TRANSCRIBING LYRICS FROM COMMERCIAL SONG AUDIO: THE FIRST STEP TOWARDS SINGING CONTENT PROCESSING

Che-Ping Tsai, Yi-Lin Tuan*, Lin-shan Lee*

National Taiwan University
Department of Electrical Engineering
r06922039@ntu.edu.tw,b02901048@ntu.edu.tw, lslee@gate.sinica.edu.tw

ABSTRACT

Spoken content processing (such as retrieval and browsing) is maturing, but the singing content is still almost completely left out. Songs are human voice carrying plenty of semantic information just as speech, and may be considered as a special type of speech with highly flexible prosody. The various problems in song audio, for example the significantly changing phone duration over highly flexible pitch contours, make the recognition of lyrics from song audio much more difficult. This paper reports an initial attempt towards this goal. We collected music-removed version of English songs directly from commercial singing content. The best results were obtained by TDNN-BLSTM with data augmentation with 3-fold speed perturbation plus some special approaches. The WER achieved (73.90%) was significantly lower than the baseline (96.21%), but still relatively high.

Index Terms— Lyrics, Song Audio, Acoustic Model Adaptation, Genre, Prolonged Vowels

1. INTRODUCTION

The exploding multimedia content over the Internet, has created a new world of spoken content processing, for example the retrieval[1, 2, 3, 4, 5], browsing[6], summarization[1, 6, 7, 8], and comprehension[9, 10, 11, 12] of spoken content. On the other hand, we may realize there still exists a huge part of multimedia content not yet taken care of, i.e., the singing content or those with audio including songs. Songs are human voice carrying plenty of semantic information just as speech. It will be highly desired if the huge quantities of singing content can be similarly retrieved, browsed, summarized or comprehended by machine based on the lyrics just as speech. For example, it is highly desired if song retrieval can be achieved based on the lyrics in addition.

Singing voice can be considered as a special type of speech with highly flexible and artistically designed prosody: the rhythm as artistically designed duration, pause and energy patterns, the melody as artistically designed pitch contours with much wider range, the lyrics as artistically authored sentences to be uttered by the singer. So transcribing lyrics from song audio is an extended version of automatic speech recognition (ASR) taking into account these differences.

On the other hand, singing voice and speech differ widely in both acoustic and linguistic characteristics. Singing signals are often accompanied with some extra music and harmony, which are noisy for recognition. The highly flexible pitch contours with much wider range[13, 14], the significantly changing phone durations in songs, including the prolonged vowels[15, 16] over smoothly varying pitch contours, create much more problems not existing in speech. The falsetto in singing voice may be an extra type of human voice not present in normal speech. Regarding linguistic characteristics[17, 18], word repetition and meaningless words (e.g.oh) frequently appear in the artistically authored lyrics in singing voice.

Applying ASR technologies to singing voice has been studied for long. However, not too much work has been reported, probably because the recognition accuracy remained to be relatively low compared to the experiences for speech. But such low accuracy is actually natural considering the various difficulties caused by the significant differences between singing voice and speech. An extra major problem is probably the lack of singing voice database, which pushed the researchers to collect their own closed datasets[13, 16, 18], which made it difficult to compare results from different works.

Having the language model learned from a data set of lyrics is definitely helpful[16, 18]. Hosoya et al.[17] achieved this with finite state automaton. Sasou et al.[13] actually prepared a language model for each song. In order to cope with the acoustic characteristics of singing voice, Sasou et al.[13, 15] proposed AR-HMM to take care of the high-pitched sounds and prolonged vowels, while recently Kawai et al.[16] handled the prolonged vowels by extending the vowel parts in the lexicon, both achieving good improvement. Adaptation from models trained with speech was attractive, and various approaches were compared by Mesaros el al.[19].

