Acoustics'08 Paris
June 29-July 4, 2008
www.acoustics08-paris.org

# Transforming modal voice into irregular voice by amplitude scaling of individual glottal cycles

Tamás Bőhm[a], Nicolas Audibert[b], Stefanie Shattuck-Hufnagel[c], Géza Németh[a] and Véronique Aubergé[b]

[a]Dept of Telecommunications and Media Informatics, Budapest Univ of Technology and Economics, Magyar Tudósok krt. 2, H-1117 Budapest, Hungary
[b]GIPSA-lab, Speech & Cognition Dept (ICP), INPG/UJF/Stendhal/CNRS UMR 5216, Université Stendhal, 1180 av. Centrale, 38040 Grenoble Cedex 9, France
[c]Massachusetts Institute of Technology, Research Laboratory of Electronics, Speech Communication Group, 77 Massachusetts Ave., Cambridge, MA 02139, USA
bohm@tmit.bme.hu

Irregular phonation can serve as a cue to segmental contrasts and prosodic structure as well as to the affective state and identity of the speaker. Thus algorithms for transforming between voice qualities, such as regular and irregular phonation, may contribute to building more natural sounding, expressive and personalized speech synthesizers. We describe a semi-automatic transformation method that introduces utterance-final irregular pitch periods into a modal speech signal by scaling the amplitude of the individual cycles. First the periods are separated by windowing, then multiplied by appropriately chosen scaling factors, and finally overlapped and added. Thus, amplitude irregularities are introduced via boosting or attenuating selected cycles. The abrupt, substantial changes in cycle lengths that are characteristic of naturally-occurring irregular phonation can be achieved by removing (scaling to zero) one or more consecutive periods. A freely available graphical tool has been developed for copying 'stylized' pulse patterns (glottal pulse spacings and amplitudes) from an irregular recording to a regular one. It allows the interactive refinement of the scaling factors and waveform regeneration. We examine the effects of the transformation on harmonic structure, and present perceptual test results showing that transformed signals are similar to natural irregular recordings in both roughness and naturalness.

# 1 Introduction

Human speakers show a wide variety of voice qualities across utterances and even within a given utterance. A common manifestation of this variation is often termed irregular phonation. In contrast to regular, quasi-periodic phonation, a region of voiced speech produced with irregular phonation displays either "an unusual difference in time or amplitude over adjacent pitch periods (exceeding normal ranges for jitter and shimmer), or an unusually wide-spacing between pitch periods for a given speaker" [1]. By specifying the term "unusual", we focus on intermittent irregularity that frequently occurs in many voices. Irregular phonation has also been referred to as glottalization, creaky voice, vocal fry, and laryngealization, and its perception is often characterized as 'rough voice'. Figure 1 shows an example for the difference between regular (1a) and irregular (1c) phonation.

Irregular phonation in normal speech may have a number of communicative functions: it can serve as a cue to underlying silence in certain languages (related to either a stop consonant or a prosodic boundary) [1] as well as to the affective state [2] and identity [3] of the speaker. As we come to understand more about the contexts in which this intermittent voice quality occurs, algorithms for transforming between regular and irregular phonation may contribute to building more natural sounding, expressive and personalized speech synthesizers.

Earlier work in this direction was either in the formant synthesis domain [4-6] or relied solely on increasing jitter and shimmer (the period-to-period variation in the glottal pulse spacings and amplitudes) in the speech signal [7-9]. Both of these approaches have major drawbacks. On the one hand, in formant synthesis, the naturalness of automatically generated speech is not yet satisfactory for many applications, while setting the synthesizer parameters manually is a laborious task. On the other hand, manipulation of jitter and shimmer often does not predict perceived voice quality [10]; as a result, ignoring other acoustic characteristics of irregular phonation may prevent the transformed speech from sounding naturally rough. Thus the development of a simple semi-automatic transformation method that introduces irregular pitch periods into a modal speech signal while avoiding some of the limitations of earlier approaches would be a useful contribution.

