PD Dr.-Ing. Tino Haderlein
Phoniatrische und pädaudiologische Abteilung in der HNO-Klinik
Universitätsklinikum Erlangen
Friedrich-Alexander-Universität Erlangen-Nürnberg
Bohlenplatz 21
91054 Erlangen
Germany

October 15, 2013

Opponent's Review for the Ph.D. Thesis of Ing. Štěpán Albrecht "Model-Based Approaches for Automatic Transcription of Music", West Bohemian University in Pilsen

a) Meaning of the Thesis for the Field

Model-based music analysis is difficult because real compositions do not contain the basic elements always in the same way and often not alone in one time interval. For identification, an automatic method must be able to compute significant, basic features from the recording in order to compare them with the reference from a sound library. Mr. Albrecht's thesis contributes to this field by identification of elements of the sound library based upon frequency and amplitude detection. A new aspect is that the sounds of the library do not have to be played completely in order to be recognized. This is achieved by the variational Bayes method which handles – among other parameters – the unknown information about truncation in a sequence of non-observable states.

b) Method of Problem Solving, Used Methods and Fulfillment of Targets

The experiments were performed on music which was supposed to be transformed into a description in the MIDI format. Similar to other fields of pattern recognition, the method tries to estimate a distribution of observable and hidden parameters from observable data features. The sequence of frames within a sound is described by a Markov chain. Due to the large number of unobserved parameters, some reasonable restrictions were made to the amplitudes and library sounds (no sudden changes in adjacent frames). Additionally, a library sound was assumed to be present just once in a time step. Unfortunately, an approach of Mr. Albrecht's publication at EUSIPCO 2010, where this restriction was not necessary, was not included in the thesis.
The Kullback-Leibler divergence served as measure for the accuracy of the estimation. The quality of the results were measured by self-defined hit measures between reference and estimate, and by the sound-to-distortion ratio measure (SDR). The presented approach was supposed to be able to compete with state-of-the-art methods. This target was reached.

c) Results of the Thesis

The data were tested on four different databases. One set of tests was performed without estimation of the amplitudes, in a second set the amplitudes were estimated. Some of the varied parameters significantly influenced the quality of the result, e.g. the number of sounds in the sound library and the test data. The accuracy in hits reached up to 94%. The comparison to results from the literature is not easy due to the different evaluation methods, the amount and composition of test data, but in general the variational Bayes modeling is among the most successful.

d) Systematics, Clarity, Formal Elaboration, and Language Level

The structure of the thesis is mostly clear. The mathematical notation and nomenclature of data and measures are sometimes a bit confusing for an unfamiliar reader. The captions of the figures sometimes give too few information which has to be looked up in the text. The use of English is mostly adequate. Some entries of the Bibliography section are truncated.

e) Publications of the Author

The publication list contains eight conference papers between 2007 and 2011. Mr. Albrecht is the only author of four of them, the others have one co-author. Two papers have been accepted at EUSIPCO (2010, 2011). Like these, one more paper (29th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Oxford, USA, 2009) is also listed in the ISI conference proceedings citation index. Three short papers were accepted at the German DAGA conference, which is also held and renowned in neighboring countries.

f) Recommendation for the Acceptance of the Thesis

The thesis and the publication list show that Mr. Albrecht is able to perform research independently. For this reason, I recommend the acceptance of his thesis for the granting of the academic title Ph.D.

Questions for the Defense of the Thesis

- p. 31, formula (3.2): Noise is assumed to be additive. Reverberation, for instance, is convolutive. How much effort would it be to extend the model to this type of noise?
- p. 32: I didn't understand from the description of the library sounds (transition between states) whether the model allows interrupted sounds, i.e. sound parts with a silent part section between them. Is this possible in the current model?
- What is the basic advantage of this method compared to the use of Hidden Markov Mixture Models (e.g. by Qi, Paisley, Carin, IEEE Trans. on Signal Processing, 2007)?
- In Chapter 3.6 (p. 45), an "online" model is presented, which, however, uses also "knowledge after [time] τ". How is this possible when the approach was "online"?
  This approach was published already in 2010. Why are there no results with it in the thesis?
  "Algorithm 3" needed 2 1/4 hours of processing time on a six-core computer for 5 1/4 minutes of data (p. 67). What was the performance of the online approach?
- Can you comment on the computational load that the other approaches of other groups had in comparison to yours?
- Chapter 4.8.4 (p. 70): Results for changes of the length of the observed signals are reported where the Mozart subpart of SD #3 was compared to the entire SD #3. But SD #3 contains 5 compositions of 5 different composers. How could it have influenced the results if the short section had not been taken from one composer only?
- Chapter 4.9.3 (p. 82): In the experiments, only the (o-bank, e-bank) pairs (SL #2, SL #1), (SL #3, SL #2), (SL #4, SL #2) were examined. Why?
- Chapter 4.9.3 (p. 82): What is the sense of estimating a common amplitude for all data?

