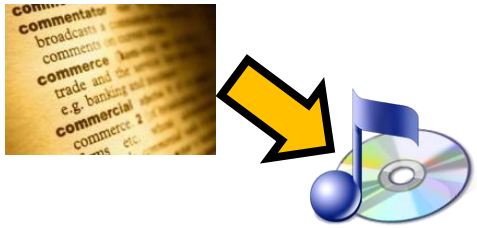
Building and Combining Document and Music Spaces for Music Query-By-Webpage System



Ryoei Takahashi, ***Yasunori Ohishi***
Norihide Kitaoka, Kazuya Takeda

Graduate School of Information Science,
Nagoya University, Japan

New music retrieval system

- Music Query-By-Webpage System
  - Songs that appropriately match Webpage automatically selected

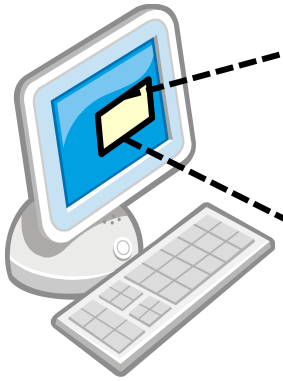
New music retrieval system

- Music Query-By-Webpage System

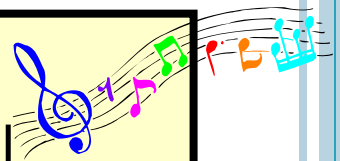
h Webpage

- Text analysis
- Ranking

Play list



Meja	「Rainbow」
The Corrs	「Breathless」
Nick Lowe	「Cruel to be kind」
Faye Wong	「Dreams」
Tom Jones	「It's not unusual」



New music retrieval system

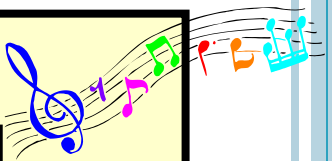
- Music Query-By-Webpage System

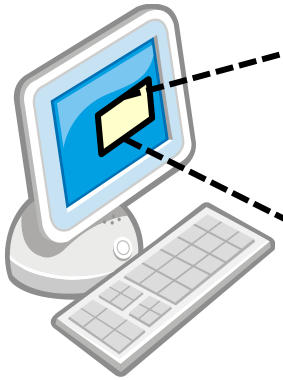
h Webpage

- Text analysis
- Ranking

Play list

Meja 「Rainbow」
The Corrs 「Breathless」





Songs associated with words in documents (Webpages)

