Single and Multiple F_0 Contour Estimation Through Parametric Spectrogram Modeling of Speech in Noisy Environments

Jonathan Le Roux, *Student Member, IEEE*, Hirokazu Kameoka, *Student Member, IEEE*, Nobutaka Ono, *Member, IEEE*, Alain de Cheveigné, *Member, IEEE*, and Shigeki Sagayama, *Member, IEEE*

***Abstract*—This paper proposes a novel F_0 contour estimation algorithm based on a precise parametric description of the voiced parts of speech derived from the power spectrum. The algorithm is able to perform in a wide variety of noisy environments as well as to estimate the F_0s of cochannel concurrent speech. The speech spectrum is modeled as a sequence of spectral clusters governed by a common F_0 contour expressed as a spline curve. These clusters are obtained by an unsupervised 2-D time-frequency clustering of the power density using a new formulation of the EM algorithm, and their common F_0 contour is estimated at the same time. A smooth F_0 contour is extracted for the whole utterance, linking together its voiced parts. A noise model is used to cope with nonharmonic background noise, which would otherwise interfere with the clustering of the harmonic portions of speech. We evaluate our algorithm in comparison with existing methods on several tasks, and show 1) that it is competitive on clean single-speaker speech, 2) that it outperforms existing methods in the presence of noise, and 3) that it outperforms existing methods for the estimation of multiple F_0 contours of cochannel concurrent speech.**

***Index Terms*—Acoustic scene analysis, expectation-maximization (EM) algorithm, harmonic-temporal structured clustering (HTC), multipitch estimation, noisy speech, spline F_0 contour.**

I. Introduction

THE analysis of complex and varied acoustic scenes is a fundamental and challenging problem for today's acoustic signal processing. At its core is the need for a method which can determine precisely and robustly the F_0 contour of harmonic signals such as speech. Robust F_0 analysis has a very broad range of applications in computational auditory scene analysis (CASA), speech recognition, prosody analysis, speech enhancement, or speaker identification. So far, many pitch determination algorithms (PDAs) have been proposed [1], some of them with very good accuracy [2]. However, they are usually limited to the clean speech of a single speaker and fail in moderate amounts of background noise or the presence of other speakers. Ideally, the performance of a PDA should stay high in as wide a range of background noises as possible (white noise, pink noise, noise bursts, music, other speech, etc.). Furthermore, the possibility to extract simultaneously the F_0 contours of several concurrent voices is also a desirable feature. Several PDAs already exist that deal with the tracking of multiple F_0s [3]–[7]. Several of these algorithms rely on an initial frame-by-frame analysis followed by post-processing to reduce errors and obtain a smooth F_0 contour, for example using hidden Markov models (HMMs) (see [3] for a review). Here, we propose to perform estimation and model-based interpolation simultaneously, based on a parametric model of the time and frequency shape of the spectral envelope of the voiced parts of speech.

To this end, we rely on a parametric description of the wavelet power spectrum that accounts for its structure simultaneously in time and frequency directions. In brief, we model the power spectrum $W(x,t)$ as a sum of K parametric source models $q_k(x,t;\mathbf{\Theta})$, where x is log-frequency, t is time, and $\mathbf{\Theta}$ is the set of model parameters: $W(x,t) \approx \sum_{k=1}^{K} q_k(x,t;\mathbf{\Theta})$. In this paper, this model is applied to the voiced parts of speech, but in other contexts it could be applied to other acoustic events, provided a mathematical description of those events. Here, we apply two constraints to the spectro-temporal model. Along the frequency axis, it consists of a series of harmonics with frequencies multiple of a common F_0. Along the temporal axis, this F_0 follows a contour modeled as a cubic spline, and the amplitude of each partial follows a smooth temporal envelope modeled as a sum of Gaussian functions. The cubic spline F_0 contour was preferred to other options such as the Fujisaki model [8] as it can be applied to a wider range of stimuli in addition to speech. Moreover, as we will explain later, this choice enabled us to obtain analytic update equations during the optimization process. To handle the slowly-varying harmonic structure of the short-term wavelet spectrum, we used a multipitch analysis method initially developed for feature extraction of music signals, the harmonic-temporal structured clustering (HTC) method [9], which gives a parametric representation of the harmonic parts of the power spectrum.

In the original formulation of HTC, each source model represents a sound stream with *constant* F_0. However, by considering speech as a succession of models that each corresponds to a phoneme (or more generally to a segment of the speech utterance with steady acoustic characteristics), we can use the HTC method to model the spectrum as a sequence of spectral cluster models with a continuous F_0 contour. Contrary to

Manuscript received July 24, 2006; revised November 22, 2006. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Simon King.

J. Le Roux, H. Kameoka, N. Ono, and S. Sagayama are with the Graduate School of Information Science and Technology, University of Tokyo, Tokyo 113–8656, Japan (e-mail: leroux@hil.t.u-tokyo.ac.jp; kameoka@hil.t.u-tokyo.ac.jp; onono@hil.t.u-tokyo.ac.jp; sagayama@hil.t.u-tokyo.ac.jp).

A. de Cheveigné is with the Centre National de la Recherche Scientifique, Université Paris V, and Ecole Normale Supérieure, 75230 Paris Cedex 05, France (e-mail: Alain.de.Cheveigne@ens.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2007.894510

HMMs, which assume discrete phonetic states, we aim here to model smooth transitions in the temporal succession of the spectral structures.

Finally, we introduce also a noise model to deal with the nonharmonic power coming from background noise. While the spectrogram of voiced speech is characterized by harmonic parts with strong relative power, noise tends to have a more flat spectrogram. Extracting voiced speech from such noise corresponds to searching for local, harmonically structured "islands" within a "sea" of unstructured noise.

The parametric model we obtain in this way is optimized using a new formulation of the EM algorithm. The spectral clusters are obtained by an unsupervised 2-D clustering of the power density, performed simultaneously with the estimation of the F_0 contour of the whole utterance. We describe our method more precisely in the following section.

II. Formulation of the Model

A. General HTC Method

Consider the wavelet power spectrum $W(x,t)$ of a signal recorded from an acoustic scene, defined on a domain of definition $D = \{x,t \in \mathbb{R} \mid \Omega_0 \leqslant x \leqslant \Omega_1, T_0 \leqslant t \leqslant T_0 + T\}$. The problem considered is to approximate the power spectrum as well as possible as the sum of K parametric source models $q_k(x,t;\mathbf{\Theta})$ modeling the power spectrum of K "objects" each with its own F_0 contour $\mu_k(t)$ and its own harmonic-temporal structure. We note that the formulation described hereafter is not limited to any particular time-frequency representation, and it could work with a standard linear-frequency short-time Fourier transform (STFT). However, we found better performance with wavelets, possibly because it offers relatively better spectral resolution for the low-frequency harmonics of the voice. We will thus use the wavelet power spectrum in the following.

As described in [9], the source models $q_k(x,t;\mathbf{\Theta})$ are expressed as a Gaussian mixture model (GMM) with constraints on the characteristics of the kernel distributions: supposing that there is harmonicity with N partials modeled in the frequency direction, and that the power envelope is described using Y kernel functions in the time direction, we can rewrite each source model in the form

$$q_k(x,t;\mathbf{\Theta}) = \sum_{n=1}^{N} \sum_{y=0}^{Y-1} S_{kny}(x,t;\mathbf{\Theta}) \tag{1}$$

where $\mathbf{\Theta}$ is the set of all parameters and with kernel densities $S_{kny}(x,t;\mathbf{\Theta})$ which are assumed to have the following shape:

$$S_{kny}(x,t;\mathbf{\Theta}) \triangleq \frac{w_k v_{kn} u_{kny}}{2\pi\sigma_k\phi_k} e^{-\frac{(x-\mu_k(t)-\log n)^2}{2\sigma_k^2} - \frac{(t-\tau_k-y\phi_k)^2}{2\phi_k^2}} \tag{2}$$

where the parameters w_k, v_{kn} and u_{kny} are normalized to unity. A graphical representation of a HTC source model $q_k(x,t;\mathbf{\Theta})$ can be seen in Fig. 1.

Our goal is to minimize the difference between $W(x,t)$ and $Q(x,t;\mathbf{\Theta}) = \sum_{k=1}^{K} q_k(x,t;\mathbf{\Theta})$ according to a certain criterion.

