SPEECH ANALYSIS OF SUNG-SPEECH AND LYRIC RECOGNITION IN MONOPHONIC SINGING

Dairoku Kawai, Kazumasa Yamamoto, Seiichi Nakagawa

Department of Computer Science and Engineering, Toyohashi University of Technology, Japan
{kawai, kyama, nakagawa}@slp.cs.tut.ac.jp

ABSTRACT

Lyric recognition in singing is challenging because of a number of problems, including a lack of singing databases, superposed musical instruments and different spectral variations. First of all, we investigated the difference of spectral variations among read speech, spontaneous speech and sung speech and we found that sung speech recognition was the most difficult. Next, we consider Japanese lyric recognition in monophonic singing that contains no musical instruments. To express singing well, we use an n-gram language model with a lyrics corpus, singing-adapted acoustic models, and plural pronunciation lexicons for vowel-lengthening. We also compare GMM-HMM and DNN-HMM acoustic models. We obtained a remarkable improvement on lyric recognition in comparison with the baseline system for spontaneous speech recognition.

***Index Terms*— Index Terms**: lyrics recognition, singing adaptation, vowel-lengthening, GMM-HMM, DNN-HMM

1. INTRODUCTION

The demand for music information retrieval is increasing, especially such desired musical information as singer, genre, melody, and lyrics. Singer identification, genre classification, and melody transcription have also been studied [1, 2, 3, 4].

Lyric recognition in singing is difficult because of several problems. The first is the lack of singing databases. An acoustic model (AM) is created by accumulating a large set of annotated singing data. Few monophonic singing databases are available for research. Previous researches [5, 6, 7, 8, 9, 10, 11, 12] mainly used closed databases. The second is different spectral variations for every phoneme, which are investigated in this paper.

Wang et al. [5] attempted lyric recognition using a read-speech database and achieved high accuracy (93.1%) by imposing a strong language constraint, including testing the texts of lyrics. Sasou et al. [6] dealt with the lack of singing databases by an AM that was trained using a large amount of read-speech data and adapted it using a small amount of sung-speech data. They also achieved high word accuracy (59.3-91.6%) because they used a closed language model. Phoneme lengthening, which occurs frequently in singing, also complicates its recognition. Sasou et al. [7] attempted to remove this lengthened part. Similar studies have also been conducted [9, 8]. In both cases, the language model (LM) was trained using only the lyrics of the test set for recognition improvement. The recognition rate was naturally improved [6, 7, 8] by the closed language model.

For large vocabulary recognition, Mesaros et al. [10] used lyrics LMs and MLLR-adapted AMs and showed phoneme accuracy of 34.9% in male monophonic English singing and word accuracy of 12.4% in male and female monophonic English singing (word accuracy of 16.1% replicating in our database). McVicar et al. [12] leveraged repeated lyric phrases and formed a consensus transcription by integrating the repeated portion of the lyrics. Their system showed phoneme accuracy of 27.0% and word accuracy of 9.5% in male and female monophonic English singing, respectively.

Recently, the Deep Neural Network-Hidden Markov Model (DNN-HMM), instead of the Gaussian Mixture Model-Hidden Markov Model (GMM-HMM), was applied to speech recognition and outperformed the conventional method [13]. The reason is that DNN-HMM is flexible for training a large dimensional multi-frame feature vector.

In this paper, we propose an adapted LM and AM and pronunciation lexicons for lyric recognition in monophonic singing. For AM, we compared the GMM-HMM and DNN-HMM acoustic models, which improved the performance.

The rest of this paper is organized as follows: In Section 2, we describe our acoustic and language database. In Section 3, we examine the acoustic differences among three types of speaking styles: read-speech, spontaneous-speech, and sung-speech. We also examine the acoustic difference of various pitches. In Section 4, we present the details of our approach, and in Section 5, we show its recognition results. We conclude the paper in Section 6.

