AUXILIARY FUNCTION APPROACH TO PARAMETER ESTIMATION OF CONSTRAINED SINUSOIDAL MODEL FOR MONAURAL SPEECH SEPARATION

Hirokazu Kameoka *

NTT Communication Science Laboratories,
3-1 Morinosato Wakamiya, Atsugi,
Kanagawa, 243-0198, Japan

Nobutaka Ono, and Shigeki Sagayama

Graduate School of Information Science
and Technology, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan

ABSTRACT

We introduce in this paper an auxiliary function approach to parameter estimation of the constrained sinusoidal model, which enables us to derive a complex-spectrum-domain EM-like multiple F_0 estimation algorithm. Through simulations, we evaluated the performance of the presented method in the ability to avoid locally optimal solutions. We implemented a monaural speech separation system based on the presented method and confirmed its performance on compound signals of real speech.

Index Terms— Acoustic signal analysis, Speech enhancement, Optimization methods, Parameter estimation

1. INTRODUCTION

Many conventional methods for multiple F_0 estimation are based on the power spectrum domain approach, in which the influence of the interferences between frequency components of different sources are assumed negligible. However, it becomes usually difficult to infer F_0s only from power spectrum when two voices are mixed with close F_0s. In such a situation not only the harmonic structure (powers of harmonics) but also the phase of each component is an important cue for precise estimation of F_0.

Generally speaking, if the compound signal were separated into single voices, then it would be a simple matter to obtain an F_0 estimate and phase estimates of the harmonics from each stream. On the other hand, if the F_0s and the phases were known in advance, these information could be very useful for any separation algorithms. This leads to a "chicken and egg" situation: F_0/phase estimation and source separation are each a prerequisite of the other. This fact leads us naturally to formulate F_0/phase estimation and source separation as a joint optimization problem. The method described in this paper performs F_0/phase estimation step and source separation step iteratively using a constrained sinusoidal model.

The range of application of sinusoidal model has widened to Text-To-Speech synthesis, speech modification, coding, etc. since McAulay et al. [1] showed that the sinusoidal signal model could be applied to Analysis-by-Synthesis systems to obtain high-quality synthesized speech. In particular, as the possibility to generate high-quality synthesized speech shows that the sinusoidal signal model represents extremely well acoustic signals such as speech and music, we can have high expectations for its application to source separation.

*The author performed the work while at the University of Tokyo.

While McAulay et al. used as the model a mixture of K pure tone signals, one can also consider the use of the superposition of K harmonic signals (source signal composed of N harmonic components, where the frequency of n-th harmonic component is n times the F_0). This model is often used for single channel source separation especially when the target mixed signal is assumed to be composed of harmonic signals [3, 4, 5]. Most of the source separation approaches using this model are based on gradient search or sampling methods[3, 4, 5]. However, this kind of numerical computation is often beset with local optimum problems and computational costs. A global optimum is not guaranteed to be obtained unless, in the case of the gradient search methods the iterative computation is led to convergence for an infinity of initial points, or in the case of the sampling methods an infinite number of trial is performed. For that reason, the problem is to know if the search for the solution can be performed with a low computational cost or if it has an ability to avoid local optima.

As explained above, albeit the sinusoidal signal model represents extremely well acoustic signals such as speech and music, room was left for discussion on how to estimate its parameters. Against this background, we describe in this paper a new optimization algorithm to obtain the maximum likelihood parameter of the constrained sinusoidal model.

