2025年3月 MUS研究会

知覚感情の不整合: 人間が作曲した音楽，音楽を記述したテキスト，テキスト楽音合成による音楽の比較

阪井 瞭介[1]，松下嶺佑[1]，福田 航希[1]，高道 慎之介[1,2]，植村あい子[3]

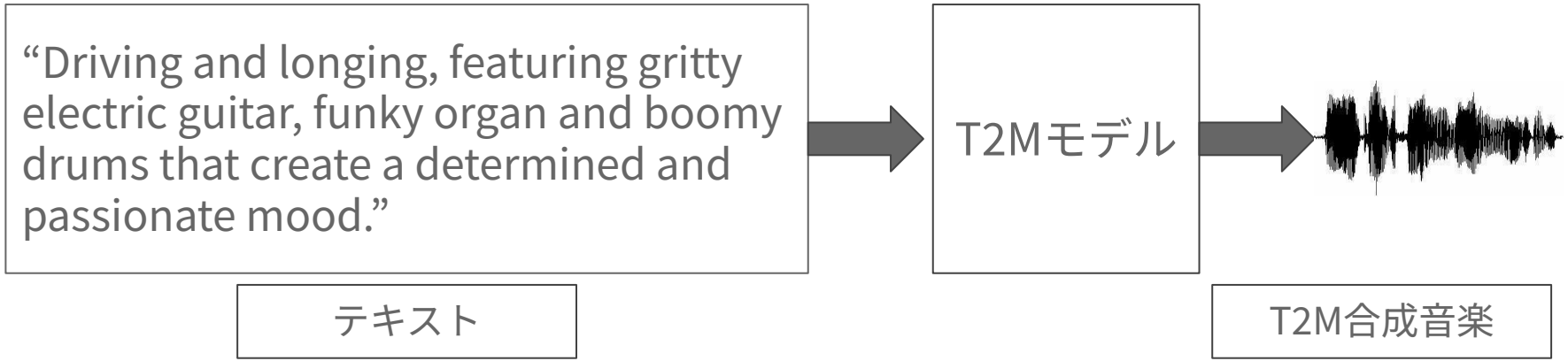
(1: 慶大, 2: 東大, 3: 津田塾大学)

Takamichi Laboratory
慶應義塾大学 高道研究室

研究背景

T2M (text-to-music) モデルとは

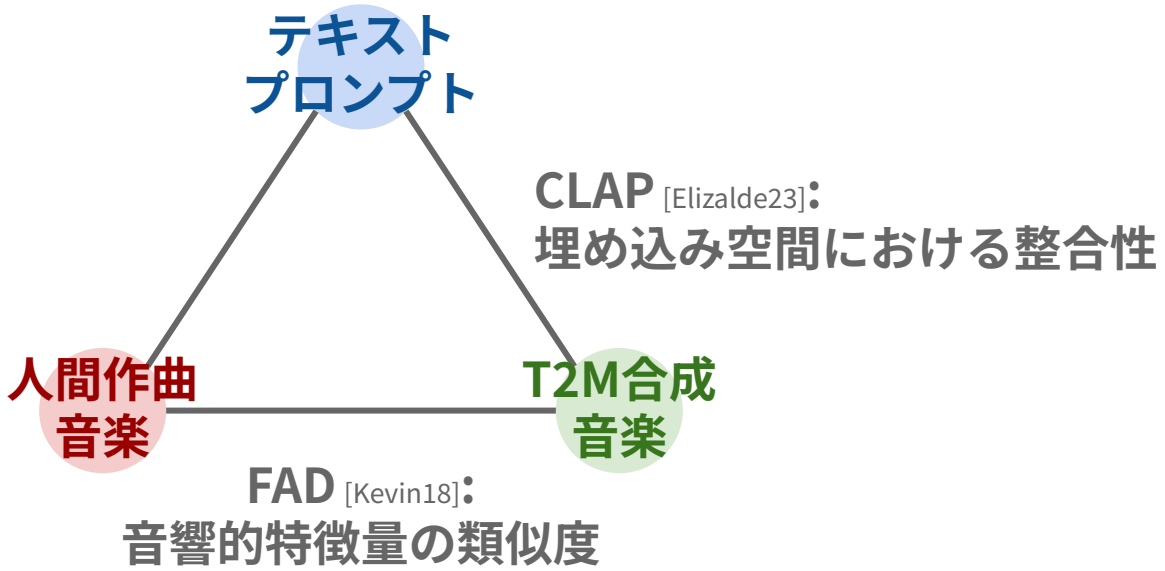
- テキストを入力として，音楽 (ボーカル無し) を生成する．
- 深層学習の発展より，複数のモデルの開発が進んでいる．
  - MusicGen [Copet23]，MusicLDM [Chen23]，Stable Audio [Zach24]など．
- T2M合成音楽は商業施設や教育分野など*1で幅広く，使用されている．



*1: 朝日新聞．“音楽授業にAI，みんな作曲家　埼玉の小学校歌詞に対応したメロディを自動生成” 2024/6/23

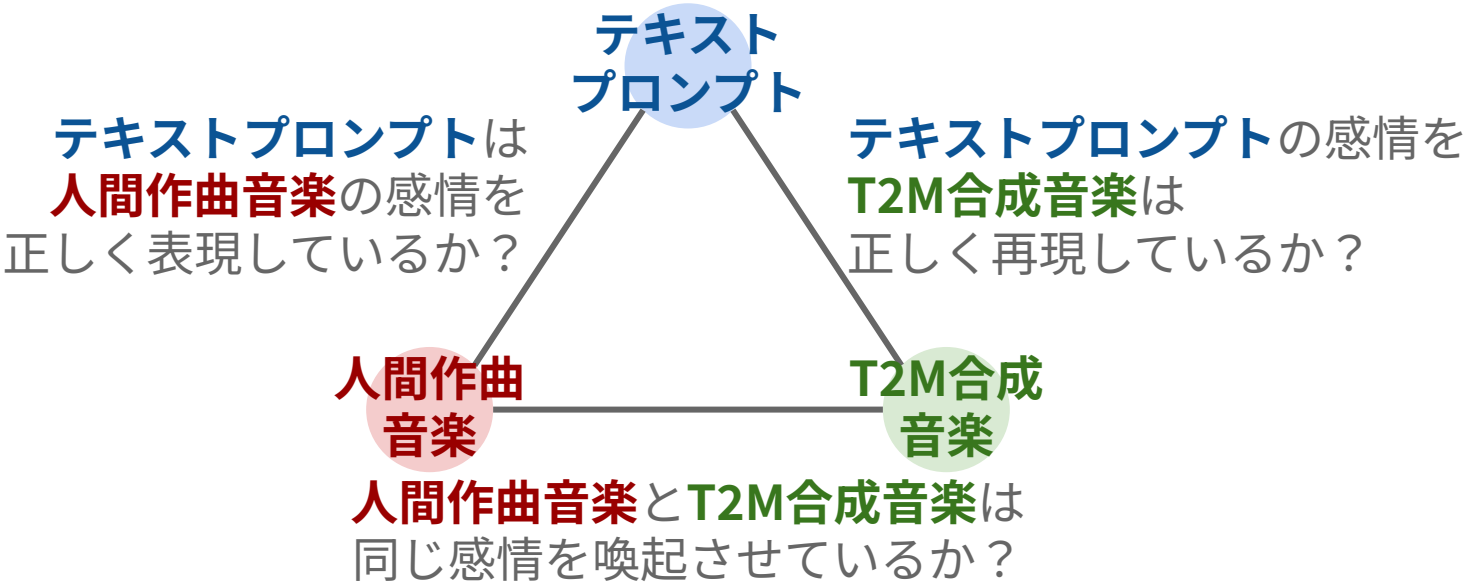
T2Mの普及に向けた，T2M合成音楽の評価

- T2M合成音楽**単体**に対する評価
  - 音楽らしさ [Cheng18]，好感度 [Ken22]，合成音楽に感じるか [Pedro18]など．
- T2M合成音楽に**対応のあるデータと**の評価 (特徴量・表層レベルの評価)



