APPROXIMATE KALMAN FILTERING FOR THE HARMONIC PLUS NOISE MODEL

Lucas Parra, Uday Jain

Sarnoff Corporation
Adaptive Signal and Image Processing Group
201 Washington Road, Princeton, NJ 08540
lparra@sarnoff.com, ujain@sarnoff.com

ABSTRACT

We present a probabilistic description of the Harmonic plus Noise Model (HNM) for speech signals. This probabilistic formulation permits Maximum Likelihood (ML) parameter estimation and speech synthesis becomes a straightforward sampling from a distribution. It also permits development of a Kalman filter that tracks model parameters such as pitch, harmonic amplitudes, and autoregressive coefficients. We focus here on pitch tracking for which the estimator is highly non-linear. As a result it is necessary to develop an approximate Kalman filter that goes beyond extended Kalman filtering.

1. THE HARMONIC PLUS NOISE MODEL

Since the work of McAulay and Quatieri [1] speech has been repeatedly modeled as the sum of harmonic sinusoids in additive noise. Based on this model speech synthesis and morphing with high perceptual quality has been achieved among other applications (see references in [2, 3]). In the Harmonic plus Noise Model (HNM) the observed data $y(t)$ is the sum of a harmonic component $h(t)$, which captures the voiced portion of the speech spectrum while a colored noise component, $n(t)$ captures the unvoiced portion of speech

$$y(t) = h(t) + n(t)\,. \tag{1}$$

The harmonic and noise processes are defined as

$$h(t) = \sum_{i=1}^{K} b_i s_i(\omega t) \qquad \Leftrightarrow \; h(t) = \mathbf{s}(\omega t)\mathbf{b} \tag{2}$$

$$n(t) = \sum_{\tau=1}^{P} a(\tau) n(t-\tau) + c e(t) \quad \Leftrightarrow \; e(t) = \mathbf{a} * n(t)/c \tag{3}$$

The harmonic basis is given by row vector $\mathbf{s}(\phi) = [\sin(\phi), \ldots, \sin(K\phi), \cos(\phi), \ldots, \cos(K\phi)]$, the fundamental frequency or pitch is ω, and the amplitude of each harmonic is given in the $2K$ coefficients of the column vector $\mathbf{b}$. The colored noise is modeled as an auto-regressive (AR) process with parameters $\mathbf{a} = [1, -a(1), \ldots, -a(P)]$ and i.i.d. zero mean Gaussian innovation $e(t)$ with standard deviation c. Convolution is represented by '*'. All parameters combined will be denoted as $\mathbf{x} = [\mathbf{a}, \mathbf{b}, c, \omega]$. These parameters change over time and in speech analysis the task is to estimate the parameters $\mathbf{x}$ from a set of T data points $\mathbf{y}(t) = [y(t+1), \ldots, y(t+T)]^T$.

With definition (1) the speech signal $y(t)$ can be interpreted as a correlated Gaussian process with a harmonically oscillating mean. This description allows us to specify the joint density function of observations $\mathbf{y}$ given the unknown HNM parameters $\mathbf{x}$ as

$$p(\mathbf{y}|\mathbf{x}) = \prod_t \mathcal{N}(e(t), 1) = \prod_t \mathcal{N}(\mathbf{a} * (y(t) - \mathbf{s}(\omega t)\mathbf{b}), c^2) \tag{4}$$

where t now extends over the length of T samples in frame $\mathbf{y}$. We ignored boundary effects of the convolution with the AR coefficients. $\mathcal{N}(n, c^2)$ represents a zero mean Gaussian distribution with variance c^2. The advantage of considering the HNM from a strictly probabilistic point of view is that one can follow the classic formalism of probabilistic modeling, such as maximum likelihood estimation, sampling, and filtering which we will discuss in the following sections.

