MULTI-MICROPHONE MAXIMUM A POSTERIORI FUNDAMENTAL FREQUENCY ESTIMATION IN THE CEPSTRAL DOMAIN

Timo Gerkmann, Rainer Martin, and Derya Dalga

Institute of Communication Acoustics (IKA)
Ruhr-Universität Bochum, 44780 Bochum, Germany

{timo.gerkmann,rainer.martin,derya.dalga}@rub.de

ABSTRACT

In this work we derive a new cepstrum based maximum likelihood fundamental frequency estimator that exploits the information of multiple microphones. The new approach results in a maximum search on the sum of the microphone cepstra. We compare the new approach to a maximum search on the cepstrum of the output signal of a delay-and-sum beamformer. We show that the new approach outperforms the beamforming approach for all considered input signal-to-noise ratios. We develop a general framework which includes the cepstral harmonics of the fundamental frequency and extend the approach towards a maximum *a posteriori* fundamental period tracker that further enhances the results and increases the robustness in noisy environments.

Index Terms— cepstral analysis, speech analysis, fundamental frequency estimation

1. INTRODUCTION

The fundamental period of voiced speech is caused by vibrations of the glottis. Its inverse, the fundamental frequency, is often simply referred to as *pitch*. As the speech fundamental period is one of the most important speech parameters, many solutions for fundamental period estimation have been proposed [1]. The fundamental period may be estimated for instance in the time domain using harmonic modelling [2], the autocorrelation function [3], or in the cepstral domain [4]. Knowledge about the speech fundamental period may be exploited for instance in speech coding [5], and speech enhancement [6, 7, 8]. As recent enhancement approaches [7, 8] operate in the cepstral domain, cepstrum based fundamental period estimators are of particular interest.

In the cepstral domain clean speech is decomposed into the lower cepstral coefficients that represent the transfer function of the vocal tract and the higher cepstral coefficients that represent the excitation of the vocal tract. For voiced sounds, the fundamental period of the excitation signal is represented by a dominant peak in the upper cepstrum, and multiples of that peak, the so-called *rahmonics* [4]. Thus, Noll suggests to simply search for the maximum peak in the cepstrum in the range of quefrencies that corresponds to the fundamental period [4].

After definition of the cepstrum in Section 2, in Section 3 we show that under certain assumptions a cepstral maximum search is optimal in the maximum likelihood (ML) sense. Further, we give the solution when R cepstral rahmonics are considered. In Section 4, we derive the ML optimal solution if multiple microphones are present. In Section 5 we incorporate a fundamental period tracking that is optimal in the maximum *a posteriori* (MAP) sense. In Section 6 we show that the new ML estimator outperforms a maximum search on the cepstrum of the output signal of a delay-and-sum beamformer for various input signal-to-noise ratios. The extension towards a MAP fundamental period tracker is shown to substantially increase the robustness in noisy environments.

2. THE CEPSTRUM

We consider the cepstral coefficients derived from the discrete short-time Fourier transform $S_k(l)$ of a discrete time domain signal $s(t)$, where t is the discrete time index, k is the discrete frequency index, and l is the segment index. The time domain signal is segmented, weighted with a window w_t, and transformed into the Fourier domain, as

$$S_k(l) = \sum_{t=0}^{K-1} w_t \, s(l\,L + t)\, e^{-\jmath 2\pi kt/K}\,, \tag{1}$$

where L is the number of samples between segments, and K is the segment size. The inverse discrete Fourier transform (IDFT) of the logarithm of the periodogram yields the cepstral coefficients

$$\widetilde{c}_q(l) = \frac{1}{K} \sum_{k=0}^{K-1} \log\left(|S_k(l)\,|^2\right) e^{\jmath 2\pi kq/K}\,, \tag{2}$$

where q is the cepstral index, the so-called *quefrency* index. As the log-periodogram is real-valued, the cepstrum is symmetric with respect to $q = K/2$. Therefore, in the

following we will only discuss the lower symmetric part $q \in \{0, 1, \ldots, K/2\}$. With the sampling frequency f_s, the fundamental period q_0/f_s of voiced speech appears as peaks at the discrete cepstral coefficients rq_0 with $r \in \mathbb{N}$.

