MULTIPITCH ESTIMATION USING BLOCK SPARSE BAYESIAN LEARNING AND INTRA-BLOCK CLUSTERING

Liming Shi, Jesper Rindom Jensen, Jesper Kjær Nielsen, and Mads Græsbøll Christensen

Audio Analysis Lab, CREATE, Aalborg University,
{ls, jrj, jkn, mgc}@create.aau.dk

ABSTRACT

Pitch estimation is an important task in speech and audio analysis. In this paper, we present a multi-pitch estimation algorithm based on block sparse Bayesian learning and intra-block clustering for speech analysis. A statistical hierarchical model is formulated based on a pitch dictionary with a fixed maximum number of harmonics for all the candidate pitches. Block sparse Bayesian learning is proposed for estimating the complex amplitudes. To deal with the problem of unknown harmonic orders and subharmonic errors, intra-block clustering structured sparsity prior is also introduced. The statistical update formulas are obtained by the variational Bayesian inference. Compared with the conventional group LASSO-type algorithms for multi-pitch estimation, experimental results indicate robustness against noise and improved estimation accuracy of the proposed method.

***Index Terms*—** Multi-pitch estimation, block sparse Bayesian learning, clustering structured sparsity, subharmonic errors.

1. INTRODUCTION

Fundamental frequency (a.k.a., pitch) estimation has diverse applications in voice disorder detection [1], automatic music transcription [2], speech enhancement [3], etc. The pitch estimation algorithms can be broadly classified as *non-parametric* and *parametric* methods. The popular Yin [4] and RAPT [5] can be categorized as non-parametric methods since they are based on the *autocorrelation* function obtained within a specified time frame. These methods are computationally simple but they are sensitive to noise and prone to *subharmonic errors* (that is, misidentifying a rational number times the actual pitch). On the other hand, the pitch estimation methods using parametric model (e.g., *harmonic model*) are less commonly used but more robust to noise. In this model, both the pitch and complex amplitudes are assumed to be invariant during a short-time period (frame) (e.g., 20-40 ms for speech signals) [6]. Various kinds of estimators, such as the nonlinear least square estimator [6], have been proposed using the harmonic model or its variants.

When multiple speakers are present or multiple instruments are mixed in a music piece, the problem of multi-pitch estimation arises. In [7], different pitches were estimated by an iterative spectral subtraction process. That is, the estimated pitch from the most prominent sound was removed from the mixture signal repeatedly. The spectral smoothness principle was used to deal with the overlapping harmonics. A statistical harmonic model-based multi-pitch estimation algorithm was proposed in [8], where spectral smoothness was also imposed by modelling the spectral envelope of overtones as an autoregressive model. More recently, a multi-pitch estimation algorithm based on a pitch dictionary and group LASSO was proposed. A convex cost function, combining the advantages of l_2, sum of l_2, and l_1 norms, was designed, which was referred to as PEBS [9]. A total variation (TV) term was further introduced to reduce the subharmonic errors (PEBS-TV). However, due to the difficulty of tuning the regularization parameters, an adaptive penalty estimator with self-regularization was proposed in [10], called PEBSI-Lite. The dictionary in this algorithm was initialized with pitch candidates estimated by frequency estimation methods (e.g., ESPRIT [11]). An iterative solution was obtained by the alternating direction method of multipliers (ADMM) [12]. Typically, these methods incorporate prior knowledge about the spectral smoothness, which can be exploited by the regularization techniques, or by the Bayesian framework with prior models on the unknown complex amplitude parameters.

In this paper, motivated by the work in Bayesian sparse signal recovery [13–15], a block sparse Bayesian learning-based multi-pitch estimation algorithm is proposed. By imposing the block sparse prior, the complex amplitudes of the active pitches in the dictionary can be recovered and thus also the corresponding pitches using block sparse Bayesian learning (BSBL) method. Moreover, to deal with an unknown number of harmonic orders and the subharmonic problem, intra-block cluster structured sparsity prior is introduced. By clustering the non-zero elements of the complex amplitudes within each block, the subharmonic errors can be reduced. Variational Bayesian inference is applied for obtaining statistical update formulas.