2. DATABASE

In such publicly available databases [14] and in most commercial music, all the tracks of a song (instruments, vocals, etc.) are combined to one track. However, we want to use only the vocal track in our experiment because this study is the first step for automatic lyric recognition in singing. Since there is no lyrics and sung-speech database for making LM and AM, we constructed a singing and lyrics database for our experiment.

Our constructed database list is shown in Table 1. We collected 130k pieces of lyrics texts uploaded by users in Piapro, a lyrics database [15]. We performed morphological analysis using MeCab [16] to segment the lyric texts into words and estimate their pronunciation. Since such texts include multiple instances of the same music, we checked their first 20 words, and if they were the same, we regarded them as a duplicate text and removed them, leaving only one lyric. The lyric texts were segmented into verses with a new line as a clue; we removed verses that contained foreign words to make a Japanese-only lyrics database.

We collected 14 pieces of commercial Japanese popular music sung by seven male singers; seven pieces of music are for the test set of lyrics recognition and seven are for the speaker adaptation set. These music vocal tracks were extracted by taking the difference between the original sound and the accompaniment track using Utagoe Rip [17]. Their overlapping verses, the non-Japanese verses, the scat (e.g., la la la), and the silent parts were removed by preprocessing. We collected 40 pieces of music sung by 40 males uploaded by

Table 1. Constructed database

(a) Language database

Title	Num. of words
Mainichi newspaper (45 months)	206.7M
Piapro (130k lyrics)	28.6M

(b) Sung-speech database

Title	Num. of music	Num. of speakers	Time length
Test set	7	7	19:01
Speaker adaptation set	7	7	19:53
Singing adaptation set	40	40	1:39:28
Read-speech set for NN	15	7	8:59
Sung-speech set for NN	15	7	25:12

Table 2. Speech DB for acoustic analysis

Title	Details
READ	Read-speech (ASJ+JNAS [18]; 133 speakers, 20k utterances, 33 hours)
SPON	Spontaneous-speech (CSJ [19]; 797 speakers, 222k utterances, 122 hours)
SUNG	Sung-speech (Singing adaptation set; 40 speakers, 40 pieces, 1.7 hours)

users in Piapro for adapting AMs [15]. For the singing adaptation set, these music sounds were preprocessed in the same way as the test set. For training a neural network that transforms read-speech MFCCs to sung-speech MFCCs, we recorded pairs of 15 pieces of the read-speech of lyrics and 15 pieces of the sung-speech of the lyrics from seven people with voice training experience. We aligned the time of the pair utterances on the phoneme-level annotation between the read- and sung-speech set for NN. We also prepared word-level transcriptions of the test set for an evaluation on large vocabulary recognition.

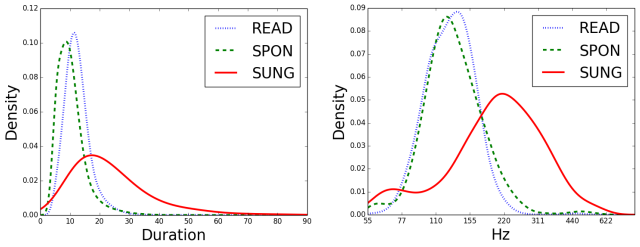
3. ANALYSIS OF SPEAKING STYLE ACOUSTIC DIFFERENCES

To examine the difficulties in sung-speech recognition, we analyzed 12 dimensional acoustic features (MFCCs) differences among the phonemes of three speech styles: read-speech, spontaneous-speech, and sung-speech. To train the full covariance Gaussian model, we used the database of three speaking styles shown in Table 2.

Figure 1(a) shows the syllable duration distribution. Sung-speech has longer duration syllables than the other speaking styles. Figure 1(b) shows the pitch (or fundamental frequency) distribution. Sung-speech has more various pitches than the other speaking styles. These properties make the sung-speech recognition be difficult.

