Artificial Fundamental Frequency Contour for Electro-Larynx Speech

Anna K. Fuchs, Martin Hagmüller, Gernot Kubin
Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria

Abstract

The electro-larynx (EL) is one of the most common substitution voices for people who have lost their larynx. A well known problem of the resulting speech is its unnaturalness. In order to improve the naturalness, we want to provide an artificial fundamental frequency (f_0) contour. To estimate this contour, machine learning strategies are introduced. In this work we propose a Gaussian mixture model (GMM) based estimation technique and compare it to a Hidden Markov model (HMM) based technique. The contours are compared in a listening test where we also include the constant EL f_0 contour, a randomly chosen f_0 contour and the natural f_0 contour taken from parallel recorded healthy speech. Stimuli are taken from 4 subjects (2 female, 2 male) and 8 listeners judged the naturalness of the sentences using comparison category rating. Moreover, the perceived gender was evaluated. The result suggests that introducing an artificial f_0 contour improves electro-larynx speech. f_0 from natural speech was rated highest whereas already random f_0 performed 1better than constant f_0. The rating of GMM and HMM based estimation did not show significant differences.

Introduction

The currently available electro-larynx device has been introduced around 1960 and can be seen in Figure 1.

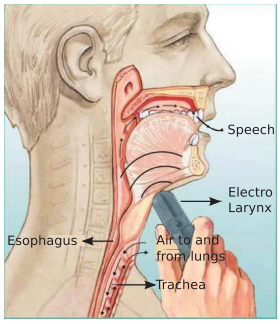

Figure 1: Anatomy and schematic of an EL user using the EL device [5].

The sound producing mechanism and the quality of the resulting speech has hardly improved since then. Prior studies investigated the problems of EL speech which include: improper source spectrum such as the reduction in low-frequency energy, lack of fine control over pitch, amplitude and voice on/offset (i.e., constant fundamental frequency; no variation in the harmonic structure) and interference of directly radiated sound from the EL device, reflections in the vocal tract due to the changed anatomy, and resulting reduced intelligibility concerning confusion between voiced and unvoiced consonants, and vowel intelligibility. Several groups worked on the enhancement of specific problems: e.g., air pressure from the stoma was used to control fundamental frequency, spectral subtraction and modulation filtering was introduced to reduce the directly radiated sound, and more recent work uses EMG based features together with voice conversion techniques.

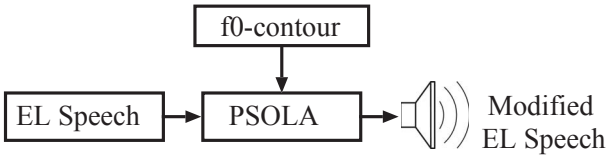

Figure 2: Approach to evaluate algorithms for f_0 estimation.

In this work we focus on the estimation of a changing f_0 and evaluate the strategies using a listening test.

Experimental Details

In order to analyze different strategies, we implement a framework as shown in Figure 2. For this purpose we used speech material from the German parallel Electro-Larynx Speech – Healthy (HE) speech corpus [3] to synthesize speech samples. Pitch synchronous overlap and add (PSOLA) was used to modify the fundamental frequency of the speech files [4]. The following strategies for estimating a changing f_0 are implemented A) constant f_0, B) HMM approach [1] (see Figure 3), C) GMM approach [2], D) random f_0 and E) natural f_0 from healthy speech. For A, D and E no training is needed and their implementation is straight forward. B and C are based on statistical models and, therefore, need prior training. Whereas C is explained in detail in [2], we want to explain B: First, EL speech files are pre-processed, i.e., the direct radiated noise from the EL device itself is removed. We use spectral subtraction in order to fulfill this task. Then, the fundamental frequency is extracted from HE speech ($f_{0,HE}$) using the auto-correlation method implemented in Praat. The next step is to time align $f_{0,HE}$ to the EL speech using a dynamic time warping algorithm. Mel-frequency cepstral coefficient (MFCC) features are extracted from EL speech. Finally we group all features to their corresponding f_0 value and estimate a GMM for each of these values. Now one HMM is fully described by its observations (MFCC features), hidden states (f_0), transition probabilities (histogram from f_0) and emission probabilities (estimated GMMs per hidden state). To estimate f_0 from an unknown EL sentence, we can use the off-line estimated HMM and estimate the best Viterbi path.