2. MAXIMUM LIKELIHOOD PARAMETER ESTIMATION

The maximum likelihood parameter estimates are given by

$$\begin{aligned} \hat{\mathbf{x}} &= \arg\max_{\mathbf{x}} \ln p(\mathbf{y}|\mathbf{x}) \\ &= \arg\min_{\mathbf{a},\mathbf{b},c,\omega} \left\{ T \ln c + \sum_t \frac{(\mathbf{a} * (y(t) - \mathbf{s}(\omega t)\mathbf{b}))^2}{2c^2} \right\} \end{aligned} \tag{5}$$

Note that this minimization is a non-linear optimization problem. However, provided all other parameters are given, the solution for either $\mathbf{a}$ or $\mathbf{b}$ is a linear least squares problem with solutions given by

$$\hat{\mathbf{a}} = \mathrm{lpc}\left(\mathbf{y} - \mathbf{S}(\hat{\omega})\hat{\mathbf{b}}, P\right) \tag{6}$$

$$\hat{\mathbf{b}} = (\hat{\mathbf{a}} * \mathbf{S}(\hat{\omega}))^{\#} (\hat{\mathbf{a}} * \mathbf{y}) \tag{7}$$

$$\hat{c}^2 = \frac{1}{T} \left\| \hat{\mathbf{a}} * \left(\mathbf{y} - \mathbf{S}(\hat{\omega})\hat{\mathbf{b}}\right) \right\|^2 \tag{8}$$

where $\mathrm{lpc}(\mathbf{n}, P)$ represents the Pth order linear prediction or AR coefficients of the noise, $\mathbf{n} = \mathbf{y} - \mathbf{S}(\omega)\mathbf{b}$, '#' represents the pseudo-inverse, $\mathbf{S}(\omega)$ is a matrix containing T row vectors $\mathbf{s}(\omega t)$ for different times t, and the convolutions '*' with $\mathbf{a}$ are over the time coordinate or columns in $\mathbf{S}$ and $\mathbf{y}$. Equations (6) and (7) can be iterated and are guaranteed to converge to the correct ML solution for a given pitch ω. Note that the close form solution for the joint optimization with respect to $\mathbf{a}$ and $\mathbf{b}$ for given pitch ω has been available for some time [4, 5]. We present equations (7) and (6) only because they may be more easely adapted to include prior distributions in the maximum *a posteriory* formalism that is used in Kalman filtering. Estimating the optimum pitch remains a challenging non-convex, and non-linear optimization problem as can be seen in figure 1. Currently we do not see any other remedy than exhaustive search to guarantee the globally optimal solution. Simple heuristics such as a hierarchical search (starting with small K) and limiting the search range based on past values can speed up the search considerably.

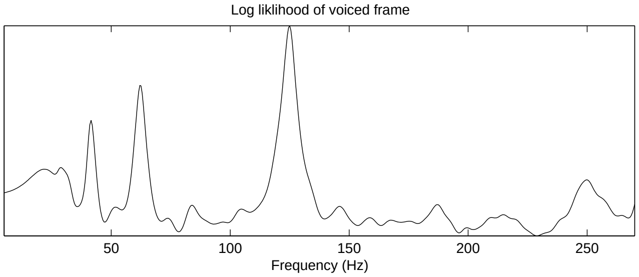

Figure 1: *Logarithm of the likelihood (4) as a function of pitch f for optimal $\hat{\mathbf{a}}, \hat{\mathbf{b}}, \hat{c}$ at that pitch for a voiced frame corresponding to figure 2*

3. SYNTHESIS

Given parameters $\mathbf{x}$, speech synthesis is a straight-forward sampling from the distribution $p(\mathbf{y}|\mathbf{x})$, whereby one generates T samples of i.i.d. zero mean unit variance Gaussian noise $e(t)$ and uses equations (1)-(3) to generate $\mathbf{y}$. The result of the ML parameter estimation and sampling for a voiced frame is shown in Figure 2.