In Section 2 of this paper, we describe such a method, one that relies on amplitude scaling of individual glottal cycles to create natural-sounding utterance-final irregularity. In order to mimic a natural utterance-final irregular pulse pattern, some of the fundamental periods are scaled to zero (i.e. removed from the waveform) and some others are either attenuated or boosted. The scaling factors can either be individually hand-selected or copied as a pattern from an utterance with irregular phonation. In Section 3 we present the results of a perceptual experiment evaluating the naturalness and roughness of the transformed stimuli. In Section 4 we discuss the acoustic characteristics of these transformed speech tokens, and in Section 5 we describe a freely available graphical program, developed by the first two authors, that implements the transformation.

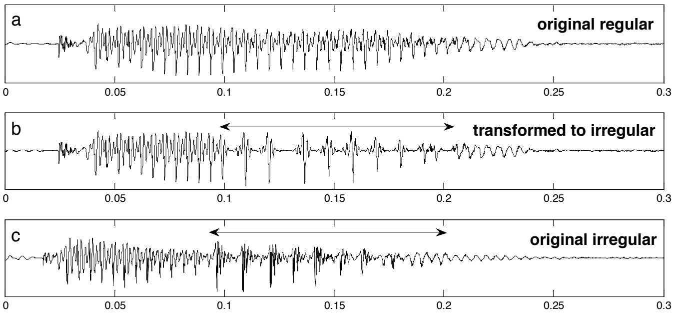


Fig. 1. A speech recording with a modal ending (a) and its transformed version (b). An originally-irregular recording is also shown (c). Horizontal arrows mark irregular regions.

# 2 Transformation method

The input to the transformation method is the speech waveform with markers for the approximate times of glottal closures (pitch marks) in the region to be transformed (see Fig. 4 for an example). As a first step, the glottal cycles are separated by applying a Hanning window in the vicinity of each glottal pulse. The peak of the window is positioned on the pitch mark and it spans from the previous to the next pitch mark. Thus it covers two fundamental periods and, if the two periods are of different lengths, it may be asymmetric. This procedure is the same as the analysis stage of the PSOLA algorithm [11] and it extracts rough estimates of the individual cycles into separate waveforms.

The sample amplitude values in each of these waveforms are then multiplied by a hand-selected scaling factor ($s_i$) and overlapped-and-added to re-synthesize the signal. The scaling factors can either boost ($s_i$>1), attenuate ($s_i$<1), remove ($s_i$=0) or leave unmodified ($s_i$=1) the individual cycles. In regions of the speech waveform where all the scaling factors are set to one, the original signal is reconstructed (apart from rounding errors), so any possible artifacts are limited to the amplitude-manipulated regions of the speech. See Fig. 1b for an example of a transformed speech waveform.

Note that the method does not move the cycles in time (which the PSOLA algorithm does, in order to change F0). In contrast to PSOLA, where the aim is to implement fine adjustments of the fundamental period, here we need abrupt, substantial changes in the glottal pulse spacings, as observed in naturally-occurring irregular phonation. We claim that this can be achieved by removing one or two consecutive cycles (and thus doubling or tripling that specific fundamental period), without the need for fine control over pulse positions.

Attenuating or zeroing a fundamental period also scales down the background noise present during the period. For example, if several consecutive cycles are removed from a recording with audible background noise, the lack of noise in the transformed region might decrease the perceived naturalness. In order to avoid this problem, background noise can be optionally added to attenuated and zeroed cycles. The noise can be windowed out from e.g. the end of the recording and scaled by 1-$s_i$ in order to compensate for the noise energy loss due to scaling. To avoid introducing periodic noise patterns, the noise signal may be reversed or inverted in successive windows (the choice between these manipulations is made randomly).

## 2.1 Setting the scaling factors

To transform a modal recording so that it is perceived as rough, one should create a pulse pattern (glottal pulse spacings and amplitudes) characteristic of natural irregular phonation. To reach this goal, the scaling factors can be modeled after a sample region of natural speech with irregular pitch periods. The factors needed to approximately match the irregular pulse pattern in that sample can be either set by hand in a trial-and-error procedure or 'copied' as a pattern from the model recording.

