Learning an Artificial F_0-Contour for ALT Speech

Anna Katharina Fuchs, Martin Hagmüller

Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria
anna.fuchs@TUGraz.at, hagmueller@TUGraz.at

Abstract

The Artificial Larynx Transducer (ALT) as a possibility to re-obtain audible speech for people who had to undergo a total laryngectomy has been known for decades. Not only the design and underlying technique but also the poor speech quality and intelligibility have not improved until now. In a world where technology rules everyday life, it is necessary to use the known technology to improve the quality of life for handicapped people.

One reason for the lack of naturalness is the constant vibration of the ALT. A method to substantially improve ALT speech is to introduce a varying fundamental frequency (F_0) - contour. In this paper we present a new method to automatically learn an artificial F_0-contour. The model used is a Gaussian mixture model (GMM) which is trained with a database containing speech of ALT users as well as healthy people. Informal listening tests suggest that this approach is a first step for a subsequent overall enhancement technique for speech produced by an ALT.

Index Terms: alaryngeal speech, Artificial Larynx Transducer (ALT), fundamental frequency, speech enhancement, laryngectomy, GMM

1. Introduction and Related Work

For people who suffer from laryngeal cancer or similar diseases, the last chance is often a total laryngectomy, which means the complete removal of the larynx (see figure 1 – (a) and (b)). According to the source-filter model, the larynx and the containing vocal folds can be seen as the source of sound energy and the vocal tract as the filter. If people lose their larynx they also lose the ability to speak.

There are three alternatives for people to re-obtain their speech. The first method is the *esophageal voice*. Within this method air is first gulped and then released in a controlled manner. Instead of the vocal folds, the tissue in the pharynx vibrates. The second method is the *tracheo-esophageal voice*, where a valve is placed between trachea and esophagus. Due to this valve speech can be produced again with the air from the lungs (see figure 1 (c)). The third method is the *artificial larynx transducer* (ALT). This is a small, hand-held and battery driven device. The device is held against the neck

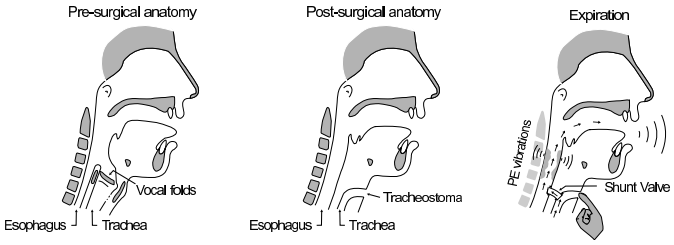

Figure 1: *Schematics of pre-surgical anatomy (a) , post surgical anatomy (b) and tracheo-esophageal voice production (c) (from [1]).*

to produce speech. Then energy is brought into the vocal tract which acts as the filter in the same way as in normal speech production.

Method one needs a lot of practice, and even then, some people are not able to learn it. Moreover, a speech valve needs continuous medical care and this cannot be afforded by people in poorer countries. The ALT device is simple and unproblematic to use and it is the best alternative to speak if the other methods fail. It is also the first post-operative possibility. There is one thing all three methods have in common: the inadequate quality of the resulting speech. Due to the fact that the existing layouts of ALTs have been available for almost six decades and that there has been no improvement of intelligibility and naturalness since then, there is an obvious need for a new generation of ALTs.

The approaches for generating an artificial F_0-contour which have been proposed so far either require manual interaction to change the fundamental frequency or only provide some predefined F_0-contours. There are some commercially available approaches such as using a switch from the standard F_0 level to a different level to mark accentuation. A more flexible approach is to use a pressure sensitive button, which allows a continuous F_0-contour to be produced [2]. Another possibility is to create an artificial F_0-contour from the speech energy envelope [3]. The energy envelope is one possibility to convey accentuation if no F_0 is available. In this paper, the energy contour has been scaled and offset so that the transformed contour matches the average fundamental frequency and the dynamic range of the speaker. This does not produce a linguistically correct intonation,

but it does give a useful F_0 which can be influenced by the speaker by modulating whisper intensity. A similar approach is to use the expiration air-pressure to control the F_0-contour. Uemi *et al.* use an air-pressure sensor that is put on the stoma so that F_0 can be controlled by lung pressure [4] . More recent work was carried out by Nakamura *et al.* [5] who also used the air-pressure sensor to improve F_0 estimation before enhancing ALT speech using a statistical voice conversion technique.

