LEVERAGING REPETITION FOR IMPROVED AUTOMATIC LYRIC TRANSCRIPTION IN POPULAR MUSIC

Matt McVicar[†] *Daniel PW Ellis* [⋆] *Masataka Goto*[†]

[†] National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan
[⋆] LabROSA, Dept. of Electrical Engineering, Columbia University, New York NY, USA

ABSTRACT

Transcribing lyrics from musical audio is a challenging research problem which has not benefited from many advances made in the related field of automatic speech recognition, owing to the prevalent musical accompaniment and differences between the spoken and sung voice. However, one aspect of this problem which has yet to be exploited by researchers is that significant portions of the lyrics will be repeated throughout the song. In this paper we investigate how this information can be leveraged to form a consensus transcription with improved consistency and accuracy. Our results show that improvements can be gained using a variety of techniques, and that relative gains are largest under the most challenging and realistic experimental conditions.

Index Terms— Music Information Retrieval, Automatic Lyric Recognition, Automatic Speech Recognition

1. INTRODUCTION

Lyrics are the set of words to a song, and are sung to form the vocal component of popular music. Whilst a large amount of research in Music Information Retrieval (MIR) has focused on content-based analysis tasks such as beat tracking [1], chord identification [2], and music segmentation [3], there is much less work on the subject of lyric analysis from audio. This is despite research that suggests that the lyrics of a song can have a profound effect on a listener's opinion of a song [4], can be indicative of style or genre [5], and can even affect behaviour over prolonged periods of time [6].

For these reasons there is a growing interest in applications in MIR involving lyrics (see [5, 7, 8]). However, the majority of these studies assume that the lyrics to a song are known in advance. The reason for this is clear: despite huge advances in Automatic Speech Recognition (ASR), Automatic Lyric Recognition (ALR) is a challenging problem, in part due to the background musical accompaniment and low similarity between the spoken and sung voices [7, 9–14].

However, there is one aspect in which 'musical speech' may be easier to transcribe than conversational speech: the use of repetition. Repetitions evoke feelings of familiarity and understanding in the listener in music [3] and help mediate expectation and novelty [15]. Specifically of interest to the current study is that the lyrics of a chorus are often approximately constant [16]. In the current work, we investigate whether the information in manually-labelled repeated choruses can be shared to yield improved lyric transcriptions: see Figure 1 for an outline of our proposed methodology. The remainder of this work is organised as follows: in Section 2, we discuss relevant work in the task of automatic lyric transcription, how structural information has previously been used to boost performance in MIR tasks, and the Recognizer Output Voting Error Reduction algorithm, which forms the basis of one of our proposed techniques. Our novel methods are described in Section 3, and evaluated in Section 4. We conclude the work and discuss areas of future research in Section 5.

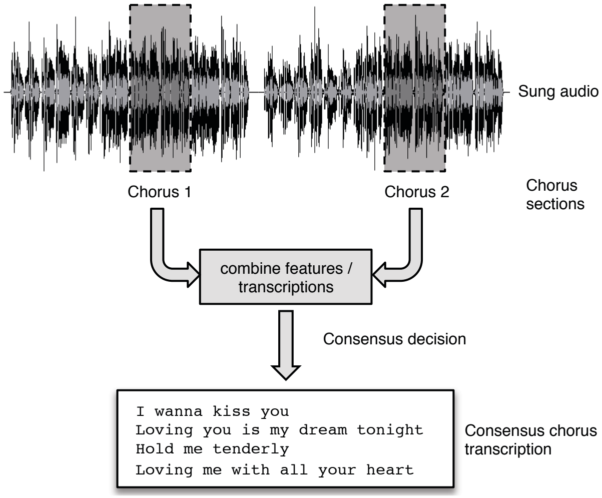
Sung audio
Chorus 1
Chorus 2
Chorus sections
combine features / transcriptions
Consensus decision
I wanna kiss you
Loving you is my dream tonight
Hold me tenderly
Loving me with all your heart
Consensus chorus transcription

Fig. 1: Outline of our proposed techniques. The chorus sections of acapella audio are aggregated at either the feature or hypothesis level to form a lyric transcription which is consistent across sections and, we hope, more accurate than a transcription computed per chorus.

