AUDIO PART MIXTURE ALIGNMENT BASED ON HIERARCHICAL NONPARAMETRIC BAYESIAN MODEL OF MUSICAL AUDIO SEQUENCE COLLECTION

Akira Maezawa[⋆†]

[⋆] Yamaha Corporation
R&D Division
203 Matsunokijima, Iwata, Shizuoka, Japan

Hiroshi G. Okuno[†]

[†] Kyoto University
Graduate School of Informatics
Yoshida Honmachi, Sakyo-ku, Kyoto, Japan

ABSTRACT

This paper proposes "audio part mixture alignment," a method for temporally aligning multiple audio signals, each of which is a rendition of a non-disjoint subset of a common piece of music. The method decomposes each audio signal into shared components and components unique to each rendition. At the same time, it aligns each audio signal based on the shared component. Decomposition of audio signal is modeled using a hierarchical Dirichlet process (Hierarchical DP, HDP), and sequence alignment is modeled as a left-to-right hidden Markov model (HMM). Variational Bayesian inference is used to jointly infer the alignment and component decomposition. The proposed method is compared with a classic audio-to-audio alignment method, and it is found that the proposed method is more robust to the discrepancy of parts between two audio signals.

Index Terms— Audio-audio alignment, Nonparametric hierarchical bayes

1. INTRODUCTION

Playing with an orchestra or top-notch musicians is among the favorite musical daydreams of amateur classical musicians. One way to realize such a dream, other than becoming a world-class artist himself, is to use MIR technology: source separation techniques can be used to generate a karaoke track of the user's favorite audio recording (henceforth referred to as "full audio"). Then, the user could play the solo part (henceforth referred to as "solo audio"), to which the karaoke track plays back in sync using alignment techniques.

In this kind of problem, *informed* source separation is the method of choice because unsupervised source separation difficult in a highly polyphonic mixture such as music, and additional information is highly helpful. For example, source separation becomes easier if the user could provide a digital music score data [1–5]. In this case, a good temporal alignment of the musical audio and the digital music score is critical [6–9]. It is also possible to aid separation by manually annotating the spectrogram [10].

Because the target user is a musician, we believe that informed source separation should be guided by the main control input of a musician: the audio of the musician's playing. Specifically, the method should take only two inputs: the solo audio played by the user, and the full audio. The method would separate from the full audio the component that corresponds to the solo audio. In this case, then, a good audio-to-audio alignment between the solo audio and the full audio is critical; users may not be necessarily skilled enough to play in sync with the full audio, as done in an existing study [11]. Solo-to-full audio alignment is a difficult task because these signals are spectrally very different. Thus, finding the alignment requires the full audio to be decomposed into solo part and the rest; however, the solo-to-full audio alignment is necessary to decompose the full audio into such components.

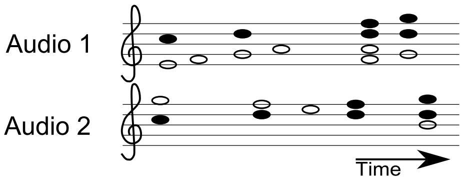

Fig. 1. Idea behind audio part mixture alignment. Assuming audio 1 and 2 play subsets of a same score, the method would align the audio signals by focusing on the notes played by both signals (shaded noteheads).

In this paper, we present an offline alignment method that is capable of aligning a solo audio and a full audio. Specifically, we define and present an *audio part mixture alignment* method, which is a generalized task that encompasses solo audio-to-full audio alignment. This objective of this task is to align multiple musical audio signals, each signal of which is an audio rendition of a unique but non-disjoint *subset* of a common music score. That is, we are interested in aligning, for example, a rendition of the violin plus the viola part, and the viola plus the cello part of a string trio. As illustrated in Fig. 1, the method would align audio signals by looking for notes that are played by two or more renditions. Our method is similar in spirit to noise-robust alignment [12], except we allow every recording to contain an audio signal unique to each recording, and signals played by two or more recordings. Hierarchical Bayes is used to model the collection of audio part mixtures, which allows us to model the notion of subsets to a common music score. Our model estimates spec-

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

tral time-slices that comprise the collection. Then, the sequence of activation of atoms is modeled using a hierarchical left-to-right HMM (LRHMM). The HMM is designed to take into account that each audio emits only some of the all possible atoms that may be emitted at a given state. Then, forced alignment is used to estimate the alignment of each audio signal.