Those approaches require manual control of the F_0-contour and an action by the user, which is hardly done in practice. Therefore, an automatic procedure for adjusting an artificial F_0-contour is a considerable advantage. In this paper, an artificial F_0-contour for ALT speech is estimated using a machine learning approach. The goal of this approach is to improve ALT speech in terms of naturalness and intelligibility by introducing variations in the F_0-contour, which means changing the frequency of the ALT excitation signal. The approach for such a variation needs to be of low complexity in order to implement it on a real-time platform. A strong requirement is that the learning procedure is automatic and that users do not have to actively change speaking parameters. Therefore, our intention is to perform an estimation of the F_0-contour using Gaussian mixture models (GMM).

2. Method: GMM and F_0 Estimation

Mixture models are probabilistic models. They are useful to model arbitrary density distributions. A GMM consists of a linear combination of K multivariate Gaussian probability density functions given by

$$P(\mathbf{x}_l) = \frac{1}{\sqrt{\det(\mathbf{\Sigma})}(2\pi)^{D/2}} e^{\left(-\frac{1}{2}(\mathbf{x}_l-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}_l-\boldsymbol{\mu})\right)}. \tag{1}$$

The mathematical representation is given by

$$P(\mathbf{x}_l|\lambda) = \sum_{m=1}^{K} b_m P_m(\mathbf{x}_l). \tag{2}$$

K is the number of Gaussian components, $\mathbf{x}_l$ is an observation of a D-dimensional random vector, b_m are the weights for each component and P_m is a single, multivariate Gaussian distribution. For b_m, equation

$$\sum_{m=1}^{K} b_m = 1 \tag{3}$$

holds. With $\lambda = \{(b_m, \boldsymbol{\mu}_m, \mathbf{\Sigma}_m); m = 1, 2, \ldots, K\}$, the whole mixture model is described. The parameters λ of a model are estimated using the expectation-maximization (EM) algorithm [6].

The estimation of an artificial F_0-contour using GMMs is inspired by Milner *et al.* [7] who used mel-frequency cepstral coefficients (MFCCs) for prediction of the F_0 and voicing in unconstrained speech. The input for training the GMM is $\mathbf{y} = [\mathbf{x}, f]^T$. x contains the MFCCs for one frame for ALT speech and f the corresponding F_0 for healthy speech. For this approach full covariance matrices $\mathbf{\Sigma}_m$ need to be trained. The mean value vector $\boldsymbol{\mu}_m$ and $\mathbf{\Sigma}_m$ can be seen as a concatenation:

$$\boldsymbol{\mu}_m = \begin{bmatrix} \boldsymbol{\mu}_m^x \\ \boldsymbol{\mu}_m^f \end{bmatrix}, \mathbf{\Sigma}_m = \begin{bmatrix} \mathbf{\Sigma}_m^{xx} & \mathbf{\Sigma}_m^{xf} \\ \mathbf{\Sigma}_m^{fx} & \mathbf{\Sigma}_m^{ff} \end{bmatrix}.$$

In order to estimate F_0, the nearest component to an unknown $\mathbf{x}_{in}$ is calculated: $m^* = \arg\max_K \{p(\mathbf{x}_{in}|c_m^x)b_m\}$, where $p(\mathbf{x}_{in}|c_m^x)$ is the marginal distribution of the MFCC vector for the mth cluster c_m^x. The final result is obtained through

$$\hat{f}_i = \boldsymbol{\mu}_{m^*}^f + \mathbf{\Sigma}_{m^*}^{fx} \cdot (\mathbf{\Sigma}_{m^*}^{xx})^{-1}(\mathbf{x}_{in} - \boldsymbol{\mu}_{m^*}^x). \tag{4}$$

3. Experimental Evaluation

3.1. Speech Database

For experimental evaluation, we built up a database containing speech material from a female as well as a male speaker. The database was recorded in a sound-treated chamber. The chosen speech material includes various types of different sentences including questions, declarations, imperatives and sentences from the newspaper. The two speakers are healthy subjects. They spoke the sentences one time with their natural, healthy voice (HE) and one time with a Servox electronic larynx (ALT). All in all 97 sentences with an average sentence length of 5 seconds were recorded per speaker. This results in a duration of 8 (for HE) - 12 (for ALT) minutes per speaker.

