Perception of synthesized voice quality in connected speech by Cantonese speakers

Edwin M-L. Yiu[a)]
Voice Research Laboratory, Department of Speech and Hearing Sciences, The University of Hong Kong, 5/F Prince Phillip Dental Hospital, Sai Ying Pun, Hong Kong

Bruce Murdoch
Department of Speech Pathology & Audiology, The University of Queensland

Kathryn Hird
School of Psychology, Curtin University of Technology

Polly Lau
Voice Research Lab, Department of Speech and Hearing Sciences, The University of Hong Kong

(Received 20 July 2000; accepted for publication 18 June 2002)

Perceptual voice analysis is a subjective process. However, despite reports of varying degrees of intrajudge and interjudge reliability, it is widely used in clinical voice evaluation. One of the ways to improve the reliability of this procedure is to provide judges with signals as external standards so that comparison can be made in relation to these ''anchor'' signals. The present study used a Klatt speech synthesizer to create a set of speech signals with varying degree of three different voice qualities based on a Cantonese sentence. The primary objective of the study was to determine whether different abnormal voice qualities could be synthesized using the ''built-in'' synthesis parameters using a perceptual study. The second objective was to determine the relationship between acoustic characteristics of the synthesized signals and perceptual judgment. Twenty Cantonese-speaking speech pathologists with at least three years of clinical experience in perceptual voice evaluation were asked to undertake two tasks. The first was to decide whether the voice quality of the synthesized signals was normal or not. The second was to decide whether the abnormal signals should be described as rough, breathy, or vocal fry. The results showed that signals generated with a small degree of aspiration noise were perceived as breathiness while signals with a small degree of flutter or double pulsing were perceived as roughness. When the flutter or double pulsing increased further, tremor and vocal fry, rather than roughness, were perceived. Furthermore, the amount of aspiration noise, flutter, or double pulsing required for male voice stimuli was different from that required for the female voice stimuli with a similar level of perceptual breathiness and roughness. These findings showed that changes in perceived vocal quality could be achieved by systematic modifications of synthesis parameters. This opens up the possibility of using synthesized voice signals as external standards or ''anchors'' to improve the reliability of clinical perceptual voice evaluation. [DOI: 10.1121/1.1500753]

PACS numbers: 43.71.Bp, 43.71.Gv [CWT]

I. INTRODUCTION

Voice quality measurements are important in characterizing or describing a voice signal. The measures provide a severity index of dysphonic voice. Despite the rapid development of instrumentation in clinical voice assessment, perceptual voice evaluation is still a popular clinical procedure in documenting the severity of abnormal voice quality (Gerratt *et al.*, 1991). The major disadvantage of perceptual voice evaluation is that it is a subjective process and reliability is an issue. A review of the literature by Kreiman *et al.* (1993) showed that the reliability and agreement in voice quality rating could be as low as 18%, although it could improve with normal or extremely deviant qualities (see Murry *et al.*, 1987). It has been suggested that individuals develop mental (internal) standards for different voice quality through their previous exposure to voice samples (Kreiman *et al.*, 1993, 1992). These internal standards, however, are unstable and vary from one individual to another (Kreiman *et al.*, 1993). It has been demonstrated that when listeners were given explicit references (external anchors) during the rating tasks, the reliability of their judgments improved (Gerratt *et al.*, 1993; Kreiman and Gerratt, 1996). For example, Gerratt *et al.* (1993) demonstrated that the agreement in rating ''roughness'' improved from 50% (with no anchor) to 70% when anchors were provided. It is now generally accepted that the use of explicit external anchors would suppress the variable influence of the internal standards that different raters might have.

Currently, there are two possible types of external anchors that can be used to facilitate perceptual voice evaluation. One is natural occurring pathological voices and the other is synthesized signals. Synthesized signals have several advantages over natural occurring voice samples. With syn-

a)Electronic mail: edwinyiu@hku.hk

thesized signals, the number of signals that can be created is theoretically unlimited and is only restricted by the specificity of the synthesis parameters. With natural voice, a large set of pathological voice samples must exist first from which the appropriate anchors can be selected. Furthermore, it is relatively difficult to find a specific natural pathological voice which varies from other voice samples in a particular way. For example, finding a voice which is ''twice'' as breathy as another voice sample would be very difficult unless there is a large database from which one can choose. A third limitation of using natural pathological voice is that they rarely exhibit a single abnormal perceptual quality, but, instead, usually show combinations of several perceived qualities. Synthesized signals, however, do not suffer from this limitation. It is almost possible to systematically vary one particular parameter to achieve different degrees of abnormality in the synthesized signals. Other advantages of synthesized signals include simplicity and reproducibility. In natural pathological voice, acoustic properties are often complex. Many studies have attempted to extract the acoustic characteristics of these ''complex'' signals and to investigate how they affect perceptual judgment (for example, Deal and Emanuel, 1978; Hirano *et al.*, 1988; Kreiman *et al.*, 1990; Martin *et al.*, 1995; Wolfe *et al.*, 1997). Although conflicting results are shown by different studies, it is generally agreed that the two most commonly rated perceptual qualities, breathiness and roughness, are indeed multidimensional. In other words, both of these two perceptual qualities are found to correlate significantly with more than one acoustic property. For example, jitter, shimmer, and noise component have all been shown to correlate with the perception of rough and breathy quality. The reported correlation coefficients were generally of moderate strength (0.4 to 0.7). Since the acoustic properties of synthesized signals are determined by the synthesis parameters, a manipulation of the specific synthesis parameter will, in theory, produce comparatively fewer acoustically complex signals than natural voice samples. This may make it easier to study the relationship between acoustic properties and perceptual quality. In summary, provided all the synthesis parameters are detailed, these signals are relatively easy to reproduce. The ease of reproducibility of synthesized signals also facilitates replication of studies.

Although synthesized voice signals have advantages over natural voice samples in many ways, there are several limitations that investigators have to overcome. The first limitation is the naturalness of the synthesized signals. Due to the difficulty in synthesizing signals that sound natural when the speech materials get longer, perceptual voice quality studies which made use of synthesized signals used only single vowels (Bangayan *et al.*, 1997; Gerratt *et al.*, 1993; Martin and Wolfe, 1996) and avoided using connected speech. Several studies have provided some general guidelines in synthesizing natural sounding signals (Karlsson, 1991, 1992; Klatt and Klatt, 1990; Price, 1989). However, these techniques are not of much use for synthesizing connected speech.

The second limitation is related to the synthesis parameters available in the synthesizer. Generally, it has been shown that a noise component is necessary to model breathiness (Childers and Ahn, 1995; Childers and Lee, 1991; Hillenbrand, 1988; Klatt and Klatt, 1990; Martin and Wolfe, 1996) while a jitter component is needed to model roughness or aperiodicity (Hillenbrand, 1988; Klatt and Klatt, 1990). For example, it is claimed that the commercially available Klatt synthesizer (Klatt and Klatt, 1990) can change the perceived breathiness (by adjusting the aspiration noise, spectral tilt, open quotient, and increased bandwidths of first and second formants) and roughness (by adjusting the flutter). Whether these parameters are sufficient to synthesize signals that could be perceived as different degrees of pathological deviation has been questioned by some investigators (e.g., Bangayan *et al.*, 1997).

