Speech Analyzer Using a Joint Estimation Model of Spectral Envelope and Fine Structure

Hirokazu Kameoka, Jonathan Le Roux, Nobutaka Ono, Shigeki Sagayama

Graduate School of Information Science and Technology, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
{kameoka,leroux,onono,sagayama}@hil.t.u-tokyo.ac.jp

Abstract

We have been working on a new speech analyzer based on a parametric representation of speech governed by the F_0 parameter, towards practical human-machine interfaces. As a precise estimation of the frequency response of the vocal tract from a real speech signal requires the power of each component of the harmonic structure to be accurately estimated, one hopes to have a high-precision estimation of F_0. At the same time, under the empirical constraint that speech spectral envelopes are usually smooth in the power domain, half pitch errors can be significantly avoided. Therefore, F_0 and the envelope should be estimated jointly rather than separately through an optimal estimation of the spectral envelope and the spectral fine structure. In this article, we introduce a new speech analysis method using a spectral model with a composite function of envelope and fine structure models.

Index Terms: parametric speech analyzer, speech synthesis, pitch estimation, spectral envelope estimation.

1. Introduction

PSOLA and waveform connection techniques [1, 2] are known to produce high-quality synthesized speech supposing there is a large enough variety of speech fragments, but their capacity to process the characteristics of speech, such as synthesizing speech with conditions out of the database or adapting to a speaker, is not very high. Even if it was possible to deal with these problems by adding speech data corresponding to the various speaking styles of speech according to what one wants to synthesize, collecting every possible fragment data would certainly be an unrealistic and pain-staking process.

On the other hand, filter-type speech synthesizers, as a representative example of the parametric speech synthesis methods, deal with that problem by (approximately) separating the spectral envelope and the spectral fine structure. One can easily produce a new speech spectrum of different vocal tract length or F_0 by controlling separately the filter characteristics and excitation signals through a small number of parameters. One can thus expect the processing capacity to be very high. Several methods are quite widely known, such as LPC (Linear Predictive Coding), Cepstrum, etc. LPC estimates the vocal tract characteristics modeled by an all-pole filter by assuming the excitation source signal of the vocal cords to be a white process [3]. MFCCs are also a well-known and widely used feature quantizer expressing the vocal tract characteristics of speech [4]. They enable a large variety of processings by working only on a small number of coefficients and parameters, and their ease of use made them become the mainstream analysis method in recent filter-type Text-To-Speech synthesizers. Meanwhile, STRAIGHT is known to enable a high-quality speech synthesis as it starts by estimating the F_0, and then, using an analysis window varying in time according to the F_0 estimate, precisely estimates the spectral envelope in a non-parametric way [5]. Making explicit use of the F_0 estimates via pitch extractor, as opposed to the LPC, is certainly one of the reasons that makes STRAIGHT such a high-quality analysis-synthesis system.

We have thus been aiming at developing a new speech model with always in mind a high-quality Text-To-Speech synthesis and analysis-synthesis systems having both these advantages (i.e., defined in a parametric way and governed by the F_0 parameter).

In the filter-type speech synthesis, the generation process of voiced speech is often assumed to be a linear system with as its input an excitation source signal consisting of a sequence of pulses at intervals of the pitch period. As the input spectrum is a sequence of pulses at intervals of the pitch frequency F_0, the power of each component of the harmonic structure should be extracted separately in order to obtain accurately the vocal tract characteristic and thus one hopes to have a high precision estimation of F_0. On the other hand, as making a half pitch error corresponds to supposing the envelope is unnaturally jagged with zero power for all the odd order harmonics, such an error could be easily corrected if we know in advance the speech spectral envelope or at least by assuming that spectral envelopes in the power domain are usually smooth. Therefore, estimation of F_0 and of the envelope, having a chicken and egg relationship, should be done jointly rather than independently with successive estimations. This is the standing point we chose in this article to formulate a joint estimation model of the spectral envelope and the fine structure.

2. Formulation of the Proposed Method

2.1. Representation of the Spectral Envelopes

In parametric modeling of speech spectrum towards synthesis systems, we must discuss before the formulation what we should employ as a model of spectral envelopes.

