The 21st International Society for Music Information Retrieval Conference (ISMIR 2020)

A Chorus-Section Detection Method for Lyrics Text

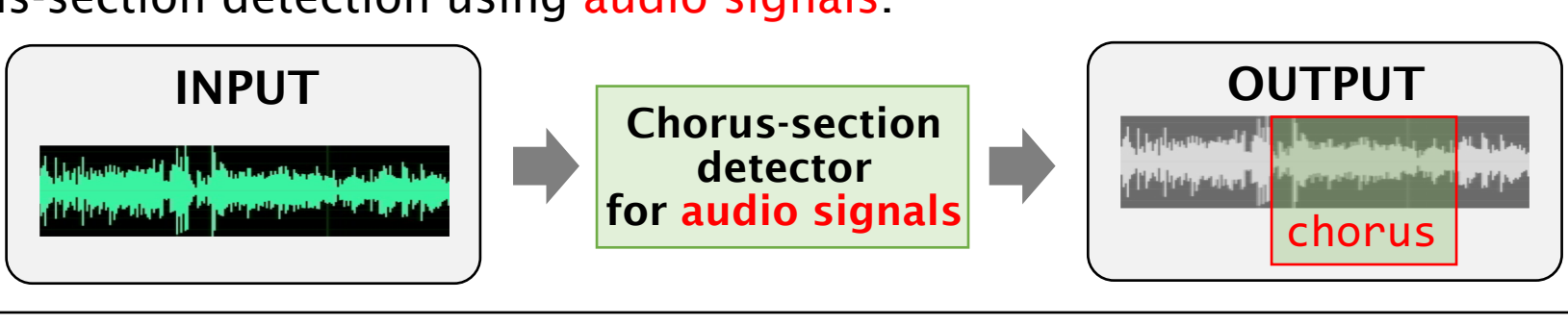
Kento Watanabe and Masataka Goto
National Institute of Advanced Industrial Science and Technology (AIST)

AIST

Background

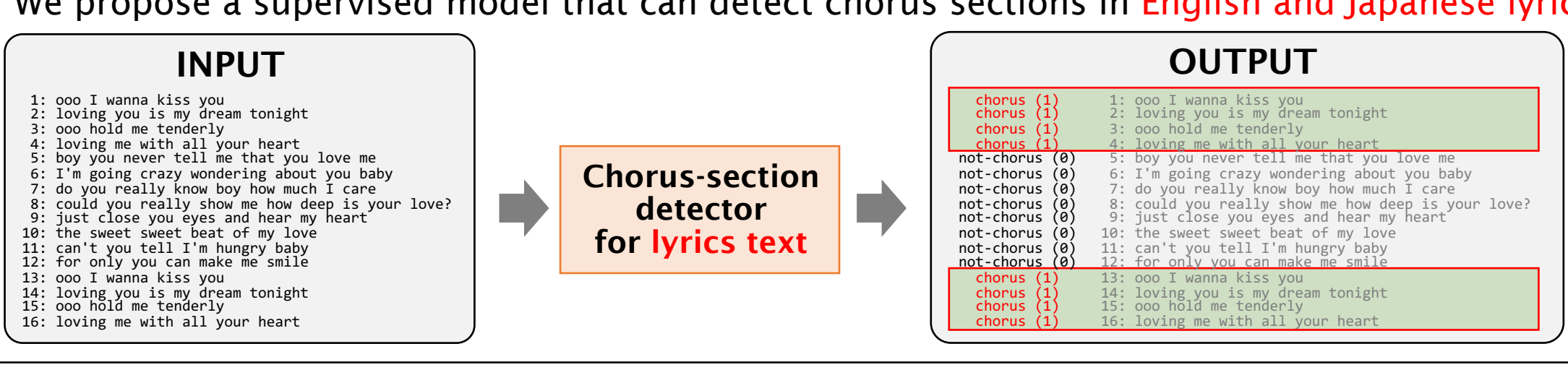
■ Previous study

- Chorus-section detection using audio signals.