本研究の立ち位置: 知覚感情での整合性

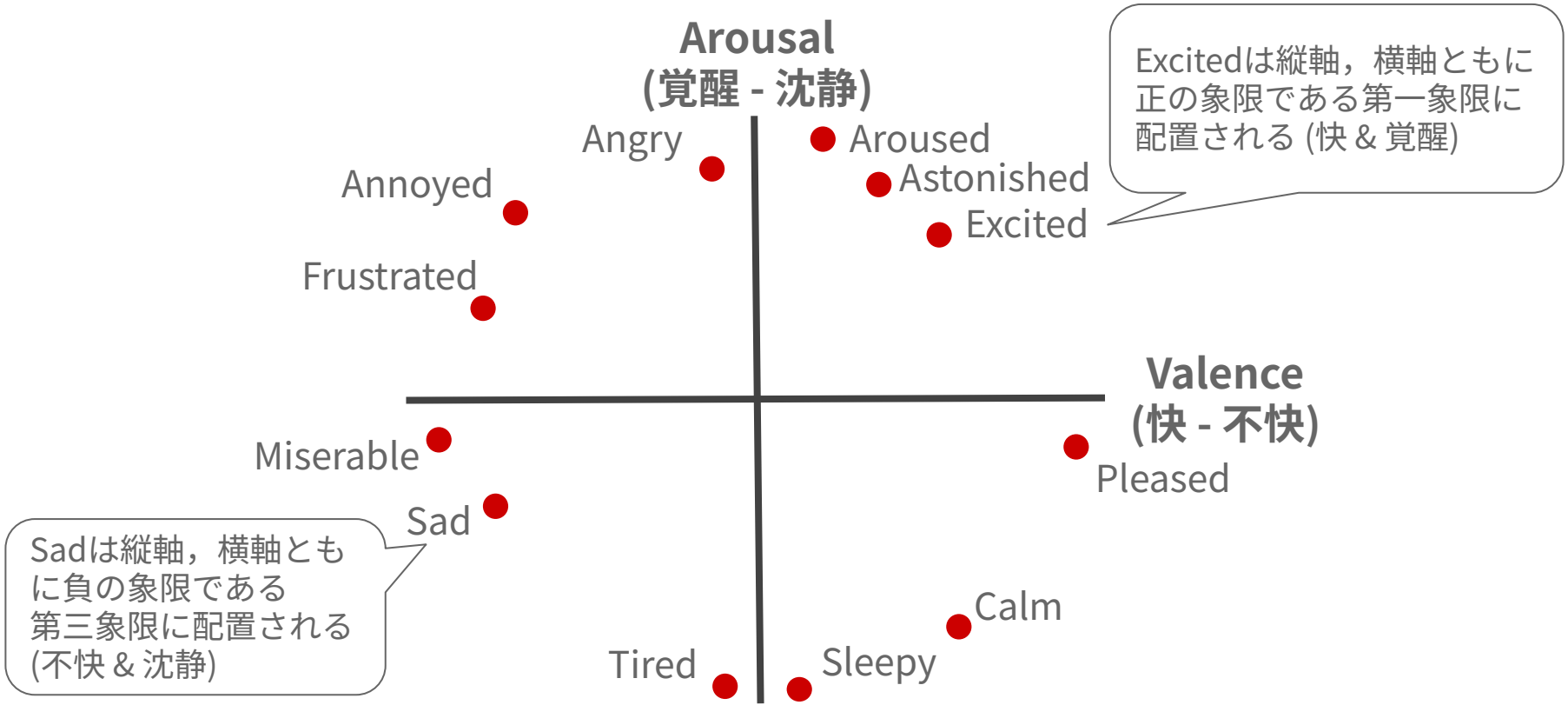
- 音楽を聴取する目的: 感情の喚起 [森 14]
  - T2M合成音楽とテキストの整合は聴取者からの**感情レベル**でも測るべき.
- 音楽から喚起される感情は属性によって異なる [En23, Xuneng24, Marcus15] .
  - 聴取者属性の要素間で評価に差異があるかを分析



提案手法

ラッセルの円環モデル[Russell 80]を用いた，聴取者による感情評価

- 離散値ではなく，連続値で感情を表現



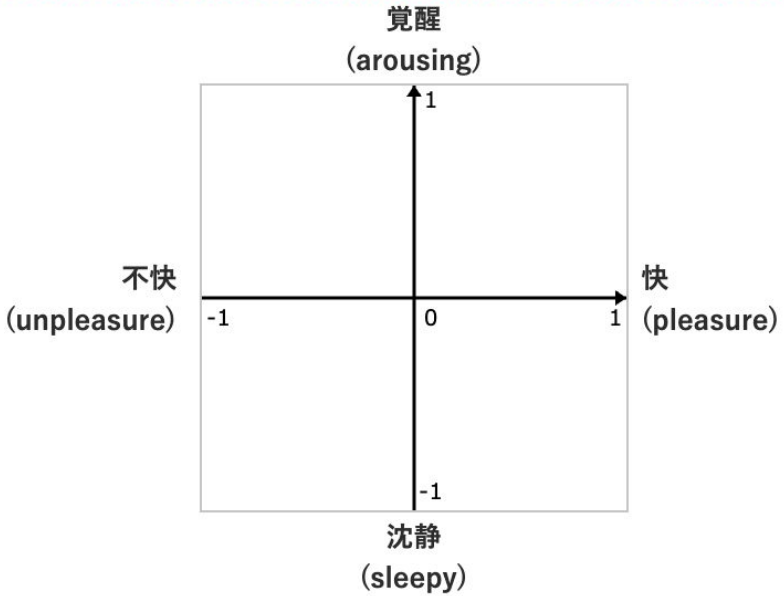
知覚感情の評価手法

- 聴取者による評価実験
  - 複数の人間作曲音楽，テキスト，T2M合成音楽の対データを提示
    - 順番はテキスト → 音楽 → 音楽 (音楽の順序はランダム)
    - 聴取者には音楽の生成元 (人間 or T2M) は明示しない．
  - それぞれに対しての知覚感情を**二次元座標空間にプロット**させる．
    - テキスト: テキストから連想される音楽の感情
    - 音楽: 音楽を再生し，聴取者が感じた感情
- 聴取者属性をアンケートにて収集

評価ページについて

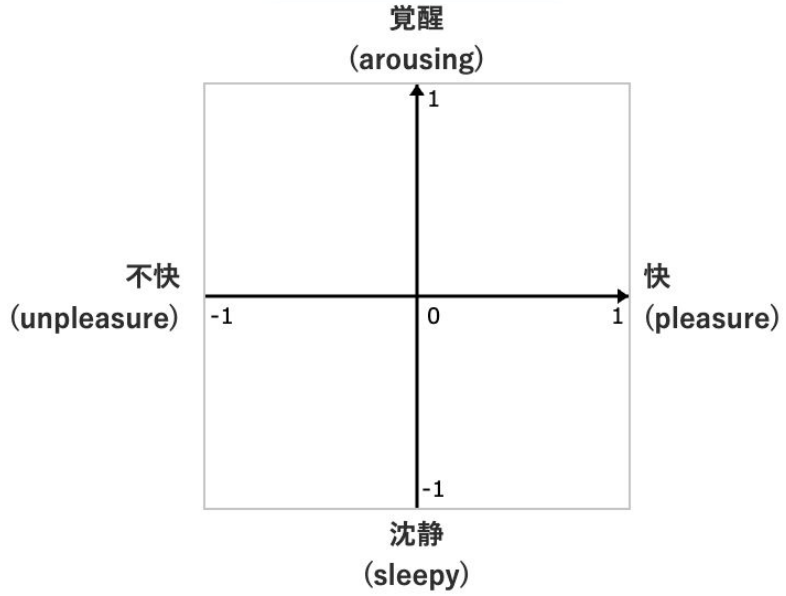
以下のテキストから連想される音楽の雰囲気を想像し、
その雰囲気に合う場所をモデルの座標上でクリックしてください

夢のようで幻想的な雰囲気で、浮遊感のあるシンセサイザーとドローン音が特徴です。
穏やかで内省的な気分を作り出します。



音楽を再生し，あなたが感じた感情を以下のモデル内の座標にクリックしてください．そして，再生した音楽が人間によるものか，生成AIによるものかをあなたが感じた方にチェックを入れてください

0:00 / 0:00



○ 人間による音楽
○ AIによる音楽

実験結果

実験条件

聴取者人数		206人
聴取者属性	性別	男性:133人，女性: 72人，回答無し: 1人
	年齢	10代: 0人，20代: 9人，30代: 45人，40代: 94人，50代: 41人， 60代: 13人，70歳以上: 4人
	好きなジャンル	ポップ: 87人，ロック: 52人，クラシック: 21人，その他: 46人
	楽器演奏経験	有り: 102人，無し: 104人
	その他	経験有り楽器の数，総演奏年数など
対データの収集	テキスト 人間作曲音楽	音楽素材配布サービス (Shutterstock[*2]) にてダウンロード
	T2M合成音楽	上記のテキストをプロンプトとして，MusicGenで合成
提示音楽の時間長		それぞれ15秒