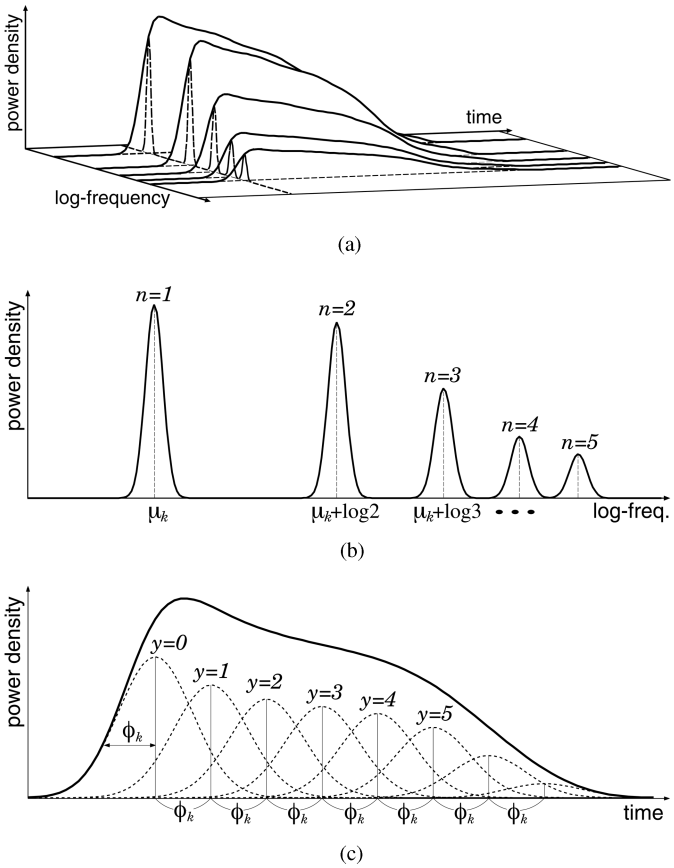


Fig. 1. Graphical representation of a HTC source model. (a) shows the time-frequency profile of the model, while (b) shows a cross section of the model at constant time, and (c) the evolution in time of the power envelope function. The harmonic structure of the model can be seen in (b), and the approximation of the power envelope in the time direction as a sum of Gaussian kernels can be seen in (c).

We use the $\mathcal{I}$-divergence [10] as a classical way to measure the "distance" between two distributions

$$\mathcal{I}\left(W|Q(\mathbf{\Theta})\right) \triangleq \iint_D \Big(W(x,t) \log \frac{W(x,t)}{Q(x,t;\mathbf{\Theta})} - \Big(W(x,t) - Q(x,t;\mathbf{\Theta})\Big) \Big) \mathrm{d}x\mathrm{d}t \tag{3}$$

and we are thus looking for $\mathbf{\Theta}_{\text{opt}} = \operatorname{argmin}_{\mathbf{\Theta}} \mathcal{I}(W|Q(\mathbf{\Theta}))$. Keeping only the terms depending on $\mathbf{\Theta}$ and reversing the sign of this expression, one defines the following function to maximize with respect to $\mathbf{\Theta}$:

$$\mathcal{J}(W,\mathbf{\Theta}) = \iint_D \Big(W(x,t) \log Q(x,t;\mathbf{\Theta}) - Q(x,t;\mathbf{\Theta})\Big) \mathrm{d}x\mathrm{d}t. \tag{4}$$

Using this function $\mathcal{J}$, one can derive the likelihood of the parameter $\mathbf{\Theta}$:

$$P(W|\mathbf{\Theta}) \triangleq e^{\mathcal{J}(W,\mathbf{\Theta}) - \iint_D \log\Gamma(1+W(x,t))\mathrm{d}x\mathrm{d}t} \tag{5}$$

where $\Gamma(\cdot)$ is the Gamma function, and the second part of the exponent ensures that we obtain a probability measure. One can indeed see this probability as the joint probability of all the variables $W(x,t)$ independently following Poisson distributions of

parameter $Q(x,t)$. This way of presenting the problem enables us to interpret it as a maximum *a posteriori* (MAP) estimation problem and to introduce prior functions on the parameters as follows, using Bayes theorem:

$$\begin{aligned}\widehat{\boldsymbol{\Theta}}_{\text{MAP}} &= \operatorname*{argmax}_{\boldsymbol{\Theta}} P(\boldsymbol{\Theta}|W) \\ &= \operatorname*{argmax}_{\boldsymbol{\Theta}} \left(\log P(W|\boldsymbol{\Theta}) + \log P(\boldsymbol{\Theta})\right) \\ &= \operatorname*{argmax}_{\boldsymbol{\Theta}} \left(\mathcal{J}(W,\boldsymbol{\Theta}) + \log P(\boldsymbol{\Theta})\right). \end{aligned} \tag{6}$$

Our goal is now equivalent to the maximization with respect to $\boldsymbol{\Theta}$ of $\mathcal{J}(W,\boldsymbol{\Theta}) + \log P(\boldsymbol{\Theta})$. In the following, we will write simply $\mathcal{J}(\boldsymbol{\Theta})$ for $\mathcal{J}(W,\boldsymbol{\Theta})$. The problem is that in the term $\iint_D W(x,t) \log \sum_{k,n,y} S_{kny}(x,t;\boldsymbol{\Theta}) \mathrm{d}x\mathrm{d}t$, there is a sum inside the logarithm, and we thus cannot obtain an analytical solution. However, if we introduce nonnegative membership degrees $m_{kny}(x,t)$ summing to 1 for each (x,t), one can write, using the concavity of the logarithm:

$$\begin{aligned}\log \sum_{k,n,y} S_{kny}(x,t;\boldsymbol{\Theta}) &= \log \sum_{k,n,y} m_{kny}(x,t) \frac{S_{kny}(x,t;\boldsymbol{\Theta})}{m_{kny}(x,t)} \\ &= \log \left\langle \frac{S_{kny}(x,t;\boldsymbol{\Theta})}{m_{kny}(x,t)} \right\rangle_m \\ &\geq \left\langle \log \frac{S_{kny}(x,t;\boldsymbol{\Theta})}{m_{kny}(x,t)} \right\rangle_m \\ &\geq \sum_{k,n,y} m_{kny}(x,t) \log \frac{S_{kny}(x,t;\boldsymbol{\Theta})}{m_{kny}(x,t)} \end{aligned} \tag{7}$$

where $\langle\cdot\rangle_m$ denotes the convex combination with coefficients m. Moreover, the inequality (7) becomes an equality for

$$\hat{m}_{kny}(x,t) = \frac{S_{kny}(x,t;\boldsymbol{\Theta})}{\sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{y=0}^{Y-1} S_{kny}(x,t;\boldsymbol{\Theta})}.$$

We can thus use an EM-like algorithm to maximize the likelihood by alternately updating $\boldsymbol{\Theta}$ and the membership degrees m, which act as auxiliary parameters, while keeping the other fixed:

$$\text{(E-step)}\ \ \hat{m}_{kny}(x,t) = \frac{S_{kny}(x,t;\boldsymbol{\Theta})}{\sum_{k=1}^{K} \sum_{n=1}^{N} \sum_{y=0}^{Y-1} S_{kny}(x,t;\boldsymbol{\Theta})}$$

$$\text{(M-step)}\ \ \widehat{\boldsymbol{\Theta}} = \operatorname{argmax}_{\boldsymbol{\Theta}} \left(\mathcal{J}(\boldsymbol{\Theta},\hat{m}) + \log P(\boldsymbol{\Theta})\right)$$

with

$$\mathcal{J}(\boldsymbol{\Theta},m) \triangleq \iint_D \left(\sum_{k,n,y} \ell_{kny}(x,t) \log \frac{S_{kny}(x,t;\boldsymbol{\Theta})}{m_{kny}(x,t)} - Q(x,t;\boldsymbol{\Theta}) \right) \mathrm{d}x\mathrm{d}t \tag{8}$$

where $\ell_{kny}(x,t) = m_{kny}(x,t) W(x,t)$. For all m, we indeed have from (7) that

$$\mathcal{J}(\boldsymbol{\Theta}) + \log P(\boldsymbol{\Theta}) \geq \mathcal{J}(\boldsymbol{\Theta},m) + \log P(\boldsymbol{\Theta}) \tag{9}$$

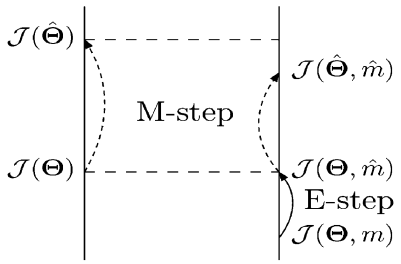


Fig. 2. Optimization through the EM algorithm. During the E-step, the auxiliary parameter m is updated to $\hat{m}$ so that $\mathcal{J}(\boldsymbol{\Theta}) = \mathcal{J}(\boldsymbol{\Theta},\hat{m})$. Then, during the M-step, $\mathcal{J}(\boldsymbol{\Theta},\hat{m})$ is optimized w.r.t. $\boldsymbol{\Theta}$, ensuring that $\mathcal{J}(\hat{\boldsymbol{\Theta}}) \geq \mathcal{J}(\hat{\boldsymbol{\Theta}},\hat{m}) > \mathcal{J}(\boldsymbol{\Theta},\hat{m}) = \mathcal{J}(\boldsymbol{\Theta})$. The maximization of $\mathcal{J}(\boldsymbol{\Theta})$ can thus be performed through the maximization of the auxiliary function $\mathcal{J}(\boldsymbol{\Theta},m)$ alternately w.r.t. m and $\boldsymbol{\Theta}$.

and $\mathcal{J}(\boldsymbol{\Theta},\hat{m})$ can be used as an auxiliary function to maximize, enabling us to obtain analytical update equations. The optimization process is illustrated in Fig. 2.

