Voice Enhancement of Male Speakers with Laryngeal Neoplasm

Martin Hagmüller, Gernot Kubin

Signal Processing and Speech Communication Laboratory
Graz University of Technology, Austria

Abstract

In this paper an approach is presented which is aimed to enhance disordered male voices. This approach aims at highpitched voices with a severe degree of hoarseness. The goal of the study is to determine whether pitch modification combined with periodicity enhancement can improve the perceived quality of a disordered speech utterance. Signal manipulation is done pitch-synchronously, so, firstly pitch marks have to be detected. Then period enhancement is performed followed by a PSOLA based pitch modification step. Finally, the period enhancement is performed once more for the utterance with the lower pitch. Perceptual evaluations were performed by both professional speech and language pathologists and naive listeners to rate the subjective perceived enhancement of the voice. Results show that the modified voice has a reduced breathiness, whereas roughness seems not to be influenced by the processing. The most significant result from the naive listener test is a reduced perceived speaking effort.

1. Introduction

A pathological voice can be a major handicap for social interaction. Problems can arise due to bad intelligibility especially in situations with higher vocal demands, such as a classroom situation or other places with high background noise. Speakers can also suffer from short speaking duration due to a quickly strained voice.

So far, only few attempts have been made to enhance dysphonic speech. Some approaches for (tracheo-) esophageal speech have been published (e.g. [1] using linear prediction analysis and a synthetic source for resynthesis) as well as reducing the breathiness in disordered speech using singular value decomposition [2] or state-space approaches [3].

We work on a method to use speech processing techniques to produce a better voice in terms of intelligibility and naturalness.

Possible applications for this method are e.g. a portable electronic device which helps the patient to cope with acoustically difficult situations in everyday life. For example for voice telephony there is almost no other possibility for augmentative communication. It could also support speech therapy, to take pressure from the voice until the patient has reached the goal of using new patterns for his voice production. Another possible application could serve as a motivation for the patient to give pressure from the voice until the patient has reached the goal of using new patterns for his voice production. Another possible application could serve as a motivation for the patient to give

In this study we investigate a special type of pathology (laryngeal neoplasm) which has two major audible characteristics. First, the voice is at a much too high pitch level, and second it has a rather high breathiness.

The paper is organized as follows. In section 2 the used speech data is described, followed by the presentation of the used algorithm in section 3. In section 4 the evaluation of the voice modification is discussed, followed by a conclusion (section 5).

2. Speech Data

The speech data was recorded at the phoniatric department at Graz Ear, Nose & Throat University Clinic. The recordings were performed in a quiet room, and were sampled at 44.1kHz. Detailed description can be found in [4].

We selected four subjects which were all male speakers with diagnosis of laryngeal neoplasm. Two professional speech and language pathologists (SLP) agreed on the rating of the four subjects (table 1). The rating is done using the main rating system for dysphonic speech in German speaking countries, the RBH (Roughness, Breathiness, Hoarseness) scale [5]. It is a compact version of the GRBAS scale [6], where the hoarseness term is used as the general disorder rating. Very often the voice quality is evaluated using utterances of sustained vowels. Since this is not a natural way of speaking and, therefore not the best choice to judge voice quality, a variety of German test phrases was used for the evaluation. The set of chosen utterances was:

- sustained vowels (‘i e a o u’)
- vowel-consonant repetitions (‘dada nini soso lolo’)
- isolated words (‘Maat Wanne Miete Minne Rose Wolle’)
- numbers (‘eins zwei drei vier fünf sechs sieben acht neun zehn’)
- days of the week (‘Montag Dienstag Mittwoch Donnerstag Freitag Samstag Sonntag’)
- short sentences (‘Nie und nimmer nimmt er Vernunft an”, “Eine Maus saust aus dem Haus’)

3. Algorithm

Table 1: Unprocessed patient data: evaluation using RBH (Roughness, Breathiness, Hoarseness) rating (0 no disorder ... 3 severe disorder).