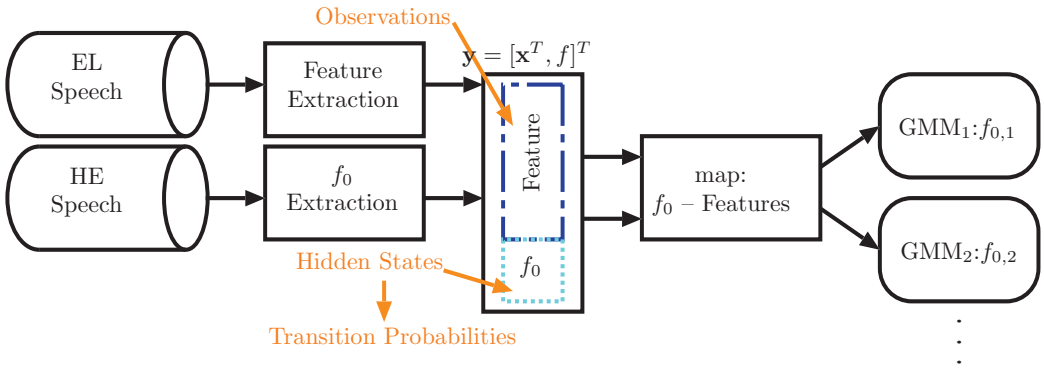

Figure 3: Feature extraction and training for B) HMM approach; estimation of f_0 based on best Viterbi path.

We used 2 sentences from 4 healthy speakers (2 female, 2 male). The sentences are phonetically balanced. The estimation of fundamental frequency was done using the strategies explained before. A listening test was carried out with 8 normal hearing listeners who rated naturalness of the 5 approaches (A - E) using comparison category rating where the listeners are presented with a pair (A, B) of speech stimuli to evaluate their naturalness. The listener has to rate if stimulus A is much better/better/slightly better/about the same/slightly worse/worse/much worse than B. Furthermore we asked for the perceived gender.

Results

The results are shown in Figure 4 and Table 1. We present the results in the order of overall preference. As expected method A (constant f_0) is rated worst and method E (natural f_0) is rated best. D is significantly better than A and significantly worse than B,C and E. We could not show any significant differences between B, C and E. We could show that the accuracy of the perceived gender is 98.24%. This means that the listeners had no problems to distinguish between gender. The results are sentence dependent and we also investigate differences between the GMM and HMM approaches depending on the gender of the speaker.

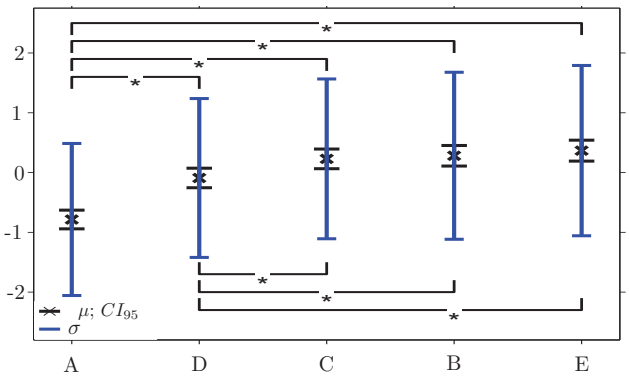

Figure 4: Overall preference; * indicates significant difference according to Wilcoxon rank sum test.

Conclusion

In this work we presented different strategies in order to enhance the naturalness of electro-larynx speech. The goal of these strategies is to estimate a changing fundamental frequency f_0. The evaluation was based on a listening test using comparison category rating. We could see that random f_0 is better than constant f_0 and that our proposed methods were rated about the same as natural f_0 and significantly better than constant f_0.

To sum this up: The artificial, changing f_0 contour improves electro-larynx speech and our proposed strategies are able to enhance EL speech in our framework.

Table 1: Overall preference $\bar{X}$ with variance σ and 95 % confidence interval CI_{95}.

rank	model	$\bar{X}$	$CI_{95} - \bar{X}$	σ
1	A (const)	-0.79	±0.16	1.27
2	D (rand)	-0.09	±0.16	1.33
3	C (GMM)	0.23	±0.16	1.34
4	B (HMM)	0.28	±0.17	1.40
5	E (nat)	0.37	±0.18	1.42

Acknowledgments

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

References

[1] Wohlmayr M., Stark M., and Pernkopf F.: "A Mixture Maximization Approach to Multipitch Tracking With Factorial Hidden Markov Models", ICASSP 2010, Dallas, pp. 5070 - 5073, 2010.

[2] Fuchs A. and Hagmüller M.: "Learning an Artificial F0-Contour for ALT speech", Interspeech 2012, Portland, Oregon, USA, 2012.

[3] Fuchs A. and Hagmüller M.: "A German Parallel Electro-Larynx Speech – Healthy Speech Corpus", 8th MAVEBA, Florence, Italy, Firenze University Press, pp. 55 - 58, 12/2013.

[4] Valbret H. Moulines E. and Tubach J.P.: "Voice transformation using PSOLA technique", ICASSP, pp. 145-148, 1992.

[5] `http://www.inhealth.com/category_s/61.htm`