In this paper, we wish our work can be compatible to more available singing content, therefore in the initial effort we collected about five hours of music-removed version of English songs directly from commercial singing content on YouTube. The descriptive term *"music-removed"* implies the background music have been removed somehow. Because many very impressive works were based on Japanese songs[13, 14, 15, 16, 17], the comparison is difficult. We analyzed various approaches with HMM, deep learning with data augmentation, and acoustic adaptation on fragment, song, singer, and genre levels, primarily based on fMLLR[20]. We also trained the language model with a corpus of lyrics, and modify the pronunciation lexicon and increase the transition probability of HMM for prolonged vowels. Initial results are reported.

2. DATABASE

2.1. Acoustic Corpus

To make our work easier and compatible to more available singing content, we collected 130 music-removed (or vocal-only) English songs from www.youtube.com so as to consider only the vocal line.The music-removing processes are conducted by the video owners, containing the original vocal recordings by the singers and vocal elements for remix purpose. [1]

After initial test by speech recognition system trained with LibriSpeech[21], we dropped 20 songs, with WERs exceeding

* indicates equal contribution.

[1] Samples of our collected data: https://youtu.be/QA6x9MLgsc8

	# songs	# singers	pop	electronic
Training set	95	49	202.2	85.8
Testing set	15	13	20.3	22.0
	rock	hiphop	R&B/soul	total
Training set	51.1	30.0	87.5	271
Testing set	17.7	8.4	9.1	42.8

Table 1. Information of training and testing sets in vocal data. The lengths are all measured in minutes.

95%. The remaining 110 pieces of music-removed version of commercial English popular songs were produced by 15 male singers, 28 female singers and 19 groups. The term *group* means by more than one person. No any further preprocessing was performed on the data, so the data preserves many characteristics of the vocal extracted from commercial polyphonic music, such as harmony, scat, and silent parts. Some pieces also contain overlapping verses and residual background music, and some frequency components may be truncated. Below this database is called **vocal data** here.

These songs were manually segmented into fragments with duration ranging from 10 to 35 sec primarily at the end of the verses. Then we randomly divided the vocal data by the singer and split it into training and testing sets. We got a total of 640 fragments in the training set and 97 fragments in the testing set. The singers in the two sets do not overlap. The details of the vocal data are listed in Table.1.

Because music genre may affect the singing style and the audio, for example, hiphop has some rap parts, and rock has some shouting vocal, we obtained five frequently observed genre labels of the vocal data from wikipedia[22] : pop, electronic, rock, hiphop, and R&B/soul. The details are also listed in Table.1. Note that a song may belong to multiple genres.

To train initial models for speech for adaptation to singing voice, we used 100 hrs of English clean speech data of LibriSpeech.

2.2. Linguistic Corpus

In addition to the data set from LibriSpeech (803M words, 40M sentences), we collected 574k pieces of lyrics text (totally 129.8M words) from *lyrics.wikia.com*, a lyric website, and the lyrics were normalized by removing punctuation marks and unnecessary words (like [CHORUS]). Also, those lyrics for songs within our vocal data were removed from the data set.

3. RECOGNITION APPROACHES AND SYSTEM STRUCTURE

Fig.1 shows the overall structure based on Kaldi[23] for training the acoustic models used in this work. The right-most block is the vocal data, and the series of blocks on the left are the feature extraction processes over the vocal data. Features I, II, III, IV represent four different versions of features used here. For example, Feature IV was derived from splicing Feature III with 4 left-context and 4 right-context frames, and Feature III was obtained by performing fMLLR transformation over Feature II, while Feature I has been mean and variance normalized, etc.

The series of second right boxes are forced alignment processes performed over the various versions of features of the vocal data. The results are denoted as Alignment a, b, c, d, e. For example, Alignment a is the forced alignment results obtained by aligning Feature I of the vocal data with the LibriSpeech SAT triphone model (denoted as Model A at the top middle).

