resonance peaks as F0 varied. The perceptual importance of amplitude modulation cannot be

### Table 3. Preference rates for different synthetic stimuli.

<table>
<thead>
<tr>
<th>Stimulus type</th>
<th>Central</th>
<th>Central ± 0.25 Hz</th>
<th>Central ± 0.5 Hz</th>
<th>Central ± 0.75 Hz</th>
<th>Central ± 1 Hz</th>
<th>Total # trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sine wave</td>
<td>245</td>
<td>39</td>
<td>35</td>
<td>43</td>
<td>38</td>
<td>400</td>
</tr>
<tr>
<td>Irregular</td>
<td>254</td>
<td>45</td>
<td>38</td>
<td>27</td>
<td>36</td>
<td>400</td>
</tr>
</tbody>
</table>
determined without formal evaluation. Consistent with anecdotal evidence (Sundberg, 1995), the excellent quality of the synthesis suggests that such modulations are not perceptually important for every tremulous voice, apart from the frequency modulations that produced them. However, some voices, including severe examples of spasmodic dysphonia, may require formal modeling of amplitude modulations.

Note also that the present study makes no claims regarding the physiological functions that produced acoustic frequency and amplitude modulations. Our goal was to determine which acoustic characteristics of the voices were perceptually important, and to examine how different acoustic characteristics interacted to determine the nature of the perceived tremor. A broader long-term goal for such studies would be to understand which physiological aspects of tremor produce perceptually important acoustic changes. However, pursuit of this desirable but ambitious goal must await the emergence of voice models that relate physiology to acoustics to perception in a unified theoretical framework. (See Titze & Story, 1997, for an example of a physiologically based model of the voice source.)

With this caveat in mind, we found no acoustic basis for formally distinguishing classes of tremors (e.g., wow and flutter) on the basis of modulation rates. No obvious discontinuities were observed in the distribution of estimated modulation rates for these 32 voices, and modulation rates were not related to the perceived severity of the tremor \( r = -0.14, \text{ns} \). It is, therefore, unclear where boundaries should be drawn between different categories of tremors. Synthesizer parameters provide continuous quantification of frequency modulation, and using such parameters to describe tremors obviates the need for categorical measurement systems (Gerratt & Kreiman, 2001).

The results of Experiment 2 suggest that difference limens for the amplitude of frequency modulation are fairly large when the target stimuli themselves vary irregularly. That is, relatively small differences in the amount of frequency variability are apparently treated as consistent with the overall variability of the original stimuli and are not noticed until they exceed some threshold level. Additional studies systematically manipulating the amplitude of frequency modulation may shed further light on how listeners perceive frequency modulations occurring in perceptually complex, changing contexts.

The present results further suggest that tremor rate, regularity, and amplitude interact, so that the perceptual importance of any one dimension depends on values of the others. Psychoacoustic studies have reported similar perceptual interactions between stimulus dimensions (Melara & Marks, 1990; Melara & Mounts, 1994).

In those studies, variation on an unattended dimension (e.g., pitch) significantly interfered with listeners’ abilities to perceive differences on a target dimension (e.g., loudness). Much further work will be necessary to determine the extent, pattern, and mechanisms of perceptual interference among dimensions of voice quality. However, to the extent that such interference occurs, it argues against the use of traditional unidimensional rating scale approaches to quality measurement. Unidimensional rating scale instruments can never adequately measure what listeners hear when the value of a stimulus on one perceptual dimension depends on the value of another dimension. (See also Van Lancker, Kreiman, & Wickens, 1985, for discussion of similar effects in the perception of personal identity from voice.) Such interactions can be modeled and studied systematically in a synthesis approach like that applied here, which is a significant advantage of this new method.

Acknowledgments

This research was supported by Grant DC01797 from the National Institute on Deafness and Other Communication Disorders. We thank Norma Antonanzas for additional programming support.

References


Received May 22, 2002
Accepted August 20, 2002
DOI: 10.1044/1092-4388(2003/016)

Contact author: Jody Kreiman, Head/Neck Surgery, The David Geffen School of Medicine at UCLA, 31-24 Rehab Center, Los Angeles, California 90095-1794. E-mail: jkreiman@ucla.edu
Perception of Vocal Tremor

Jody Kreiman, Brian Gabelman, and Bruce R. Gerratt

*J Speech Lang Hear Res* 2003;46;203-214
DOI: 10.1044/1092-4388(2003/016)

This article has been cited by 1 HighWire-hosted article(s) which you can access for free at:

http://jslhr.asha.org/cgi/content/abstract/46/1/203#otherarticles

This information is current as of April 1, 2011

This article, along with updated information and services, is located on the World Wide Web at:

http://jslhr.asha.org/cgi/content/abstract/46/1/203