SMC SAMPLERS FOR MULTIRESOLUTION AUDIO SEQUENCE ALIGNMENT

Dogac Basaran [*], *A. Taylan Cemgil* [†], *Emin Anarım* [*]

Boğaziçi University
Electrical and Electronics Engineering Department [*]
Computer Engineering Department [†]
İstanbul, Turkey
{dogac.basaran,taylan.cemgil,anarim}@boun.edu.tr

ABSTRACT

In our previous work, we formulated multiple audio sequence alignment in a probabilistic framework [1]. Here, we extend the model for multi resolution alignment and focus on pairwise cases. We defined a similarity based approach for binary feature sequences and integrate it into a new generative model. We modify the model for multi resolution case and the matching is achieved with a Sequential Monte Carlo Sampler (SMCS) which uses low resolution models as bridge distributions. The simulation results on real data sets suggest that our method is very robust and efficient under very noisy conditions with proper choices of model parameters.

Index Terms— Audio alignment, Audio matching, Probabilistic Model,Sequential Monte Carlo Sampler

1. INTRODUCTION

Audio alignment or fingerprinting is defined in the literature as matching an unknown audio signal to a large dataset. Some popular use cases are identifying the metadata of an unknown audio signal such as song title or artist name and monitoring radio broadcasts for copyright purposes. There are several audio fingerprinting methodologies with high matching performance [2]-[7]. In [1], we viewed the common audio alignment from a different angle where there are several unsynchronised recordings i.e., each microphone starts and stops recording at different times independent of each other, and the aim is to align these sequences on a generic time line according to each other. The difficulty of the problem rises from the facts that the sequences may or may not overlap, none of the sequences have to cover all the timeline and there is no clean original source database.

Alignment, from this point of view, is applicable to several other problems such as synchronisation of video clips with no offsets [8] or restoration of an audio scene from its noisy recordings. A possible application is restoring a recording of concert from the recordings of the audience [9]. Similar approaches exist in different fields such as genetics where DNA strands are assembled from shorter sequences [10] and image stiching where a panoramic view is assembled from multiple partially overlapping images [11].

There are two important performance criteria for the alignment problem; it should be fast and robust. For both purposes, the alignment is usually applied in feature space rather than on raw audio data. The majority of the framework rely on spectral representation of the signal such as local peaks on the magnitude of short-time Fourier Transform (STFT) [2],[8], thresholded energy of first difference through time and frequency in the STFT [3], mel-frequency ceptstral coefficients (MFCC) [4],positive spectral difference [5],[12] and constant Q transform (CQT) [6].

Most state of the art methods employ hashing algorithms that reduces the amount of data, and then apply search strategies that works on all possible pairs [2],[3],[6],[8]. In [1], we proposed a model based approach where we are able to match an unknown sequence against a group of sequences with known relative shifts. In this work, we extend the model for multi resolution alignment and focus on pairwise cases.

The pairwise alignment problem can also be tackled with deterministic approaches such as cross-correlation or any similarity based approach but it is not always clear how to apply these methods when the sequences do not overlap or there is some missing data. In this work, we used a similarity measure based on Hamming distance for binary sequences and defined a generative model following [1] for which the posterior is similar to this measure. For the search strategy, we propose a SMC sampler based method to compute the optimum alignment without explicitly evaluating score function for all alignments. The main idea is to use low resolution bridge distributions that guides samples through the modes of target posterior distribution. The model is slightly modified for the multi resolution case. Our main motivation is to extend the SMCS based multi resolution model to multiple alignment cases and this work is an initial phase that

DB is supported by the Turkish State Planning Organization (DPT) under the TAM Project, number 2007K120610

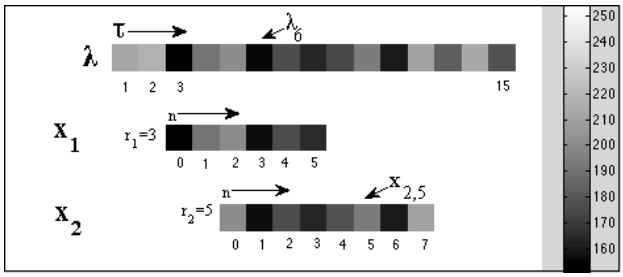
ATC is supported by TÜBİTAK project number 110E292, Bayesian matrix and tensor factorisations (BAYTEN) and Bogazici University BAP 6882.

considers only pairwise scenarios.