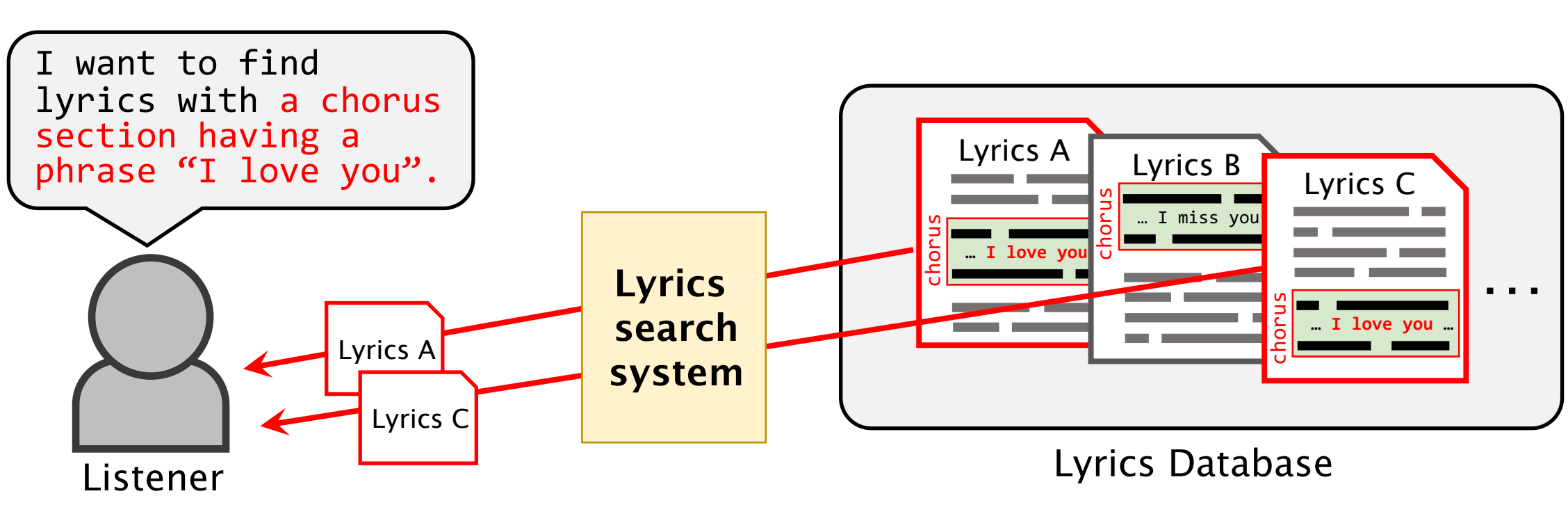
■ This study

- Whether chorus sections can be detected from text-only lyrics is an open issue.
- We propose a supervised model that can detect chorus sections in English and Japanese lyrics.