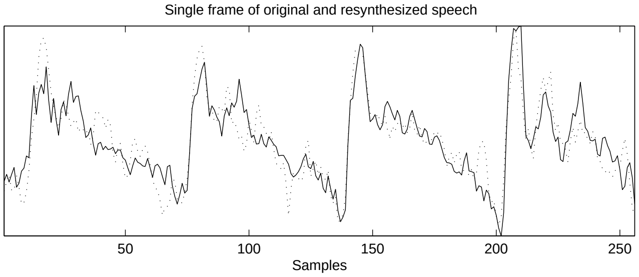

Figure 2: *Voiced speech and re-synthesized version based on ML estimates of HNM parameters.*

In a frame based description, the kth frame of speech $\mathbf{y}_k = \mathbf{y}(kS)$ is associated with parameters $\mathbf{x}_k$. Conventionally, an overlap of half a frame with $S = T/2$ is used. When data is re-synthesized the overlapping frames can be combined with a conventional overlap-add procedure. Since the harmonic basis captures phase information we do find that overlapping frames blend in smoothly. Also note that there is no need for a distinction between voiced and unvoiced speech. The parameters will simply adjust to represent varying magnitudes of harmonic versus noise powers. In fact we can define a Harmonic to Noise Ratio as, $HNR = 20\log_{10}||\mathbf{h}||/||\mathbf{n}||$. An example spectrograms for a original and synthesized speech segment are shown in figure 3. The estimated pitch and associated HNR for the same segment of speech are shown in Figure 4.

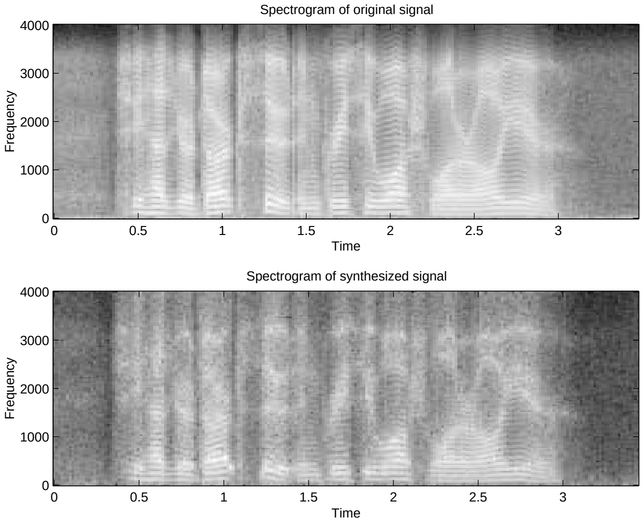

Figure 3: *Spectrogram of speech and re-synthesized version based on ML estimates of HNM parameters.*

4. TRACKING USING KALMAN FILTERING

The purpose of filtering is to obtain a smooth estimate of unknown parameters $\mathbf{x}_k$, often called hidden states, given past and current observations $\mathbf{y}_k, \mathbf{y}_{k-1}, \ldots$, etc. The hidden states are assumed to represent a Markov chain, $p(\mathbf{x}_1, \mathbf{x}_2, \ldots) \propto \prod_k p(\mathbf{x}_k|\mathbf{x}_{k-1})$, emitting at each step k an observation $\mathbf{y}_k$ with probability $p(\mathbf{y}_k|\mathbf{x}_k)$. The most general form of Kalman filtering can be derived from the following maximum *a posteriori* (MAP) estimation [6],

$$\hat{\mathbf{x}}_k = \arg\max_{\mathbf{x}_k} p(\mathbf{x}_k|\mathbf{y}_k, \mathbf{y}_{k-1}, \ldots) \quad (9)$$

$$= \arg\max_{\mathbf{x}_k} \{\ln p(\mathbf{y}_k|\mathbf{x}_k) + \ln p(\mathbf{x}_k|\mathbf{y}_{k-1}, \mathbf{y}_{k-2}, \ldots)\} \quad (10)$$