2. BACKGROUND & RELEVANT WORK

2.1. Automatic lyric alignment/recognition

Perhaps due to the challenging nature of performing full transcription of the sung voice, researchers have mostly in the past concentrated on the task of aligning/synchronising lyrics to audio, where the task is to assign timestamps to a set of lyrics given the corresponding audio (see, for example, [12, 17–20]).

However, there are clearly situations in which ALR is required. For example, an accurate ALR system could be used in situations where the lyrics are not available, or to aid an alignment-based system when untimed lyrics (perhaps obtained from the Web) are inaccurate. ALT techniques could also give rise to ASR systems which are highly robust to noise. The literature on unaided recognition of lyrics is limited to just a handful of studies, most notably by Mesaros and Virtanen [7, 21], who attempted the recognition of phonemes and words from both acapella and polyphonic music, the latter via automatically extracting the vocal line from the polyphonic and re-synthesizing into a clean mix. Acoustic and language model adaptation was conducted in order to boost performance, which peaked at 12.4% / 20.0%

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

word/phoneme accuracy (see Section 4) for 49 fragments of 12 songs, each between 20 and 30 seconds in duration. Further improvements were seen by adapting specifically to the male/female sung voice.

Following this, the authors went on to show that their system could be used for two interesting applications, despite the relatively low performance: lyric alignment and lyric retrieval. In the former, they achieved an average alignment error from reference to hypothesis of 1.27s, whilst in the latter, precision-at-k for retrieving a song based on automatically transcribed lyrics revealed precisions of 57%, 67%, 71% at $k = 1, 5, 10$ respectively.

2.2. Structural information to aid MIR

Previous studies have shown the benefits of utilising structural information in MIR tasks. For example, Dannenberg [22] improved beat tracking performance by incorporating structural labels, whilst Rafii and Pardo [23] used structural cues to aid source separation. Mauch et. al [24] noticed an improvement in chord recognition accuracy by exploiting structural similarity. Audio features in this last work belonging to the same segment (*chorus*, for example) were averaged before being fed to the classifier (a dynamic Bayesian network), which led to an improvement in accuracy in a majority of songs. The authors also mentioned that an additional boon of this technique was that it ensured sections were consistently labelled, regardless of if an improvement in performance was noted.

2.3. Recognizer Output Voting Error Reduction (ROVER)

Modern ASR systems are sensitive to a number of parameters, and it is difficult to know in advance which parameter set or algorithm will be optimal for a given task. Also, it is possible that some systems are better at transcribing certain words, or perform better under different conditions within an utterance (such as alternative pronunciations or types of noise). In order to take advantage of the varying strengths of multiple ASR systems, Recogniser Output Voting Error Reduction (ROVER) was developed by Jonathan G. Fiscus in 1997 [25].

ROVER takes as input multiple transcriptions of the same audio from different ASR systems and combines them to form a consensus transcription, often with lower error rate. It performs this by Dynamic Time Warping (DTW) two transcriptions together, keeping track of the optimal insertions, deletions and substitutions required to transform one transcription to the next in a Word Transition Network (WTN). The output of the DTW procedure is then warped to the third transcription, and so on, until the list of transcriptions has been exhausted. This procedure is illustrated in Figure 2. Next, an optimal left-to-right path through the final WTN is formed. Majority vote at each node can be used to achieve this, but ties with equal number of counts must be broken arbitrarily. To resolve this, Fiscus suggests using word confidence scores to make a more informed decision. To this end, the score for a given word w with confidence $C(w) \in [0, 1]$ and which occurs with relative frequency $F(w) \in [0, 1]$ is defined:

$$\text{Score}(w) = \alpha F(w) + (1 - \alpha)C(w), \tag{1}$$

where α defines the balance between trusting word frequency and word confidence. Setting $\alpha = 1$ is equivalent to using majority vote at each node in the WTN, whilst $\alpha = 0$ corresponds to only using word confidence. A parameter C_{INS} is used to define the confidence of inserting a silence into the WTN (deletions have no cost).