2. PROBLEM SETTING

Consider as the time-varying acoustic signal the sum of pseudoperiodic signals given in an analytic signal representation:

$$s(t) = \sum_{k=1}^{K} \sum_{n=1}^{N} A_{k,n}(t) e^{jn\theta_k(t)}, \ \ t \in (-\infty, \infty), \quad (1)$$

where the instantaneous phase $\theta_k(t)$ of the fundamental component, and the instantaneous complex amplitude $A_{k,n}(t)$ of the n-th partial component are the unknown parameters. $\mu_k(t) = \dot{\theta}_k(t)$ amounts to the instantaneous F_0 and $a_{k,n}(t) = |A_{k,n}(t)|$ the instantaneous amplitude, which are both assumed here to change gradually over time. These are the free parameters that one wants to estimate, which we denote for convenience by $\mathbf{\Theta}(t)$. Now letting $y(t)$ be the observed signal of interest, we assume the following model:

$$y(t) = s(t) + n(t), \ \ t \in (-\infty, \infty), \quad (2)$$

where $n(t)$ is a Gaussian white noise. The maximum likelihood estimate of $\mathbf{\Theta}(t)$ can thus be obtained by minimizing

the L^2 norm of the error signal in $t \in \mathbb{R}(-\infty, \infty)$:

$$\int_{\mathbb{R}} \Big\| y(t) - s(t) \Big\|^2 \mathrm{d}t. \tag{3}$$

We now show that this time domain objective can be equivalently formalized in the time-frequency domain defined by the Gabor transform (STFT).

Lemma 1. *The time-frequency components of $y(t)$ and $s(t)$ by Gabor transform is by definition given by*

$$G_y(\omega, t) \triangleq \Big\langle y(u), \psi_{\omega,t}(u) \Big\rangle_{u \in \mathbb{R}}, \tag{4}$$

$$G_s(\omega, t) \triangleq \Big\langle s(u), \psi_{\omega,t}(u) \Big\rangle_{u \in \mathbb{R}}, \tag{5}$$

where $\psi_{\omega,t}(u)$ is the Gabor function, which is a nonorthogonal basis used to measure the component of frequency ω at time t, and defined as the product of the complex sinusoid with frequency of ω and the Gaussian window centered at time t:

$$\psi_{\omega,t}(u) = e^{-\frac{\omega_0^2}{2}(u-t)^2 + j\omega(u-t)}, \tag{6}$$

where ω_0 is a time spread parameter of the Gaussian window, that can be chosen arbitrarily. Though trivial, we then have

$$\int_{\mathbb{R}} \Big\| y(t) - s(t) \Big\|^2 \mathrm{d}t = \eta \iint_{\mathbb{R}^2} \Big\| G_y(\omega, t) - G_s(\omega, t) \Big\|^2 \mathrm{d}\omega \mathrm{d}t,$$

where η is a constant.

If we now assume that $\theta_k(u)$ and $A_{k,n}(u)$ are respectively piecewise linear and piecewise constant, then

$$G_s(\omega, t) = \sum_{k=1}^{K} \sum_{n=1}^{N} A_{k,n}(t) e^{-\frac{(\omega - n\mu_k(t))^2}{2\omega_0^2}}. \tag{7}$$

In the case of discrete-time observations, the problem is to minimize

$$\Phi(\boldsymbol{\Theta}) = \int_{\mathbb{R}} \left\| Y(\omega) - \sum_{k,n} A_{k,n} e^{-\frac{(\omega - n\mu_k)^2}{2\omega_0^2}} \right\|^2 \mathrm{d}\omega, \tag{8}$$

at each discrete time point with respect to $A_{k,n}$ and μ_k. For clarity of notations the time index of $A_{k,n}(t)$, $\mu_k(t)$ and $\boldsymbol{\Theta}(t)$ is omitted and $Y(\omega)$ is a simplified notation of $G_y(\omega, t)$.

3. PARAMETER OPTIMIZATION ALGORITHM

The parameter optimization algorithm we propose in this paper is based on a principle called the *auxiliary function method*. We first define the auxiliary function and then show how the iterative algorithm is performed.