2. PROPOSED MODEL

In this section, we summarize the model given in [1] and show how to modify it such that it is applicable to low resolution signals. In Figure 1, a toy example is given to illustrate the model. The features are positive coefficients and color of each coefficient depends on its value. The main idea of the model is; *Properly aligned feature sequences are noisy realizations or functions of a common but unobserved feature sequence* [1]. The unobserved feature sequence is denoted by λ_τ where $\tau = 1 \dots T$ is a global time frame index. In this example, two sequences are observed which are denoted by x_k, where k is the sequence index. The length of each observation is denoted by N_k and n is a local time index. The alignment variable for each sequence is denoted by r_k. Here, the lengths of the sequences are $N_1 = 6$, $N_2 = 8$ and their starting points are $r_1 = 3$, $r_2 = 5$. In this scenario, the sequences overlap with each other at several points, i.e., $x_{1,2}$ and $x_{2,0}$ coincide at global time $\tau = 5$.

Fig. 1. Toy example

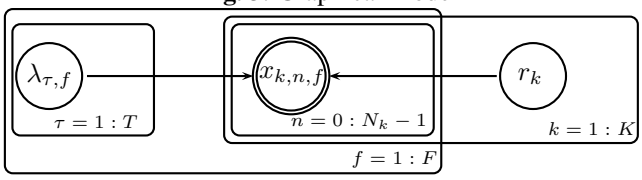


It can be observed from the Figure 1 that $x_{1,2}$ and $x_{2,0}$ values are close to each other since they are observations of a common source λ_5. Intuitively, the overlapping parts of such sequences should be similar to each other at the exact alignment point. Therefore by applying such a similarity measure, one can find the best alignment between two sequences. In binary case, a bitwise comparison in the overlapping parts of the signals can be used as a similarity measure. In Figure 2, an example of such a situation is shown. If two coefficients of sources x_1 and x_2 i.e.,$x_{1,1,1}$ and $x_{2,0,1}$ that are aligned to the time $\tau = 1$, are equal to each other then they are counted as 1, otherwise they are not counted. The ratio of this count to the total number of overlapping bits acts as a similarity measure since at the exact alignment, this ratio should be highest. In this scenario, there are 4 overlapping bits and 3 of them are equal to each other therefore the ratio is computed as 3/4. This similarity measure acts as a strong scoring function even in low SNR cases. As mentioned before, following the template generative model in [1], we propose the following generative

Fig. 2. Similarity of two sequences

τ =	1	2	3	4	5
$r_1 = 2$		$x_{1,0,1} = 1$	$x_{1,1,1} = 1$	$x_{1,2,1} = 0$	
		$x_{1,0,2} = 0$	$x_{1,1,2} = 1$	$x_{1,2,2} = 1$	
$r_2 = 3$			$x_{2,0,1} = 1$	$x_{2,1,1} = 0$	$x_{2,2,1} = 1$
			$x_{2,0,2} = 1$	$x_{2,1,2} = 0$	$x_{2,2,2} = 1$