In general, the difference between peaks and dips in the speech spectral envelope (spectral dynamic range) is often as big as several dozens of dB. In LPC system for instance, all-pole filters are used to try to express this kind of envelope with a small number of parameters. However, the tendency of the Q-value of the res-

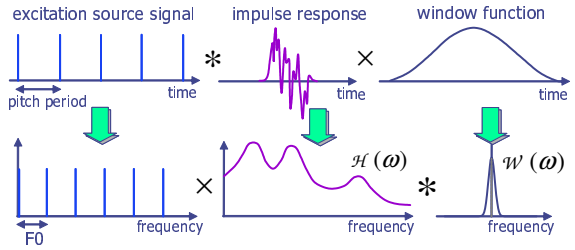

Figure 1: The linear system approximation model in the power domain

onance characteristics of each pole to be larger than that of the actual vocal tract characteristics may be considered to be one of the reasons that LPC system often generates an inarticulate synthesized speech due to the successive superpositions of slow decaying filter outputs, that results in sounding like an echoed speech.

As an alternative to the all-pole filter models having such inherent problems, we model a spectral envelope with a Gaussian mixture function similar to [6] and thoroughly discuss how to estimate its model parameters. Being able to approximate a spectral envelope with a Gaussian mixture, the mean parameter estimate of each Gaussian would ideally correspond to the formant (resonance) frequency and the diffusion parameter estimate to the bandwidth of each formant. We found this is certainly advantageous in terms of being able to incorporate all kinds of knowledge derived from phonetics as well as the classic formant synthesis framework. Furthermore, using as a model of resonance characteristics a mixture of Gaussians, which are quadratic functions in the dB domain, one may be able to keep the Q-value relatively low when creating peaks and dips. Gaussian mixture functions having such merits and being convenient for modeling, we chose to use them for our spectral envelope model, which will be embedded in the formulation described below.

2.2. Power Spectrum Model of Speech

Now if we suppose the excitation source signal is a pulse sequence, its spectrum is again a pulse sequence

$$S(\omega) = \sum_{n=-\infty}^{\infty} \delta(\omega - n\mu), \tag{1}$$

where ω is the frequency, μ the pitch frequency parameter, $\delta(\cdot)$ the Dirac delta function, and n runs over the integers. Multiplying $S(\omega)$ by the vocal tract characteristic $H(\omega)$ and then taking the convolution with the frequency response $W(\omega)$ of the window function gives the complex spectrum of the voiced parts of speech:

$$\begin{aligned} Y(\omega) &= \Big(S(\omega)H(\omega)\Big) * W(\omega) \\ &= \left(\sum_{n=-\infty}^{\infty} H(n\mu)\delta(\omega - n\mu)\right) * W(\omega) \\ &= \sum_{n=-\infty}^{\infty} H(n\mu)W(\omega - n\mu). \end{aligned} \tag{2}$$

We will use as a model of the speech spectrum the approximation of its power spectrum (Fig. 1):

$$\big|Y(\omega)\big|^2 \approx \sum_{n=-\infty}^{\infty} \underbrace{\big|H(n\mu)\big|^2}_{\triangleq \mathcal{H}(n\mu)} \underbrace{\big|W(\omega - n\mu)\big|^2}_{\triangleq \mathcal{W}(\omega - n\mu)} \triangleq \mathcal{Y}(\omega). \tag{3}$$

This approximation is justified under the assumption that the power spectrum of the sum of multiple signal components is approximately equal to the sum of the power spectra generated independently from the components. The smaller the spectral leakage from adjacent components, so that the cross term $W(\omega - n\mu)W^*(\omega - n'\mu)$ with $n \neq n'$ is sufficiently small, the higher the accuracy of this approximation. We then consider the definition interval of ω to be limited by the Nyquist frequency, and n is thus bounded. If we now suppose the analysis window to be a Gaussian window, $\mathcal{W}(\omega)$ can be written as a Gaussian distribution function with diffusion parameter σ:

$$\mathcal{W}(\omega) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\omega^2}{2\sigma^2}\right). \tag{4}$$