2. FORMULATION

We assume that musical audio collection is comprised of a collection of symbols, each symbol of which is associated with an audio spectral time-slice. Moreover, we model music as a sequence of the set of symbols to emit, constrained such that every musical audio traverses the state sequence in the same order. Finally, we assume that for each state, each musical audio emits a subset of the set of symbols associated with the state. Then, the problem of part mixture alignment becomes that of inferring the state sequence associated with each musical audio.

Let us formalize this concept. Let $X(d,t,f)$ be the power spectrogram of dth audio signal in a collection of D audio part mixtures, evaluated at time t of duration T_d and frequency bin f of F bins. We interpret X as the number of times time-frequency bin (t,f) was observed for the dth signal, as in [13]. Let us introduce a variable $C(d,c,f,t)=1$, where c is defined for $c \in [1, X(d,f,t)]$. In other words, $\sum_c C(d,c,f,t) = X(d,t,f)$.

We assume that each count C is generated from a spectral "atom," which is an element from the set of all possible spectral time-slice that can be generated. An atom may resemble the spectral time-slice of, say, a note played by an instrument, or that of a percussion. Denote the kind of the distribution of atoms that can be emitted in the collection $X(d,t,f)$ as G_0. The number of atoms in G_0 should grow with the complexity of data. To this end, G_0 is modeled as a draw from the Dirichlet process (DP) with concentration parameter α and base measure H:

$$G_0 \sim \mathrm{DP}(\alpha, H) \tag{1}$$

In other words, H defines a measure over the space of distribution of atoms, and α controls the growth of the effective number of atoms in G_0. In our case, we choose H to be a Dirichlet distribution of length F, $\mathrm{Dir}(g_0(f))$, though other choices, such ones that incorporate harmonicity constraints [13], are possible as well. In effect, G_0 defines a countable collection of distribution of a spectral time-slice, whose effective size is governed by α and the observation.

Next, we consider the generative process of a music score. We formalize music score as a state sequence such that every signal that plays it traverses the sequence in the same order. Therefore, we model music score as a LRHMM:

$$Z(d, 1 \cdots T_d) \sim \mathrm{LRHMM}(\pi, \tau) \tag{2}$$

Here, $\mathrm{LRHMM}(\pi,\tau)$ denotes LRHMM with initial state pdf π, and state transition pdf τ. By definition, τ is constrained such that it either stays in the current state or moves to the next state. Note that we ignore structural discrepancies between two signals, such as repeats and cuts. Each state of the music score is associated with a subset of G_0, called a "chord." Each audio signal is allowed to emit, at each state, a subset of the chord. We denote the chord at state s as G_s, and assume it is drawn from a DP with concentration parameter β, and base measure G_0:

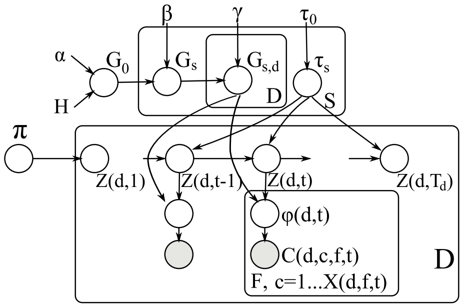

Fig. 2. Graphical model of our method.

$$G_s \sim \mathrm{DP}(\beta, G_0) \tag{3}$$

Intuitively, chords G_s correspond to the notated notes in the music score, and is a subset of G_0, which roughly corresponds to all notes notated in the piece of music; state sequence Z corresponds to sequence of positions in the music score.

Finally, we consider how each audio signal renders the music score. At each state s in the score, each audio signal chooses a subset of the chord G_s; for example, an audio signal of a violin solo that plays a violin/piano duo chooses to emit only the atoms corresponding to the violin part. Therefore, we model the actual atoms emitted by signal d at state s, $G_{s,d}$, as a draw from a DP with concentration parameter γ and base measure $G_{s,d}$:

$$G_{s,d} \sim \mathrm{DP}(\gamma, G_s) \tag{4}$$

We refer to the atoms of $G_{s,d}$ as "signal-specific" atoms at state s. At time t, the state sequence Z is referenced to find the state s, and then the corresponding $G_{s,d}$ generates a spectral atom $\phi_f(d,c,t)$:

$$\phi_f(d,c,t) \sim G_{Z(d,t),d}(f) \tag{5}$$

Finally, $C(d,c,f,t)$ is drawn from $\phi_f(d,c,t)$:

$$C(d,c,f,t) \sim \mathrm{Mult}(\phi_f(d,c,t)) \tag{6}$$

Here, $\mathrm{Mult}(\cdot)$ denotes the multinomial distribution. Fig. 2 shows the graphical model.