Potential Application

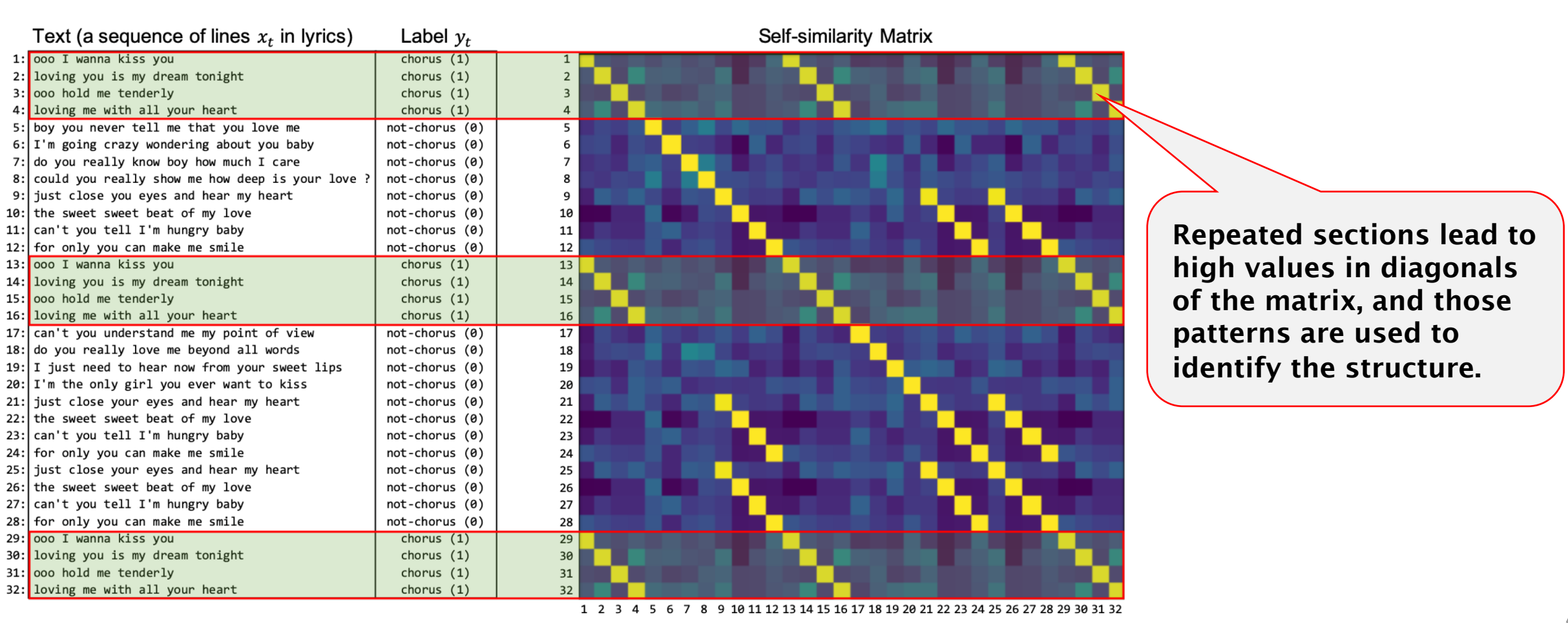
■ A lyrics search system



It is necessary for a lyrics search system to automatically detect which lines of the lyrics are included in chorus sections.