2. FUNDAMENTALS

We aim to fit the observed speech signals to an over-complete harmonic model with harmonic series including P candidate pitches and each pitch have up to $L_{\max}$ harmonics, i.e.,

$$y_n = \sum_{p=1}^{P} \sum_{l=1}^{L_{\max}} a_{p,l} e^{j\omega_p l n} + m_n, \tag{1}$$

where $a_{p,l}$ denotes the complex amplitude of the l^{th} harmonic of the p^{th} pitch in the dictionary, n is the time index, m_n is the complex Gaussian white noise, $\omega_p = 2\pi f_p / F_s$, f_p denotes the p^{th} pitch, and F_s is the sampling rate. Collecting N observed samples and writing (1) to a matrix form, we have

$$\mathbf{y} = \mathbf{Z}\mathbf{a} + \mathbf{m}, \tag{2}$$

where $\mathbf{y} = [y_0, y_1, \cdots, y_{N-1}]^T$, the noise vector is given by $\mathbf{m} = [m_0, m_1, \cdots, m_{N-1}]^T$, the complex amplitude vector by $\mathbf{a} = \left[\mathbf{a}_1^T, \mathbf{a}_2^T, \cdots, \mathbf{a}_P^T\right]^T$, $\mathbf{a}_p = [a_{p,1}, a_{p,2}, \cdots, a_{p,L_{\max}}]^T$, $1 \leq$

This work was funded by the Independent Fund Denmark, grant ID: DFF 4184-00056

$p \leq P$, the dictionary $\mathbf{Z}$ is a $N \times PL_{\max}$ matrix denoted as $\mathbf{Z} = [\mathbf{Z}(\omega_1), \mathbf{Z}(\omega_2), \cdots, \mathbf{Z}(\omega_P)]$ and $\mathbf{Z}(\omega_p)$, for $1 \leq p \leq P$, has a Vandermonde structure as follows:

$$\mathbf{Z}(\omega_p) = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ e^{j\omega_p} & e^{j2\omega_p} & \vdots & e^{jL_{\max}\omega_p} \\ \vdots & \vdots & \vdots & \vdots \\ e^{j\omega_p(N-1)} & e^{j2\omega_p(N-1)} & \cdots & e^{jL_{\max}\omega_p(N-1)} \end{bmatrix}.$$

A key assumption in the over-complete harmonic model (2) is that the complex amplitude vector $\mathbf{a}$ is block sparse. However, when the actual number of harmonics of the p^{th} pitch candidate is less than $L_{\max}$, $\mathbf{a}_p$ also contains zeros. The sum of norms and L_1-norm regularization terms are introduced in [9] to impose both the block sparse and sparse priors for multi-pitch estimation. However, only using these two regularization terms may lead to subharmonic errors. For example, if the true pitch of an observed sinusoidal signal is 100 Hz and we have 50 Hz pitch in the dictionary, we may wrongly estimate the pitch as 50 Hz. This is because the observed signal can be fitted well with a block sparse complex amplitude vector estimate $\hat{\mathbf{a}}$ (e.g., $\hat{\mathbf{a}} = [\cdots, \mathbf{0}, \hat{\mathbf{a}}_p, \mathbf{0}, \cdots]^T$) and a sparse sub-block estimate $\hat{\mathbf{a}}_p$ that corresponds to the 50 Hz pitch (e.g., $\hat{\mathbf{a}}_p = [0, a_1, 0, a_2, 0 \cdots]^T$). To counter this problem, a total variation term is further added to the cost function to impose smoothness to the complex amplitudes.

3. PROPOSED BLOCK SPARSE BAYESIAN LEARNING AND INTRA-BLOCK CLUSTERING

As noted before, when subharmonic errors occur, the complex amplitude vector estimates of the subharmonics contain zeros. Instead of using the sparse and smoothness priors like the PEBS-TV, an alternative approach is to identify the complex amplitudes as cluster structured sparsity around the first several elements and up to the actual number of harmonics, which can be easily verified from the spectrogram of speech signals. In this paper, we impose both the block sparse prior and intra-block clustered structured sparse prior to the first several elements of each $\mathbf{a}_p$ for multipitch estimation. Block sparse prior is applied for estimating the complex amplitudes of the active pitches in the dictionary. Intra-block clustered structured sparse prior is exploited to counter the problem of unknown harmonic orders and subharmonic errors. In this section, we first formulate the problem using the hierarchical model and then give the update formulas using the variational Bayesian inference.