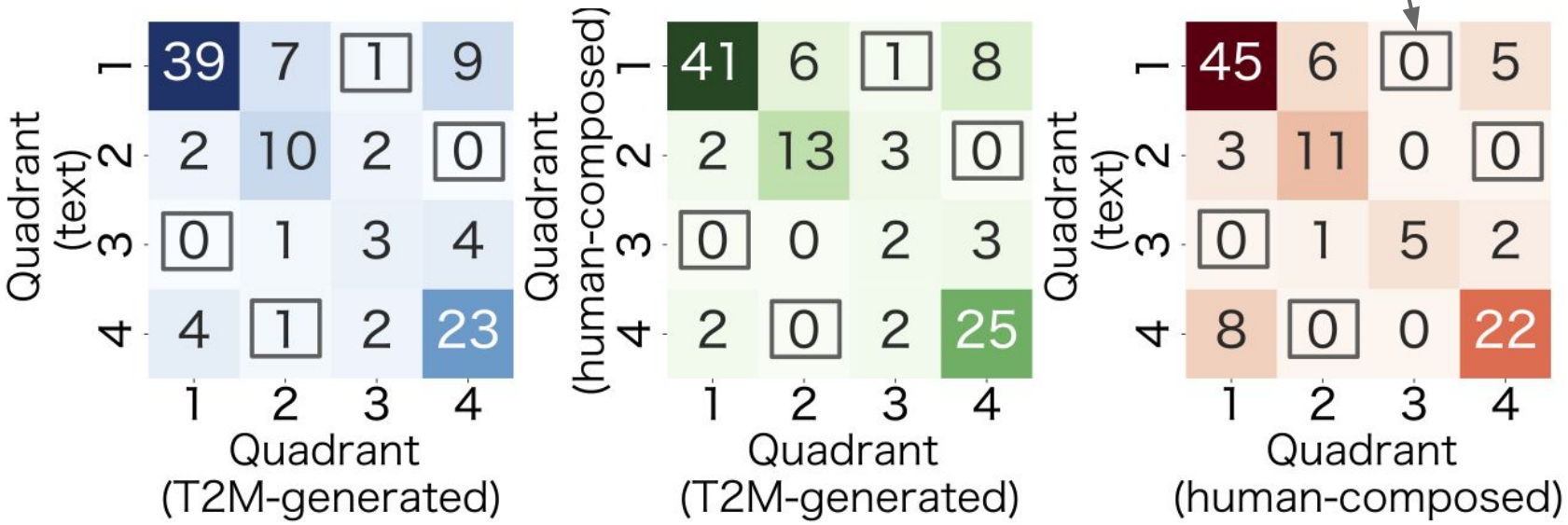
*2: https://www.shutterstock.com/ja/music

分析内容

- テキスト，{人間作曲, T2M合成} 音楽の象限間での差異はあるのか.
- テキスト，{人間作曲, T2M合成} 音楽の感情座標分布の差異はあるのか.
  - 座標を横軸 (x座標)と縦軸 (y座標)に分けて，それぞれで分析
- 対データにおけるテキスト座標と {人間作曲, T2M合成} 音楽座標のズレに傾向はあるのか.
- 聴取者属性は感情座標分布に影響するか.
  - 例) 性別属性: 男性 vs 女性 vs 男女混合
  - 性別，年齢，好きなジャンル，楽器演奏経験，経験楽器数，演奏経験年数
  - 本発表ではT2M合成音楽に限定
    - その他は論文参照

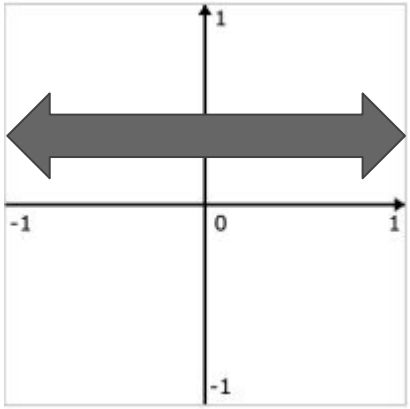
象限間における知覚感情の混同はほとんど存在しない．

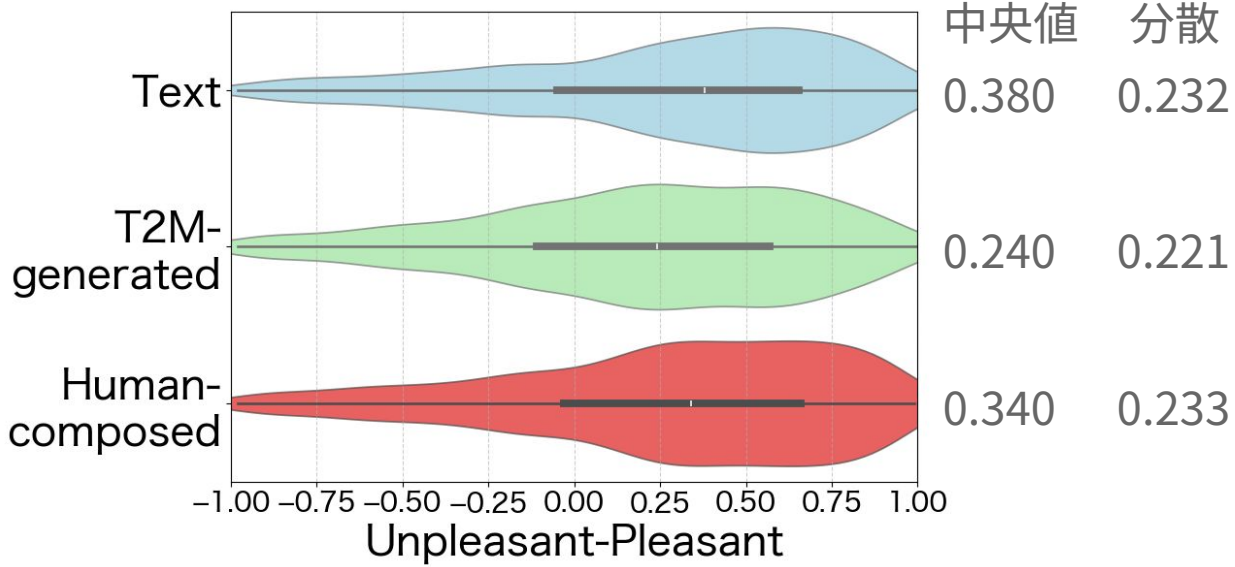
- テキストは第一象限と評価されたが，人間作曲音楽は第三象限と評価される，など．
  - 最も大きい区分での評価では，知覚感情はある程度整合する．



対データ間での感情座標分布 (快 - 不快軸; 横軸) の差異

- 等分散性[*3]: テキスト ≈ T2M合成音楽 ≈ 人間作曲音楽
  - T2M合成音楽が喚起する感情は他の2つと同程度に多様である.
- 中央値[*4]: T2M合成音楽 < {人間作曲音楽, テキスト}
  - T2M合成音楽は他の2つよりも平静寄りの感情を喚起する.