Fig. 3. Graphical Model



model for binary sequences;

$$\lambda_{\tau,f} \sim \mathcal{BE}(\lambda_{\tau,f}; \alpha_\lambda)$$

$$r_k \sim \prod_{\tau=1}^{T} \pi_{k,\tau}^{[r_k=\tau]}$$

$$x_{k,n,f}|r_k, \lambda_{\tau,f} \sim \prod_{\tau=1}^{T} \mathcal{P}(x_{k,n,f}|r_k, \lambda_{1:T,f})^{[n=\tau-r_k]}$$

where $\mathcal{P}(x_{k,n,f}|r_k, \lambda_{1:T,f})$ is a conditional Bernoulli distribution which is defined as,

$$\mathcal{P}(x_{k,n,f}|r_k, \lambda_{1:T,f}) = (w_{i,j})^{\sum_{i=0}^{1}\sum_{j=0}^{1}[x_{k,n,f}=i][\lambda_{\tau,f}=j]}$$

Here the $w_{i,j}$ is the probability that the $\lambda_{\tau,f} = j$ and $x_{k,n,f} = i$. f is frequency sub band index. $[\cdot]$ is the indicator function which is equal to one if the expression inside is true. In this work, we assumed $w_{i,j} = w$ if $i \neq j$ and $w_{i,j} = 1 - w$ if $i = j$, and the parameter of prior, $\alpha_\lambda = 0.5$. The hidden coefficients λ_τ are assumed to be a-priori independent and each r_k is uniformly distributed. Here, the $[n = \tau - r_k]$ expression in the observation model indicates that if $x_{k,n,f}$ is aligned to time τ, then it only depends on the hidden coefficient $\lambda_{\tau,f}$, hence each observation coefficient is conditioned on a different hidden coefficient. The graphical model is shown in Figure 3.

The aim is to find most likely alignments of observed sequences denoted by $r^*_{1:2}$, which is actually the prime mode of the joint conditional posterior probability $p(r_{1:2}|x_{1:2,0:N_k-1})$. Assuming no prior information, likelihood, posterior and joint distribution are proportional. Hence, one can use $\Phi(r_{1:2}) = p(x_{1:2,0:N_k-1}, r_{1:2})$ as a scoring function. By choosing prior and likelihood distributions as conjugate pairs ,i.e., Gamma-Inverse Gamma, Bernoulli Bernoulli, analytical derivation of

$\Phi(r_{1:2})$ is possible by summing over $\lambda_{\tau,f}$. Then the optimum alignment is the one that maximizes the logarithm of $\Phi(r_{1:2})$, i.e., $\mathcal{L}(r_{1:2}) = \log \Phi(r_{1:2})$, This formulation can also be viewed as a Bayesian model selection problem [13]. We are comparing different configurations of $r_{1:2}$ to find the 'model' that describes the data best.

In the model as given in [1], each observation coefficient $x_{k,n,f}$ depends on only one of the hidden coefficients $\lambda_{\tau,f}$, if it is aligned to time τ. To obtain lower resolution data, we modify this idea such that L number of consecutive observation coefficient depends on one hidden coefficient $\lambda_{\tau,f}$. To illustrate the idea, the toy example in Figure 1 is also modified in Figure 4 where $L = 2$. The length of each sequence is halved and as it can be observed the coefficients $x_{1,4}, x_{1,5}, x_{2,2}$ and $x_{2,3}$ are aligned to time $\tau = 4$, hence they are noisy realizations of λ_4. We can also interpret the second row of each sequence like a new sequence that has to be exactly aligned with the first row. From this point of view, there are 4 sequences aligned at time $\tau = 4$. We define $n_l = \lfloor \frac{n}{L} \rfloor$ where $\lfloor . \rfloor$ is the floor operation and switch the local time index with n_l in the generative model which modifies the model for low resolution case. It is important to mention that there are other ways to obtain low resolution sequences rather than modifying the model such as increasing window size in feature extraction or downsampling before or after feature extraction. In this work,we just modify the structure of data without changing the actual resolution.