3.1. Hierarchical Model

We proceed by assigning a circular, symmetric white complex Gaussian to the observed noises, i.e.,

$$p(\mathbf{m}|\gamma) = C\mathcal{N}(\mathbf{m}|\mathbf{0}, \gamma^{-1}\mathbf{I}_N), \tag{3}$$

where a complex Gaussian variable $\mathbf{x}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ is defined as

$$C\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\pi^N|\boldsymbol{\Sigma}|}\exp\{-(\mathbf{x}-\boldsymbol{\mu})^H\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}. \tag{4}$$

A Gamma distribution is assigned to the precision γ of the complex Gaussian (conjugate prior), i.e.,

$$p(\gamma) \sim \Gamma(\gamma|c, d). \tag{5}$$

To motivate block sparsity and intra-block clustered sparsity for the complex amplitude vector $\mathbf{a}$, we first introduce a latent variable $\theta_{p,l}$ (the l^{th} element of the p^{th} block of $\boldsymbol{\theta}$) to indicate the zero/nonzero status of the corresponding complex amplitude coefficients $a_{p,l}$, i.e. $\mathbf{a} = \mathbf{u} \odot \boldsymbol{\theta}$, where $\odot$ denotes element-wise multiplication and

$$\begin{aligned} p(\mathbf{u}|\boldsymbol{\alpha}) &= C\mathcal{N}(\mathbf{u}|\mathbf{0}, \boldsymbol{\Lambda}^{-1}), \\ \boldsymbol{\Lambda} &= \text{diag}(\boldsymbol{\alpha}) \otimes \mathbf{I}_{L_{\max}}. \end{aligned} \tag{6}$$

The hyperparameter α_p (p^{th} element of $\boldsymbol{\alpha}$) is the precision of the p^{th} block, and when it is infinite, the p^{th} block will be zero [13]. A Gamma distribution is also assigned to the hyperparameter α_p as

$$p(\alpha_p) \sim \Gamma(\alpha_p|g, h). \tag{7}$$

Besides, the latent variable $\theta_{p,l}$ is drawn from Bernoulli distribution with success probability $\pi_{p,l}$, i.e.,

$$\theta_{p,l} \sim \text{Bernoulli}(\pi_{p,i}). \tag{8}$$

Three different patterns for clustered sparse recovery were introduced in [16,17], i.e., $P0$: "strongly eliminate", when the two neighbours are both zeros; $P1$: "weakly eliminate", when one of the neighbour are zero; $P2$: "strongly plump", when both of the neighbours are non-zeros. However, in pitch estimation, non-zero clusters are formed around the first several elements of $\mathbf{a}_p$ of true pitches. Therefore, we propose to use the following four-pattern model for the latent variable $\theta_{p,l}, 1 \leq p \leq P, 1 < l < L_{\max}$, i.e., *P0*: "strongly elimination", when $\theta_{p,1} = 0$ (fundamental frequency is missing); *P1*: "mildly eliminate", when the two neighbours are both zeros and $\theta_{p,1} = 1$; *P2*: "weakly eliminate", when one of the neighbour is zero and $\theta_{p,1} = 1$; *P3*: "strongly plump", when both of the neighbours are non-zeros and $\theta_{p,1} = 1$. According to these clustering patterns, the success probability for $1 < l < L_{\max}$ is chosen by

$$\pi_{p,l} = \begin{cases} \pi^0, \text{ if } P0 \\ \pi^1, \text{ if } P1 \\ \pi^2, \text{ if } P2 \\ \pi^3, \text{ if } P3 \end{cases}, \quad \pi^j \sim \text{Beta}(\pi^j|e^j, f^j),\ 0 \leq j \leq 3, \tag{9}$$