Previous work

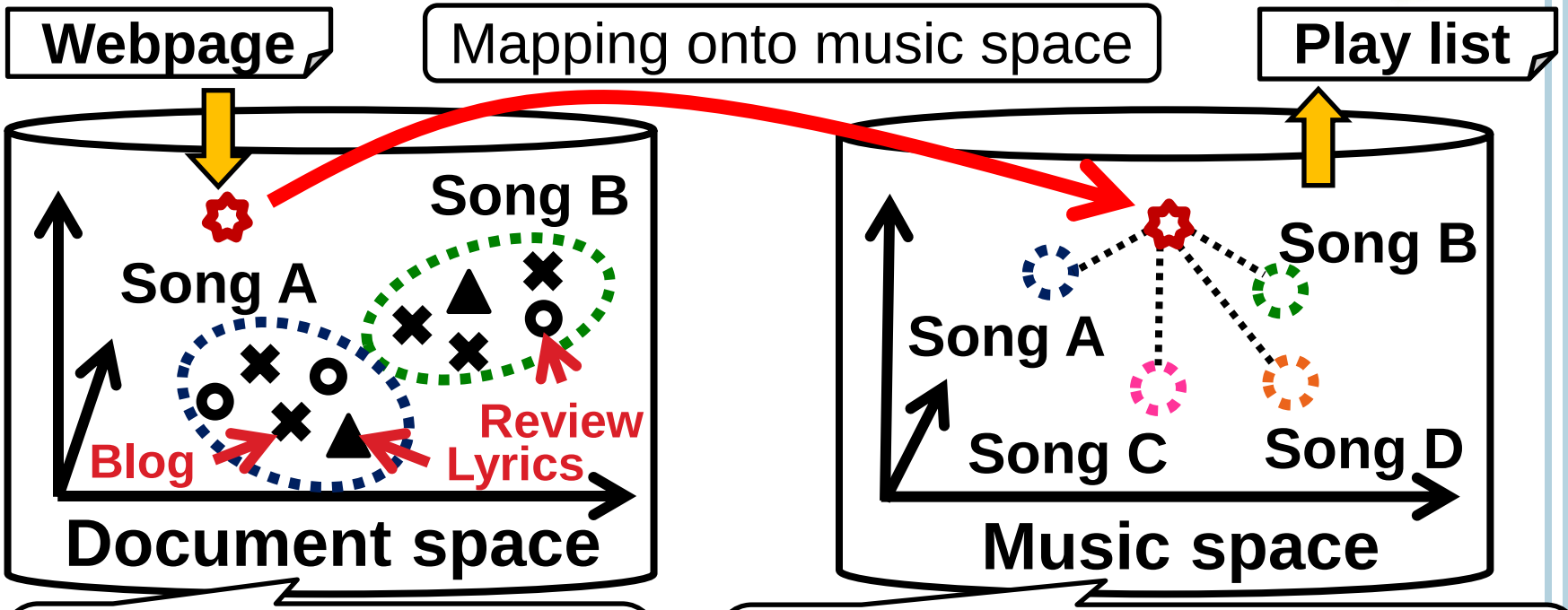
- Building similarity measures between songs
  - Query-by-keyword systems
    - Word similarities defined by titles, artist names ...
  - Song classification tasks
    - Acoustic similarities between songs
  - Music annotation systems
    - Words expressing impressions and acoustic cues [Kumamoto *et al.*]
    - Musically informative words and acoustic cues [Whitman *et al*., Turnbull *et al*.]

But

Individual correspondences between songs and words

Proposed method

- Two different vector spaces



Latent semantic space
defined by documents related to songs

Signal acoustic space
defined by song's spectral/temporal characteristics

Document space $\{\mathrm{d}\}$ and music space $\{\mathrm{a}\}$ associated by linear transformation $\mathrm{a} = W\mathrm{d}$

Basic algorithm (1)

- Building document space $\{\mathbf{d}\}$
  - Term frequency-inverse document frequency (TF-IDF)
    - TF-IDF weight matrix X $(I \times J)$ $(X \equiv [\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_J])$

$$X_{i,j} = \frac{tf_{ij}}{\sum_j tf_{ij}} \times \log \frac{J}{df_i}$$

tf : Term (word) frequency

df : Document frequency

      - I : Number of words, J : Numbers of songs

Document for Song *j*

Up-tempo rock tune filled with riffs and pop sensibility

Document vector $\mathbf{x}'_j$

"tempo"	"rock"	"techno"	...	"guitar"	"riff"
[0.08	0.3	0	...	0.7	0.2]'

  - Singular value decomposition

$$X = USV^{\top}$$

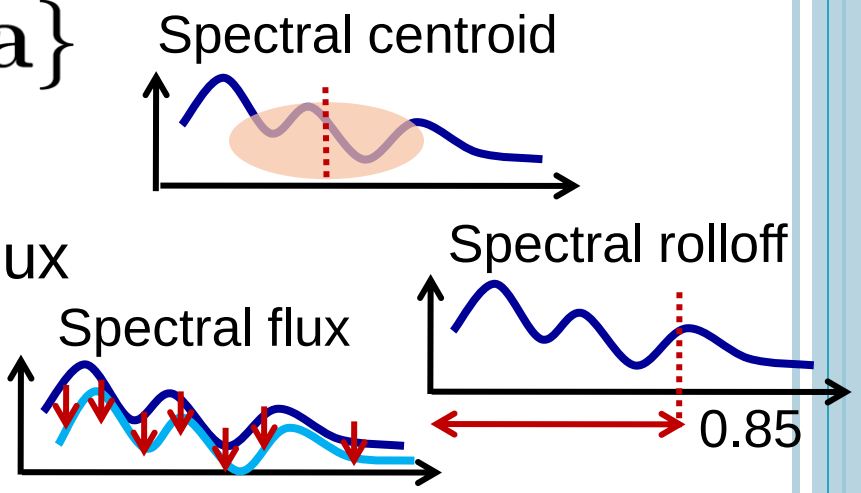
    - Reduced dimensions of document vector $\mathbf{x}$