When setting the factors manually, the following principles should be considered:

- If a glottal cycle is substantially longer in the irregular recording than in the modal one (e.g. two or three times as long), then one or two cycles at the corresponding location in the regular waveform should be zeroed out. Since a naturally-occurring irregular cycle length is not always an integer multiple of the corresponding regular cycle length, this method of period removal usually cannot match the exact length of the irregular cycles, but the results of the perceptual evaluation (see later) suggest that this imprecision is not perceptually critical. The abrupt, substantial cycle length changes introduced by the transformation seem to be sufficient to achieve a rough-sounding voice quality, while the exact lengths of the cycles are apparently of less importance.
- The relative amplitudes of the irregular pulses in the sample should also be reproduced, in order to mimic amplitude irregularities.
- The transformation can be iteratively fine-tuned by removing more or fewer cycles and adjusting the scaling factors until the transformed speech is judged to be both natural sounding and a perceptually salient example of utterance-final irregularity.

When the scaling factors are set by pattern copying, one has to select both the regular region to be manipulated in the signal and the irregular region to be copied. Then a 'stylized' pulse pattern is extracted from the selected irregular region, consisting of the scaling factors to be used in transforming the regular region (i.e. not the absolute pulse positions and amplitudes). This stylized pulse pattern is initially constructed as a vector containing the relative amplitudes of the glottal pulses in the sample irregular region. The amplitude of each period is measured as the peak amplitude (either positive or negative) around the pitch mark. The values in the stylized pulse pattern are expressed relative to the mean amplitude of some regular periods preceding the irregular region.

When an irregular cycle is substantially longer than a reference cycle length (e.g. two or three times or more than the reference, $T0_{ref}$, that is calculated as the mean of some preceding regular cycles), zeros are inserted in the stylized pulse pattern since, at these points, periods need to be removed from the regular recording. The number of zeros to be inserted between two consecutive scaling values (i.e. the number of periods to be removed) is determined by the rounded ratio of the actual cycle length to the reference: $n_i=round(T0_i/T0_{ref})$. Cycle lengths are measured as time differences between consecutive pitch marks. The number of periods used to calculate the reference cycle length and reference amplitude is 5 by default, but can be set as a parameter. Fig. 4 shows an example of copying a pulse pattern by means of the graphical program.

# 3 Perceptual evaluation

A major factor in determining the practical usefulness of the proposed transformation method is its acceptability to human listeners. This was measured in a listening test that aimed to evaluate the degree of perceived roughness and any degradation in naturalness introduced in the speech signal by the method. As the perception of irregular phonation is usually described as rough voice, measuring perceived roughness allows us to assess whether the transformed recordings sound like irregular phonation. Apart from roughness, a significant degradation in naturalness would signal audible artifacts introduced by the method. Such artifacts might appear, for example, due to the fact that the impulse responses induced by consecutive glottal pulses usually overlap, so it is not possible to achieve their complete separation by windowing.

In the evaluation we focused on irregular phonation in utterance-final position, because irregularities are very likely to occur in this position, and usually span a considerable region of speech, possibly making their occurrence perceptually more salient. We have not yet attempted to create non-final irregularities, but it may also be possible to do so using the same method. We compared the results of transformed speech samples to unmanipulated, natural utterances, both those with regular and those with irregular phonation at the end.

## 3.1 Methods

To create the natural speech stimuli, we recorded four American English speakers uttering four words and short phrases (*ready*, *Debby*, *yesterday*, and *your paper*) at a 16 kHz sampling rate. Two of the speakers frequently produced utterance-final irregular phonation and were able to utter the words in both voice qualities on request. This allowed us to obtain pairs of recordings of each word – one with a regular and the other with an irregular ending. The other two speakers seldom produced glottalization in these recordings. They were unable to consistently produce the words with irregular phonation, so we could obtain only regular recordings from them. Thus in total there were 16 natural recordings with regular endings and 8 with irregular endings.