where π^j, for $0 \leq j \leq 3$ is drawn from the Beta distribution. Note that, the model for $l \in \{1, L_{\max}\}$ are not shown here for simplicity. However, we can follow the above definitions but use two patterns for $l = 1$ and three patterns for $l = L_{\max}$ because of their single neighbour characteristic. Using patterns $P1$, $P2$ and $P3$, we can expect that the non-zero elements within each block will be clustered together. Moreover, an all-zero cluster will be formed in the rear of the block since a large $L_{\max}$ is used. By introducing the pattern $P0$, a nonzero cluster around the first several elements of complex amplitude vector $\mathbf{a}_p$ is encouraged if the p^{th} pitch in the dictionary is active. We refer to the proposed algorithm as pitch estimation using block sparse Bayesian learning and intra-block clustering (PE-BSBL-Cluster). Note that if we set the latent variable $\theta_{p,l} = 1$, $1 \leq p \leq P$, $l \leq l \leq L_{\max}$, the intra-block clustering scheme will be dropped and only block sparse prior will be applied, which we refer to as PE-BSBL.

3.2. Variational Bayesian Inference

The exact joint posterior distribution can not be derived analytically. Instead, we resort to an approximation method, i.e., variational Bayesian inference [18]. For completeness, we give the update formulas in section 6. A detailed derivation and the results for $l \in \{1, L_{\max}\}$ is given in the technical report [19].

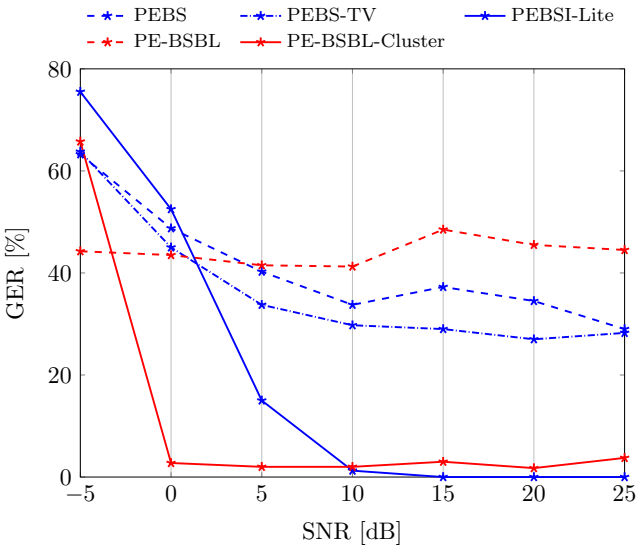

Fig. 1: Gross error ratio for synthetic signal in different SNRs, Q=2, $f_1^0 = 160$ and $f_2^0 = 240$ Hz ($240 = \frac{3}{2} \times 160$) .

4. RESULTS

We test the proposed PE-BSBL and PE-BSBL-Cluster in both synthetic and mixed speech signals scenarios [1]. All the modeling parameters are fixed as follows: $c = d = h = 10^{-6}$, $g = 1$, $(e^0, f^0) = (1, 10^6)$, $(e^1, f^1) = (1/L_{\max}, 1 - 1/L_{\max})$, $(e^2, f^2) = (1/L_{\max}, 1/L_{\max})$, $(e^3, f^3) = (1 - 1/L_{\max}, 1/L_{\max})$. The proposed algorithms are terminated if $\frac{\|\boldsymbol{\alpha}^i - \boldsymbol{\alpha}^{i-1}\|_2}{\|\boldsymbol{\alpha}^i\|_2} \leq 10^{-3}$ or 1000 iterations are reached, where i denotes the iteration number. Pitch estimates are obtained by choosing the pitches that have the largest posterior energies defined as $\tilde{\boldsymbol{\mu}}_p^H \tilde{\boldsymbol{\mu}}_p + \mathrm{Tr}(\tilde{\boldsymbol{\Sigma}}_p)$, where $\tilde{\boldsymbol{\mu}}_p$ and $\tilde{\boldsymbol{\Sigma}}_p$ denote the posterior mean and covariance of $\mathbf{a}_p$, and $\mathrm{Tr}(\cdot)$ is the trace operator. We compare the proposed algorithms with the PEBS [9], PEBS-TV [9] and PEBS-Lite [10]. For the PEBS and PEBS-TV, the regularization parameters are set to the same as in [10].