The series of blocks in the middle of Fig.1 are the different versions of trained acoustic models. For example, model B is a

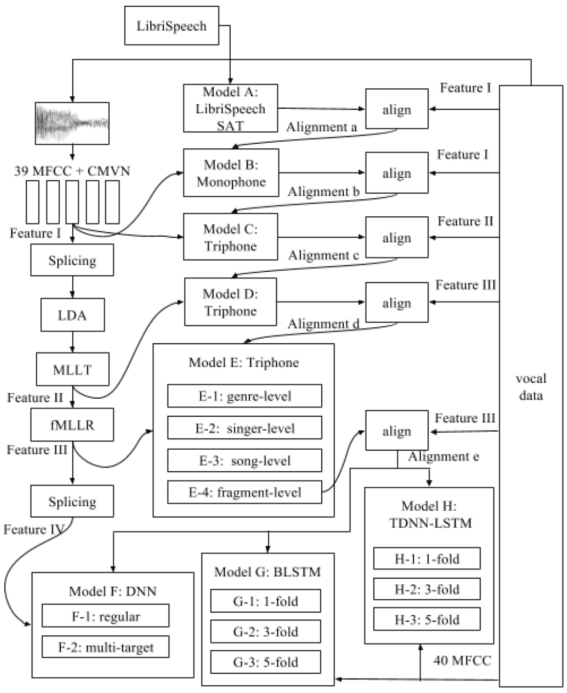

Fig. 1. The overall structure for training the acoustic models.

monophone model trained with Feature I of the vocal data based on alignment a. Model C is very similar, except based on alignment b which is obtained with Model B, etc. Another four sets of Models E, F, G, H are below. For example Model E includes models E-1, 2, 3, 4, Models F,G and H include F-1,2 , G-1,2,3, and H-1,2,3.

We take Model E-4 with fragment-level adaptation within model E as the example. Here every fragment of song (10-35 sec long) was used to train a distinct fragment-level fMLLR matrix, with which Feature III was obtained. Using all these fragment-level fMLLR features, a single Model E-4 was trained with Alignment d. Similarly for Models E-1, 2, 3 on genre, singer and song levels. The fragment-level Model E-4 turned out to be the best in model E in the experiments.

3.1. DNN, BLSTM and TDNN-LSTM

The deep learning models (Models F,G,H) are based on alignment e, produced by the best GMM-HMM model. Models F-1,2 are respectively for regular DNN and multi-target, LibriSpeech phonemes and vocal data phonemes taken as two targets. The latter tried to adapt the speech model to the vocal model, with the first several layers shared, while the final layers separated.

Data augmentation with speed perturbation[24] was implemented in Models G, H to increase the quantity of training data and deal with the problem of changing singing rates. For 3-fold, two copies of extra training data were obtained by modifying the audio speed by 0.9 and 1.1. For 5-fold, the speed factors were empirically obtained as 0.9, 0.95, 1.05, 1.1. 1-fold means the original training data.

Models G-1,2,3 used projected LSTM (LSTMP)[25] with 40 dimension MFCCs and 50 dimension i-vectors with output delay of 50ms. BLSTMs were used at 1-fold, 3-fold and 5-fold.

Models H-1,2,3 used TDNN-LSTM[26], also at 1-fold, 3-fold and 5-fold, with the same features as Model G.

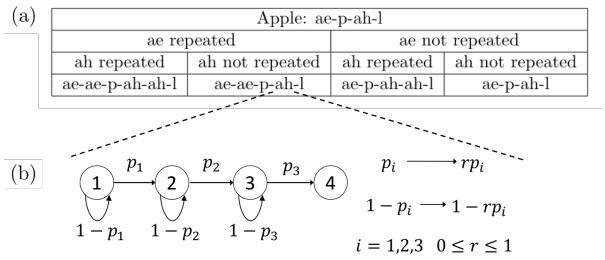

Fig. 2. Approaches for prolonged vowels: (a) extended lexicon (vowels can be repeated or not), (b) increased self-loop transition probabilities (transition probabilities to the next state reduced by r).

3.2. Special Approaches for Prolonged Vowels

Consider the many errors caused by the frequently appearing prolonged vowels in song audio, we considered two approaches below.