From Eq. (3), one can see that with this model each frequency component power is not free but determined at once through $\mathcal{H}(\omega)$, each component power being dependent on the rest of the components. As we want $\mathcal{H}(\omega)$ to be a smooth and non-negative function of ω, we introduce the following Gaussian mixture function as discussed beforehand:

$$\mathcal{H}(\omega) \triangleq \eta \sum_{m=1}^{M} \frac{\theta_m}{\sqrt{2\pi}\nu_m} \exp\left(-\frac{(\omega - \rho_m)^2}{2\nu_m^2}\right), \tag{5}$$

where $\sum_{\forall m} \theta_m = 1$. We could have employed other types of function, but choosing a Gaussian mixture function as the envelope model for the reason mentioned before also had the advantage to enable a prompt application to the speech synthesis method called "Composite Wavelet Model (CWM)" developed by our group [8]. From Eqs. (3)–(5), the speech spectrum can now be written as:

$$\begin{aligned} \mathcal{Y}(\omega) &= \sum_{n=0}^{N} \frac{\mathcal{H}(n\mu)}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\omega - n\mu^2)}{2\sigma^2}\right) \\ &= \frac{\eta}{2\pi\sigma} \sum_{n=0}^{N} \left(\sum_{m=1}^{M} \frac{\theta_m}{\nu_m} e^{-\frac{(n\mu - \rho_m)^2}{2\nu_m^2}}\right) e^{-\frac{(\omega - n\mu)^2}{2\sigma^2}}. \end{aligned} \tag{6}$$

One notices from Eq. (6) that the spectral model we present here is a composite function of two Gaussian mixtures each of which represents the spectral envelope and the spectral fine structure.

So far we have only discussed voiced speech with a harmonic structure, but by making the up to now constant σ in Eq. (6) a free parameter, the model can also be used to approximate reasonably an unvoiced speech spectrum. White noise is indeed generally used as excitation source to synthesize unvoiced speech, but as its power spectrum is a uniform distribution, if in Eq. (6) σ becomes large enough such that the tails of adjacent Gaussians cover each other, the harmonic structure disappears and the model appears as a white spectrum. However, as the approximation given in Eq. (3) in this case becomes less accurate, a more careful modeling for unvoiced speech should be investigated in the future.

The free parameters of the model are $\boldsymbol{\Theta} = (\mu, \sigma, \eta, \rho_1, \cdots, \rho_M, \nu_1, \cdots, \nu_M, \theta_1, \cdots, \theta_{M-1})^T$, and their optimal estimation from a real speech signal is the goal of the following subsection.

2.3. Parameter Optimization

Denoting by $F(\omega)$ the observed speech power spectrum, the problem we are solving is the minimization of the Kullback-Leibler (KL) divergence between $\mathcal{Y}(\omega)$ and $F(\omega)$:

$$J \triangleq \int_D F(\omega) \log \frac{F(\omega)}{\mathcal{Y}(\omega)} d\omega, \tag{7}$$

which henceforth allows us to derive an elegant parameter optimization algorithm. When $\int_D F(\omega)d\omega = \int_D \mathcal{Y}(\omega)d\omega$, this measure is non-negative and gives a notion of distortion between distributions, and η is the normalization parameter that ensures this equality. Since the model $\mathcal{Y}(\omega)$ is characterized by both the parameters for envelope and fine structures, this optimization leads to a joint estimation of F_0 and the envelope. Now if we write $\mathcal{Y}(\omega)$ in the form of the sum over n and m of

$$\mathcal{L}_{n,m}(\omega) \triangleq \frac{\eta\theta_m}{2\pi\sigma\nu_m} e^{-\frac{(\omega-n\mu)^2}{2\sigma^2} - \frac{(n\mu-\rho_m)^2}{2\nu_m^2}}, \tag{8}$$

for any weight functions $\lambda_{n,m}(\omega)$ such that $0 \leqslant \lambda_{n,m}(\omega) \leqslant 1$ and $\forall\omega : \sum_{\forall n,m} \lambda_{n,m}(\omega) = 1$, from Jensen's inequality based on the concavity of the logarithm function we have

$$J_\lambda^+ \triangleq \sum_{n=0}^{N}\sum_{m=1}^{M} \int_D \lambda_{n,m}(\omega)F(\omega)\log\frac{\lambda_{n,m}(\omega)F(\omega)}{\mathcal{L}_{n,m}(\omega)}d\omega \geqslant J \tag{9}$$

and equality $J_\lambda^+ = J$ holds if and only if

$$\forall n, \forall m, \forall\omega : \ \lambda_{n,m}(\omega) = \mathcal{L}_{n,m}(\omega)/\mathcal{Y}(\omega). \tag{10}$$