Here we have used Bayes' rule, dropped the term $p(\mathbf{y}_k|\mathbf{y}_{k-1}, \mathbf{y}_{k-2}, \ldots)$ as it is independent of the parameters, and applied the logarithm. The first term corresponds to the conventional ML problem, while the prior distribution in the second term biases the estimate based on past observations. The key problem in filtering is to efficiently compute this prior distribution, which following the Markov assumption is given by,

$$p(\mathbf{x}_k|\mathbf{y}_{k-1}, \mathbf{y}_{k-2}, \ldots) = \int d\mathbf{x}_{k-1} p(\mathbf{x}_k|\mathbf{x}_{k-1}) p(\mathbf{x}_{k-1}|\mathbf{y}_{k-1}, \mathbf{y}_{k-2}, \ldots) \,. \quad (11)$$

Notice that the second term is the posterior distribution that has already been maximized for the previous step, $k-1$, indicating the recursive operation of the filtering process.

4.1. An approximate Kalman filter

In conventional Kalman filtering all distributions are assumed Gaussian. As a result all past evidence is captured in a single Gaussian prior [7, 6], and the key step in filtering consists in estimating the mean and variance of the prior $p(\mathbf{x}_k|\mathbf{y}_{k-1}, \mathbf{y}_{k-2}, \ldots)$. For simplicity we will assume the transition probabilities $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ to be zero mean Gaussian [1] with covariance Σ,

$$p(\mathbf{x}_k|\mathbf{y}_{k-1},\mathbf{y}_{k-2},\dots) = \int d\mathbf{x}_{k-1}\mathcal{N}(\mathbf{x}_k - \mathbf{x}_{k-1}, \Sigma) p(\mathbf{x}_{k-1}|\mathbf{y}_{k-1},\mathbf{y}_{k-2},\dots) \tag{12}$$

The emission probability, $p(\mathbf{y}_k|\mathbf{x}_k)$, of the HNM is given by (4) and in terms of parameters $\mathbf{x}$ it is not Gaussian. As a result also the posterior distributions are non-Gaussian. Note that for the optimization in (10) the value of the integral (12) is needed only in the vicinity of the optimum value $\hat{\mathbf{x}}_k$. Due to our Gaussian assumption for the transition probability, $\hat{\mathbf{x}}_k$ is likely to be close to the previous optimum value $\hat{\mathbf{x}}_{k-1}$. Since the Gaussian distribution far from $\mathbf{x}_k$ is basically zero the structure of the posterior far from $\hat{\mathbf{x}}_k$ and therefore $\hat{\mathbf{x}}_{k-1}$ is irrelevant. It is therefore fair to approximate the posterior distribution by a Gaussian around $\hat{\mathbf{x}}_{k-1}$, with a covariance $\hat{\Sigma}_{k-1}$ that captures the curvature of the posterior at its maximum $\hat{\mathbf{x}}_{k-1}$.

$$p(\mathbf{x}_k|\mathbf{y}_{k-1},\mathbf{y}_{k-2},\dots) \tag{13}$$

$$\approx \int d\mathbf{x}_{k-1}\mathcal{N}(\mathbf{x}_k - \mathbf{x}_{k-1}, \Sigma)\mathcal{N}(\mathbf{x}_{k-1} - \hat{\mathbf{x}}_{k-1}, \hat{\Sigma}_{k-1}) \tag{14}$$

$$= \mathcal{N}(\mathbf{x}_k - \hat{\mathbf{x}}_{k-1}, \Sigma + \hat{\Sigma}_{k-1}) \tag{15}$$

In the special case that the noise in the model is additive to the hidden state $\mathbf{x}$, this approximation corresponds exactly to the approximation used by the extended Kalman filter [7]. In the present case, however, the noise is partly multiplicative (parameters $\mathbf{a}$ multiply the noise). The more general derivation above, allows an extension of Kalman filtering to cases of non-additive noise.