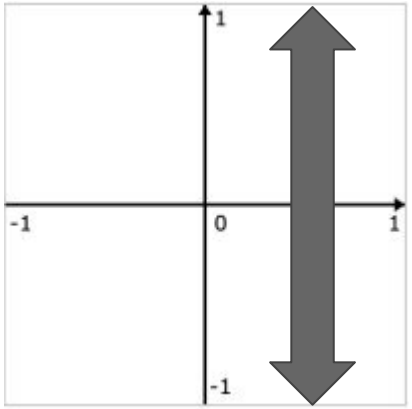


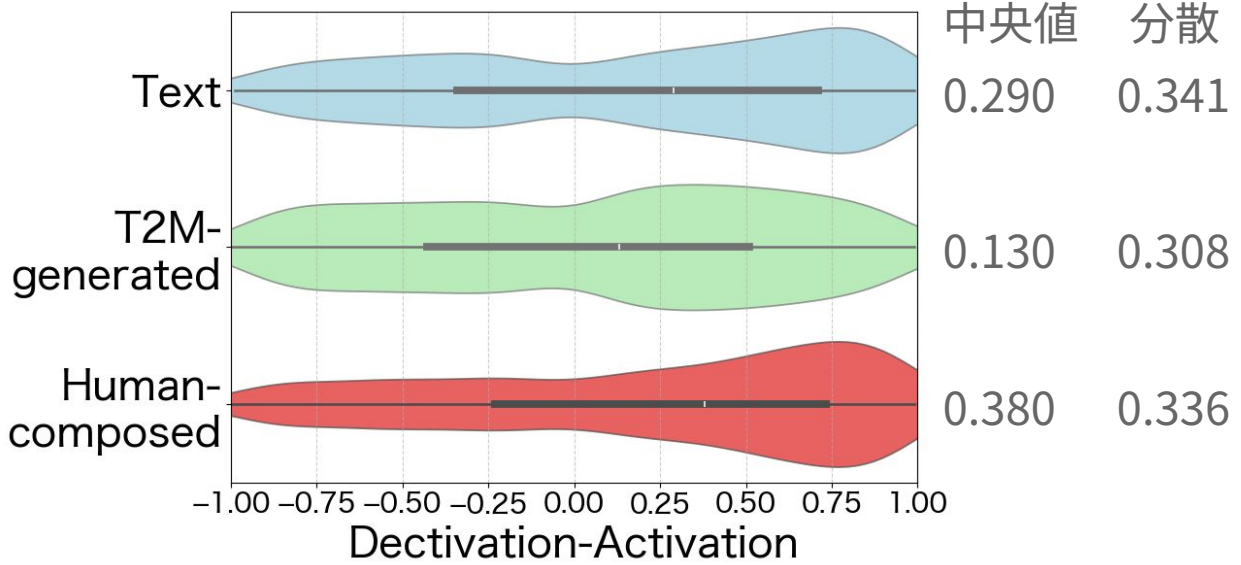
	中央値	分散
Text	0.380	0.232
T2M-generated	0.240	0.221
Human-composed	0.340	0.233

*3: Brown-Forsythe検定
*4: Friedman検定

対データ間での感情座標分布 (覚醒 - 沈静軸; 縦軸) の差異

- 等分散性: T2M合成音楽 ＜ テキスト
  - T2M合成音楽はテキストほどの喚起する感情の多様さが無い.
- 中央値: T2M合成音楽 ＜ テキスト ＜ 人間作曲音楽
  - テキストに比べ，T2M合成音楽は沈静に，人間作曲音楽は覚醒に寄る.

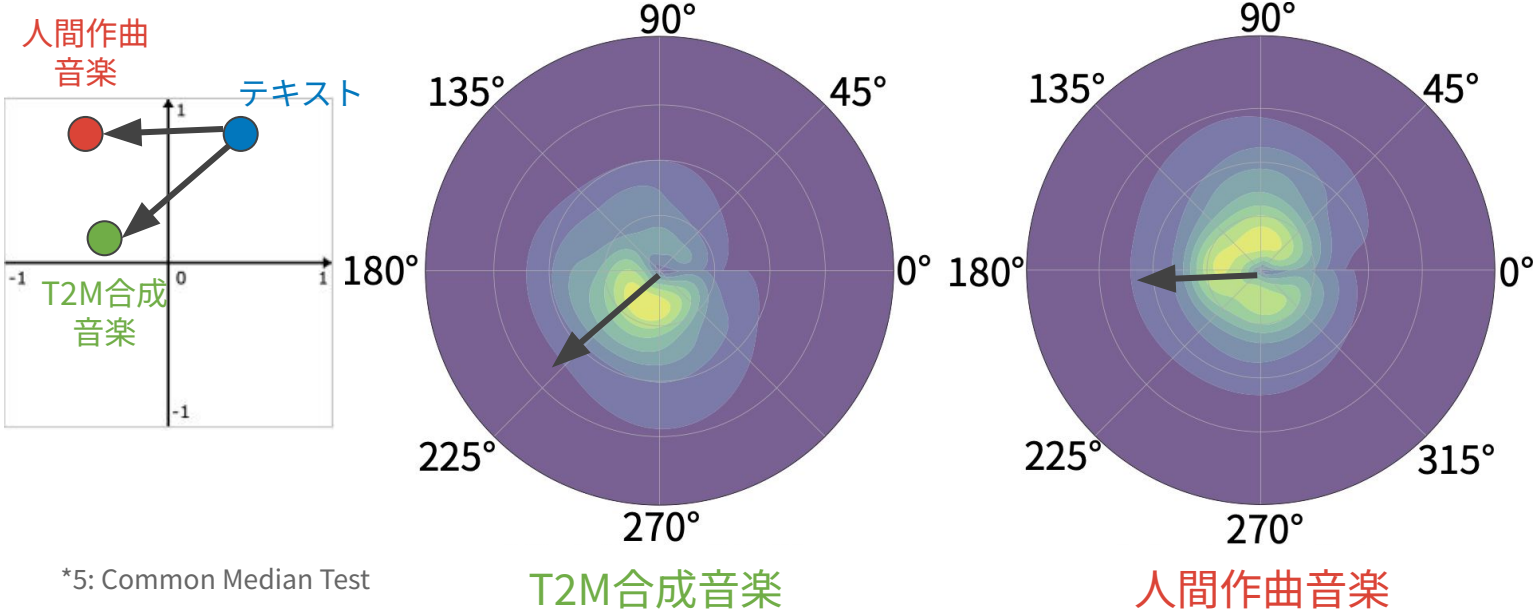




	中央値	分散
Text	0.290	0.341
T2M-generated	0.130	0.308
Human-composed	0.380	0.336

テキストを基準とした2種の音楽の評価の違い

- 2種の音楽データの中央値は有意に異なる[*5].
  - T2M合成音楽はテキストよりも不快・沈静向きに評価された.
  - 人間作曲音楽はテキストよりも不快・覚醒向きに評価された.



*5: Common Median Test

T2M合成音楽　　人間作曲音楽

T2M合成音楽評価に対する聴取者属性の影響

- 快 - 不快軸について．
  - 等分散性: 経験有り楽器数が多いほど感情の広がりが狭い．
  - 中央値: 聴取者属性は影響しない．
- 覚醒 - 沈静軸について．
  - 等分散性: 好きなジャンル，楽器演奏経験，経験楽器数属性で有意差あり．
    - 人間作曲音楽でも同様の属性で有意差有り．
    - これらの属性は音楽の生成元に依らず，同様の傾向を持つ．
  - 中央値: 聴取者属性は影響しない．

更なる調査: T2Mのアーキテクチャやパラメタ数は知覚感情評価に影響するのか.

T2Mモデルの違いによる影響

- 使用モデル:
  - MusicGen-Small, MusicGen-Large (Transformer基盤)
  - MusicLDM (拡散モデル基盤)
- 結果:

	等分散性	中央値
快 - 不快軸	モデルの違いは影響しない.	MusicGenでは大規模の方が快寄り. パラメタ数が同程度の場合, LDMの方が快寄り.
覚醒 - 沈静軸	MusicGen: パラメタ数は影響しない. LDM: より多様な表現を持つ.	MusicGen: パラメタ数は影響しない.LDM: より沈静の感情を喚起させる.

まとめ

まとめ

- 目的:
  - 人間作曲音楽，テキスト，T2M合成音楽の知覚感情に差異があるのか.
- 提案手法:
  - 聴取者を募り，上記の3つの対データに対して同一聴取者に評価させた.
- 評価結果:
  - T2M 合成音楽が喚起する感情は，
    - 快 - 不快軸においてテキスト，人間作曲音楽と同程度の広がりを持つ.
    - 快 - 不快軸と覚醒 - 沈静軸の両方でその中央値が原点方向に縮退する.
  - 楽器演奏経験などの聴取者属性が，知覚感情の広がりに影響を与える.
- 今後の予定:
  - 人間の知覚感情に基づいたT2Mモデルの開発
  - ユーザ特性に基づいてパーソナライズされたT2Mモデルの開発

発表終わり

モデルの違い

- 横軸 (不快 - 快)
  - 分散について
    - T2Mモデルの違いによる影響はない．
  - 中央値について
    - MusicGenでは大規模の方が快寄りの感情を喚起．
      同数程度のパラメタの場合，LDMの方がMusicGenよりも快寄り．
- 縦軸 (沈静 - 覚醒)
  - 分散について
    - Transformerはパラメタ数の影響を受けない．
    - パラメタ数に関わらず，LDMの方がMusicGenよりも多様な表現を持つ．
  - 中央値について
    - Transformerはパラメタ数の影響を受けない．
    - パラメタ数に関わらず，LDMの方がMusiGenよりも沈静よりの感情を喚起させる．

