ELSEVIER

Speech Communication 22 (1997) 343–368

SPEECH
COMMUNICATION

Analysis by synthesis of pathological voices using the Klatt synthesizer

Philbert Bangayan [a], Christopher Long[a] [1], Abeer A. Alwan [a,*], Jody Kreiman [b], Bruce R. Gerratt [b]

[a] *Department of Electrical Engineering, 66-147E Engr. IV, School of Engineering and Applied Sciences, UCLA, 405 Hilgard Avenue, Box 951594, Los Angeles, CA 90095-1594, USA*

[b] *Division of Head and Neck Surgery, UCLA School of Medicine, CHS 62-132, Los Angeles, CA 90095, USA*

Received 11 September 1995; revised 18 September 1996; accepted 30 June 1997

Abstract

The ability to synthesize pathological voices may provide a tool for the development of a standard protocol for assessment of vocal quality. An analysis-by-synthesis approach using the Klatt formant synthesizer was applied to study 24 tokens of the vowel /a/ spoken by males and females with moderate-to-severe voice disorders. Both temporal and spectral features of the natural waveforms were analyzed and the results were used to guide synthesis. Perceptual evaluation indicated that about half the synthetic voices matched the natural waveforms they modeled in quality. The stimuli that received poor ratings reflected failures to model very unsteady or "gargled" voices or failures in synthesizing perfect copies of the natural spectra. Several modifications to the Klatt synthesizer may improve synthesis of pathological voices. These modifications include providing jitter and shimmer parameters; updating synthesis parameters as a function of period, rather than absolute time; modeling diplophonia with independent parameters for fundamental frequency and amplitude variations; providing a parameter to increase low-frequency energy; and adding more pole-zero pairs. © 1997 Elsevier Science B.V.

Résumé

Pouvoir synthétiser des voix pathologiques peut être un outil utile pour le développement de protocoles standards pour l'évaluation de la qualité vocale. On a appliqué une méthode d'analyse-synthèse, utilisant le synthétiseur de Klatt, à 24 énoncés de la voyelle /a/ prononcés par des locuteurs hommes et femmes ayant des troubles de la parole allant de modéré à sévère. Les indices tant spectraux que temporels des formes d'onde naturelles ont été analysés et les résultats utilisés pour guider la synthèse. Une évaluation perceptive montre qu'environ la moitié des voix synthétiques sont considérées comme identiques en qualité aux voix naturelles qu'elles sont sensées modéliser. Les stimuli considérés comme différents reflètent les échecs de la modélisation de voix très instables ou "gargarisées" ou d'imperfection dans la reproduction des spectres naturels. Diverses modifications apportées au synthétiseur de Klatt pourraient améliorer la synthèse des voix pathologiques. Ces modifications concernent l'introduction de paramètres de jitter et de shimmer; la mise à jour des paramètres de synthèse en fonction de la période plutôt qu'en fonction d'une durée absolue; la modélisation de la diplophonie par des paramètres

[*] Corresponding author. E-mail: alwan@icsl.ucla.edu.

[1] Present Address: Speech and Hearing Sciences Program, HST, MIT, 77 Mass. Ave., Cambridge, MA 02139, USA.

PII S0167-6393(97)00032-0

indépendants pour les variations de fréquence fondamentale et d'amplitude; la disponibilité d'un paramètre permettant d'augmenter l'énergie dans les basses fréquences; et l'ajoût de plus de paires pôle-zéro.

Keywords: Analysis-by-synthesis; Voice quality; Pathological voices

1. Introduction

No accepted standard system exists for describing pathological voice qualities (e.g., Jensen, 1965; Yumoto et al., 1982; see Kreiman and Gerratt, 1996, for a review). Qualities are labeled based on the perceptual judgments of individual clinicians, a procedure plagued by inter-and intra-rater inconsistencies and terminological confusions. Synthetic pathological voices could be useful in a standard protocol for quality assessment (Gerratt et al., 1993; Kreiman and Gerratt, 1996). This paper describes a pilot study of the mechanics of synthesizing moderately-to-severely pathological voices and provides guidelines for synthesizing some kinds of pathological voice qualities.

Speech synthesizers with the ability to model a range of vocal qualities have many applications, including improved vocal prostheses (Qi et al., 1995), analysis and coding of natural-sounding speech (e.g., Price, 1989; Karlsson, 1991), and modeling phonation types in languages with voice-quality contrasts (Ladefoged, 1995). Accordingly, source models have received increasing attention in the literature (e.g., Ananthapadmanabha, 1984; Fant et al., 1985; Fujisaki and Ljungqvist, 1986). Recent studies (Gobl, 1988; Klatt and Klatt, 1990; Carlson et al., 1991; Imaizumi et al., 1991; Gobl and Ní Chasaide, 1992; Karlsson, 1992; Lofqvist et al., 1995; Ladefoged, 1995) have focused on variations in normal quality, rather than on pathology. With the exception of studies by Childers and colleagues (Childers and Lee, 1991; Lalwani and Childers, 1991; Childers and Ahn, 1995), attempts to synthesize pathological voices have not been reported, and synthesis of such voices is not well developed. Childers and colleagues modeled modal, fry, falsetto and breathy phonation in patients with a variety of diagnoses; other types of pathological voices were not examined. Their work revealed limitations of existing source models and suggested that a turbulent noise component and a pitch perturbation generator were necessary to model breathy voices. These features have proved useful for modeling normal voices as well, and have been added to implementations of the popular Liljencrants/Fant (LF) model (Fant et al., 1985) in other laboratories (e.g., Carlson et al., 1991; Karlsson, 1992).

Despite the predominant focus on normal speakers, previous synthesis studies provide some insight into pathologic voices, because many of the qualities examined occur in pathology. Further, modeling continuous speech resembles modeling pathologic voices, in that both tasks require a dynamic source model to mimic changes over the course of an utterance.

However, from our perspective these studies have significant limitations. They typically used small numbers of speakers (often as few as one or two). They examined a limited range of qualities (typically breathiness, creak, hoarseness, harshness and modal voice, following Laver's classification (Laver, 1980)), as produced by normal speakers. Finally, formal perceptual evaluation of the resulting synthesis has been very limited or absent in studies of both normal and pathological voice. Most authors determine which LF parameters best sort voices into a priori perceptual categories, or merely report whether synthesis quality is "good" or "improved". Lack of detailed perceptual data also makes it difficult to determine the necessary and sufficient parameters to control a synthesizer. Although the LF source model (Fant et al., 1985) specifies 4 timing parameters, many different combinations of these parameters can be used to control synthesis. Authors differ considerably in how they define control parameters, largely because perceptual data to guide standardization are lacking. Modeling of voice quality in these studies has not been driven by an interest in acoustic-perceptual relations. Thus, they have generated little insight into the perceptual importance of different features of glottal pulses, or of different synthesizer control parameters (e.g., Ananthapadmanabha, 1984). This information is critical for the development of efficient and standardized synthesis strategies for both pathological quality and for variations in normal quality.

The present study used the Sensyn 1.1 (Sensimetrics, Cambridge, MA) version of the Klatt formant synthesizer (Klatt and Klatt, 1990) to synthesize a random sample of moderately-to-severely pathological voices. The Klatt synthesizer was chosen because it is commercially available, widely used, and often referenced. In addition, the synthesizer includes a turbulent noise component, pole and zero pairs that can be used to model tracheal and/or nasal coupling, a provision for time-varying parameters to model unsteady qualities, and a "diplophonia" parameter to model bifurcated phonation. However, the synthesizer was originally designed for synthesizing normal voices, and questions remain about its suitability for producing acceptable pathologic stimuli. In fact, the experiments reported in this study led to a number of suggested modifications to the synthesizer that would facilitate synthesis of pathological voices.

2. Analysis-by-synthesis

2.1. Stimuli

Twenty-four samples of the vowel /a/ were selected from a library of voice recordings. Signals were recorded with a miniature head-mounted microphone (AKG C410) placed 4 cm away from the speaker's lips (Winholtz and Titze, 1997). Use of vowel stimuli has a number of advantages. First, isolated vowels are routinely used in clinical practice for evaluation of pathological voice quality. Second, acoustic analysis and synthesis are more straightforward for vowels than for continuous speech. Study of continuous speech is the ultimate goal and an obvious next step. However, valid results based on less complex stimuli are first required.

Signals were low-pass filtered at 8 kHz, digitized at 20 kHz, and then downsampled to 10 kHz, the maximum sampling rate at which all synthesizer parameters could be manipulated. One second segments were excerpted from the middle portion of each natural sample.

Each voice was given an informal severity rating by authors JK and BG, who are experienced in perceptual ratings of pathological voices. Ratings were made on a 6-point EAI (Equal-Appearing Interval) scale, where 1 represented near-normal voice quality and 6 represented extremely severe pathology. Because this study focused on moderately-to-severely pathological voices, only samples rated 3 or higher were chosen.

2.2. Acoustic analysis

Time-and frequency-domain analyses of each voice sample were undertaken to guide synthesis efforts. Most of the effort was directed at matching the time-varying spectra of the natural utterances. Analyses included measuring the fundamental frequency (F_0), strengths of the first three harmonics, formant frequencies, and any additional resonances. Most analyses were performed using SpeechStation (version 3.1 for the IBM PC; Sensimetrics, Cambridge, MA), because it is compatible with the Sensimetrics synthesizer, and because it can display the natural and synthesized speech files simultaneously. WAVES software (version 5.0 for the Sun SparcStation; Entropic Research Laboratory, Washington, DC) was also used, especially for time-domain analysis.

Time-domain analyses included measuring the amplitude and fundamental frequency (F_0) of the voices. As a first pass, the SpeechStation F_0 tracking algorithm, which consists of center clipping followed by autocorrelation and parabolic interpolation, was used. If the F_0 tracker failed to produce a reasonable F_0 contour, as in the case of voices with high jitter, then F_0 was measured manually from the time waveform.

Frequency-domain analyses included computing 14th-order LPC spectra to measure the formant frequencies and using DFT spectra to determine the overall spectral shape, the strength of the first three harmonics, and the locations of poles and zeros due to nasal and/or tracheal coupling. The analysis window was a Hamming window whose length was varied as necessary to measure variations in quality over the duration of a sample. For example, steady-state segments can be measured with a longer-duration window than rapidly-varying segments. Spectrograms were used throughout this process to visualize the time course of the waveform.

2.3. Synthesis

Synthetic waveforms were modeled after each of the natural tokens using Sensyn 1.1. All samples were synthesized with a sampling rate of 10 kHz, using the cascade branch of the synthesizer and a version of the LF source model (SS = 3).[2] The Klatt synthesizer is based on the source-filter theory of speech production and, for the cascade implementation, consists of 34 modifiable parameters. Default values for these parameters are listed in Appendix A. The parameters used to control the glottal pulse shape and timing are the open quotient (OQ), defined as the percentage of time the glottis is open in one fundamental period; the speed quotient (SQ), defined as the ratio of the duration of the rising portion of the glottal pulse to that of the falling portion; the fundamental frequency (F_0); the tilt of the voicing source spectrum (TL); the amplitude of aspiration noise (AH); the amplitude of voicing (AV); the flutter parameter (FL), which adds a quasi-random component to the nominal F_0 value; and the degree of diplophonic double pulsing (DI). Synthesis was aimed at matching the spectro-temporal details of the natural waveforms. It was undertaken by authors PB and CL and supervised by author AA, who has more than 10 years experience in speech synthesis with the Klatt synthesizer. Synthesis proceeded as follows.

Step 1: Match parameters in the frequency domain
The first step in the synthesis was to match frequency domain parameters, such as F_0, formant frequencies (parameters F_1–F_6), and formant bandwidths (parameters B_1–B_6). Bandwidths were chosen such that the amplitudes of the natural and synthetic formants matched. The parameter SQ was adjusted if necessary to match the overall spectral slope. The synthetic sample was then played back to check for vowel quality.

Step 2: Adjust amplitude of voicing (AV and GV) and amplitude of aspiration noise (AH and GH)
Next, the amplitude of voicing was adjusted to match the intensity of the natural waveform as closely as possible. This was important because loudness affects the perceived similarity of two voices (Kempster et al., 1991). Likewise, the amplitude of the aspiration noise was adjusted to match the degree of aspiration or breathiness in the natural sample. This was done by adjusting AH to match the amount of noise present in the spectrogram of the natural voice. Fine-tuning of GH and GV was determined by perceptual evaluation. When synthesizing pathological voices, careful manipulation of aspiration noise is as important as that of the amplitude of voicing because, for example, the degree of aspiration noise can be an important factor in the perception of rough and breathy qualities (Kreiman et al., 1993, 1994).