Table 3 shows the distance between speaking styles, where A_B denotes the distance between A and B. Comparing READ_SPON with READ_SUNG, READ_SPON is smaller than READ_SUNG. This means that SUNG varies greatly from the other speaking styles. Comparing READ_SUNG with SPON_SUNG, SPON_SUNG is smaller than READ_SUNG. This means that when the sung-speech are insufficient, spontaneous-speech might be used instead of sung-speech for sung-speech recognition.

Table 4 shows the Bhattacharyya distance between Gaussians for vowels in each speaking style. The largest distance between vowels is READ, followed by SPONTANEOUS and SUNG. The bigger the



(a) Syllable duration (b) Pitch

Fig. 1. Distributions of syllable duration and pitch (fundamental frequency)

Table 3. Distance between speaking styles

Vowel	READ_SPON	READ_SUNG	SPON_SUNG
a	0.14	0.87	0.68
i	0.14	0.64	0.58
u	0.22	0.70	0.30
e	0.30	0.96	0.53
o	0.22	0.99	0.48
Ave.	0.21	0.83	0.51

distance is, the easier their recognition is. This result shows that it is most difficult to recognize sung-speech.

Splitting SUNG into three pitch ranges : low pitch (55∼174 Hz), middle pitch (174∼311 Hz) and high pitch (311∼600 Hz), we also analyzed the acoustic differences among the phonemes of three pitch ranges. Table 5 shows the Bhattacharyya distance between Gaussians for vowels in each pitch range. By splitting the data into three pitch ranges, the distance between vowels got larger than before (compared to Table 4(c)). This indicates that the pitch information has effective information for sung-speech recognition.

4. LYRICS RECOGNITION SYSTEM

4.1. N-gram Language model

We used Palmkit [20] to make word-based n-gram LMs and Witten-Bell smoothing for insufficient n-grams. The vocabulary of these LMs is restricted to the top 20k frequent appearance words. We compared LMs based on newspaper and lyrics corpora.

4.2. GMM-HMM

GMM-HMM consists of 116 context-independent syllables having five states and 64 mixture components/state. We did not use the context-dependent model because context-independent model showed better performance after being adapted to sung-speech. We used GMM-HMM-based AMs consisting of 39 dimensional features: 12 MFCCs, 12 delta MFCCs, 12 delta-delta MFCCs, log energy, delta log energy, delta-delta log power. To investigate the efficiency of pitch information, we also attempted 40 dimensional features containing the above 39 dimensional features and a pitch feature. The pitch feature is normalized in from -0.5 to 0.5 corresponding to fundamental frequency from 55 Hz to 800 Hz.

We adapted the read-speech AMs to sung-speech AMs. One is MAP adaptation, and the other is a neural network-based feature transformation. For comparison with conventional approach [10], we also attempted MLLR adaptation.

Table 4. Distance between vowels in each speaking style

(a) READ

	i	u	e	o	Ave.
a	1.28	1.49	1.37	1.61	1.44
i		0.96	0.30	2.80	1.33
u			1.00	0.82	1.07
e				2.94	1.40
o					2.04
Ave.					1.46

(b) SPONTANEOUS

	i	u	e	o	Ave.
a	1.07	1.17	1.20	1.08	1.13
i		0.54	0.31	1.55	0.87
u			0.75	0.79	0.81
e				1.60	0.97
o					1.26
Ave.					1.01

(c) SUNG

	i	u	e	o	Ave.
a	1.10	0.71	0.53	0.62	0.74
i		0.88	0.47	1.44	0.97
u			0.72	0.63	0.74
e				1.07	0.70
o					0.94
Ave.					0.82

Table 5. Distance between vowels in each pitch range

(a) $\mathrm{SUNG}_{\mathrm{low}}(55 \sim 174\mathrm{Hz})$

	i	u	e	o	Ave.
a	1.66	1.06	1.00	1.07	1.20
i		0.82	0.55	1.49	1.13
u			0.75	0.89	0.88
e				1.60	0.98
o					1.26
Ave.					1.09