Given that a word may occur multiple times in each hypothesis, each with different confidence, setting $C(w)$ is non-trivial. Fiscus suggests two methods for setting $C(w)$: either choosing the maximum confidence over all occurrences of w, or averaging the scores.

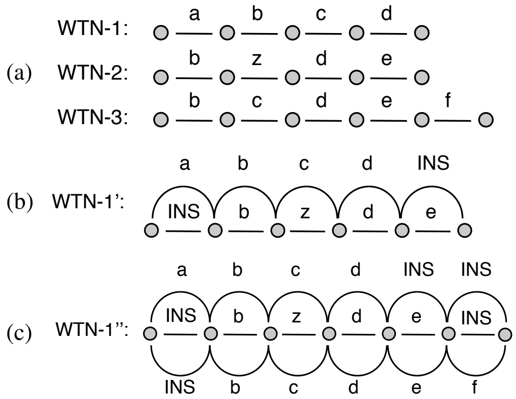

Fig. 2: Example of the ROVER Word Transition Network algorithm (from [25]) for three simple hypothesis transcriptions, WTN-1, WTN-2, WTN-3. (b): WTN-2 is first Dynamic Time Warped to WTN-1, resulting in one insertion (INS) at the start of the sequence, forming WTN-1'. (c): WTN-3 is then DTW'd to the resulting sequence, requiring one INS at the end of the sequence, forming WTN-1". Majority vote & tie-breaking schemes are then used to find a consensus transcription from WTN-1", (majority vote consensus for this example being "b, c, d, e")

Finally, the consensus output for the algorithm is computed by computing the path through the final WTN which greedily chooses the maximum score at each node.

3. PROPOSED TECHNIQUES

In this Section, we describe the main contributions of this work. As stated in Section 1, our main idea is to exploit the fact that the lyrics for popular songs are often repeated in a chorus, meaning we have multiple utterances which may be used to aid transcription. There are many ways in which the multiple choruses may be aggregated; we describe three methods in the remainder of this Section.

3.1. Average MFCC consensus chorus transcription

As stated in Section 2, averaging of audio features in identical sections has previously shown to improve performance in chord estimation. A natural analogue of this in ALR is to try averaging MFCC features which belong to the chorus of the song (assuming, as in [24] that the choruses are of equal duration) and feeding this to an ASR system.

It is not immediately obvious that this simple technique will yield any improvement, although visualising the MFCCs for two realisations of a chorus (Figure 3), we do indeed notice a high degree of correlation between the two signals, though subtle variations exist, consisting perhaps of variations in pitch, volume, or phone duration. Note also in Figure 3 that the rhythmic nature of popular music means that the word onsets in choruses are likely to be well-aligned.

3.2. Maximum likelihood consensus chorus transcription

Another methodologically simple technique to resolving multiple chorus transcriptions into one consensus is to transcribe each chorus individually, and choose the transcription with the highest model likelihood. This method is appealing in that the audio features remain unchanged and therefore well-matched to the acoustic model

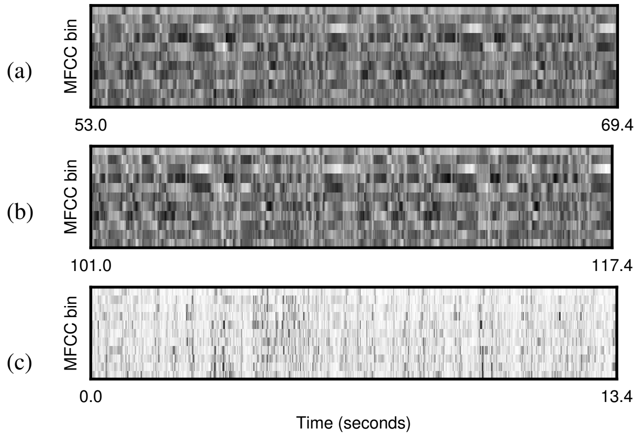

Fig. 3: MFCC features (first 12 coefficients plus energy) for two realisations of a chorus, (a) and (b), from the RWC Music Database [26]. The final subfigure (c) shows their absolute difference, indicating good alignment a high degree of similarity.

parameters, although it suffers from having to make a 'hard' decision over which chorus transcription to use, neglecting the fact that each transcription may have individual strengths and weaknesses.