A slightly different approach was presented in [9], but leads to the same optimization process as the one we present here. The E-step is straightforward and is dealt with in exactly the same way as in [9]. However, the possibility to obtain analytical update equations during the M-step depends on the actual expression of $\mu_k(t)$, and it could be performed there in the special case of a piece-wise flat F_0 contour $\mu_k(t) = \mu_k$. More generally, if the other HTC parameters (w_k, τ_k, u_{ky}, v_{kn}, ϕ_k, σ_k) do not enter in the expression of $\mu_k(t)$, then the update equations obtained in [9] for these parameters can be used as is, and we only need to obtain the M-step update equations for the F_0 contour parameters. We will explain how we proceed in the following subsection.

B. Speech Modeling

In the following, in order to model the spectrum of a speech utterance, we will make several assumptions. First, we assume that the F_0 contour is smooth and defined on the whole interval: we will not make voiced/unvoiced decisions, and F_0 values are assumed continuous. Second, we fix the harmonic structure of each HTC source model so that it corresponds to segments of speech with steady acoustic characteristics, and assume that a speech segment is a succession of such steady segments sharing a common F_0 contour.

1) Spline F_0 Contour: In our previous work [9], the HTC method has only been applied to piece-wise flat F_0 contours, which is relevant for certain instruments like the piano for example, but of course not in speech. Looking for a smooth F_0 contour, we chose to use cubic spline functions as a general class of smooth functions, so as to be able to deal in the future with a wide variety of acoustic phenomena (background music, phone ringing, etc). Moreover, their simple algebraic formulation enables us to optimize the parameters through the EM algorithm, as update equations can be obtained analytically. It may happen that the smooth F_0 assumption is invalid, such as at the end of utterances where F_0-halving can occur, but this problem is faced by all algorithms that exploit continuity of F_0, and the assumption is justified empirically in that including it tends to reduce overall error rates. Moreover, failure to track F_0-halving at the end of utterances is perhaps not too serious, as the halving is

usually not salient perceptively, other than as a roughness of indeterminate pitch. Furthermore, it is one of a wider class of irregular voicing phenomena (diplophony, creak) for which F_0 is hard to define [11].

The analysis interval is divided into subintervals $[t_i, t_{i+1})$ which are assumed of equal length. Following [12], the parameters of the spline contour model are then the values z_i of the F_0 at each bounding point t_i. The values z_i'' of the second derivative at those points are given by the expression $\boldsymbol{z}'' = M\boldsymbol{z}$ for a certain matrix M which can be explicitly computed offline, under the hypothesis that the first-order derivative is 0 at the bounds of the analysis interval. We can assume so if we set the bounds of the interval outside the region where there is speech. One can then classically obtain a simple algebraic formula for the contour $\mu(t;\boldsymbol{z})$ on the whole interval. For $t \in [t_i, t_{i+1})$,

$$\mu(t;\boldsymbol{z}) \triangleq \frac{1}{t_{i+1}-t_i}\left(z_i(t_{i+1}-t)+z_{i+1}(t-t_i) -\frac{1}{6}(t-t_i)(t_{i+1}-t)\left[(t_{i+2}-t)z_i''+(t-t_{i-1})z_{i+1}''\right]\right). \quad (10)$$

2) Optimization of the Model: To design our model, we further make the following assumptions. For simplicity, we first describe here the case of a single speaker, and the multiple speakers case will be presented in Section II-B.4.

We make all the source models inside the HTC method share the same F_0 contour: $\mu_k(t) = \mu(t)$, $\forall k$, by plugging the analytical expression (10) of the spline F_0 contour into (2), such that all the source models are driven by the same F_0 expression. Our intention is to have a succession in time of slightly overlapping source models which correspond if possible to successive phonemes, or at least to segments of the speech utterance with steady acoustic characteristics. As the structure is assumed harmonic, the model takes advantage of the voiced parts of the speech utterance, which it uses as anchors. When used on natural speech, if the unvoiced/silent parts are too long, it may happen that the spline contour becomes unstable, which can deteriorate the accuracy of the F_0 contour extraction immediately before or after a section of unvoiced speech, especially if the neighboring voiced parts are not strongly enough voiced. If they are not, we believe there is no particular incidence on the accuracy of the F_0 contour near unvoiced parts of the speech. The results of the experimental evaluations will show that this assumption is justified.

We also assume that inside a source model the same power envelope is used for all harmonics, as we want to isolate structures in the speech flow with stable acoustic characteristics. The model allows source models to overlap, so a given spectral shape can merge progressively into another, which allows it to fit arbitrary spectro-temporal shapes. The subscript n can thus be excluded in u_{kny}. We note however that the following discussion on the optimization of the model is independent of this particular assumption and that the algorithm is general enough to allow separate power envelope functions.

The optimization process goes as follows: we start by updating the HTC parameters which do not enter in the spline model (namely w_k, τ_k, u_{ky}, v_{kn}, ϕ_k, σ_k) through analytical update equations which are the same as in [9]. Once these parameters have been updated, we compute the derivatives of $\mathcal{J}(\boldsymbol{\Theta}, \hat{m})$ with respect to the spline parameters z_j. A global maximum is difficult to obtain, but we can update the z_j one after the other, starting for example from z_0 and using the already updated parameters for the update of the next one. This way of performing the updates is referred to as the coordinate descent method [13], and can be summarized as

$$\begin{cases} z_0^{(p)} \leftarrow \underset{z_0}{\operatorname{argmax}}\, \mathcal{J}\left(z_0, z_1^{(p-1)}, \ldots, z_n^{(p-1)}, \boldsymbol{\Theta}_{-\boldsymbol{z}}, \hat{m}\right) \\ z_1^{(p)} \leftarrow \underset{z_1}{\operatorname{argmax}}\, \mathcal{J}\left(z_0^{(p)}, z_1, z_2^{(p-1)}, \ldots, z_n^{(p-1)}, \boldsymbol{\Theta}_{-\boldsymbol{z}}, \hat{m}\right) \\ \quad\vdots \\ z_n^{(p)} \leftarrow \underset{z_n}{\operatorname{argmax}}\, \mathcal{J}\left(z_0^{(p)}, \ldots, z_{n-1}^{(p)}, z_n, \boldsymbol{\Theta}_{-\boldsymbol{z}}, \hat{m}\right) \end{cases} \quad (11)$$

where $\boldsymbol{\Theta}_{-\boldsymbol{z}}$ denotes the set of whole parameters except the z_j. The corresponding optimization procedure, called the expectation-constrained maximization algorithm (ECM) [14], does not ensure the maximization in the M-step but guarantees the increase of the function $\mathcal{J}(\boldsymbol{\Theta}, \hat{m})$. Putting the derivatives with respect to z_j to 0, one then finds update equations analytically at step p:

$$z_j^{(p)} = \frac{\sum_{k,n,y} \iint_D \left(x - \hat{\mu}_j^{(n)}\left(t;\boldsymbol{z}^{(j,p)}\right)\right)\frac{\partial\mu}{\partial z_j}(t)\frac{\ell_{kny}^{(p-1)}(x,t)}{{\sigma_k^{(p)}}^2}\mathrm{d}x\mathrm{d}t}{\sum_{k,n,y}\iint_D \left(\frac{\partial\mu}{\partial z_j}(t)\right)^2 \frac{\ell_{kny}^{(p-1)}(x,t)}{{\sigma_k^{(p)}}^2}\mathrm{d}x\mathrm{d}t} \quad (12)$$

where $\boldsymbol{z}^{(j,p)} = (z_0^{(p)}, \ldots, z_{j-1}^{(p)}, z_j^{(p-1)}, z_{j+1}^{(p-1)}, \ldots, z_n^{(p-1)})$ and $\hat{\mu}_j^{(n)}(t;\boldsymbol{z}^{(j,p)}) = \mu(t;\boldsymbol{z}^{(j,p)}) - (\partial\mu/\partial z_j)(t) z_j^{(p)} + \log n$ does not depend on z_j and $(\partial\mu/\partial z_j)(t)$ only depends on t and the fixed matrix M.

