PITCH CONTOUR FROM FORMANTS FOR ALARYNGEAL SPEECH

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

Abstract: People without a larynx have to communicate with a substitute voice, where all possibilities have major shortcomings. A common problem is the lack of a natural pitch contour. This is either because of a constant pitch in the case of an electro-larynx or very limited control over pitch in case of esophageal or tracheo-esophageal speech. To introduce a more natural pitch contour, we propose to use the speech formants as a source to generate an artificial pitch contour. Earlier offline methods to introduce a natural pitch contour to AL speech have shown, that this significantly improves the speech quality. For voice modification, we use the voice pulse model which enables us to place voice pulses at arbitrary positions even when no harmonic voice source is available, while preserving the voice identity of the speaker. Informal perceptual evaluations showed that using a formant based fundamental frequency yields a reasonable pitch contour and is perceived as an improvement for alaryngeal speech.

Keywords: Alaryngeal speech, pitch, formants, enhancement

I. INTRODUCTION

In case of laryngeal cancer at an advanced stage the last possibility to stop the further advancement of the cancer, and therefore save a patients life is to remove the entire larynx. This results in the loss of the usual voice production mechanism, based on vibration of the vocal folds. In addition the trachea is surgically moved to an opening at the neck, called the tracheostoma. As a result, the air is not passing through the vocal tract anymore.

Alaryngeal patients have then to rely on a substitute voice production mechanism. There are three major methods available:

Electro-larynx (EL): A hand held device, which is held against the neck, produces a buzz-like sound and excites the vocal tract. Through usual articulation movements the sound source is modulated and voiced speech sounds can be formed. Unvoiced sounds are produced as in healthy speech by using the air reservoir available in the mouth.

Esophageal voice (ES): Air is swallowed into the esophagus and is released again in a controlled manner. The false vocal folds are then excited and produce the source of the speech sound. In healthy, laryngeal speech, the false vocal folds are not used for phonation, so the patient has to be trained to use this substitute voice source.

Tracheo-esophageal voice (TE): A valve between the trachea and the esophagus is surgically inserted. The valve allows to let the air from the lungs through the vocal tract if the tracheostoma is closed when exhaling. The air through the pharynx excites the false vocal folds and some kind of oscillation is produced.

All three of those voices have major shortcomings. The electro-larynx voice sounds very mechanical, due to the monotonous sound, which is strictly periodic at a constant pitch. The methods which excite the false vocal folds (ES and TE) have very unstable oscillation patterns, which result in a very rough voice. Therefore the fundamental frequency ($F_0$) cannot be reliably extracted by means of digital signal processing methods. Further, the oscillation cannot be controlled very well, which leads to an inconsistent pitch contour. To improve the voice one approach would be to introduce a more natural pitch contour, but this prosodic information has to be derived from feature other than the fundamental frequency, since this is either constant, or not measurable. The esophageal voice also suffers from timing problems, because the amount of air that can be swallowed limits the duration of speech phrases considerably.

This paper will first look into related work concerning alaryngeal voice enhancement and prosody in alaryngeal speech. Then we will present a method to introduce a pitch contour derived from the speech formants.

II. BACKGROUND AND RELATED WORK

A. Alaryngeal speech

Previous approaches were introducing an artificial voicing source as a substitute for the bad voicing source of alaryngeal speech [2], [1]. The voicing source is based on voicing models such as the Liljencrants-Fant (LF) [5] model. A different approach for voicing substitution would be to use prerecorded voice samples of sufficient length to avoid audible loops. Most beneficial for the patient is achievable in case it was possible to do extensive voice recordings prior to surgery and even better prior to the voice degradation, which in most cases is already severe in case of laryngectomy. This would be the best way to preserve the voice identity of the speaker after the operation [7].

In [13] the prosody of alaryngeal speech has been investigated. Experiments have been performed to solve the
question whether alaryngeal speakers are able to convey prosodic information without being able to produce an $F_0$ contour. It was shown that alaryngeal speakers do convey prosodic information, which can be interpreted by a listener. Further, it was investigated which cues communicate the pitch-like information. Features considered were high frequency intensity and spectral tilt. A majority of the alaryngeal speakers were able to convey accent information, without using pitch cues. The study did not show, though, which features are used for this ‘alternative’ pitch.

Meltzer et al. [10] performed a study to find out which type of enhancement brings the most improvement to electrolarynx speech. They investigated combinations of low-frequency boosting, noise-reduction and natural pitch. Listening tests were used to determine the preferred modification method. For the natural pitch experiment the same sentence was uttered by one person with healthy speech, then using the vocal cords and then holding the breath and using an EL. The natural pitch contour was then applied to the EL speech utterance. Earlier Ma et al. [9] have also performed similar experiments. Both experiments show that the substitution of the monotonous EL pitch with a natural pitch contour significantly improves EL speech. For hoarse speech, we have previously shown, that a pitch contour in the expected frequency range a male speaker improves the perceived quality of disordered speech [6].

Recent approaches have used energy as a feature to provide a pitch contour for generating voicing for whispered [11] and esophageal speech [8]. While the energy contour may provide reasonable results for whispered and ES speech, it does not make sense for EL speech. Electro-larynx speech is very limited in expression and energy modulation is not possible or only very limited. Commercially available ELs, if at all, do only provide two intensity positions. A high energy position has to be activated manually by pressing a button. Therefore, other features have to be used to calculate a pitch contour from the speech signal. Taking the formants as a source for the pitch contour seems to be a useful approach.

