TIME AND FREQUENCY SYNTHESIS PARAMETERS OF SEVERELY PATHOLOGICAL VOICE QUALITIES

A. Alwan, P. Bangayan, J. Kreiman, and C. Long***

Department of Electrical Engineering, UCLA
405 Hilgard Ave.
Los Angeles, CA 90024

ABSTRACT

This paper describes a pilot study into the mechanics of synthesizing severely pathological voices. Successful synthesis of such voices may ultimately provide a quantitative method for evaluating and documenting voice qualities. An analysis-by-synthesis approach using the formant synthesizer KLSYN was used to model the voices of 24 patients suffering from voice disorders. Results suggest a number of modifications to KLSYN that would facilitate synthesis of these voices.

1. INTRODUCTION

No standard system of description exists for pathological voice qualities. 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 creating a standard protocol for quality assessment.

A serious limitation of past studies on synthesizing pathological voices is the focus on single aspects of quality and/or of the acoustic signal. Accordingly, previous studies provide little insight into the techniques necessary to generate reasonable copies of natural pathological voices.

The present study used Sensyn 1.1, the Sensimetrics version of the Klatt formant synthesizer KLSYN [1]. 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 or nasal coupling, a provision for time-varying parameters to model unsteady quality, and a "diplophonia" parameter to model bicyclic (period doubled, bifurcated) phonation. However, KLSYN was designed for synthesizing normal voices, and questions remain about its suitability for producing acceptable pathological stimuli.

Supported in part by NIH. *Head and Neck Surgery, UCLA, **HST-MIT, Cambridge, MA.

2. METHODOLOGY

Twenty-four voice samples of the vowel /a/ were selected from a library of recordings. Signals were 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. Stimuli were informally grouped (based on the perceptual judgment of author JK) into the following categories: rough and rough-breathy (11 tokens), bicyclic (8 tokens), rough-bicyclic (1 token), strained-breathy (2 tokens), and strained-rough (2 tokens).

Time and frequency domain analyses of each voice sample were undertaken to guide synthesis efforts. In the time domain, we tracked long-term amplitude and frequency modulations. In the frequency-domain, the fundamental frequency (F0), formant frequencies, strengths of the first three harmonics, and any additional resonances were measured.

3. SYNTHESIS PROCEDURES

Synthetic waveforms were modeled after each of the natural tokens using the cascade branch of the synthesizer. Synthesis proceeded as follows.

Step 1: Match Formant Frequencies and Mean F0: As a first step, the formants' frequencies and bandwidths were matched. In addition, the mean value of F0 was used.

Step 2: Adjust Amplitude of Voicing (AV) and Amplitude of Aspiration Noise (AH): When synthesizing

pathological voices, matching AH is as important as matching AV because increased breathiness often enhances the perception of rough and bicyclic qualities.

Step 3: Adjust Open Quotient (OQ): The degree of the strained quality in a voice, if present, was matched by altering OQ, which defines the percentage of the pitch period in which the glottis is open.

Step 4: Boost Low Frequency Components: It was often difficult to match the amplitudes of harmonics below F1 in the synthetic voice to those of the natural waveforms. This harmonic mismatch resulted in synthetic voices which did not sound as "rich" as the natural voices. In these cases, additional pole/zero pairs (the nasal and/or tracheal) were placed around the frequencies of the first formant. Typically, one pole/ zero pair was placed at the first harmonic, and the other pair, at the second harmonic.

Step 5: Alter F0: F0 was varied to model the natural utterances. For voices with high jitter, F0 values were generated such that they followed a Gaussian distribution, given a mean and variance calculated from the natural sample. In other cases (particularly with bicyclic voices), values were measured manually from the natural waveform and imported into the synthesizer. For bicyclic voices where F0 values were not entered manually, the diplophonia (DI) and flutter (FL) parameters were used. DI was useful for synthesizing some bicyclic voices, but failed to capture the pattern of F0 and amplitude alterations for others. Flutter creates slowly varying and regularly repeating F0 values, as described by:

Step 6: Alter AV as a Time-Varying Parameter for Amplitude-Modulated Voices: The parameter AV was altered in a time-varying fashion to model shimmer.

Step 7: Add Additional Pole/Zero Pairs if Necessary: Some voices required additional pole/zero pairs to model nasal and/or tracheal coupling.

In some cases, the speed quotient of the glottal waveform (parameter SQ) was adjusted to match the overall spectral shape.

These steps were repeated as necessary to fine-tune the synthesis, until the synthesized voice was judged to be a reasonable match to the original waveform.

4. SYNTHESIS RESULTS

Rough and Rough-Breathy Voices: Eleven rough and rough-breathy voices (4 female, 7 male) were analyzed. Seven samples had fairly steady qualities, but four voices varied considerably. Capturing the variation in F0 proved critical for successful synthesis of these voices. Seven of the ten voices required time-varying AV, and the rough voices were generally accompanied by turbulent noise. Eight out of the ten voices had OQ ¿ 50

Bicyclic Voices: Bicyclic voices (also referred to as diplophonia or bifurcated phonation) present a pattern of cycles that alternate in frequency, amplitude, or both, in a large-small-large-small pattern. Eight bicyclic voices (4 female, 4 male) were analyzed. None showed a perfect pattern of periods alternating in an ABAB fashion. Instead, three patterns emerged: (1) three voices had fundamental frequencies alternating among a small number of values (typically 5 to 9); (2) F0 was bimodally distributed for 4 voices; and (3) one voice had increasing bicyclicity with time.