質疑応答ありそうなの

- 属性内要素によって対象人数が異なる．それの影響が大きいのでは？
  - どう返答する？
- 曲は一曲だけなの？合成音楽の話
- 音楽の中身に依存する部分が多いのでは？
  - ただの反復している音ではないか？音楽性は保証できているのか
- 秒数的に音楽と言えるのか？
  - 15秒は短くない？
- 評価セットの音楽はラッセルから均等にとったよ，と言う話は付録でいいもの？

発表内容: 目次

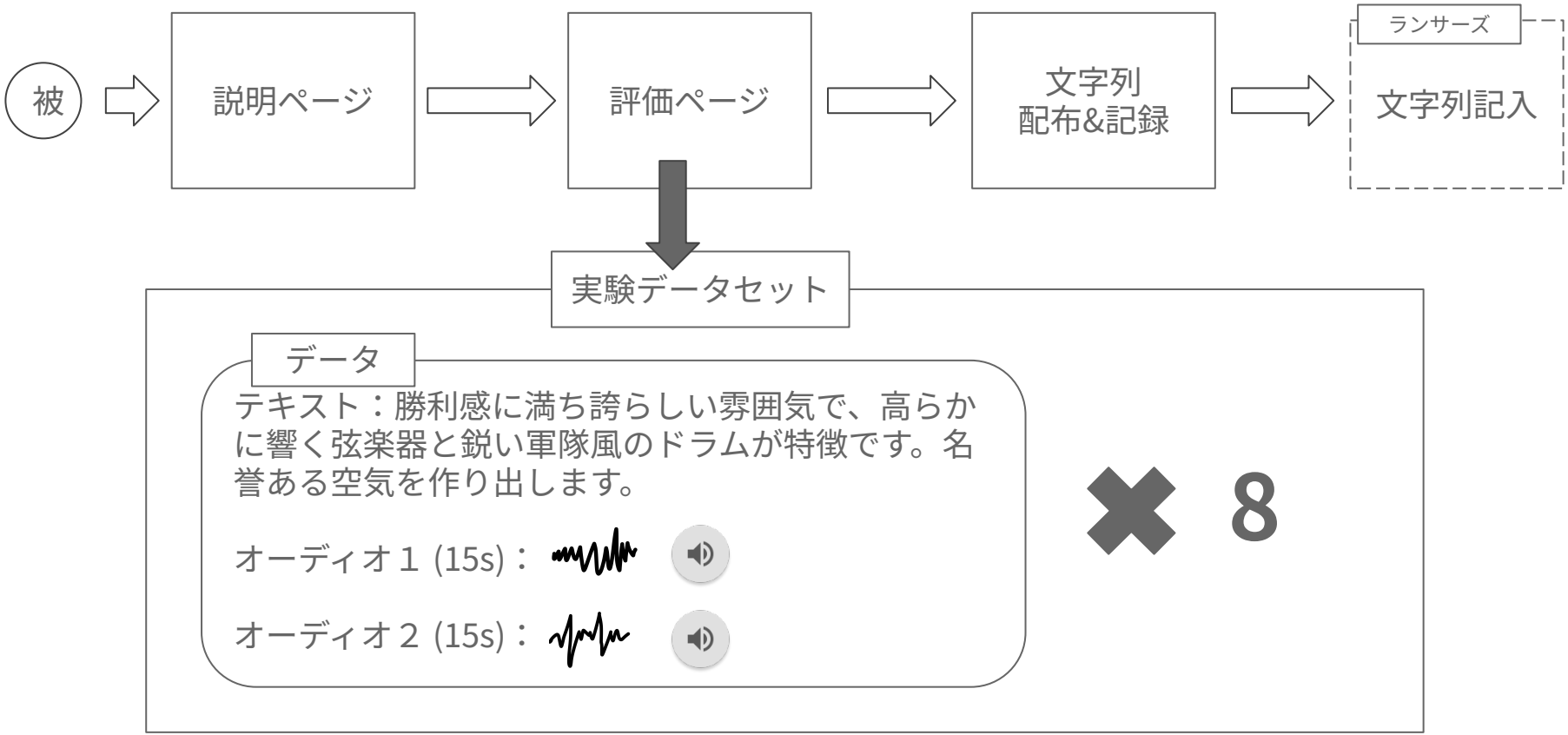
- 研究背景
  - T2M (text-to-music) モデルとは．
  - T2Mモデルの評価方法
  - 本研究の立ち位置
- 提案手法
  - 感情へのアプローチ．ラッセルの円環モデルについて．
  - 評価手法，評価ページについて．
- 実験結果
  - 実験条件
  - 対データ間での分析
  - テキストを基準にした際の2種の音楽評価の違い
  - 聴取者属性が音楽評価に与える影響
- まとめ

付録

モデルの違い

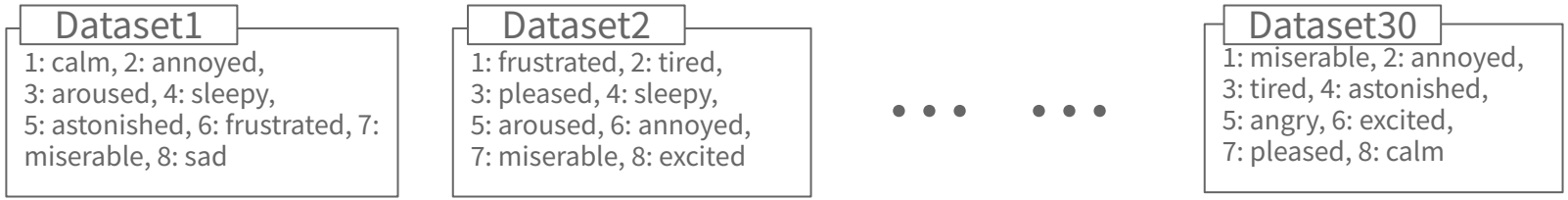
- 横軸 (不快 - 快)
  - 分散について
    - T2Mモデルの違いによる影響はない．
    - 人間作曲音楽に比べて表現の多様性はない．
  - 中央値について
    - モデルの違いによる影響あり．GenではLの方がSより快寄りの感情を喚起．同数程度のパラメタの場合，LDMの方がTransformerベースよりも快寄り．
    - LDMは人間作曲音楽と同傾向．
- 縦軸 (沈静 - 覚醒)
  - 分散について
    - レイヤーの層数の違いに影響はされないが，パラメタ数に関わらず，LDMの方がTransformerベースよりも多様な表現を持つ．
    - LDMが生成する音楽は人間作曲音楽よりも多様に喚起させる．
  - 中央値について
    - レイヤーの層数の違いに影響はされないが，パラメタ数に関わらず，LDMの方がTransformerベースよりも沈静よりの感情を喚起させる．

実験手順について


被
説明ページ
評価ページ
文字列
配布&記録
ランサーズ
文字列記入
実験データセット
データ
テキスト：勝利感に満ち誇らしい雰囲気で、高らかに響く弦楽器と鋭い軍隊風のドラムが特徴です。名誉ある空気を作り出します。
オーディオ１ (15s)：
オーディオ２ (15s)：
× 8

実験データセットの制作

- 制作
  - 実験データセットは8つのデータからなる
    - 12個の感情タグから重複が無いよう8個の感情タグを選択
      - 各象限からそれぞれ2個ずつ選択
  - 実験データセットは30個用意



- 振り分け
  - 30個のデータセットからランダムで1つ選ばれる
    - データの順番はランダム
    - オーディオの順番もランダム

リサーチクエスチョンの内容 (ランサーズで)

RQ	内容
1 (選択)	性別を教えてください (男性 or 女性 or 回答しない)
2 (選択)	年齢を教えてください (19歳以下，10-70は10歳ごと，70歳以上)
3 (記述)	過ごした時間が一番長い国を教えてください
4 (選択)	一番好きな音楽のジャンルを教えてください (10ジャンル提示)
5 (選択)	今の調子を教えてください (良 or 普通 or悪)
6 (選択)	何時ごろに実験を行ったかを教えてください (24時間を3時間ごとに分割)
7 (記述)	どこで実験を行ったかを教えてください
8 (選択)	楽器演奏経験の有無を教えてください (有 or 無)
9 (選択)	演奏経験のある楽器の種類を教えてください (回答無しを含めて14種類提示)
10 (選択)	楽器演奏の経験年数 (合計)を教えてください (回答無しを含めて6種類提示)
11 (記述)	実験ページの最後に出力された文字列を記入してください