3.2. Experimental Setup

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. Due to this analysis-by-synthesis method, results can be evaluated using informal listening tests. The pre-processing step, as shown in figure 2, resamples the data (HE and ALT speech) to a sampling frequency of f_s =16000 Hz. A high pass filter (HPF) removes DC and very low frequency components and the directly radiated noise is removed from the ALT speech using a modulation filtering technique [8]. In the training step 28 MFCCs (including 0'th order cepstral coefficient, the log energy and delta coefficients) features are extracted from the ALT speech using a hanning window with a frame length of 50 ms (FE). 23 filters in the filterbank are used. The natural F_0 of the HE speech is tracked by the algorithm provided by the Praat speech software [9]. The features and the F_0 values are time aligned using Dynamic Time Warping (DTW) [10] because people need more time to articulate with an ALT than with

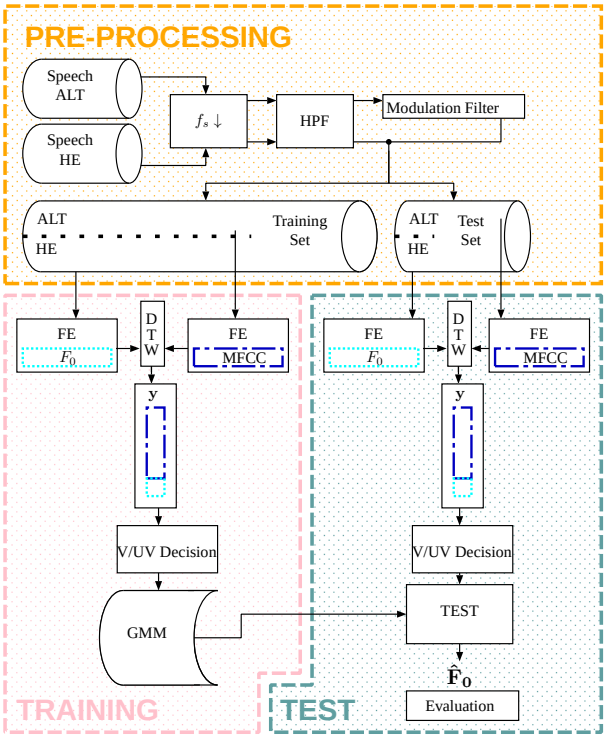

Figure 2: *Block diagram of experimental setup: PRE-PROCESSING with downsampling to $f_s = 16000$ Hz, high pass filter (HPF) and removal of directly radiated noise (modulation filter framework); TRAINING and TEST with feature extraction (FE), dynamic time warping (DTW), concatenation of $\mathbf{y} = [\mathbf{x}, f]^T$ and voiced/unvoiced (V/UV) decision.*

healthy speech. Then, the frame-wise extracted MFCC feature vector concatenated with the F_0 value for the corresponding frame creates feature vectors with the dimension $D = 29$ as described in section 2. Only voiced frames are taken for further processing. Afterwards, GMMs with full covariance matrix and $K = 16$ components are used to train the artificial F_0-contour. The statistical model is speaker- and gender-dependent. In the test scenario, the MFCC features from an unknown ALT input sentence are calculated in the same way as in the training scenario. Afterwards, a voiced/unvoiced decision is carried out and the F_0 value is estimated only for the voiced frames according to the method described in section 2. For evaluation resulting F_0-contours are compared with healthy F_0-contours.

3.3. Results

The best method to evaluate speech enhancement methods is to carry out listening tests. Due to the fact that listening tests are time and cost consuming the results are presented in terms of numerical values. Furthermore, informal perceptual evaluations verified the numerical values and confirmed the improvement of naturalness. Figure 3 shows an example of a male speaker.