$$\mathbf{d} = U_N^{\top}\mathbf{x}$$

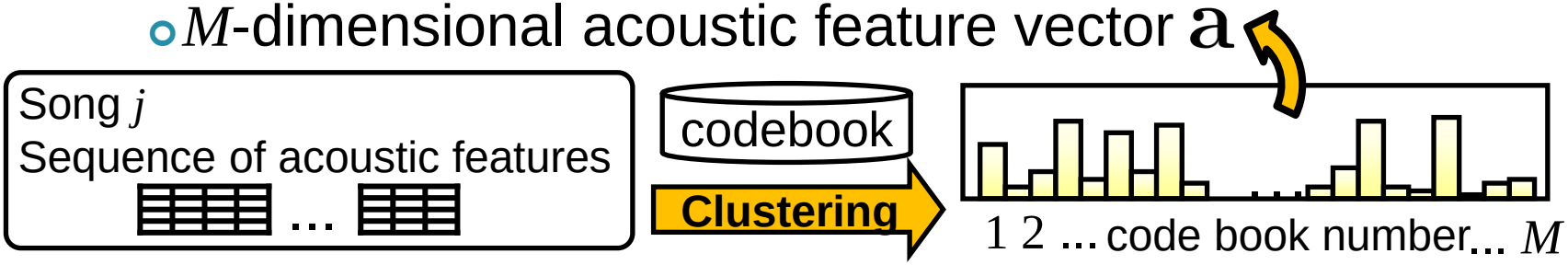
U_N : 1st to Nth columns of U

Basic algorithm (2)

- Building music space $\{\mathbf{a}\}$
  - Acoustic characteristics
    - Spectral centroid, rolloff, flux
    - Zero-crossing rate
    - Energy, rhythm



  - Vector quantization (VQ) codebook (Size M)
    - Feature vectors of training data
    - LBG Algorithm
  - Normalized code histogram of VQ results
    - M-dimensional acoustic feature vector $\mathbf{a}$



Basic algorithm (3)

- Associating document and acoustic vectors through linear transformation

$$\hat{\mathbf{a}} = W\mathbf{d}$$

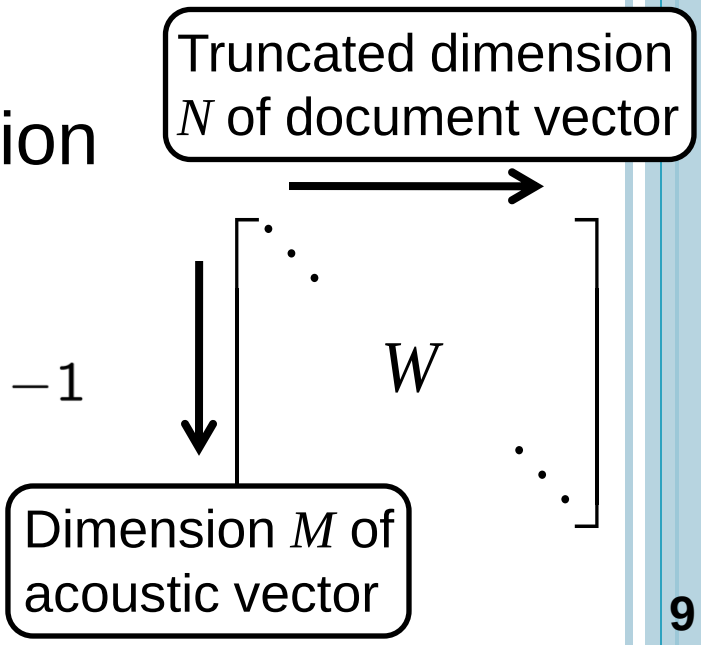
  - Transformation matrix W ($M \times N$) trained using pairs of document and acoustic vectors