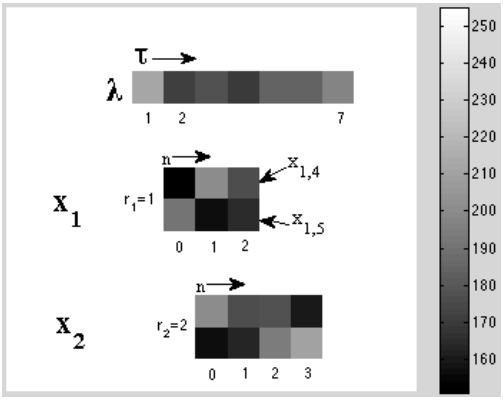


Fig. 4. Modified Toy Example From Figure 1

The posterior $\mathcal{L}(r_{1:2})$, would be equal for the alignments where the sequences do not overlap or where the amount of overlap between sequences is the same. Hence, if we fix the first sequence to $r_1 = N_2 + 1$, then the posterior becomes one dimensional $\mathcal{L}(r_2)$ and of length $N_1 + N_2$. Note that $\mathcal{L}(r_2 = 1)$ accounts for the score of not overlapping. For the ease of representation, we will use r instead of r_2 in the rest of the paper.

3. SEQUENTIAL MONTE CARLO SAMPLER

In this section, we introduce a SMC sampler based algorithm that uses low resolution $\Phi(r)$ as bridges. Here, the aim is to find the optimum alignment r^* without explicitly visiting all possible alignments. To achieve this, one needs a sampling mechanism that samples from $\Phi(r)$ and if some of the samples would eventually hit the mode of the distribution the optimum alignment would be found.

SMCS is a popular sampler due to its flexibility in design and ability to sample from rough and high dimensional densities. It samples from a sequence of distributions, denoted by γ_i, which are called intermediate distributions [14]. At each step, the algorithm samples from the next intermediate distribution and in the last step, the resulting samples would be drawn from the target distribution which is $\Phi(r)$ in our case. The main idea behind SMCS is that if the intermediate distributions in the consecutive steps are close enough to each other, they would act like a bridge and guide the samples through modes of the target density. At each step, new samples $r_s^{(i+1)}$ are drawn from a forward Markov transition kernel $K_{i+1}(r_s^{(i+1)}, r_s^{(i)})$ where s is the sample index and i is the dimension index. Then the discrepancy between the sampling distribution and intermediate distribution is corrected using importance sampling [14]. The weight of each sample is computed as,

$$w_i(r_s^{1:i}) = w_{i-1}(r_s^{1:i-1}) \frac{B_{i-1}(r_s^i, r_s^{i-1}) \gamma_i(r_s^i)}{K_i(r_s^i, r_s^{i-1}) \gamma_{i-1}(r_s^{i-1})}$$

where $B_{i-1}(r_s^i, r_s^{i-1})$ is a backward Markov kernel. The increase in variance of weights indicates that some of the samples have much higher importance weights than others. Hence, a resampling stage is applied to get rid off the samples with small weights and replicate the ones with higher weights. A common criteria to measure this degeneracy is the effective sample size (ESS) which is defined as $\left(\sum_{s=1}^{S} \left(w_s^{(i)} \right)^2 \right)^{-1}$ [14].

We choose the intermediate distributions as low resolution posterior distributions denoted by $\Phi_L(r)$ where $L = 2^l$, $l = 8, 7, \cdots, 1$. Note that the length of each $\Phi_{L/2}(r)$ is twice the length of one step lower resolution $\Phi_L(r)$, i.e., length of $\Phi_{64}(r)$ is twice the length of $\Phi_{128}(r)$. Hence, we need to design a forward kernel such that samples are moved from lower resolution to higher resolution. In SMC sampler framework, the choice of the forward and backward kernels are flexible so that any proposal mechanism is possible at any step of the algorithm, i.e., $K_i(.)$ do not have to be equal to $K_j(.)$.