Definition 1. *(Auxiliary function) Let $\Phi(\boldsymbol{\Theta})$ be the objective function that one wants to minimize with respect to the parameter $\boldsymbol{\Theta} = (\Theta_1, \cdots, \Theta_I)$. We then define $\Phi^+(\boldsymbol{\Theta}, m)$ as the auxiliary function of $\Phi(\boldsymbol{\Theta})$, and $m = (m_1, \cdots, m_J)$ as the auxiliary parameter if $\Phi^+(\boldsymbol{\Theta}, m)$ satisfies*

$$\Phi\big(\boldsymbol{\Theta}\big) = \min_{m} \Phi^+\big(\boldsymbol{\Theta}, m\big) \Rightarrow \Phi\big(\boldsymbol{\Theta}\big) \leqq \Phi^+\big(\boldsymbol{\Theta}, m\big). \tag{9}$$

Lemma 2. *Denoting by $\Phi(\boldsymbol{\Theta})$ the objective function, and by $\Phi^+(\boldsymbol{\Theta}, m)$ the auxiliary function of $\Phi(\boldsymbol{\Theta})$, then the objective function $\Phi(\boldsymbol{\Theta})$ can be decreased monotonically by minimizing $\Phi^+(\boldsymbol{\Theta}, m)$ iteratively with respect to $m = (m_1, \cdots, m_J)$ and with respect to $\boldsymbol{\Theta}$:*

$$\widehat{m} = \operatorname*{argmin}_{m} \Phi^+\big(\boldsymbol{\Theta}, m\big) \tag{10}$$

$$\widehat{\boldsymbol{\Theta}} = \operatorname*{argmin}_{\boldsymbol{\Theta}} \Phi^+\big(\boldsymbol{\Theta}, \widehat{m}\big) \tag{11}$$

If $\Phi(\boldsymbol{\Theta})$ is bounded below, then the parameter $\boldsymbol{\Theta}$ converges to a stationary point.

Proof. Suppose we set $\boldsymbol{\Theta}$ to an arbitrary value $\widetilde{\boldsymbol{\Theta}}$. We will prove that $\Phi(\boldsymbol{\Theta})$ is non-increasing after the update Eq. (10) and Eq. (11). From Eq. (10), one obtains $\Phi(\widetilde{\boldsymbol{\Theta}}) = \Phi^+(\widetilde{\boldsymbol{\Theta}}, \widehat{m})$, and it is obvious from Eq. (11) that $\Phi^+(\widetilde{\boldsymbol{\Theta}}, \widehat{m}) \geqq \Phi^+(\widehat{\boldsymbol{\Theta}}, \widehat{m})$. By definition, one sees from Eq. (9) that $\Phi^+(\widehat{\boldsymbol{\Theta}}, \widehat{m}) \geqq \Phi(\widehat{\boldsymbol{\Theta}})$. Therefore, we can immediately prove that $\Phi(\widetilde{\boldsymbol{\Theta}}) = \Phi^+(\widetilde{\boldsymbol{\Theta}}, \widehat{m}) \geqq \Phi^+(\widehat{\boldsymbol{\Theta}}, \widehat{m}) \geqq \Phi(\widehat{\boldsymbol{\Theta}})$. □

It should be emphasized here that the EM algorithm can be considered as a special case of this method.

One possible auxiliary function of $\Phi(\boldsymbol{\Theta})$ can be derived using the lemma suggested for example in [2].

Lemma 3. *If a complex function $m_i(x)$ satisfies $\sum_i m_i(x) = \sum_i m_i^*(x) = 1$, then for $x \in \mathbb{R}(-\infty, \infty)$*

$$\int_{\mathbb{R}} \Big\| y(x) - \sum_i s_i(x) \Big\|^2 \mathrm{d}x \leqq \sum_i \frac{1}{\beta_i} \int_{\mathbb{R}} \Big\| m_i(x) y(x) - s_i(x) \Big\|^2 \mathrm{d}x, \tag{12}$$

where β_i is a constant such that $\sum_i \beta_i = 1,\ \beta_i \in (0, 1)$.