    $\{(\mathbf{d}_j, \mathbf{a}_j)\}_{j=1,2,\ldots}$

  - Minimum squared error criterion

$$\hat{W} = \underset{W}{\operatorname{argmin}} \sum_j ||\mathbf{a}_j - W\mathbf{d}_j||^2$$

$$\Rightarrow \hat{W} = \left(\left(\sum_j \mathbf{d}\mathbf{d}^\top \right)^{-1} \sum_j \mathbf{d}\mathbf{a}^\top \right)^{-1}$$



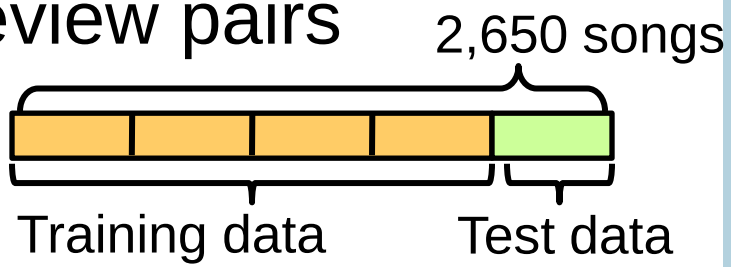
Evaluation of baseline system

- Implemented with 2,650 pop songs

  - Japanese music download site: “ Mora”
    - 30 s previews and reviews
  - Document space
    - Morphological analysis by Chasen ver. 2.3.3
    - Nouns, Adjectives and Verbs (I : 10,578)
  - Music space
    - Applied 32-ms analysis window every 16 ms
    - Calculated acoustic feature vector at each frame
  - Size of transformation matrix W
    - 1,024 (Dimension M of acoustic vector)
      x 1,024 (Truncated dimension N of document)

Experimental setup

- Evaluation under ‘open’ condition
  - Divided 2,650 song and review pairs into five sets



- Evaluation measure
  - Mapped document vector $\mathbf{d}$ of query onto acoustic space $\hat{\mathbf{a}}$ through transformation W
  - Generated rank-ordered song list based on distance between $\hat{\mathbf{a}}$ and $\{\mathbf{a}_j\}_{j=1,2,\ldots}$
  - Mean reciprocal rank (MRR)

$$\text{MRR} = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{r_k}$$

N : number of test samples

r_k : rank order of kth song for which review was given

Evaluation results

- MRR of 0.21 obtained
- Comparison with previous results
  - Music query-by-text system based on **Naive Bayes** [Turnbull *et al*.]
  - 3 times better than previous system, with mean average precision (mAP)

Proposed method	**(open)** **(open)** **(closed)**	**MRR = 0.210** **mAP = 0.351** **mAP = 0.816**
Naive Bayes	**(open)**	**mAP = 0.109**

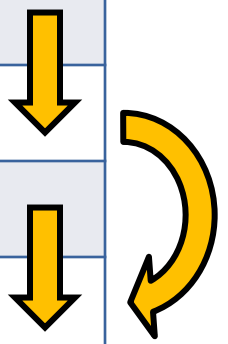
3 times

Results clarified effectiveness of combining document and music spaces

Improving document space

- Using Web text for training
  - Trained document space using Web texts collected from **top 100 Web pages** of Google search results for song title and artist name as query key words
- Using word bigrams
  - Trained document space using word bigrams while considering **word sequence information**

Document vector	Training corpus	MRR
TF-IDF	Review texts	0.210
TF-IDF	Web texts	0.739
Bigrams	Review texts	0.312
Bigrams	Web texts	0.794