For the forward kernel, we propose to move samples from lower resolution $(2L)$ to higher resolution (L) through some smoothed distributions of Φ_L. Defining Q as a smoothing kernel, one can obtain these distributions by applying Q several times to $\Phi_L(.)$, i.e., $Q^n\Phi_L, Q^{n-1}\Phi_L, \cdots, Q\Phi_L, \Phi_L$. Illustration of the smoothed distributions through each stage

and movement of a sample is shown in Figure 5. Note that smoothing kernel is chosen to be sparse so that one does not need to explicitly compute all values in $Q^n\Phi_L$,i.e., computation of a few values in $Q^n\Phi_L$ would be enough. We applied averaging kernel for smoothing purposes and backward kernel is chosen to be equal to forward kernel in the weight update.

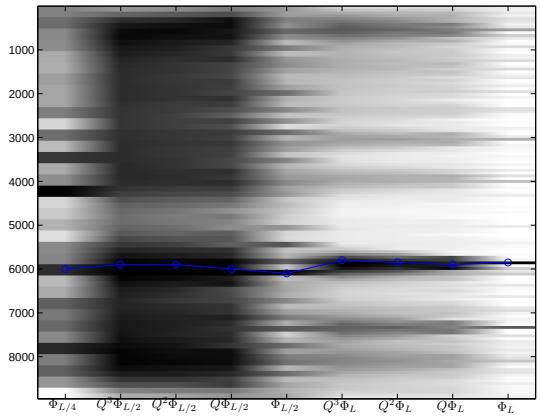


Fig. 5. Smoothed Bridge Distribution through each stage

One issue in the design of proposal is that proposal mechanism should be different for moving samples between smooth distributions ($Q^n\Phi_L$,$Q^{n-1}\Phi_L$) where resolution stays the same and for moving samples from low resolution (L) to high resolution (L/2) ($\Phi_L, Q^n\Phi_{L/2}$). In the latter case, a sample in the $(i-1)$'th stage in L resolution approximately corresponds to $r_s^{(i)} \approx 2 * r_s^{(i-1)} - 1$ in the i'th stage in $L/2$ resolution. Hence, proposed samples at these stages are chosen around $2 * r_s^{(i-1)} - 1$.

Note that none of the samples represent the case $\Phi(r = 1)$ which is the score for the sequences not overlapping. Simply by computing this value in the last step of SMC sampler where other samples are also drawn from $\Phi(r)$ and compare with the sample of highest score, one can easily decide whether or not the sequences overlap.

4. RESULTS AND CONCLUSION

In simulations, 20 datasets that include both speech and music recordings around 2 hours were used with hand labeled ground-truth. Each dataset consists of two overlapping or non-overlapping audio signals of varying length (from 30-40 seconds to 20-25 minutes), amplitude levels and noise content. The binary features are extracted by following the method in [3], which is basically taking the first difference of STFT on both time and frequency and then applying a threshold. The STFT resolution is 0.04ms and 32 sub bands are used.

In SMC sampler framework, intermediate distributions are usually annealed so that they become more similar [14]. Different annealing strategies are possible. Here, we anneal the intermediate distributions by adjusting the w parameter. When w is close to 0.5, the effect of data decreases therefore sequences could be aligned with less similarity. For lower resolution models, we choose smaller values for w and increase as the resolution increases. One of the major advantages of the algorithm is that, even if the corresponding alignment of the prime mode in lower resolutions is a local mode, the SMC sampler is still able to hit the prime mode in high resolution.

Another implementation issue is that the size of averaging kernel and/or number of appliance on the current target distributions can change over the steps of SMC sampler according to the resolution. As the resolution increases, we increase the number of appliance, hence have more smooth intermediate distributions for higher resolution steps which is observed to enhance the performance of the algorithm.

The number of samples used in SMCS is determined according to the length of Φ_L where L is the lowest resolution. For example if the length of the sequences $N_1 = 6500, N_1 = 7000$ and we start with a low resolution with $L = 256$, the length of sequences become $\lfloor 6500/256 \rfloor = 25$ and $\lfloor 7000/256 \rfloor = 27$ respectively. Then the number of samples is determined as 25+27-1=51.