3.2.1. Extended Lexicon

The previously proposed approach [16] was adopted here as shown by the example in Fig.2(a). For the word "apple", each vowel within the word (but not the consonants) can be either repeated or not, so for a word with n vowels, 2^n pronunciations become possible. In the experiments below, we only did it for words with $n \leq 3$.

3.2.2. Increased Self-looped Transition Probabilities

This is also shown in Fig.2. Assume an vowel HMM have $m+1$ states (including an end state). Let the original self-looped probability of state i is denoted $1-p_i$ and the probability of transition to the next state is p_i, $i = 1, 2, ..., m$. We increased the self-looped transition probabilities by replacing p_i by rp_i. This was also done for vowel HMMs only but not for consonants.

4. EXPERIMENTS

4.1. Data Analysis

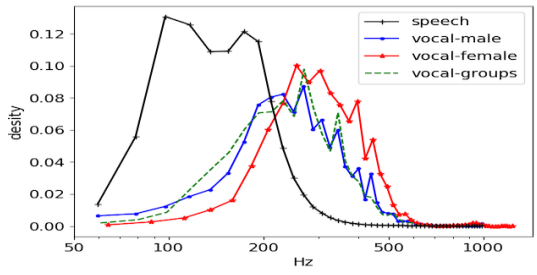

Fig. 3. Histogram of pitch distribution.

4.1.1. Language Model (LM) statistics

We analyzed the perplexity and out-of-vocabulary(OOV) rate of the two language models (trained with LibriSpeech and Lyrics respectively) tested on the transcriptions of the testing set of vocal data. Both models are 3-gram, pruned with SRILM with the same threshold. LM trained with lyrics was found to have a significantly lower perplexity(123.92 vs 502.06) and a much lower OOV rate (0.55% vs 1.56%).

	Acoustic Models	WER(%)	PER(%)
Libri Speech LM	(1) Model A: LibriSpeech(SAT)	96.21	87.17
	(2) Model E-4: fragment-level	88.26	77.18
Lyrics Language Model	(3) Model E-4: fragment-level	80.40	68.80
Extended Lexicon	(4) Model B: Monophone	86.57	76.10
	(5) Model C: Triphone	81.58	71.11
	(6) Model D: Triphone	82.02	72.10
	(7) Model E-4: fragment-level	77.08	66.04
	(8) Model E-4: fragment-level +Increased Trans. Prob.	76.62	65.79
	(9) Model F-1 DNN (regular)	75.56	65.64
	(10) Model F-2 DNN (multi-target)	75.84	65.56
	(11) Model G-1 BLSTM (1-fold)	79.94	70.27
	(12) Model G-2 BLSTM (3-fold)	74.32	63.86
	(13) Model G-3 BLSTM (5-fold)	75.35	65.50
	(14) Model H-1 TDNN-LSTM (1-fold)	79.01	69.20
	(15) Model H-2 TDNN-LSTM (3-fold)	**73.90**	64.33
	(16) Model H-3 TDNN-LSTM (5-fold)	74.53	**63.70**

Table 2. Word error rate (WER) and phone error rate (PER) over the test set of vocal data.

4.1.2. Pitch Distribution

Fig.3 depicts the histogram for pitch distribution for speech and different genders of vocal. We can see the pitch values of vocal are significantly higher with a much wider range, and female singers produce slightly higher pitch values than male singers and groups.

4.2. Recognition Results

The primary recognition results are listed in Table.2. Word error rate (WER) is taken as the major performance measure, while phone error rate (PER) is also listed as references. Rows (1)(2) on the top are for the language model trained with LibriSpeech data, while rows (3)-(16) for the language model trained with lyrics corpus. In addition, in rows (4)-(16) the lexicon was extended with possible repetition of vowels as explained in subsection 3.2.1. Rows (1)-(8) are for GMM-HMM only, while rows (9)-(16) with DNNs, BLSTMs and TDNN-LSTMs.

