MUSIC ACOUSTIC FEATURES: DO MACHINE PREDICTIONS CORRESPOND TO HUMAN JUDGMENTS?

Maya B. Flannery, Department of Psychology, Neuroscience & Behaviour, McMaster University
Matthew H. Woolhouse, School of the Arts, McMaster University
Contact: flannerm@mcmaster.ca

BACKGROUND

Researchers have great difficulty describing music objectively.
Musical *genre* is often used in research but is highly subjective and arbitrary[1].
Potential solutions include methods of Music Information Retrieval (MIR)[2] and the use of structural and expressive *musical cues*[3].

OBJECTIVES

Our goal was to establish *Music Acoustic Features* (MAFs) as a reliable method of music classification and description for experimental research.
MAFs are a combination of MIR methods (as used by the *Essentia* library) and musical cues and satisfy three criteria. MAFs must be:

1. *Manipulable*: such that produced music varies along a given measure.
2. *Measurable*: objective analysis can determine the level of a feature.
3. *Readily perceivable*: 1 and 2 correspond with perceived differences in stimuli.

METHODS

MAFs were investigated though a four-step procedure.

1. Training data were generated consisting of labelled MAFs.
   - Six MAFs were selected: Articulation, Dynamic, Register, Tempo, Texture, and Timbre.
2. Features were extracted from the training dataset with the MIR tool Essentia; models were trained to predict MAFs from Essentia features.
   - `MusicExtractor()` returned approximately 450 features
   - Linear regression initially used; other models under investigation.
3. Models were applied to predict MAFs in real-world musical excerpts.
   - Forty-four excerpts from 7 to 25 seconds long.
4. A listening experiment was conducted.
   - Participants (N = 43) provided ratings for MAFs for the same real-world excerpts.

A final analysis compared predictions from models to participants' ratings.

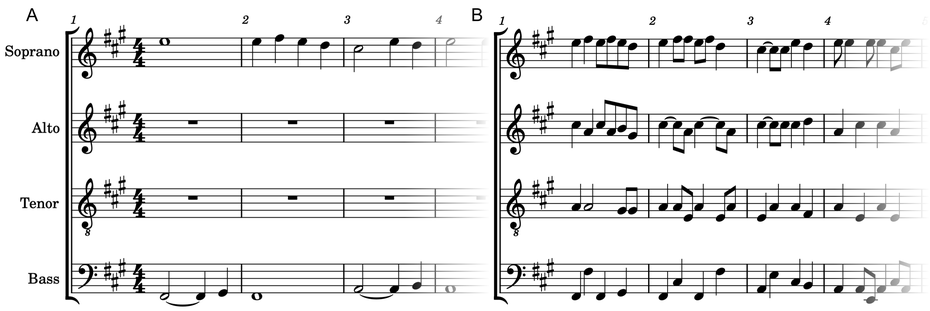

Figure 1: [Step 1] Example of texture manipulation (5 levels). Two levels are shown: very sparse texture (A) and very dense texture (B). Articulation (4), Dynamic (3), Register (5), Tempo (4), and Timbre (4) were programmatically manipulated, resulting in 4800 stimuli.

RESULTS

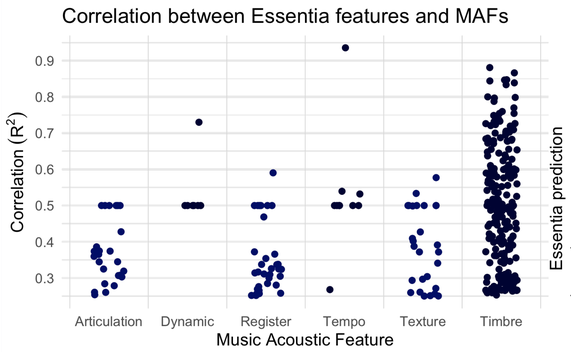

Figure 2: [Step 2] Linear regression was used to predict MAFs from Essentia features. Each point represents an Essentia feature's strength (y-axis, R^2) in predicting a particular MAF shown on the x-axis.