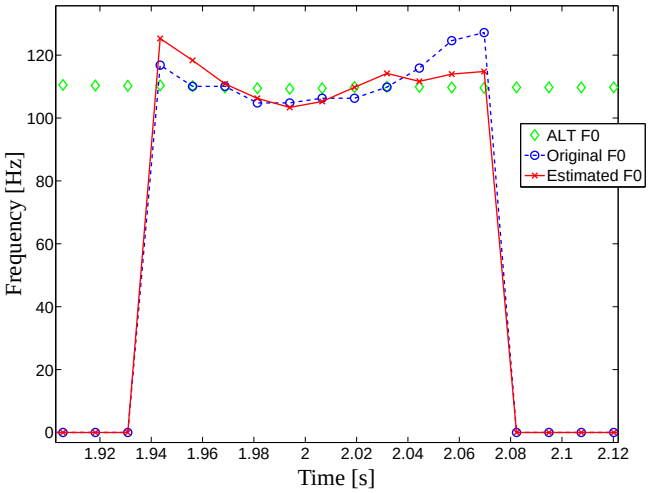

Figure 3: *One voiced phrase of a male speaker: ◊ – original constant F_0 from ALT (without processing); F_0 from healthy speech* (o)*; estimated F_0 (×).*

The correlation coefficient r is calculated using the Matlab function `corr`. r compares the natural F_0-contour and the original F_0-contour for the ALT speech (r_{orig}) on the one hand, and the artificial F_0-contour from the enhanced speech (r_{est}) on the other hand. For one sentence from the male speech database r_{orig} is 0.41 and the enhanced speech reaches a very good value of $r_{est} = 0.91$.

As minor variations from the natural healthy F_0-contour have only little effect on the quality of the speech, other measures which map the perception of humans need to be found. The absolute mean error $|\epsilon|$ between two different F_0-contours ($F_{0,1}$ and $F_{0,2}$) is calculated as

$$|\epsilon| = \frac{\sum_{n=1}^{n=N} |F_{0,1} - F_{0,2}|}{N}. \quad (5)$$

For evaluation f values of the F_0-contours are transformed into midi (semitones) values according to

$$midi = 69 + 12 \cdot \log_2\left(\frac{f}{440}\right). \quad (6)$$

For the above sentence, the absolute mean error between the natural and the original ALT F_0-contour is $|\epsilon|_{midi,orig} = 12.2$ and decreases to $|\epsilon|_{midi,est} = 0.7$ for the estimated F_0-contour.

Furthermore, we used the 4-folds cross validation technique. The training data is randomly split into four sets. Three parts are then used in training and the fourth for validations. Then three others are picked and so on. Following this, results could be obtained separately for the male and the female database. r and $|\epsilon|$ were calculated for each sentence and the mean value over all sentences was determined. The results for r are shown in table 1. Figure 4 shows a Matlab boxplot of $|\epsilon|$. On each box, the central mark is the median ($q2$), the edges of the box are the 25th ($q1$) and 75th ($q3$) percentiles, the whiskers extend to the most extreme data points and the

	Male	Female
r_{orig}	0.29	0.23
r_{est}	0.92	0.90

Table 1: *Correlation coefficient of 4-folds cross validation.*

outliers are plotted individually. The values are listed in table 2. It can be seen that the enhancement in terms of the absolute mean error is significant.

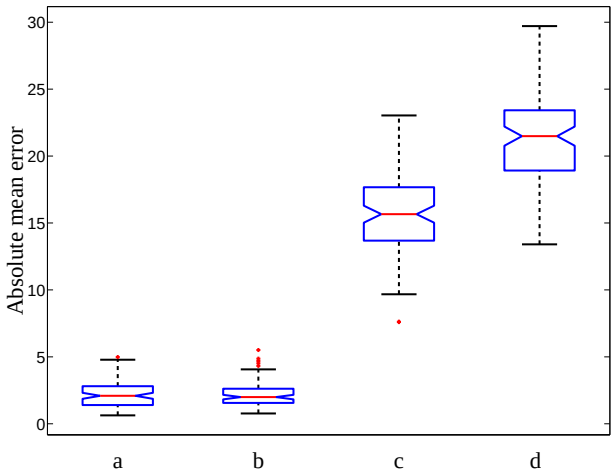

Figure 4: *Absolute mean error of 4-folds cross validation for male (a and c) and female (b and d) database: a and b: $|\epsilon|_{midi,est}$ – error between natural and estimated F_0-contour; c and d: $|\epsilon|_{midi,orig}$ – error between natural and original F_0-contour.*