To create the transformed speech stimuli, we used our transformation method to convert the endings of all 16 regular recordings into irregular; the waveform of one such stimulus can be seen on Fig. 1b. Pitch marks were automatically placed on negative peaks by Praat and hand-corrected where necessary. We set the scaling factors manually, with the goal of matching the irregular counterpart, i.e. the same word uttered by the same speaker with irregular final phonation, where possible. The transformation was iteratively fine-tuned (for about 15 minutes on average) by removing more or fewer periods and adjusting the amplitude scaling factors until the first author judged it to be a perceptually salient example of final irregularity.

The stimulus set also included formant synthetic stimuli (with the same linguistic content) with both regular and irregular final phonation and irregularly-ended speech samples transformed to regular by concatenating a natural regular ending. These stimulus types were used for another study [12] and will be discussed only briefly here.

The perceptual experiment consisted of two tests. In one, listeners rated the naturalness of the speech samples, while in the other they judged the roughness of the same samples. For both tests, responses were given on a 5-point scale using a mouse to click on a button. The endpoints of the roughness scale were labeled on a monitor screen as 'not rough at all' (1) and 'very rough' (5), while the extremes of the naturalness scale were denoted as 'very unnatural' (1) and 'very natural' (5). Before starting the naturalness test, the entire stimulus set was played for a listener, to demonstrate the range of naturalness that would be encountered during the test. Before the roughness test, listeners heard some examples of natural speech (not used in the test) both with and without irregular pitch periods, to clarify the meaning of the term 'rough'.

All the stimuli were rated twice in both tests. Presentation order was re-randomized for each listener and each test, with the order of the two tests counterbalanced across listeners. The 12 listeners who participated in the experiment were all native speakers of American English and were not familiar with the talkers' voices. In addition, as described below, we partially analyzed the data for 5 pilot listeners.

## 3.2 Results

The first five pilot subjects judged the irregularly-ending natural utterances as having about the same (or even less) roughness as the ones with modal endings. This surprising result may have arisen because of misunderstanding the task. In further sessions involving 12 new listeners, we attempted to clarify the term 'rough' more thoroughly during a short conversation before the experiment, and this change in the method was effective: these 12 subjects rated the natural irregular stimuli to be 0.98 scale points rougher on average than the natural regular tokens.

For these 12 listeners, a one-way ANOVA for naturalness ratings and a separate one for roughness ratings both showed a significant effect of stimulus type ($F=442.4$, $p<0.0005$ and $F=87.4$, $p<0.0005$, respectively). Average scores for each stimulus type can be seen in Fig. 2. All differences reported below are significant at the 5% level, according to Tukey's post hoc tests, unless noted otherwise.

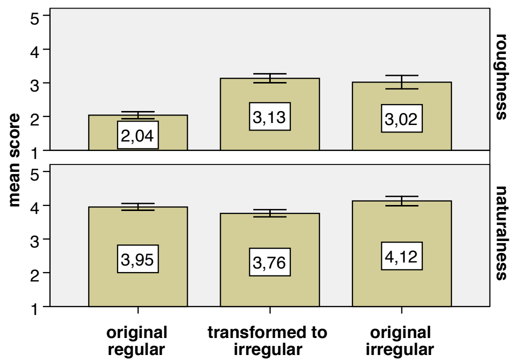


Fig. 2. Mean scores with 95% confidence intervals

Comparing mean ratings across the three conditions can test whether the method created natural sounding roughness from modal stimuli. When a natural utterance with regular utterance-final phonation was transformed to irregular voice with the proposed method the roughness ratings increased by 1.09 point ($p<0.0005$), but this transformation caused only a non-significant, 0.19 point decrease in naturalness scores ($p=0.11$, not significant). Not only did the transformation substantially increase perceived roughness, but this increase matches the roughness of natural irregularly-phonated speech: the difference in the mean ratings was only 0.11 points ($p=0.91$, n. s.). Thus, listeners perceived the increased roughness of the transformed utterances, considered them natural, and heard no difference in the degree of roughness between the originally-irregular and transformed-irregular stimuli.