Key Idea

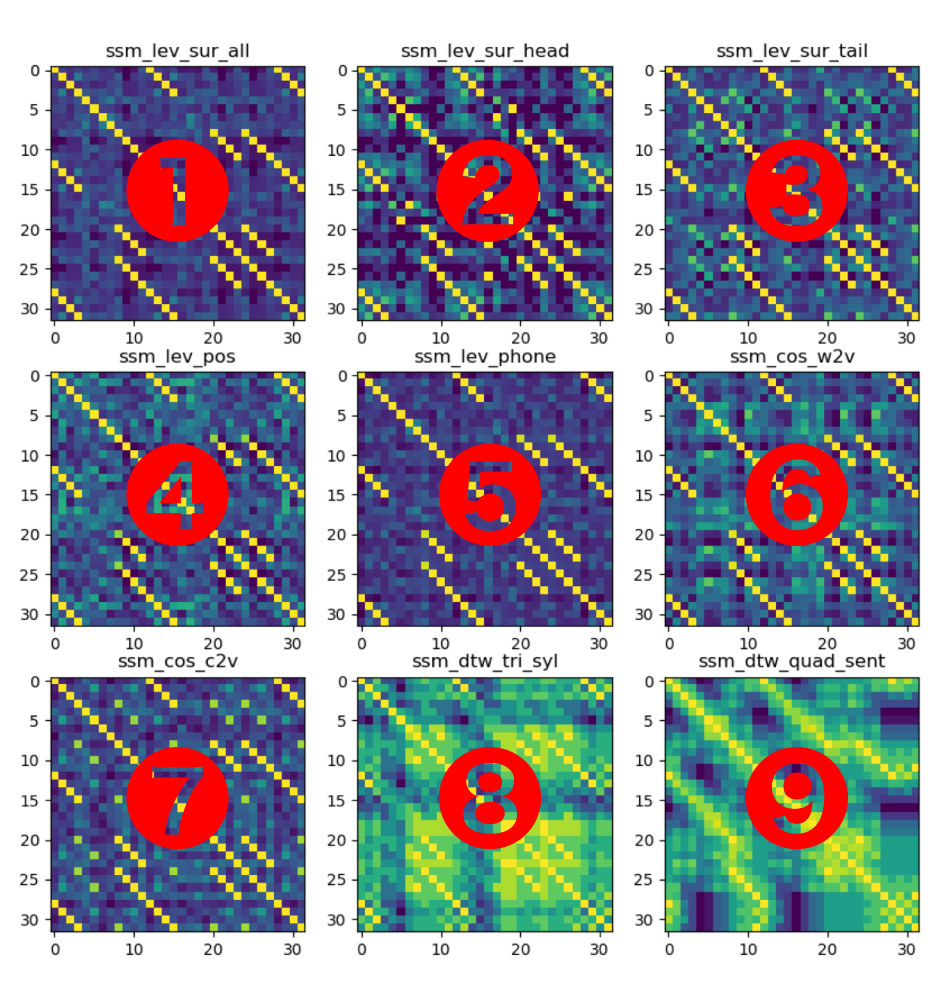
- We compute the **Self-similarity matrix (SSM)** from lyrics text.
- SSM representations are widely used in computational music structure analysis.