The present study had two objectives. First, it aimed to investigate whether a commercially available Klatt parallel/cascade speech model synthesizer could be used to create different pathological voice qualities using its available parameters. Second, it aimed to determine how the acoustic properties of the synthesized signals, as measured by jitter, shimmer, and noise to harmonic ratio, would affect perceptual voice quality judgment. If pathological voice quality could be synthesized successfully using a Klatt synthesizer, and was shown to correlate with perceptual ratings, this could ultimately provide a framework for creating ''reference'' voice qualities for evaluation and documenting abnormal voices.

The present study attempted a further step by synthesizing connected speech. The investigators of the present study, like other researchers (e.g., Hammarberg *et al.*, 1980; Kreiman and Gerratt, 2000), questioned the degree to which sustained vowels were representative in describing voice quality. We believe that connected speech should be used in perceptual voice evaluation because it is more representative of the voice used by speakers in daily speech tasks. Therefore, if one is to synthesize perceptual anchors with different voice qualities, connected speech should be used. In this study, we chose a simple subject–verb–object structure as the target connected speech. The Klatt synthesizer was chosen as it is commercially available and can be run on a personal computer with either a Macintosh or Window platform. This choice therefore makes it possible to allow other investigators to further explore this area without requiring more sophisticated instrument or special programming skills (cf. Hillenbrand, 1988).

II. METHODS

A. Preparation of the prototype stimuli

Synthesized signals based on a Cantonese sentence were created to simulate male and female voices. The signals were created using Sensimetrics' HLSyn Speech Synthesis System in a Microsoft Window platform. The prototype sentence used was

/baba	da	bɔ/	
father	hit	ball	(''father hits the ball'')

The HLSyn system is essentially a Klatt synthesizer (Klatt and Klatt, 1990) with the addition of some ''high-level'' synthesis parameters. In the present study, only the original, or

TABLE I. Percentage of stimuli within each synthesis parameter that were perceived as roughness, breathiness, and vocal fry. AH—amplitude of aspiration, AV—amplitude of voicing, DI—Diplophonia, FL—Flutter, OQ—Open quotient, and TL—spectral tilt. In some cases, more than one descriptor was used for the same stimulus; therefore, they may add up to more than 100%.

	AH	AV	DI	FL	TL	OQ	AV+DI
Rough	55%	29%	75%	52%	16%	15%	64%
Breathy	66%	53%	94%			19%	78%
Vocal fry			60%	8%			42%

''low-level,'' synthesis parameters were used. The average values of the synthesis parameters for the male and female prototype sentences were determined from analyzing sentences produced by six native Cantonese speakers (three males and three females) using fast Fourier transform (FFT) and linear predictive coding (LPC) analyses in the Kay Elemetric's Computerized Speech Lab 4300B system. The average values of the fundamental frequency (f_0), the first four formant frequencies (F1, F2, F3, and F4), and the duration of the vowels were used to synthesize the two prototype sentences. The fundamental frequency of the female signal was between 181 and 270 Hz, while that of the male signal was between 92 and 133 Hz. The variation in the fundamental frequency was due to the fact that the third word (/da_2/) of the sentence is a falling-rising tone. The values of these synthesis parameters were varied slightly by trial and error so that natural sounding prototype sentences, as determined by two native Cantonese speakers (authors EY and PL), were synthesized.

B. Pilot study

After the male and female prototype sentences were generated, seven synthesis parameters associated with voice qualities were varied independently with nine levels of severity to create 63 stimuli for each gender voice (a total of 126 stimuli). These seven parameters included amplitude of aspiration (AH) in dB, amplitude of voicing (AV) in dB, diplophonic double pulsing % (DI), flutter % (FL), open quotient % (OQ), spectral tilt of voicing source (TL) in dB, and amplitude of voicing in dB mixed with diplophonic double pulsing % (AV+DI). When one synthesis parameter was varied, the other parameters were all held constant at the Klatt's recommended default values. A pilot experiment was carried out using these 126 stimuli to determine (1) what perceptual voice qualities were to be included in the main study, and (2) which synthesis parameters were to be used in varying these voice qualities in the synthesized signals.

Five speech pathologists, each with at least two years of experience in assessing and treating voice disorders, were asked to serve as judges to listen to these synthesized stimuli. The judges were told that the stimuli were synthesized signals which represented different voice qualities. They were asked to label each signal with a descriptor which would best represent the voice quality. No specific instruction was given to the judges as to what descriptors were to be used.

Roughness, breathiness, and vocal fry were the three descriptors used overwhelmingly by the judges to describe the 126 stimuli. More than 75% of the stimuli were covered under these three descriptors. This was taken as an indicator that the Klatt synthesis parameters for voice quality could create signals primarily perceived as rough, breathy, or fry. These three descriptors were therefore used in the main study.

The data were further examined to determine which synthesis parameters were primarily responsible for signaling these three perceptual voice qualities. Table I lists the percentage of stimuli (with the male and female stimuli combined) within each synthesis parameter group. It was decided that the synthesis parameter which had 50% or more of its stimuli being perceived as rough, breathy, or fry were to be used in the main study to create stimuli with varying degree of roughness, breathiness, and fry. Therefore, the amplitude of aspiration (AH), diplophonia (DI), flutter (FL), amplitude of voicing (AV), and amplitude of voicing mixed with diplophonia (AV+DI) parameters were chosen to be used in the main study.

C. Main study

The objective of the main study was to investigate how the perception of different voice quality was determined by the synthesis parameters and the corresponding acoustic properties.

1. Preparation of stimuli with varying degree of abnormal voice quality

Based on the results of the pilot study (see Table I), the parameters AH, DI, FL, AV, and AV+DI were varied independently to synthesize different degree of voice quality. The incremental steps were 5 dB for the AH and AV, 10% for DI, and 20% for FL. For the stimuli which were varied in both AV and DI, each incremental step for AV was 5% (with the DI value set at 0%) until it reached the maximum value, i.e., 80%. From then onwards, the DI value was varied with 0.5% steps. Together with the prototype stimulus, this resulted in a total of 36 stimuli for each gender voice. Table II lists the synthesis parameters and the range of manipulation. When one synthesis parameter was varied, the other parameters were held constant at the Klatt's recommended default values.

Acoustic measures of jitter, shimmer, and noise-to-harmonic ratio using Kay's Computerized Speech Lab 4300B and Multidimensional Voice Program were carried out on extracted segments of these signals. Each extracted signal included the onset of the first word (/ba/) and the offset of the last word (/bɔ/). The Computerized Speech Lab has been shown to be tolerant to the fluctuation in acoustic properties in connected speech and provide valid acoustic results (Yiu *et al.*, 2000).