The mean naturalness ratings for the regular and irregular formant-synthetic stimuli were 1.68 and 1.75, while their roughness was judged to be 3.41 and 3.77, respectively. The speech samples transformed from irregular to regular received an average roughness score of 2.58 and naturalness score of 4.05.

Taken together, these results show that the proposed transformation method successfully mimicked utterance-final irregularity. Unlike formant synthesis, this less labor-intensive method did not introduce artifacts that significantly distorted the signal, and the transformed speech sounded just as rough as a naturally-occurring irregular voice.

# 4 Acoustic evaluation

Considering how simple this transformation method is, one may ask how it could produce such promising results in the perceptual evaluation. It seems that the method captures some perceptually-critical acoustic characteristics of irregular phonation. Although the perceptual structure of irregular phonation is still unclear, this voice quality has a number of acoustic correlates consistently reported in the literature. If the transformed utterances match most of these correlates well, then that might provide a plausible explanation for their perceptual acceptability. In this section, we summarize the acoustic characteristics of irregular phonation, speculate on how well the transformed utterances may match these characteristics, and test some of these speculations by acoustical measurements.

Irregular phonation, compared to regular, is usually characterized by the following properties (detailed explanations can be found in [4] and [13]):

1. The time elapsed between successive glottal pulses is longer and substantially more irregular, i.e. lower F0 and higher jitter.
2. The overall vocal intensity is lower.
3. The proportion of the glottal cycle where the glottis is open, i.e. the open quotient OQ, is lower.
4. More acoustic losses at the glottis, i.e. the first formant bandwidth B1 is increased.
5. The closure of the vocal folds is more abrupt, i.e. the spectral tilt TL is lower.

It is clear that, because it involves the removal of individual pitch periods, the transformation method substantially increases and perturbs the spacing of the glottal pulses, and also decreases the overall intensity level. Due to period removal, the length of the cycles is doubled or tripled, but the open phase keeps its original duration. Thus we can expect that transformed recordings are perceived by listeners similarly to recordings with a decreased OQ. When removing a period, the damping of the previous period may become quicker due to applying a tapered window, resembling to the effect of lowering B1. It is unclear how TL is affected by the transformation. In order to check these expectations, we conducted a set of acoustic measurements focusing on these last three parameters.

## 4.1 Methods

Although all of the three voice source characteristics investigated here can be interpreted as time domain parameters, it is more convenient to consider them in the frequency domain. It has been shown [14] that OQ is proportional to the first harmonic amplitude relative to the second harmonic amplitude (i.e. the difference H1-H2 in dBs), that B1 is correlated with H1 relative to the first formant amplitude (H1-A1), and that TL is related to H1 relative to the third formant amplitude (H1-A3).

Thus by measuring changes in H1-H2, H1-A1 and H1-A3, we can draw conclusions about the three parameters in question. Because these amplitude values can be biased by the effects of the formants, we used the equations described in [15] to calculate the value of H1 and H2 corrected for the effect of the first formant frequency (H1*, H2*), and the value of A3 corrected for the first and second formant frequency and the third formant bandwidth (A3*).

We conducted these measurements on the stimuli used in the perceptual evaluation, i.e. original regular and irregular recordings and the former ones transformed to irregular. We first resampled the wave files to 8 kHz. Then, in the final region of each utterance, three points (roughly uniformly spaced, aligned with the pitch marks) were selected. The 512-point FFT magnitude spectrum, calculated using a Hanning window, was displayed at these locations and the parameters were graphically measured. In many cases, strong subharmonics appeared (due to highly irregular phonation) that were hard to distinguish from the original harmonics. In these cases, we considered the lowest two of all these spectral peaks to be the first and second harmonics.

## 4.2 Results

In one-way ANOVAs, stimulus type had a significant effect on all the three measured parameters (Fig. 3; $p<0.0005$). The mean parameter values of each stimulus type were compared by Tukey's post hoc tests at the 5% level.