Improving document space

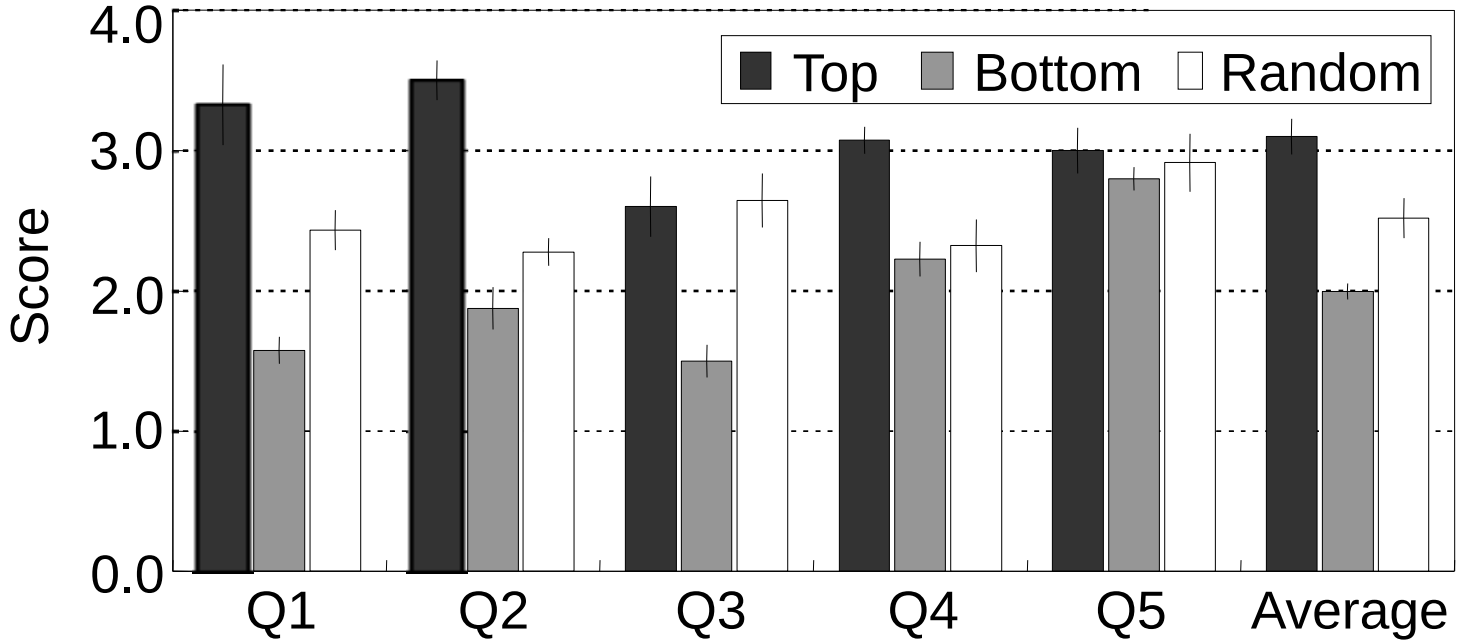
- Using Web text for training
  - Trained document space using Web texts collected from **top 100 Web pages** of Google search results for song title and artist name as query key words
- Using word bigrams
  - Trained document space using word bigrams while considering **word sequence information**

Document vector	Training corpus	MRR
TF-IDF	Review texts	0.210
TF-IDF	Web texts	0.739
Bi		
Bi		