	$q1$	$q2$	$q3$
	Male		
$\|\epsilon\|_{midi,est}$	1.40	2.09	2.81
$\|\epsilon\|_{midi,orig}$	13.68	15.66	17.69
	Female		
$\|\epsilon\|_{midi,est}$	1.56	1.99	2.62
$\|\epsilon\|_{midi,orig}$	18.92	21.49	23.42

Table 2: *25th ($q1$), median ($q2$) and 75th ($q3$) percentiles of 4-folds cross validation for male and female database; $|\epsilon|_{midi,est}$ – error between natural and estimated F_0-contour; $|\epsilon|_{midi,orig}$ – error between natural and original F_0-contour.*

Additionally, the ALT speech is manipulated with the estimated F_0-contour using the well known PSOLA method [11], which results in an improved listening quality.

4. Conclusions

In this paper, a method for automatically learning the F_0-contour for speech produced by an artificial larynx transducer is presented. Based on a database consisting of speech material of healthy and ALT speech, spectral features are extracted and a statistical model is trained. While the energy source, the larynx, is removed for laryngectomees, the vocal tract is still (fairly) unimpaired. Therefore, the assumption that spectra carry information about the speech prosody is valid. The results demonstrate that fundamental frequency estimation based on a machine learning procedure is possible and in terms of real-time application preferable.

The final aim is to improve ALT speech in general. Improving the F_0-contour is only a first step to reach this aim. Informal perceptual evaluations and earlier offline methods lead to the conclusion that an artificial F_0-contour considerably improves the naturalness and the speech quality of ALT speech.

Future research will investigate voiced/unvoiced decision of speech produced by the ALT. Furthermore, there is a strong need to improve the gender variation, which means that women are actually perceived as women. Further work is necessary to improve the sound quality, while preserving the identity of the voice and to give the user an individual voice characteristic.

5. Acknowledgements

The authors would like to thank HEIMOMED Heinze GmbH & Co.KG for their support.

6. References

[1] J. Lohscheller, “Dynamics of the Laryngectomy Substitute Voice Production”, Shaker-Verlag, Aachen, Germany, 2003.

[2] C. J. Griffin, “Artificial larynx with frequency control”, US Patent 5.812.681, Sept 22 1998.

[3] A. Loscos and J. Bonada, “Esophageal voice enhancement by modeling radiated pulses in frequency domain”, In *Proceedings of 121st Convention of the Audio Engineering Society*, San Francisco, CA, USA, Oct 3-6 2006.

[4] N. Uemi, T. Ifukube, M. Takahashi and J. Matsushima, “Design of a new electrolarynx having a pitch control function”, In *Proceedings of 3rd IEEE International Workshop on Robot and Human Communication*, RO-MAN p. 198 – 203, Nagoya, Japan, July 18-20 1994.

[5] K. Nakamura, T. Toda, H. Saruwatari and K. Shikano, “The use of air-pressure sensor in electrolaryngeal speech enhancement”, INTERSPEECH, p. 1628 – 1631, Makuhari, Japan, Sept 26-30 2010.

[6] C. Bishop, “Pattern Recognition and Machine Learning (Information Science and Statistics)”, Springer, 2007.

[7] B. Milner and X. Shao, “Prediction of Fundamental Frequency and Voicing From Mel-Frequency Cepstral Coefficients for Unconstrained Speech Reconstruction”, IEEE Trans. on Audio, Speech & Language Proc., 15(1): 24–33, 2007.

[8] Martin Hagmüller, “Speech Enhancement for Disordered and Substitution Voices”, Dissertation, Graz University of Technology, 2009.

[9] P. Boersma and D. Weenink “Praat ver 4.06”, software, downloaded from http://www.praat.org, 2007.

[10] D. Ellis, “Dynamic Time Warp (DTW) in Matlab”, http://www.ee.columbia.edu/dpwe/resources/matlab/dtw/, 2003.

[11] E. Moulines and F. Charpentier, “Pitch-synchronous waveform processing techniques for test-to-speech synthesis using diphones”, Speech Communication 9, p. 453 – 467, 1990.