Putting $S_{k,n}(\omega) \triangleq A_{k,n} e^{-(\omega - n\mu_k)^2/2\omega_0^2}$ for simplicity of notation, then by the Lemma 3 and from Eq. (8) we have the following inequality:

$$\Phi(\boldsymbol{\Theta}) = \int_{\mathbb{R}} \left\| Y(\omega) - \sum_{k,n} S_{k,n}(\omega) \right\|^2 \mathrm{d}\omega \leqq \sum_{k,n} \frac{1}{\beta_{k,n}} \int_{\mathbb{R}} \Big\| m_{k,n}(\omega) Y(\omega) - S_{k,n}(\omega) \Big\|^2 \mathrm{d}\omega, \tag{13}$$

where $\beta_{k,n} \in (0, 1),\ \sum_{k,n} \beta_{k,n} = 1$ and equality holds if

$$m_{k,n}(\omega) = \frac{1}{Y(\omega)} \left[S_{k,n}(\omega) + \beta_{k,n} \left(Y(\omega) - \sum_{k,n} S_{k,n}(\omega) \right) \right]. \tag{14}$$

Let us denote by $\Phi^+(\boldsymbol{\Theta}, m)$ the right-hand side of Eq. (13). By Definition 1, $\Phi^+(\boldsymbol{\Theta}, m)$ is an auxiliary function of the objective $\Phi(\boldsymbol{\Theta})$, and $m_{k,n}(\omega)$ is an auxiliary parameter. By Lemma 2, we consider next to minimize $\Phi^+(\boldsymbol{\Theta}, m)$ with

respect to $\mathbf{\Theta}$. Using the result of the Gaussian integral: $\int_{\mathbb{R}} \|S_{k,n}(\omega)\|^2 \mathrm{d}\omega = \sqrt{\pi}\omega_0 \|A_{k,n}\|^2$, one obtains

$$\Phi^+(\mathbf{\Theta}, m) = \sqrt{\pi}\omega_0 \sum_{k,n} \frac{\left\|A_{k,n}\right\|^2}{\beta_{k,n}} + \sum_{k,n} \frac{1}{\beta_{k,n}} \int_{\mathbb{R}} \left(\left\|m_{k,n}(\omega)Y(\omega)\right\|^2 - 2e^{-\frac{(\omega - n\mu_k)^2}{2\omega_0^2}} \mathrm{Re}\Big[A_{k,n} m_{k,n}^*(\omega) Y^*(\omega)\Big] \right) \mathrm{d}\omega. \quad (15)$$

However, one notices from Eq. (15) that one still cannot obtain analytically the update equation for μ_k because of the nonlinear part $e^{-(\omega - n\mu_k)^2/2\omega_0^2}$ in Eq. (15). One may want to derive another auxiliary function such that the update equation for μ_k can be obtained analytically. In order to derive such an auxiliary function, we focused on two points: one is that $-e^{-x}$ is a continuously differentiable concave function, and second is that we have the following theorem about continuously differentiable concave function.

Lemma 4. *Let $f(x)$ be a real function of x that is continuously differentiable and concave. Then, for any point $\alpha \in \mathbb{R}$,*

$$f(x) \leqq f(\alpha) + (x - \alpha) f'(\alpha). \quad (16)$$

Since $-e^{-x}$ is a differentiable concave function, using Lemma 4 we have the inequality $-e^{-x} \leqq (x - \alpha - 1)e^{-\alpha}$, for any point $\alpha \in \mathbb{R}$. Replacing x with $(\omega - n\mu_k)^2/2\omega_0^2$ and α with a real function $\alpha_{k,n}(\omega)$, then from Eq. (15),

$$\Phi^+(\mathbf{\Theta}, m) \leq \sqrt{\pi}\omega_0 \sum_{k,n} \frac{\left\|A_{k,n}\right\|^2}{\beta_{k,n}} + \sum_{k,n} \frac{1}{\beta_{k,n}} \int_{\mathbb{R}} \left[\left\|m_{k,n}(\omega)Y(\omega)\right\|^2 + 2\mathrm{Re}\Big[A_{k,n} m_{k,n}^*(\omega) Y^*(\omega)\Big] e^{-\alpha_{k,n}(\omega)} \left(\frac{(\omega - n\mu_k)^2}{2\omega_0^2} - \alpha_{k,n}(\omega) - 1 \right) \right] \mathrm{d}\omega. \quad (17)$$