The partial derivatives with respect to the other parameters (w_k, τ_k, u_{ky}, v_{kn}, ϕ_k, σ_k) are the same as in [9], as mentioned above.

3) Prior Distribution: As seen in Section II-A, the optimization of our model can be naturally extended to a MAP estimation by introducing prior distributions $P(\boldsymbol{\Theta})$ on the parameters, which work as penalty functions that try to keep the parameters within a specified range. The parameters which are the best compromise with empirical constraints are then obtained through equation (6).

By introducing such a prior distribution on v_{kn}, it becomes possible to prevent half-pitch errors, as the resulting source model would usually have a harmonic structure with zero power for all the odd order harmonics, which is abnormal for speech. We apply the Dirichlet distribution, which is explicitly given by

$$p(\boldsymbol{v}_k) \triangleq \frac{\Gamma\left(\sum_n (d_v \bar{v}_n + 1)\right)}{\prod_n \Gamma(d_v \bar{v}_n + 1)} \prod_n v_{kn}^{d_v \bar{v}_n} \quad (13)$$

where $\bar{v}_n$ is the most preferred "expected" value of v_{kn} such that $\sum_n \bar{v}_n = 1$, d_v the contribution degree of the prior and $\Gamma(\cdot)$ the Gamma function. The maximum value for $p(\boldsymbol{v}_k)$ is taken when $v_{kn} = \bar{v}_n$ for all n. When d_v is zero, $p(\boldsymbol{v}_k)$ become uniform

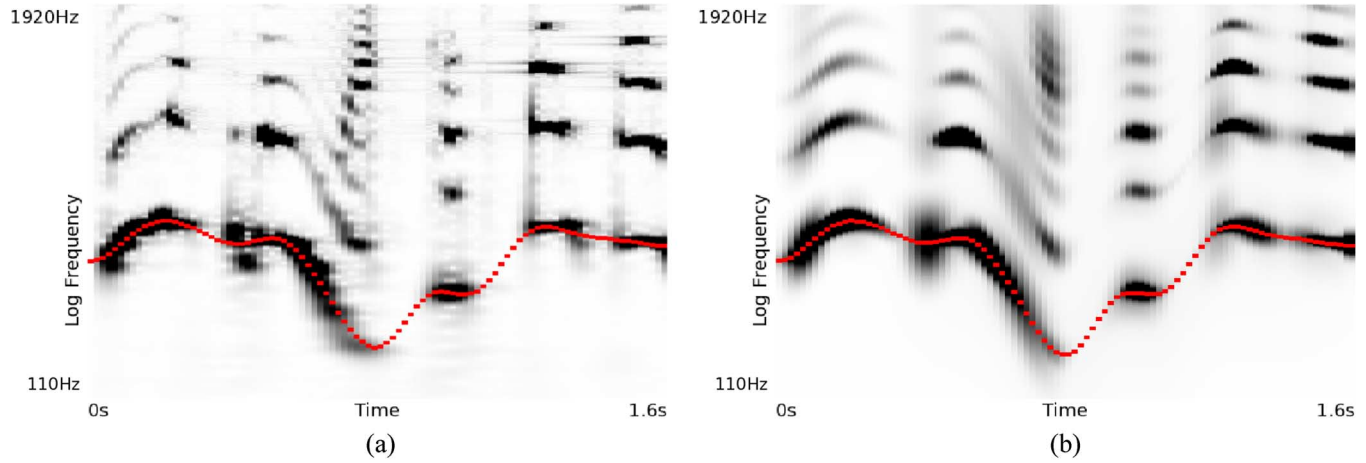


Fig. 3. Comparison of observed and modeled spectra ("Tsuuyaku denwa kokusai kaigi jimukyoku desu," female speaker). The estimated F_0 contour is reproduced on both the observed and modeled spectrograms to show the precision of the algorithm. (a) Observed spectrogram and estimated F_0 contour. (b) Modeled spectrogram and estimated F_0 contour.

distributions. The choice of this particular distribution allows us to give an analytical form of the update equations of v_{kn}.

Although the spline model can be used as is, one can also introduce in the same way a prior distribution on the parameters z_j of the spline F_0 contour, in order to avoid an overfitting problem with the spline function. Indeed, spline functions have a tendency to take large variations, which is not natural for the F_0 contour of a speech utterance. Moreover, the F_0 contour might also be hard to obtain on voiced parts with relatively lower power or poor harmonicity. The neighboring voiced portions with higher power help the estimation over these intervals by providing a good prior distribution.

To build this prior distribution, we assume that the z_j form a Markov chain, such that

$$P(z_0,\ldots,z_n) = P(z_0)\prod_{j=1}^{n} P(z_j|z_{j-1})$$

and assume furthermore that z_0 follows a uniform distribution and that, conditionally to z_{j-1}, z_j follows a Gaussian distribution of center z_{j-1} and variance σ_s^2 corresponding to the weighting parameter of the prior distribution:

$$P(z_j|z_{j-1}) = \frac{1}{\sqrt{2\pi}\sigma_s} e^{-\frac{(z_j - z_{j-1})^2}{2\sigma_s^2}}.$$

In the derivative with respect to z_j used above to obtain (12) add up two new terms

$$\frac{\partial \log P(z_j|z_{j-1})}{\partial z_j} + \frac{\partial \log P(z_{j+1}|z_j)}{\partial z_j}$$

and the update equation (12) then becomes

$$z_j^{(p)} = \frac{\frac{2}{\sigma_s^2}\cdot\frac{z_{j-1}^{(p)}+z_{j+1}^{(p-1)}}{2} + A_j^{(p)}}{\frac{2}{\sigma_s^2} + B_j^{(p)}} \qquad (14)$$

where $A_j^{(p)}$ and $B_j^{(p)}$ are, respectively, the numerator and denominator of the right term of equation (12). The update equation for the boundary points is derived similarly.

An example is presented in Fig. 3, based on the Japanese sentence "Tsuuyaku denwa kokusai kaigi jimukyoku desu" uttered by a female speaker. Shown are 2-D representations of the observed and modeled spectra (after 30 iterations of the estimation algorithm). The F_0 contour estimated through our method is reproduced on both the observed and modeled spectrograms to show the precision of our algorithm. One can see that the model approximates well the spectrum and that the F_0 contour is accurately estimated.

4) Multiple F_0 Estimation: The multiple-speakers case is a simple extension of the single-speaker one. While for a single F_0 estimation all the source models share the same F_0 contour, for multiple speakers, according to the number of F_0 contours that we want to estimate, we group together source models into subsets such that source models inside a subset share a common F_0 contour. For example, if we use $K = 10$ models in total for two speakers, the models with index $k \in \{1,\ldots,5\}$ will be attributed to the first speaker, while the others will be attributed to the second one. We thus have pools of source models driven by a single F_0 contour for each of the pools and corresponding to one of the speakers, and we only need to introduce a set of spline parameters for each of the F_0 contours. These sets can be optimized independently and simultaneously in the exact same way as in the single-speaker case: the E-step is unchanged, and the M-step is performed by first updating the HTC parameters which do not enter in the spline model, and then updating the spline parameters of each F_0 contour through the ECM algorithm. This last update can be done independently as the derivatives of $\mathcal{J}(\boldsymbol{\Theta}, \hat{m})$ with regards to the parameters of one of the F_0 contours do not include any term depending on the parameters of the other contours.

The method handles overlapping harmonics by, at each iteration of the algorithm, re-estimating simultaneously the contributions of each voice in the spectrum. It relies on the assumption that the spectra of the contributing sources combine additively, and neglects phase-dependent vector summation. This is

of course a rough approximation in the view of perfect source separation. However, the speech spectrum is usually relatively sparse, and the F_0 estimation can rely mostly on the components that are isolated from each other. This is the reason why we think this rough approximation is not a serious problem, as F_0 estimation is our main objective in this paper.