3.3. ROVER-based consensus chorus transcription

The two methods listed above each suffer from a flaw: averaging MFCCs (Subsection 3.1) may yield poor-quality features not well-match to the model parameters, maximum-likelihood chorus selection (Subsection 3.2) ignores the interactions between the choruses and forces a hard decision. In our final methods, we will use ROVER to combine multiple outputs into a consensus transcription, which we anticipate will have neither of the drawbacks listed above.

To our knowledge, ROVER has only previously been used to combine the output of *multiple ASR systems* for a *single audio segment*. The contribution of using this technique to transcribe *multiple audio segments* using a *single ALR system* is therefore a novel contribution of this work. Crucially, it is the repetitive verse-chorus nature of popular music which allows us to exploit this; the same technique will not yield advances in the transcription of conversational speech.

4. EXPERIMENTS

To test the methods described in Section 3, unaccompanied song samples (solo vocal) were taken from the RWC Music Database (Popular Music, RWC-MDB-P-2001 No. 81-100 [26]), containing 20 songs sung in English by professional male and female singers at a 16kHz sampling rate, mono Microsoft Wave audio format. Although separation and isolation of the vocal melody is clearly required for real-world applications, we feel performing recognition on solo sung voice is challenging enough as a research problem, and an interesting starting point to establish an upper bound for our techniques. For these reasons, the audio used in this work consists of unaccompanied sung utterances. Since our proposed method is concerned with chorus sections, all other sections were stripped from the data manually.

Lyrics for these songs were obtained from the same source and manually checked to be consistent with the audio. We defined the chorus of each song to be the longest set of consecutive words which were at some point repeated. Two songs were discarded at this stage as they were judged to contain no chorus sections. A summary of the test dataset after the processing steps mentioned above can be found in Table 1. Methods 3.1–3.3 were tested across each song in the database, with performance measured by calculating the total number of insertions I, deletions D, and substitutions S required to convert every transcript to its corresponding reference annotation. The percentage of correct words and the accuracy was then calculated at the word or phoneme level as

$$\text{Accuracy} = \frac{N - I - D - S}{N} \times 100\% \tag{2}$$

n	duration	words	phones	choruses
18	19:17 / 00:58	1930 / 107	5710 / 291	57 / 3

Table 1: Total number of songs, total/median duration, number of words, phones and choruses for data used in this paper.

4.1. Baseline ALR system

For our baseline system we used a cross-word triphone HMM trained on the WSJ corpus [27] with 8 Gaussians per phone and 16 silence Gaussians. The model had approximately 2750 tied states. It has been noted in previous research [7] that the language models required for decoding singing differ from those trained on speech. Concerned that this would result in such poor performance that our system would not see any improvement, we constructed a bigram language model per test set song (which we call 'Song'), to act as an upper bound on performance. This assumption was then relaxed by constructing a bigram model from the entire test set, ('Test'), and furthermore by using a general-purpose bigram model trained on the WSJ transcriptions ('WSJ'). Each of the models were formed from the unstressed pronunciation from the CMU set of 40 phones.

Finally, suspecting that this model would not match our sung audio, following [21] we performed acoustic adaptation via two rounds of Maximum Likelihood Linear Regression on a set of ten held-out (English) training songs from the RWC Music Database (Royalty-Free Music, RWC-MDB-R-2001 No.6-15 [26]). All decoding was performed using the HDecode command within HTK [28].