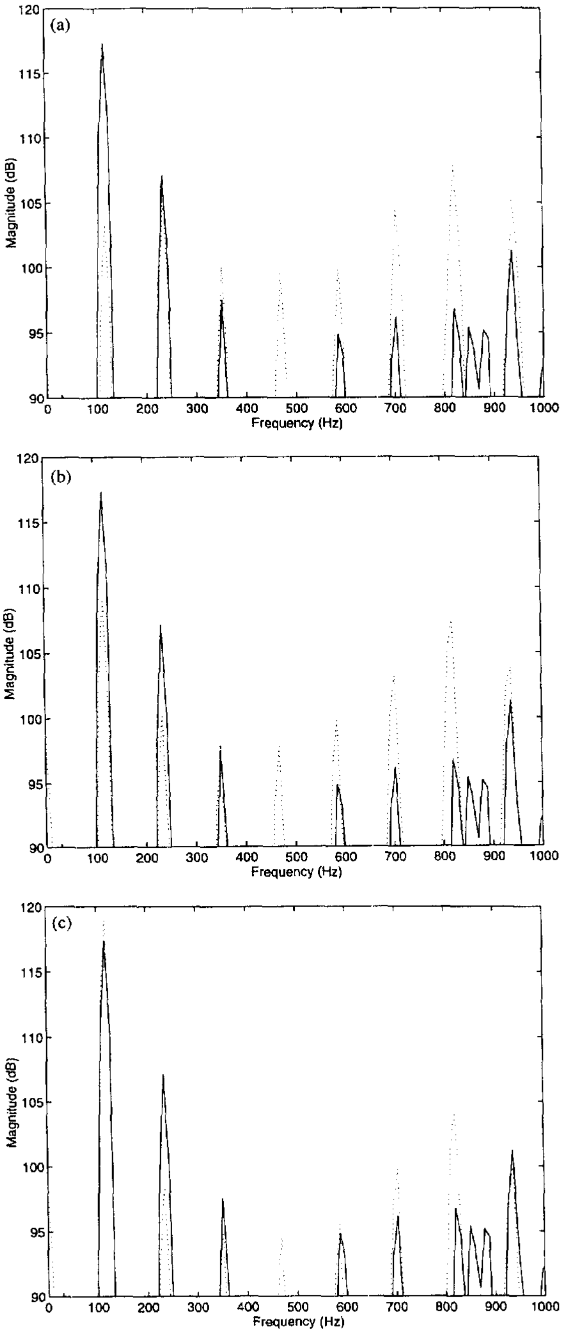
Step 3: Adjust open quotient (OQ)
The third step was to match the degree of strain or breathiness in a voice, if present, by altering OQ. Normal voices are characterized by an OQ of about 50%; strained voices have OQ < 50%, and breathy voices typically have OQ > 50% (Klatt and Klatt, 1990). In the frequency domain, increasing OQ strengthens the amplitude of the first harmonic.

Step 4: Adjust low-frequency harmonics
It was often difficult to match the amplitudes of harmonics below F_1 in the synthetic vowels to those of the natural samples. This harmonic mismatch resulted in synthetic voices that did not sound as "rich" as the natural voices. The synthesizer provides two pole-zero pairs. One pair (FNP, FNZ) is intended to model a pole and zero that may arise from coupling to the nasal tract, while the other pair (FTP, FTZ) models a pole-zero pair that may arise from coupling to the trachea. The default values of the synthesizer restrict both pairs to frequencies below 3000 Hz. To increase the amplitude of a harmonic, the nasal and/or tracheal pole-zero pairs were placed at that harmonic, keeping the bandwidth of the pole narrower than that of the zero. Similarly, particular frequency regions were attenuated by placing a pole-zero pair in that region and adjusting the pole and zero bandwidths such that the bandwidth of the zero was narrower than that of the pole.

Fig. 1 provides an example in which the amplitude of the first harmonic is manipulated by changing OQ and placing a pole-zero pair at the frequency of the harmonic. Fig. 1(a) shows the energy of the

[2] The LF source is implemented as a filtered impulse.

first harmonic when OQ = 50%. In this case, the first harmonic of the synthetic stimulus is 14 dB less than that of the natural stimulus. Increasing OQ to 90% (Fig. 1(b)) increases the strength of the first harmonic to 8 dB below that of the natural stimulus. Notice that increasing OQ, in this case, decreases the amplitude of the second harmonic (Klatt and Klatt, 1990). With OQ = 90% and a pole-zero pair placed at the frequency of the first harmonic, the amplitudes of the natural and synthetic first harmonic are about the same (Fig. 1(c)). Likewise, we can compensate for the second harmonic amplitude decrease by placing a pole-zero pair at the frequency of that harmonic.



Step 5: Alter fundamental frequency (F_0)

Next, F_0 was varied to model the natural utterances. Four approaches were used. In the first, the Klatt parameters FL and DI modulated the F_0 value used in Step 1. FL slowly and regularly varies F_0 values as described by

$$\Delta F_0 = \frac{\mathrm{FL}}{50}\frac{F_0}{100}(\sin(2\pi 12.7t) + \sin(2\pi 7.1t) + \sin(2\pi 4.7t))\ \mathrm{Hz}.$$

DI varies F_0 by delaying every other pulse and decreasing its amplitude. As a result, the pitch period alternates between $T_0 - \Delta T_0$ and $T_0 + \Delta T_0$, where

$$\Delta T_0 = \frac{\mathrm{DI}}{100} T_0 \left(1 - \frac{\mathrm{OQ}}{100}\right).$$

The shorter pitch period is attenuated by

$$\Delta \mathrm{AV} = \mathrm{AV}\left(1 - \frac{\mathrm{DI}}{100}\right).$$

Fig. 2 shows the effect of changing DI. Part (a) shows a glottal waveform with DI = 0%; in part (b), DI = 50%. This technique improved the naturalness of some voices with relatively steady F_0 values, and of some bifurcated voices.

Fig. 1. Discrete Fourier Transform (DFT) spectra of a natural (solid line) and synthetic (dashed line) voice. (a) With OQ = 50%, the first harmonic amplitude for the natural utterance is 14 dB higher than that of the synthesized voice. (b) Using OQ = 90%, the difference is 8 dB. (c) Adding a pole/zero pair (FNP, FNZ) at the frequency of the first harmonic and using OQ = 90% yields a better match. The parameters used were FNP = FNZ = 116 Hz, BNP = 40 Hz, BNZ = 180 Hz. Placing a pole-zero pair near any harmonic affects the amplitude of that, and possibly adjacent, harmonic(s).

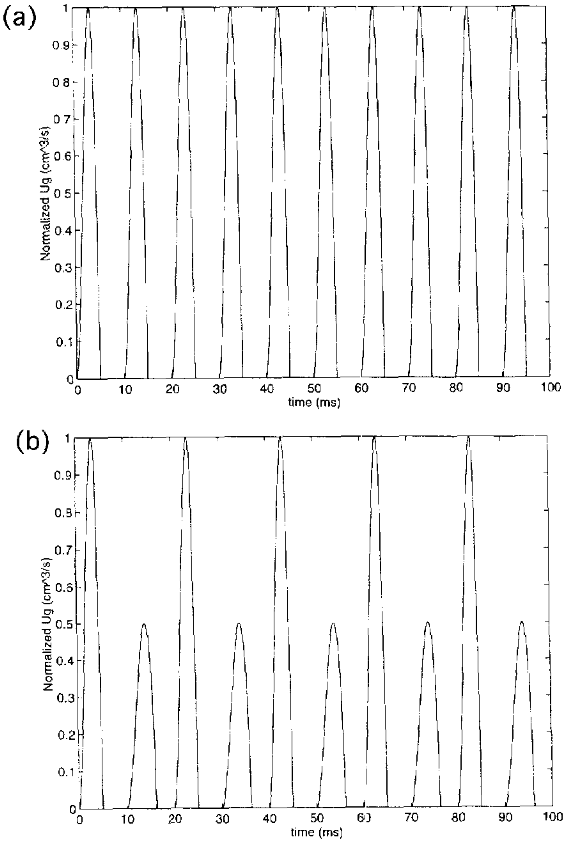


Fig. 2. Plot of the glottal waveform (volume velocity versus time) with OQ = 50% and (a) DI = 0%, (b) DI = 50%.

The second technique involved importing the F_0 contour calculated with the SpeechStation pitch tracker. This technique worked well for voices with an F_0 that changed slowly enough to be tracked by the pitch tracker, with maximum variations of 20 Hz.

A third technique was to model F_0 as a Gaussian random variable (Hillenbrand, 1987) with mean and variance derived from the natural sample. This technique worked well for some rough and rough-breathy voices.

When none of the above techniques produced an acceptable sounding F_0 contour, manually measured F_0 values were used. This technique, while cumbersome, improved the synthesis of some bifurcated voices.

Special care was taken to set the update interval (UI) equal to the greatest integer (in ms) less than the average period. This is important because F_0 values, unlike some other synthesis parameters, are updated at the beginning of the period rather than at each UI. Thus, with an alternating T_0, or a randomly varying F_0, the proper UI must be specified.

Step 6: Alter AV as a time-varying parameter for amplitude-modulated voices

The parameter AV was time-varied to model shimmer (period-to-period amplitude variations) and amplitude modulation (longer-term variations) in some voices.

Step 7: Add additional pole-zero pairs if necessary

Finally, some voices required pole-zero pairs to model nasal and/or tracheal coupling. This step was performed if the synthesizer's pole-zero pairs were not both used to boost or attenuate the energy in certain frequency regions (Step 4).

These seven steps were repeated and fine-tuned using visual comparisons of spectrograms and short time spectra as well as perceptual evaluations in the following manner. In the first iteration, visual cues were used to guide the operator for five of the seven steps. The formant frequencies were determined by matching spectrograms of the synthetic and natural stimuli. The bandwidths and speed quotient were determined by comparing short-time Fourier spectra. The parameters AV, GV, AH and GH were determined by matching the intensity of the spectrograms as well as the energy in the waveform. When changing the voicing and noise source amplitudes, it was necessary to alter the formant bandwidths to obtain the appropriate formant amplitude. Steps involving the adjustment of the first few harmonics (steps 3, 4 and 7) were performed by comparing short time spectra. Detailed measurements of period-to-period variations (fundamental frequency and amplitude) were taken in order to measure jitter and shimmer. These values were used to guide the adjustment of the time varying parameters F_0 and AV. Once these steps were performed and repeated several times, a decent synthetic voice was obtained. Next, the synthesis was fine-tuned using perceptual evaluations. Two synthetic voices were compared to the natural utterance in the pattern: synthetic attempt *A*, natural, synthetic attempt *B*. Synthetic voice *A* represented the best synthesis thus far, and synthetic voice *B* differed, in most cases, from the former by simply one or two parameters. The three voices were played; the better synthetic voice was deemed the new synthetic attempt *A*, and alterations were made to create a new synthetic voice *B*. This process was repeated until either all attempts to improve the voice resulted in a perceptually worse voice or these attempts resulted in a synthetic voice *B* that was perceptually identical to synthetic voice *A*. At that point, synthetic attempt *A* was determined to be the best possible synthesis. In this fashion, the synthetic voice was matched to the natural voice as closely as possible. Synthesis of a single token took from 1–20 hours, depending on the severity of the vocal pathology.

3. Perceptual evaluation

Some of our attempts to synthesize pathological voices were subjectively more acceptable than others. The following experiment was undertaken to evaluate the overall quality of the synthesis, and to determine which voices listeners considered good matches to the original samples.

3.1. Methods

Ten expert listeners (6 speech-language pathologists, three otolaryngologists, and one phonetician, including authors JK and BG) participated in this experiment. Each had a minimum of one year postgraduate experience evaluating voice quality, and none reported any speech, language or hearing difficulties. The 24 voices described in Section 2 were used as stimuli. Each sample (synthetic and natural) lasted 1 second. Stimuli were normalized for peak voltage, and onsets and offsets were multiplied by 25 ms ramps to eliminate click artifacts.

All listening tests took place in sound-treated booths. To mimic actual clinical listening conditions as closely as possible, all testing was done in free field. Listeners were seated 3 feet from a high fidelity loudspeaker (Boston Acoustics A40). Stimuli were low-pass filtered at 8 kHz and played through a 16-bit D/A converter at a constant listening level (approximately 80 dB SPL). Responses were recorded and stored by a computer.