TABLE II. Incremental steps and range of manipulation of the values of synthesis parameters. (Default values for prototype stimulus: DI-0, AH40, AV60, FL0).

Parameter modified	Incremental steps	Range of manipulation
Amplitude of aspiration (AH)	8 Steps: AH5 for each step	AH45 to AH80
Diplophonia (DI)	10 Steps: DI-10 for each step	DI-10 to DI-100
Flutter (FL)	6 Steps: FL20 for each step	FL20 to FL100
Amplitude of voicing (AV)	4 Steps: AV5 for each step	AV65 to AV80
Amplitude of voicing at 80 dB plus diplophonia (AV80+DI)	7 Steps: DI-0.5 for each step	AV80DI-1 to AV80DI-4

2. Subjects

Twenty speech pathologists (17 females and 3 males) participated in the main study. They were all native Cantonese speakers. All had at least three years of experience in assessing and treating voice disorders on a daily basis.

3. Procedure

The synthesized stimuli were presented using a program written in Microsoft Visual Basic. The hardware system used included a Creative Sound Blaster Gold sound card and a pair of Sony SRS-PC51 speakers. The stimuli were presented in a random order to the listeners in a quiet room. Each stimulus was repeated twice, resulting in a total of 144 trials (72 female and 72 male stimuli). Precautions were taken, however, to prevent the same stimulus from being presented in a sequential manner. Half of the subjects were presented with the male stimuli first and the other half were presented with the female stimuli first. The subjects were asked whether the voice quality of each stimulus was normal, rough, breathy, or fry. Definitions of the three descriptors for abnormal quality were given to the subjects in writing during the procedure (see Table III). Subjects were given three trial items as practice before each set of stimuli was presented. The subject could choose to listen to each stimulus as many times as they would like in practice as well as in all trials.

III. RESULTS

For the acoustic measures of the female signals, the fundamental frequency was around 240 to 250 Hz, with the exception of the DI signals, which showed a frequency of around 127 Hz. This was approximately half the values of the other signal series. This halving of fundamental frequency, as pointed out by Klatt and Klatt (1990), could happen in signals where the alternate pulses disappear in extreme cases. The female AH and FL series showed a steady stepwise increase in all five acoustic measures (see Fig. 1). The female DI series also showed a general stepwise increase in the jitter (RAP and PPQ) and shimmer (Shim% and APQ) values, with the exception of DI-10 (which showed higher values in the jitter and shimmer measures when compared to those of the DI-20 signal) and DI-100 (which showed smaller values than those of DI-90). The female AV+DI series also demonstrated a general increase in the jitter (RAP and PPQ), shimmer (Shim% and APQ), and NHR, but the increase was not even for the whole AV+DI signal series. The last three signals showed relatively higher jitter and shimmer values.

For the acoustic measures of the male signals, the fundamental frequency was around 113 Hz, with the DI signals showing also about half of the values at 65 Hz. A general stepwise increase in all five acoustic measures was noticed in the male AH and FL series (see Fig. 1). Interestingly, the acoustic analysis of the male DI signal series showed a general decrease from the signal DI-20 to DI-90 in the jitter (RAP and PPQ) and shimmer (Shim% and APQ) values. This unusual finding may be due to the fact that by lowering the fundamental frequency below 60 Hz, the signal pulses contained less perturbation with the alternate pulses gone. A gentle and steady increase in RAP was noticed with the male AV+DI series.

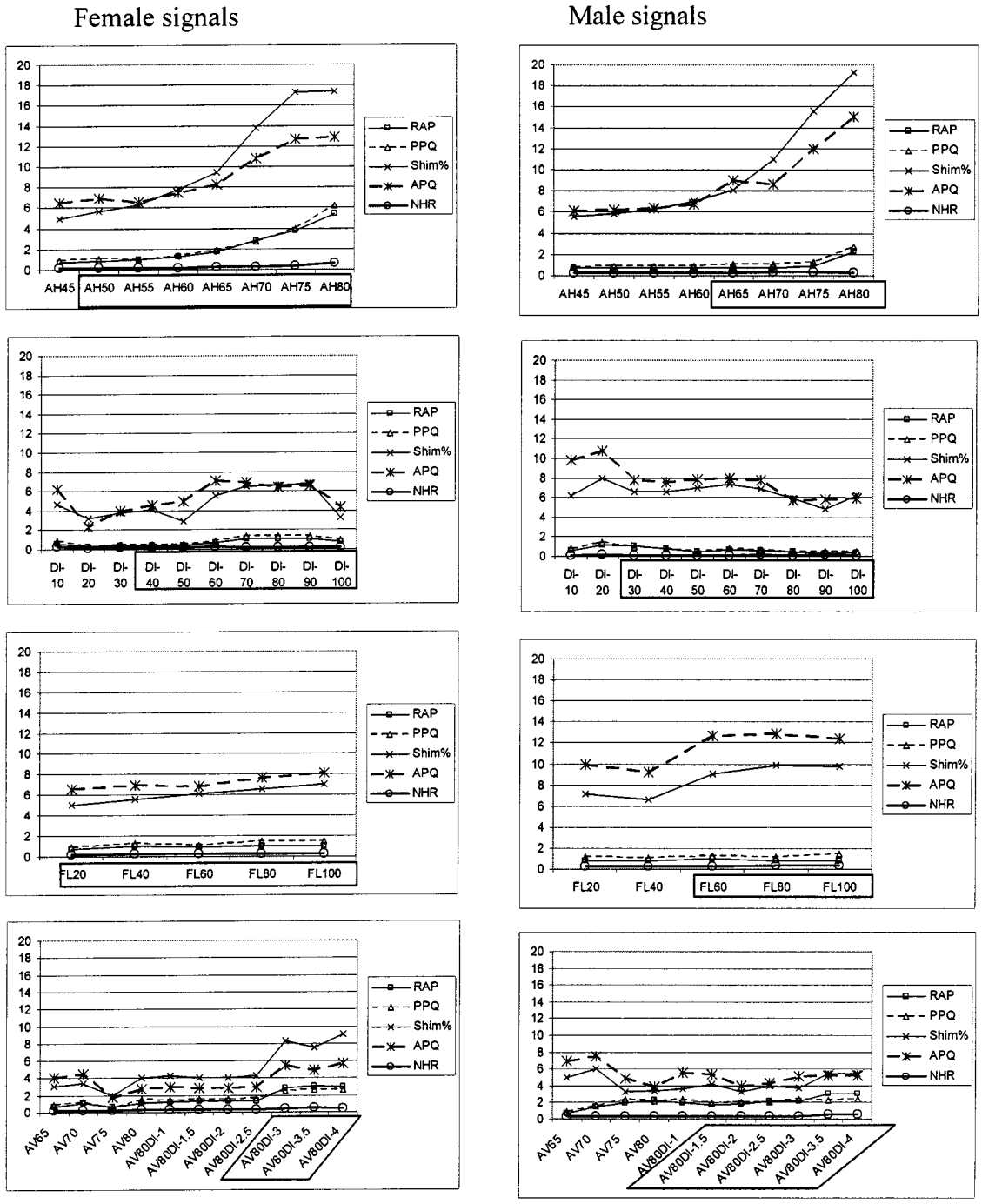
The responses of the subjects on each set of stimuli are given in Figs. 2–5. The figures show clearly that the number of subjects who perceived the stimuli as normal decreased

TABLE III. Definitions of abnormal voice qualities.