4.1. Synthetic Signal Analysis

The first experiment examines the performance for synthetic signals sampling of 8000 Hz, as shown in Fig. 1. Two pitches with 160 and 240 Hz are used. The data length N is set to 240. Uniform grid ranging from 50 to 500 Hz with grid interval 2 Hz and $L_{\max} = 10$ is used for all the experiments. To simulate the off-grid effect, for each trial, the true pitches are drawn from the uniform distribution, i.e., $f_{0,q} \sim \mathrm{Unif}(f_q^0 - d/2, f_q^0 + d/2), 1 \leq q \leq Q$. The deviation d is the grid interval and f_q^0 denotes a pitch on the grid. The number of harmonics are uniformly drawn over the integer interval $[3, 10]$ in each simulation. The amplitude of each harmonic is set to unit magnitude and the phase is drawn uniformly on $[0, 2\pi)$ [10]. The performance is measured by the gross error rate (GER), defined by calculating the number of pitch estimates that is differed by more than a certain percentage from the ground truth [20, 21]. In this paper, we use 5% for all the experiments. The experimental results are obtained by the ensemble averages over 200 Monte Carlo simulations. As can be seen, the GERs of the PEBS, PEBS-TV and PEBSI-Lite, are lower than the PE-BSBL, especially in high SNRs. This is because that the PE-BSBL only exploits the block sparse prior, and thus it is prone to subharmonic errors. Moreover, the PEBS-TV presents a better performance than the PEBS due to the TV term in the cost function. Furthermore, PEBSI-Lite obtains the lowest GER in high SNRs due to the built-in refining process and good performance of the ESPRIT in high SNRs. But its performance degenerates severely in low SNRs. By exploiting the block sparse and intra-block clustering structured priors together, the proposed PE-BSBL-Cluster achieves the lowest GER compared with PEBS, PEBS-TV and PE-BSBL in 0 to 25 dB SNRs. Although the PEBSI-Lite presents a slightly better performance than the proposed PE-BSBL-Cluster in high SNRs (10 to 25 dB), the proposed PE-BSBL-Cluster has a much lower GER in low SNRs (-5 to 5 dB). Thus, it is more robust to noise. Note that, high-resolution estimates for the proposed algorithm can be found by refining methods, such as gradient ascend method [6].

[1] An implementation of the proposed algorithms using MATLAB may be found in `https://tinyurl.com/y8orkosc`

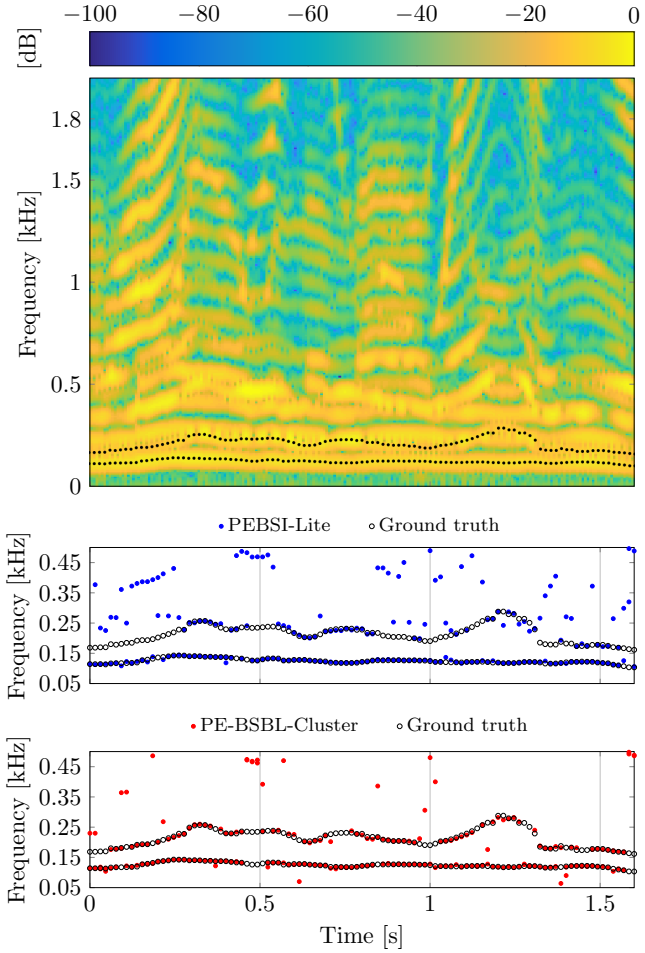

Fig. 2: Pitch estimates of real mixed speech of the spoken sentences "Why were you away a year?" from a female and "Our lawyer will allow your rule." from a male speaker, $F_s = 8000$ Hz, SNR=5 dB.