(b) $\mathrm{SUNG}_{\mathrm{mid}}(174 \sim 311\mathrm{Hz})$

	i	u	e	o	Ave.
a	1.91	1.62	0.81	1.05	1.35
i		1.12	0.57	1.64	1.31
u			0.98	0.92	1.16
e				1.29	0.91
o					1.22
Ave.					1.19

(c) $\mathrm{SUNG}_{\mathrm{high}}(311 \sim 600\mathrm{Hz})$

	i	u	e	o	Ave.
a	1.91	1.02	1.12	0.91	1.24
i		1.23	0.65	2.19	1.50
u			1.14	0.95	1.08
e				1.93	1.21
o					1.49
Ave.					1.30

4.2.1. MAP adaptation

MAP adaptation [21] maximizes the *a posteriori* distribution of the singing for the given *a priori* model parameters trained by read-speech. We first trained AMs using a large amount of read-speech and adapted them to sung-speech using a small amount of sung-speech.

4.2.2. Speech to singing transformation

From the results of Table 3, we found a larger distance between the read-speech and the sung-speech than between the read-speech and the spontaneous-speech. This means that read-speech and spontaneous speech are relatively similar. On the other hand, sung-speech is very different from the other two types of speaking styles. Thus, we transformed the read-speech MFCCs to sung-speech MFCCs using a neural network trained by pairs of read-speech MFCCs and sung-speech MFCCs. The neural network consisted of four layers; the input layer has 12 inputs (12 MFCCs), the two hidden layers have 24 sigmoid units each, and the output layer has 12 linear units. The input and hidden layers have one bias unit each. We used pairs of read-speech MFCCs and sung-speech MFCCs that were uttered by the same speaker and that appeared in the same context. The pairs are aligned by using dynamic time warping. After transforming the read-speech data using the neural network, the AM was trained using the transformed data.

4.3. DNN-HMM

The DNN consists of five layers; the input layer has 429 units, the three hidden layers have 1024 rectified linear units each, and the output layer has 580 units. The inputs are 11 frames of 39 dimensional features: 12 MFCCs, 12 delta MFCCs, 12 delta-delta MFCCs, log energy, delta log energy, and delta-delta log power. We also used 40 dimensional features with an additional pitch feature.

The number of output layer units corresponds to the context independent acoustic states of the HMMs: five states $\times$ 116 syllables. The DNN was trained by fine-tuning without pre-training.

To make a singing-adapted DNN, we simply trained the DNN using a large amount of spontaneous-speech data and a small amount of sung-speech data.

4.4. Pronunciation dictionary

Insertion errors caused by phoneme lengthening often appear in the form of consecutive vowels. We added modified pronunciation to the pronunciation lexicon to capture longer uttered vowel. For instance, the word "cho u cho (butterfly)" can be extended to "cho u cho", "cho o u cho", "cho u u cho", ..., "cho o u u cho o". In this case, the number of extended pronunciations was eight ($= 2^3$).

5. EXPERIMENTS

5.1. Setup

We evaluated our test set that contained seven pieces of commercial music sung by seven males, as explained in Section 2. The test set is manually classified into four classes on the basis of whether containing reverberation or not and whether containing chorus or not (Table 6). These effects are present in many popular music.

We used 39 dimensional features extracted from 25 ms Hamming-windowed frames with a 10 ms frame shift. We also used 40 dimensional features with an additional pitch feature. The GMM-HMM baseline AM was trained using the Corpus of Spontaneous Japanese (CSJ), which contained 222k utterances (121 hours) [19]. The DNN was also trained using CSJ. Mesaros et al. transformed all the speech acoustic models by MLLR-adaptation using ten pieces of sung-speech data [10]. To compare their approach with our method, we also prepared MLLR-adapted AMs transformed by a matrix that was trained using singing adaptation set. The MAP adaptation data were singing adaptation set, as explained in Section 2. The neural