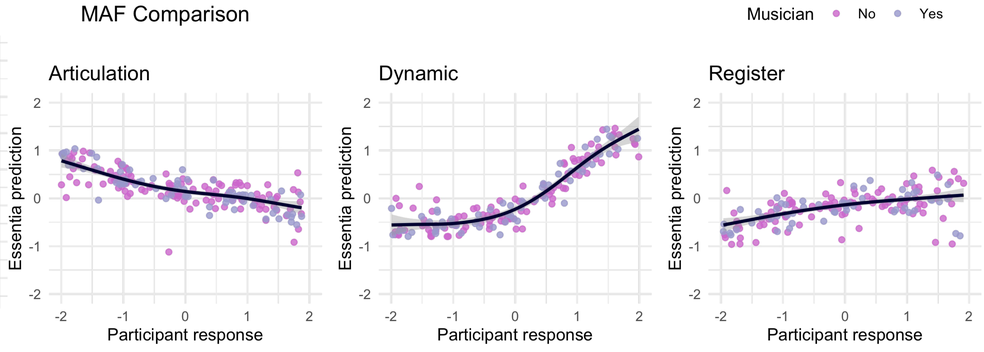

Figure 4: [Analysis] Participants' responses (x-axis) were compared to the Essentia model predictions (y-axis). Both axes are converted to z-scores. Each point represents a participant's rating for one excerpt and the corresponding model prediction.

1. All possible combinations of stimuli were produced (4800 in total).
   - Articulation (4 levels, staccato–legato), Dynamic (3 levels, soft–loud), Register (5 levels, -8ve to +8ve), Tempo (4 levels, slow–fast), Texture (5 levels, very sparse to very dense), and Timbre (4 levels, dark–bright).
2. Analyses of Essentia features returned 315 features with R^2 greater than 0.25.
   - A single optimal feature was selected to predict each MAF in Step 3.
   - This produced a model in the form of $MAF_{i,j} = \beta_{0i} + \beta_{1i} \times EF_{i,j}$ (i = MAF; j = musical excerpt)
3. Models were applied and predicted MAFs from real-world stimuli.

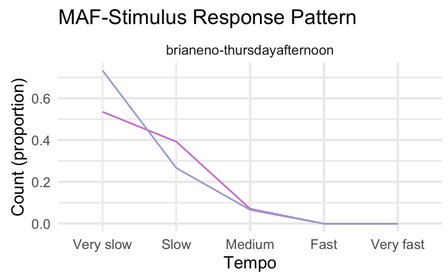

Figure 3: [Step 4] Participants rated MAFs while listening to musical excerpts. The number of times each rating (x-axis) was selected was counted and is shown on the y-axis as a proportion of total ratings.

4. Participant responses were counted for each MAF-excerpt to determine if responses were consistent.

The correlation between Participant responses and Model predictions were measured to determine if MAFs were identified consistently:

- Articulation, $R^2 = 0.086$, $p < 0.001$
- Dynamic, $R^2 = 0.301$, $p < 0.001$
- Register, $R^2 = 0.066$, $p < 0.001$
- Tempo, $R^2 = 0.431$, $p < 0.001$
- Texture, $R^2 = 0.129$, $p < 0.001$
- Timbre, $R^2 = 0.051$, $p < 0.001$

CONCLUSIONS

MAFs can be reliably produced and manipulated, effectively measured within audio stimuli, and readily perceived by listeners.
Since MAFs can be measured from an audio signal, they can be widely applied to categorize and describe existing music.
Furthermore, since MAFs are manipulable, they can be reliably used in experimental contexts.
These use cases will greatly improve the conclusions and interpretations of research in music psychology.

References: 1. Aucouturier, J.-J., & Pachet, F. (2003). Representing Musical Genre: A State of the Art. Journal of New Music Research, 32(1), 83–93. **2.** Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J., & Serra, X. (2013). Essentia: an open-source library for sound and music analysis. Proceedings of the 21st Acm International Conference on Multimedia, 855–858. **3.** Eerola, T., Friberg, A., & Bresin, R. (2013). Emotional expression in music: Contribution, linearity, and additivity of primary musical cues. Frontiers in Psychology, 4, 487. **Ethics:** Approved by McMaster Research and Ethics Board: #2524. **Funding:** SSHRC Insight Development Grant (#430 2012 0835) and McMaster Arts Research Board.