Biomedical Signal Processing and Control xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Fundamental frequency tracking in diplophonic voices

P. Aichinger[a,*], M. Hagmüller[b], I. Roesner[a], B. Schneider-Stickler[a], J. Schoentgen[c], F. Pernkopf[b]

[a] Division of Phoniatrics-Logopedics, Department of Otorhinolaryngology, Medical University of Vienna, Waehringer Guertel 18-20, 1090 Vienna, Austria
[b] Signal Processing and Speech Communication Laboratory, Graz University of Technology, Inffeldgasse 16c/EG, 8010 Graz, Austria
[c] BEAMS (Bio-, Electro- And Mechanical Systems), Faculty of Applied Sciences, Université Libre de Bruxelles, 50, Av. F.D. Roosevelt, B-1050 Brussels, Belgium

ARTICLE INFO

Article history:
Received 19 April 2016
Received in revised form 7 October 2016
Accepted 10 October 2016
Available online xxx

Keywords:
Laryngeal high-speed videos
Diplophonia
f_o Tracking
Voice disorders
Audio signal processing

ABSTRACT

Background and objectives: Fundamental frequency (f_o) extraction in disordered voices is a prerequisite for many types of clinical analyses. Special attention must be paid if multiple oscillators with different f_os are active simultaneously. Two independent approaches to f_o tracking in diplophonic voices are proposed and compared with a benchmark from the literature.
Material and methods: Six samples of sustained phonations were analyzed. High-speed videos were obtained in addition to audio recordings. Video-based f_o tracks were obtained from cycle marks that report maximal vocal fold deflection in digital kymograms. Audio waveform modeling based extraction involved candidate tracking, oscillator waveform synthesis and track selection. Audio subband auto-correlation based extraction served as a benchmark.
Results and discussion: Promising qualitative and quantitative agreement of audio waveform modeling based estimates with kymogram-based tracks was observed. With reference to the kymogram-based tracks, audio waveform modeling based extraction had a median total error rate of 1.9%, which is an improvement over the benchmark method (17.7%).
Conclusion: The results illustrate that f_os of diplophonic voices may be validly obtained from kymogram cycle marks, as well as via audio waveform modeling. The acquisition of two simultaneous f_o tracks in diplophonic voices may increase the validity of clinical voice analysis procedures in the future.

1. Introduction

Voice disorders may lead to missed job opportunities, loss of quality of life, and social isolation, and thus need to be clinically understood and cared for. In the clinical care of voice disorders, voice assessment is essential to evidence-based decision making. In particular, voice assessment aids the indication, selection, evaluation and optimization of treatment techniques.

Voice assessment usually involves f_o-based techniques that fail if f_o measurement is wrong, which may lead to clinical misinterpretations. These techniques include audio analyzers that obtain jitter [1], shimmer [2,3], Harmonics-to-Noise-Ratios [4], voice range profiles [5], but also stroboscopy [6]. Attention must be paid when no unique f_o exists in the voice. For instance, if two f_os occur simultaneously in diplophonic voice, single f_o estimators fail. Two examples of possible clinical misinterpretations if analyses fail are the following. First, procedures that are based on f_o extraction provide invalid results if no single f_o exists. For instance, a measured jitter value of 0.4% may either be representing the "true" jitter of a signal if a single f_o exists, or it may be a quasi-random number if no single f_o exists [7]. Another example of misinterpretation exists with regard to voice range profiling if no single f_o exists in the voice. In such cases, artifacts that cannot be easily distinguished from valid results may be recorded. Thus, the correctness of f_o measurement is important for voice assessment. We perceive a need for improving the validity of clinical methods, which is confirmed by reports on limited reliability, accuracy, robustness, and validity of acoustic analyses of disordered voices [8–11]. As a consequence, not much is known about voices that involve no unique f_o. Thus, methods for f_o extraction need to be investigated on such voices. This study

Abbreviations: AWM, audio waveform modeling; ASA, audio subband auto-correlation; DFT, discrete Fourier transformation; FIR, finite impulse response; f_o, fundamental frequency; FS, Fourier synthesis; HMM, hidden Markov model; SPP, spectral peak picking.

* Corresponding author at: Department of Otorhinolaryngology, Waehringer Guertel 18-20, 1090 Vienna, Austria.
E-mail addresses: philipp.aichinger@meduniwien.ac.at (P. Aichinger), hagmueller@tugraz.at (M. Hagmüller), imme.roesner@meduniwien.ac.at (I. Roesner), berit.schneider-stickler@meduniwien.ac.at (B. Schneider-Stickler), jschoent@ulb.ac.be (J. Schoentgen), pernkopf@tugraz.at (F. Pernkopf).

http://dx.doi.org/10.1016/j.bspc.2016.10.002
1746-8094/

advances our understanding of diplophonic voices by introducing a probabilistic model of the dynamics of observed f_o tracks.

Dynamics of f_o tracks may be modeled by HMMs, which had been proposed for mixtures of two simultaneously talking speakers, but not for diplophonic voices. Wohlmayr et al. [12] published an audio-based f_o tracking procedure based on factorial HMMs. The HMM outperforms audio subband auto-correlation (ASA) based tracking [13]. Algorithmic extensions of Wohlmayr's model improved computational efficacy and handles mismatches between training and testing data, using model adaption [14]. Computational efficacy was increased via a variant of the Viterbi algorithm [15,16], which considered a restricted set of f_o combinations only. Unfortunately, the tracker requires amounts of training data, which are not available for diplophonic voices.