Table 6. Test set classification (Reverberation/Chorus)

	noR/noC	noR/C	R/noC	R/C
Time length	2:38	6:31	4:33	4:01

Table 7. N-gram LM performance; LM_N=N-gram LM

LM_N	OOV rate[%]					Perplexity				
	noR noC	noR C	R noC	R C	All	noR noC	noR C	R noC	R C	All
$News_3$	14.9	12.4	15.3	11.4	13.3	97	307	160	192	198
$News_4$	14.9	12.4	15.3	11.4	13.3	108	333	178	192	213
$Lyrics_3$	1.3	1.5	2.3	3.2	2.0	57	129	154	114	113
$Lyrics_4$	1.3	1.5	2.3	3.2	2.0	60	145	170	114	122

network that transformed the read-speech MFCCs to sung-speech MFCCs was trained using 15 pairs of read- and sung-speech lyrics, as explained in Section 2. After transforming the read-speech data using the neural network, the AM was trained using this transformed training data. The word-based n-gram LMs were trained using a news corpus and a lyrics corpus, as explained in Section 2. The vocabulary of these LMs was restricted to the top 20k frequent appearance words. The LM weight on the decoder was chosen from 1, 10, 15, 20, 25, and 30. An insertion penalty was chosen from -30, -20, -10, and 0.

5.1.1. Evaluation of LM

Table 7 lists the perplexity and the OOV rate for the two models using transcriptions of the test set. The word-based 3-gram LM trained by the lyrics database has a 2% OOV rate, perplexity of 113, and better performance on test set All. This perplexity might be a sufficient result because the perplexity of a text of newspapers is about 50-100 on an LM trained by newspapers [22].

5.1.2. Evaluation of large vocabulary recognition

Table 8 shows the results of LVCSR on the test set. We used the word-based 3-gram LMs trained using newspaper and lyrics texts. The AM of "GMM" denotes the GMM-HMM and "DNN" denotes the DNN-HMM. The subscript means "training data". "SPON" denotes spontaneous-speech [19]. "SUNG" denotes singing adaptation set. "SPK" denotes speaker adaptation set. "NN" denotes voice converted spontaneous-speech. In the following experiments we use unclassified test set All for a comparison of accuracy unless otherwise mentioned. Our baseline system only showed accuracy of 3.8%. The system using the lyrics LMs showed 4.9% absolute improvement compared to the system using the newspaper LMs. The MLLR singing-adapted model, which corresponds to the previous research [10], showed accuracy of 16.1%. The system using an extended pronunciation dictionary showed improvement compared to the system using only the original pronunciation (8.7% $\rightarrow$ 11.9%).

The neural network based transformation model ($\mathrm{GMM_{NN}}$) showed better performance than the base model ($\mathrm{GMM_{SPON}}$). Furthermore, by adapting to singing and speaker, the NN model performed better and showed accuracy of 29.2% ($\mathrm{NN + MAP_{SUNG} + MAP_{SPK}}$). The GMM model with pitch feature showed better performance than the model without pitch feature (14.6% $\rightarrow$ 16.0%). The DNN model with pitch feature showed slightly better performance than the model without pitch feature. In the test set case without reverberation and chorus, in particular, the addition of pitch information was effective. The DNN-HMM ($\mathrm{DNN_{SPON}}$) showed accuracy of 21.7%. By adding 40 pieces of music data to the training data of DNN-HMM, $\mathrm{DNN_{SPON+SUNG}}$ showed better accuracy

Table 8. Word Accuracy of LVCSR [%]