Rough
- Synonymous with "Harshness" or "Hoarseness"
- Perceptual correlates:
  - (1) Irregular quality
  - (2) Random fluctuations of glottal pulse
  - (3) Lack of clarity
  - (4) Uneven quality
- Acoustic correlates:
  - (1) Aperiodic mode of vibration
  - (2) Perturbation of the spectrum

Breathy
- Synonymous with "Whispery voice" or "Whisperiness"
- Perceptual correlates:
  - (1) Audible sound of expiration
  - (2) Audible air escape
  - (3) Audible friction noise
- Acoustic correlates:
  - (1) Related to a significant component of noise due to turbulence

Fry
- Synonymous with "Creaky"
- Perceptual correlates:
  - (1) Creaky, sounds like a creaking door
  - (2) Also sounds rough and low in pitch
- Acoustic correlate
  - (1) A complex pattern of subharmonics and modulations



RAP- Relative Average Perturbation,
PPQ- Pitch Perturbation Quotient, Shim%- Shimmer Percent,
APQ- Amplitude Perturbation Quotient, NHR- Noise to Harmonic Ratio

Synthesized values enclosed in boxes refer to signals being perceived by 75% of listeners as abnormal

FIG. 1. Acoustic measurements of synthesized signals.

with increasing values of the synthesis parameters. In other words, the higher the values of the synthesis parameters, the higher the number of subjects who perceived the signals as abnormal.

In order to determine the cutoff point for a set of stimuli to be perceived as abnormal, a binomial distribution was employed using a 95% confidence level. Since each subject had to decide first whether the stimulus was normal or abnormal, the chance level of making any judgment is 0.5. With a total of 20 listeners, a binomial distribution table indicated that at least 15 of them had to agree on the judgment in order to reach the 95% confidence level (Runyon *et al.*, 1996).

The signals which were determined by at least 15 or

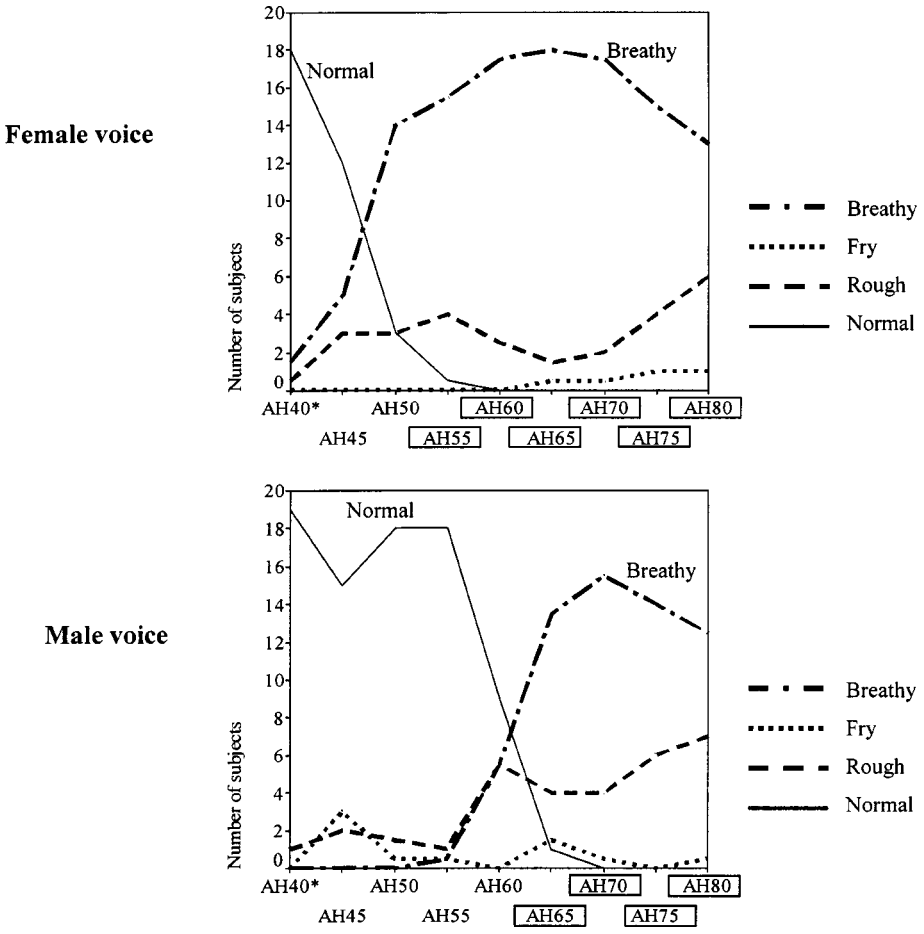


FIG. 2. The use of different descriptors in labeling stimuli with varying degree of amplitude of aspiration (AH).

* is the prototype stimulus

Synthesized values enclosed in boxes refer to signals being perceived by 75% of listeners as abnormal

more listeners to be abnormal are marked in Figs. 1–5 inside boxes. We further assumed that subjects would have one out of three chances to label the stimulus as either breathy, rough, or fry after having identified a stimulus as abnormal. By using the binomial distribution again, it was determined that at least 10 subjects would have to agree on a particular voice quality descriptor with a confidence level of 95%, assuming, of course, that at least 15 subjects had decided that a particular stimulus was abnormal.

For the stimuli that varied in AH and FL, the results clearly showed that they were perceived as breathy and rough, respectively, in both the female and male stimuli (see Table IV). For the DI parameter, male stimuli with high values of DI were all perceived as having a fry quality, while for the female stimuli the results were less clear. Female stimuli with DI values of 60 and 70 were perceived as rough, but the descriptor changed to vocal fry when the DI value increased to 100 (Table IV). When the DI parameter was varied in conjunction with a high AV, the male stimuli were generally perceived as rough. However, when the DI value increased up to 3%, a vocal fry quality was perceived. When the DI value increased to 4%, almost as many listeners perceived the stimuli as vocal fry as perceived them as rough (Table IV). For the female stimuli, only the AV80DI-3 stood out as having a rough quality. When the DI values increased, both rough and fry qualities were reported (Table IV).

It should be noted that female stimuli did not require so much aspiration (AH) as did the male stimuli to be perceived as breathy. The female stimuli were perceived as breathy starting at AH50 while the male stimuli were not perceived as breathy until AH had risen to 65 or above. When DI was added to the male stimuli, they were perceived as vocal fry when the DI value reached 30%. However, when DI was added to female stimuli, the perceived quality was less distinct. Apart from being perceived as vocal fry, roughness was also reported. Only when the DI was increased to 100% were the stimuli perceived distinctively as fry.