ここからは削除でいい

音楽評価における聴取者属性の違い

- 快 - 不快軸について
  - 等分散性: 経験楽器数属性で有意差が認められた
    - T2M合成音楽・人間作曲音楽関係なく，経験のある楽器数が増えるほど音楽から知覚される感情の広がりが狭い
  - 中央値: T2M合成音楽・人間作曲音楽ともに有意差がある属性はなかった
    - 人間作曲音楽に対して，好きなジャンル属性で有意差があったのみ
    - 聴取者属性は音楽から喚起される快 - 不快軸における感情に特定の傾向を生まない
- 覚醒 - 沈静軸について
  - 等分散性: 好きなジャンル，楽器経験の有無，経験楽器数属性で有意差
    - ポップを好むをグループは他のジャンル (ロックやクラシック)のグループよりも喚起される感情に広がりがある
    - 経験のある楽器数が増えるほど音楽から知覚される感情の広がりが狭い (楽器経験のないグループが一番広がりが広い)

T2M合成音楽評価に対する聴取者属性の影響

- 快 - 不快軸について
  - 等分散性: 経験楽器数属性で有意差あり
    - 経験のある楽器数が多いほど音楽から喚起される感情の広がりが狭い
  - 中央値: 全属性で有意差なし
    - 聴取者属性は快 - 不快軸における知覚感情の傾向に影響しない
- 覚醒 - 沈静軸について
  - 等分散性: 好きなジャンル，楽器演奏経験，経験楽器数属性で有意差あり
    - ポップを好むグループは他のジャンル (ロックやクラシック)のグループと比べて，喚起される感情の広がりが広い
    - 経験のある楽器数が多いほど音楽から喚起される感情の広がりが狭い (楽器経験のないグループが一番広がりが広い)
  - 中央値: 全属性で有意差なし
    - 聴取者属性は覚醒 - 沈静軸における知覚感情の傾向に影響しない

	等分散性	ペワイズ	結論	中央値	ペアワイズ	結論
快 - 不快	○	無し(斜め線	T2M合成音楽が喚起する感情はテキストや人間作曲音楽が喚起する感情と同程度に多様である.	❌	T2M合成音楽 - テキスト T2M合成音楽 - 人間作曲音楽	T2M合成音楽はテキストや人間作曲音楽よりも平静寄りの感情を喚起する.
覚醒 - 沈静	❌	T2M合成音楽 - テキスト	T2M合成音楽が喚起する感情はテキストが喚起する感情ほど表現の多様さがない	❌	T2M合成音楽 - テキスト T2M合成音楽 - 人間作曲音楽	

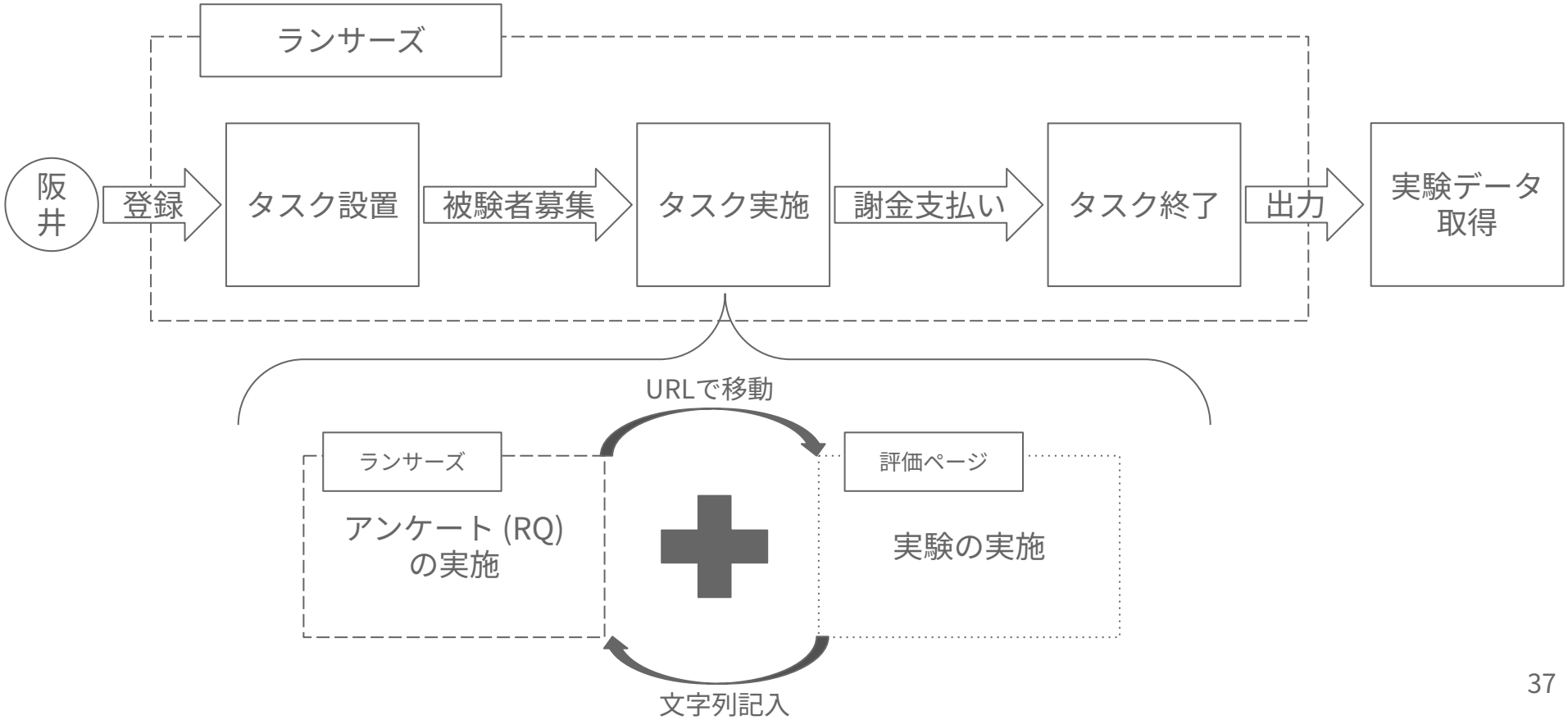
	快 - 不快軸	覚醒 - 沈静軸
等分散性		
ペアワイズ		
結論	T2M合成音楽が喚起する感情はテキストや人間作曲音楽が喚起する感情と同程度に多様である	T2M合成音楽が喚起させる感情はテキストが喚起する感情ほど表現の多様さがない
中央値		
ペアワイズ	T2M合成音楽 - テキスト T2M合成音楽 - 人間作曲音楽	
結論	T2M合成音楽はテキストや人間作曲音楽よりも平静寄りの感情を喚起する	テキストに比べ，T2M合成音楽は沈静寄りに，人間作曲音楽は覚醒寄りに表現している.

本研究の立ち位置：概要と先行研究

- 概要：T2Mにおけるテキスト-音楽間，また人間による音楽 (既存データセットから抽出)と生成AIによる音楽に対するリスナーの感情の違いの解析
  - 感情へのアプローチ：ラッセルの円環モデルを使用 [James80]

- 先行研究
  - 円環モデルを用いた生成AIによる音楽の評価と真値の感情タグの比較[Imasato23]
  - 生成AIによる音楽とデータセット音楽の比較 (フィードバック式) [Xuneng24]
  - 年齢や音楽知識 (楽器演奏できる or できない)は感情認識の違いに繋がる．文化的背景は感情認識の一致に繋がる (同じ文化を持つ人々は同じ感情認識になる) [Harin21, En22]