A common assumption for the cepstral coefficients is that they are Gaussian distributed with a fixed variance. Assuming a complex Gaussian distribution for the frequency domain coefficients S_k, this variance can be shown to be $\mathrm{var}\{c_q\} = \pi^2/(6K)$ for $0 < q < K/2$ where a rectangular spectral analysis window w_t is assumed [9, 10]. With $\widetilde{c}_q$, S_k being realizations of the random variables $\widetilde{\mathcal{C}}_q, \mathcal{S}_k$, the mean of the cepstral coefficients can be shown to be [10]

$$\mathrm{E}\left\{\widetilde{\mathcal{C}}_q\right\} = \mathrm{IDFT}\left\{\log\left(\mathrm{E}\left\{|\mathcal{S}_k|^2\right\}\right)\right\} - \epsilon_q\,, \tag{3}$$

with $\epsilon_q = \gamma + \frac{2\log 2}{K}$ for $q = 0$, $\epsilon_q = \frac{2\log 2}{K}$ if q is even, and $\epsilon_q = 0$ if q is odd, where $\gamma = 0.5772$ is the Euler constant. Note, that the case $q = 0$ is not treated properly in [10]. For white signals we have $\mathrm{IDFT}\left\{\log\left(\mathrm{E}\left\{|\mathcal{S}_k|^2\right\}\right)\right\} = 0$ for $q > 0$. However, due to ϵ_q, even for white signals the cepstral coefficients $q > 0$ do not have zero mean. To obtain cepstral coefficients that have zero mean for white signals, we define

$$c_q = \widetilde{c}_q + \epsilon_q\,. \tag{4}$$

For non-white signals, the expected value of the cepstrum, $\mathrm{E}\{\mathcal{C}_q\}$, carries the information about the spectral shape. We find speech to be compactly represented by few lower cepstral coefficients $q < q_l$ representing the speech spectral envelope, the fundamental period peak q_0, and its rahmonics [4, 8]. Thus, for the cepstrum of noisy speech, we assume that cepstral coefficients at $q \geq q_l$ have zero mean except for the coefficients at $q = rq_0$ with $r \in \mathbb{N}$ that represent the fundamental period. Typically q_l corresponds to 1-2 ms.

In [10] it has been shown, that the cepstral coefficients are asymptotically uncorrelated for large K. Thus, Gaussian distributed cepstral coefficients are asymptotically independent.

3. ML FUNDAMENTAL PERIOD ESTIMATOR

In this section we derive a maximum likelihood (ML) estimator for the fundamental period in the cepstral domain.

Because the mean of the cepstrum is zero for $q \neq rq_0$ and $q \geq q_l$, the distribution of a noisy cepstral observation vector $\mathbf{c} = \left[c_{q_l}, c_{q_l+1}, ..., c_{K/2-1}\right]^T$ given the speech fundamental period index q_0 can be written as

$$\begin{aligned} p(\mathbf{c}|q_0) &= \prod_{q=q_l}^{K/2-1} \frac{1}{(2\pi\sigma_q^2)^{\frac{1}{2}}} \exp\left(-\frac{(c_q - \mathrm{E}\{\mathcal{C}_q\})^2}{2\sigma_q^2}\right) \\ &= \frac{1}{(2\pi\sigma_q^2)^{\frac{K/2-q_l}{2}}} \exp\left(-\frac{1}{2\sigma_q^2}\sum_{q=q_l}^{K/2-1} c_q^2\right) \\ &\cdot \exp\left(\frac{\sum_{r=1}^{R+1} 2c_{rq_0}\mathrm{E}\{\mathcal{C}_{rq_0}\} - \left(\mathrm{E}\{\mathcal{C}_{rq_0}\}\right)^2}{2\sigma_q^2}\right). \end{aligned} \tag{5}$$

For simplicity we neglect the Nyquist bin $q = K/2$, as it has a different variance than the coefficients $q_l < q < K/2$ [10].