<table>
<thead>
<tr>
<th>ID</th>
<th>age</th>
<th>R</th>
<th>B</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>57</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>B</td>
<td>52</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>C</td>
<td>49</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>D</td>
<td>71</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Figure 1 shows an overview of the performed steps of the algorithm. First the pitch marks are determined, then the pitch is lowered and as post-processing a periodicity enhancement algorithm is computed. Lack of periodicity is a common feature of dysphonic speech, so the periodicity enhancement step is performed to reduce the breathiness in the voice. Since breathiness is usually perceived noise like, one would think noise reduction
algorithms would be suitable to enhance this kind of voice disorder. Breathiness in hoarse voice is usually highly correlated with the speech signal and therefore the required condition of independence of signal and noise is not satisfied. Pitch modification is done to lower the voice of the subjects since for male voices they are much too high pitched. After the PSOLA pitch modification, the periodicity enhancement is performed again to reduce processing artefacts introduced by PSOLA.

### 3.1. Pitch Mark Determination

Both the pitch modification (PSOLA) and the periodicity enhancement work pitch synchronously, so the first step is to determine every pitch cycle correctly. Due to PSOLA pitch modification, the best performance can be gained when the pitch marks are set at the energy maximum of a cycle. This step is carried out using the “Praat” (a software package for speech processing [7]) pitch mark determination function. For sections which were considered unvoiced by ‘Praat’, pitch marks at a chosen period are set. To avoid an unwanted periodicity, a jitter of 10\% has been superposed on the calculated pitch mark in unvoiced regions.

### 3.2. Pitch-synchronous overlap-add

The time-domain pitch-synchronous overlap-add algorithm [8] cuts the signal into overlapping windowed frames of single pitch cycles and then rearranges the cycles according to a specified new pitch contour (fig. 2). The newly arranged signal pieces are then added together to get the modified signal.

The new pitch contour is simply calculated by multiplying the pitch marks with a factor $\alpha$:

$$P_{\text{new}} = P \alpha,$$

where the vector $P$ contains the time instances of the pitch marks and $\alpha$ is the pitch modification factor. Resynthesis is done by choosing the pitch cycle with a pitch mark close to the target pitch mark. In case of lowering the fundamental frequency some pitch cycles drop out, in case of raising the pitch some have to be used twice. Since this algorithm operates at pitch-cycle level, the correct identification of the energy maximum of each cycle is a crucial point for the performance of the algorithm.

### 3.3. Periodicity enhancement

Recently, Kleijn [9] proposed a method for periodicity enhancement for voiced speech at the output of a digital speech coder. This enhancement is performed under two constraints.

First, a signal-norm (energy) preservation condition is given,

$$||\tilde{x}_j|| = ||x_j||,$$

where $\cdot, \cdot$ denotes the Euclidean norm, and $\tilde{x}_j$ is the sample sequence that replaces the signal frame $x_j$ for enhancement. The second constraint is to keep the difference between $\tilde{x}_j$ and $x_j$ relatively small (modification constraint):

$$||x_j - \tilde{x}_j||^2 \leq \beta ||x_j||^2,$$

where $\beta \in [0,1]$. This constraint prevents that the enhancement procedure modifies the signal more than desirable.

The periodicity of the enhanced signal can be measured by

$$\eta_\mathcal{J} = \sum_{j \in \mathcal{J}} \sum_{m \in \mathbb{Z} - \{0\}} \alpha_m \langle \tilde{x}_j, \tilde{x}_{j,m} \rangle,$$

where $\alpha_m$ is a discrete window function, $\mathcal{I}$ is a set of integers that describes the support of this window and $\langle \cdot, \cdot \rangle$ is the Euclidean inner product and $\mathcal{J}$ is a set of block indices, which are centered around consecutive pitch marks. Kleijn’s algorithm maximizes the periodicity criterion (4) under the constraints (2) and (3).

### 3.4. Results

The spectrum and the waveform of the original and the modified signal are compared in fig. 3 and fig. 4, respectively. One can see the influence of the period modification in the modified distance between the signal harmonics in the spectral representation and the increased length of the pitch period in the waveform representation. The periodicity enhancement can be seen in the clarification of the signal harmonics in the spectrum and reduction of the signal noise in the signal waveform. Fig. 5 shows the original and the lowered pitch contour.

### 4. Subjective Evaluation

#### 4.1. Preliminary Considerations

The evaluation of this kind of disordered speech enhancement has to be done in three different ways.

1. Of course, professional speech and language pathologists should judge the original versus the modified voice to assess the modification of the voice from a medical point of view.
2. Naive listeners should be included because the results may not necessarily be of medical relevance, but have to be seen in the context of everyday life.

3. The patients should express their opinion about their own modified voice, whether they feel comfortable with this new voice and could imagine to use a device which modifies their own voice.

For the present study, the subjective evaluation was performed only by trained speech-language pathologists and naive listeners.

