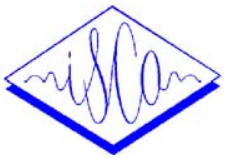
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Source Separation using Particle Filters

Mital A. Gandhi, Mark A. Hasegawa-Johnson

Department of Electrical and Computer Engineering
University of Illinois, Urbana-Champaign
{magandhi,jhasegaw}@uiuc.edu

10.21437/Interspeech.2004-710

Abstract

Our goal is to study the statistical methods for source separation based on temporal and frequency specific features by using *particle filtering*. Particle filtering is an advanced state-space Bayesian estimation technique that supports non-Gaussian and nonlinear models along with time-varying noise, allowing for a more accurate model of the underlying system dynamics. We present a system that combines standard speech processing techniques in a novel method to separate two noisy speech sources. The system models the pitch and amplitude over time separately, and adopts particle filtering to reduce complexity by generating a discrete distribution that approximates well the desired continuous distribution. Preliminary results that demonstrate the separation of two noisy sources using this system are presented.

1. Introduction

One of the primary questions in source separation has been how to handle the processing in noisy environments. Stochastic analyses [1, 2] often involve a Markov source chain and an associated observation sequence together with a noise process. The goal is to determine the optimal filter based on the observations, the Markov transition probabilities, and the distribution models (for the noise especially).

The work of Meddis and Hewitt [3] for the first time demonstrated a quantitative model of methods by which the human ability to identify two simultaneous vowels may improve if the fundamental frequencies of the vowels are different. The first part of their system simulated the human auditory periphery via a bank of bandpass filters and inner hair-cell simulators. Pitch and timbre information was extracted over all the channels and a pooled ACF was used to estimate a pitch. The individual ACFs were then grouped accordingly and used to determine one of the two vowels. The second vowel was assumed to be represented by the remaining channels. Essentially, the authors used the pitch and timbre information from the mixed signal with a template matching procedure to generate information about the two source vowels. We have incorporated this idea of tracking the periodicity information from the observation signal over both source chain reconstructions. However, we extend the idea to full speech source reconstructions instead of just vowel identifications.

In addition to the pitch information, we have also incorporated the magnitude spectra into the system, similar to the work of Nix *et. al.* [1]. The authors in this work proposed an algorithm that performed a source separation based on magnitude spectra and direction, tracked by a particle filtering procedure on a frame-by-frame basis. However, according to the authors, the two-voice experiments led to spectral (magnitude) estimates that differed significantly from the true spectra of the sources. In our work, we do not incorporate any directional information.

Finally, another important work in the literature on source separation is the re-filtering approach proposed by Roweis [2]. The approach is based on the idea of selectively re-weighting across sub-bands via a set of varying masking signals. Different sources are estimated from mixtures by changing the masking signals. Roweis infers the masking signals using a factorial HMM structure.

We note two important concepts: first, for two clean speech signals that are mixed additively in time, the log spectrogram of the mixture is represented very well by the *maximum* of the log spectrograms of the individual sources [4]. Operating over small time-frequency regions, this approximation holds well only if both sources in question are not large and equal. In general, speech is such that it is very unlikely that two sources will contain a substantial amount of energy over a narrow frequency sub-band and hence the approximation holds.

Second, we note that some of the previous works in source separation incorporated factorial-HMMs in the algorithms [2, 5, 6]. In our proposed work, we generalize the factorial HMM to a much higher dimensional search space using the sequential Monte Carlo (SMC) scheme, which is a generalization of the traditional Kalman filtering methods. We estimate a Markov chain X (source signal) from its noisy dependent variable Y, where the transition probability kernel for the chain generally depends on a specified set of parameters. The *jointly optimal particle filter* is then that which maximizes the conditional distribution $\mathbb{P}\{X_n, \cdots, X_1 | Y_n, \cdots, Y_1\}$. Unlike the Hidden Markov Model, the particle filter does not evaluate every possible X_n though, and thus, handles the case where X_n

is drawn from a very large search space.