Repeated sections lead to high values in diagonals of the matrix, and those patterns are used to identify the structure.

Nine SSMs

- The **design of the similarity measure** to compute each cell of the SSM is important.
- We propose to use the following nine variations of similarity measures.



Edit distance

① String similarity
② Head similarity
③ Tail similarity
④ Part-of-speech similarity
⑤ Phonetic similarity

Cosine similarity

⑥ Word vector similarity
⑦ Context vector similarity

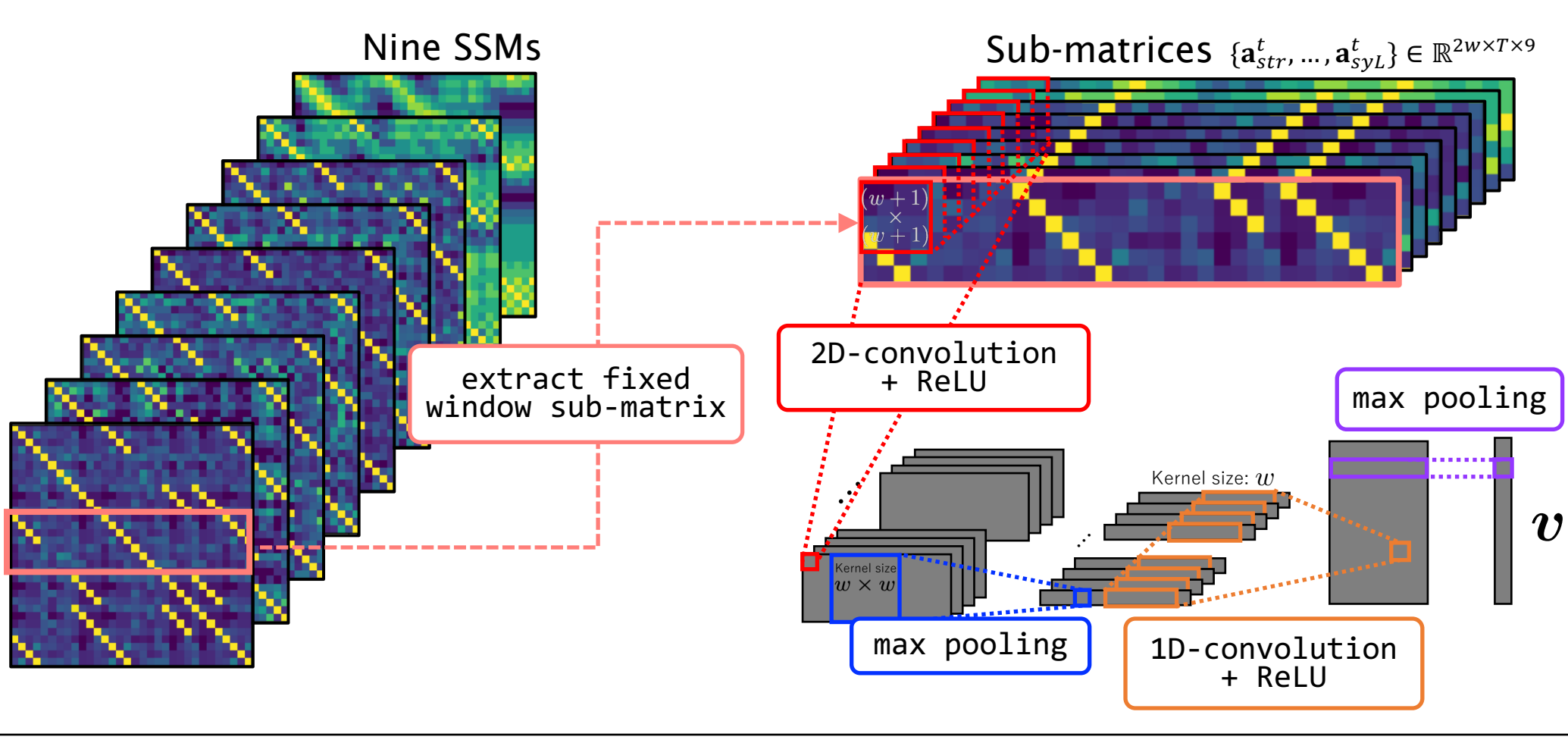
Dynamic time warping

⑧ Word syllable count similarity
⑨ Lyric line syllable count similarity

Structural Feature

* [Fell et al. 2018]

To calculate feature vectors v_t from the above nine SSMs, we use a **CNN architecture*** to detect textual macro structures from various patterns in SSMs regardless of their locations and relative sizes in SSMs.



Some expressions of lyrics tend to appear in chorus sections?

Chorus-section Tri-gram	$P_c - P_n$	Non-chorus-section Tri-gram	$P_n - P_c$
I'm	0.12%	there's	0.04%
don't	0.11%	I've	0.03%
oh oh oh	0.05%	's a	0.03%
I'll	0.05%	I'd	0.02%
we're	0.04%	but I'	0.02%
you're	0.04%	's not	0.01%
'll be	0.04%	what's	0.01%
I don'	0.04%	na na na	0.01%
Let's	0.03%	yeah yeah yeah	0.01%
you got ta	0.03%	've been	0.01%
I can'	0.03%	't take	0.01%
can't	0.03%	didn't	0.01%

P_c and P_n denote word tri-gram probabilities in the chorus and non-chorus sections, respectively.