AM	noR noC	noR C	R noC	R C	All
LM=$News_3$, Lexicon=Original, Feature=Base					
$\mathrm{GMM_{SPON}}$	4.5	3.2	4.7	6.6	3.8
LM=$Lyrics_3$, Lexicon=Original, Feature=Base					
$\mathrm{GMM_{SPON}}$	20.9	8.4	7.5	6.2	8.7
+$\mathrm{MLLR_{SUNG}}$	35.2	15.6	11.3	14.8	16.1
+$\mathrm{MAP_{SUNG}}$	37.3	26.2	14.2	26.2	24.3
LM=$Lyrics_3$, Lexicon=Extension, Feature=Base					
$\mathrm{GMM_{SPON}}$	33.6	10.3	9.7	10.5	11.9
+$\mathrm{MAP_{SUNG}}$	42.6	26.0	16.4	27.2	26.6
$\mathrm{GMM_{NN}}$	30.7	11.7	11.3	13.1	13.7
+$\mathrm{MAP_{SUNG}}$	43.0	29.8	19.8	25.9	28.1
+$\mathrm{MAP_{SUNG}}$+$\mathrm{MAP_{SPK}}$	47.1	28.5	19.2	30.2	29.2
LM=$Lyrics_3$, Lexicon=Extension Feature=Base or Base+Pitch					
$\mathrm{GMM_{SUNG}}$ (Base)	25.9	15.9	11.9	15.1	14.6
$\mathrm{GMM_{SUNG}}$ (+Pitch)	29.1	17.4	10.7	15.7	16.0
LM=$Lyrics_3$, Lexicon=Extension, Feature=Base$\times$11 frames					
$\mathrm{DNN_{SPON}}$	51.2	18.9	20.1	19.7	21.7
$\mathrm{DNN_{SPON+SUNG}}$	55.7	31.2	19.8	34.8	32.2
$\mathrm{DNN_{SPON+SUNG+SPK}}$	56.6	32.6	20.8	39.1	32.5
$\mathrm{DNN_{SPON+SUNG+NN}}$	**59.0**	30.5	21.4	35.7	32.9
LM=$Lyrics_3$, Lexicon=Extension Feature=Base$\times$11 frames or Base$\times$11 frames+Pitch					
$\mathrm{DNN_{SUNG}}$ (Base$\times$11)	46.7	31.5	15.1	33.4	30.4
$\mathrm{DNN_{SUNG}}$ (+Pitch)	48.8	31.1	14.8	34.1	30.9

Table 9. Accuracy of syllable recognition using syllable-based lyrics 3-gram LMs and extended pronunciation dictionary [%]

AM	Syllable	Phoneme
$\mathrm{GMM_{SPON}}$	30.5	45.8
$\mathrm{GMM_{NN}}$+$\mathrm{MAP_{SUNG}}$+$\mathrm{MAP_{SPK}}$	51.1	63.7
$\mathrm{DNN_{SPON+SUNG+SPK}}$	51.2	62.1

of 32.2%. Furthermore, the $\mathrm{DNN_{SPON+SUNG+SPK}}$ showed better accuracy of 32.5%. The $\mathrm{DNN_{SPON+SUNG+NN}}$ showed the best accuracy of 32.9%. For the test set case without reverberation and chorus, it achieved the word accuracy of 59.0%, which is the best among all published papers.

We also conducted syllable recognition experiments using the extended pronunciation dictionary and the syllable-based lyrics 3-gram LMs (see Table 9). In this experiment, $\mathrm{DNN_{SPON+SUNG+SPK}}$ and $\mathrm{GMM_{NN} + MAP_{SUNG} + MAP_{SPK}}$ showed comparable syllable accuracy and phoneme accuracy, respectively.

6. CONCLUSION

By analyzing the acoustic differences among three speaking styles, we showed that sung-speech recognition is the most difficult. To improve lyric recognition, we proposed a singing-adapted LM and AM as well as a pronunciation dictionary. The word-based 3-gram LM trained using a Japanese pop music lyrics corpus showed a 2% OOV rate and perplexity of 113, and outperformed the LMs trained by a newspaper corpus on Japanese. MAP adaptation using singing data improved the recognition performance. The transformation by a neural network followed by MAP adaptation outperformed a traditional MAP adaptation. DNN-HMM showed better performance than GMM-HMM. The best result of our proposed models showed a word accuracy of 59.0%. To the best of our knowledge, this accuracy is the best among all published papers.

7. REFERENCES

[1] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Singer identification based on accompaniment sound reduction and reliable frame selection," in *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, 2005, pp. 329–336.