Listeners heard each natural sample paired with its synthetic copy, and were asked to judge how well the copy matched the original (on a 7-point scale, where 1 indicated a perfect match in quality). Complete listener instructions are given in Appendix B. Voice pairs were always presented in the order natural/synthetic. Three additional pairs consisting of two identical natural stimuli were also included. Tokens within a pair were separated by 500 ms. Each of the 27 voice pairs was presented twice (although listeners were not told this); stimuli were played out, re-randomized, and played out again. Different random orders were used for each listener. Listeners controlled the rate of presentation, and were able to replay the voice pairs as often as necessary. The experiment lasted approximately 10 minutes.

3.2. Perceptual results

Test-retest reliability was acceptably high for all listeners. Across listeners, Pearson's *r* for the first versus second rating of a voice pair was 0.83 (range of individual values = 0.66–0.89); also across listeners, the first and second rating of a voice differed by 0.74 scale value on average. Because most listeners were unfamiliar with synthesized speech, the first set of judgments was treated as practice and discarded.

Performance on trials where voices were identical was also satisfactory. Only 1 listener failed to rate these voices as being identical in quality ("1"). However, that listener used the category "1" much less frequently than other listeners (3/56 trials versus an average of 12.6 trials rated "1" for the other listeners). Given that this listener rated voices consistently (test-retest Pearson's $r = 0.84$; 89.3% of ratings within 1 scale value), we concluded that these ratings probably represent response bias, rather than lapses of attention, and data from this subject were retained.

Interrater reliability was also acceptable. Ratings for 8 of the 10 listeners were consistently correlated at $r = 0.7$ or better (mean Pearson's $r = 0.79$, sd = 0.05, range = 0.7–0.88). The remaining two listeners used a limited range of values when making their ratings. One gave 18 of 28 pairs a rating of "1" (identical qualities), and never rated a pair above "4". The second, as described above, rarely used the value "1". Ratings for these listeners were less well correlated with the remainder of the group (average Pearson's $r = 0.59$, sd $= 0.14$, range $=$ 0.35–0.87). However, because they rated voices consistently, their data were retained.

Listeners unanimously reported being pleased by the overall quality of the synthesis. Table 1 shows the average rating for each voice, which ranged from 1.3 to 6.3. On the whole, copies of voices with milder pathology were more acceptable than those of more severely disordered voices (Pearson's r comparing mean rating and severity $= 0.60$, $p < 0.05$). Copies of male voices were more acceptable overall than were copies of female voices (males: mean $=$ 2.99 females: mean $= 4.19$; $F(1,23) = 5$, $p < 0.05$).

3.3. Discussion

Our efforts to synthesize moderately-to-severely pathological voices were variably successful. Less severely pathological voices were synthesized best, and male voices were synthesized more successfully than female voices: of the 13 stimuli rated 3.5 or better, only 3 modeled female voices. This gender difference has been noted previously (Klatt and Klatt, 1990), and suggests that the synthesizer's glottal-source model is better suited to synthesize male voices than female voices. In addition, efforts to synthesize voices having slow, unsteady time variations, in addition to their period-to-period fluctuations, in amplitude of voicing or fundamental frequency were often less successful due to the additional complexity of this task.

Although the success of the synthesis is better predicted by severity or gender than by the "qualities" a voice might possess, we found some synthesis parameters that were common to voices with prominent turbulence noise (rough or rough-breathy voices) and to voices with bifurcated/bicyclic phonation. The following sections describe the analysis-by-synthesis procedures employed for these

Table 1
Results of perceptual evaluation

Token	Category	Gender	Mean rating	Std. dev.	Voice severity
bim2	bifurcated	male	1.3	0.483	4
rm2	rough	male	1.7	0.675	5
bim1	bifurcated	male	1.9	0.994	5
rbrm4	rough-breathy	male	1.9	0.738	5
bif1	bifurcated	female	2.0	0.943	4
rbrm1	rough-breathy	male	2.3	0.823	3
rf1	rough	female	2.3	0.949	6
rbrm2	rough-breathy	male	2.5	0.850	5
rbim	rough-bifurcated	male	2.8	0.789	5
rbrm3	rough-breathy	male	2.9	0.994	5
rm1	rough	male	3.0	1.054	4
bim4	bifurcated	male	3.4	1.174	5
bif3	bifurcated	female	3.5	1.269	5
rf2	rough	female	3.8	1.317	6
bif2	bifurcated	female	4.0	1.826	4
bif4	bifurcated	female	4.0	1.700	4
bim3	bifurcated	male	4.3	1.252	5
sbrf1	strained-breathy	female	4.5	1.581	6
srf	strained-rough	female	4.6	1.430	6
rbrm5	rough-breathy	male	5.4	1.174	6
srm	strained-rough	male	5.5	1.354	6
rbrf1	rough-breathy	female	5.8	1.874	6
rbrf2	rough-breathy	female	5.9	1.197	6
sbrf2	strained-breathy	female	6.3	1.252	6

Mean rating, across listeners, for each voice is shown along with the standard deviation, gender of the speaker, and type and severity of the pathology. Listeners judged the similarity between natural and synthetic voices on a scale from one to seven, one implies that the natural and synthetic voice qualities sounded identical, while seven indicates that the voices were not similar.

voices and speculate as to why we were unable to model some tokens adequately.

3.3.1. Rough and rough-breathy voices

Perceived roughness has been traditionally associated with amplitude and/or pitch perturbation (Heiberger and Horii, 1982), and by some degree of additive noise (Hillenbrand, 1987). Breathy quality is traditionally associated with several cues, including aspiration noise and increases in the amplitude of the first harmonic relative to the second (Bickley, 1982; Klatt and Klatt, 1990). These acoustic cues are often correlated, and breathiness and roughness are perceptually-related multidimensional constructs (Kreiman et al., 1994).

Table 2
Synthesis parameters for the rough female voices

Tokens	rf1 (6)	rf2 (6)
Time-varying parameters		
F_0 (Hz)	*G*	*G*
avg	205	185
min/max	155/248	162/204
AV (dB)	*G*	*G*
avg	58	57
min/max	51/64	52/72
Constant parameters		
NF	4	4
GV (dB)	55	53
GH (dB)	55	51
OQ (%)	70	90
AH (dB)	60	73
F_1 (Hz)	900	780
B_1 (Hz)	150	130
F_2 (Hz)	1330	1367
B_2 (Hz)	90	160
F_3 (Hz)	2700	3000
B_3 (Hz)	300	250
F_4 (Hz)	3700	3700
B_4 (Hz)	425	450
FNP (Hz)	210	180
BNP (Hz)	40	30
FNZ (Hz)	210	180
BNZ (Hz)	100	90

The parameters are either time-varying (the parameter varies through the segment) or constant. If a parameter is not specified, it is set to the default of the synthesizer. TV = time-varying, *K* = constant, NA = not activated. The symbol *G* refers to using a Gaussian random variable to model perturbation. The severity rating of the voice is given in parentheses next to the token's name.

Table 3
Synthesis parameters for the rough male voices (see Table 2 for further details)

Tokens	rm1(4)	rm2 (5)
Time-varying parameters		
AV (dB)		*G*
avg	63	65
min/max	61/65	62/68
Constant parameters		
NF	4	4
GV (dB)	56	57
GH (dB)	56	57
F_0 (Hz)	124	88
OQ (%)	90	80
SQ (%)	250	200
DI (%)	20	0
AH (dB)	73	64
F_1 (Hz)	700	740
B_1 (Hz)	60	130
F_2 (Hz)	1115	1230
B_2 (Hz)	90	120
F_3 (Hz)	2750	2695
B_3 (Hz)	200	175
F_4 (Hz)	3600	3500
B_4 (Hz)	200	350
FNP (Hz)	124	290
BNP (Hz)	35	30
FNZ (Hz)	124	290
BNZ (Hz)	80	100
FTP (Hz)	372	1900
BTP (Hz)	100	200
FTZ (Hz)	372	1900
BTZ (Hz)	280	100

A total of eleven rough and rough-breathy voices (2 rough female, 2 rough male, 2 rough-breathy female and 5 rough-breathy male voices) were analyzed. Parameters used to synthesize these voices are listed in Tables 2–5. Capturing the variation in F_0 proved critical for successful synthesis of these voices. As shown in Tables 2–5, ten of the eleven voices required some form of F_0 variation; this was achieved by either modeling the F_0 variations with a Gaussian distribution (seven voices), using the flutter parameter (FL) (one voice), the diplophonia parameter (DI) (one voice), or by hand-copying the F_0 contour of the natural waveform (one voice).

Eight of the eleven voices (all four rough voices, and four of the seven rough-breathy voices) were synthesized with a time-varying AV that emulated amplitude modulation. All voices were synthesized with some degree of aspiration noise. In fact, only two voices (rbrm1 and rm2) were synthesized with a greater amplitude of voicing than aspiration (AV + GV > AH + GH), while all the other voices had a greater amplitude of aspiration noise than voicing. To match the increased amplitude of the first harmonic observed in the natural spectra, nine of the synthetic voices required an OQ > 50%. With the exception of one voice (rbrm1), all voices were synthesized with pole-zero pairs placed below F_1 to boost or attenuate the amplitude of the harmonics in that frequency region. When matching the amplitudes of the formant frequencies, formant bandwidths had to be, in general, wider for female voices than they were for male voices.

Table 4
Synthesis parameters for the rough-breathy female voices (see Table 2 for further details)

Tokens	rbrf1 (6)	rbrf2 (6)
Time-varying parameters		
F_0 (Hz)	*G*	*G*
avg	160	200
min/max	75/211	175/220
AV (dB)		
avg	45	$K = 60$
min/max	42/48	
Constant parameters		
NF	5	4
GV (dB)	63	55
GH (dB)	60	60
AV (dB)	TV	60
OQ (%)	50	90
SQ (%)	160	200
AH (dB)	50	60
F_1 (Hz)	735	850
B_1 (Hz)	650	200
F_2 (Hz)	1100	1240
B_2 (Hz)	250	200
F_3 (Hz)	2400	3470
B_3 (Hz)	550	300
F_4 (Hz)	3300	3860
B_4 (Hz)	500	400
F_5 (Hz)	4000	NA
B_5 (Hz)	600	NA
FNP (Hz)	150	240
BNP (Hz)	250	40
FNZ (Hz)	150	240
BNZ (Hz)	120	120
FTP (Hz)	480	NA
BTP (Hz)	300	NA
FTZ (Hz)	480	NA
BTZ (Hz)	100	NA

Perceptual ratings for voices in this category ranged from 1.7 (very close match) to 5.9 (poor match). Seven synthesized voices were considered good matches to the natural voices (ratings of 3.5 or better), while four voices were rated above 3.5. Spectrograms of the natural and synthetic tokens for a rough male voice (rm2) that received the best rating in the rough/rough-breathy category are shown in Fig. 3(a). Fig. 3(b) shows short-time DFT spectra of the natural stimulus superimposed on those of the synthetic copy at two places in the time waveform. As can be seen in Fig. 3(a) and (b), the spectral characteristics of the synthetic token matched those of the natural one well. Using the tracheal and nasal pole-zero pairs (Table 3) was critical in achieving a good copy of this voice. Fig. 3(c) shows a portion of the time waveform of the natural token. The shimmer observed in this voice was mimicked by modeling the AV parameter as a Gaussian random variable.

Fig. 4(a) shows spectrograms of the natural and synthetic tokens of an unsuccessfully modeled rough-breathy female voice (rbrf1), which received a poor rating of 5.8. The spectral match between the natural and synthetic copy was not very good in the mid- and high-frequency regions. This can be seen in Fig. 4(b) where DFT spectra of the natural and synthetic copy at 2 different time intervals are superimposed. The natural voice had significant period-to-period fluctuations, as can be seen in a portion of the voice's time waveform shown in Fig. 4(c). We attempted to mimic these fluctuations by modeling F_0 as a Gaussian random variable and hand-copying AV (Table 4). The slight mismatch in the spectral domain and the random and large temporal variations in the natural voice may have resulted in the low rating of the synthesized copy.