For a quefrency independent cepstral variance σ_q^2, only the second exponential function has to be evaluated. As the exponential function is monotonically increasing, the ML estimator is given by

$$q_0^{\mathrm{ML}} = \arg\max_{q_0} p(\mathbf{c}|q_0) = \sum_{r=1}^{R+1} 2c_{rq_0}\mathrm{E}\{\mathcal{C}_{rq_0}\} - \left(\mathrm{E}\{\mathcal{C}_{rq_0}\}\right)^2$$

As speech is highly non-stationary, the instantaneous value is used to estimate the expected value, $\widehat{\mathrm{E}}\{\mathcal{C}_{rq_0}\} = |c_{rq_0}|$, where we may write the absolute value operator as, due to the structure of spectral harmonics, the fundamental period peak is always positive. For rectangular spectral analysis windows this also holds for the rahmonics at rq_0 with $r \in \{2, 3, \ldots\}$. The ML fundamental period estimation results in

$$q_0^{\mathrm{ML}} \quad = \quad \arg\max_{q_0} \sum_{r=1}^{R+1} |c_{rq_0}|\left(2c_{rq_0} - |c_{rq_0}|\right)\,, \tag{6}$$

which results in a cepstral peak detection if no rahmonics are considered, *i.e.* $R = 0$

$$q_0^{\mathrm{ML},R=0} = \arg\max_{q} c_q\,. \tag{7}$$

Thus, for rectangular spectral analysis windows and $R = 0$ a cepstral peak detection is an optimal fundamental period estimator in the ML sense.

4. EXTENSION TO MULTIPLE MICROPHONES

To extend the ML optimal solution when M microphones are present, we assume that the cepstral coefficients, given q_0, of the M microphones are independent. Thus we can write

$$p(\mathbf{C}|q_0) = \prod_{m=1}^{M} p(\mathbf{c}_m|q_0)\,, \tag{8}$$

with $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_M]$. For $R = 0$ and a quefrency and microphone independent cepstral variance σ_q^2, the ML estimator for multiple microphones results in a maximum search on the sum or mean of the microphone cepstra:

$$q_0^{\mathrm{ML},R=0} = \arg\max_{q} \sum_{m=1}^{M} c_{q,m}\,. \tag{9}$$

We refer to this approach as multi-microphone cepstral ML (MM-CML). Another approach that exploits the information of multiple microphones is to apply a ML fundamental period estimation on the output of a beamformer (BF-CML). The output of a beamformer has an increased signal-to-noise ratio as compared to each single microphone channel. This results in more prominent spectral harmonics and thus in an

increased cepstral peak. However, the variance of the non-speech cepstral coefficients stays unchanged, as it is independent of the signal power.

The MM-CML estimator (9), is fundamentally different to the BF-CML approach. Taking the mean of the microphone cepstra decreases the variance of cepstral coefficients, while the cepstral fundamental period peak stays approximately the same. While both approaches, MM-CML and BF-CML, increase the estimation performance, the superiority of the cepstral averaging approach is demonstrated in Section 6.

5. MAP FUNDAMENTAL PERIOD TRACKING

To decrease the amount of estimation errors, it is common practice to track the fundamental period over time. For this, we extend the ML fundamental period estimator towards a MAP fundamental period estimator similar to [2]. Assuming that the cepstral coefficients of consecutive signal segments given q_0 are independent, and treating the *a priori* probability of q_0 as a first order Markov chain, the MAP estimator including the information of the last Λ signal segments is given by [2]

$$q_0^{\text{MAP}}(l) = \arg\max_{q_0} \sum_{\lambda=0}^{\Lambda-1} \Big(\underbrace{\log\big(p(\mathbf{C}(l-\lambda)|q_0)\big)}_{L_{q_0}(\mathbf{C}(l-\lambda))} + \underbrace{\log\big(p(q_0|q_0^{\text{MAP}}(l-\lambda-1))\big)}_{B_{q_0}\left(q_0^{\text{MAP}}(l-\lambda-1)\right)} \Big). \quad (10)$$

Here, $p(q_0|q_0^{\text{MAP}}(l-1))$ is the *a priori* transition probability of the fundamental period q_0, when the MAP fundamental period estimate of the previous frame is $q_0^{\text{MAP}}(l-1)$. This *a priori* distribution can be chosen to be Gaussian, *i.e.*

$$p(q_0|q_0^{\text{MAP}}(l-1)) = \frac{1}{\sqrt{2\pi\sigma^2_{\text{tracking}}}} \exp\left(-\frac{\left(q_0 - q_0^{\text{MAP}}(l-1)\right)^2}{2\sigma^2_{\text{tracking}}} \right)$$

whereas the standard deviation σ_{tracking} can be found using labelled training data, *e.g.* [11].