[2] C.-Y. Sha, Y.-H. Yang, Y.-C. Lin, and H. H. Chen, "Singing voice timbre classification of chinese popular music.," in *ICASSP*. 2013, pp. 734–738, IEEE.

[3] D. Giannoulis, E. Benetos, A. Klapuri, and M. D. Plumbley, "Improving instrument recognition in polyphonic music through system integration," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 5222– 5226.

[4] E. Benetos, S. Ewert, and T. Weyde, "Automatic transcription of pitched and unpitched sounds from polyphonic music," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, 2014, pp. 3131–3135.

[5] C. Kai Wang, R.-Y. Lyu, and Y.-C. Chiang, "An automatic singing transcription system with multilingual singing lyric recognizer and robust melody tracker.," in *INTERSPEECH*. 2003, pp. 1197–1200, ISCA.

[6] A. Sasou, M. Goto, S. Hayamizu, and K. Tanaka, "An arhmm-based speech analysis method and an evaluation of a singing-voice recognition," *Report of IEICE. SP, Speech*, vol. 105, no. 199, pp. 19–24, jul 2005, (in Japanese).

[7] A. Sasou and M. Goto, "Japan patent, jp4576612b," 2007, (in Japanese).

[8] H. Fujihara, M. Goto, J. Ogata, K. Komatani, T. Ogata, and H.G. Okuno, "Automatic synchronization between lyrics and music cd recordings based on viterbi alignment of segregated vocal signals," in *ISM'06. Eighth IEEE International Symposium on Multimedia*, Dec 2006, pp. 257–264.

[9] T. Hosoya, M. Suzuki, A. Ito, S. Makino, L. A. Smith, D. Bainbridge, and I. H. Witten, "Lyrics recognition from a singing voice based on finite state automaton for music information retrieval," in *Proc. ISMIR*, 2005, pp. 532–535.

[10] A. Mesaros and T. Virtanen, "Automatic recognition of lyrics in singing," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, no. 1, pp. 546047, 2010.

[11] A. Mesaros, "Singing voice identification and lyrics transcription for music information retrieval invited paper," in *Speech Technology and Human - Computer Dialogue (SpeD), 2013 7th Conference on*, Oct 2013, pp. 1–10.

[12] M. McVicar, D. Ellis, and M. Goto, "Leveraging repetition for improved automatic lyric transcription of popular music," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 3141–3145.

[13] H. Seki, K. Yamamoto, and S. Nakagawa, "Comparison of syllable-based and phoneme-based dnn-hmm in japanese speech recognition," in *Advanced Informatics: Concept, Theory and Application (ICAICTA), 2014 International Conference of*, Aug 2014, pp. 249–254.

[14] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "Rwc music database: Popular, classical, and jazz music databases," in *Proc. 3rd International Conference on Music Information Retrieval*, 2002, pp. 287–288.

[15] Crypton Future Media, "Piapro," http://piapro.jp/, (accessed 24th Spt 2014).

[16] T. KUDO, "Mecab : Yet another part-of-speech and morphological analyzer," *http://mecab.sourceforge.net/*, 2005.

[17] TODAKEN, "Utagoe lip.," http://www.vector.co.jp/soft/win95/art/se127635.html, (accessed 24th Spt 2014).

[18] K. Itou, M. Yamamoto, K. Takeda, T Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "The design of the newspaper-based Japanese large vocabulary continuous speech recognition corpus," in *Proc. Int. Conf. on Spoken Language Processing*, 1998, pp. 722–725.

[19] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of japanese.," in *Proc. 2nd LREC*. 2000, pp. 947–952, European Language Resources Association.

[20] A. Ito, "Palmkit," http://palmkit.sourceforge.net/, (accessed 11th Nov 2014).

[21] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*, Cambridge University Engineering Department, Cambridge, UK, 2006.

[22] W. Naptali, M. Tsuchiya, and S. Nakagawa, "Topic-dependent-class-based n -gram language model," in *Audio, Speech, and Language Processing, IEEE Transactions on, vol. 20, no.5*, July 2012, pp. 1513–1525.