Tino Haderlein

Prof. Ing. Jan Nouza, CSc.
Technická univerzita v Liberci
Ústav informačních technologií a elektroniky
Studentská 2, 461 17 Liberec

OPONENTSKÝ POSUDEK DISERTAČNÍ PRÁCE ING. ŠTĚPÁNA ALBRECHTA „MODELOVĚ ORIENTOVANÉ PŘÍSTUPY PRO AUTOMATICKOU HUDEBNÍ TRANSKRIPCI"

A. Význam disertační práce pro obor

Disertační práce Ing. Štěpána Albrechta se týká oblasti počítačové analýzy záznamu hudby a možnostmi její transkripce, tj. převedení polyfonní sekvence hudebních tónů do detailního zápisu odpovídajícímu např. standardu MIDI. Práce patří do specifického a úzce zaměřeného oboru zpracování hudebních signálů, jemuž se věnuje - i v celosvětovém měřítku - poměrně malá skupina vysoce specializovaných odborníků. Sám takovým specializovaným odborníkem nejsem. Při posuzování práce nicméně mohu vycházet ze svých zkušeností s alespoň částečně příbuzným oborem zpracování řeči a z poznatků získaných svými dřívějšími pokusy s elektronickou hudbou a z vedení dvou diplomových prací na podobné, byť ne tak komplexně pojaté téma.

Z přehledu literatury i z podrobné rešerše v úvodních kapitolách jsem nabyl přesvědčení, že autor práce tématu velmi dobře rozumí, je seznámen se současným stavem v oboru, je schopen kriticky zhodnotit práci jiných autorů a v neposlední řadě přispět svými vlastními návrhy na rozpracování a vylepšení existujících řešení.

B. Postup řešení, cíle práce

Autor si dal za cíl řešit úlohu transkripce záznamu polyfonní hudby (tedy hudby, v níž může současně znít více tónů). Obtížnost této úlohy spočívá zejména v tom, že hudební tóny obsahují velký podíl vyšších harmonických (podle typu nástroje) a při určitých souzvucích je proto těžké identifikovat jednotlivé tóny či nástroje. Autor svůj přístup definuje jako úlohu inverzního sekvenceru, který ze signálu vytváří pokud možno úplný symbolický zápis (např. typu MIDI). Vychází z toho, že pro nástroje, které se v záznamu objevují, má k dispozici knihovnu jejich zvuků. Úlohu přepisu pak řeší pravděpodobnostním přístupem, kdy hledá nejpravděpodobnější sekvenci a kombinaci tónů odpovídajících dané knihovně. V práci probírá a porovnává několik typů pravděpodobnostních modelů a v experimentální části pak ověřuje jejich úspěšnost. Využívá zde veřejně dostupnou databázi knihoven nástrojů a skladeb jimi nahraných. Cíle, které si vytyčil, splnil.

C. Výsledky a přínosy práce

Autor zformalizoval úlohu inverzního sekvenceru a navrhl její řešení metodou pravděpodobnostních modelů založených na bayesovském učení a maximálně věrohodném odhadu neznámých parametrů. Svůj přístup porovnal s metodami jiných autorů, přičemž na experimentálních datech ukázal, že jím navržený přístup může vést k přesnějším výsledkům při určování časových, hlasitostních i dalších atributů tónů.