Since the speech utterances are modified by digital signal processing algorithms the usual evaluation methods for dysphonic speech are not sufficient to adequately describe the effect of the enhancement algorithms. Therefore, in addition to the RBH rating scale, other evaluation attributes (including such as recommended for telephone transmission quality or synthetic speech [10, 11]) are used. The speaking effort parameter was included due to a suggestion of an SLP.

For the comparison of the modified with the original speech sample the following features were questioned:

**Naturalness:** How would you judge the naturalness of the utterance?

**Listening effort:** How much effort does it take to listen to and understand the utterance?

**Speaking effort:** How much effort do you assume does the speaker need for this utterance?

**Noise:** What about noise or artefacts either due to the voice disorder or signal processing?

**Acceptability:** Is the sound of the voice acceptable to you?

**General opinion:** What is the general opinion of the samples, concerning the overall impression?

All test subjects were native Austrian German speakers. Different utterances were presented to them in the original and the modified version. They could listen to the test examples as often as they wanted. They were asked to rate the difference between the modified and the original utterance according to the following differential scale:

- much worse ...– 3
- worse ...– 2
- slightly worse ...– 1
- about the same ... 0
- slightly better ... 1
- better ... 2
- much better ... 3

4.2. Results

In table 2 the results for the SLP evaluation of the processed speech samples are presented (again two SLPs agreed on the results). The overall hoarseness index and in particular the breathiness of the voices was reduced. The roughness of the voice was not influenced by the processing.

The subjective evaluation was performed with 11 naive listeners. Results for the investigated features averaged over all four speakers are presented in table 3 and figure 6. The features speaking effort and acceptability have the lowest variance and a similar rating. Both might be related to each other since one feels more comfortable listening when a speaker does not seem to be needing a lot of effort to speak. The other features show rather low changes between the processed and unprocessed versions, but high standard deviation.

An interesting observation was that all results for isolated vowels had very high standard deviations, which means that the test persons did not agree on the questioned characteristics. This supports the above mentioned concerns about the use of isolated vowels only. Results for speaker ‘D’ also showed very high standard deviation and mainly negative ratings.

### Table 2: Processed patient data: evaluation by SLP using RBH rating (0 no disorder … 3 severe disorder).

<table>
<thead>
<tr>
<th>ID</th>
<th>age</th>
<th>R</th>
<th>B</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>57</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>52</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>49</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>D</td>
<td>71</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
The noise feature received rather negative ratings, with high variance, though. This points to the fact, that humans are very sensitive concerning processing artefacts. This is an important point for future work.

All used samples of the original and the processed speech utterances can be found at: http://spsc.inw.tugraz.at/hagmueller/icslp04/.

5. Conclusions

Even though the processing did not turn a disordered voice into a healthy voice the characteristics of the voice have definitely been improved. The breathiness rating is lower in all of the processed voices compared to the original speech samples. In addition other subjective features have improved. The modified speech was found much more relaxed than the original. The algorithm is very dependent on the actual speaker as seen for speaker ‘D’ with unreliable, negative ratings. But specially voice disorders have very different acoustic outcomes, so the development of a fit-all algorithm is rather challenging.

The acceptance of such a modification of the voice has still to be determined.

The correct determination of pitch marks is crucial for the successful performance of the algorithm. For dysphonic speech, state-of-the-art pitch determination algorithms only offer limited success. Therefore improvement of the pitch mark determination is an important future work. Currently, we are working on pitch mark determination methods using state-space embed-

Table 3: Subjective evaluation of the pathologic speech enhancement. Comparison category rating (-3 ... 0 ... 3) averaged over all speakers before and after processing by naive listeners

<table>
<thead>
<tr>
<th>Feature</th>
<th>Mean</th>
<th>Std. dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naturalness</td>
<td>0.43</td>
<td>1.28</td>
</tr>
<tr>
<td>Listening effort</td>
<td>0.07</td>
<td>1.39</td>
</tr>
<tr>
<td>Speaking effort</td>
<td>0.82</td>
<td>0.81</td>
</tr>
<tr>
<td>Noise</td>
<td>-0.33</td>
<td>1.15</td>
</tr>
<tr>
<td>Acceptability</td>
<td>0.7</td>
<td>1.09</td>
</tr>
<tr>
<td>General opinion</td>
<td>0.09</td>
<td>1.44</td>
</tr>
</tbody>
</table>

6. References