Eq. (10) is obtained by setting the variation of the functional J_λ^+ with respect to $\lambda_{n,m}(\omega)$ equal to 0. By looking at J_λ^+, one can see that, if $\lambda_{n,m}(\omega)$ is fixed, the minimization of J_λ^+ w.r.t Θ:

$$\widehat{\Theta} = \operatorname*{argmin}_{\Theta} J_\lambda^+ \tag{11}$$

with $\sum_{\forall m}\theta_m = 1$ (the Lagrange multiplier terms are omitted), can be done analytically, which is impossible with J.

When $\lambda_{n,m}(\omega)$ is given by Eq. (10) with arbitrary Θ, the original objective function J is equal to J_λ^+. Then, the parameter Θ, by which both J and J_λ^+ are governed, that minimizes J_λ^+ with $\lambda_{n,m}(\omega)$ fixed according to Eq. (10) necessarily decreases J, since the original objective function is always guaranteed by the inequation (9) to be even smaller than the minimized J_λ^+. Therefore, by repeating the update of $\lambda_{n,m}(\omega)$ by Eq. (10) and the update of Θ by Eq. (11), the objective function, bounded by below, decreases monotonically and converges to a stationary point. Notice that this is our original interpretation of the well-known EM algorithm without using the Bayes rule and thus implies that practically the same algorithm can also be used even though $F(\omega)$ and $\mathcal{Y}(\omega)$ are not probability density functions.

Now the parameter update equations obtained through Eq. (11) for μ, ρ_m, θ_m, σ and ν_m are derived as follows:

$$\begin{pmatrix} \mu^{(i)} \\ \rho_1^{(i)} \\ \vdots \\ \rho_M^{(i)} \end{pmatrix} \leftarrow \left(\begin{array}{c|ccc} a & b_1 & \cdots & b_M \\ \hline b_1 & c_1 & & \multirow{2}{*}{\Large 0} \\ \vdots & & \ddots & \\ b_M & \multicolumn{2}{c}{\Large 0} & c_M \end{array}\right)^{-1} \begin{pmatrix} d \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{12}$$

$$a \triangleq \sum_{n=0}^{N} n^2 \sum_{m=1}^{M} \left(\frac{1}{\sigma^{(i-1)2}} + \frac{1}{\nu_m^{(i-1)2}}\right)\Phi_{n,m},$$

$$b_m \triangleq -\frac{1}{\nu_m^{(i-1)2}} \sum_{n=0}^{N} n\Phi_{n,m},$$

$$c_m \triangleq \frac{1}{\nu_m^{(i-1)2}} \sum_{n=0}^{N} \Phi_{n,m},$$

$$d \triangleq \frac{1}{\sigma^{(i-1)2}} \sum_{n=0}^{N} n \sum_{m=1}^{M} \int_D \lambda_{n,m}(\omega)F(\omega)\omega d\omega,$$

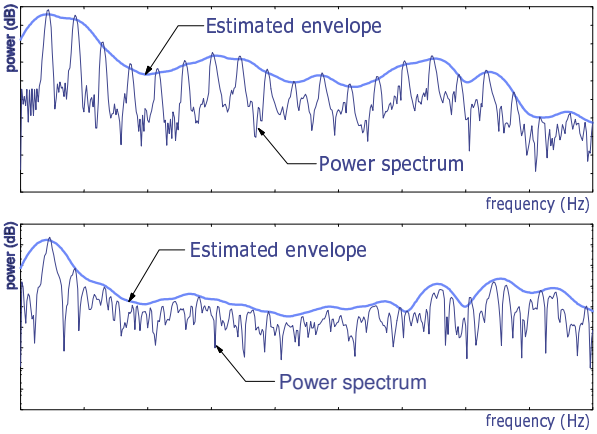

Figure 2: Power spectra of voiced (top) and unvoiced (bottom) speech and the estimated envelopes

$$\theta_m^{(i)} \leftarrow \frac{\sum_{\forall n}\Phi_{n,m}}{\sum_{\forall n}\sum_{\forall m}\Phi_{n,m}} \tag{13}$$