D. Formální stránka práce

Práce je napsána v angličtině a tvoří ji 107 stran textu. Řazení textu je přehledné, od úvodní motivační části, přes rešerši současného stavu, podrobný přehled metod a matematického aparátu až po část popisující experimentální výsledky. Angličtina je velmi dobrá s minimem překlepů či chyb (např. "signing" místo "singing" na str. 2), případně nepřesných vyjádření (ve větě "There are 12 tones in one octave .." by spíše mělo být "12 semitones"). Větší problém jsem měl s pochopením některých klíčových obrázků. Není mi např. jasné jak interpretovat diagram na str. 34, který se později objevuje v dalších částech textu. Chápu, že na vodorovné ose je vynášen diskrétní čas, na svislé ose diskrétní frekvence tónů. Jaký význam mají v této diskrétní 2D mapě černé linie ležící mimo vyhrazené hodnoty? Není mi rovněž úplně jasný význam grafů na straně 61 až 65, znázorňujících statistiky skladeb používaných v experimentech. Grafy na str. 72 až 80 obsahují takové množství informací a detailů, že je neumím ani po přečtení stručného doprovodného textu interpretovat, natož mezi sebou porovnat. Zde bych - nejen jako oponent, ale zejména jako čtenář - uvítal podrobnější vysvětlení každého typu grafu, srozumitelné i pro toho, kdo není zrovna expertem v daném úzkém oboru. Různé varianty těchto grafů by pak mohly být spíše součástí příloh než hlavního textu.

E. Publikační aktivity autora

Tři články vážící se přímo k tématu práce byly již publikovány na prestižních konferencích indexovaných v registru ISI, dalších 5 článků lze nalézt ve sbornících mezinárodních regionálních konferencí, což považuji za přiměřené požadavkům na disertační práci.

F. Závěr

Disertační práci považuji za kvalitní a přispívající k rozvoji oboru počítačového zpracování hudby. Témata v ní rozvedená byla publikována na mezinárodní úrovni. Doporučuji proto konání obhajoby a udělení titulu Ph.D.

Otázky k obhajobě:

1) Skladby použité v testech byly vytvořeny na klavíru. Klavír při vyhodnocování často používají i jiní autoři (např. Klapuri). Oproti jiným nástrojům, např. dechovým, má výhodu ve snazší identifikaci nasazení tónu a v příznivém poměru amplitud základní harmonické a vyšších harmonických. Byly by navržené metody aplikovatelné i na další, např. dechové nástroje, případně jaký by mohl být dopad na úspěšnost?

2) Pokud jsem dobře rozuměl popisu experimentů, byl stejný typ klavíru používán jak při učení systému (z banky tónů), tak i při testování. Jaké jsou výsledky při použití dvou různých fyzických nástrojů (klavírů)?

3) Metody byly testovány na datech, které byly získány stejným (pravděpodobně značně ideálním) způsobem snímání zvuku nástroje. Jaký vliv na úspěšnost metod by měla situace, kdyby se hudba snímala běžným mikrofonem umístěným v prostoru sálu nebo kdyby se analyzovala nahrávka např. z hudebního CD nosiče?

V Liberci 15.12.2013

Jan Nouza

TAMPERE UNIVERSITY OF TECHNOLOGY
Department of Signal Processing

To University of West Bohemia, Faculty of Applied Sciences

Evaluation statement about the PhD thesis manuscript "Model-based Approaches for Automatic Transcription of Music" of Štěpán Albrecht

The doctoral thesis manuscript of Štěpán Albrecht deals with automatic transcription of music, which means converting a musical audio signal to a symbolic representation by computational algorithms. The research topic is important since such algorithms can be used in many applications, ranging from music education software to automatic indexing and retrieval of music databases. Automatic music transcription is a challenging and unsolved research problem, since the acoustic properties of natural music sounds have a wide range of characteristics, and multiple instruments can play simultaneously multiple tones.

The thesis manuscript suggests a clever approach where an audio signal to be analyzed is decomposed into units from a library of sounds. Since music is decomposed of atomic units such as individual notes, it is desirable to represent them as a sum of such basic units. This kind of approaches have recently been used successfully not only in music analysis, but for processing of other types of sounds. The manuscript describes a novel framework that allows selecting parts of sounds in the library. It is based on a probabilistic framework that allows combining different types of information.

The thesis consists of five chapters. The first chapter includes a review of previous music transcription systems which is claimed to include state of the art, but actually appears to be outdated. The most recent system that is included in the review is from 2008, but the field of automatic music transcription has made significant progress after then. State-of-the-art systems can be easily found since they typically participate in the annual MIREX evaluation. The manuscript should be update to include a description of the state of the art approaches. The chapter contains formulas of NMF estimation algorithms that are not related to the work done in the thesis, and should therefore not be included. In Section 1.1 the definition of "musical key" is inaccurate as it is not clear what is meant by "first note of a chord"; many instruments can play chords in the way that e.g. the note first in time or the note having the lowest in f0 does not correspond to the musical key. In Section 1.2 the description of the calculation of MFCCs is not correct since the discrete cosine transform is not applied on the segmented signal. Also the imaginary part of the DFT is not completely discarded, but an absolute value of the DFT is taken. MFCC are typically not obtained by taking the IDFT of the log-mel energies, but the IDCT of them.