The covariance $\hat{\Sigma}_k$ captures the curvature of the posterior around the maximum value. It can be computed by fitting a Gaussian to the posterior around its maximum $\mathbf{x}_k$,

$$\hat{\Sigma}_k^{-1} = -\frac{\partial^2}{\partial \mathbf{x}_k \partial \mathbf{x}_k^T} \left(\ln p(\mathbf{x}_k|\mathbf{y}_k, \mathbf{y}_{k-1}, \dots)\right)\big|_{\mathbf{x}_k = \hat{\mathbf{x}}_k} \tag{16}$$

$$= -\frac{\partial^2}{\partial \hat{\mathbf{x}}_k \partial \hat{\mathbf{x}}_k^T} \ln\left(p(\mathbf{y}_k|\hat{\mathbf{x}}_k) p(\hat{\mathbf{x}}_k|\mathbf{y}_{k-1}, \mathbf{y}_{k-2}, \dots)\right) \tag{17}$$

$$\approx \frac{\partial^2 \ln p(\mathbf{y}_k|\hat{\mathbf{x}}_k)}{\partial \hat{\mathbf{x}}_k \partial \hat{\mathbf{x}}_k^T} + \left(\Sigma + \hat{\Sigma}_{k-1}\right)^{-1} \tag{18}$$

In (17) we have used the same manipulations as for Equation (10) and in (18) we have used approximation (15). The intuitive interpretation of (18) is that the iteration step from $k-1$ to k increases the covariance by Σ, while the evidence, $\mathbf{y}_k$, introduced in frame k reduces it.

Note that the derivation of this approximate Kalman iteration is completely independent of the current HNM model.

4.2. Pitch tracking

In principle, the Hessian in (18) can be computed for all HNM parameters. For simplicity, in this work, we focus on pitch only. If the parameters $\mathbf{a}, \mathbf{b}$, and c change sufficiently fast from frame to frame so that the corresponding entries in Σ are very large, it is in fact justified to ignore the effect of past values on the current estimate. The covariance matrix $\hat{\Sigma}_k$ then reduces to a one dimensional variance $\hat{\sigma}_k^2$ and is updated for each frame using Equation (18). The first term in (18) can be computed from (4). Omitting subscript k we can write,

$$\frac{\partial^2 \ln p(\mathbf{y}|\hat{\mathbf{x}})}{\partial \hat{\omega}^2} = \frac{\partial^2}{\partial \hat{\omega}^2} \sum_t \frac{e(t;\hat{\mathbf{x}})^2}{2} \tag{19}$$

$$= \frac{1}{\hat{c}^2} \sum_t \left((\hat{\mathbf{a}} * (\hat{\mathbf{b}}^T \dot{\mathbf{s}}(\hat{\omega} t) t))^2 - e(t;\hat{\mathbf{x}})(\hat{\mathbf{a}} * (\hat{\mathbf{b}}^T \ddot{\mathbf{s}}(\hat{\omega} t) t^2)) \right) \tag{20}$$

With the approximate prior (15) and variance computed through (18) and (20) we can now compute the MAP estimate of the pitch according to (10).

$$\hat{\omega}_k = \arg\max_{\omega_k} \left\{ -T \ln \hat{c}_k - \sum_t \frac{e_k(t)^2}{2} - \frac{(\omega_k - \hat{\omega}_{k-1})^2}{2\hat{\sigma}_k^2} \right\} \tag{21}$$

$$= \arg\min_{\omega_k} \left\{ \ln \hat{c}_k(\omega_k) + \frac{(\omega_k - \hat{\omega}_{k-1})^2}{2\hat{\sigma}_k^2 T} \right\} \tag{22}$$

In computing $e_k(t)$ the optimal $\hat{c}_k$ from (8) has been inserted. Note also that the ML optimal $\hat{\mathbf{a}}_k, \hat{\mathbf{b}}_k$ for given pitch ω_k have to be used when computing $\hat{c}(\omega_k)$ which in fact we emphasized by writing the dependency on ω_k explicitly.