3.3.2. *Bifurcated phonation*

Bifurcated voices (also labeled "diplophonic" (Klatt and Klatt, 1990), "bicyclic" (Gerratt et al., 1988), or "dicrotic dysphonia" (Moore and Von Leden, 1958)) are characterized by a pattern of cycles that alternate in fundamental period, amplitude, or both, in a large-small-large-small (AbAb) pattern. Eight bifurcated voices (four female and four male) were analyzed. None showed a perfect pattern of periods alternating in an AbAb fashion. Instead, three patterns emerged: (1) four voices had fundamental frequencies varying randomly between 3 to 9 different F_0 values; (2) three voices had F_0 bimodally distributed; and (3) one voice was increasingly bifurcated with time (see Kreiman et al., 1993, for a detailed discussion of the acoustics and perception of such voices).

Synthesis parameters for the eight bifurcated voices are given in Tables 6 and 7. As this table shows, three methods of modeling F_0 were used. F_0 contours were carefully matched to those of the natural waveforms for three of the eight voices (bif3, bif4, bim4). DI and FL were used with a constant F_0 parameter for three voices (bim1, bim2, bim3), while time-varying F_0 was combined with DI for the remaining two voices (bif1, bif2). One voice (bim2) demonstrated considerable shimmer, so a time-varying AV was used. In contrast to rough and rough-breathy voices, male and female bifurcated voices differed in OQ values. The male voices sounded more strained and had weaker first harmonics than their female counterparts. Hence, OQ was less than 50% for all the male voices (but for only one female voice).

Seven of the eight voices (four female and three male) required a low-frequency energy boost using the nasal and/or tracheal pole-zero pairs, and in one case (bif4), the tracheal pole-zero pair was used to

Table 5
Synthesis parameters for the rough-breathy male voices (see Table 2 for further details)

Tokens	rbrm1 (3)	rbrm2 (5)	rbrm3 (5)	rbrm4 (5)	rbrm5 (6)
Time-varying parameters					
F_0 (Hz)		*G*		*G*	*G*
avg	119	143	$K = 117$	71	210
min/max	112/125	105/215		65/78	198/220
AV (dB)				*G*	*G*
avg	$K = 58$	52	$K = 60$	66	60
min/max		30/57		55/72	58/62
Constant parameters					
NF	4	4	4	4	4
GV (dB)	61	64	61	53	57
GH (dB)	60	64	60	56	60
F_0 (Hz)	TV	TV	117	TV	TV
AV (dB)	58	TV	60	TV	TV
OQ (%)	60	80	90	50	70
SQ (%)	200	150	200	200	200
FL (%)	0	0	5	0	0
AH (dB)	57	59	70	70	57
F_1 (Hz)	680	870	815	595	640
B_1 (Hz)	80	125	180	80	70
F_2 (Hz)	1280	1110	1260	1140	1200
B_2 (Hz)	140	125	200	90	300
F_3 (Hz)	2425	2700	2655	2350	2290
B_3 (Hz)	170	150	180	150	300
F_4 (Hz)	3760	3830	2470	3650	3380
B_4 (Hz)	200	250	300	250	300
FNP (Hz)	NA	NA	116	290	200
BNP (Hz)	NA	NA	40	40	50
FNZ (Hz)	NA	NA	116	290	200
BNZ (Hz)	NA	NA	180	120	150
FTP (Hz)	NA	560	NA	NA	NA
BTP (Hz)	NA	100	NA	NA	NA
FTZ (Hz)	NA	560	NA	NA	NA
BTZ (Hz)	NA	280	NA	NA	NA

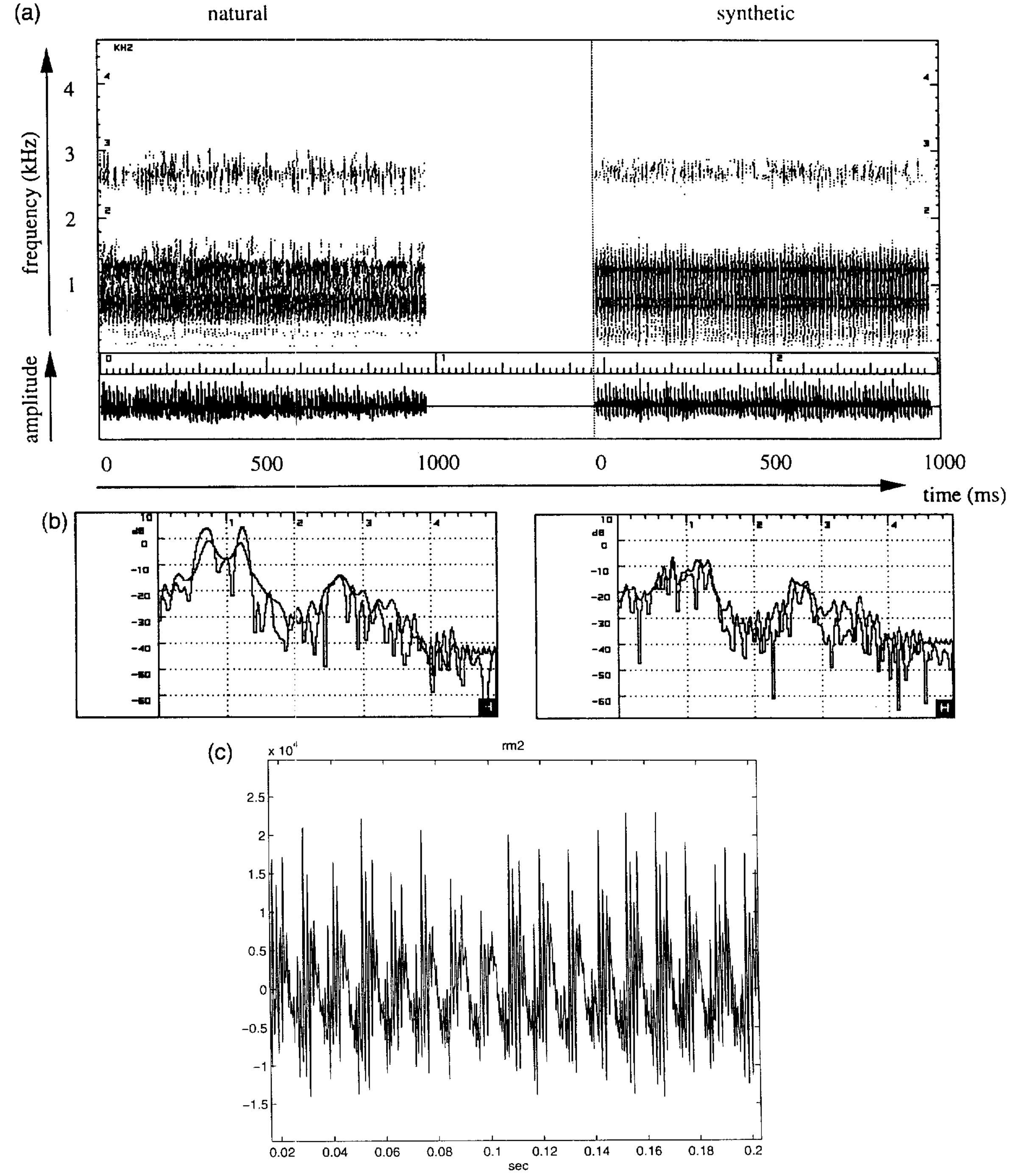


Fig. 3. (a) Spectrograms and time waveforms of the natural and synthetic tokens of a rough male voice (rm2). (b) DFT spectra (calculated using a 12.8 ms Hamming window) of the natural token superimposed with those of the synthetic token at 2 different time intervals. (c) A portion of the time waveform of the natural voice.

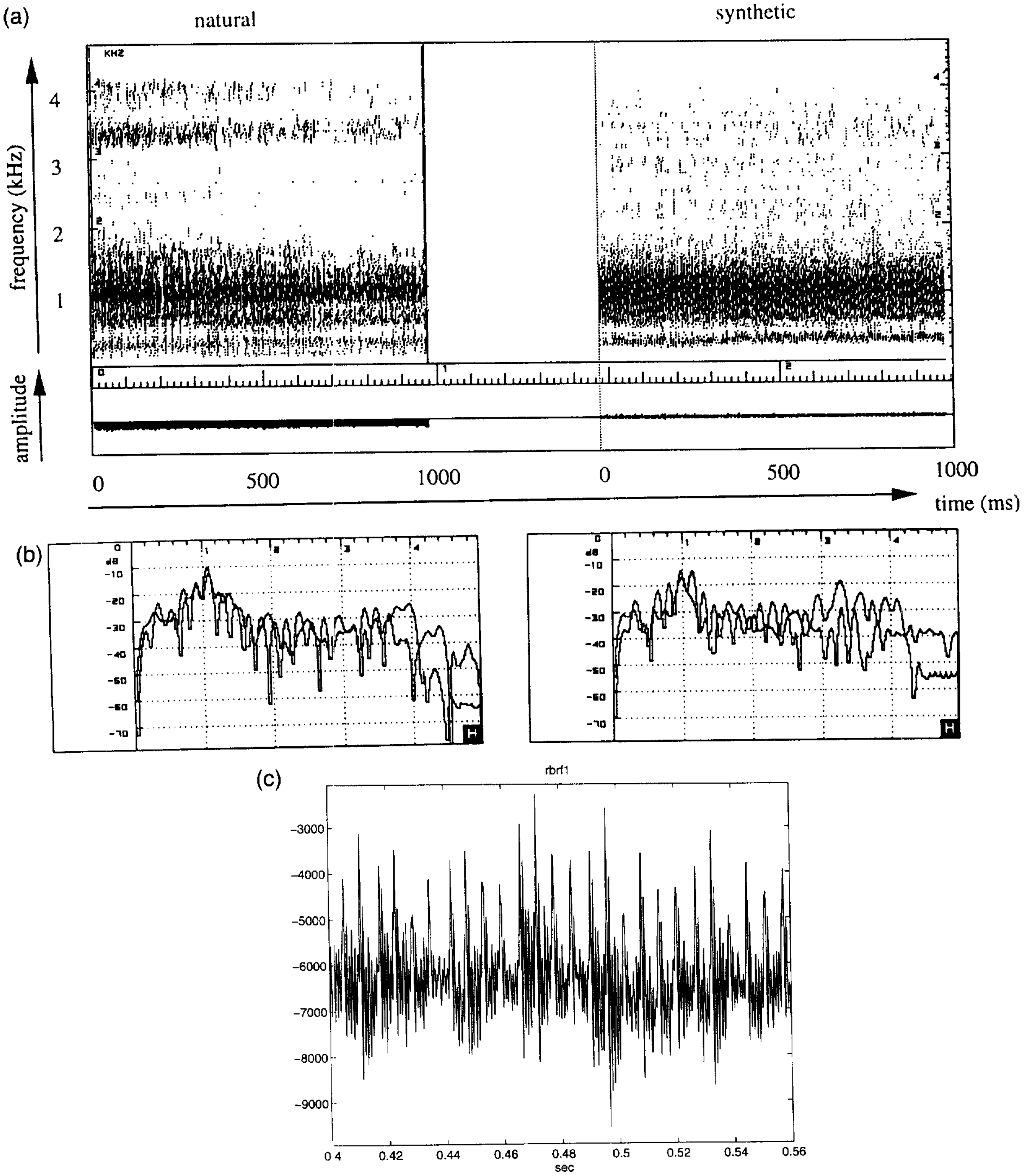


Fig. 4. (a) Spectrograms and time waveforms of the natural and synthetic tokens of a rough female voice (rbrf1). (b) DFT spectra of the natural token superimposed with those of the synthetic token at 2 different times intervals. (c) A portion of the time waveform of the natural voice.

weaken the second harmonic. Only one voice (bif2) sounded breathy, and it was modeled with 5 dB more aspiration than voicing. All other voices had more voicing than aspiration.

Perceptual ratings ranged between 1.3 to 4.3; five synthesized voices were considered good matches to the natural voices (ratings of 3.5 or better), while three voices were rated above 3.5. Fig. 5(a) shows spectrograms of natural and synthetic tokens for a bifurcated male voice (bim2) which was the most successfully-synthesized in the bifurcated category. The spectral match between the natural and synthetic voice was very good, as can be seen in the DFT spectra of the natural and synthetic voices shown in Fig. 5(b). This voice showed period-to-period fluctuations which were mimicked well by hand-copying the amplitude of the waveform and by setting the DI parameter to 8% and FL parameter to 5%. F_0 bimodality in the natural and synthetic copies is shown in Fig. 5(c).