Feasibility of music recommendation system using arbitrary Web texts as input

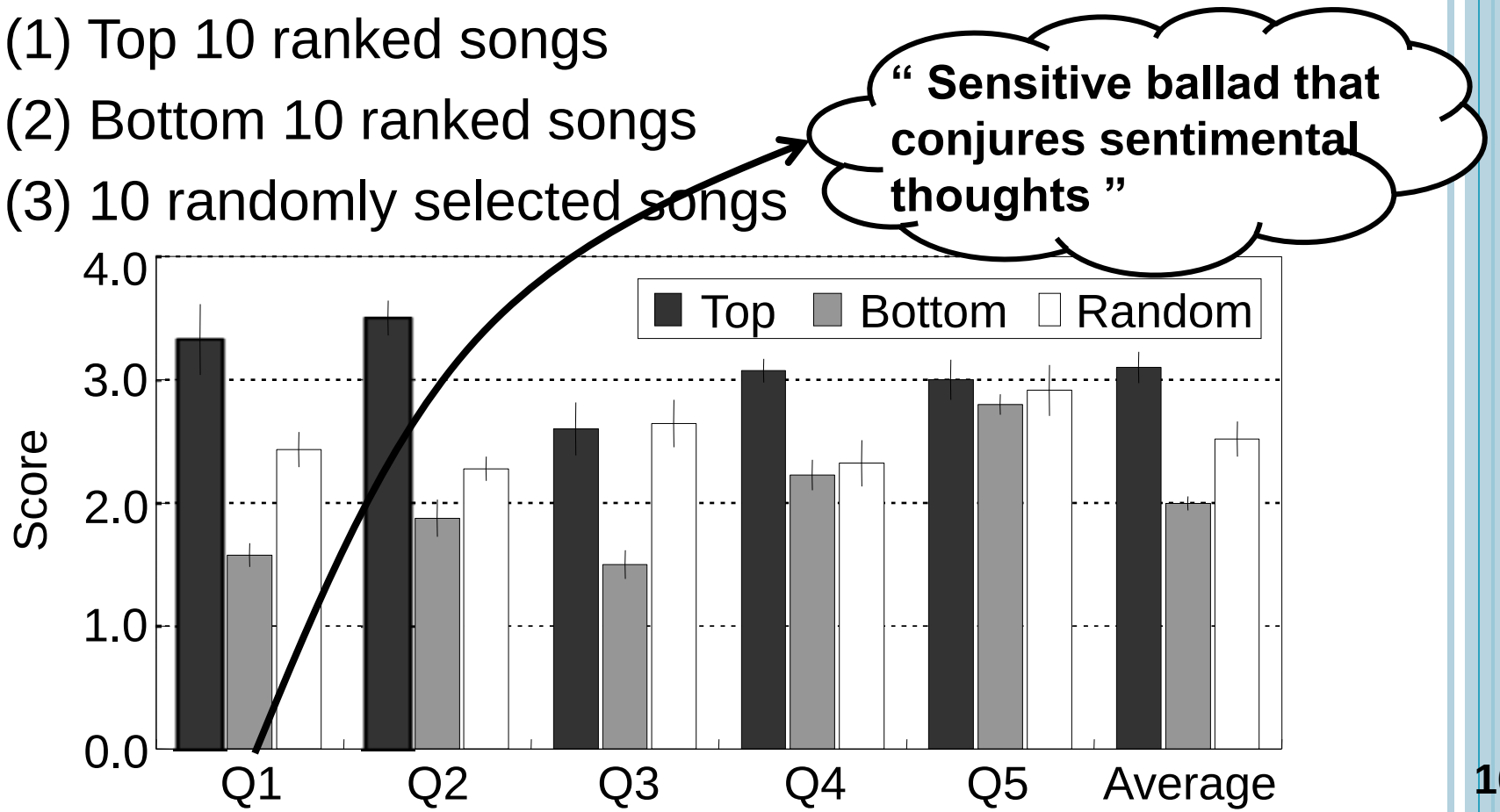
Subjective test

- Four subjects evaluated three sets of songs to determine how appropriately they corresponded to input query sentences on **a scale of 0 to 4**
  - (1) Top 10 ranked songs
  - (2) Bottom 10 ranked songs
  - (3) 10 randomly selected songs