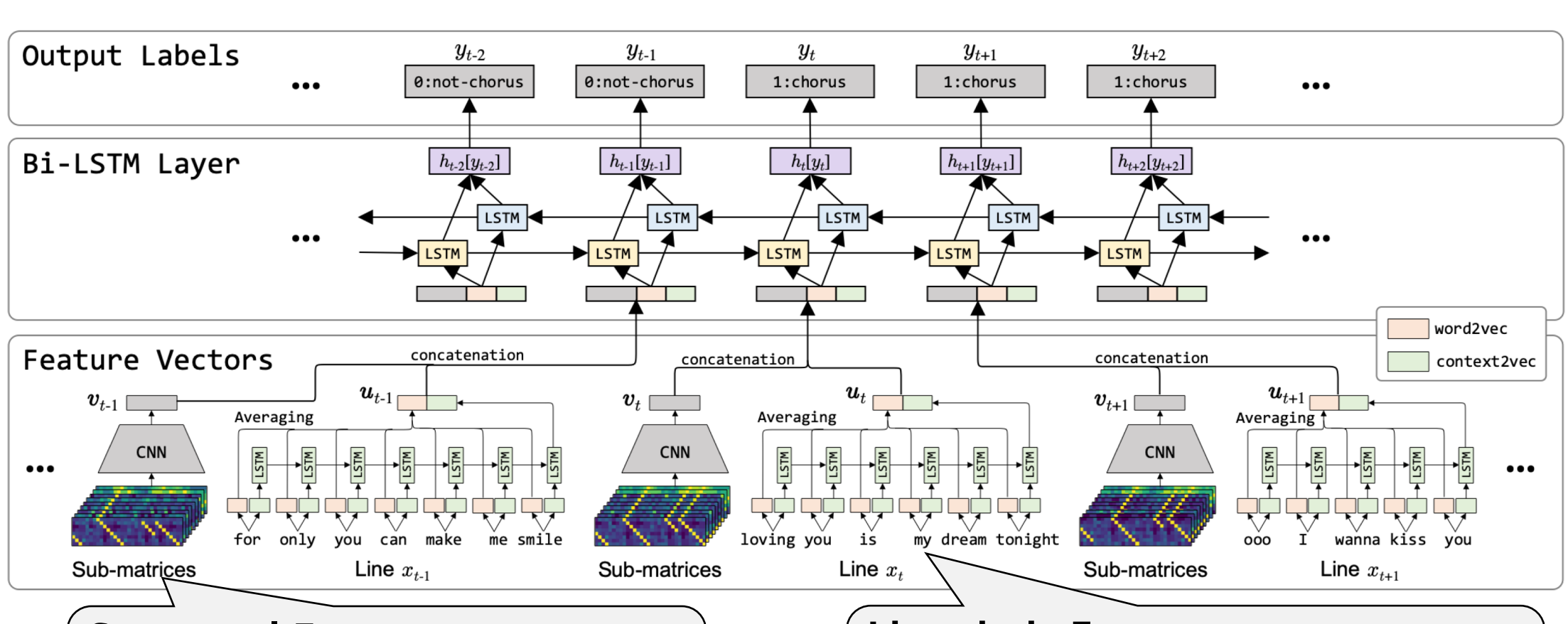
■ Findings

Phrases about the future (e.g., *I'll* and *Let's*) tend to appear in chorus sections more often than do phrases about the past (e.g., *have been* and *didn't*).

It can be expected that some expressions will contribute to the chorus-section detection.

Chorus-section Detection Model

- We propose a Bidirectional LSTM-based model using **two types of feature representations**.



Structural Feature
We exploit a CNN architecture to detect textual macro structures from nine SSMs.

Linguistic Features
Our model utilizes linguistic features such as word vectors and sentence vectors.

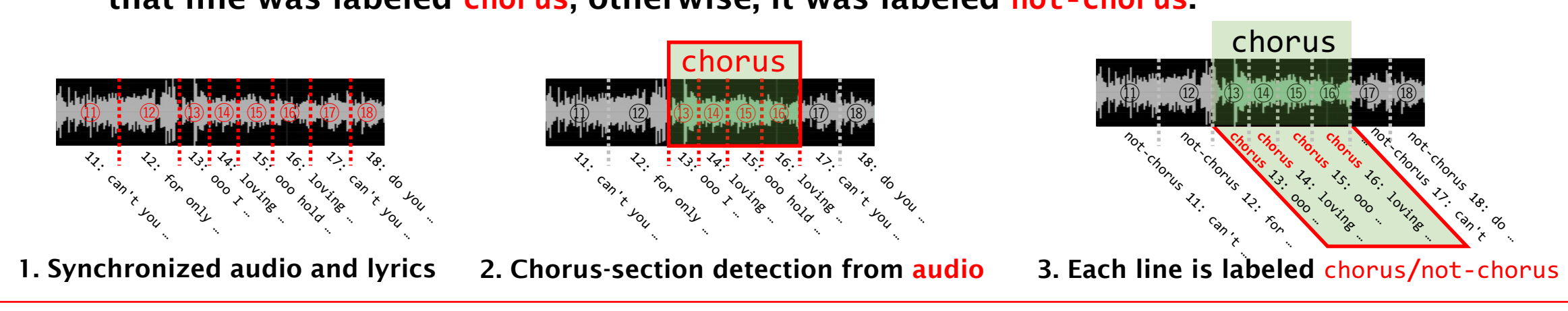
Training Data

■ Technical Problem for training our model

No lyrics data with line-level chorus-section annotations are available.