For the stimuli with variation of DI in combination with high levels of AV, the male stimuli were perceived as rough when the DI value started at 1.5% while the female required a DI of 3% to be perceived as rough. Once the DI values increased further to 3.5%, the fry quality began to appear in the perception of some listeners.

Table V shows the significant correlation coefficients between the number of judges that used a particular perceptual voice quality descriptor and the acoustic properties of the

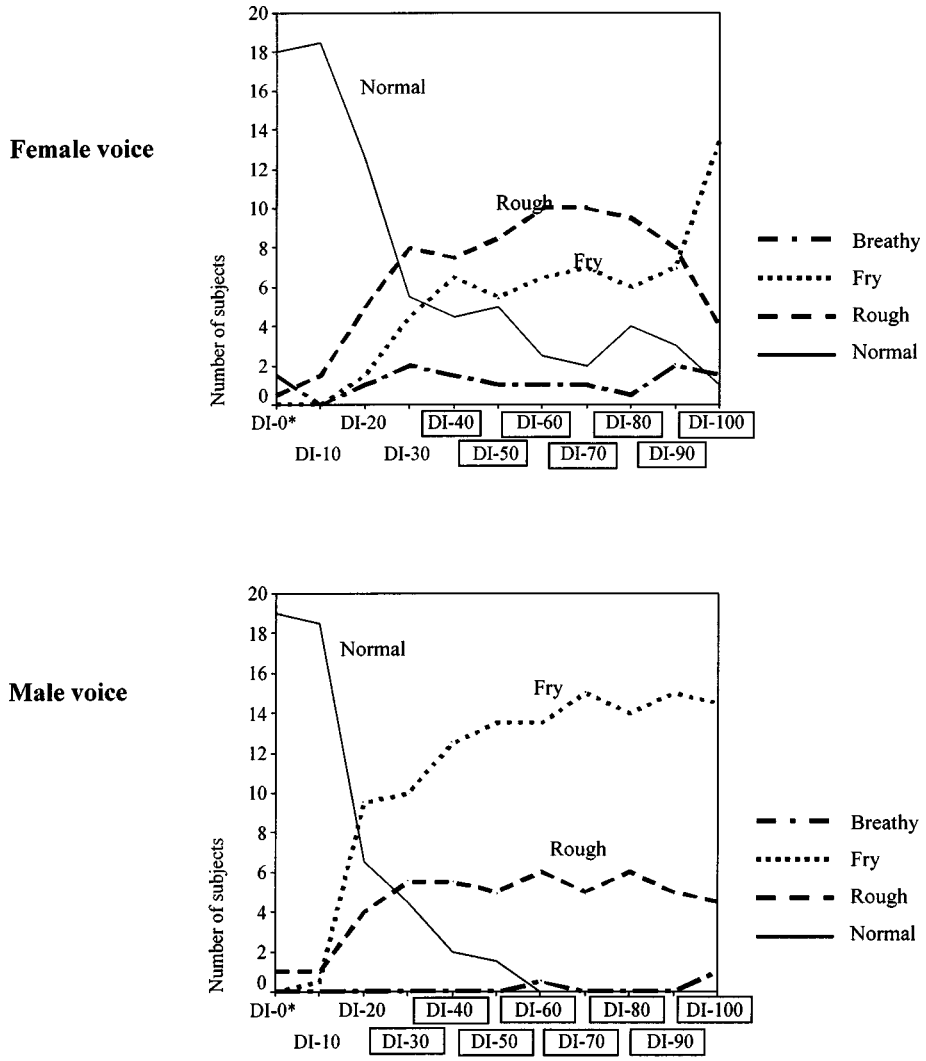


FIG. 3. The use of different descriptors in labeling stimuli with varying degree of diplophonia (DI).

* is the prototype stimulus

Synthesized values enclosed in boxes refer to signals being perceived by 75% of listeners as abnormal.

synthesized signals. The female AH stimuli showed no significant correlation while the male AH stimuli showed significant correlation between breathiness and three acoustic parameters (PPQ, Shim%, and APQ). The female DI stimuli showed a significant correlation between perceptual roughness and APQ whereas the male DI stimuli showed a significant negative correlation between perceptual roughness and shimmer percent. With the female FL stimuli, no significant correlation was found between perceptual roughness and any of the acoustic variables. The male FL stimuli, however, showed a significant correlation between roughness and shimmer percent. For the AV+DI stimuli, the female set showed significant correlation between perceptual rough and fry qualities with the RAP, whereas the male stimulus set demonstrated significant correlations between fry quality and the RAP as well as the NHR.

IV. DISCUSSION

This study shows the Klatt synthesizer can be used to create signals with different perceptual voice qualities by varying the synthesis parameters. It also shows that different degrees of synthesis values would be required to create a similar degree of perceptual quality in voices of different genders.

Synthesizing speech using a high values of the amplitude of aspiration (AH) parameter, as Klatt and Klatt (1990) contended, results in the perception of a breathy quality. In Fig. 1, it is clearly shown that increasing AH values resulted in a sharper increase in shimmer values (Shim% and APQ) and a moderate increase in jitter values (RAP and PPQ). These changes in acoustic properties appeared to account for the perception of breathiness. Klatt and Klatt (1990) suggested the default AH value be set at 40 dB so that a synthesized stimulus would sound natural. This study shows that relatively higher aspiration noise (AH) is needed in the male stimuli (at AH65) than in the female stimuli (at AH50) in order to produce a similar degree of perceptual breathiness. A closer examination of the acoustic properties of the signals (see Fig. 1) showed that the jitter (RAP and PPQ) had increased to a higher degree in the female signals than in the male signals with identical AH values. Therefore, this higher