The performance of the SMCS depends on the initial number samples and number of intermediate stages of same resolution level. By starting with enough number of samples and choosing proper w parameters for each stage, the SMCS is able find the ground truth for all data sets. The number of resolution levels may vary for different datasets, it is chosen manually such that minimum number of samples in a set is not below 20 in lowest resolution.

Rather than robustness, the computational efficiency of multi resolution model over naive computation of $\Phi(r)$ can be illustrated with an example ignoring the effect of smoothing operation. Defining the computation time for $\Phi(r)$ for any sample r as T_0, the computation time for the $\Phi_L(r)$ is $T_L = \frac{1}{L}T_0$ since the length of each sequence also decreases to $1/L$ of it. For each sample r_s^{i-1}, 2 samples are proposed for r_s^i, hence the number of required computation of smooth distributions $Q^n\Phi_L$ is 2. Assuming there are 4 stages of the same resolution, the number of required computations is 8 between each resolution change. For $L = 256$, the number of increase in resolution $\log_2 256$ is 8. Hence for one sample, the time elapsed in the end is, $8*(\frac{T_0}{128} + \frac{T_0}{64} + \frac{T_0}{32} + \cdots + T_0) = 14.5T_0$. Since the number of samples is approximately $1/256$ times of the original length N_1+N_2, the computational time for SMCS is computed as $\frac{14.5}{256} * T_0 * (N_1 + N_2) = 0.0566T_0$ which is lower compared to computing the $\Phi(r)$ for all possible alignments, i.e., $(N_1 + N_2) * T_0$. Hence, it can be concluded that SMC sampler with multi resolution intermediate distributions is both robust and computationally efficient and extending the framework to multiple cases rests as a future work.

5. REFERENCES

[1] D. Basaran, A. T. Cemgil and E. Anarm, Model Based Multiple Audio Sequence Alignment, in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics WASPAA '11, pp. 13-16, 2011.

[2] Wang, A.L, "An Industrial-Strength Audio Search Algorithm", InProc. ISMIR, Baltimore, USA, 2003.

[3] Haitsma J., Kalker T., "A Highly Robust Audio Fingerprinting System". in Proc. ISMIR Paris, France, 2002

[4] E. Weinstein and P. Moreno, Music identification with weighted finite-state transducers, in ICASSP 07, IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2, (Honolulu, HI), pp. 689692, April 2007.

[5] S. Dixon and G. Widmer, "Match: A music alignment tool chest", in Proc. ISMIR, London, GB, 2005

[6] S. Fenet, G. Richard, and Y. Grenier, "A Scalable Audio Fingerprint Method with Robustness to Pitch-Shifting", ;in Proc. ISMIR, 2011, pp.121-126.

[7] M. Muller, F. Kurth and M. Clausen, "Audio Matching via Chroma-based statistical features", In Proc. Int. Conf. on Music Info. Retr. ISMIR-05, pages 288-295, London, 2005.

[8] Bryan, N.J., P. Smaragdis, G.J. Mysore, "Clustering and Synchronizing Multi-Camera Video via Landmark Cross-Correlation", in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Kyoto, Japan. March 2012.

[9] L. Kennedy and M. Naaman, "Less talk, more rock: automated organization of community-contributed collections of concert videos," in Proc. 18th Int. Conf. on World Wide Web, 2009.

[10] Weber J. L., Myers E. W., "Human Whole-Genome Shotgun Sequencing", Genome Res. 1997 7: 401-409

[11] Brown, M. and Lowe, D., "Automatic Panoramic Image Stitching using Invariant Features", IJCV: Vol. 74, pp.59-73, 2007

[12] J.P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, "A tutorial on onset detection in musical signals", IEEE Transactions on Speech and Audio Processing, vol. 13, no. 5, pp. 1035-1047, 2005.

[13] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006

[14] P. Del Moral,A. Doucet,A. Jasra, "Sequential Monte Carlo Samplers " Journal of the Royal Society of Statistics, Series B. vol. 68, No. 3, pp. 411-436 (2006)