4.2. Mixed Speech Signal Analysis

We also examine the performance of the PE-BSBL and PE-SBL-Cluster for a mixed speech signal of the spoken sentences "Why

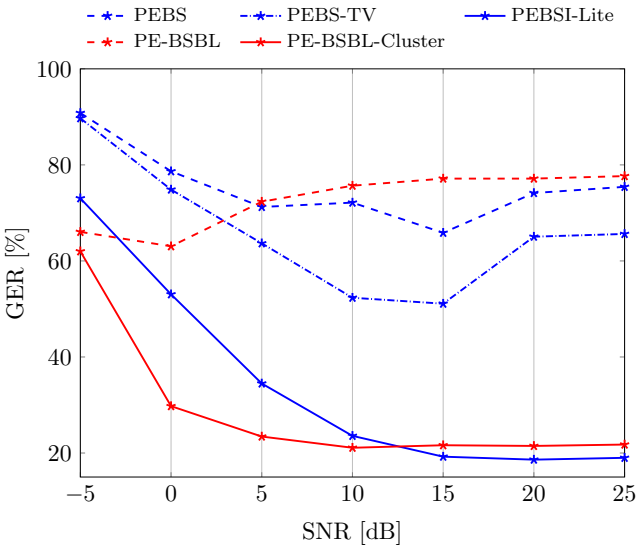

Fig. 3: Gross error ratio for real mixed speech in different SNRs.

were you away a year?" from a female speaker and "Our lawyer will allow your rule." from a male speaker. The ground truth pitch estimates of each sentence are obtained by Yin in noise-free scenario. The sampling rate is 8000 Hz. The spectrogram of mixed speech (noise-free), pitch estimates of PEBSI-Lite and the proposed PE-BSBL-Cluster under 5 dB SNR are shown in Fig. 2. On the spectrogram, the two black dotted lines (from top to bottom) denote the ground truth pitch estimates of the female and male sentences, respectively. The GER versus different SNRs, computed using 10 Monte-Carlo simulations, is shown in Fig. 3. Analysis is performed every 30 ms with 50% overlap. As can be seen from Fig. 2, the proposed PE-BSBL-Cluster has less estimation errors than the PEBSI-Lite. From the plots of both algorithms, the estimated pitch tracks of the male speaker can be clearly seen. However, it is easier to see the the estimated pitch track of the female speaker using the proposed PE-BSBL-Cluster than PEBSI-Lite. Similar conclusions to Fig. 1 can be drawn from Fig. 3. The proposed PE-BSBL-Cluster achieves the lowest GER in low SNRs (-5 to 10 dB) and has a comparable performance with the PEBSI-Lite in high SNRs (15-25 dB). Above all, due to the usage of the block sparse and clustering structured priors, compared with group-LASSO type algorithms, the proposed PE-BSBL-Cluster can deal with the problems of unknown harmonic orders and subharmonic errors, and presents a good performance even in low SNRs.

5. CONCLUSION

A multi-pitch estimation algorithm using block sparse Bayesian learning and intra-block clustering has been proposed. Using a block sparse prior model, the complex amplitude vectors corresponding to the true pitches in the pitch dictionary can be recovered. Moreover, to deal with unknown number of harmonic orders and subharmonic errors, intra-block clustering structured sparsity are encouraged by imposing a clustering prior. Update equations are obtained by the variational Bayesian inference. Simulation results using both synthetic and real mixed speech show that the proposed PE-BSBL-Cluster has improved multipitch estimation accuracy in terms of GER and robustness against noise.