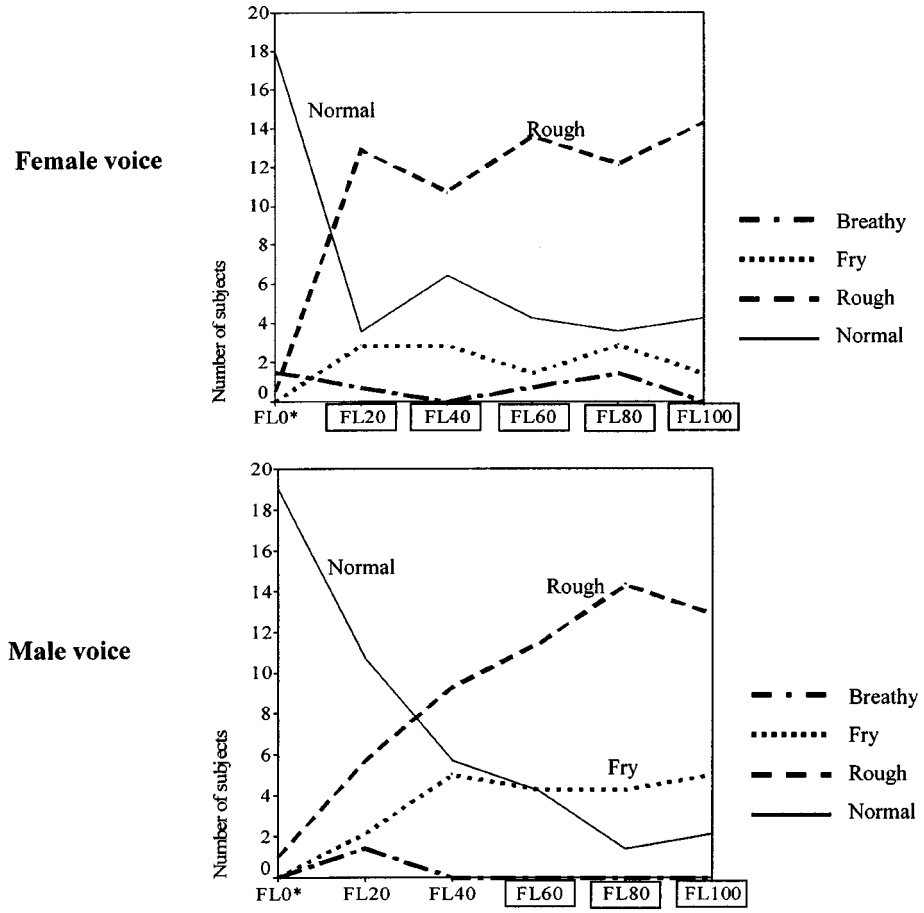


FIG. 4. The use of different descriptors in labeling stimuli with varying degree of flutter (FL).

* is the prototype stimulus

Synthesized values enclosed in boxes refer to signals being perceived by 75% of listeners as abnormal

degree of jitter in the female signals might have resulted in more breathy signals. The difference between the male and female signals was basically in the fundamental and formant frequencies. Therefore, the source of jitter might have come from the interaction of the frequency parameter with the aspiration parameter of the Klatt synthesizer. The correlation between the AH values in the male signals and the three acoustic measures (PPQ, Shim%, and APQ) are rather high. They are at least 0.78 or higher (see Table V). For the female AH signals, the ceiling effect might have accounted for the lack of correlation between the perceptual breathiness and any of the acoustic variables. This could be attributed to the high number of judges which perceived the female AH stimuli as breathy even with an AH value as low as 50.

The diplophonia (DI) parameter, according to Klatt and Klatt (1990), uses two glottal pulses in slightly different phases. In the present study, it was demonstrated that high values of this parameter are associated with perceptual roughness and vocal fry quality (see Fig. 3). When the DI value was increased beyond 5%, signals were primarily perceived as vocal fry in the male stimuli. However, in the female stimuli, an increase in the DI values was equally perceived as roughness or vocal fry quality. With the female DI signal series, one of the shimmer measures, APQ, correlated significantly with the perception of roughness (Spearman rho $= 0.73$, $p = 0.04$). However, no significant correlation was found between the acoustic measures and vocal fry quality. Indeed, the apparent negative correlation between the perceptual roughness and RAP might have been due to the error in extracting the perturbation measurements within the male DI stimuli as a result of the disappearance of alternate pulses in the signals.

Increasing values of the flutter (FL) parameter were found to produce a rough quality (see Fig. 4). Relatively higher flutter value was needed in the male stimuli (FL60) than in the female stimuli (FL20) in order to make the stimuli sound rough. Indeed, the male FL signal series already demonstrated relatively higher Shim% and APQ values than the female signals with the same FL values. The male FL stimuli showed a significant correlation between perceptual roughness and Shim% while there is a lack of correlation of any kind in the female stimuli. This lack of correlation appeared to be attributed to the ceiling effect of the subjects perceiving the female FL stimuli as perceptually rough. A number of subjects reported that the perception of roughness due to high FL values was very different from that produced by high AV + DI values. They reported a trembling quality in the stimuli synthesized with increased flutter (FL) values. Indeed, a reexamination of the data from the pilot

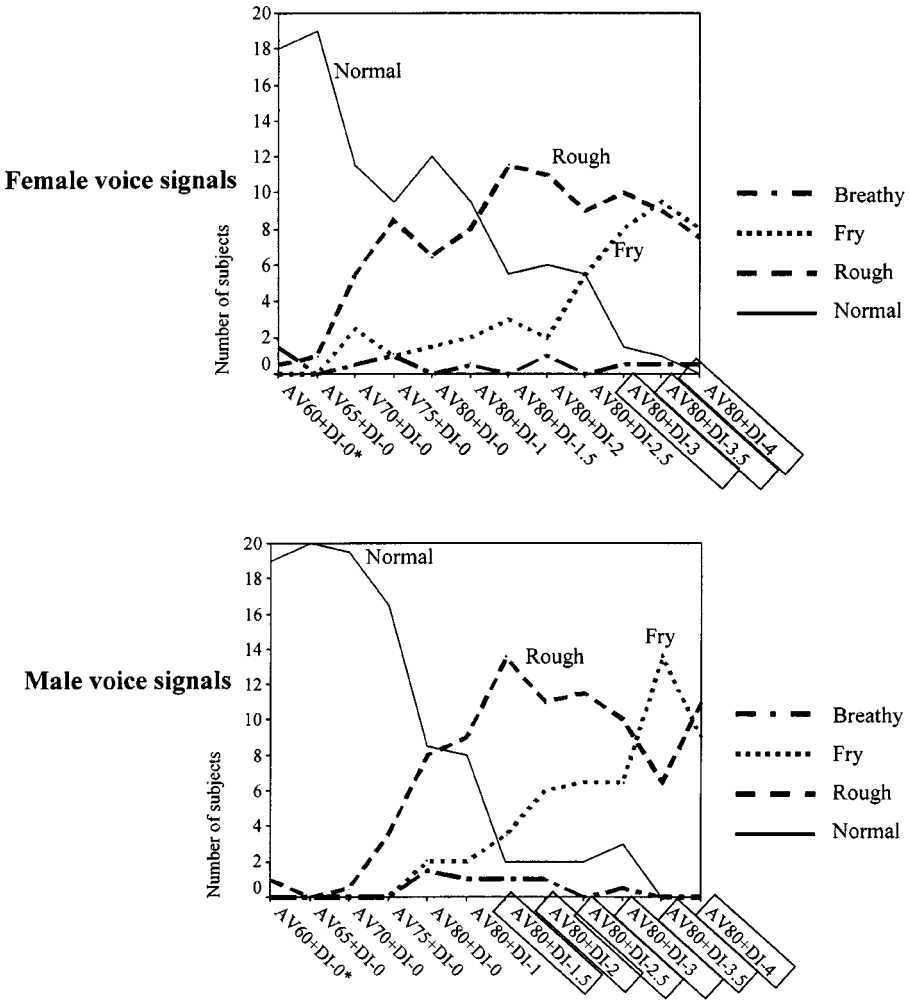


FIG. 5. The use of different descriptors in labeling stimuli with high amplitude of voicing and varying degrees of diplophonia (AV+DI).