■ Key Idea： We generate training data with chorus-section annotation

1. **We prepared 100,772 pairs of musical audio signals and their corresponding manually time-aligned (temporally synchronized) lyrics.**
2. **We detected chorus sections of every song automatically by using its audio signals.**
3. **If the start time of a lyric line was within any chorus section detected in audio signals, that line was labeled chorus; otherwise, it was labeled not-chorus.**



1. Synchronized audio and lyrics 2. Chorus-section detection from audio 3. Each line is labeled chorus/not-chorus

Dataset

■ Dataset for training comparison

- EN_auto: English lyrics data with the generated chorus-section annotations. (9,313 songs)
- JA_auto: Japanese lyrics data with the generated chorus-section annotations. (91,459 songs)
- JA_man: Japanese lyrics data with the gold chorus-section annotations. (1,103 songs)

■ Dataset for tuning model parameters

We annotated chorus-section label manually.

- EN_RWC: English lyrics data in RWC Music Database. (21 songs)
- JA_RWC: Japanese lyrics data in RWC Music Database. (79 songs)

■ Dataset for testing

We annotated chorus-section label manually.

- EN_test: English lyrics data. (118 songs)
- JA_test: Japanese lyrics data. (128 songs)

Experimental Results (1)

■ Question

To confirm **the effectiveness of our Bi-LSTM model** that can learn dependencies between adjacent lyric lines, we compared its performance with that of two baseline methods.

■ Result

Method	Training and test data (# of training songs) English (9,313 songs) F-measure	Japanese (91,459 songs) F-measure
Heuristic (extract repeated lines as chorus sections)	57.8 %	57.1 %
Multi-Layer Perceptron	74.2 %	80.6 %
Bi-LSTM (Proposed)	**78.1 %**	**83.4 %**

■ Findings

- **Methods based on supervised learning are better than a rule-based method.**
- **The proposed method is the best for the chorus-section detection task.**

Experimental Results (2)

■ Question

To investigate **the effectiveness of structural and linguistic features**, we compared their use individually and in combination.

■ Result

Feature	Training and test data (# of training songs) English (9,313 songs) F-measure	Japanese (91,459 songs) F-measure
Structural feature	77.9 %	81.2 %
Linguistic feature	57.4 %	55.2 %
Both	**78.1 %**	**83.4 %**

■ Findings

- **The model with only the structural features greatly outperformed.**
- **The additional use of linguistic features is helpful for detecting chorus sections.**

Experimental Results (3)

■ Question

We confirm that **our generated data is reliable enough for training purposes** by comparing the performance of the model trained on JA_auto with that of the model trained on JA_man.

■ Result

Training Data	F-measure (Japanese test data)
JA_auto: generated training data (91,459 songs)	**83.4 %**
JA_man: human-annotated training data (1,103 songs)	80.3 %

■ Findings

- **The model trained using generated data (91,459 songs) outperformed the model trained using human-annotated data (1,103 songs).**
- **Even if generated annotations are not perfect, they are reliable enough for training.**

Experimental Results (4)

■ Question

- Can a model trained on a large amount of Japanese data detect English chorus sections?
- Can a model trained on both EN_auto and JA_auto perform better than one trained on only EN_auto or JA_auto?

★ Structural features based on the SSMs can be language independent because our SSMs simply represent patterns of repeating lyric lines, which could be universal in music. So we use the model without linguistic features in this experiment.

■ Result

Training Data	F-measure (English test data)
EN_auto: generated training English data (9,313 songs)	77.9 %
JA_auto: generated training Japanese data (91,459 songs)	80.3 %
EJ_auto: EN_auto + JA_auto (100,772 songs)	**81.0 %**

■ Findings

- **The SSM-based model can detect chorus sections regardless of the language.**
- **English and Japanese SSMs (i.e., patterns of repeating lyric lines) have similar structures.**
- **Mixing different language data allows the model to learn the general structure of chorus sections and thereby perform better.**

Conclusion

■ Contributions

1. We designed a variety of features to capture **structural and linguistic properties of chorus sections**.
2. We proposed a sequence labeling model that can detect chorus sections in lyrics.
3. We showed how to **generate a large training dataset** of lyrics with chorus-section annotations.
4. We demonstrated that **our Bi-LSTM-based method outperforms** alternative baseline methods.
5. We thoroughly investigated this detection task and the nature of chorus sections of lyrics from different perspectives such as the **importance of features**, the **amount of training data**, and **language dependency**.