As the pitch tracking algorithm is meant to provide pitch estimates for low-delay applications, no major look ahead is possible and an instantaneous decision is needed in each signal segment. To emphasize the information in recent signal segments, we realize (10) via a recursive averaging as

$$W_{q_0}(l) = \alpha W_{q_0}(l-1) + (1-\alpha)\Big(B_{q_0}(\hat{q}_0^{\text{MAP}}(l-1)) + L_{q_0}(\mathbf{C}(l)) \Big) \quad (11)$$

with $W_{q_0}(0) = L_{q_0}(\mathbf{C}(0))$ and the MAP fundamental period estimate

$$\hat{q}_0^{\text{MAP}}(l) = \arg\max_{q_0} W_{q_0}(l)\,, \quad (12)$$

where the log *a priori* transition probability $B_{q_0}(\cdot)$ and the log likelihood $L_{q_0}(\cdot)$ are defined in (10).

6. EVALUATION

We compare the MM-CML estimator based on the summation of the microphone cepstra (9) to a cepstral ML estimation on the output signal of a beamformer (BF-CML). Further, we give the results for the multi-microphone MAP fundamental period estimator MM-CMAP for $R = 0$ and $R = 1$. For the evaluation we use the Keele database [11] that consists of 5 male and 5 female speakers and up to 40 s of speech per speaker. The sampling rate is 20 kHz, the segment size 25.6 ms and the frame shift 10 ms. This corresponds to $K = 512$ in (2) and $L = 200$ in (1). We choose a rectangular spectral analysis window w_t in (2) and $q_l = 40$ (2 ms) in (5). For the MAP algorithm we choose the smoothing constant $\alpha = 0.8$ in (11). The standard deviation of the *a priori* probability is determined based on the labelled training data [11] and set to $\sigma_{\text{tracking}} = 23$ bins which corresponds to 1.1 ms.

To decouple the evaluation of the fundamental period estimators from the problem of automatic voiced/unvoiced classification, a fundamental period estimation is applied only on those signal segments that are marked as voiced in the Keele database. The estimated fundamental frequency $\hat{f}_0$ is compared to the reference fundamental frequency f_0 of the Keele database. For the evaluation we use the gross error rate (GER) and the relative root mean square error (RMSE) according to [12]. The GER is given as the percentage of signal segments that have a fundamental frequency estimate that deviates by more than $\theta\%$ of the reference fundamental period.

$$\text{GER}(\theta) = \frac{1}{K_{\text{v}}} \sum_{l=1}^{K_{\text{v}}} \left\{ \frac{|\hat{f}_{0_l} - f_{0_l}|}{f_{0_l}} > \theta\% \right\}, \quad (13)$$

where K_{v} is the number of voiced signal segments. The relative RMSE

$$\text{RMSE}(\theta) = \sqrt{\frac{1}{K_\theta} \sum_{l \in \Omega(\theta)} \left(\frac{\hat{f}_{0_i} - f_{0_i}}{f_{0_i}} \right)^2}. \quad (14)$$

is evaluated only for those K_θ signal segments of the set $\Omega(\theta)$, which have a relative fundamental frequency estimation error smaller than $\theta\%$. It can be seen as a measure for the fine fundamental frequency estimation error [12].

We now want to demonstrate the possible performance gain of the proposed multi-microphone approach in comparison to a delay-and-sum beamformer under ideal conditions. For this, we generate ten microphone signals with uncorrelated additive white Gaussian noise at several segmental signal-to-noise ratios (SNR). For the BF-CML approach, the ten microphone signals are summed in time domain and the ML estimator is applied on the cepstrum of the sum. We thus simulate the case of a source at the broadside of the array with its location perfectly known. For the MM-CML algorithm, the cepstrum is computed for each microphone signal, and the maximum of the sum of the cepstra is searched for (9). Note,

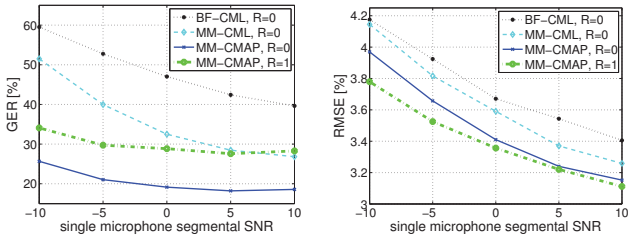

Fig. 1. GER (left) and RMSE (right) for various input segmental SNRs for $\theta = 10\%$, $M = 10$, and uncorrelated white Gaussian noise. The proposed MM-CML approach outperforms a delay-and-sum-beamformer (BF-CML) in terms of GER and RMSE for all considered input SNRs. Maximum a posteriori fundamental period tracking (MM-CMAP) further enhances the estimation performance.