2. Particle Filtering

2.1. The Bayesian Approach

In the problem of estimating a set of hidden variables (states) based on observations of the system, prior knowledge of the unknown quantities can be exploited in a Bayesian approach. If available, prior distributions and the likelihood functions relating these distributions to the observations can be integrated with Bayes' theorem to create a posterior distribution for estimating the states.

The model can be fundamentally expressed by a state and an observation equation:

$$\mathbf{x}_t = f_t(\mathbf{x}_{t-1}, \mathbf{v}_{t-1}) \tag{1}$$

$$\mathbf{y}_t = h_t(\mathbf{x}_t, \mathbf{w}_t) \tag{2}$$

In the state equation, $\mathbf{x}_t$ is the current state, f_t is a transition function (possibly non-linear), and $\mathbf{v}_t$ is the associated noise process. The second equation represents similar quantities to generate the observation $\mathbf{y}_t$.

Given a set of observations $\mathbf{y}_t$ and the set of unknown sources $\mathbf{x}$, the following is the posterior distribution [4]:

$$p(\mathbf{x}_t|\mathbf{Y}_t) = \frac{p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{Y}_{t-1})}{p(\mathbf{y}_t|\mathbf{Y}_{t-1})} \tag{3}$$

where

$$p(\mathbf{y}_t|\mathbf{Y}_{t-1}) = \int p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{Y}_{t-1})d\mathbf{x}_t \tag{4}$$

We note that $p(\mathbf{y}_t|\mathbf{x}_t)$ and $p(\mathbf{x}_t)$ are the likelihood and the prior distribution, respectively. Also, $\mathbf{Y}_t$ above represents $\mathbf{y}_{1:t}$. With this posterior distribution, optimal MMSE and MAP estimators can be defined. For a linear Gaussian system, the state-space approach can be solved analytically via a Ricatti equation (or iteratively by Kalman-Bucy filters). The Extended Kalman Filter allows for non-linearity, by first linearizing the system using Taylor series expansions. Unfortunately, this filter has a possibility of divergence when the non-linear functions are poorly approximated. Furthermore, real world data is generally high dimensional, non-linear, non-stationary, and non-Gaussian, leaving the above intractable.

2.2. Monte-Carlo Particle Filters

The hidden state $\mathbf{x}_t$ is modeled as a Markov process with initial distribution $p(\mathbf{x}_0)$ and transition probabilities $p(\mathbf{x}_t|\mathbf{x}_{t-1})$. The goal is to estimate the posterior distribution $p(\mathbf{X}_t|\mathbf{Y}_t)$ and associated features useful in computing the following step's distribution. The following set of equations summarize the stochastic setup, beginning with the formulae for posterior distribution given by Bayes' Theorem and its recursive version:

$$p(\mathbf{X}_t|\mathbf{Y}_t) = \frac{p(\mathbf{Y}_t|\mathbf{X}_t)p(\mathbf{X}_t)}{\int p(\mathbf{Y}_t|\mathbf{X}_t)p(\mathbf{X}_t)d\mathbf{X}_t} \tag{5}$$

$$p(\mathbf{X}_{t+1}|\mathbf{Y}_{t+1}) = p(\mathbf{X}_t|\mathbf{Y}_t)\frac{p(\mathbf{y}_{t+1}|\mathbf{x}_{t+1})p(\mathbf{x}_{t+1}|\mathbf{x}_t)}{p(\mathbf{y}_{t+1}|\mathbf{Y}_t)} \tag{6}$$

The marginal distributions are obtained recursively from the following "prediction and update" equations:

$$p(\mathbf{x}_t|\mathbf{Y}_{t-1}) = \int p(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{Y}_{t-1})d\mathbf{x}_{t-1} \tag{7}$$

$$p(\mathbf{x}_t|\mathbf{Y}_t) = \frac{p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{Y}_{t-1})}{\int p(\mathbf{y}_t|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{Y}_{t-1})d\mathbf{x}_t} \tag{8}$$