2.1. Rewriting as a conjugate model

Having defined the model, our goal is to determine the posterior distribution of the model. To this end, we seek to derive an inference scheme using variational Bayesian [14] (VB) method. Because VB inference is easy when the probabilistic model is conjugate, we rewrite our model to an equivalent conjugate model, based on Sethuraman's stick-breaking construction [15].

First, we rewrite G_0. Let us draw $I \to \infty$ spectral atoms from the base measure H using a stick-breaking process (SBP). Specifically, we draw $g_f(i) \sim \mathrm{Dir}\,(g_{f,0}(i))$ and let $w^{(g)} \sim \mathrm{GEM}\,(\alpha)$. $\mathrm{GEM}\,(\alpha)$ represents the SBP, which draws $w_i^{(g)}$ by first drawing $\xi_i^{(g)} \sim \mathrm{Beta}(1, \alpha)$ and letting $w_i^{(g)} = \xi_i^{(g)} \prod_{i'}^{i-1}(1 - \xi_{i'}^{(g)})$. In other words, the SBP replaces the DP by an equivalent representation, where we successively draw from the base measure and assign to the draw a multinomial likelihood of being chosen later on. The multinomial is generated by iteratively breaking off a stick with proportion $\xi^{(g)}$, and setting the proportion of the broken stick relative to the original length as the likelihood.

Next, we rewrite G_s by defining a DP over g for each state s. We draw $J \to \infty$ indicator variables $Z^{(A)}(s, j)$, jth of which indicates to which global atom the jth atom within a chord in state s refers. Specifically, we draw $Z^{(A)}(s, j) \sim \mathrm{Mult}(w^{(g)})$, and $w^{(A)}(s) \sim \mathrm{GEM}\,(\beta)$, by letting $\xi_j^{(A)}(s) \sim \mathrm{Beta}(1, \beta)$ and setting $w_j^{(A)}(s) = \xi_j^{(A)}(s) \prod_{j'}^{j-1}(1 - \xi_{j'}^{(A)}(s))$. In other words, atoms of G_s is realized by drawing j *indices* ranging from 1 to I according to $w_i^{(g)}$, the length of the stick broken for the top-level DP, and the result of the draw is stored by setting $Z^{(A)}(s, j) := i$. Then, each of the J draws are associated with a J-dimensional multinomial likelihood, drawn by iteratively breaking a piece of stick.

Next, we rewrite $G_{s,d}$ by drawing $K \to \infty$ indicator variables $Z^{(L)}(d, s, k)$, kth of which indicates to which state-atom the kth signal-specific atom refers. Specifically, we draw $Z^{(L)}(d, s, k) \sim \mathrm{Mult}(w^{(A)}(s))$, and $w^{(L)}(d, s) \sim \mathrm{GEM}\,(\gamma)$, by drawing $\xi_k^{(L)}(d, s) \sim \mathrm{Beta}(1, \gamma)$ and setting $w_k^{(L)}(d, s) = \xi_k^{(L)}(d, s) \prod_{k'}^{k-1}(1 - \xi_{k'}^{(L)}(d, s))$. In other words, we draw K indices according to $w^{(A)}(s)$, and associating with the draws a multinomial likelihood generated from the SBP.