Waveform modeling is an approach involved in f_o extraction and has been investigated in the context of disordered voice analysis. Finite impulse response (FIR) filtering of unit pulse trains was used to determine doubled cycle marks in glottal area waveforms [17,18]. It was assumed that the number of oscillators and their f_os were known a priori from spectral video analysis. Joint f_o estimation and source separation with an unknown number of oscillators was used for distinguishing diplophonia from other types of dysphonia via audio signals only [18,19]. Spectral analysis of pixel intensity time series was used for extracting f_o tracks from the video and for evaluating f_o tracks extracted from diplophonic voices via waveform modeling without the use of HMM [20].

We here report the comparison of three methods for f_o tracking, which do not rely on the uniqueness of f_o. First, we obtained kymo-f_o-tracks from digital kymograms via manual marking of cycle boundaries. The tracks were acquired from cycle marks in kymograms by differencing, temporal smoothing, temporal interpolation, and calculation of the cycle length's reciprocals. Second, from audio signals via audio waveform modeling (AWM), we obtained AWM-f_o-tracks. The track candidates were obtained first via HMMs applied to audio spectrograms. Waveforms were then Fourier synthesized for each track candidate. The synthesizer used Fourier coefficients that were obtained by cross-correlating unit pulse trains with the audio signal. The best tracks were heuristically selected by evaluating the corresponding waveform model error. Third, we compared the two proposed methods with ASA-f_o-tracks [13].

The work is organized as follows. Voice samples are described in Section 2.1. Then, the obtainment of kymo-f_o-tracks and AWM-f_o-tracks is described in Sections 2.2 and 2.3. The obtainment of the ASA-f_o-tracks is explained subsequently in Section 2.4. The tracks are compared with each other and with audio spectrograms. The error measures are introduced in Section 2.5. The results are presented in Section 3. In Section 4, our choices are motivated and the limitations of the study are discussed. The text is concluded with a summary, inference of possible clinical consequences and an outlook.

2. Material and methods

2.1. Data collection

Six voice samples of sustained phonations from clinically diplophonic patients were analyzed. Three voice samples contained diplophonic intervals (D1, D2, and D3) and three were quasi-modal (M1, M2, and M3). Table 1 lists labels, ages, sexes, medical diagnoses, and the sample lengths.

The voice samples were recorded with a rigid endoscope laryngeal high-speed video camera HRES ENDOCAM 5562 (Richard Wolf GmbH) at 4000 video frames per second. The tip of the tongue was lightly held by a medical doctor when the endoscope was inserted into the mouth of the subject way back to the pharynx. The larynx was illuminated and filmed by the endoscopic camera and the pictures were previewed on a computer screen. The position of the camera was adjusted for optimal sight of the vocal folds. Once an acceptable position was achieved the subject was instructed to phonate an [i], which positioned the epiglottis so as to allow sight of the vocal folds. The produced sounds were schwa-like, due to the lowered position of the tongue [21].

Table 1
List of the label, age, sex, medical diagnosis, and length of the analyzed voice samples for each patient. Three samples were diplophonic (D1, D2, and D3) and three were quasi-modal (M1, M2, and M3).

Label	Age (yrs)	Sex	Medical diagnosis	Length (s)
D1	74	Female	Paresis	0.87
D2	72	Female	Sulcus	1.17
D3	72	Female	Paresis	0.95
M1	42	Female	Functional dysphonia	1.43
M2	68	Male	Cyst	0.30
M3	55	Male	Paresis	0.30

In parallel to the video recording, audio recordings were obtained with a head-worn condenser microphone AKG HC 577 L and a portable recorder TASCAM DR-100. The microphone was used with the original cap (no presence boost), windscreen AKG W77 MP and a phantom power adapter AKG MPA V L (linear response setting). The sampling rate was 48 kHz and the quantization resolution was 24 bits.

2.2. Kymogram-marking based f_o tracking

The proposed kymogram-marking based approach to obtaining kymo-f_o-tracks involved visual identification of oscillators in the video, extraction of digital kymograms, manual marking of cycle boundaries, determination of cycle lengths, temporal smoothing of cycle lengths, temporal interpolation of cycle lengths, and calculation of the cycle lengths' reciprocal (Fig. 1).

First, all glottal oscillators with distinct f_os were visually identified in the videos. The diplophonic samples were characterized by the occasional simultaneity of two glottal oscillators vibrating at different f_os. Such oscillators have been frequently observed in the past [22–28]. Second, the oscillators were visualized by digital kymography or multi-plane kymography [29,30]. If the right and the left vocal fold were vibrating at different f_os, one kymogram was used. If the anterior and the posterior parts of the vocal folds were vibrating at different f_os, two kymograms which visualized both oscillators were used. In modal voices, only one kymogram was used.

All cycles of vocal fold vibration were identified in the kymograms and their boundaries were marked graphically. Marks at cycle boundaries were manually positioned in the kymograms so as to agree with time instances of maximal vocal fold deflections. Fig. 2 shows a fragment of the kymogram of voice sample D1 and its cycle boundary marks. The shown cycle boundaries are located at maximal deflections towards the top (right vocal fold) and bottom (left vocal fold) of the kymogram.

Kymo-f_o-tracks were obtained from cycle boundaries. Cycle length sequences T_0 were obtained by subtracting video frame numbers of pairs of consecutive marks. Frame numbers were converted into seconds via division by the video frame rate, i.e. 4000 frames/s. For instance, a typical cycle length of 20 video frames corresponded to a cycle length of 5 ms. The cycle length sequences were temporally smoothed by averaging over 16 consecutive cycles to decrease noise in the kymogram markings. The noise stemmed from nonexact manual positioning of the markers and the finite temporal resolution of the kymograms. It had been assumed that this measurement noise is Gaussian, and that cycle lengths are con-

Please cite this article in press as: P. Aichinger, et al., Fundamental frequency tracking in diplophonic voices, Biomed. Signal Process. Control (2016), http://dx.doi.org/10.1016/j.bspc.2016.10.002