C. Noise Modeling

We introduce a noise model to cope with the background noise that can be a disturbance in the process of clustering the harmonic portions of speech. Indeed, it would be more rewarding, in the purpose of decreasing the $\mathcal{I}$-divergence, for the harmonic source models to take very large variances in the log-frequency direction and have the centers of the Gaussian distribution go on portions of the spectrogram with strong broad power, even though there is no harmonic structure corresponding to these portions, especially if the noise power is comparable to or even larger than the speech signal power.

The spectrogram of a quasi-periodic signal such as voiced speech consists of a large number of line spectrum components and has spikes that are strongly marked, while the spectrogram of a white or pink noise has a tendency to be more flat, without significant spiky regions. The idea to design the noise model was thus that detecting the harmonic parts of the spectrogram in a noisy background corresponds to searching for thin and harmonically distributed "islands" which rise out of a "sea" of noise. We thus chose to model the noise using a mixture of Gaussian distributions with large fixed variance and with centers fixed on a grid, the only parameters of the model being the respective weights of the Gaussians in the model and the ratio of noise power inside the whole spectrogram. This noise-cancelling approach can be considered quite close to spectral subtraction in the sense that the power spectrum is in both cases assumed additive. However, while spectral subtraction generally needs voiced/unvoiced decision to obtain the power spectrum of the background noise, which furthermore has to be assumed stationary, our approach estimates adaptively the noise part of the spectrum even when there is speech, taking advantage of the valleys of the spectral structure. The only assumption we make is that the noise spectrum is smooth in both time and frequency direction, whereas speech spectrum is more spiky in the frequency direction. Therefore, we can expect our model's performance to be close to the best performance that spectral subtraction could reach.

Let N_c be the number of columns of the grid (in the time direction) and N_r the number of rows (in the log-frequency direction). The noise model is defined as

$$\mathcal{N}(x,t;\boldsymbol{\Theta}) = \rho \sum_{n=1}^{N_r}\sum_{y=1}^{N_c} w_{ny}^{\mathcal{N}} \frac{1}{2\pi\sigma_r^{\mathcal{N}}\sigma_c^{\mathcal{N}}} e^{-\frac{(x-\alpha_n)^2}{2(\sigma_r^{\mathcal{N}})^2}} e^{-\frac{(t-\beta_y)^2}{2(\sigma_c^{\mathcal{N}})^2}} \tag{15}$$

where ρ is the proportion of the noise model in the total model, the $w_{ny}^{\mathcal{N}}$ are the weights of the Gaussian functions in the mixture and add up to 1, $\boldsymbol{\sigma}^{\mathcal{N}} = (\sigma_r^{\mathcal{N}}, \sigma_c^{\mathcal{N}})$ the variances of the Gaussian functions, α_n the log-frequency index of the centers of the nth row, and β_y the time index of the centers of the yth column.

If we note $\mathcal{N}(x,t;\boldsymbol{\Theta}) = \sum_{n=1}^{N_r}\sum_{y=1}^{N_c} S_{0ny}(x,t;\boldsymbol{\Theta})$ and introduce $\ell_{0ny}(x,t)$ and $m(0,n,y;x,t)$ as in Section II-A, the EM algorithm can be applied in the same way as described above, just differing in the range of the summations on k, n, and y. We only need to specify the update equations of the M-step for the noise model parameters

$$\rho^{(p+1)} = \sum_{n=1}^{N_r}\sum_{y=1}^{N_c} \iint_D \ell_{0ny}^{(p)}(x,t)\mathrm{d}x\mathrm{d}t \tag{16}$$

$$w_{ny}^{\mathcal{N}\,(p+1)} = \frac{1}{\rho^{(p+1)}} \iint_D \ell_{0ny}^{(p)}(x,t)\mathrm{d}x\mathrm{d}t \tag{17}$$

where the superscript (p) refers to the iteration cycle. The update equations for the other parameters remain the same, the only changes coming from the E-step where the noise model now enters in the summation.

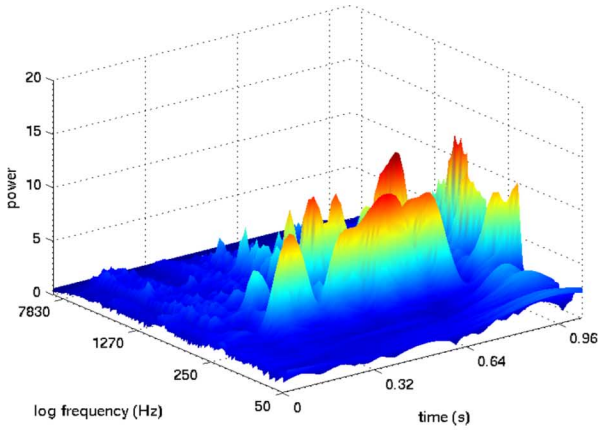
The noise cancelling performed through the introduction of the noise model is illustrated in Fig. 4: Fig. 4(a) gives a 3-D view of the original clean spectrogram of a part of the sentence "It was important to be perfect since there were no prompts" uttered by a female speaker, and Fig. 4(b) shows the spectrogram of the same part of the utterance to which white noise at a signal-to-noise ratio (SNR) of -2 dB has been added. The estimation of the spectrogram where the noise has been cancelled, shown in Fig. 4(c), is obtained through the optimized masking functions $\hat{m}_{kny}(x,t)$ as in the following formula:

$$\sum_{k=1}^{K}\sum_{n=1}^{N}\sum_{y=0}^{Y-1} \hat{m}_{kny}(x,t)W(x,t). \tag{18}$$

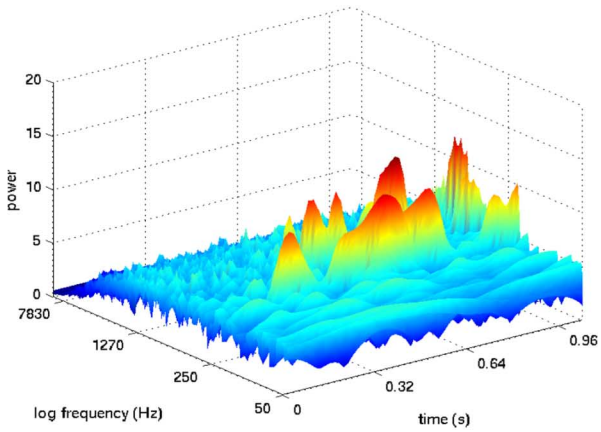
In the course of the optimization of the model, the F_0 contour estimation is performed on this "cleaned" part of the spectrogram, which enables our F_0 estimation algorithm to perform well even in very noisy environments, as we will show in the next section.

D. Parametric Representation and Potential Applications

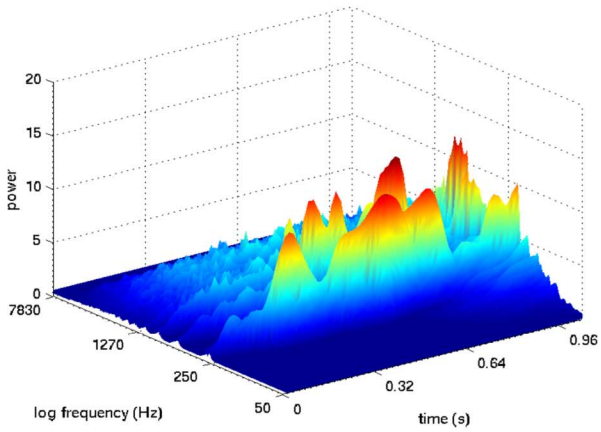
We would like to stress the fact that our algorithm not only estimates the F_0 contour, but also gives a parametric representation of the voiced parts of the spectrogram. This can be useful especially in the analysis of cochannel speech by multiple speakers, as one can get a parametric representation of the harmonic parts of the separated spectrograms of each utterance, as shown in Fig. 5: Fig. 5(b) represents the modeled spectrogram of the Japanese utterance "oi wo ou" by a male speaker, and Fig. 5(c) the one of the utterance "aoi" by a female speaker, extracted from the mixed signal shown in Fig. 5(a) (the F_0 contour near the boundary is not relevant as there is no sound by the second speaker there). These parametric representations could be used for example to cluster the spectrogram of the mixed sound and separate the speakers, as well as for noise cancelling, as we showed in Fig. 4. The reader might observe that the harmonic tracks at the end of the utterance in Fig. 5(b) possess greater energy than their original counterparts. This might be related to the fact that the power envelope functions of all the harmonics are linked together. We shall note however that this