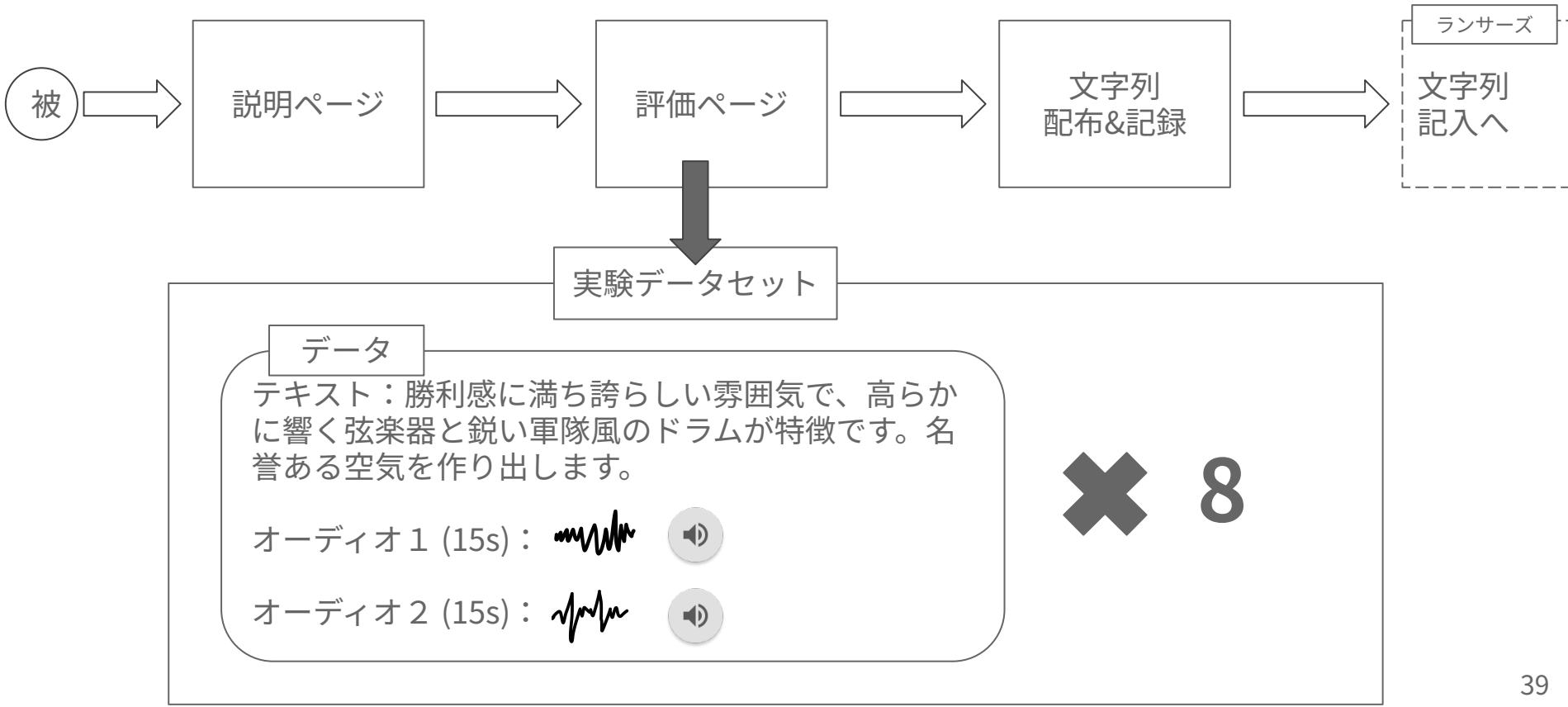
実験の流れ


ランサーズ
阪井
登録
タスク設置
被験者募集
タスク実施
謝金支払い
タスク終了
出力
実験データ取得
URLで移動
ランサーズ
アンケート (RQ) の実施
評価ページ
実験の実施
文字列記入

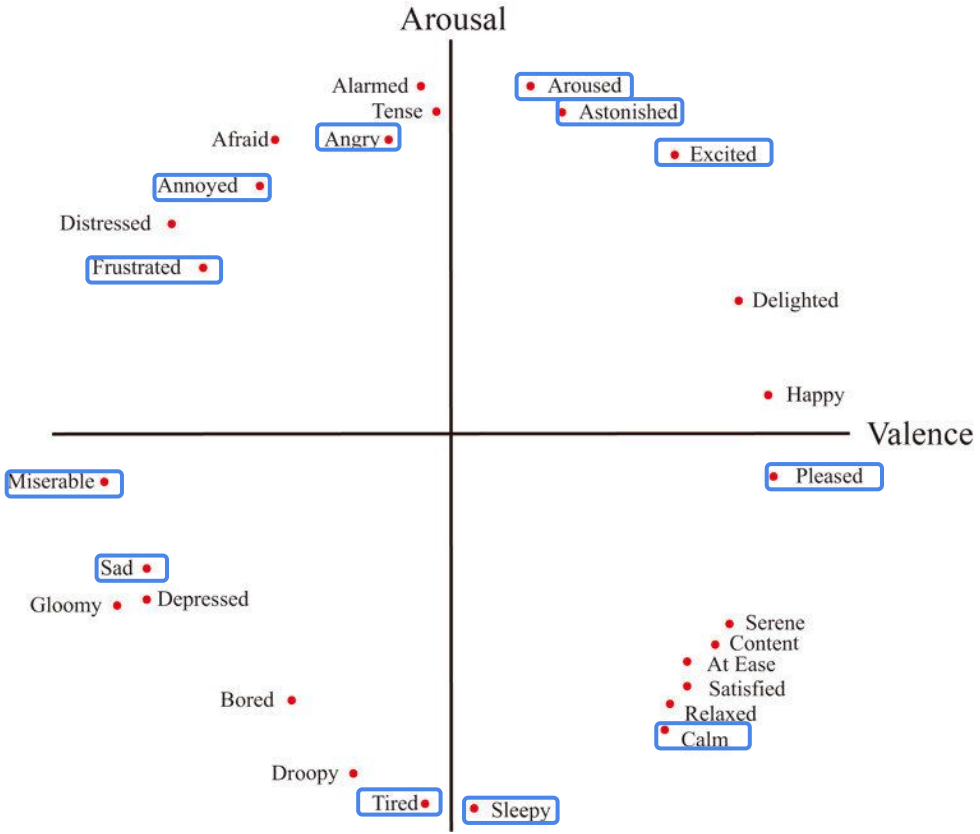
リサーチクエスチョンの内容 (ランサーズで)

RQ	内容
1 (選択)	性別を教えてください (男性 or 女性 or 回答しない)
2 (選択)	年齢を教えてください (19歳以下，10-70は10歳ごと，70歳以上)
3 (記述)	過ごした時間が一番長い国を教えてください
4 (選択)	一番好きな音楽のジャンルを教えてください (10ジャンル提示)
5 (選択)	今の調子を教えてください (良 or 普通 or悪)
6 (選択)	何時ごろに実験を行ったかを教えてください (24時間を3時間ごとに分割)
7 (記述)	どこで実験を行ったかを教えてください
8 (選択)	楽器演奏経験の有無を教えてください (有 or 無)
9 (選択)	演奏経験のある楽器の種類を教えてください (回答無しを含めて14種類提示)
10 (選択)	楽器演奏の経験年数 (合計)を教えてください (回答無しを含めて6種類提示)
11 (記述)	実験ページの最後に出力された文字列を記入してください

実験ページについて


被
説明ページ
評価ページ
文字列
配布&記録
ランサーズ
文字列
記入へ
実験データセット
データ
テキスト：勝利感に満ち誇らしい雰囲気で、高らかに響く弦楽器と鋭い軍隊風のドラムが特徴です。名誉ある空気を作り出します。
オーディオ１ (15s)：
オーディオ２ (15s)：
× 8

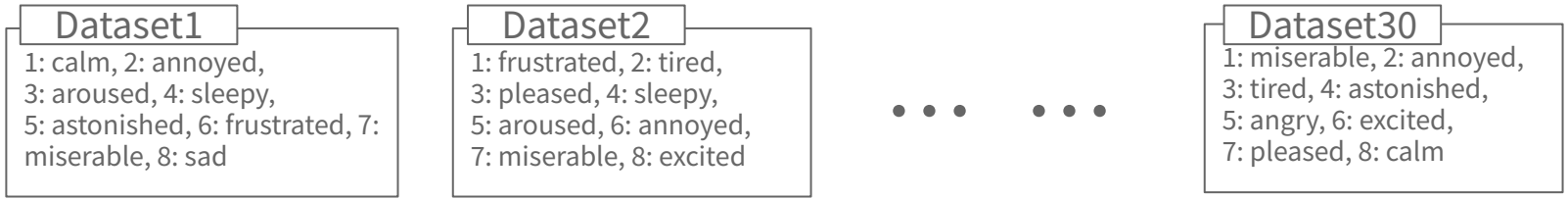
人間による音楽の購入について



1. ラッセルの円環モデル [Russell 80]の二次元座標平面の各象限からそれぞれ3つの感情タグを抽出
2. shutterstockにて感情タグを検索クエリとして検索
3. 人気順の上から10曲を購入
   a. 声が含まれているものをスキップ
4. 必要な情報を保存
   a. 音楽データ
   b. 音楽の説明文 (キャプション)
   c. 音楽のbpm
   d. 音楽の感情タグ