* is the prototype stimulus

Synthesized values enclosed in boxes refer to signals being perceived by 75% of listeners as abnormal.

study showed that 46% of stimuli with increased FL values were perceived as showing tremor. However, as ''tremor'' was not an option given in the labeling task in the main study, the subjects might have been forced to choose roughness as the closest descriptor.

When the amplitude of voicing (AV) was increased to 80 dB and a few percent of DI was added, roughness was perceived. It should also be noted that none of the stimuli that varied only in the AV parameter was perceived as abnormal. Only when the DI was varied (even in small degree in the order of 1.5% to 3%) in combination with a high value of AV were the stimuli perceived as abnormal. Relatively higher degrees of AV plus DI are needed in the female stimuli (AV80DI-3) than in the male stimuli (AV80DI-1.5) in order to produce a rough or fry quality. Stimuli with high values of DI were perceived distinctively as vocal fry when synthesized as a male voice but equivocally as vocal fry and rough when synthesized as female voice. When AV80 was used to synthesize the signals, a relatively higher degree of DI was needed in the female signal (3% of DI) than in the male signals (1.5% of DI) in order to make the signals perceptually rough. The correlation of RAP and NHR with perceptual roughness and fry in these stimuli (see Table IV) showed the multidimensional nature of perceptual voice qualities and acoustic properties.

In summary, the Klatt synthesizer was found to be capable of synthesizing different degrees of breathiness, vocal fry, and roughness. Signals generated with a small degree of aspiration noise (AH) were perceived as breathy while small degrees of double pulsing (DI) or flutter (FL) were perceived as roughness. When the double pulsing (DI) and flutter (FL) increased, vocal fry was perceived instead of roughness.

Although some investigators (e.g., Klatt and Klatt, 1990; Bangayan *et al.*, 1997) contend that the Klatt synthesizer is better at synthesizing male voices than female voices, the present study demonstrated that it is possible to synthesize female voice with reasonably high quality. Nevertheless, the amount of AH, DI, or FL required to produce the perception of a similar level of pathological voice qualities was different for male and female voice stimuli.

The first objective of the study was to determine whether the Klatt synthesis parameters could be used to create signals with different types and degrees of voice quality. The findings from the present study show that the Klatt synthesizer

TABLE IV. Voice quality descriptors used by at least ten judges for particular synthesis parameters. AH—amplitude of aspiration, AV—Amplitude of voicing, DI—Diplophonia, and FL—Flutter.

Stimuli	Descriptors
Female	
AH50, AH55, AH60, AH65, AH70, AH75, AH80	Breathy
DI-40, DI-50	Rough and fry[a]
DI-60, DI-70,	Rough
DI-80, DI-90,	Rough and fry[a]
DI-100	Fry
FL20, FL40, FL60, FL80, FL100	Rough
AV80DI-3	Rough
AV80DI-3.5, AV80DI-4	Rough and fry[a]
Male	
AH65, AH70, AH75, AH80	Breathy
DI-30, DI-40, DI-50, DI-60, DI-70, DI-80, DI-90, DI-100	Fry
FL60, FL80, FL100	Rough
AV80DI-1.5, AV80DI-2, AV80DI-2.5, AV80DI-3, AV80DI-4	Rough
AV80DI-3.5	Fry

[a]None of the two descriptors was statistically more significant than the other, i.e., they did not reach the ''ten judges'' criterion. However, since similar numbers of judges were found in using these two descriptors, both descriptors are therefore reported here.

can be used to create synthesized voice signals with breathy, rough, and fry qualities. However, there are still some limitations with the Klatt synthesizer. First, the synthesized signals might not be exact matches to naturally occurring dysphonic qualities. This may have happened because when the fundamental frequency of the stimulus is not a whole multiple of the sampling rate, artifacts will be created and contribute to perceived roughness.[1] Therefore, uneven roughness could have distributed across the connected speech as each syllable had different fundamental frequency. A second limitation relates to the variation of voice quality across an utterance when connected speech material is used as the stimuli. It is known that voice quality would vary due to consonant articulation, use of different vowels (e.g., tensed versus laxed), or prosody changes (such as glottalization at phrase endings). In the present study, the quality settings were held constant across the whole utterance and these possible variations were not taken into consideration. These probably accounted for some of the ''unnaturalness'' in the synthesized dysphonic stimuli. Third, it is not known whether the Klatt synthesizer is capable of synthesizing all pathological voice qualities found in clinical situations using its current available synthesis parameters. As the Klatt synthesize is originally based on models derived from normal voices and is not designed to readily accommodate pathological qualities, such a question is a valid one. Indeed, more research is needed to develop appropriate models for pathological voice quality. A recent report by Bangayan *et al.* (1997) has explored some of the alternatives and has made two suggestions. The first is to include jitter and shimmer parameters in the Klatt synthesizer, and the second is to modify the DI parameter of the Klatt synthesizer so that fundamental frequency and amplitude could be varied separately. The DI parameter operates by truncating and reducing the amplitude of the closed phase of every second pulse. This is very different from natural signals. Therefore, the DI parameter produced effect which is not just perceived as diplophonia but as rough as well (see Table V). The fourth problem with the Klatt synthesizer, as noted by Hermes (1991), is that when noise is added up to a certain level, the noise is perceived as a separate noise stream rather than as a further increase in the breathiness of the noise signal. Finally, in the present study, it has been shown that the FL does not produce jitter appropriately as Klatt and Klatt (1990) claimed. Although increasing FL does alter the fundamental frequency

TABLE V. Correlation coefficients between values of acoustic parameters and the number of judges that used a particular descriptor. AH—Amplitude of aspiration, AV—amplitude of voicing, DI—diplophonia, FL—flutter, RAP— relative average perturbation, PPQ—pitch perturbation quotient, Shim%—shimmer percent, APQ—amplitude perturbation quotient, NHR—noise to harmonic ratio.

Varied synthesis parameters	Acoustic parameters	Perceptual descriptors	Spearman rho	Two-tailed *p* level
Female AH		Breathy	No significant correlation	
Female DI	APQ	Rough	0.73	0.04
		Fry	No significant correlation	
Female FL		Rough	No significant correlation	
Female AV+DI	RAP	Fry	0.75	0.03
	RAP	Rough	0.80	0.02
Male AH	PPQ	Breathy	0.78	0.02
	Shim%	Breathy	0.83	0.01
	APQ	Breathy	0.78	0.01
Male DI	RAP	Fry	−0.77	0.009
Male FL	Shim%	Rough	0.90	0.04
Male AV+DI	RAP	Fry	0.75	0.008
	NHR	Fry	0.82	0.002

in the time domain, this results in the perception of tremor, not roughness. Provided one takes these limitations into consideration when synthesizing pathological voice stimuli, the Klatt synthesizer is a useful signal synthesizer for researchers who want to study the perception of voice quality.