Next, we model each count $C(d, c, f, t)$ as having originated from one of k signal-specific atoms, given $Z^{(S)}(d, t)$. We introduce a latent variable $Z^{(X)}$, which indicates the signal-specific atom that generated $C(d, c, f, t)$:

$$Z^{(X)}(d, c, f, t) \sim \mathrm{Mult}\Big(w^{(L)}(d, Z^{(S)}(d, t))\Big) \tag{7}$$

Then, the observation likelihood can be modeled as follows:

$$C(d,c,f,t) \sim \mathrm{Mult}\Big[g\Big(Z^{(A)}\Big(s, Z^{(L)}\Big(d, s, Z^{(X)}(d,c,f,t)\Big)\Big)\Big)\Big] \tag{8}$$

where $s = Z^{(S)}(d, t)$. Notice that variable ϕ_f introduced in the previous section has been replaced with $g\Big(Z^{(A)}\big(s, Z^{(L)}\big(d, s, Z^{(X)}(d, c, f, t)\big)\big)\Big)$. This means that ϕ_f is constrained to be one of draws from G_0, and the hierarchy of DP's used to arrive at ϕ_f is realized by traversing the index associated with an element of each layer's DP.

We shall now represent the latent variables as binary, 1-of-K variables. For example, $Z^{(S)}(d, t) = s'$ is denoted as $Z_s^{(S)}(d, t) = 1$ for $s = s'$ and 0 otherwise. Then, the complete log-joint likelihood is given as follows, up to a constant normalization factor:

$$\begin{aligned}
&\sum_{i,f,j,s,k,d,c,t} Z_i^{(A)}(s, j) Z_j^{(L)}(d, s, k) Z_k^{(X)}(d, c, f, s) Z_s^{(S)}(d, t) \log g_f(i) \\
&\quad + \sum_{s,i,j} Z_i^{(A)}(s, j) \log w_i^{(g)} + \sum_{d,s,j,k} Z_j^{(L)}(d, s, k) \log w_j^{(A)}(s) \\
&\quad + \sum_{t,c,f,k,d,s} Z_k^{(X)}(d, c, f, s) Z_s^{(S)}(d, t) \log w_k^{(L)}(d, s) \\
&+ \sum_{t,s,s',d} Z_s^{(S)}(d, t-1) Z_{s'}^{(S)}(d, t) \log \tau_{s,s'} + \sum_{t,s,d} Z_s^{(S)}(d, 0) \log \pi_s \\
&\quad + \log \mathrm{GEM}\left(w_i^{(g)} | \alpha\right) + \sum_s \log \mathrm{GEM}\left(w_j^{(A)}(s) | \beta\right) \\
&\quad + \sum_{d,s} \log \mathrm{GEM}\left(w_k^{(L)}(d, s) | \gamma\right) + \log \mathrm{Dir}\,(\pi | \pi_0) \\
&\quad + \sum_s \log \mathrm{Dir}\,(\tau_s | \tau_0) + \sum_i \log \mathrm{Dir}\,(g_f(i) | g_{f,0})
\end{aligned} \tag{9}$$

Again, this model is fully conjugate, so standard variational Bayesian method [14] and truncation approximation to Sethuraman's stick-breaking representation [15] can be used to arrive at an approximate posterior distribution. The derivation is omitted due to space constraints.

After the algorithm has converged, audio d can be aligned to d' by first finding the MAP state sequence $Z_s^{(S)}(d, t)$ of each signal. Then, a mapping from d to d' can be generated by finding the MAP state of d, $\hat{s}$, and finding the time at signal d' such its MAP state is $\hat{s}$.

3. EVALUATION

We compare our method against an alignment method based on dynamic time warping (DTW) that uses the cosine distance, one of the popular choices for similarity measure used in audio alignment [16] [1]. We first compare the two methods for full audio-to-full audio alignment task (i.e., typical audio-audio alignment). Then, we compare the two methods when the full audio is aligned against an audio part mixture. The baseline is used to see how a "symmetric" distance behaves[2].

For standard MIDI files to selected pieces of music as shown in Table 1, we synthesized each part using a software synthesizer using FreePats patch [17]. For the piano part, the right hand (RH) and the left hand (LH) were synthesized separately. For each music, we then prepared a time-stretched version of the full audio, by adding all parts to a given song, and time-stretching the resulting audio such that it is 20% slower than the original. Then, the alignment to the full audio was computed for stretched audio of the the melodic part and the stretched full audio. Finally, for each alignment computed, we computed the cumulative distribution of the absolute alignment error.

$X(d, t, f)$s were computed by evaluating the spectrogram of each audio signal using sampling frequency of 44.1kHz,

[1] We found cosine distance to be robust in a typical audio-score formulation, and is hence used as the baseline, instead of other possible distributions such as l1 or l2 norm.