Fig. 6(a) shows the spectrograms of the natural and synthetic copies of a bifurcated female voice (bif2) which received a lower rating of 4.0. The fundamental frequency of the natural voice was measured manually, and a plot of F_0 for the first 500 ms is shown in Fig. 6(b). F_0 varied randomly between 5 values in the range 230–250 Hz and was difficult to mimic properly. In addition, the spectral match in the

Table 6
Synthesis parameters for the bifurcated female voices (see Table 2 for further details)

Tokens	bif1 (4)	bif2 (4)	bif3 (5)	bif4 (4)
Time-varying parameters				
F_0 (Hz)				
avg	227	245	182	209
min/max	220/232	237/257	177/187	196/222
DI (%)				
avg	10	$K = 7$	$K = 0$	$K = 0$
min/max	7/14			
Constant parameters				
NF	4	4	4	4
GV (dB)	57	59	62	59
GH (dB)	57	60	60	58
AV (dB)	62	63	62	65
OQ (%)	70	77	60	40
SQ (%)	200	140	200	150
DI (%)	TV	7	0	0
AH (dB)	60	67	55	60
F_1 (Hz)	930	790	645	990
B_1 (Hz)	200	180	60	200
F_2 (Hz)	1450	1220	1620	1410
B_2 (Hz)	150	90	100	130
F_3 (Hz)	2800	2270	3240	3795
B_3 (Hz)	250	130	250	350
F_4 (Hz)	3740	3150	4200	4245
B_4 (Hz)	150	150	350	450
FNP (Hz)	230	225	194	205
BNP (Hz)	50	40	30	60
FNZ (Hz)	230	225	194	205
BNZ (Hz)	150	100	100	400
FTP (Hz)	460	450	2050	414
BTP (Hz)	50	100	60	150
FTZ (Hz)	460	450	2050	414
BTZ (Hz)	120	200	350	80

mid- and high-frequency region was not perfect as can be seen in the DFT spectra of the natural and synthetic voices shown in Fig. 6(c).

3.3.3. Other voices: rough-bifurcated, strained-rough and strained-breathy qualities

A number of voices did not fall into traditional perceptually- or acoustically-based categories. Synthesizing these perceptually-complex stimuli presented particular challenges.

One male voice (rbim) was severely rough, breathy, and intermittently bifurcated. The fundamental frequency fluctuated between about 100–200 Hz and aspiration noise was present in the F_3 and F_4 regions. The synthesis parameters for this voice, which received a reasonable rating of 2.8, are given in Table 8, and spectrograms of the natural and synthetic tokens are shown in Fig. 7(a). The synthesis involved matching the overall F_0 contour to that of the natural token and using the DI parameter, much like a severe bifurcated voice. Unlike the bifurcated voices, however, the rough-bifurcated voice was synthesized with 10–14 dB more aspiration than voicing. The spectral details of this voice

Table 7
Synthesis parameters for the bifurcated male voices (see Table 2 for further details)

Tokens	bim1 (5)	bim2 (4)	bim3 (5)	bim4 (5)
Time-varying parameters				
F_0 (Hz)				
avg	$K = 164$	$K = 177$	$K = 140$	147
min/max				146/152
AV (dB)				
avg	$K = 60$	59	$K = 63$	$K = 63$
min/max		57/60		
Constant parameters				
NF	5	5	4	4
GV (dB)	57	62	58	60
GH (dB)	56	60	58	60
F_0 (Hz)	164	177	140	TV
AV (dB)	60	TV	63	63
OQ (%)	45	27	30	30
SQ (%)	150	200	200	200
FL (%)	10	5	5	0
DI (%)	23	8	10	0
AH (dB)	60	50	60	55
F_1 (Hz)	755	580	700	790
B_1 (Hz)	60	120	60	150
F_2 (Hz)	1095	1120	1060	1240
B_2 (Hz)	90	30	140	100
F_3 (Hz)	2370	2700	2480	2920
B_3 (Hz)	300	250	90	150
F_4 (Hz)	2980	3330	3000	760
B_4 (Hz)	200	400	90	150
F_5 (Hz)	3780	3720	NA	NA
B_5 (Hz)	200	500	NA	NA
FNP (Hz)	170	170	NA	280
BNP (Hz)	30	45	NA	60
FNZ (Hz)	170	170	NA	280
BNZ (Hz)	100	90	NA	280
FTP (Hz)	565	340	NA	3760
BTP (Hz)	40	95	NA	60
FTZ (Hz)	565	340	NA	3760
BTZ (Hz)	100	180	NA	300

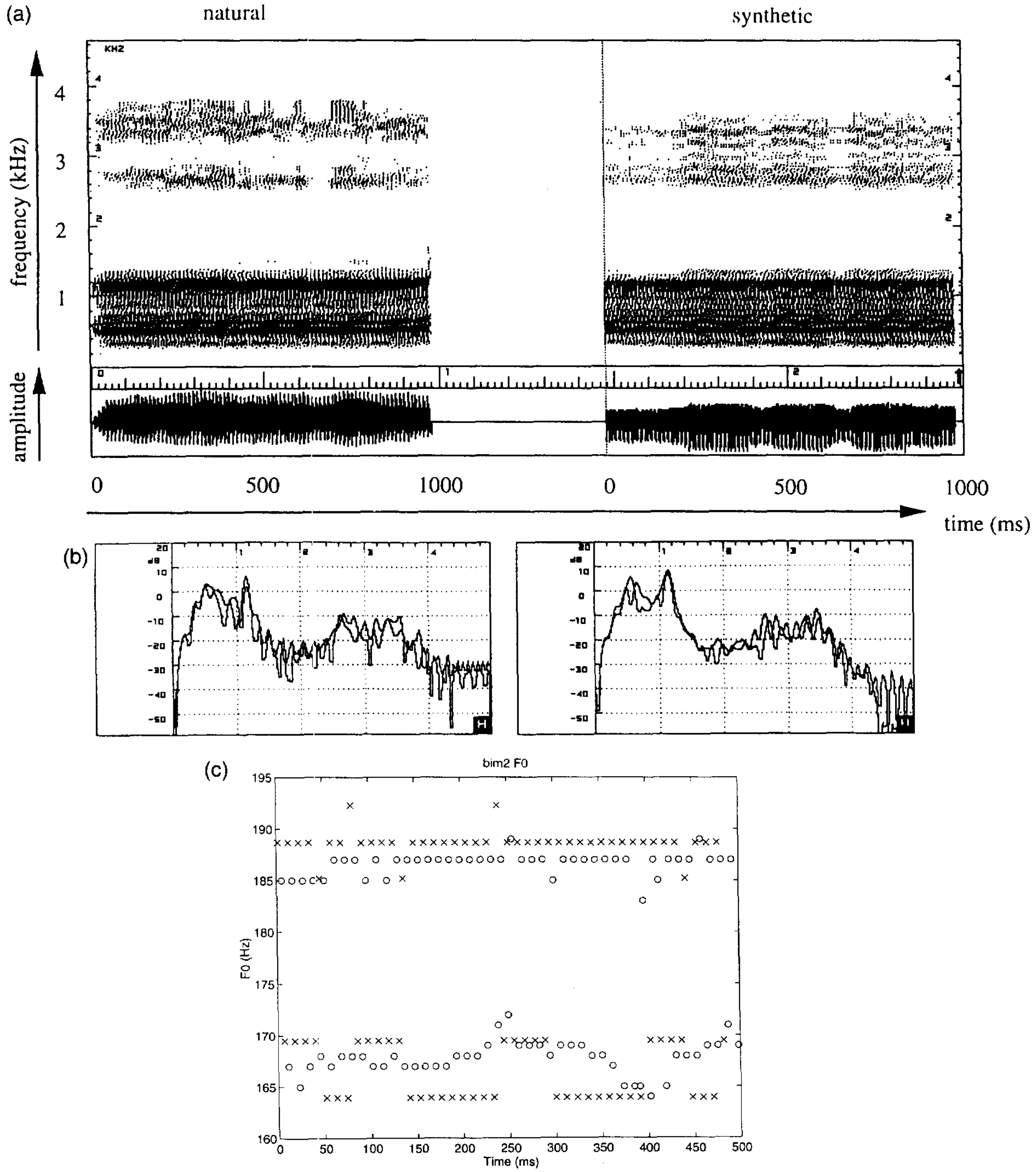


Fig. 5. (a) Spectrograms and time waveforms of the natural and synthetic tokens of a bifurcated male voice (bim2). (b) DFT spectra of the natural token superimposed with those of the synthetic token at 2 different time intervals. (c) Plot of F_0 versus time for the natural (denoted by circles) and synthetic (denoted by the x symbol) tokens.

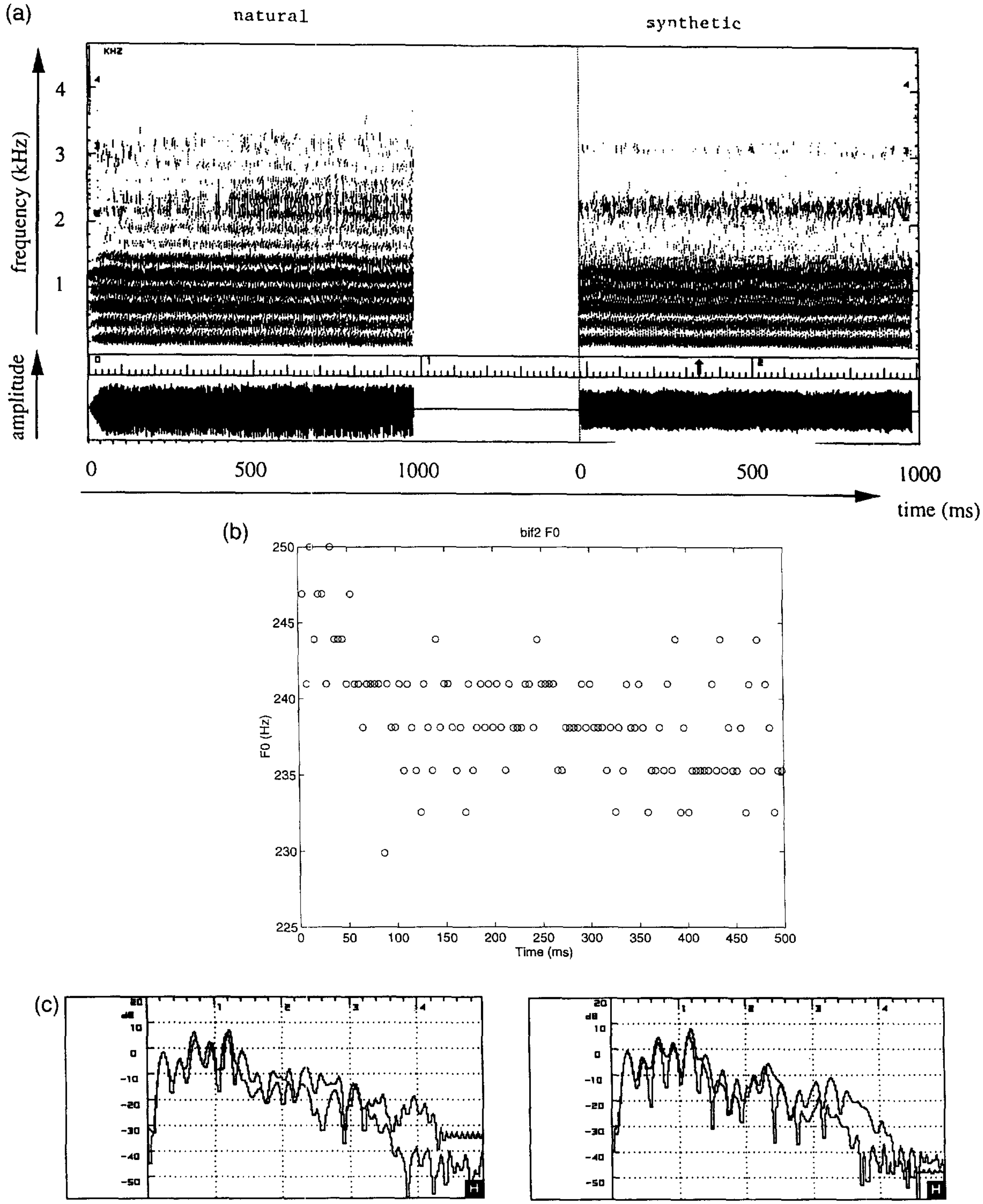


Fig. 6. (a) Spectrograms and time waveforms of the natural and synthetic tokens of a bifurcated female voice (bif2). (b) Plot of F_0 versus time for the natural token. (c) DFT spectra of the natural token superimposed with those of the synthetic token at 2 different time intervals.

varied approximately every 100 ms. Fig. 7(b) shows DFT spectra of the natural voice superimposed with the spectra of the synthetic copy; the spectral match between the natural and synthetic token was good.