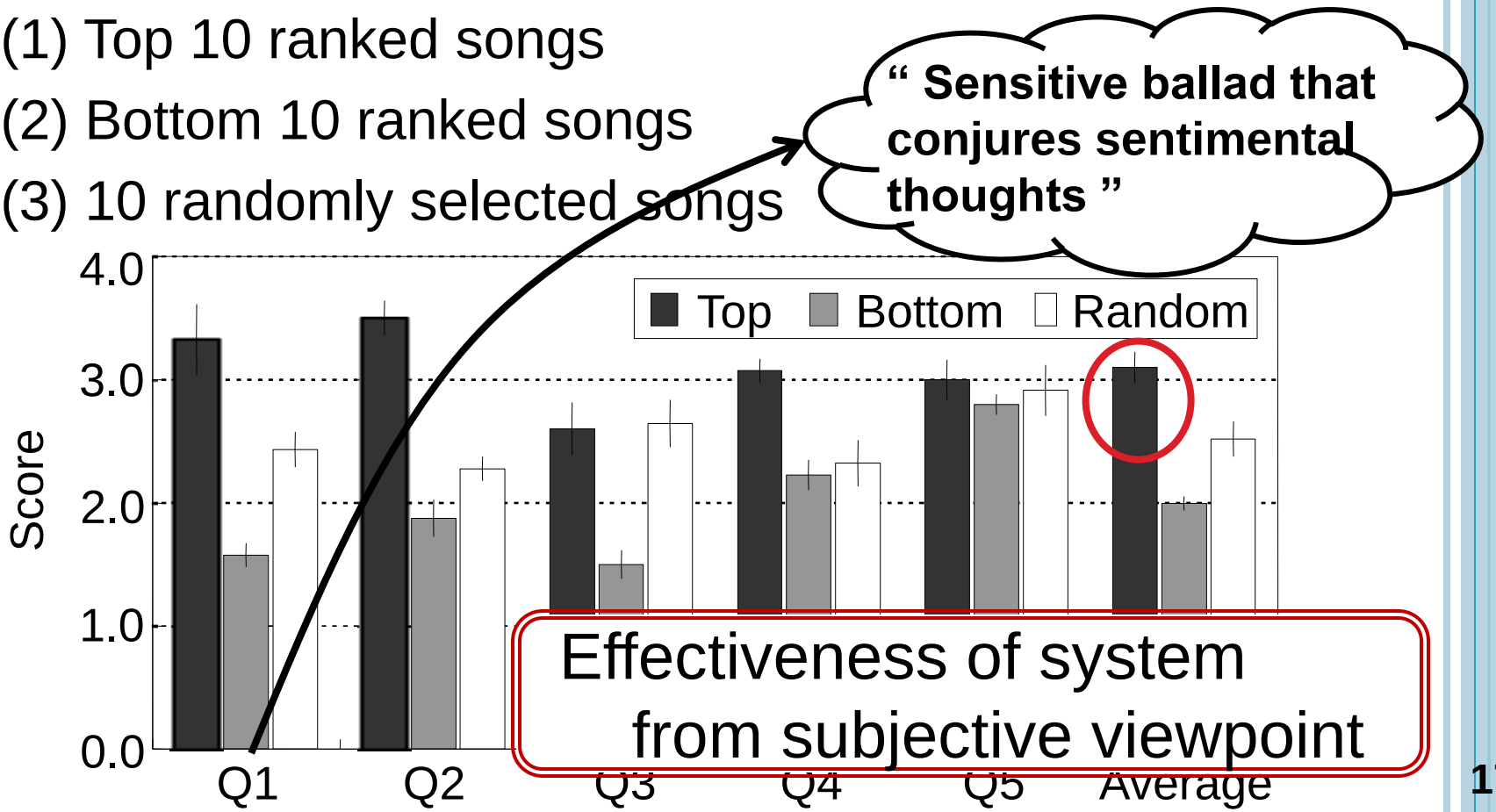
Subjective test

- Four subjects evaluated three sets of songs to determine how appropriately they corresponded to input query sentences on **a scale of 0 to 4**
  - (1) Top 10 ranked songs
  - (2) Bottom 10 ranked songs
  - (3) 10 randomly selected songs