$$\sigma^{(i)} \leftarrow \left(\frac{\sum_{\forall n}\sum_{\forall m}\int_D \lambda_{n,m}(\omega)F(\omega)\left(\omega - n\mu^{(i)}\right)^2 d\omega}{\sum_{\forall n}\sum_{\forall m}\Phi_{n,m}}\right)^{1/2} \tag{14}$$

$$\nu_m^{(i)} \leftarrow \left(\frac{\sum_{\forall n}\left(n\mu^{(i)} - \rho_m^{(i)}\right)^2\Phi_{n,m}}{\sum_{\forall n}\Phi_{n,m}}\right)^{1/2} \tag{15}$$

$$\eta^{(i)} \leftarrow \frac{\sqrt{2\pi}\int_D F(\omega)d\omega}{\sum_{\forall n}\sum_{\forall m}\frac{\theta_m^{(i)}}{\nu_m^{(i)}}\exp\left(-\frac{(n\mu^{(i)} - \rho_m^{(i)})^2}{2\nu_m^{(i)2}}\right)}. \tag{16}$$

where $\Phi_{n,m} = \int_D \lambda_{n,m}(\omega)F(\omega)d\omega$ and the superscript i refers to the iteration cycle. Each update of η adjusts the total power of the model to $\int_D F(\omega)d\omega$. Some examples of the estimated envelope $\mathcal{H}(\omega)$ with $M = 15$ can be seen in Fig. 2.

3. Experimental Evaluations

3.1. Pitch Extraction and Evaluation of the Speech Model

To confirm its performance as a pitch extractor, we tested our method on 10 Japanese speech data of male ('`myi`') and female ('`fym`') speakers from the ATR speech database and chose the well-known pitch extractor "YIN"[7] for comparison. All power spectra were computed with a sampling rate of 16kHz, a frame length of 32ms and a frame shift of 10ms. The spectral model was made using $N+1 = 60$ Gaussians, and the envelope model was made using $M = 15$ Gaussians. The number of free parameters is thus $3 + 15 \times 3 = 48$. The initial values of μ were set to 47Hz, 94Hz and 141Hz, respectively, and among these conditions, the converged parameter set that gave the minimum of J was considered as the global optimum. The initial values of θ_m were determined uniformly, and σ and ν_m were initialized to 31Hz and 313Hz, respectively. For an F_0 estimation task, we defined two error criteria: deviations over 5% and 20% from the hand-labeled F_0

Table 1: *Results of the evaluation*

Speech File	F_0 accuracy (%)		Cosine (%)
	$\pm 5\%$	$\pm 20\%$	
myisda01	98.4 (85.3)	98.6 (98.6)	96.7
myisda02	93.3 (82.6)	97.8 (97.8)	98.0
myisda03	94.2 (79.9)	97.5 (96.9)	96.0
myisda04	98.0 (86.3)	99.0 (95.1)	96.8
myisda05	93.7 (71.7)	97.8 (96.1)	95.9
fymsda01	97.2 (87.0)	98.0 (98.0)	98.3
fymsda02	96.8 (88.5)	98.1 (98.1)	97.6
fymsda03	95.4 (84.6)	98.5 (98.5)	98.2
fymsda04	97.0 (88.2)	98.1 (98.1)	98.2
fymsda05	95.7 (86.5)	99.2 (98.5)	98.1

reference as fine and gross errors, respectively. The former criterion shows how precisely the proposed analyzer is able to estimate F_0 and the latter shows the robustness against the double/half pitch errors. The areas where reference F_0s are given by zero were not considered in the computation of the accuracy. As a second evaluation, we took the average of the cosine measures between $\mathcal{Y}(\omega)$ and $F(\omega)$ on the whole analysis interval to verify how well the choices of the distortion measure to minimize and of the model for expressing actual speech power spectra are. These results can be seen in Table 1. The numbers in the brackets in Table 1 are the results obtained with YIN. Its code was kindly provided to us by its authors. One can verify from the results that our method is as accurate as YIN when it comes to roughly estimate F_0 and significantly outperforms YIN for precise estimation. Thus, our method would be especially useful for situations in which a highly precise F_0 estimate is required. We should note however that the parameters used for YIN may not do it full justice. The results seem to be rather good for a frame-by-frame algorithm, which encourages us to embed this envelope structured model into the parametric spectrogram model proposed in [9, 10] to exploit the temporal connectivity of speech attributes for a further improvement.