that for the MM-CML a source localization is not needed, as the phase of the complex spectra is neglected when computing the cepstrum via (2). The results are given in Figure 1. The proposed MM-CML approach (9) clearly outperforms a delay-and-sum beamformer approach in terms of GER and RMSE. When the fundamental period is tracked over time (MM-CMAP), the results are further enhanced both in terms of a lower GER and RMSE. The estimation performance can be seen to be much more robust in noisy environments when MAP tracking is applied. While for all above simulations we chose $R = 0$, for the MM-CMAP we also present the results for $R = 1$. As the rahmonics are often much smaller than the fundamental period peak [4], it may happen that the sum of two noise bins is larger than the sum of the fundamental period peak and its rahmonic. Additionally, especially for male speakers, incorporating the rahmonics increases fundamental period halving errors. In that cases, estimation errors occur that result in an increased GER. However, the fine pitch estimation error is reduced if a rahmonic is included, as can be seen by a decreased RMSE in Figure 1.

We also conducted experiments with a microphone array in a reverberant room and different noise sources, namely white noise, speech-shaped noise, and babble noise. Due to the correlation in the low-frequencies, the performance gain achievable by using multiple microphones is reduced, and so is the difference between BF-CML and MM-CML. While the performance gain in terms of the RMSE became negligible in our setup, the proposed MM-CML approach still clearly outperformed the BF-CML for all SNRs and noise types in terms of the GER.

7. CONCLUSION

We derive the maximum likelihood and maximum a posteriori estimators for a fundamental frequency estimation in the cepstral domain and by this have also motivated the well known approach by Noll [4]. When no rahmonics are considered and rectangular spectral analysis windows are used, a simple maximum search is optimal in the maximum likelihood sense. We extend the likelihood function towards multiple microphones. The maximum likelihood solution results in a maximum search on the sum of all microphone cepstra. We show that this approach outperforms a cepstral maximum search on the cepstrum of the output of a delay-and-sum beamformer in terms of the gross error rate (GER) and root mean square fundamental frequency estimation error (RMSE) for all considered signal-to-noise ratios. Further, MAP fundamental period tracking substantially improves the robustness in noisy environments.

8. REFERENCES

[1] W. J. Hess, *Pitch Determination of Speech Signals*. Berlin: Springer Verlag, 1983.

[2] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum a posteriori probability pitch tracking in noisy environments using harmonic model," *IEEE Trans. on Speech and Audio Proc.*, vol. 12, no. 1, pp. 76–87, Jan. 2004.

[3] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

[4] A. M. Noll, "Cepstrum pitch estimation," *J. Acoust. Soc. Am.*, vol. 41, pp. 293–309, Feb. 1967.

[5] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding And Error Concealment*. John Wiley & Sons, 2006.

[6] J. Tilp, "Verfahren zur Verbesserung gestörter Sprachsignale unter Berücksichtigung der Grundfrequenz stimmhafter Laute," Ph.D. dissertation, Universität Darmstadt, Darmstadt, Germany, Jul. 2002.

[7] C. Breithaupt, T. Gerkmann, and R. Martin, "Cepstral smoothing of spectral filter gains for speech enhancement without musical noise," *IEEE Signal Proc. Letters*, vol. 14, no. 12, pp. 1036–1039, Dec. 2007.

[8] ——, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," *IEEE ICASSP*, pp. 4897–4900, Apr. 2008.

[9] P. Stoica and N. Sandgren, "Total-variance reduction via thresholding: Application to cepstral analysis," *IEEE Trans. on Signal Proc.*, vol. 55, no. 1, pp. 66–72, Jan. 2007.

[10] Y. Ephraim and M. Rahim, "On second-order statistics and linear estimation of cepstral coefficients," *IEEE Trans. on Speech and Audio Proc.*, vol. 7, no. 2, pp. 162–176, Mar. 1999.

[11] "Keele pitch database," *University of Liverpool, School of Psychology*, http://www.liv.ac.uk/Psychology/hmp/projects/pitch.html.

[12] F. Flego, "Fundamental frequency estimation techniques for multi-microphone speech input," Ph.D. dissertation, University of Trento, Italy, Mar. 2006.