The second objective of this study was to determine how the acoustic properties affected perceptual voice evaluation. Previous studies have shown that the correlation between acoustic variable and perceptual qualities varies between 0.4 and 0.7 and a particular perceptual quality may correlate with several acoustic measures (Kreiman and Gerratt, 2000). The results from our present study also support these findings. The significant correlation coefficients were moderately high (>0.73; see Table V). Furthermore, many perceptual qualities were also found to correlate with more than one acoustic variable. Nevertheless, we are unable to make a direct comparison between our data and those from the previous studies as the coefficients from the present study were based on the number of judges agreeing on a particular quality rather than on the severity of each quality. It would be more appropriate to investigate the correlation between the acoustic variables and the severity ratings made by the judges on different perceptual qualities. Nevertheless, it is clear from the results that perceptual voice quality is multidimensional in nature. This means that it is determined not by a single acoustic variable, but more likely by a set of acoustic variables.

A pertinent question that many voice clinicians may ask is whether these synthesized signals have any clinical significance. Chan and Yiu (2002) employed the synthesized signals developed in the present study as anchors to investigate whether they could improve the reliability of perceptual voice rating. Their findings showed that the use of synthesized signals as anchors facilitated a better reliability than natural voice anchors (Chan and Yiu, 2002). Therefore, these synthesized signals are of clinical importance as they are the pertinent materials for investigating perceptual voice evaluation. It is hoped that the present study serves to generate more interest in investigating the process of voice quality perception.

ACKNOWLEDGMENTS

This study was supported by a grant from the Hong Kong Research Grant Council (HKU7196/98H). The authors would like to thank Dr. Alex Francis, Dr. Christopher Turner, and two anonymous reviewers for their constructive comments on the initial draft of the manuscript.

[1]This particular issue related to the use of Klatt synthesizer was brought to the attention of the authors by an anonymous reviewer of the manuscript.

Bangayan, P. T., Long, C., Alwan, A. A., Kreiman, J., and Gerratt, B. R. (**1997**). ''Analysis by synthesis of pathological voices using the Klatt synthesizer,'' Speech Commun. **22**, 343–368.

Chan, K. M.-K., and Yiu, E. M. L. (**2002**). ''The effect of anchors and training on the reliability of perceptual voice evaluation,'' J. Speech Lang. Hear. Res. **45**, 111–126.

Childers, D. G., and Ahn, C. (**1995**). ''Modeling the glottal volume-velocity waveform for three voice types,'' J. Acoust. Soc. Am. **97**, 505–519.

Childers, D. G., and Lee, C. K. (**1991**). ''Vocal quality factors: analysis, synthesis, and perception,'' J. Acoust. Soc. Am. **90**, 2394–2410.

Deal, R., and Emanuel, F. (**1978**). ''Some waveform and spectral features of vowel roughness,'' J. Speech Hear. Res. **21**, 250–264.

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

Gerratt, B. R., Till, J. A., Rosenbek, J. C., Wertz, R. T., and Boysen, A. E. (**1991**). ''Use and perceived value of perceptual and instrumental measures in dysarthria management,'' in *Dysarthria and Apraxia of Speech: Perspective on Management*, edited by C. A. Moore, K. M. Yorkston, and D. R. Beukelman (Brookes, Baltimore), pp. 77–93.

Hammarberg, B., Fritzell, B., Gauffin, J., Sundberg, J., and Wedin, L. (**1980**). ''Perceptual and acoustic correlates of abnormal voice qualities,'' Acta Otolaryngol. Suppl. (Stockh) **90**, 441–451.

Hermes, D. J. (**1991**). ''Synthesis of breathy vowels: Some research methods,'' Speech Commun. **10**, 497–502.

Hillenbrand, J. (**1988**). ''Perception of aperiodicities in synthetically generated voices,'' J. Acoust. Soc. Am. **83**, 2361–2371.

Hirano, M., Hibi, S., Yoshida, T., Hirade, Y., Kasuya, H., and Kikuchi, Y. (**1988**). ''Acoustic analysis of pathological voice,'' Acta Otolaryngol. (Stockh) **105**, 432–438.

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

Karlsson, I. (**1992**). ''Modeling voice variations in female speech synthesis,'' Speech Commun. **11**, 491–495.

Klatt, D. H., and Klatt, L. C. (**1990**). ''Analysis, synthesis, and perception of voice quality variations among female and male talkers,'' J. Acoust. Soc. Am. **87**(2), 820–857.

Kreiman, J., and Gerratt, B. R. (**1996**). ''The perceptual structure of pathologic voice quality,'' J. Acoust. Soc. Am. **100**, 1787–1795.

Kreiman, J., and Gerratt, B. R. (**2000**). ''Measuring vocal quality,'' in *Voice Quality Measurement*, edited by M. Kent and M. Ball (Singular-Thomson Learning, San Diego), pp. 73–102.

Kreiman, J., Gerratt, B. R., Kempster, G. B., Erman, A., and Berke, G. S. (**1993**). ''Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research,'' J. Speech Hear. Res. **36**, 21–40.

Kreiman, J., Gerratt, B. R., and Precoda, K. (**1990**). ''Listener experience and perception of voice quality,'' J. Speech Hear. Res. **33**, 103–115.

Kreiman, J., Gerratt, B. R., Precoda, K., and Berke, G. S. (**1992**). ''Individual differences in voice quality perception,'' J. Speech Hear. Res. **35**, 512–520.

Martin, D. P., and Wolfe, V. I. (**1996**). ''Effects of perceptual training based upon synthesized voice signals,'' Percept. Mot. Skills **83**(3) (part 2), 1291–1298.

Martin, D., Fitch, J., and Wolfe, V. (**1995**). ''Pathological voice type and the acoustic prediction of severity,'' J. Speech Hear. Res. **38**, 765–771.

Murry, T., Brown, W. S. J., and Rothman, H. (**1987**). ''Judgments of voice quality and preference: Acoustic interpretation,'' J. Voice **1**, 252–257.

Price, P. J. (**1989**). ''Male and female voice source characteristics: Inverse filtering results,'' Speech Commun. **8**, 261–277.

Runyon, R. P., Harber, A., Pittenger, D. J., and Coleman, K. A. (**1996**). *Fundamentals of Behavioural Statistics* (McGraw-Hill, New York).

Wolfe, V., Fitch, J., and Martin, D. (**1997**). ''Acoustic measures of dysphonic severity across and within voice types,'' Folia Phoniatr. Logop. **49**, 292–299.

Yiu, E., Worrall, L. E., Longland, J., and Mitchell, C. (**2000**). ''Analysing vocal quality of connected speech using Kay's Computerized Speech Lab: A preliminary finding,'' Clin. Linguist. Phonetics **14**(4), 295–305.