3.2. Analysis and Synthesis

We compared through a psychological experiment the processing capacity and the intelligibility of the synthesized speech restored from the parameters obtained via the proposed and LPC analyzers. The parameters extracted via the proposed analyzer were transformed to a synthesized speech using the [1]CWM method [8]. As a test set, we used speech data of 5 vowels (/a/, /i/, /u/, /e/, /o/) and 40 randomly chosen words uttered by a female speaker excerpted from the same database. Analyses were done with a sampling rate of 16kHz, a frame shift of 10ms and a frame length of 32ms for the proposed method and 30 ms for the LPC. The dimension of the parameters for the proposed model and the LPC's were both set to 45. For the LPC analysis, the F_0s were extracted via the supplementary pitch extraction tool included in the Snack Sound Toolkit. Each synthesized speech used for the evaluation was excited with an estimated vocal tract characteristic by a pulse sequence at intervals of a different pitch period from the original one. The pitch periods were modified to 80% and 120% of the pitch periods obtained from the original speech. We let 10 listeners choose the one they thought was more intelligible and obtained a preference score of the results via the proposed analyzer. The preference score, shown in Table 2, shows that the processing capacity and the intelligibility of the synthesized speech generated through the proposed analyzer are higher than that from through LPC analyzer.

[1]CWM synthesizes speech by spacing composite Gabor functions, transformed from a Gaussian mixture envelope, by a pitch period interval.

Table 2: *Preference score(%) of the synthesized speech generated by CWM[8] using the parameter estimates of the proposed model.*

listener	vowel	word
A	60	84
B	60	83
C	40	68
D	80	80
E	60	95
F	80	96
G	100	100
H	40	64
I	80	94
J	60	88
Ave.	66	83

4. Concluding Remark and Future Work

In this article, we formulated the estimation of F_0 and the spectral envelope as a joint optimization of a composite function model of the envelope and the fine structures, and confirmed through experiments the effectiveness of this method. Encouraged by the results, we are planning to embed this envelope structured model into a 2D time-frequency power spectrum model, towards a novel computational acoustic scene analysis as discussed in [9, 10].

5. References

[1] E. Moulines and F. Charpentier, "Pitch-synchronous Waveform Processing Techniques for Text-to-speech Synthesis Using Diphones," Speech Communication, No. 9, pp. 453–467, 1985.

[2] N. Campbell, "A High-definition Speech Resequencing System," In *Proc. 3rd ASA/ASJ Joint meeting*, pp.1223–1228, 1996.

[3] F. Itakura and S. Saito, "Analysis Synthesis Telephony based upon the Maximum Likelihood Method," In *Proc. 6th ICA*, C-5-5, C17–20, 1968.

[4] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," IEEE Trans., ASSP-28(4), pp. 357–366, 1980.

[5] H. Kawahara, "Speech Respresentation and Transformation using Adaptive Interpolation of Weighted Spectrum: Vocoder Revisted," In *Proc. ICASSP '97*, Vol. 2, pp. 1303–1306, 1997.

[6] P. Zolfaghari, S. Watanabe, A. Nakamura and S. Katagiri, "Bayesian Modelling of the Speech Spectrum Using Mixture of Gaussians," In *Proc. ICASSP 2004*, Vol. 1, pp. 553–556, 2004.

[7] A. de Cheveigné and H. Kawahara, "YIN, a Fundamental Frequency Estimator for Speech and Music," J. Acoust. Soc. Am., 111(4), pp. 1917–1930, 2002.

[8] T. Saikachi, K. Matsumoto, S. Sako and S. Sagayama, "Speech Analysis and Synthesis based on Composite Wavelet Model," Technical Report of IEICE, to appear, in Japanese, 2005.

[9] H. Kameoka, T. Nishimoto and S. Sagayama, "A Multipitch Analyzer based on Harmonic-Temporal Structured Clustering," IEEE Trans., Speech and Audio Processing, submitted, 2006.

[10] J. Le Roux, H. Kameoka, N. Ono and S. Sagayama, "Parametric Spectrogram Modeling of Single and Concurrent Speech with Spline Pitch Contour," In *Proc. ICSLP 2006*, submitted, 2006.