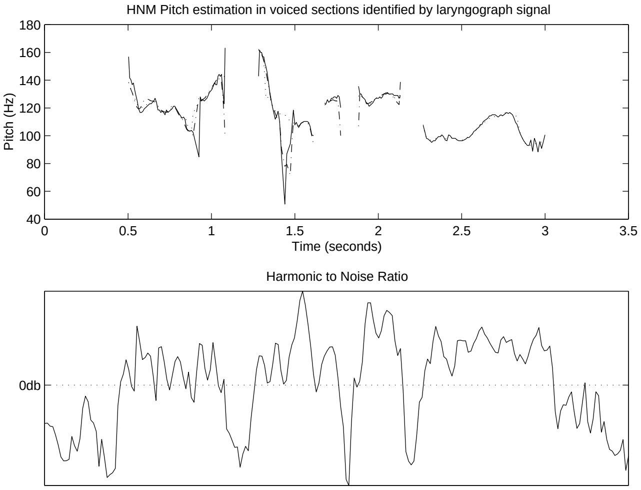

Figure 4: *Pitch estimate using ML estimator and approximate Kalman pitch tracking shown together with 'truth' data obtained from laryngographic recordings. The corresponding HNR shows negative values during noise sections for which pitch information is not meaningfully defined and for which in fact no pitch truth data is available.*

[1] In making a Gaussian assumption on transition probabilities we are roughly stating that the parameters change from frame to frame symmetrically around their current value and that the distribution of those changes is unimodal and can be captured well by the square magnitude of the changes. For example we assume with this that the pitch is equally likely to go up or down and the amount of change is well captured by the standard deviation. In reality the pitch may in average have a net drift and it is reasonable to assume different behaviors within voiced and unvoiced sections. Due to space limitations a more thorough discussion of the assumptions and more complex alternatives have to be omitted.

4.3. Experiments

We ran the approximate Kalman filter pitch tracking and ML estimator on 60 seconds of speech data for which truth data on the pitch was available in the form of laryngograph recordings [8]. In figure 4 the results of the ML estimator and the approximate Kalman filtering are shown. We search for the optimal pitch within one standard deviation, $\hat{\sigma}_k$, of the prior expected pitch $\hat{\omega}_{k-1}$. The adaptive $\hat{\sigma}_k$ guarantees that we search in an appropriate region. If the Harmonic to Noise Ratio is sufficiently small (HNR $<$ -6 dB) $\hat{\omega}_{k-1}$ is set to some default value as it is meaningless to track pitch during a noise only section. Both estimators accurately track the laryngograph data. More detail is shown in Figure 5 where one can see that the Kalman filter smoothes the estimated pitch. The standard deviation, σ, for the transition probabilities in (18) was assumed to be 5 Hz for voiced frames with HNR $>$ 0 dB and 25 Hz otherwise. One should note that the laryngograph data may not always lead to correct pitch estimates. For instance, in the example shown on figure 5 it is very unlikely that the true pitch goes down to 50 Hz. Nevertheless, we used the pitch resulting from laryngograph recordings as truth data to quantify the estimation error. The distribution of the error for both estimators is shown in Figure 6. We note that the approximate Kalman filter makes less gross errors while maintaining the magnitude of small deviations essentially the same than the ML estimator (standard deviation of 2.8 Hz for Kalman vs. 2.6 Hz for ML). The estimation bias (0.5 Hz) lies within the step-size we used for the optimum pitch search (1 Hz).