6. APPENDIX

The approximated posteriors are listed as follows:
(1) **the indicator variable** $\theta_{p,l},\ 1 \le p \le P,\ 1 \le l \le L_{\max}$:

$$q(\theta_{p,l}) = \text{Bernoulli}(\tilde{\pi}_{p,l}), \tag{10}$$

where

$$\begin{aligned}&\tilde{\pi}_{p,l}\\ =&[1+\exp\{\langle\log(1-\pi_{p,l})\rangle - \langle\log(\pi_{p,l})\rangle + \langle\gamma\rangle[\langle u^*_{p,l}u_{p,l}\rangle \mathbf{z}^H_{p,l}\mathbf{z}_{p,l}\\ &- 2\text{Re}(\langle u_{p,l}\rangle^* \mathbf{z}^H_{p,l}(\mathbf{y} - \sum_{(i,j)\neq(p,l)} \langle\theta_{i,j}\rangle\langle u_{i,j}\rangle \mathbf{z}_{i,j})]\}]^{-1},\end{aligned}$$

where $\langle\cdot\rangle$ denotes the expectation operator, $(\cdot)^*$ denotes the conjugate and $(\cdot)^H$ denotes conjugate transpose.
(2) **the complex amplitude** $\mathbf{u}$:

$$q(\mathbf{u}) = C\mathcal{N}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}), \tag{11}$$

where

$$\begin{aligned}\tilde{\boldsymbol{\Sigma}} &= (\langle\boldsymbol{\Lambda}\rangle + \langle\gamma\rangle\langle\text{diag}(\boldsymbol{\theta})\mathbf{Z}^H\mathbf{Z}\text{diag}(\boldsymbol{\theta})\rangle)^{-1},\\ \tilde{\boldsymbol{\mu}} &= \langle\gamma\rangle\tilde{\boldsymbol{\Sigma}}\langle\text{diag}(\boldsymbol{\theta})\rangle\mathbf{Z}^H\mathbf{y},\end{aligned}$$

and $\langle\text{diag}(\boldsymbol{\theta})\mathbf{Z}^H\mathbf{Z}\text{diag}(\boldsymbol{\theta})\rangle = (\mathbf{Z}^H\mathbf{Z}) \odot (\langle\boldsymbol{\theta}\rangle\langle\boldsymbol{\theta}\rangle^T + \text{diag}(\langle\boldsymbol{\theta}\rangle \odot (1 - \langle\boldsymbol{\theta}\rangle)))$.
(3) **the noise precision** γ:

$$q(\gamma) = \Gamma(\gamma|\tilde{c}, \tilde{d}), \tag{12}$$

where

$$\begin{aligned}\tilde{c} =& c + N,\\ \tilde{d} =& d + \|\mathbf{y} - \mathbf{Z}(\langle\mathbf{u}\rangle \odot \langle\boldsymbol{\theta}\rangle)\|^2 + \text{Tr}\{\mathbf{Z}^H\mathbf{Z}(\langle\mathbf{u}\mathbf{u}^H\rangle \odot (\langle\boldsymbol{\theta}\boldsymbol{\theta}^T\rangle)\\ &- (\langle\mathbf{u}\rangle \odot \langle\boldsymbol{\theta}\rangle)(\langle\mathbf{u}\rangle \odot \langle\boldsymbol{\theta}\rangle)^H)\}.\end{aligned}$$

(4) **the precision** $\alpha_p,\ 1 \le p \le P$ **of the complex amplitudes**:

$$q(\alpha_p) = \Gamma(\alpha_p|\tilde{g}_p, \tilde{h}_p), \tag{13}$$

where

$$\tilde{g}_p = g + L_{\max}, \quad \tilde{h}_p = h + \langle\mathbf{u}_p^H\mathbf{u}_p\rangle.$$