Row(1) is for Model A in Fig.1 taken as the baseline, which was trained on LibriSpeech data with SAT, together with the language model also trained with LibriSpeech. The extremely high WER (96.21%) indicated the wide mismatch between speech and song audio, and the high difficulties in transcribing song audio. This is taken as the baseline of this work. After going through the series of Alignments a, b, c, d and training the series of Models B, C, D, we finally obtained the best GMM-HMM model, Model E-4 in Model E with fMLLR on the fragment level, as explained in section 3 and shown in Fig.1. As shown in row(2) of Table.2, with the same LibriSpeech LM, Model E-4 reduced WER to 88.26%,

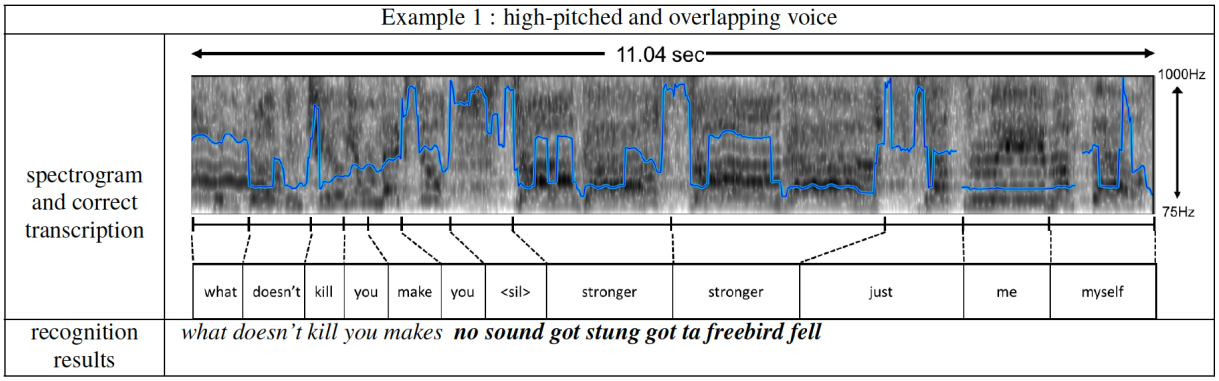

Fig. 4. Sample recognition errors produced by Model E-4 : fragment-level in row(7) of Table.2.

and brought an absolute improvement of 7.95% (rows (2) vs. (1)), which shows the achievements by the series of GMM-HMM alone. When we replaced the LibriSpeech language model with Lyrics language model but with the same Model E-4, we obtained an WER of 80.40% or an absolute improvement of 7.86% (rows (3) vs. (2)). This shows the achievement by the Lyrics language model alone.

We then substituted the normal lexicon with the extended one (with vowels repeated or not as described in subsection 3.2.1), while using exactly the same model E-4, the WER of 77.08% in row (7) indicated the extended lexicon alone brought an absolute improvement of 3.32% (rows (7) vs. (3)). Furthermore, the increased self-looped transition probability ($r = 0.9$) in subsection 3.2.2 for vowel HMMs also brought an 0.46% improvement when applied on top of the extended lexicon (rows (8) vs. (7)). The results show that prolonged vowels did cause problems in recognition, and the proposed approaches did help.

Rows (4)(5)(6) for Models B, C, D show the incremental improvements when training the acoustic models with a series of improved alignments a, b, c, which led to the Model E-4 in row (7). Some preliminary tests with p-norm DNN with varying parameters were then performed. The best results for the moment were obtained with 4 hidden layers, 600 and 150 hidden units for p-norm nonlinearity[27]. The result in rows (9) shows absolute improvements of 1.52% (row (9) for Model F-1 vs. row (7)) for regular DNN. Rows(10) is for Models F-1 DNN (multi-target).