2.3. Steps of Particle Filtering

- Sampling - Draw N_s particles, $1 \leq i \leq N_s$, from the prior [7]:

$$\mathbf{x}_t^{*i} \sim \pi(\mathbf{x}_t|\mathbf{X}_{t-1}^i, \mathbf{Y}_t) \sim \pi(\mathbf{x}_t|\mathbf{x}_{t-1}^i, \mathbf{y}_t) \tag{9}$$

- Re-Sampling - Assign the particle a weight w_t^i:

$$w^i = \frac{p(\mathbf{y}_t|\mathbf{x}_t^{*i})}{\sum_{j=1}^{N_s} p(\mathbf{y}_t|\mathbf{x}_t^{*j})} \tag{10}$$

  Re-sample N_s times from the discrete distribution to generate samples $\mathbf{x}_t^i$ such that $p(\mathbf{x}_t^j = \mathbf{x}_t^{*i}) = w_i$. Set the following and repeat the above steps:

$$\mathbf{X}_t^i = (\mathbf{X}_{t-1}^i, \mathbf{x}_t^i) \tag{11}$$

3. System Description

The model of the system is based primarily on the particle filtering algorithm along with various speech processing primitives built into the procedure (figure 1):

Pitch Extraction: An open-loop pitch value is extracted from the input over each frame, to be used as one of the pitch values from which the adaptive pitch codebook is generated.

MFCC Codebook: The VQ codebook design is based on the LBG algorithm. The codebook consists only of MFCC vectors, along with a transition probability matrix $p(x_t = \mu_k|x_{t-1} = \mu_i)$ that specifies the probability of a transition from centroid μ_i to μ_k.

Particle Design: Each particle consists of two quantized MFCC vectors. Pitch, voicing, and turbulence amplitude information could be included as other source states, but we have pursued alternate methods of including this information.

Initialization/Sampling: For the first frame, the MFCC vectors in the particles can be initialized randomly or in a predefined manner. In a simplified experiment, random versus structured initialization resulted in negligible differences. As the size of the codebook grows,

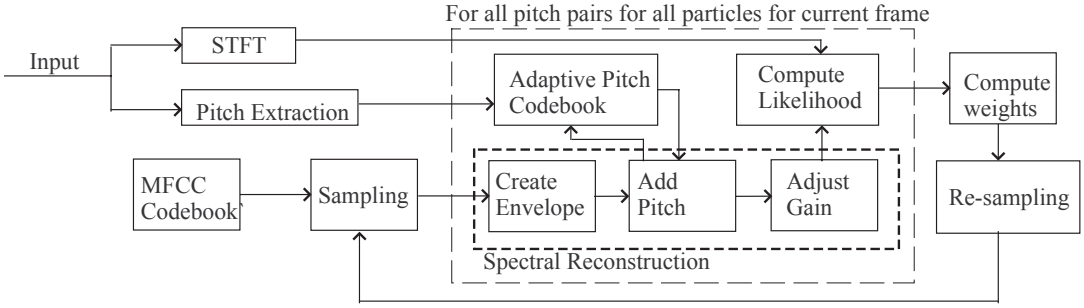


Figure 1: System Block Diagram

the need for some structured initialization might become necessary (or if the training quality for each of the quantized vector decreases). For subsequent frames, sampling is performed based on the transition probability matrix from the codebook.

Create Envelope: The spectral envelopes, $\tilde{X}(w)$ and $\tilde{Y}(w)$, corresponding to reconstruction chains is obtained by inverting the quantized MFCCs.

Adaptive Pitch Codebook: All pitch frequencies within 50Hz of the open-loop pitch frequency, or within 50Hz of the previous frame's pitch in the same chain, are considered to be pitch candidates.

Add Pitch: Harmonic window spectra of these pitches are generated. The log DFT spectrum is created by adding this harmonic window spectrum to the spectral envelope.