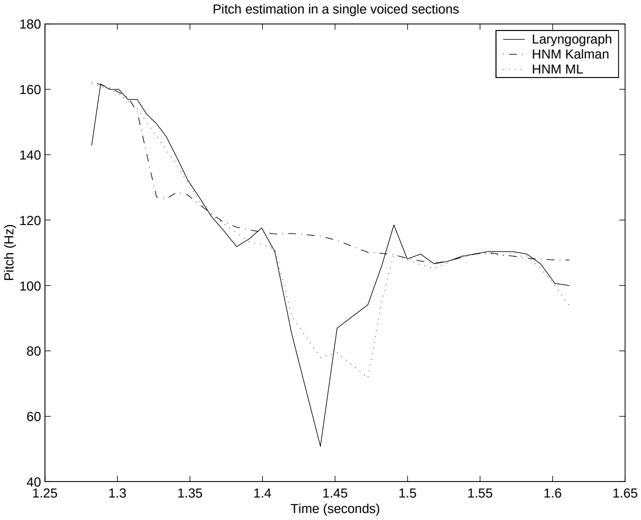

Figure 5: *Detail from figure 4 on pitch estimate using ML estimator and approximate Kalman pitch tracking.*

5. CONCLUSION

We have demonstrated that a rigorous probabilistic treatment of the harmonic plus noise model allows straightforward parameter estimation without the need to classify frames of data into voiced or unvoiced nor to divide the spectrum into some arbitrary way. It also allows the application of the Kalman filter formalism. The non-linearity of the model made it necessary to develop a more general Kalman formalism that goes beyond conventional extended Kalman filtering. In this work the tracking parameters are assumed known but a more complete analysis may allow the estimation of tracking parameters using an EM algorithm [6]. In principle this can set the foundation for formant segmentation and recognition based on perceptually meaningful HNM parameters.

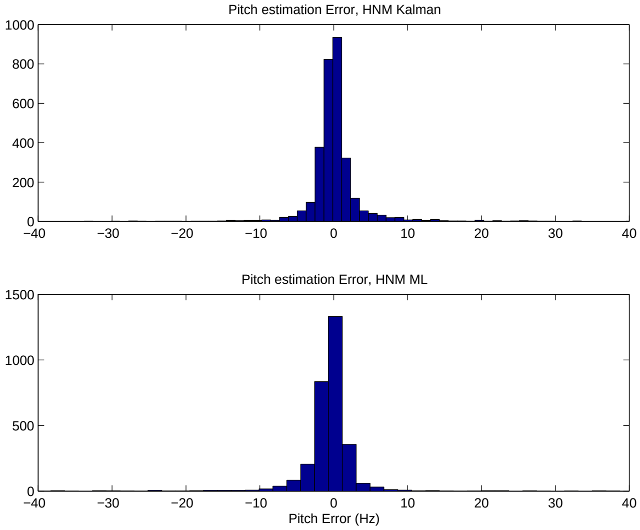

Figure 6: *Histogram of pitch estimation error for the ML and approximate Kalman estimator.*

6. REFERENCES

[1] Robert J. McAulay and Thomas F. Quatieri, “Pitch Estimation and Voicing Detection based on a Sinusoidal Speech Model”, in *ICASSP*, 1990, pp. 249–252.

[2] Yannis Stylianou, “Applying the Harmonic Plus Noise Model in Concatenative Speech Synthesis”, *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, Jan 2001.

[3] Darragh O’Brien and A. I. C. Monaghan, “Concatenative Synthesis based on a Harmonic Model”, *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, Jan 2001.

[4] Steven M. Kay and Venkatesh Nagesha, “Extraction of Periodic Signals in Colored Noise”, in *ICASSP*, 1992, vol. 5, pp. 281–284.

[5] Steven M. Kay and Venkatesh Nagesha, “Maximum Likelihood Estimation of Signals in Autoregressive Noise”, *IEEE Transactions on Signal Processing*, vol. 42, no. 1, Jan 1994.

[6] Sam Roweis and Zoubin Ghahramani, “A Unifying Review of Linear Gaussian Models”, *Neural Computation*, vol. 11, no. 2, Feb 1999.

[7] Steven M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, chapter 13, Prentice Hall, 1993.

[8] Center for Speech Technology Research University of Edinburgh, “http://www.festvox.org/dbs/dbs_kdt.html”.