Two strained-breathy female voices (sbrf1, sbrf2) were analyzed. Both voices began with strong voicing, and then the voicing amplitude decreased in the last 400 ms of the sample. The synthesis parameters for these voices are listed in Table 9, and spectrograms of the natural and synthetic utterances for one token (sbrf1) are shown in Fig. 8(a).

These voices were very difficult to synthesize and received poor ratings of 4.5 and 6.3. Techniques used to model the time-varying nature of these voices included sequentially altering OQ to capture the breathy quality at one time (by increasing OQ) and the strained quality at another time (by decreasing OQ); time-varying AH and AV to enhance breathiness perception when appropriate; and utilizing FL and time-varying F_0. The difficulty in synthesizing these voices was matching the widely-varying spectral and temporal details of the voice. Fig. 8(b) shows DFT spectra at three points in the natural waveform; note the significant variation in the spectra. Fig. 8(c) shows DFT spectra of the natural and synthetic copy; note the spectral mismatch at high frequencies. Fig. 8(d) shows a portion of the time waveform of this voice which demonstrated significant period-to-period variability.

Finally, two strained-rough voices (one female, one male) were analyzed and synthesized. Both voices (srf, srm) were described as "gargly" because of their unsteadiness when the voices became strained, and both synthetic versions received poor

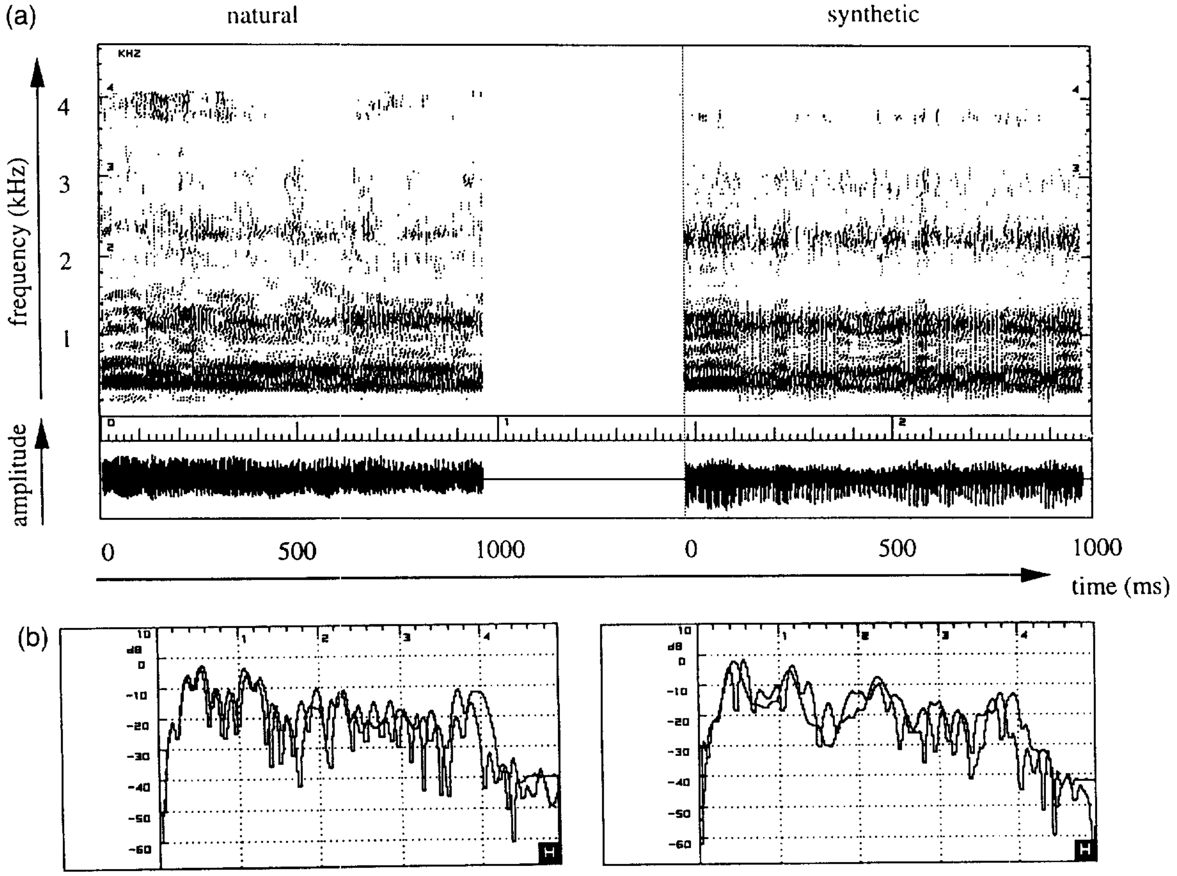


Fig. 7. (a) Spectrograms and time waveforms of the natural and synthetic tokens of a rough-bifurcated male voice (rbim). (b) DFT spectra of the natural voice superimposed with those of the synthetic voice at 2 times in the waveform.

Table 8
Synthesis parameters for the rough-bifurcated male voice (see Table 2 for further details)

Token	rbim (5)
Time-varying parameters	
F_0 (Hz)	
avg	152
min/max	80/210
AV (dB)	
avg	61
min/max	58/62
Constant parameters	
NF	5
GV (dB)	59
GH (dB)	62
OQ (%)	45
SQ (%)	400
DI (%)	2
AH (dB)	69
F_1 (Hz)	450
B_1 (Hz)	60
F_2 (Hz)	1150
B_2 (Hz)	100
F_3 (Hz)	2240
B_3 (Hz)	150
F_4 (Hz)	3000
B_4 (Hz)	350
F_5 (Hz)	3800
B_5 (Hz)	150
FTP (Hz)	1860
BTP (Hz)	140
FTZ (Hz)	1860
BTZ (Hz)	200

ratings of 4.6 and 5.5. The synthesis parameters for these voices are listed in Tables 10 and 11, and spectrograms of the natural and synthetic female token (srf) are shown in Fig. 9(a). The "gargly" period consisted mainly of the last 400 ms of this voice, and several parameters (F_0, AV, SQ and FL) were time-varied to capture the unsteadiness and alternating strained/rough percept in the natural token. Our attempts, however, were not very successful in capturing these variations. In particular, it was not possible to significantly attenuate the high-frequency energy in the synthetic copy especially since both pole-zero pairs (tracheal and nasal) were used to match the spectra below 3 kHz. Fig. 9(b) shows DFT spectra of the natural and synthetic copies at two time intervals; in one interval the spectral match was good but in another, there was a clear mismatch at high frequencies.

3.3.4. Why exact matches of some voices could not be achieved

In general, voices with significant amplitude and/or frequency perturbations were difficult to synthesize with the Klatt synthesizer. The flutter parameter (FL), which alters F_0 in a slow time-varying fashion, is available, but does not model jitter appro-

Table 9
Synthesis parameters for the strained-breathy female voices (see Table 2 for further details)

Tokens	sbrf1 (6)	sbrf2 (6)
Time-varying parameters		
F_0 (Hz)		
avg	216	211
min/max	204/250	202/222
AV (dB)		
avg	53	61
min/max	41/57	58/65
OQ (%)		
avg	60	17
min/max	45/70	10/25
FL (%)		
avg	$K = 20$	13
min/max		10/20
AH (dB)		
avg	77	61
min/max	76/80	58/63
Constant parameters		
NF	4	4
GV (dB)	58	57
GH (dB)	56	73
F_0 (Hz)	TV	202
SQ (%)	180	200
FL (%)	20	TV
DI (%)	6	0
F_1 (Hz)	750	820
B_1 (Hz)	230	400
F_2 (Hz)	1020	1450
B_2 (Hz)	230	200
F_3 (Hz)	2675	2950
B_3 (Hz)	200	300
F_4 (Hz)	3990	3850
B_4 (Hz)	250	600
FNP (Hz)	NA	606
BNP (Hz)	NA	200
FNZ (Hz)	NA	606
BNZ (Hz)	NA	60

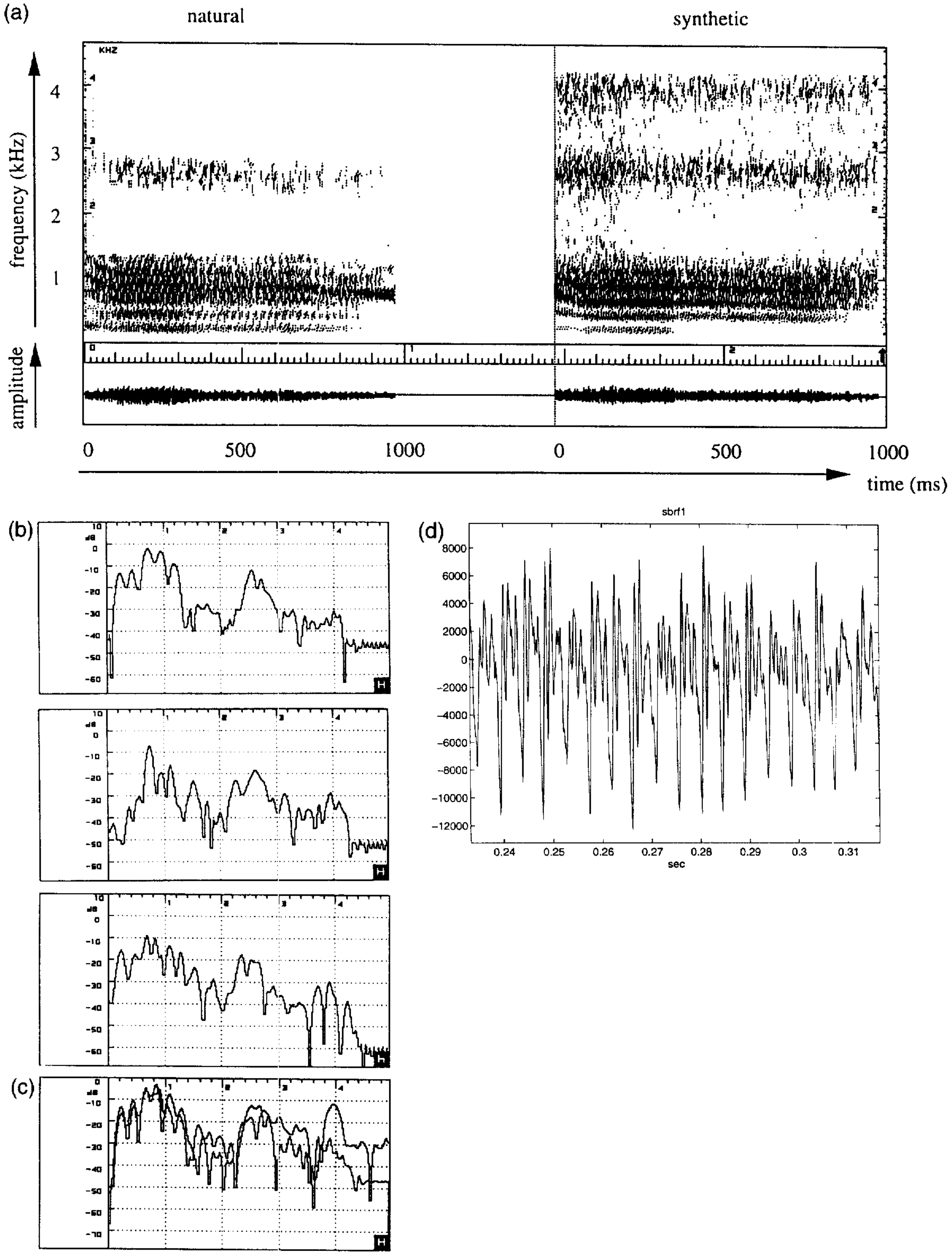

(a)
natural
synthetic
frequency (kHz)
4
3
2
1
amplitude
KHZ
0
500
1000
0
500
1000
time (ms)
(b)
dB
10
0
-10
-20
-30
-40
-50
-60
1
2
3
4
H
(c)
-70
(d)
sbrf1
8000
6000
4000
2000
0
-2000
-4000
-6000
-8000
-10000
-12000
0.24
0.25
0.26
0.27
0.28
0.29
0.3
0.31
sec

Table 10
Synthesis parameters for the strained rough female voice (see Table 2 for further details)