Amplitude Gain: The amplitude gain is modeled as a three-way MMSE estimator, modeling the amplitudes of two periodic window spectra and a high pass filtered noise spectrum. The gain values α, β, and η are multiplied with the two periodic window spectra and the noise spectrum, respectively, and the components are finally summed to give an estimate $\hat{Z}(w)$ of the observed log magnitude spectrum $Z(w)$.

Given observation $Z(w)$, voiced components $V_1(w), V_2(w)$, and noise component $N(w)$ in power spectral domain, compute α, β, η such that

$$\hat{Z} = \alpha \cdot V_1 + \beta \cdot V_2 + \eta \cdot N \tag{12}$$

Minimization of $\sum_w (Z - \hat{Z})^2$ gives us the following result:

$$\begin{pmatrix} \alpha \\ \beta \\ \eta \end{pmatrix} = \mathbf{R}^{-1}_{V_1 V_2 N} \begin{pmatrix} 2R_{ZV_1} \\ 2R_{ZV_2} \\ 2R_{ZN} \end{pmatrix} \tag{13}$$

where

$$\mathbf{R}_{V_1 V_2 N} = \begin{pmatrix} R_{V_1V_1} & R_{V_1V_2} & R_{V_1N} \\ R_{V_2V_1} & R_{V_2V_2} & R_{V_2N} \\ R_{NV_1} & R_{NV_2} & R_{NN} \end{pmatrix} \tag{14}$$

and

$$R_{V_1V_2} = \sum_w V_1(w)V_2(w) \tag{15}$$

The fine structure components of the spectra are given as follows:

$$\hat{X}(w) = \alpha \cdot V_1 + \eta \cdot N \tag{16}$$
$$\hat{Y}(w) = \beta \cdot V_2 + \eta \cdot N \tag{17}$$

The model log magnitude spectra as given as follows:

$$X(w) = log(\tilde{X}(w)) + log(\hat{X}(w)) \tag{18}$$
$$Y(w) = log(\tilde{Y}(w)) + log(\hat{Y}(w)) \tag{19}$$

Compute likelihood: The reconstructed spectra are merged into a single log magnitude output spectrum $\hat{Z}(w)$ via a *max* operation over the frequencies in each frame:

$$\hat{Z}(w) = max(X(w), Y(w)) \tag{20}$$

The distance between the observed and reconstructed spectra is then simply $D = (Z - \hat{Z})^2$.

We also apply a penalization to the reconstruction candidates whose pitch values have changed by more than 10% relative to the previous frame. We suppose that the pitch transition penalty kernel can be described by

$$P_t = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{q^2}{2\sigma^2}} \tag{21}$$

where

$$q = F_0(t) - F_0(t-1) \tag{22}$$

Hence, we obtain the following:

$$log(P^i_{t,k}) = -\frac{q_k^2}{2\sigma^2} - log(\sqrt{2\pi}) - log(\sigma) \tag{23}$$
$$log(P^{i,j}_{net}) = D + log(P^i_{t,1}) + log(P^j_{t,2}) \tag{24}$$

where $P^i_{t,k}$ is the cost associated with the pitch change q_k in the k^{th} source. Furthermore, $i, j \in \{1, \cdots, R\}$ for some arbitrary values of R, which represents the number of values in the pitch range.