Denoting by $\widetilde{\Phi}^+(\mathbf{\Theta}, m, \alpha)$ the right-hand side of this inequation, $\widetilde{\Phi}^+(\mathbf{\Theta}, m, \alpha)$ can also be considered as an auxiliary function of $\Phi(\mathbf{\Theta})$ because

$$\Phi(\mathbf{\Theta}) \leqq \Phi^+(\mathbf{\Theta}, m) \leqq \widetilde{\Phi}^+(\mathbf{\Theta}, m, \alpha). \quad (18)$$

In this case both $m_{k,n}(\omega)$ and $\alpha_{k,n}(\omega)$ are the corresponding auxiliary parameters. The equality $\Phi(\mathbf{\Theta}) = \widetilde{\Phi}^+(\mathbf{\Theta}, m, \alpha)$ holds when $m_{k,n}(\omega)$ is given by Eq. (14) and $\alpha_{k,n}(\omega)$ by

$$\alpha_{k,n}(\omega) = \frac{(\omega - n\mu_k)^2}{2\omega_0^2}. \quad (19)$$

The advantage worth mentioning of deriving this auxiliary function is that it enables the analytical expression of the update equation for the F_0 parameter μ_k, allowing us a complex-spectrum-domain EM-like multiple F_0 estimation algorithm.

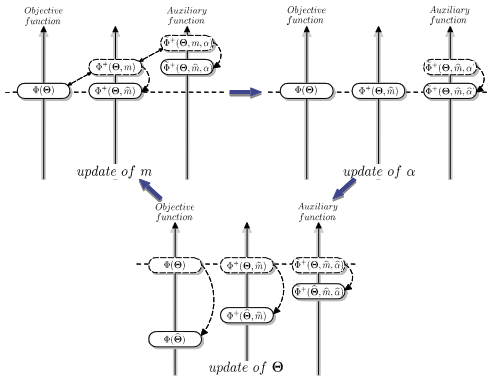

Fig. 1. An illustration of the iterative algorithm

Setting to 0 the partial derivative of $\widetilde{\Phi}^+(\mathbf{\Theta}, m, \alpha)$ with respect to μ_k, one obtains the F_0 parameter update rule:

$$\mu_k = \frac{\displaystyle\sum_n \frac{n}{\beta_{k,n}} \int_{\mathbb{R}} e^{-\alpha_{k,n}(\omega)} \mathrm{Re}\Big[A_{k,n} m_{k,n}^*(\omega) Y^*(\omega)\Big] \omega \mathrm{d}\omega}{\displaystyle\sum_n \frac{n^2}{\beta_{k,n}} \int_{\mathbb{R}} e^{-\alpha_{k,n}(\omega)} \mathrm{Re}\Big[A_{k,n} m_{k,n}^*(\omega) Y^*(\omega)\Big] \mathrm{d}\omega}. \quad (20)$$

Setting to 0 the partial derivative of $\widetilde{\Phi}^+(\mathbf{\Theta}, m, \alpha)$ with respect to $A_{k,n}^*$, the update equation for $A_{k,n}$ can also be derived as

$$A_{k,n} = \frac{1}{\sqrt{\pi}\omega_0} \int_{\mathbb{R}} m_{k,n}(\omega) Y(\omega) e^{-\alpha_{k,n}(\omega)} \left(\frac{(\omega - n\mu_k)^2}{2\omega_0^2} - \alpha_{k,n}(\omega) - 1 \right) \mathrm{d}\omega. \quad (21)$$

The presented algorithm is summarized as follows (Fig. 1).