Mailing address:
P.O.Box 553
FI-33101 Tampere
FINLAND

Street address:
Korkeakoulunkatu 1
33720 Tampere
FINLAND

Telephone:
+358 401981308
Fax:
+358 3 3115 4989

E-mail:
tuomas.virtanen@tut.fi

www.cs.tut.fi/~tuomasv

TAMPERE UNIVERSITY OF TECHNOLOGY
Department of Signal Processing

Chapter 2 presents a review of relevant probabilistic modeling tools that are used in the proposed algorithms. Part of the review is heavily based on reference [26], and Theorem 1 has been copied from the reference exactly. The reference is appropriately cited, but the copied part should be more clearly marked, for example using quotation marks. Section 2.6.1. is heavily based on reference [26], only with some minor modifications made. Since the main contributions of the section originate from [26], it would have been more appropriate to put the original text without modifications to an appendix, with an appropriate reference and explanation of the source of the the material. There is at least two minor errors in the chapter: Eq. 2.9 is not correct since the two leftmost terms $p(\Theta_\tau \mid D_\tau)$ and $p(\Theta_\tau \mid d_\tau , D_{\tau-1})$ are equal, and the product of two rightmost terms is not unity. In Eq. 2.12 the integration variable should be $\theta_{\tau-1}$, not θ_τ.

Chapter 3 presents the proposed algorithms. The chapter is mostly well written, but there as some minor unclarities and minor errors. Firstly, it is not clear what is the purpose of the Poisson model given in Eq. 3.7, since the model is actually not used at all. Would it not be simpler to define the model straightforwardly for the Gaussian distribution that is used in the actual method? Furthermore, Eq. (3.8) that is defined to describe a Gaussian distribution model is not complete. A Gaussian distribution is defined by a mean and variance, and based on Eq. 3.8 it is not clear what these are, and what is the random variable. Equations 3.4 and 3.5 do not match with Equation 3.6. In Eq. 3.4 variable $\boldsymbol{l}$ should be a column vector in order that it can be left multiplied with a matrix, but it is defined to be a row vector. If vector $\boldsymbol{l}_{s,\tau}$=[1,0,0,...] defined in the first row of Eq. 3.4 is multiplied with matrix $\mathbf{T}$ defined in Eq. 3.6, $\boldsymbol{l}_{s,\tau+1}$ becomes equal to the leftmost column of $\mathbf{T}$, i.e., $[t_{sil}\ t_{start}\ t_{start}...]$. This does not match with the second row of Eq. 3.5.

Chapter 4 presents the evaluation of the proposed method, which is based on synthesizing material from MIDI, analyzing the obtained audio, and comparing the results to the original MIDI. The evaluation is extensive, but there are some unclarities. On pages 61-65 where the simulation data is explained, it should be described how and exactly from which source the ground truth note onset times were acquired. The algorithms' ability to separate sources is used as an evaluation metric, but the test material consists of piano only. What does the source-to-distortion ration metric measure in this case? On pages 72-80 where the results are presented it is not clear which of the algorithm configurations 1-3 on pages 56-57 were used. The above missing information does not allow replicating the results according to good scientific practices.

Variable s in Eq. 4.3 has not been properly defined. The manuscript uses variable s simultaneously to denote the source index, which is confusing. The same variable should be used only for one purpose. For a reader it is also confusing that variable \hat{s} in Eq. 4.3 is not defined until on the next page. When variables are used the first time, they should be defined soon after. In the evaluations, it should be explained what principles were used to choose the values of the parameters. For example on page 68 $\sigma_{a,0}$ is close to zero, $\mu_{hyp,a,0}$=0.65, no description has been given how the values were set. Were these values chosen for example to maximize the performance on the test data? In Section 4.1, several observations are described starting from text "having t > 80..." to "...F converges to F^{est}". It has not been described how these observations have been made. On page 52 STFT bins higher than the Nyquist frequency are claimed to be affected by aliasing. This is not correct since aliasing occurs in the sampling of signals, not in the calculation of STFT.