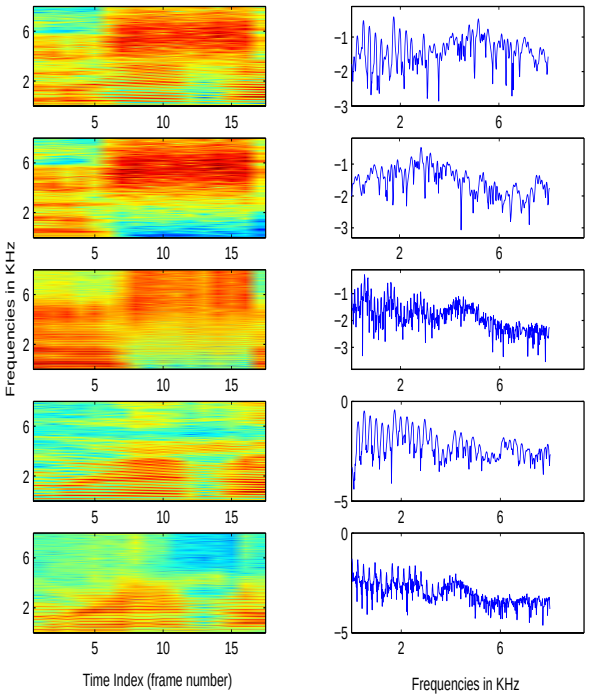


Figure 2: *Left column: Top plot shows the mixture observation. Plots 2 & 4 are original sources; 3 & 5 are reconstructions. Right column: spectra of one particular frame from each corresponding spectrogram.*

Compute normalized weights: We then compute normalized weights for each of the particles using equation 10, with the likelihood as given in equation 24.

Re-sampling: New particles are defined as those containing a new set of MFCC vectors, one for each chain tracked by the particle (each chain tracked by the particle corresponds to explaining one of the two sources involved in the mixture). We also store back pointers to the particles from which the transition occurred for later use.

Final Reconstruction: At the end of the above computations for all frames of the observation, a Viterbi type back-traversal procedure is used to reconstruct the sequence of frames that provides the optimal probability (or equivalently, the two source sequences that are most likely to explain the observation sequence).

4. Results

The system was tested with a mixture of two sources, one male and one female. Pitch frequencies for the adaptive codebooks were centered around the two pitch frequencies with the largest average autocorrelation peak, averaged over the whole utterance. Pitch penalization was applied as described previously. The MFCC codebook was generated from the original source vectors. This simplification essentially implies a very good training in the general case. The algorithm was tested with five different segments of about 20 frames each. Source reconstructions from one such segment are shown in figure 2. These first results demonstrate the system identifying the sources quite precisely from the mixture shown. Other segments for both sources gave similar successful results. The most noticeable difference between the original and synthesized spectra is the shape of each harmonic peak. Each peak in the synthesized spectrum is the modulated transform of a Hamming window, but peaks in the original spectrum are much wider. The pitch estimates are decent, but frames were observed often that contained a non-negligible pitch deviation. More study is required to track the pitch more successfully in such cases.

5. Conclusions

We have demonstrated that it is possible to separate two sources via particle filtering, and by modeling the pitches and amplitude gains separately. There is room for further improvement perhaps by better modeling the shape of the harmonic spectral peak.

6. References

[1] J. Nix, M. Kleinschmidt, & V. Hohmann, "Computational Auditory Scene Analysis by using statistics of high-dimensional speech dynamics and sound source direction", *Proc. of Eurospeech*, pp. 1441-1444, 2003.

[2] Sam T. Roweis, *Neural Information Processing Systems*, 13, pp. 793-799, 2000.

[3] R. Meddis & M. Hewitt, "Modeling the identification of concurrent vowels with different fundamental frequencies", *JASA*, Vol. 91, pp. 233-245, 1992.

[4] A. Nadas, D. Nahamoo, & M.A. Picheny, Speech Recognition Using Noise-Adaptive Prototypes," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 37, No. 10, October 1999.

[5] A.P. Varga & R.K. Moore, "Hidden Markov Model decomposition of speech and noise", *ICASSP*, 1990.

[6] A. Deoras & M.A. Hasegawa-Johnson, "Automatic Recognition of Simultaneous Spoken Digits," *ICASSP*, 2004.

[7] N.J. Gordon, D.J. Salmond, & A.F.M. Smich, "Novel Approach to Nonlinear/NonGaussian Bayesian State Estimation", *IEE Proc.-F*, Vol. 140, No. 2, pp. 107-113, 1993.

[8] E.-K. Kim, W.-J. Han, & Y.-H. Oh, "A score function of splitting band for two-band speech model", Speech Communication, vol. 41, no. 4, pp. 663-674, 2003.