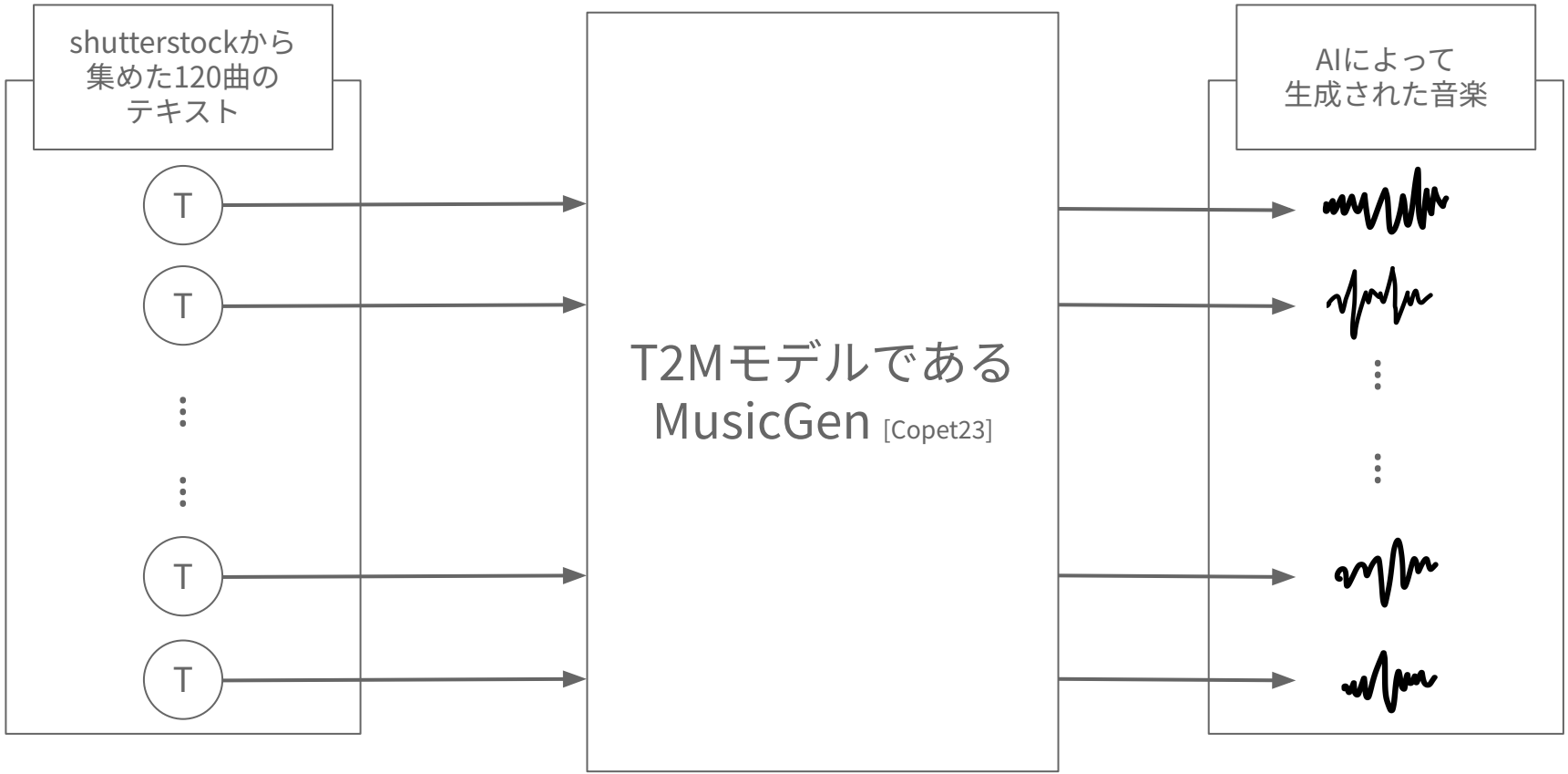
実験データセットの制作

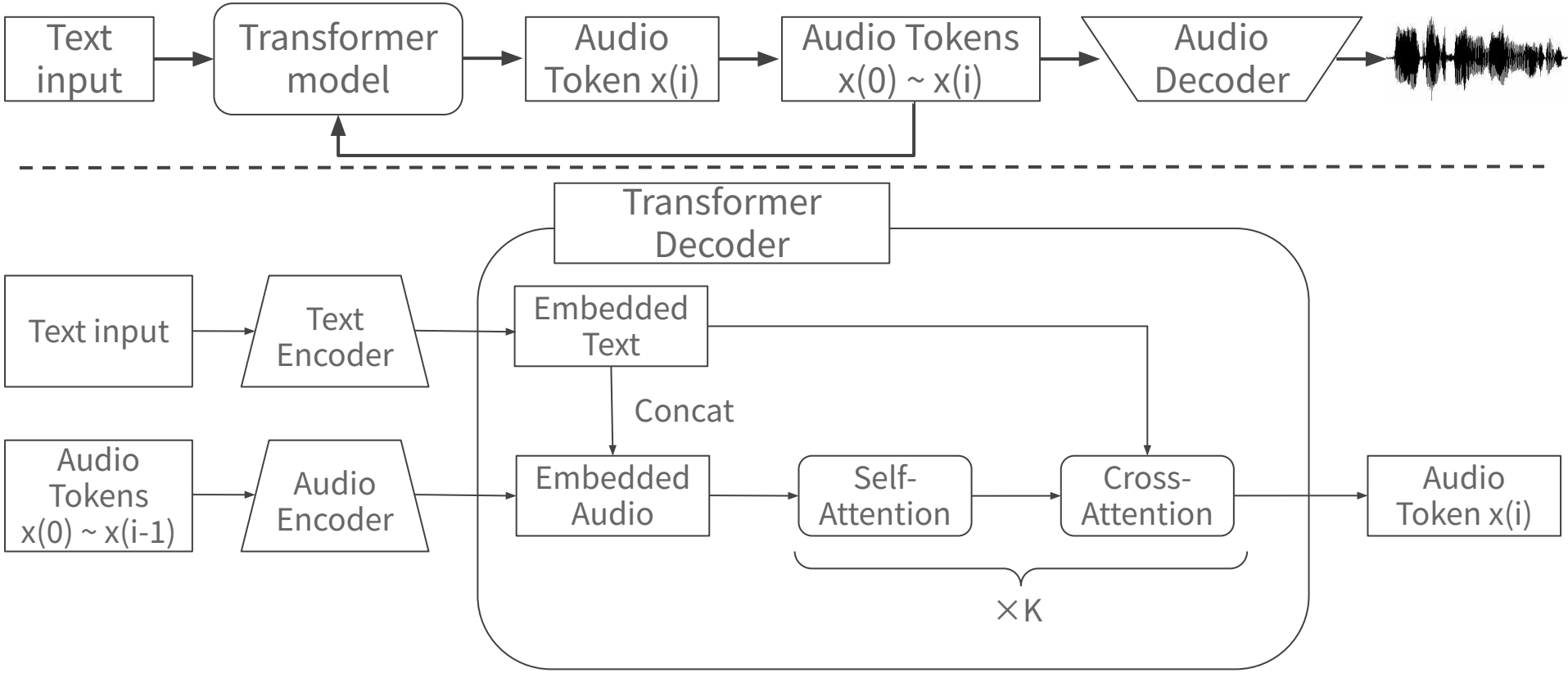
- 制作
  - 実験データセットは8つのデータからなる
    - 12個の感情タグから重複が無いよう8個の感情タグを選択
      - 各象限からそれぞれ2個ずつ選択
  - 実験データセットは30個用意



- 振り分け
  - 30個のデータセットからランダムで1つ選ばれる
    - データの順番はランダム
    - オーディオの順番もランダム

AIによる音楽の生成


shutterstockから
集めた120曲の
テキスト
T
T
T
T
T2Mモデルである
MusicGen [Copet23]
AIによって
生成された音楽


Text input
Transformer model
Audio Token x(i)
Audio Tokens x(0) ~ x(i)
Audio Decoder
Transformer Decoder
Text input
Text Encoder
Embedded Text
Concat
Audio Tokens x(0) ~ x(i-1)
Audio Encoder
Embedded Audio
Self-Attention
Cross-Attention
Audio Token x(i)
×K