4.2. Averaging/max-likelihood consensus chorus transcription

Our first methods for exploiting repetitions in songs involve taking the mean of MFCC features or choosing the chorus sequence with the highest likelihood. Results for these experiments together with the baseline method can be seen in Table 2. Inspecting these results,

	Performance / Language Model					
	Word Accuracy (%)			Phone Accuracy (%)		
Method	Song	Test	WSJ	Song	Test	WSJ
Baseline	43.78	23.47	03.40	53.33	38.95	24.40
Mean	41.97	21.81	05.39	51.24	38.35	**26.99**
Likel.	**48.70**	**27.15**	**05.91**	**56.08**	**41.56**	**26.99**

Table 2: Baseline autmatic lyric recognition method and two preprocessing techniques which leverage structural information. Best results for each column are shown in boldface, statistically significant improvements from the baseline are underlined.

we at first see that the performance of the proposed ALR system is not as high as for ASR systems [29], even for the case when we have full knowledge of the language model per song (columns 1 and 4). This indicates that significant research effort needs to be invested into acoustic model creation if performance in ALR is to approach the accuracy of ASR systems. Performance is highest when the most information about the language model is given, with a decrease in accuracy from 43.87% to 3.40% at the word level and 53.33% to 24.40% at the phone level. Although these figures are low in magnitude, they are similar to results from previous studies (see Subsection 2.1), at least at the phoneme level.

Moving on to the first of the proposed methods (row 2), we see that simple averaging of MFCC features only shows an improvement with the WSJ language model. We believe the reason that the same improvements reported for chord recognition using the same technique cannot be replicated, is that the rate at which phonemes change is far faster than musical chords meaning that the salient phoneme onsets are 'blurred out' by the averaging process. We also tested the improvements seen using this technique for significance using the paired non-parametric Wilcoxon signed-rank test, using the per song accuracy as a statistic, finding no significant improvement.

Finally, row 3 of Table 2 shows that choosing the maximum-likelihood utterance yields improvements over the baseline in all cases and evaluations. All improvements were found to be significant at the 5% level using the signed-rank test per song. Interestingly, it seems that the largest relative improvements for word accuracy are found when baseline performance is lowest (relative word improvements: 11%, 16%, 74%), contrary to our assumptions (see Subsection 4.1) and indicating that leveraging multiple choruses is most advantageous in the most realistic, challenging scenarios.

4.3. ROVER-based consensus chorus transcription

Next, we implemented ROVER on our chorus audio, granting it access to each baseline chorus transcription and using the scoring methods listed in Subsection 2.3. We found that $C_{\text{INS}} = 1$ and values of α between 0.4 and 0.8 were required to produce optimal Word Accuracy. Results can be seen in Table 3. Inspecting Table 3, we see that ROVER offers substantial improvements over the baseline in every regard. Word accuracy increases from around (43%, 23%, 3%) to (50%, 31%, 9%) with (song, test-set, WSJ) language models respectively, representing relative improvements of (16%, 35%, 200%). All improvements at the word level were found to be significant. There appears to be little difference between the scoring methods (rows 2–4) , which was confirmed by a signed-rank statistical test. Performance is also higher than each of the results from Subsection 4.2, confirming our hypothesis that the optimal transcript comes from different parts of competing hypotheses.

	Performance / Language Model					
	Word Accuracy (%)			Phone Accuracy (%)		
Method	Song	Test	WSJ	Song	Test	WSJ
Baseline	43.78	23.47	03.40	53.33	38.95	24.40
Majority	**50.05**	**30.98**	**09.48**	55.06	42.75	25.53
Max con.	49.69	30.73	**09.48**	55.62	41.52	**25.64**
Av. con.	49.95	30.78	**09.48**	**55.71**	**43.06**	**25.64**

Table 3: Word/phoneme accuracies for the baseline and ROVER-based approaches. Best results for each column are in boldface.

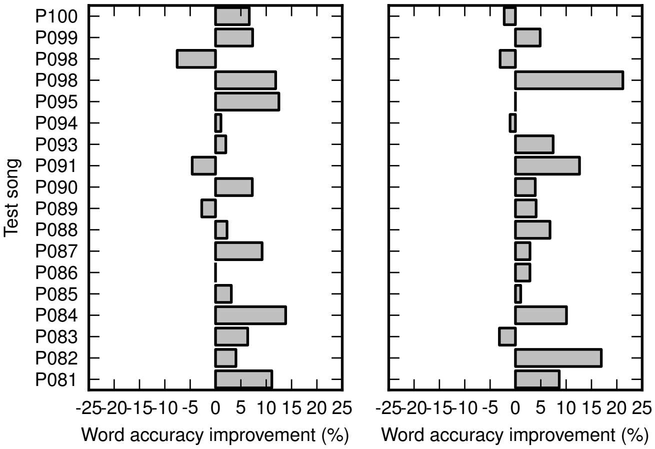

Fig. 4: Absolute word accuracy improvements using our proposed techniques (Likelihood, left; ROVER majority vote, right) using the 'Song' language model.