Mailing address: P.O.Box 553 FI-33101 Tampere FINLAND
Street address: Korkeakoulunkatu 1 33720 Tampere FINLAND
Telephone: +358 401981308
Fax: +358 3 3115 4989
E-mail: tuomas.virtanen@tut.fi
www.cs.tut.fi/~tuomasv

Section 4.10 describes that the comparison is done to the state of the art, but the tested methods do not actually include the best performing algorithms from recent years. Each algorithm in the comparison uses different material, so comparing their accuracies does not give any reliable figures. The observation that the proposed solution competes well with the state of the art is not justified, since state of the art methods are not included, and the materials used are different.

In computer science, a good scientific practice requires comparing new computational models to existing ones in a fair evaluation. In the field of automatic music transcription, ways to obtained comparable results include sharing resources (signals and software), implementing previous methods based on publications, or asking fellow scientists to run experiments. This thesis work does not fairly evaluate the accuracy of the proposed computational models in comparison to other methods, which limits the credibility of the work. The thesis, however, evaluates the performance of the developed computational model in comparison to its different variants, which are scientifically novel results.

Chapter 5 presents the conclusions of the work.

The manuscript has been written using good-quality English. A minor error includes "can be performed unsupervised", which should be e.g. "can be performed in an unsupervised manner". Small typographical errors include "signing transcription" (should be "singing transcription"), "Kullaback-Leibler" (Kullback-Leibler), "gradient descend" (gradient descent), "bayesian" (Bayesian), "markov" (Markov), "kalman", (Kalman), "midi" (MIDI), "mpeg" (MPEG), "label matrix L excerpt ." (an extra space before the period). The role of each chapter of the manuscript is clear, but the structure within each chapter is in some places somewhat difficult to follow. Sentence "the estimation of all sounds in time" in the abstract is unclear as it does not describe what "estimation of sounds" means.

The manuscript follows mostly good scientific citation practices, with some exceptions. The first two sentences of Chapter 1 are directly copied from reference [1, page 3]. There is a citation to the reference, but any directly copied text should be more explicitly marked as a quotation. Similarly, on page 7 sentence "The divergence cost (3) of an individual observation Y is linear as a function of the scale of the input, since $D(\alpha p, \alpha q) = \alpha D(d,p)$ for any positive scalar α, whereas for the Euclidean cost the dependence is quadratic." has directly been copied from reference [1, Chapter 9], without marking it as an exact quotation. I did not search for all cases similar to the ones above, but gave them just examples. In any similar cases where direct quotations are used, the quotations should be marked more explicitly. There are several references to books, especially [1]. Exact page numbers should be given whenever a specific part in the book is referred to, for example just before Eq. (3.3). Book [1] is an edited one, and several chapters are written by authors that are not editors of the book. Therefore any reference to a chapter that is not written by the editors should be its own reference that includes the chapter author names.

There is plenty of missing information from the references. Complete publication information is missing from references 20, 36, 48, 57, 80, 84, and 91. From online references 92 and 93 the date when they have been accessed is missing. References 32 and 33 are duplicates. The title of the reference 36 is partially missing. Full publication names should be used instead of acronyms, for example in references 40 (ICA), and references 64-65. In reference 47 work "in" appears twice.

Mailing address:
P.O.Box 553
FI-33101 Tampere
FINLAND

Street address:
Korkeakoulunkatu 1
33720 Tampere
FINLAND

Telephone:
+358 401981308
Fax:
+358 3 3115 4989

E-mail:
tuomas.virtanen@tut.fi

www.cs.tut.fi/~tuomasv

TAMPERE UNIVERSITY OF TECHNOLOGY
Department of Signal Processing

To conclude my review, the manuscript presents novel scientific results that are supported by sufficient experiments. However, because of the shortcomings that I describe above, the manuscript needs to be revised. Not all the issues pointed out are necessary for an acceptable doctoral thesis, but at least the issues related to missing state-of-the-art references, referencing and quotation style, any technical and language errors, and critical information about the evaluation procedure and material needs to be addressed. Many of the above remarks should have been addressed by the thesis supervisor, not an external examiner. I feel I have already contributed sufficiently to this thesis work as an external examiner, and would not like to review revised versions of the manuscript.

In Tampere 10.1.2014

Dr. Tuomas Virtanen
Academy Research Fellow / Adjunct Professor
Tampere University of Technology

Mailing address:
P.O.Box 553
FI-33101 Tampere
FINLAND

Street address:
Korkeakoulunkatu 1
33720 Tampere
FINLAND

Telephone:
+358 401981308
Fax:
+358 3 3115 4989

E-mail:
tuomas.virtanen@tut.fi

www.cs.tut.fi/~tuomasv