(a) Original clean spectrogram

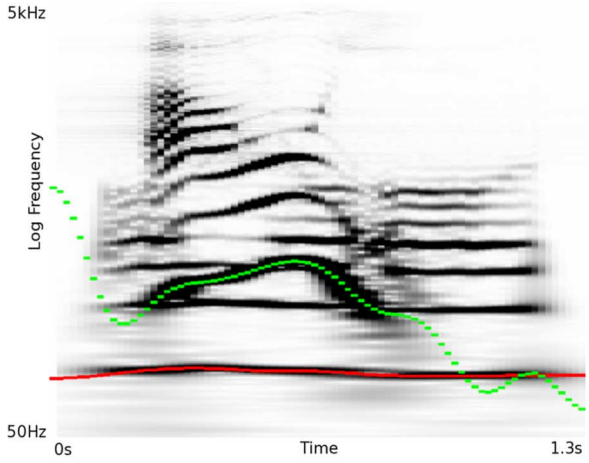


(b) Noisy spectrogram (clean speech mixed with white noise, SNR=-2dB)

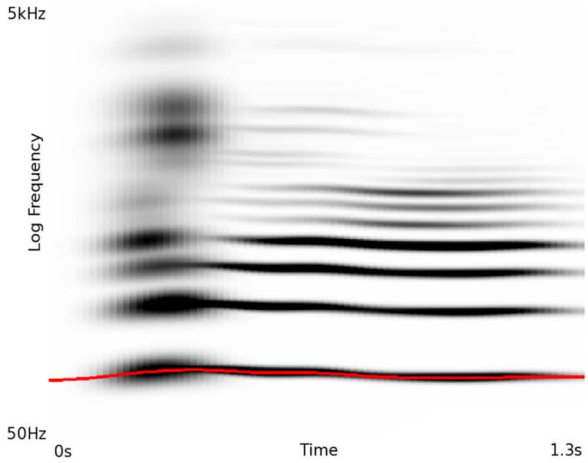


(c) Estimated noise-cancelled part of the spectrogram

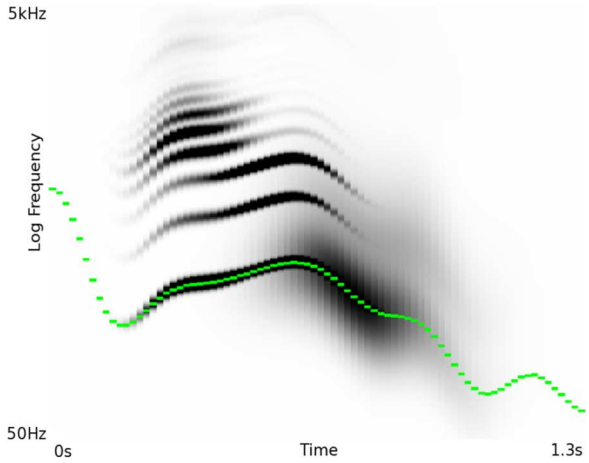
Fig. 4. Estimation of the clean part of a noisy spectrogram. (a) shows a 3-D view of the original clean spectrogram of a part of the sentence "It was important to be perfect since there were no prompts" uttered by a female speaker. (b) shows the spectrogram of the same part of the utterance to which white noise at an SNR of -2 dB has been added. (c) shows the estimated noise-cancelled part of the spectrogram of (b).



(a) Original Spectrogram and estimated F_0 contour



(b) Modeled spectrogram, speaker 1



(c) Modeled spectrogram, speaker 2

Fig. 5. Parametric representation of separated spectrograms. (a) shows the spectrogram of a signal obtained by mixing the two Japanese utterances "oi wo ou" by a male speaker and "aoi" by a female speaker, together with the F_0 contours estimated by our method. (b) and (c) show the parametric representations of the spectrograms of the utterances by the male and female speaker, respectively, extracted from the mixed signal shown in (a).

is not an important issue for the applications that we foresee, such as speech enhancement or clustering, because we plan to use the parametric models only to obtain the proportions of each audio object inside the original spectrogram, and then multiply them by the original spectrogram to extract the desired object, as in (18). If the original spectrogram had low energy at a certain time-frequency point, the obtained spectrogram will have low energy as well.

III. Experimental Evaluation

A. Single-Speaker F_0 Estimation in Clean Environment

We evaluated the accuracy of the F_0 contour estimation of our model on a database of speech recorded together with a laryngograph signal [15], consisting of one male and one female speaker who each spoke 50 English sentences for a total of 0.12 h of speech, for the purpose of evaluation of F_0-estimation algorithms.

The power spectrum $W(x,t)$ was calculated from an input signal digitized at a 16-kHz sampling rate (the original data of the database was converted from 20 kHz to 16 kHz) using a Gabor wavelet transform with a time resolution of 16 ms for the lowest frequency subband. Higher subbands were downsampled to match the lowest subband resolution. The lower bound of the frequency range and the frequency resolution were, respectively, 50 Hz and 14 cent. The spline contour was initially flat and set to 132 Hz for the male speaker and 296 Hz for the female speaker. These values were tuned in frequency bins after a few preliminary experiments. We shall note that, as we were able to obtain very good performance with the same initial conditions on various data, we believe that the performance is not very sensitive to the priming of the F_0 contour parameters. The length of the interpolation intervals was fixed to four frames. For HTC, we used $K = 10$ source models, each of them with $N = 10$ harmonics. This is enough for F_0 estimation. For a better modeling of the spectrogram, one can use 40 or 60 harmonics for example. Temporal envelope functions were modeled using $Y = 3$ Gaussian kernels. The w_k were set to $1/K$, u_{ky} to $1/Y$, τ_k to $T_0+(k-1)T/K$, ϕ_k to 32 ms, and σ_k to 422 cents. For the prior functions, σ_s was fixed to 0.4, d_v to 0.04, and $(\overline{v}_n)_{1\leq n\leq N} = (1/N)(8,8,4,2,1,\ldots,1)$. The algorithm was run on the utterances from which the initial and final silence parts were manually removed. We note that the minimum time window on which the algorithm can work is one frame, in which case the algorithm works frame-by-frame. This gives acceptable results as described in [16], but perfomance improves with longer windows, as documented in [9]. It is thus used here on the whole time interval. The computational cost is also discussed in [9].

We used as ground truth the F_0 estimates and the reliability mask derived by de Cheveigné *et al.* [2] under the following criteria: 1) any estimate for which the F_0 estimate was obviously incorrect was excluded; and (2) any remaining estimate for which there was evidence of vocal fold vibration was included. Frames outside the reliability mask were not taken into account during our computation of the accuracy, although our algorithm gives values for every point of the analysis interval by construction. As the spline function gives an analytical expression for the F_0 contour, we compare our result with the reference values at a sampling rate of 20 kHz although all the analysis was performed with a time resolution of 16 ms.

Deviations over 20% from the reference were deemed to be gross errors. The results can be seen in Table I, with for comparison the results obtained by de Cheveigné *et al.* [2] for several other algorithms. Notations stand for the method used, as follows. **ac**: Boersma's autocorrelation method [17], [18], **cc**: cross-correlation [18], **shs**: spectral subharmonic summation [18], [19], **pda**: eSRPD algorithm [15], [20], **fxac**: autocorrelation function (ACF) of the cubed waveform [21], **fxcep**: cepstrum [21], **additive**: probabilistic spectrum-based method [22], **acf**: ACF [2], **nacf**: normalized ACF [2], **TEMPO**: the TEMPO algorithm [23], **YIN**: the YIN algorithm [2]. More details concerning these algorithms can be found in [2]. We can see that our model's accuracy for clean speech is comparable to the best existing single-speaker F_0 extraction algorithms designed for that purpose.