Performance improvements (in absolute word accuracy %) for two of our methods can be seen in Figure 4, showing an improvement in 14 of 18 songs for method 3.2, and an improvement of 12 of 18 songs (of larger magnitude) for method 3.3. Interestingly, it seems that in some cases one model improves where the other fails, indicating that a further round of ROVER on the outputs may yield further improvements, which has been noted in the ASR literature [29].

5. CONCLUSIONS AND FUTURE WORK

In this paper, we investigated how repeated utterances phrases in chorus songs can be combined to form a consistent consensus transcription. We tested one existing method (feature averaging) and introduced two new methods to achieve this goal: maximum likelihood chorus selection and ROVER. The application of ROVER to multiple audio example using a single ASR system is a novel contribution of this work. Experimental results showed that averaging of MFCC features did not offer an improvement in most cases, but that significant gains can be made using the two proposed methods. Particularly surprising to us was that ROVER-based methods offered the greatest relative improvement when baseline performance was lowest, highlighting the potential of this method in real-world tasks.

The techniques presented in this work may be of use in other MIR scenarios in which a consensus annotation from many candidates is desired, which has been the subject of at least two recently-published papers on beat tracking [30] and chord detection [31]. We would like to explore this in future work, as well as create high-quality acoustic and language models for ALR. Finally, we would like to investigate if the emission probabilities from multiple choruses can be effectively modelled using a multiple-emission Hidden Markov Model framework. All of the above we hope will boost baseline ALR accuracy, bringing performance closer to the figures seen in ASR.

6. ACKNOWLEDGEMENTS

The authors would like to thank Keith Vertanen for making his WSJ HTK training recipes available online [32] and Jun Ogata for his assistance with acoustic model adaptation within HTK. This work was supported in part by CREST, JST and NSF grant IIS-1117015.

7. REFERENCES

[1] M. Davies and M. Plumbley, "Context-dependent beat tracking of musical audio," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 1009–1020, 2007.

[2] M. Mauch and S. Dixon, "Simultaneous estimation of chords and musical context from audio," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 6, pp. 1280–1289, 2010.

[3] J. Paulus, M. Müller, and A. Klapuri, "State of the art report: Audio-based music structure analysis," in *Proc. ISMIR*, 2010, pp. 625–625–636.

[4] B. Anderson, D. Berger, R. Denisoff, K. Etzkorn, and P. Hesbacher, "Love negative lyrics: Some shifts in stature and alterations in song," *Communications*, vol. 7, no. 1, pp. 3–20, 1981.

[5] R. Mayer, R. Neumayer, and A. Rauber, "Rhyme and style features for musical genre classification by song lyrics," in *Proc. ISMIR*, 2008, pp. 337–342.

[6] C. Anderson, N. Carnagey, and J. Eubanks, "Exposure to violent media: The effects of songs with violent lyrics on aggressive thoughts and feelings," *Journal of personality and social psychology*, vol. 84, no. 5, pp. 960–971, 2003.

[7] A. Mesaros and T. Virtanen, "Automatic recognition of lyrics in singing," *European Association for Signal Processing Journal on Audio, Speech, and Music Processing*, vol. 2010, 2010.

[8] X. Hu and B. Yu, "Exploring the relationship between mood and creativity in rock lyrics," *Proc. ISMIR*, pp. 789–794, 2011.

[9] M. Goto, T. Saitou, T. Nakano, and H. Fujihara, "Singing information processing based on singing voice modeling," in *ICASSP*. IEEE, 2010, pp. 5506–5509.