(5) **the success probability** $\pi_{p,l},\ 1 \le p \le P,\ 1 < l < L_{\max}$:

$$q(\pi^j_{p,l}) = \text{Beta}(\pi^j_{p,l}|\tilde{e}^j_{p,l}, \tilde{f}^j_{p,l}), \tag{14}$$

where for $j \in \{0, 1, 2, 3\}$,

$$\begin{aligned}\tilde{e}^j_{p,l} =& e^j + p(Pj)\langle\theta_{p,l}\rangle,\\ \tilde{f}^j_{p,l} =& f^j + p(Pj)(1 - \langle\theta_{p,l}\rangle),\end{aligned}$$

and

$$\begin{aligned}p(P0) =& 1 - \langle\theta_{p,1}\rangle,\\ p(P1) =& \langle\theta_{p,1}\rangle(1 - \langle\theta_{p,l-1}\rangle)(1 - \langle\theta_{p,l+1}\rangle),\\ p(P2) =& \langle\theta_{p,1}\rangle(\langle\theta_{p,l-1}\rangle(1 - \langle\theta_{p,l+1}\rangle) + \langle\theta_{p,l+1}\rangle(1 - \langle\theta_{p,l-1}\rangle)),\\ p(P3) =& \langle\theta_{p,1}\rangle\langle\theta_{p,l-1}\rangle\langle\theta_{p,l+1}\rangle.\end{aligned}$$

The Expectation of logarithm function can be calculated as

$$\langle\log\pi_{p,l}\rangle = \sum_{j=0}^{3} p(Pj)\langle\log\pi^j_{p,l}\rangle,$$

$$\langle\log(1-\pi_{p,l})\rangle = \sum_{j=0}^{3} p(Pj)\langle\log(1-\pi^j_{p,l})\rangle.$$

7. REFERENCES

[1] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 4, pp. 884–893, 2010.

[2] M. Muller, D. P. Ellis, A. Klapuri, and G. Richard, "Signal processing for music analysis," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1088–1110, 2011.

[3] M. K. Becker and T. Gerkmann, "Fundamental frequency informed speech enhancement in a flexible statistical framework," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 24, no. 5, pp. 940–951, 2016.

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

[5] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech coding and synthesis*, W. B. Kleijn and K. K. Paliwal, Eds: Elseview Science B. V., vol. 495, 1995.

[6] M. G. Christensen and A. Jakobsson, *Multi-pitch estimation*, ser. Synthesis Lectures on Speech & Audio Processing, B. Juang, Ed. Morgan and Claypool Publishers, vol. 5, no. 1, pp. 1–160, 2009.

[7] A. P. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 804–816, 2003.

[8] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 18, no. 6, pp. 1643–1654, 2010.

[9] S. I. Adalbjörnsson, A. Jakobsson, and M. G. Christensen, "Multi-pitch estimation exploiting block sparsity," *Signal Process.*, vol. 109, pp. 236–247, 2015.

[10] F. Elvander, T. Kronvall, S. I. Adalbjörnsson, and A. Jakobsson, "An adaptive penalty multi-pitch estimator with self-regularization," *Signal Process.*, vol. 127, pp. 56–70, 2016.

[11] P. Stoica, R. L. Moses, *et al.*, *Spectral analysis of signals*, vol. 452, Pearson Prentice Hall, Upper Saddle River, NJ, 2005.

[12] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[13] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, no. Jun, pp. 211–244, 2001.

[14] T. Park and G. Casella, "The bayesian lasso," *J. Amer. Stat. Assoc.*, vol. 103, no. 482, pp. 681–686, 2008.

[15] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2346–2356, 2008.

[16] L. Yu, H. Sun, J. P. Barbot, and G. Zheng, "Bayesian compressive sensing for cluster structured sparse signals," *Signal Process.*, vol. 92, no. 1, pp. 259–269, 2012.

[17] L. Yu, C. Wei, J. Jia, and H. Sun, "Compressive sensing for cluster structured sparse signals: Variational bayes approach," *IET Signal Process.*, vol. 10, no. 7, pp. 770–779, 2016.

[18] C. M. Bishop, *Pattern recognition and machine learning*, springer, 2006.

[19] L. Shi, J. R. Jensen, J. K. Jensen, and M. G. Christensen, "Technical report," `https://tinyurl.com/y8orkosc`.

[20] F. Flego and M. Omologo, "Robust F0 estimation based on a multi-microphone periodicity function for distant-talking speech," in *Proc. European Signal Processing Conf.* IEEE, 2006, pp. 1–4.

[21] M. W. Hansen, J. R. Jensen, and M. G. Christensen, "Estimation of multiple pitches in stereophonic mixtures using a codebook-based approach," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2017, pp. 186–190.