[2] Note that the baseline is not intended to make claims about the performance of our method in comparison to current audio alignment methods, which incorporate features more sophisticated than simple power spectrum.

Table 1. Comparison of alignment error percentages with audio alignment based on cosine-distance DTW ("Cos.") and the proposed method ("Prop."), for alignment between full and full audio ("Full-Full"), and alignment between full and a melodic part audio ("Solo-Full"), average over pairs of audio parts ("Part-Part"), and average over parts pairs and full audio ("Part-Full").

Music	Part composition	Condition	err<0.05s		err<0.1s		err<0.3s		err<0.5s		err<1.0s	
			Cos.	Prop.	Cos.	Prop.	Cos.	Prop.	Cos.	Prop.	Cos.	Prop.
J.S. Bach, BWV847	Piano RH+LH	Full-Full	61%	52%	91%	83%	100%	99%	100%	100%	100%	100%
Fugue		Part-Full	49%	35%	75%	58%	92%	93%	93%	98%	93%	100%
F. Chopin, Op.22	Piano RH+LH	Full-Full	58%	57%	94%	90%	100%	100%	100%	100%	100%	100%
Bars 1-16 of Polonaise		Solo-Full	54%	42%	89%	70%	100%	97%	100%	100%	100%	100%
J. Brahms, Op.40	French Horn +	Full-Full	70%	53%	96%	84%	99%	100%	100%	100%	100%	100%
Bars 1-32 of Mvt. 1	Violin +	Part-Full	32%	40%	48%	59%	55%	87%	57%	92%	60%	96%
	Piano RH+LH	Part-Part	32%	35%	47%	53%	57%	84%	60%	93%	64%	99%
P. Tchaikovsky, Op.35	Violin solo +	Full-Full	64%	51%	98%	84%	98%	98%	98%	100%	100%	100%
Bars 8-34 of Mvt. 2	Orchestra	Solo-Full	46%	39%	74%	66%	80%	96%	81%	99%	84%	100%

frame length of 8192 samples with a 50% overlap, and windowed by the Bartlett-Hanning window. Then, frequency components greater than 2kHz were discarded. α, β and γ were initially set to 100, 50 and 50, respectively, and were optimized using empirical Bayes. S, I, J and K were set to $\min(T_1, T_2)$, 95, 20 and 10, respectively. $w_{f,0}(i)$ was set to $100e^{-\left(f-440\times 2^{\frac{i-69}{12}}\right)^2}$ so that each global atom, a priori, is assigned to unique pitch. π_0 was set such that it was 1 for the first index and 0 for all other indices. All variables except for π and τ are updated. For cosine-DTW method, we evaluated the same spectrogram, and used dynamic time warping (DTW) based on cosine distance to measure the dissimilarity between each spectral time-slice[3]. Because the typical path constraints used in DTW-based alignment [16] has worse worst-case error than a LRHMM, we changed the path constraint of the DTW method such that it is equivalent to a LRHMM. This makes the DTW method equivalent to Viterbi decoding of a LRHMM with von Mises-Fisher emission likelihood; therefore, the evaluation keeps the sequential model fixed, and compares only the model of the spectral time-slice.

Table 1 shows the alignment error for full audio-to-full audio alignment and alignment error when aligning a melodic part to the full audio. For Brahms Op. 40, we evaluated the average alignment over all possible combination of parts, as the melodic part interchanges among instruments. We also evaluated the average alignment over possible pairs of two or three parts mixtures. Note that with full audio-to-full audio alignment, our method performs similarly or worse than cosine-DTW, perhaps for two reasons. First, our method is susceptible to local optima whereas DTW is globally optimal. Second, cosine distance is a good spectral dissimilarity measure between two audio signals that play the identical score.

On the other hand, when aligning between two audio part mixtures, our method outperforms cosine-DTW in many cases, especially improving severe errors greater than 0.5 seconds. Cosine distance fails because full audio and audio part mixture may be spectrally quite different, even though they share same components. Our method absorbs such discrepancies, and hence maintains robustness against the comprising parts. This point is evident in Fig. 3, which shows alignment error percentile of part-to-part alignment of a piano trio. Cosine-DTW fails to align two part mixture audio signals because they are spectrally very different, even though they share some common instruments. On the other hand, our method is capable of decomposing the spectrum into constituent instruments, and hence is much more robust.