Token	srf (6)
Time-varying parameters	
F_0 (Hz)	
avg	200
min/max	195/201
AV (dB)	
avg	61
min/max	58/63
SQ (%)	
avg	370
min/max	300/500
FL (%)	
avg	20
min/max	15/30
Constant parameters	
NF	5
GV (dB)	55
GH (dB)	60
OQ (%)	35
AH (dB)	52
F_1 (Hz)	790
B_1 (Hz)	100
F_2 (Hz)	1200
B_2 (Hz)	130
F_3 (Hz)	2590
B_3 (Hz)	400
F_4 (Hz)	3800
B_4 (Hz)	375
F_5 (Hz)	4200
B_5 (Hz)	375
FNP (Hz)	402
BNP (Hz)	60
FNZ (Hz)	402
BNZ (Hz)	180
FTP (Hz)	3000
BTP (Hz)	180
FTZ (Hz)	3000
BTZ (Hz)	120

Table 11
Synthesis parameters for the strained rough male voice (see Table 2 for further details)

Token	srm (6)
Time-varying parameters	
F_0 (Hz)	
avg	170
min	160/175
AV (dB)	
avg	56
min	53/58
OQ (%)	
avg	50
min/max	45/60
SQ (%)	
avg 237	
min/max	200/300
Constant parameters	
NF	4
GV (dB)	60
GH (dB)	60
FL (%)	20
AH (dB)	55
F_1 (Hz)	700
B_1 (Hz)	60
F_2 (Hz)	1030
B_2 (Hz)	90
F_3 (Hz)	3100
B_3 (Hz)	100
F_4 (Hz)	3800
B_4 (Hz)	100

priately (Klatt and Klatt, 1990). The current implementation of the diplophonia parameter (DI) is also inadequate for modeling jitter, shimmer or bifurcated voices. Further, DI produces patterns of amplitude and frequency variation that do not match measurements of natural bifurcated waveforms, for which there is no consistent correlation between amplitude and F_0 (Kreiman et al., 1993). Synthesis would be improved by allowing amplitude to be changed independent of delay, and/or by allowing amplitude to be specified for each individual period.

One technique that was successful for modeling F_0 in voices with high jitter and shimmer (especially rough and rough-breathy voices) was the use of a Gaussian distribution (Hillenbrand, 1988). Using this technique was cumbersome; it involved exiting the synthesizer, generating new random numbers, and importing them back into the synthesizer. A synthe-

Fig. 8. (a) Spectrograms and time waveforms of the natural and synthetic tokens of a strained-breathy female voice (sbrf1). (b) DFT spectra at the three different times in the natural token. (c) DFT spectra of the natural voice superimposed with those of the synthetic voice. (d) A portion of the time waveform of the natural voice.

sis parameter that allows F_0 (and possibly AV) to be modeled by a Gaussian random variable with a given mean and variance would greatly facilitate this process.

Most "unacceptable" ratings reflected failures to model unsteady or gargled qualities. Our failure to capture the gargly nature of some voices is due to several factors. First, the update interval (UI) is implemented as a function of time, not period. Some parameters, such as F_0, are updated at the end of each period. However, most time-varying parameters (such as AV) are specified at multiples of UI, and linear interpolation is used to determine time-varying values between update intervals. This poses a problem when one needs to change attributes of one glottal pulse without affecting other pulses. For example, it would be much easier to mimic spikes that occur in the time waveforms of some pathological voices, if it were possible to specify AV at a single period rather than at an update interval. Spikes can usually be modeled adequately if UI = 1 ms, but then we can only synthesize 400 ms due to a limitation imposed by the synthesizer.

It was often difficult to match the low-frequency energy of the natural samples (especially energy below the first formant). A partial solution is to adjust the open quotient (OQ), which primarily affects the amplitude of the first harmonic. As pointed out earlier, this is often inadequate. Additional pole-zero pairs (nasal and/or tracheal) can be placed at particular frequency regions. This solution works well as long as the nasal and tracheal pole-zero pairs

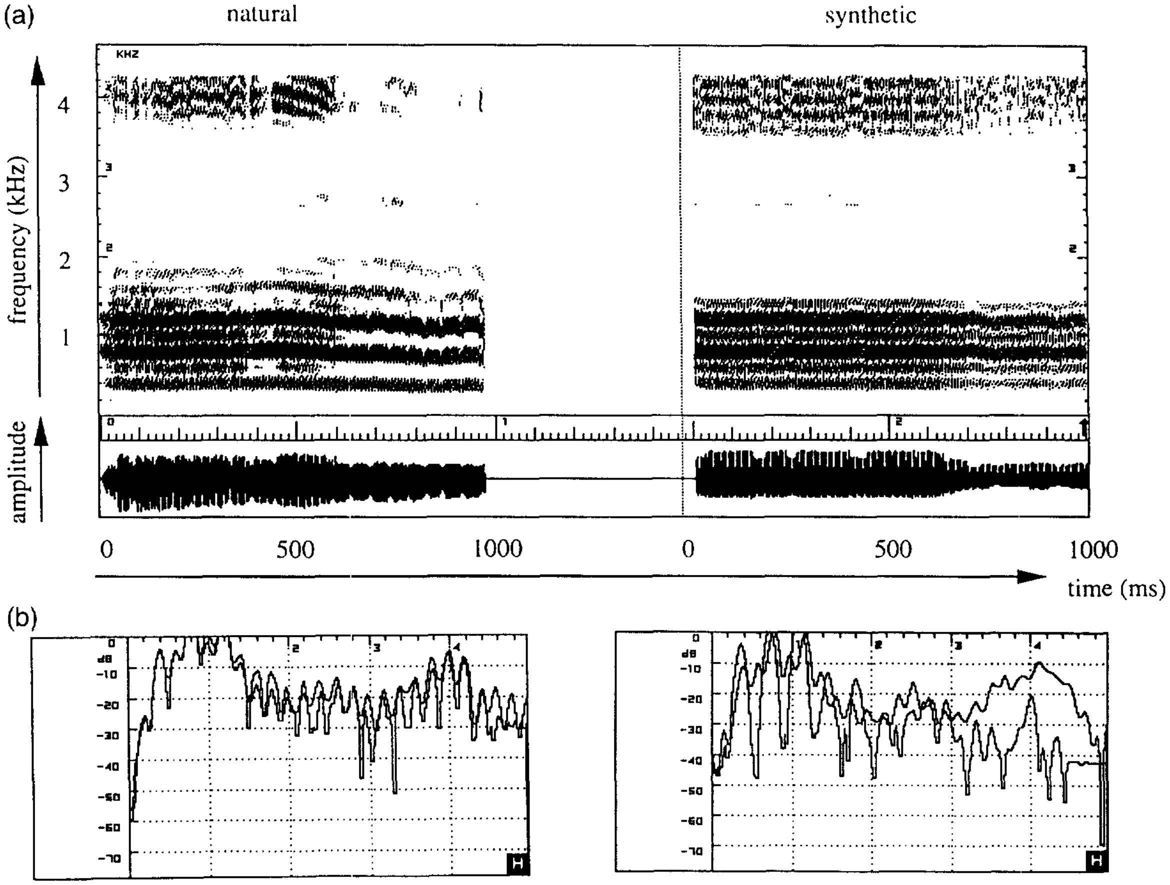


Fig. 9. (a) Spectrograms and time waveforms of the natural and synthetic tokens of a strained-rough female voice (srf). (b) DFT spectra of the natural voice superimposed with those of the synthetic voice at 2 different time intervals in the waveform.

are not used elsewhere to model source-tract interactions. Providing more pole-zero pairs, which can be placed at any frequency, would not only alleviate this problem but would also aid in creating a better spectral match between synthetic and natural voices at all frequencies. Alternatively, a new parameter that adjusts the amplitudes of individual harmonics (especially those below F_1) would help.

4. Summary and discussion

This paper describes a pilot study into the mechanics of synthesizing moderately-to-severely pathological voices. Successful synthesis of such voices may ultimately provide a method for evaluating and documenting voice qualities. An analysis-by-synthesis approach using a Klatt formant synthesizer was applied to study 24 tokens of the vowel /a/ spoken by males and females with voice disorders. Voice qualities included rough and rough-breathy; bifurcated; rough-bifurcated; strained-rough; and strained-breathy. Both temporal and spectral features of the natural waveforms were analyzed and the results were used to guide synthesis.

Ten expert listeners found about half the synthetic voices well-matched to the natural waveforms they modeled. Synthesis parameters common to all rough and rough-breathy voices included a time-varying fundamental frequency (F_0) (achieved mainly by modeling F_0 variations using a Gaussian random variable); amplitude of aspiration noise that was large relative to that of voicing; and a relatively high low-frequency energy, achieved by setting the open quotient (OQ) > 50% and/or placing a pole-zero pair at low frequencies. Most bifurcated voices required varying F_0, either by hand-copying the natural F_0 contours or by using the diplophonia parameter (DI); an OQ < 50% reflecting a strained quality (especially for the male voices); and an amplitude of voicing greater than that of aspiration noise. Rough-bifurcated quality was synthesized by time-varying F_0, using DI, and using a larger amplitude of aspiration noise than of voicing. Strained-breathy and strained-rough voices were not successfully synthesized.

Our results indicate that some modifications to the Klatt synthesizer are necessary to successfully synthesize pathological voices. Modifications include providing a parameter to increase the low frequency energy below F_1; adding more pole-zero pairs; providing jitter and shimmer parameters; changing the update interval parameter to work in periods rather than in absolute time; and modifying the diplophonia parameter so that fundamental frequency and amplitude variations can be independently controlled. Modifying the DI parameter and increasing the number of formants and the pole-zero pairs are straightforward operations. It is less clear how jitter and

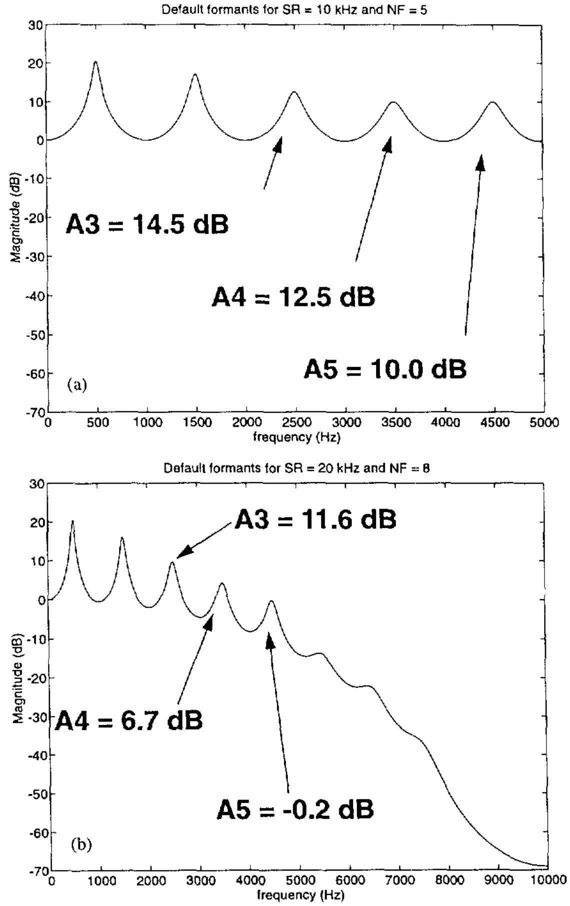


Fig. 10. Linear Predictive Code (LPC) spectra of a synthetic neutral vowel like /uh/ for two different sampling rates. The transfer function is calculated using formants at 500 Hz, 1500 Hz, etc. and the synthesizer's default values for the bandwidths. (a) The vocal tract transfer function for a sampling rate of 10 kHz and using 5 formants. (b) The vocal tract transfer function for a sampling rate of 20 kHz and using 8 formants.