TABLE I
Gross Error Rates for Several F_0 Estimation Algorithms on Clean Single-Speaker Speech

Method	Gross error (%)
pda	19.0
fxac	16.8
fxcep	15.8
ac	9.2
cc	6.8
shs	12.8
acf	1.9
nacf	1.7
additive	3.6
TEMPO	3.2
YIN	1.4
HTC (proposed)	3.5

B. Single F_0 Estimation on Speech Mixed With White and Pink Noise

We performed F_0 estimation experiments on speech to which a white noise, band-passed between 50 and 3300 Hz, was added, with SNRs of 0, -2, and -10 dB. These SNRs were selected because 0 dB corresponds to equal power, -2 dB is used in a study [24] that used the same database as we use in III-C, and -10 dB is a relatively noisy condition to illustrate the effectiveness of our algorithm in that case. The database mentioned above [15] was again used, and the white noise added was generated independently for each utterance. We also performed experiments with pink noise, band-passed between 50 and 3300 Hz, with an SNR of -2 dB. The spectrum of pink noise is closer to that of speech than white noise. The noise model was initialized with $\rho = 0.1$ and the $w_{ny}^{\mathcal{N}}$ all equal to $1/(N_c N_r)$, while the variances were fixed to $\sigma_r^{\mathcal{N}} = 1120$ cent and $\sigma_c^{\mathcal{N}} = \sigma_r^{\mathcal{N}}/3$, and the centers (β_y, α_n) of the Gaussian distributions of the noise model were fixed on a grid such that the distances between them in the time and log-frequency directions were all equal to $\sigma_c^{\mathcal{N}}$ and $\sigma_r^{\mathcal{N}}$, respectively, to ensure a good overlap between the Gaussian distributions. The determination of the variances was made after a few experiments while keeping in mind that $\sigma_r^{\mathcal{N}}$ should be significantly larger than the typical variance in the log-frequency direction of the Gaussian of the harmonic model but small enough to still be able to model fluctuations in the noise power.

As a comparison, we present results obtained on the same database using YIN [2] and the algorithm of Wu, Wang, and Brown [4], specifically designed for F_0 tracking in a noisy environment, and that can also handle the estimation of two simultaneous F_0s. Their code is made available on the Internet by their respective authors. The algorithm of Wu, Wang, and Brown will be referred to as the WWB algorithm. According to

TABLE II
ACCURACY (%) OF THE F_0 ESTIMATION OF SINGLE-SPEAKER SPEECH MIXED WITH WHITE AND PINK NOISES

	HTC (YIN,WWB)			
	White noise			Pink noise
	SNR=0dB	SNR=-2dB	SNR=-10dB	SNR=-2dB
Female speaker	95.8 (83.5, 56.1)	96.3 (77.8, 48.8)	88.2 (36.7, 09.0)	91.9 (46.1, 44.6)
Male speaker	92.1 (82.5, 69.3)	92.2 (77.2, 59.2)	79.7 (41.5, 19.2)	74.0 (58.1, 37.6)
Total	94.0 (83.0, 62.5)	94.3 (77.5, 53.8)	84.1 (39.0, 13.9)	83.2 (51.9, 41.2)

TABLE III
CATEGORIZATION OF INTERFERENCE SIGNALS

	Interference signals
Category 1	White noise, noise bursts
Category 2	1kHz tone, "cocktail party" noise, rock music, siren, trill telephone
Category 3	Female utterance 1, male utterance, female utterance 2

TABLE IV
ACCURACY (%) OF THE F_0 ESTIMATION OF VOICED SPEECH WITH SEVERAL KINDS OF INTERFERENCES

	HTC	WWB	YIN
Category 1	99.7	90.8	93.1
Category 2	98.6	96.1	75.7
Category 3	99.5	97.8	87.1

TABLE V
F_0 ESTIMATION OF CONCURRENT SPEECH BY MULTIPLE SPEAKERS, GROSS ERROR FOR A DIFFERENCE WITH THE REFERENCE HIGHER THAN 20% AND 10%

Gross error threshold	20%		10%	
	HTC	WWB	HTC	WWB
Male-Female	93.3	81.8	86.8	81.5
Male-Male	96.1	83.4	87.9	69.0
Female-Female	98.9	95.8	95.6	90.8
Total	96.1	87.0	90.2	83.5

its authors, the parameters of this algorithm could be tuned on a new database to obtain the best performances, but they mention [4] that it is supposed to work fine in the version made available (trained on a corpus [24] that we will use later).

We obtained good results, presented in Table II, showing the robustness of our method on noisy speech, when noise is not harmonic. Note that the level of the noise added is greater than that of the original signal, and that at -10 dB it is even difficult for human ears to follow the pitch of the original signal. The harmonic structure of our model is effective for detecting speech in the presence of background noise. YIN and the WWB algorithm were both outperformed, although we should note again that their code was used as is, whereas ours was developed with the task in mind. Thus, this comparison may not do them full justice.

C. Validation on a Corpus of Speech Mixed With a Wide Range of Interferences

In order to show the wide applicability of our method, we also performed experiments using a corpus of 100 mixtures of voiced speech and interference [24], commonly used in CASA research. In [4], half of the corpus is used for model parameter estimation and the other half for system evaluation. As it is not specified in that paper which part of the corpus was used for which purpose, we decided to use the full corpus as the evaluation set for comparison of the algorithms, which can only be an advantage for the WWB algorithm. The results we present for the WWB algorithm differ from the ones given in [4] as the criterion we use is different. To be able to compare it with our method, which does not perform a voiced/unvoiced decision, we do not take into account errors on the estimation of the number of F_0s, but only look at the accuracy of the output of the pitch determination algorithm. Moreover, we focus on the F_0 estimation of the main voiced speech, as we want here to show that our algorithm robustly estimates the F_0 in a wide range of noisy environments. The ten interferences are grouped into three categories: 1) those with no pitch; 2) those with some pitch qualities; and 3) other speech, as shown in Table III. The reference F_0 contours for the ten voiced utterances were built using YIN on the clean speech and manually corrected.

The experiments were performed in the same conditions as described in Section III-B for HTC, and the results are presented in Table IV. One can see that our algorithm again outperforms YIN and the WWB algorithm in all the interference categories.

D. Multipitch Estimation

We present here results on the estimation of the F_0 contour of the cochannel speech of two speakers speaking simultaneously with equal average power. We used again the database mentioned above [15] and produced a total of 150 mixed utterances, 50 for each of the "male–male," "female–female," and "male–female" patterns, using each utterance only once and mixing it with another such that two utterances of the same sentence were never mixed together. We used our algorithm in the same experimental conditions as described in Section III-A for clean single-speaker speech, but using two spline F_0 contours. The spline contours were initially flat and set to 155 and 296 Hz in the male–female case, 112 and 168 Hz in the male–male case, and 252 and 378 Hz in the female–female case.

The evaluation was done in the following way: only times inside the reliability mask of either of the two references were counted; for each reference point, if either one of the two spline F_0 contours lies within a criterion distance of the reference, we considered the estimation correct. We present scores for two criterion thresholds: 10% and 20%. For comparison, tests using the WWB algorithm [4] introduced earlier were also performed, using the code made available by its authors. YIN could not be used as it does not perform multipitch estimation. Results summarized in Table V show that our algorithm outperforms the WWB algorithm on this experiment.

IV. Conclusion and Future Work

We introduced a new model describing the spectrum as a sequence of spectral cluster models governed by a common F_0 contour function, with smooth transitions in the temporal succession of the spectral structures. The model enables an accurate estimation of the F_0 contour on the whole utterance by taking advantage of its voiced parts in clean as well as noisy environments. We explained how to optimize its parameters efficiently and performed several experiments to evaluate the accuracy of our model. On single-speaker clean speech, we obtained good results which we compared with existing methods specifically designed for that task. On cochannel concurrent speech, single-speaker speech mixed with white noise, pink noise, and on a corpus of single-speaker speech mixed with a variety of interfering sounds, we showed that our algorithm outperforms existing methods.

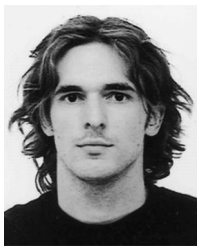
We are currently working on using the precise parametric expression and clustering of the spectrogram we obtained for noise cancelling, speech enhancement, and speech separation. We also intend to improve the modeling of the spectral structure using a new method that we describe in [25] in order to embed the formant structure into our model.

Acknowledgment

The authors would like to thank Dr. E. McDermott and P. Masurel for fruitful discussions and comments and the anonymous reviewers for their valuable comments which greatly helped improve the quality of this article.