H1*-H2* of the transformed recordings was significantly lower than that of the original speech samples ($p<0.0005$) and was approximately the same as the mean value of the original irregulars ($p=0.97$, n.s.). Thus, in terms of the open quotient, the transformed utterances closely matched the values for natural irregular phonation. H1*-A3* (and correspondingly spectral tilt) also decreased significantly due to the transformation ($p=0.001$), moving toward the mean value for irregular speech but still being different from it ($p=0.033$).

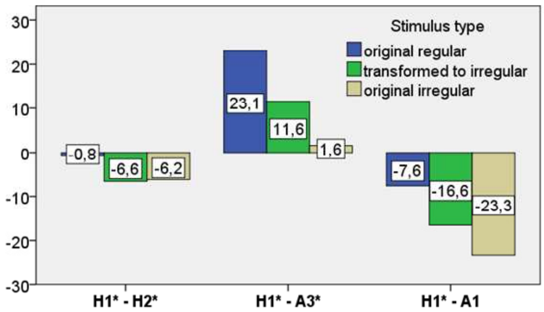


Fig. 3. Averages of the three spectral parameters (in dB)

B1, measured as H1*-A3, showed significantly lower values after the transformation than before ($p<0.0005$). Although this result is somewhat unexpected (it is contrary to our speculations), even lower B1 values are measured for original irregular recordings ($p=0.001$). The reason behind this may lie in the variation of H1 values. H1*-A3 is intended to measure variation in A3, with H1* serving as the basis of comparison. But in many irregular recordings, a subharmonic was chosen as H1 that typically has lower amplitude than the first harmonic in regular recordings. Thus the decrease in H1* might have been larger than the decrease in A3, resulting in a first formant bandwidth mistakenly measured as narrower than in regular speech.

We can conclude that, besides lowering F0 and overall energy and introducing jitter, the proposed transformation method also reproduces some additional acoustical correlates of perceived roughness, such as a decreased open quotient and spectral tilt. This represents a potential improvement over earlier manipulation methods that aimed to increase the roughness of a speech signal by focusing exclusively on jitter and shimmer [7-9].

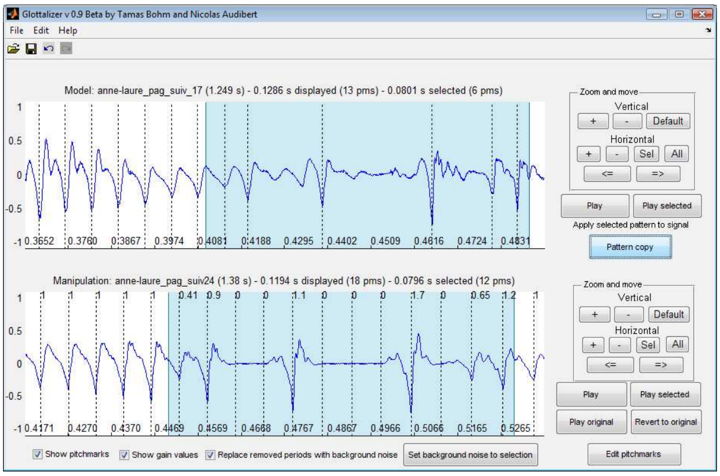


Fig. 4. The screen of the program after copying the pulse pattern in the selected region of the model recording to the other waveform; scaling factors are shown above the pitch marks (dashed lines)

# 5 Graphical program

A graphical tool named *Glottalizer* has been developed to allow fast and convenient application of the transformation method. It runs under Windows and it is freely available for non-commercial use[1]. It provides means for a) the parallel display of both the waveform to be modified and the model waveform; b) copying stylized pulse patterns; and c) convenient iterative refinement of the scaling factors, because the effects of the parameter changes are immediately visible and audible. The program also has the usual sound displaying and playing functionalities as well as a command history.

Fig. 4 shows a screenshot of the program in operation. The bottom panel displays the waveform of the recording to be manipulated. The top panel depicts the model waveform that can be used to guide the transformation (either manually or by copying its pulse pattern); for creating irregular phonation, a model recording that contains irregular phonation can be loaded into this panel. Note that the model recording cannot be manipulated. In order to open a wave file in either one of the two panels, a corresponding pitch mark file must also be available (e.g. in Praat .PointProcess format). The pitch marks can be overlaid on the waveform and can be edited and saved.