Rows (11)(12)(13) show the results of BLSTMs with different factors of data augmentation described in 3.1. Models G-1,2,3 used three layers with 400 hidden states and 100 units for recurrent and projection layer, however, since the amount of training data were different, the number of training epoches were 15, 7 and 5 respectively. Data augmentation brought much improvement of 5.62% (rows (12) v.s.(11)), while 3-fold BLSTM outperformed 5-fold by 1.03%. Trend for Model H (rows (14)(15)(16)) is the same as Model G, 3-fold turned out to be the best. Row (15) of Model TDNN-LSTM achieved the lowest WER(%) of 73.90%, with architecture $T^{130}T^{130}L^{130}T^{520}T^{520}L^{130}T^{520}T^{520}L^{130}$, while T^n and L^m denotes that the size of TDNN layer was n and the size of hidden units of forward LSTM was m. The WER achieved here are relatively high, indicating the difficulties and the need for further research.

4.3. Different Levels of fMLLR Adaptation

In Fig.1 Model E includes different models obtained with fMLLR over different levels, Models E-1,2,3,4. But in Table.2 only Model E-4 is listed. Complete results for Models E-1,2,3,4 are listed in Table.3, all for Lyrics Language Model with extended lexicon.

	Acoustic Model	WER(%)	PER(%)
Lyrics Language Model Extended Lexicon	(1) Model E-1, genre-level	84.24	68.92
	(2) Model E-2, singer-level	78.53	68.48
	(3) Model E-3, song-level	78.80	68.24
	(4) Model E-4, fragment-level	**77.08**	**66.04**

Table 3. Model E : GMM-HMM with fMLLR over different levels.

Row (4) here is for Model E-4, or fMLLR over fragment level, exactly row (7) of Table.2. Rows (1)(2)(3) are the same as row (5) here, except over levels of genre, singer and song. We see fragment level is the best, probably because fragment(10-35 sec long) is the smallest unit and the acoustic characteristic of signals within a fragment is almost uniform (same genre, same singer and the same song).

4.4. Error Analysis

From the data, we found errors frequently occurred under some specific circumstances, such as high-pitched voice, widely varying phone duration, overlapping verses (multiple people sing simultaneously), and residual background music.

Figure 4 shows a sample recognition results obtained with Model E-4 as in row(7) of Table.2, showing the error caused by high-pitched voice and overlapping verses. At first, the model successfully decoded the words, *"what doesn't kill you makes"*, but afterward the pitch went high and a lower pitch harmony was added, the recognition results then went totally wrong.

5. CONCLUSION

In this paper we report some initial results of transcribing lyrics from commercial song audio using different sets of acoustic models, adaptation approaches, language models and lexicons. Techniques for special characteristics of song audio were considered. The achieved WER was relatively high compared to experiences in speech recognition. However, considering the much more difficult problems in song audio and the wide difference between speech and singing voice, the results here may serve as good references for future work to be continued.

6. REFERENCES

[1] Lin-shan Lee, James Glass, Hung-yi Lee, and Chun-an Chan, "Spoken content retrieval-beyond cascading speech recognition with text retrieval," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 9, pp. 1389–1420, 2015.

[2] Ciprian Chelba, Timothy J Hazen, and Murat Saraclar, "Retrieval and browsing of spoken content," *IEEE Signal Processing Magazine*, vol. 25, no. 3, 2008.

[3] Martha Larson, Gareth JF Jones, et al., "Spoken content retrieval: A survey of techniques and technologies," *Foundations and Trends® in Information Retrieval*, vol. 5, no. 4–5, pp. 235–422, 2012.

[4] Anupam Mandal, KR Prasanna Kumar, and Pabitra Mitra, "Recent developments in spoken term detection: a survey," *International Journal of Speech Technology*, vol. 17, no. 2, pp. 183–198, 2014.

[5] Hung-Yi Lee and Lin-Shan Lee, "Improved semantic retrieval of spoken content by document/query expansion with random walk over acoustic similarity graphs," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 80–94, 2014.

[6] Lin-shan Lee and Berlin Chen, "Spoken document understanding and organization," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 42–60, 2005.

