Microsoft Corporation
One Microsoft Way
Redmond, WA 98052-6399

Tel 425 882 8080
Fax 425 936 7329
http://www.microsoft.com/

Microsoft

November 13, 2012

Dr. Carlos Agon
Institut de Recherche et Coordination Acoustique Musique
Universite Pierre et Marie Curie
Paris, France

Dear Dr. Agon:
This is my review of the thesis titled "Multiobjective time series matching and classification" by Phillipe Esling.

Most importantly, I learned something from this thesis. While I do not agree with the utility of all the algorithms presented in this thesis, I am intrigued with the general ideas presented by Mr. Esling. This is the most important factor I use when evaluating a thesis. The thesis is a pleasure to read. The work is well written and researched.

The thesis explores many idea around using multi-objective optimization algorithms. The thesis claims that this approach is important when there is no single objective function, which is often the case when dealing with perceptual issue. Often there might be a number of optimal solutions, and it is up to later processing to figure out what to do with the multitude of solutions.

The thesis is a well-written compendium of issues related to content-based audio and music similarity. It has an extensive introduction detailing many of the related technologies. It describes using multi-objective optimization for classification and identification. It finishes with a description of how these algorithms can be used to compose new music.

The thesis starts with a discussion that many perceptual factors are not quantifiable with a single distance metric. This is especially true with an acoustic measure such as timbre. Still people want to have a distance metric, and the author should acknowledge that content-based methods are problematic [Slaney, Does Content Matter?, IEEE Multimedia Magazine, Spring 2011]

Most importantly, this thesis emphasizes time-series recognition, but yet ignores the wealth of knowledge resulting from speech recognition. Speech recognition is definitely a time-series recognition problem, and has met with much success. The thesis describes many types of time-series algorithms, but ignores modern speech recognition. (I don't consider vowel recognition to be a serious attempt at speech.)

I have to admit I'm confused about how this multi-objective approach is used for pattern classification. Classification is a single-objective problem. I think the thesis talks about finding the Pareto boundary, which makes sense. And then finding the distance to these boundaries. Why is this a good classification scheme? I believe the Pareto boundary is based on the features used to describe the data. But this is an arbitrary set of descriptors.

Microsoft Corporation
One Microsoft Way
Redmond, WA 98052-6399

Tel 425 882 8080
Fax 425 936 7329
http://www.microsoft.com/

Microsoft

compared to modern problems, since there is a direct comparison to other algorithms on a range of existing data.

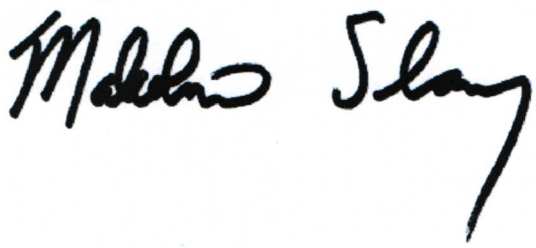
Music composition seems like another good use for the multi-objective approach. I know very little about composition, and it seems useful to be able to pick a number of different directions for the next segment of music. But in the end, unless there is a human in the loop, isn't there just a single objective? How does this system decide the best direction to take, from all the possible solutions on the Pareto boundary? Doesn't this just simplify to a single-objective, conventional music-similarity decision?

Conclusions

Not withstanding these questions and comments, which are a healthy response to a good piece of research, I am happy with the research presented here, and intrigued enough with the approach that I want to learn more. That is the mark of a good thesis!

For these reasons, I think that this PhD is ready to be defended on December, 5th, 2012.

Sincerely,



Malcolm Slaney
Principal Scientist, Microsoft Research Converational Systems Laboratory
(Consulting) Professor, Stanford CCRMA

Biography: Malcolm Slaney is interested in building computational models of users, sounds, images, and video in order to better connect users and signals. For the last 20 years he has organized the Stanford CCRMA Hearing Seminar, where he is a (Consulting) Professor. Before joining Microsoft he was a researcher at Yahoo and IBM's Almaden Research Center, working on multimedia analysis and user models. He has also been employed by Interval Research, Apple's Advanced Technology Group, Schlumberger's Palo Alto Research Laboratory, and Bell Labs. He is the coauthor of the book "Principles of Computerized Tomographic Imaging," which was recently republished by SIAM as a "Classics in Applied Mathematics." He is coeditor of the book "Computational Models of Auditory Function." He has served as an associate editor for IEEE Transactions on Audio, Speech and Language Processing; IEEE Multimedia Magazine; ACM Transactions on Multimedia Computing, Communications, and Applications; and the Proceedings of the IEEE. He is a Fellow of the IEEE.