[10] J. Sundberg, "Formant structure and articulation of spoken and sung vowels," *Folia Phoniatrica et Logopaedica*, vol. 22, no. 1, pp. 28–48, 1970.

[11] D. Lundy, S. Roy, R. Casiano, J. Xue, and J. Evans, "Acoustic analysis of the singing and speaking voice in singing students," *Journal of voice*, vol. 14, no. 4, pp. 490–493, 2000.

[12] H. Fujihara, M. Goto, J. Ogata, and H. Okuno, "LyricSynchronizer: Automatic synchronization system between musical audio signals and lyrics," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1252–1261, 2011.

[13] Y. Ohishi, M. Goto, K. Itou, and K. Takeda, "Discrimination between singing and speaking voices.," in *Proceedings of INTERSPEECH*, 2005, pp. 1141–1144.

[14] Y. Ohishi, M. Goto, K. Itou, and K. Takeda, "On human capability and acoustic cues for discriminating singing and speaking voices.," in *Proc. ICMPC*, 2006, pp. 1831–1837.

[15] E. Narmour, *The analysis and cognition of melodic complexity: The implication-realization model*, University of Chicago Press, 1992.

[16] C. Xu, X. Shao, N. C Maddage, and M. Kankanhalli, "Automatic music video summarization based on audio-visual-text analysis and alignment," in *Proceedings of the 28th ACM SIGIR conference on research and development in information retrieval*. ACM, 2005, pp. 361–368.

[17] K. Lee and M. Cremer, "Segmentation-based lyrics-audio alignment using dynamic programming.," in *Proc. ISMIR*, 2008, pp. 395–400.

[18] A. Mesaros and T. Virtanen, "Automatic alignment of music audio and lyrics," in *Proc. DAFx*, 2008, pp. 321–324.

[19] Y. Wang, M. Kan, T. Nwe, A. Shenoy, and J. Yin, "Lyrically: automatic synchronization of acoustic musical signals and textual lyrics," in *Proceedings of the 12th annual ACM international conference on Multimedia*. ACM, 2004, pp. 212–219.

[20] M. Mauch, H. Fujihara, and M. Goto, "Integrating additional chord information into HMM-based lyrics-to-audio alignment," *Audio, Speech, and Language Processing, IEEE Transactions on*, pp. 200–210, 2012.

[21] A. Mesaros and T. Virtanen, "Recognition of phonemes and words in singing," in *ICASSP*. IEEE, 2010, pp. 2146–2149.

[22] R. Dannenberg, "Toward automated holistic beat tracking, music analysis, and understanding," in *Proc. ISMIR*, 2005, pp. 366–373.

[23] Z. Rafii and B. Pardo, "REpeating Pattern Extraction Technique (REPET): A simple method for music/voice separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 1, pp. 73–84, 2013.

[24] M. Mauch, K. Noland, and S. Dixon, "Using musical structure to enhance automatic chord transcription," in *Proc. ISMIR*, 2009, pp. 231–236.

[25] J. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proceedings of Automatic Speech Recognition and Understanding, IEEE Workshop on*. IEEE, 1997, pp. 347–354.

[26] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC Music Database: Popular, classical and jazz music databases.," in *Proc. ISMIR*, 2002, pp. 287–288.

[27] D. Paul and J. Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[28] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK book (for HTK version 3.2)," Tech. Rep., Cambridge university engineering department, 2002.

[29] H. Schwenk and J. Gauvain, "Improved ROVER using language model information," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[30] A. Holzapfel, M. Davies, J. R Zapata, L. Oliveira, and F. Gouyon, "Selective sampling for beat tracking evaluation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 9, pp. 2539–2548, 2012.

[31] Y. Ni, M. McVicar, R. Santos-Rodriguez, and T. De Bie, "Understanding effects of subjectivity in measuring chord estimation accuracy," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 12, pp. 2607 – 2615, 2013.

[32] K. Vertanen, "Baseline WSJ acoustic models for HTK and Sphinx: Training recipes and recognition experiments," Tech. Rep., Cavendish Laboratory, University of Cambridge, 2006.