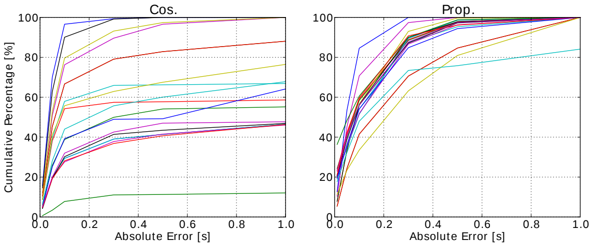

Fig. 3. Cumulative percentages of aligning part mixture combinations in Brahms Op. 40.

Our method suffers from small errors less than 0.5 seconds. We observed that this kind of error occurred when the HMM stayed in one state for too long, and "fast forwarded" to compensate for it. Because this kind of problem is inherent to the Markovian nature of the state sequence, employing a semi-Markovian state dynamics [6] might mitigate this problem.

4. CONCLUSION

This paper presented an audio part mixture alignment method. It is a new MIR task that seeks to align two audio signals, each audio signal of which plays a unique, non-disjoint subset of a common music score. The problem was tackled by modeling a collection of audio of musical part mixtures as a combination of a three-level HDP and a LRHMM. The essense of our method is simple: sequences can be aligned, if they share common parts. We believe that such an essense, conveyed thorough our model, is useful in other MIR tasks such as score following. Future work includes better temporal model and application to informed source separation.

[3] We also aligned two signals using their chromagrams. We omit the result since its part mixture alignment performance was worse than that using the magnitude spectrogram. This is presumably because the chromagram drops octave information, which is a valuable cue for identifying the parts played in common between two signals.

5. REFERENCES

[1] C. Raphael. A classifier-based approach to score-guided source separation of musical audio. *CMJ*, 32(1):51–59, March 2008.

[2] S. Ewert and M. Müller. Using score-informed constraints for NMF-based source separation. In *Proc. ICASSP*, pages 129–132, 2012.

[3] Y. Han and C. Raphael. Informed source separation of orchestra and soloist. In *Proc. ISMIR*, pages 315–320, 2010.

[4] K. Itoyama et al. Simultaneous processing of sound source separation and musical instrument identification using Bayesian spectral modeling. In *Proc. ICASSP*, pages 3816–3819, 2011.

[5] Romain Hennequin, Bertrand David, and Roland Badeau. Score informed audio source separation using a parametric model of non-negative spectrogram. In *Proc. ICASSP*, 2011.

[6] A. Cont. A coupled duration-focused architecture for real-time music-to-score alignment. *IEEE PAMI*, 32(6):974–987, 2010.

[7] R. B. Dannenberg and C. Raphael. Music score alignment and computer accompaniment. *CACM*, 49(8):38–43, August 2006.

[8] M. Müller and S. Ewert. Towards timbre-invariant audio features for harmony-based music. *IEEE TASLP*, 18(3):649–662, 2010.

[9] C. Joder, S. Essid, and G. Richard. A conditional random field viewpoint of symbolic audio-to-score matching. In *Proc. ACMM*, pages 871–874, 2010.

[10] Nicholas Bryan and Gautham Mysore. An efficient posterior regularized latent variable model for interactive sound source separation. In *Proc. ICML*, pages 208–216, 2013.

[11] P. Smaragdis and G. J. Mysore. "Separation by Humming:" user guided sound extraction from monophonic mixtures. In *Proc. WASPAA*, 2009.

[12] Brian King, Paris Smaragdis, and Gautham J Mysore. Noise-robust dynamic time warping using PLCA features. In *Proc. ICASSP*, pages 1973–1976, 2012.

[13] K. Yoshii and M. Goto. A nonparametric Bayesian multipitch analyzer based on infinite Latent Harmonic Allocation. *IEEE TASLP*, 20(3):717–730, March 2012.

[14] M. J. Beal. *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, University College London, 2003.

[15] C. Wang, J. W. Paisley, and D. M. Blei. Online variational inference for the hierarchical Dirichlet process. *JMLR*, 15:752–760, 2011.

[16] R. B. Dannenberg and N. Hu. Polyphonic audio matching for score following and intelligent audio editors. In *Proc. ICMC*, September 2003.

[17] E. A. Walsh. Freepats project. Retrieved from http://freepats.zenvoid.org/ on 8/1/2013, 2013.