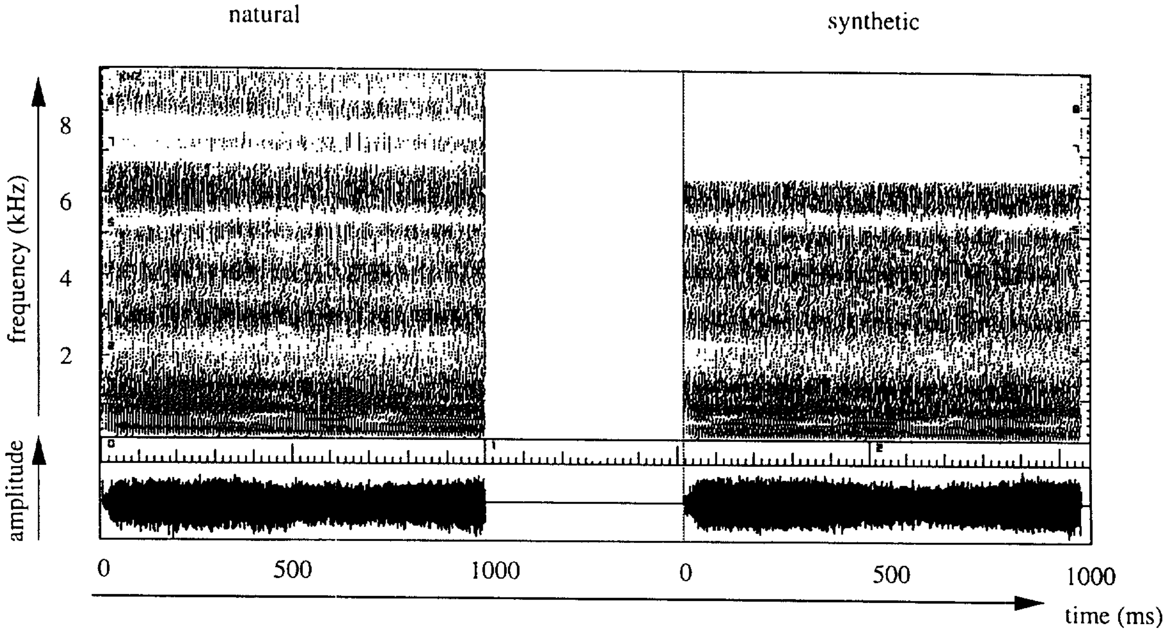


Fig. 11. Consequences of the limited number of formants when applied to synthesizing a breathy female voice sampled at 20 kHz. (a) The natural utterance has energy at high frequencies (up to 9 kHz in this example.) (b) The synthetic utterance has most of the energy at frequencies below 6.5 kHz.

shimmer parameters should be implemented. One possibility is to implement a Gaussian random variable, like the one used in this study.

In this study we were obliged to limit our analysis-by-synthesis to waveforms sampled at 10 kHz because the synthesizer provides only six *variable* formants, limiting the maximum usable sampling rate to 10–12 kHz. Fig. 10 shows the high frequency spectral roll off for a vowel synthesized at sampling rates of 10 kHz and 20 kHz, using the default values for formant frequencies and bandwidths. Fig. 10(a) shows a spectrum synthesized with five formants and a sampling rate of 10 kHz while Fig. 10(b) shows a synthetic vowel with eight formants and a sampling rate of 20 kHz. The resulting spectrum for the higher sampling rate slopes downward starting at the fourth formant, due to the lack of formants near the Nyquist frequency (10 kHz in Fig. 10(b)). Providing more than six formants with variable frequency and bandwidth would alleviate this difficulty. This spectral slope has noticeable effects in the resultant synthetic voice quality. For example, Fig. 11(a) shows a spectrogram of a natural breathy voice sampled at 20 kHz and of the synthetic stimulus generated with formant frequencies and amplitudes measured from the original sample. The synthetic stimulus matches the natural one for frequencies below 5 kHz, but not for higher frequencies. Although these high frequencies may not be important for speech perception or intelligibility, they are important aspects of voice quality.

Finally, more acoustic modeling of severe vocal pathology is necessary. As discussed above, most models are based on variations in normal speech, and do not easily accommodate pathologic cases. Improved models and synthesizers are essential for improved pathological voice quality evaluation and for the creation of well-matched synthesized voices.

Acknowledgements

This research was supported in part by NIDCD grant DC 01797. We thank Dr Sorin and the reviewers for their valuable comments.

Appendix A. Parameters for the cascade branch of the SENSYN synthesizer

Unless otherwise noted, the parameters in Table 12 were set to these default values.

Appendix B. Listener instructions

You are about to hear a series of voice pairs. The first voice in each pair was recorded from a dysphonic patient. The second voice is a copy of the dysphonic voice, made with a speech synthesizer. Some of the copies are much more successful at capturing the quality of the original voice than others are. We would like you to listen to each pair and tell us just how successful each synthesized copy is. You may listen to each pair as many times as you like. Please rate the goodness of the copy on a 7 point scale, where "1" means the copy is identical to the original, and "7" means it does not sound like the same person. You need not use the entire scale; if all

Table 12

Symbol	Default	Description
DU	1000	Duration of the utterance, in ms
SR	10000	Output sampling rate, in samples/sec
NF	5	Number of formants in cascade branch
SS	3	Source switch (1 = impulse, 2 = natural, 3 = LF model)
GV	60	Overall gain scale factor for AV, in dB
GH	60	Overall gain scale factor for AH, in dB
F_0	100	Fundamental frequency, in Hz
AV	60	Amplitude of voicing, in dB
OQ	50	Open quotient (voicing open-time/period), in %
SQ	200	Speed quotient (rise/fall time of open period, LF model only), in %
TL	0	Extra tilt of voicing spectrum, dB down at 3 kHz
FL	0	Flutter (random fluctuation in F_0), in % of maximum
DI	0	Diplophonia (pairs of periods migrate together), in % of maximum
AH	0	Amplitude of aspiration, in dB
F_1	500	Frequency of the 1st formant, in Hz
B_1	60	Bandwidth of the 1st formant, in Hz
F_2	1500	Frequency of the 2nd formant, in Hz
B_2	90	Bandwidth of the 2nd formant, in Hz
F_3	2500	Frequency of the 3rd formant, in Hz
B_3	150	Bandwidth of the 3rd formant, in Hz
F_4	3250	Frequency of the 4th formant, in Hz
B_4	200	Bandwidth of the 4th formant, in Hz
F_5	3700	Frequency of the 5th formant, in Hz
B_5	200	Bandwidth of the 5th formant, in Hz
F_6	4990	Frequency of the 6th formant, in Hz
B_6	500	Bandwidth of the 6th formant, in Hz
F_7	6500	Frequency of the 7th formant, in Hz (not modifiable)
B_7	500	Bandwidth of the 7th formant, in Hz (not modifiable)
F_8	7500	Frequency of the 8th formant, in Hz (not modifiable)
B_8	600	Bandwidth of the 8th formant, in Hz (not modifiable)
FNP	280	Frequency of nasal pole, in Hz
BNP	90	Bandwidth of nasal pole, in Hz
FNZ	280	Frequency of nasal zero, in Hz
BNZ	90	Bandwidth of nasal zero, in Hz
FTP	2150	Frequency of tracheal pole, in Hz
BTP	180	Bandwidth of tracheal pole, in Hz
FTZ	2150	Frequency of tracheal zero, in Hz
BTZ	180	Bandwidth of tracheal zero, in Hz

the copies are rated 1 or 7, that is fine. Please try to judge each pair independently of the others. You will hear the entire set of voice pairs before the study starts to give you an idea of how much they vary.

When judging the pairs, try to focus on the overall quality. Please ignore differences in the loudness of the stimuli as much as possible. Thank you for participating in this study.

References

Ananthapadmanabha, T.V., 1984. Acoustic analysis of voice source dynamics. STL-QPSR 2-3, 1–24.

Bickley, C., 1982. Acoustic analysis and perception of breathy vowels. Speech Communication Group Working Papers I, MIT, pp. 71–80.

Carlson, R., Granström, B., Karlsson, I., 1991. Experiments with voice modeling in speech synthesis. Speech Communication 10, 481–489.

Childers, D.G., Lee, C.K., 1991. Vocal quality factors: analysis, synthesis and perception,. J. Acous. Soc. Amer. 90, 2394–2410.

Childers, D.G., Ahn, C., 1995. Modeling the glottal volume-velocity waveform for three voice types. J. Acoust. Soc. Amer. 97, 505–519.

Fant, G., Liljencrants, J., Lin, Q.G., 1985. A four parameter model of glottal flow. STL-QPSR 4, 1–13.

Fujisaki, H., Ljungqvist, M., 1986. Proposal and evaluation of models for the glottal source waveform. In: Proc IEEE Internat. Conf. Acoust. Speech Signal Process., pp. 1605–1608.

Gerratt, B.R., Precoda, K., Hanson, D.G., Berke, G.S., 1988. Source characteristics of diplophonia. Paper presented at the 115th Meeting of the Acoustical Society of America, Seattle.

Gerratt, B.R., Kreiman, J., Antonanzas-Barroso, N., Berke, G.S., 1993. Comparing internal and external standards in voice quality judgments. J. Speech Hear. Res. 36, 14–20.

Gobl, C., 1988. Voice source dynamics in connected speech. Speech Transmission Laboratory Quarterly Status and Progress Report 1, 123–159.

Gobl, C., Ní Chasaide, A., 1992. Acoustic characteristics of voice quality. Speech Communication 11, 481–490.

Heiberger, V. L., Horii, Y., 1982. Jitter and shimmer in sustained phonation. In: Lass, N. (Ed.), Speech and Language: Advances in Basic Research and Practice. Academic Press, New York, pp. 299–322.

Hillenbrand, J., 1987. A methodological study of perturbation and additive noise in synthetically generated voice signals. J. Speech Hear. Res. 30, 448–461.

Hillenbrand, J., 1988. Perception of aperiodicities in synthetically generated voices. J. Acoust. Soc. Amer. 83, 2361–2371.

Imaizumi, S., Kiritani, S., Saito, S., 1991. Perceptual evaluation of a glottal source model for voice quality control. In: Gauffin, J., Hammarberg, B. (Eds.), Vocal Fold Physiology: Acoustic, Perceptual and Physiological Aspects of Voice Mechanisms, Singular, San Diego, pp. 225–232.

Jensen, P.J., 1965. Adequacy of terminology for clinical judgment of voice quality deviation. The Eye, Ear, Nose and Throat Monthly 44, 77–82.

Karlsson, I., 1991. Female voices in speech synthesis. J. Phonetics 19, 111–120.

Karlsson, I., 1992. Modelling voice variations in female speech synthesis. Speech Communication 11, 491–495.

Kempster, G.B., Kistler, D.J., Hillenbrand, J., 1991. Multidimensional scaling analysis of dysphonia in two speaker groups. J. Speech Hear. Res. 34, 534–543.

Klatt, D.H., Klatt, L.C., 1990. Analysis, synthesis and perception of voice quality variation among female and male talkers. J. Acoust. Soc. Amer. 83, 820–857.

Kreiman, J., Gerratt, B.R., 1996. The perceptual structure of pathologic voice quality. J. Acoust. Soc. Amer. 100, 1787–1795.

Kreiman, J., Gerratt, B.R., Precoda, K., Berke, G.S., 1993. Perception of supraperiodic voices. Paper presented at the 125th Meeting of the Acoustical Society of America.

Kreiman, J., Gerratt, B.R., Berke, G.S., 1994. The multidimensional nature of pathologic vocal quality. J. Acoust. Soc. Amer. 96, 1291–1302.

Ladefoged, P., 1995. A phonation-type synthesizer for use in the field. In: Fujimura, O., Hirano, M. (Eds.), Vocal Fold Physiology: Voice Quality Control, Singular, San Diego, pp. 61–76.

Lalwani, A.L., Childers, D.G., 1991. Modeling vocal disorders via formant synthesis. In: Proc. IEEE, pp. 505–508.

Laver, J., 1980. The Phonetic Description of Voice Quality. Cambridge University Press, Cambridge, MA.

Lofqvist, A., Koenig, L., McGowan, R., 1995. Voice source variations in running speech: A study of Mandarin Chinese tones. In: Fujimura, O., Hirano, M. (Eds.), Voice Fold Physiology: Voice Quality Control, Singular, San Diego, pp. 3–22.

Moore, P., Von Leden, H., 1958. Dynamic variations of the vibratory pattern in the normal larynx. Folia Phoniatrica 10, 205–238.

Price, P.J., 1989. Male and female voice source characteristics: Inverse filtering results. Speech Communication 8, 261–277.

Qi, Y., Weinberg, B., Bi, N., 1995. Enhancement of female esophageal and tracheoesophageal speech. J. Acoust. Soc. Amer. 98, 2461–2465.

Winholtz, W., Titze, I., 1997. Iniature head mount microphone for acoustic analysis, J. Speech Hear. Res. 40, 894–899.

Yumoto, E., Gould, W.J., Baer, T., 1982. Harmonics-to-noise ratio as an index of the degree of hoarseness. J. Acoust. Soc. Amer. 71, 1544–1550.