Step 0 Initial setting of $\{\mu_k, \{A_{k,n}\}_{1 \leq n \leq N}\}_{1 \leq k \leq K}$.
Step 1 Update $m_{k,n}(\omega)$ by Eq. (14).
Step 2 Update $e^{-\alpha_{k,n}(\omega)}$ by Eq. (19).
Step 3 Update $A_{k,n}$ by Eq. (21).
Step 4 Update μ_k by Eq. (20) and return to Step 1.

The algorithm described above generates the suboptimal estimates not only of μ_k and $A_{k,n}$ but also of the separate signal $m_{k,n}(\omega)Y(\omega)$. This fact shows that the algorithm performs parameter estimation and source separation simultaneously.

4. EXPERIMENTAL EVALUATION

We first compared the dependency on the initial parameter and the convergence speed of the gradient search method and the proposed method. In this comparative experiment, we used a synthetic signal as test data which was created by adding together two periodic signals (with F_0s of 200Hz and 270Hz) composed of 10 harmonic components, with each component's amplitude and phase determined by random generation. In the sinusoidal signal model, we set $K = 2$ and $N = 10$. A Gabor transform with diffusion parameter $\omega_0 = 0.366$ was performed on this synthetic signal (16kHz sampling frequency) to obtain $Y(\omega)$. The courses of the update of the F_0 parameters μ_1, μ_2 at each step through the

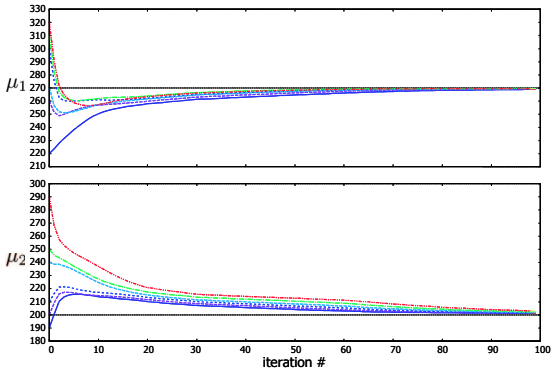

Fig. 2. *Course of the update of μ_1, μ_2 (the proposed method)*

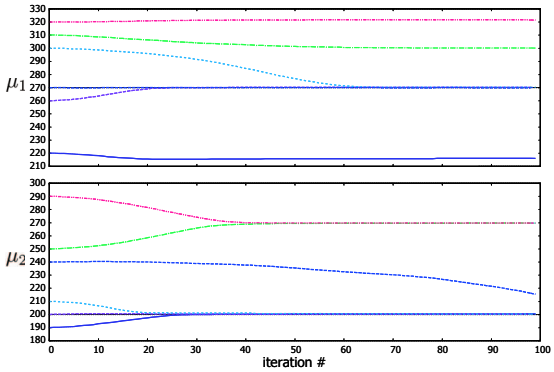

Fig. 3. *Course of the update of μ_1, μ_2 (the gradient method)*

proposed method and the gradient method, starting from various initial parameter conditions, are shown in Fig. 2 and Fig. 3. The transitions of the update values of μ_1 and μ_2 corresponding to the same iterative computation are shown in each figure respectively in the upper and lower part with the same color and same line type. The initial value for the amplitude $A_{k,n}$ was set to 0. One sees from Fig. 2 and Fig. 3 that the gradient method often gets trapped into stationary points different from the true values for initial values of μ_1, μ_2 which are not sufficiently close to the true values (200Hz, 270Hz), while the proposed method converges quickly from any initial points in a large interval to the true values.