### B. Radiated Voiced Pulse Modeling

A method to pitch modification is of course the TD-PSOLA approach [12]. While this works for EL speech, it cannot be used for ES and TE speech, since no reliable pitch mark estimation can be performed. If we want a pitch modification system to serve as a framework for alaryngeal speech in general, not only EL speech, we need a different approach. We have chosen the radiated voiced pulse modeling approach by Bonada [4], which is briefly described below.

If the input signal $y[n]$ consists of the sum of $R$ identical input signals $x[n]$ which are delayed by multiples of $\Delta n$,

$$y[n] = x[n] + x[n - \Delta n] + x[n - 2\Delta n] + \ldots + x[n - (R - 1)\Delta n]$$

then, after some calculation, which can be found in [4], we find that

$$Y(e^{j\Omega}) = X(e^{j\Omega})e^{jR\Delta n} \frac{\sin(0.5\Omega \Delta n R)}{\sin(0.5\Omega \Delta n)} \equiv X(e^{j\Omega}) \text{sinc}_R(\Omega \Delta n).$$

The effect of the $\text{sinc}_R$ term is that the spectrum of $X(e^{j\Omega})$ is sampled. If we assume that the $X(e^{j\Omega})$ only varies slowly we can estimate $X'(e^{j\Omega})$ from $Y(e^{j\Omega})$ by interpolating the harmonic peaks (see Figure 1). The full derivation of this assumption including how the phase is dealt with can be found in [4].

So if the harmonic peaks are interpolated and transformed back into the time domain, one can reconstruct the voice pulses, which were filtered by the vocal tract and radiated by the mouth (see Figure 2). The reconstructed pulses can be placed at arbitrary positions, similar to TD-PSOLA [12], while introducing the possibility of complex modifications. Another advantage of the voiced pulse model is, that no pitch marks are needed for the analysis of the signal.

The above method formed the basis of an enhancement approach for esophageal speech [8]. Since in esophageal speech harmonic peaks cannot be reliably determined the spectral envelope is determined by using a bank of constant bandwidth filters. The phase is derived by smoothing, shifting and scaling the magnitude envelope of the spectrum.

The next section will describe the procedure, how the enhancement system for electro-larynx speech is implemented.
III. DESCRIPTION THE ALGORITHM

While the system is intended to be working in real time, at the current stage it is implemented in MatLab, by loading wav files, which are then processed frame-by-frame. The proposed algorithm works with a sampling frequency of 16kHz, so if necessary the sound file is resampled. After a high pass filter which removes DC and very low frequency components, the pitch of the speech utterance is tracked with Praat. Since the pitch is usually constant for EL speech, the processing frame length is fixed to 3 times the pitch period. The pitch tracking can be omitted once the fundamental frequency of the EL is known, or in case of ES or TE speech, where pitch determination will not work reliably.

The following steps are performed frame-wise: The signal is transformed to the spectral domain, by calculating an FFT. The spectral envelope is derived from interpolating the peaks, which in case of EL speech will be the harmonics. The phase is calculated as proposed in [8], by smoothing, scaling and offsetting the interpolated magnitude envelope.

Since for EL speech energy modulation is very limited, the generation of the pitch contour relies on the formants. The formants are tracked with the algorithm provided by the Praat speech software [3]. At the current stage different methods to calculate the pitch contour from the formants have been tried and informally evaluated. The smoothed difference between the 1st and 2nd formant has been chosen to generate the pitch contour:

\[ f_0(t) = \text{smooth}(F_1(t) - F_2(t))/\alpha + \beta, \]

(1)

where \( F_1(t) \) and \( F_2(t) \) are the 1st and second formant and \( \alpha \) and \( \beta \) are constants for offsetting and scaling. They have to be chosen to match the target average pitch and the pitch range of the patient. Voicing is switched on and off with a simple energy based voice activity detector (VAD). The voice pulse model enables us to place the voiced pulse at arbitrary positions, i.e. at the pitch marks derived from the pitch contour determined using Eq. 1.

To avoid the influence of other enhancement methods for the evaluation of the improvement due to a pitch contour, only the pitch was modified. Informal perceptual evaluation has been performed with five listeners using sound samples from EL, ES and TE speech. A clear preference has been indicated for the modified speech utterances using the proposed pitch modification. The negative effect of unexpected pitch movements – within a certain boundary – is compensated by the existence of a reasonable pitch contour at all.

IV. CONCLUSION

One of the major shortcomings of electro-larynx speech is the lack of a normal pitch contour. Previous publications showed that adding a natural pitch contour was the most important modification to improve the perceptual quality of EL speech. We presented an approach that exchanges the monotonous pitch with a more natural \( F_0 \) contour. While it may not necessarily be linguistically correct at all times, it does improve the perceived naturalness of the speech and reduces the impression of robot-like
characteristics especially for EL speech.

**Further work**

Further research on correct prosody has still to be carried out and is expected to yield an improved understanding of how ‘pitch’ accent is conveyed in alaryngeal speech.

A further shortcoming of EL speech is, that there is no appropriate distinction between voiced and unvoiced sounds. At the moment the whole speech utterance is treated as voiced. A step further would be to introduce a distinction between voiced and unvoiced and treating them accordingly. A V/UV classifier is needed which is able to correctly label EL speech. Then unvoiced sounds can be left unmolded.

Further work is necessary to improve the sound quality, while preserving the identity of the voice. This includes noise reduction to suppress the directly radiated noise. This is the energy which is omitted from the EL directly in the air and is not modulated by the vocal tract.

**V. ACKNOWLEDGMENT**

This research has been partially supported by the European COST Action 2103.

**REFERENCES**