[7] Sz-Rung Shiang, Hung-yi Lee, and Lin-shan Lee, "Supervised spoken document summarization based on structured support vector machine with utterance clusters as hidden variables.," in *INTERSPEECH*, 2013, pp. 2728–2732.

[8] Hung-yi Lee, Yu-yu Chou, Yow-Bang Wang, and Lin-shan Lee, "Unsupervised domain adaptation for spoken document summarization with structured support vector machine," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8347–8351.

[9] Bo-Hsiang Tseng, Sheng-syun Shen, Hung-Yi Lee, and Lin-Shan Lee, "Towards machine comprehension of spoken content: Initial TOEFL listening comprehension test by machine," *Interspeech 2016*, pp. 2731–2735, 2016.

[10] Wei Fang, Juei-Yang Hsu, Hung-yi Lee, and Lin-Shan Lee, "Hierarchical attention model for improved machine comprehension of spoken content," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 232–238.

[11] Hung-yi Lee, Sz-Rung Shiang, Ching-feng Yeh, Yun-Nung Chen, Yu Huang, Sheng-Yi Kong, and Lin-shan Lee, "Spoken knowledge organization by semantic structuring and a prototype course lecture system for personalized learning," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 5, pp. 883–898, 2014.

[12] Sheng-syun Shen, Hung-yi Lee, Shang-wen Li, Victor Zue, and Lin-shan Lee, "Structuring lectures in massive open online courses (moocs) for efficient learning by linking similar sections and predicting prerequisites.," in *INTERSPEECH*, 2015, pp. 1363–1367.

[13] Akira Sasou, Masataka Goto, Satoru Hayamizu, and Kazuyo Tanaka, "An auto-regressive, non-stationary excited signal parameter estimation method and an evaluation of a singing-voice recognition," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*. IEEE, 2005, vol. 1, pp. I–237.

[14] Dairoku Kawai, Kazumasa Yamamoto, and Seiichi Nakagawa, "Lyric recognition in monophonic singing using pitch-dependent DNN," .

[15] Akira Sasou, "Singing voice recognition considering high-pitched and prolonged sounds," in *Signal Processing Conference, 2006 14th European*. IEEE, 2006, pp. 1–4.

[16] Dairoku Kawai, Kazumasa Yamamoto, and Seiichi Nakagawa, "Speech analysis of sung-speech and lyric recognition in monophonic singing," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 271–275.

[17] Toru Hosoya, Motoyuki Suzuki, Akinori Ito, Shozo Makino, Lloyd A Smith, David Bainbridge, and Ian H Witten, "Lyrics recognition from a singing voice based on finite state automaton for music information retrieval.," in *ISMIR*, 2005, pp. 532–535.

[18] Annamaria Mesaros and Tuomas Virtanen, "Recognition of phonemes and words in singing," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 2146–2149.

[19] Annamaria Mesaros and Tuomas Virtanen, "Adaptation of a speech recognizer for singing voice," in *Signal Processing Conference, 2009 17th European*. IEEE, 2009, pp. 1779–1783.

[20] Mark JF Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.

[21] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5206–5210.

[22] Wikipedia, "Plagiarism — Wikipedia, the free encyclopedia," 2004, [Online; accessed 22-July-2004].

[23] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The Kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number EPFL-CONF-192584.

[24] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio augmentation for speech recognition.," in *INTERSPEECH*, 2015.

[25] Haşim Sak, Andrew Senior, and Françoise Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[26] Vijayaditya Peddinti, Yiming Wang, Daniel Povey, and Sanjeev Khudanpur, "Low latency acoustic modeling using temporal convolution and LSTMs," *IEEE Signal Processing Letters*, 2017.

[27] Xiaohui Zhang, Jan Trmal, Daniel Povey, and Sanjeev Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 215–219.

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

[29] Anna M Kruspe and IDMT Fraunhofer, "Bootstrapping a system for phoneme recognition and keyword spotting in unaccompanied singing," in *17th International Conference on Music Information Retrieval (ISMIR), New York, NY, USA*, 2016.