The male voices tended to sound more strained and also had weaker first harmonics than their female counterparts. Hence, the OQ was less than 50the male voices, but only for one female voice.

Strained-Breathy and Strained-Rough Voices: These were the most difficult voices to synthesize. Attempts were made to capture the strained yet breathy quality by changing the speed quotient, changing bandwidths, sequentially altering OQ, time-varying AH to modulate breathiness, utilizing FL and varying F0. None of these techniques proved entirely successful.

The strained-rough voices were unsteady during the periods when the voices become strained. Techniques used to model these voices included time-varying SQ, OQ, and AV. The gargly nature and unsteadiness of the voices were not captured well in the synthesized versions.

5. PERCEPTUAL EVALUATION

As suggested above, some attempts at synthesizing pathological voices were subjectively more successful 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.

5.1. Methods

Ten expert listeners participated in this experiment. The 24 voices described above were used as stimuli.

Stimuli were normalized for peak voltage, and onsets and offsets were multiplied by 25 ms ramps to eliminate click artifacts.

Listeners heard each natural sample paired with its synthetic copy, and were asked to judge how well the copy matched the original on a 1-7 scale (1: perfect match). Stimuli were presented in free field at a comfortable listening level.

5.2. Results

Listeners unanimously reported being pleased by the overall quality of the synthesis. Mean ratings ranged from 1.30 to 6.30. As Figure 1 shows, listeners agreed well in their ratings when they thought that a copy was nearly identical to the original sample. For less successful copies, both the mean rating and the variability in ratings increased.

On the whole, copies of male voices (filled circles in Fig. 1,) were more successful than were copies of female voices (asterisks in Fig. 1). Most "unsuccessful" ratings reflect failures to model unsteady or gargled qualities or failures in modeling voices with strong low frequency components and a muffled quality. The spectrogram of an unsteady voice which was difficult to synthesize is shown in Fig. 2.

6. SUMMARY AND DISCUSSION

On the whole, our efforts at synthesizing severely pathological voices were fairly successful. Less severe pathological voices (in particular, male voices with steady F0 contours) were synthesized best. Our results suggest that several modifications to the Klatt synthesizer would improve the quality of synthesis, and would facilitate the production of a wide range of pathological qualities. (1) More than six formants are needed to synthesize voices at high sampling rates. KLSYN provides enough variable formants to support sampling rates of 10-12 kHz. Providing more formants, with variable frequencies and bandwidths, would alleviate this difficulty. (2) A parameter is needed to increase the spectral energy below F1. KLSYN-synthesized voices often lack energy at frequencies below the first formant. One solution is to increase the open quotient (OQ), which increases the first harmonic energy. Another is to add pole/zero pairs below F1. Both solutions were often inadequate. A new parameter that boosts the harmonics below F1 would provide more low-frequency energy. (3) More pole/zero pairs are needed to account for increased coupling to the nasal tract or to the trachea in pathological cases. (4) Jitter and shimmer parameters are needed to facilitate modeling of perturbations in natural voices. In this study, jitter and shimmer were modeled by manually entering the time-varying values of F0 and AV. This approach is cumbersome to use. KLSYN does offer a flutter parameter (FL) which slowly time-varies F0, but this does not model jitter appropriately. (5) The diplophonia parameter (DI) should be split into separate F0 and amplitude perturbation parameters. DI was designed to model bicyclic (bifurcated, period doubled) phonation, and as presently implemented it attenuates and delays every other glottal pulse. The resulting patterns of amplitude and frequency variation do not match measurements of natural bicyclic waveforms, for which there is no consistent correlation between amplitude and F0. Modeling would be improved by allowing amplitude to be changed independent of delay, and/or by allowing amplitude to be specified for each individual period. (6) The update interval (UI) should be implemented as a time-varying function that could be updated at any time instant and not necessarily at the beginning/end of a pitch period. Some parameters, such as F0, are updated at the end of each period while most other time-varying parameters (such as AV) are specified at multiples of UI, and linear interpolation is used to determine values between updates. This is problematic when one needs to change attributes of one glottal pulse without affecting other pulses.

Finally, more acoustic modeling of severe vocal pathology is necessary. As discussed above, most acoustic models are based on variations in normal speech, and do not easily accommodate pathologic cases. Effective synthesis of strained/ breathy voices and gargled qualities in particular must await improvements in modeling as well as improvements in synthesizers.

ACKNOWLEDGMENT:

This research was supported in part by NIDCD grant DC 01797.

7. REFERENCES

[1] Klatt, D.H., and Klatt, L.C. (1990). "Analysis, Synthesis and Perception of Voice Quality Variation among Female and Male Talkers," JASA, 820-857.

Fig. 1: Variability in perceptual ratings versus mean similarity of the synthetic tokens to the natural voices, as judged by ten expert listeners.

Fig. 2: Spectrogram of an unsteady female voice which was the most difficult voice to synthesize.