References

[1] W. J. Hess, *Pitch Determination of Speech Signals*. New York: Springer, 1983.
[2] A. de Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 111, no. 4, pp. 1917–1930, 2002.
[3] A. de Cheveigné, "Multiple F_0 estimation," in *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*, D.-L. Wang and G. J. Brown, Eds. New York: IEEE Press/Wiley, 2006.
[4] M. Wu, D. Wang, and G. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 229–241, May 2003.
[5] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 708–716, Nov. 2000.
[6] Y. H. Gu and W. M. G. van Bokhoven, "Co-channel speech separation using frequency bin nonlinear adaptive filter," in *Proc. ICASSP*, 1991, pp. 949–952.
[7] A. Klapuri and M. Davy, *Signal Processing Methods for Music Transcription*. New York: Springer, 2005.
[8] H. Fujisaki and S. Nagashima, "A model for synthesis of pitch contours of connected speech," Faculty of Eng., Univ. Tokyo, Tokyo, Japan, 1969, vol. 28, pp. 53–60, Annu. Rep. Eng. Res. Inst..
[9] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 982–994, Mar. 2007.
[10] I. Csiszar, "I-divergence geometry of probability distributions and minimization problems," *Ann. Probability*, vol. 3, no. 1, pp. 146–158, 1975.
[11] P. Hedelin and D. Huber, "Pitch period determination of aperiodic speech signals," in *Proc. ICASSP*, 1990, pp. 361–364.
[12] W. H. Press, W. T. Vetterling, S. A. Teukolsky, and B. P. Flannery, *Numerical Recipes in C++: The Art of Scientific Computing*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2002.
[13] W. I. Zangwill, *Nonlinear Programming: A Unified Approach*. Englewood Cliffs, NJ: Prentice-Hall, 1969.
[14] X. L. Meng and D. B. Rubin, "Maximum likelihood estimation via the ECM algorithm," *Biometrika*, vol. 80, no. 2, pp. 267–278, 1993.
[15] P. C. Bagshaw, S. M. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of F_0 contours for computer and intonation teaching," in *Proc. Eurospeech*, 1993, pp. 1003–1006.
[16] H. Kameoka, T. Nishimoto, and S. Sagayama, "Separation of harmonic structures based on tied Gaussian mixture model and information criterion for concurrent sounds," in *Proc. ICASSP*, 2004, vol. 4, pp. 297–300.
[17] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proc. Inst. Phonetic Sci.*, 1993, vol. 17, pp. 97–110.
[18] P. Boersma and D. Weenin, "Praat System," [Online]. Available: http://www.fon.hum.uva.nl/praat/
[19] D. J. Hermes, "Measurement of pitch by subharmonic summation," *J. Acoust. Soc. Amer.*, vol. 83, pp. 257–264, 1988.
[20] Edinburgh Speech Tools Library. [Online]. Available: http://www.cstr.ed.ac.uk/
[21] Speech Filing System. [Online]. Available: http://www.phon.ucl.ac.uk/resource/sfs/
[22] B. Doval, "Estimation de la fréquence fondamentale des signaux sonores," Ph.D. dissertation, Univ. Pierre et Marie Curie, Paris, France, 1994.
[23] H. Kawahara, H. Katayose, A. de Cheveigné, and R. D. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F_0 and periodicity," in *Proc. Eurospeech*, 1999, vol. 6, pp. 2781–2784.
[24] M. P. Cooke, "Modeling auditory processing and organisation," Ph.D. dissertation, Univ. Sheffield, Sheffield, U.K., 1993.
[25] H. Kameoka, J. Le Roux, N. Ono, and S. Sagayama, "Speech analyzer using a joint estimation model of spectral envelope and fine structure," in *Proc. ICSLP*, 2006, pp. 2502–2505.



Jonathan Le Roux (S'06) received the degree in mathematics from the Ecole Normale Supérieure, Paris, France, the M.Sc. degree in partial differential equations from the University of Paris XI, Paris, in 2001, and the M.Sc. degree in stochastic processes from the University of Paris VI, Paris, in 2003. He is currently pursuing the Ph.D. degree at the Graduate School of Computer Science, Telecommunications and Electronics of Paris, University of Paris VI, and at the Graduate School of Information Science and Technology, Department of Information Physics and Computing, University of Tokyo, Tokyo, Japan.

His research interests include acoustic signal processing, speech processing, computational acoustic scene analysis and language acquisition modeling.

Mr. Le Roux is a student member of the Acoustical Society of Japan (ASJ) and the International Speech Communication Association (ISCA).

Hirokazu Kameoka (S'05) received the B.E. and M.E. degrees from the University of Tokyo, Tokyo, Japan, in 2002 and 2004, respectively. He is currently pursuing the Ph.D. degree at the Graduate School of Information Science and Technology, University of Tokyo.

His research interests include acoustic signal processing, speech processing, and music processing.

Mr. Kameoka is a student member of the Institute of Electronics, Information and Communication Engineers (IEICE), Information Processing Society of Japan (IPSJ), Acoustical Society of Japan (ASJ), and International Speech Communication Association (ISCA). He was awarded the Yamashita Memorial Research Award from IPSJ, Best Student Paper Award Finalist at IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), and the 20th Telecom System Technology Student Award from the Telecomunications Advancement Foundation (TAF), all in 2005.

Nobutaka Ono (M'02) received the B.E., M.S., and Ph.D degrees in mathematical engineering and information physics from the University of Tokyo, Tokyo, Japan, in 1996, 1998, 2001, respectively.

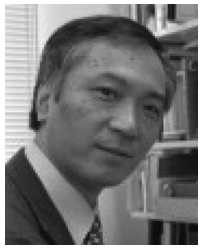
He joined the Graduate School of Information Science and Technology, University of Tokyo, in 2001 as a Research Associate and became a Lecturer in 2005. His research interests include acoustic signal processing, speech processing, music processing, sensing and measurement, and auditory modeling. He has been a Secretary of the technical committee of psychological and physiological acoustics in Japan since 2005.

He is a member of the Institute of Electrical Engineers of Japan (IEEJ), the Acoustical Society of America (ASA), the Acoustical Society of Japan (ASJ), the Society of Instrument and Control Engineers (SICE), and the Japan Society of Applied Physics (JSAP). He received the Sato Paper Award from the ASJ in 2000 and the Igarashi Award at the Sensor Symposium on Sensors, Micromachines, and Applied Systems from the IEEJ in 2004.

Alain de Cheveigné (M'06) received the M.Sc. degree in electronics from the University of Paris VI, Paris, France, in 1977, the Ph.D. degree in linguistics from the University of Paris VII in 1982, and the Habilitation in Neurosciences from Paris VI in 2000.

He is currently a CNRS Senior Researcher with the Audition Group, Laboratoire de Psychologie de la Perception, affiliated with CNRS, University of Paris V and Ecole Normale Supérieure. His research interests include psychophysics, modeling of hearing processes such as pitch and auditory scene analysis, speech and music signal processing such as F0 and spectral estimation, audio indexing and retrieval, and signal processing of magnetoencephalography signals.

Dr. de Cheveigné is member of the Acoustical Society of America (ASA), the Acoustical Society of Japan (ASJ), the Society for Neuroscience (SFN), and the Association for Research in Otolaryngology (ARO).



Shigeki Sagayama (M'82) was born in Hyogo, Japan, in 1948. He received the B.E., M.E., and Ph.D. degrees from the University of Tokyo, Tokyo, Japan, in 1972, 1974, and 1998, respectively, all in mathematical engineering and information physics.

He joined Nippon Telegraph and Telephone Public Corporation (currently, NTT) in 1974 and started his career in speech analysis, synthesis, and recognition at NTT Laboratories, Musashino, Japan. From 1990 to 1993, he was Head of the Speech Processing Department, ATR Interpreting Telephony Laboratories, Kyoto, Japan, pursuing an automatic speech translation project. From 1993 to 1998, he was responsible for speech recognition, synthesis, and dialog systems at NTT Human Interface Laboratories, Yokosuka, Japan. In 1998, he became a Professor of the Graduate School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), Ishikawa, Japan. In 2000, he was appointed Professor of the Graduate School of Information Science and Technology (formerly Graduate School of Engineering), University of Tokyo. His major research interests include processing and recognition of speech, music, acoustic signals, hand writing, and images. He was the leader of anthropomorphic spoken dialog agent project (Galatea Project) from 2000 to 2003.

Prof. Sagayama is a member of the Acoustical Society of Japan (ASJ), Institute of Electronics, Information and Communications Engineers, Japan (IEICEJ), and Information Processing Society of Japan (IPSJ). He received the National Invention Award from the Institute of Invention of Japan in 1991, the Chief Official's Award for Research Achievement from the Science and Technology Agency of Japan in 1996, and other academic awards including Paper Awards from IEICEJ in 1996 and from IPSJ in 1995.