In the bottom panel, individual periods can be scaled, removed (scaled to zero) and reverted to their original form (i.e. resetting the scaling factor to 1) by simple mouse clicks. The applied scaling factors are shown above the manipulated waveform, and can be saved in a separate file that can be reloaded later.

The transformation can also be carried out by copying a stylized pulse pattern. In order to do this, one has to select the region in the model waveform that the target pulse pattern is to be extracted from, and the region in the bottom panel where the pattern is to be applied. To enable pattern copy (Fig. 4), there should also be enough pitch marks preceding the model selection to calculate the reference values.

[1] http://www.bohm.hu/glottalizer.html

# 6 Conclusions

We have presented a simple method for transforming utterance-final modal voice into irregular voice by windowing out the fundamental periods and individually amplitude-scaling them. The scaling factors can be set either one-by-one or 'copied' as a pattern from irregularly phonated speech. A freely available graphical program was developed to speed up the transformation process.

Results of the evaluations presented here illustrate that this transformation method reproduces most of the well-known acoustic characteristics of irregular phonation, and that listeners perceive the output to be acceptable as rough, natural-sounding speech. We believe that this algorithm will make it possible to generate experimental stimuli to test the perceptual effects of irregular pitch periods in a number of domains. Other applications, e.g. in speech synthesis, would require automatic setting or copying of the scaling factors. Preliminary results with copying stylized pulse patterns are promising. Such an automated procedure can serve as a tool for synthesizing expressive speech or the voices of different individual speakers.

# References

[1] J. Slifka, "Irregular phonation and its preferred role as cue to silence in phonological systems", *Proc. ICPhS 2007*, 229-232 (2007)

[2] C. Gobl, A. Ní Chasaide, "The role of voice quality in communicating emotion, mood and attitude", *Speech Comm.* 40, 189-212 (2003)

[3] T. Bőhm, S. Shattuck-Hufnagel, "Listeners recognize speakers' habitual utterance-final voice quality", *Proc. ParaLing07*, 29-34 (2007)

[4] D.H. Klatt, L.C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers", *JASA* 87, 820-857 (1990)

[5] J.B. Pierrehumbert, S. Frisch, "Synthesizing Allophonic Glottalization", in *Progress in Speech Synthesis*, J.P.H. van Santen et al. (eds) New York: Springer, 9-26. (1997)

[6] M. Edgington, "Investigating the limitations of concatenative synthesis", *Proc. Eurospeech97*, 593-596 (1997)

[7] A. Verma, A. Kumar, "Introducing rougness in individuality transformation through jitter modeling and modification", *Proc. ICASSP2005*, 5-8 (2005)

[8] A.V. McCree, T.P. Barnwell, "A mixed excitation LPC vocoder model for low bit rate speech coding", *IEEE Tr. Speech and Audio Proc.* 3, 242-249 (1995)

[9] A. Loscos, J. Bonada, "Emulating rough and growl voice in the spectral domain", *Proc. DAFx'04*, 2004.

[10] J. Kreiman, B.R. Gerratt, "Perception of aperiodicity in pathological voice", *JASA* 117, 2201-2211 (2005)

[11] E. Moulines, F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Speech Comm.* 9, 453-467 (1990)

[12] T. Bőhm, S. Shattuck-Hufnagel, "Does a speaker's habitual utterance-final voice quality help listeners recognize the voice?", submitted (2008)

[13] C.G. Henton, A. Bladon, "Creak as a sociophonetic marker", in *Language, Speech and Mind*, L.M. Hyman and C.N. Li (eds) London: Routledge, 3-29 (1987)

[14] E.B. Holmberg, R.E. Hillman, J.S. Perkell, P. Guiod, S. Goldman, "Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice", *JSHR* 38, 1212-1223 (1995)

[15] H.M. Hanson, "Glottal characteristics of female speakers: Acoustic correlates", *JASA* 101, 466-481 (1997).