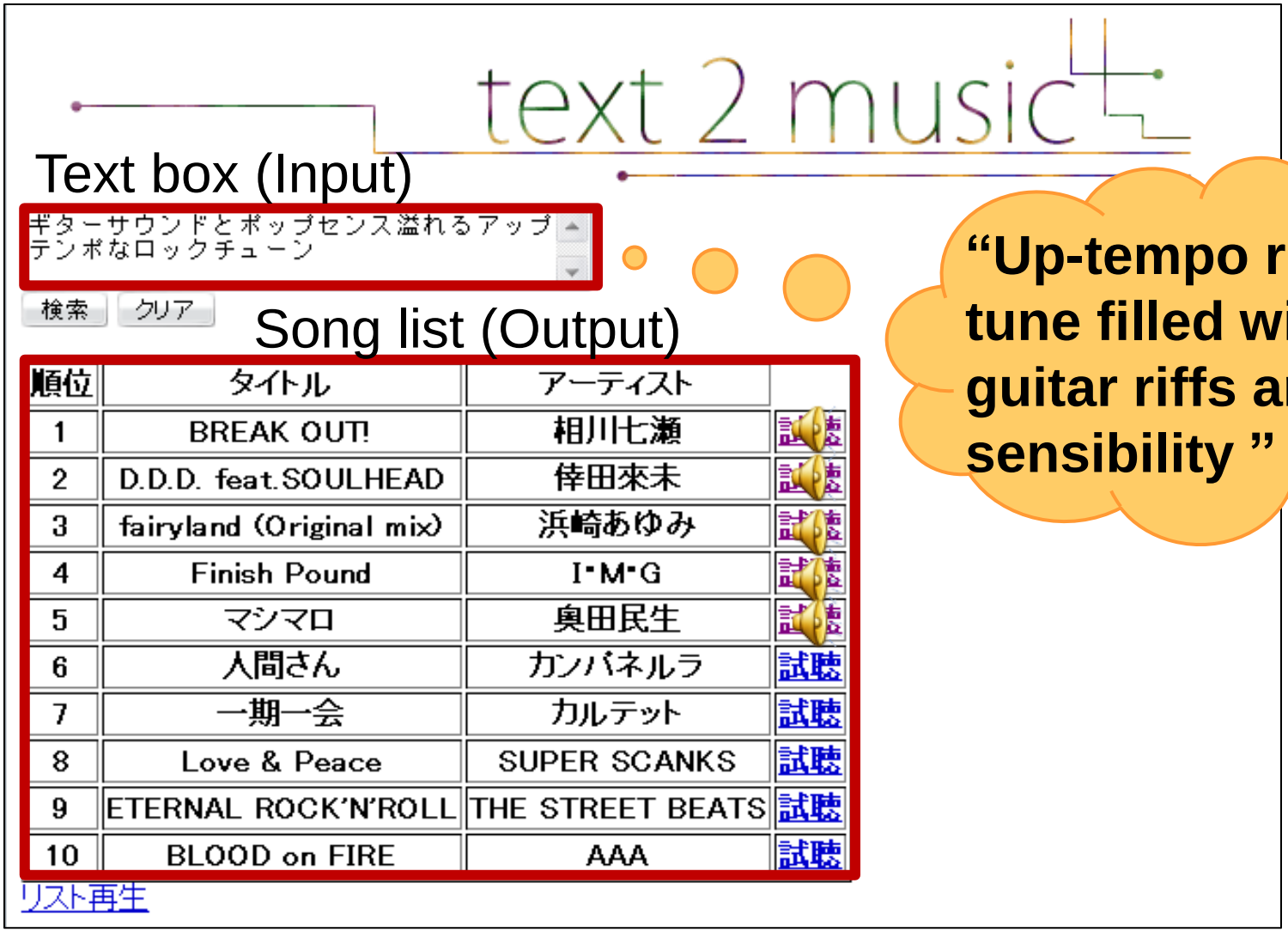
Subjective test

- Four subjects evaluated three sets of songs to determine how appropriately they corresponded to input query sentences on **a scale of 0 to 4**
  - (1) Top 10 ranked songs
  - (2) Bottom 10 ranked songs
  - (3) 10 randomly selected songs

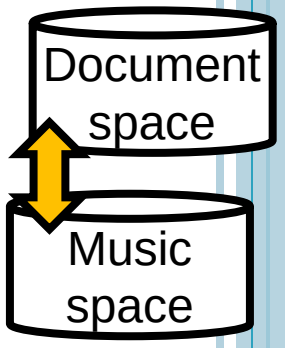


Let's listen

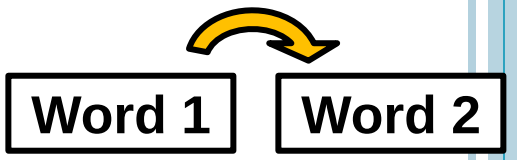
- Demo system " **text 2 music** "



Conclusion



- Built document and music spaces on which “closeness” among songs and texts can be defined and combined
- Implemented music query-by-Webpage system based on combined vector space
  - Proposed system was effective, having a mAP three times higher than previous system
  - Improving document space
    - Use of Web texts as training corpus
    - Use of bigrams as document representation



- Future work
  - Use very large Web documents for training higher order n-grams

Thank you !

Yasunori Ohishi
ohishi@sp.m.is.nagoya-u.ac.jp