Next, we confirm here the performance for single channel source separation. We use the ATR B-set speech database to build the mixed signals by adding together the waveforms of utterances from two male speakers, two female speakers, or a male speaker and a female speaker. For all the speech data the sampling rate was 16kHz, and the frequency analysis was done using a Gabor transform with a frame interval of 10ms. ω_0 was set to 0.366 and N was set to 30. The initial values for μ_k were obtained by finding all the frequencies giving a minimum or a maximum of the real part or the imaginary part of $Y(\omega)$, and selecting the frequencies with the 10 largest powers. K was initially set to 10 and after the parameter converged, the source models with the 2 largest total-powers were chosen as the final source signal estimates. In this experiment, in order to confirm the basic source separation performance in the situation where the permutation problem would be dealt with, we determine to which source the separated signals correspond by looking at their proximity to each signal prior to the mixing. Under the above conditions, an example of results of the separation of the mixed signal shown in Fig. 4 is shown in Fig. 5. After separation performed on the mixed signal of the male speaker A and the female speaker B (with a SNR of -0.3dB seen from the male speaker A), the SNRs for the speakers were 7.2dB and 6.4dB. On the mixed signal of the female speaker A and the female speaker B (with a SNR of 1.5dB seen from the female speaker A), we obtained the SNRs of 6.0dB and 4.8dB, and on the mixed signal of the male speaker A and the male speaker B (with a SNR of -0.3dB seen from the male speaker A), we obtained the SNRs of 4.8dB and 4.3dB after the separation was performed.

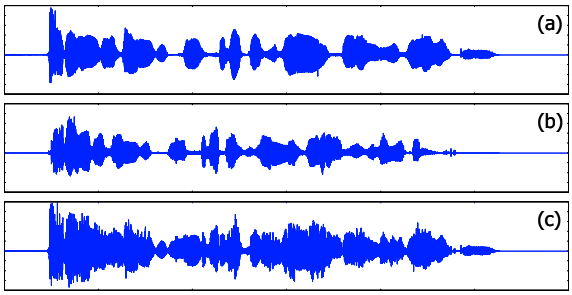

Fig. 4. *Utterance by a female speaker (a), a male speaker (b) and their mixed signal (c).*

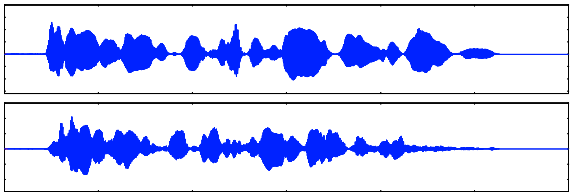

Fig. 5. *Separated signals corresponding to the female (top) and the male speaker (bottom).*

Although the method presented in this paper estimates the parameter independently for each frame, one may expect a substantial reduction of the musical noise and an improvement of the SNR if a coordinated parameter estimation across several adjacent frames could be performed. This shall be one of our future works.

5. SUMMARY

In this paper, we introduced the auxiliary function method for parameter optimization of constrained sinusoidal model, which enabled us to derive a complex-spectrum-domain EM-like multiple F_0 estimation algorithm. We confirmed the ability to avoid local solutions and the convergence speed of the presented method and the performance on speech separation.

6. REFERENCES

[1] R. J. McAulay and T. F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE Trans. Acoust., Speech, Signal Process.*, Vol. ASSP-34, No. 4, pp. 744–754, 1986.

[2] M. Feder and E. Weinstein, "Parameter Estimation of Superimposed Signals Using the EM Algorithm," *IEEE Trans. Acoust., Speech, Signal Process.*, Vol. ASSP-36, No. 4, pp. 477–489, 1988.

[3] D. Chazan, Y. Stettiner and D. Malah, "Optimal Multi-Pitch Estimation Using the EM Algorithm for Co-Channel Speech Separation," In *Proc. ICASSP'93*, Vol. 2, pp. 728–731, 1993.

[4] P. Jinachitra, "Constrained EM Estimates for Harmonic Source Separation," In *Proc. ICASSP2003*, Vol. 6, pp. 609–612, 2003.

[5] S. Godsill and M. Davy, "Bayesian Harmonic Models for Musical Pitch Estimation and Analysis," In *Proc. IEEE ICASSP2002*, Vol. 2, pp. 1769–1772, 2002.