INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Automatic Lyric Transcription from Karaoke Vocal Tracks: Resources and a Baseline System

Gerardo Roa Dabike, Jon Barker

Department of Computer Science, University of Sheffield, UK
groadabike1,j.p.barker@sheffield.ac.uk

Abstract

Automatic sung speech recognition is a relatively understudied topic that has been held back by a lack of large and freely available datasets. This has recently changed thanks to the release of the DAMP Sing! dataset, a 1100 hour karaoke dataset originating from the social music-making company, Smule. This paper presents work undertaken to define an easily replicable, automatic speech recognition benchmark for this data. In particular, we describe how transcripts and alignments have been recovered from Karaoke prompts and timings; how suitable training, development and test sets have been defined with varying degrees of accent variability; and how language models have been developed using lyric data from the LyricWikia website. Initial recognition experiments have been performed using factored-layer TDNN acoustic models with lattice-free MMI training using Kaldi. The best WER is 19.60% – a new state-of-the-art for this type of data. The paper concludes with a discussion of the many challenging problems that remain to be solved. Dataset definitions and Kaldi scripts have been made available so that the benchmark is easily replicable.

Index Terms: Lyrics, Singing, Speech Recognition, Lyrics Transcription, DAMP.

1. Introduction

Until recently, automatic recognition of sung speech has received little attention from the speech research community. The lack of work in this area is partly due to sung speech recognition applications having less immediate importance, but also – and perhaps not unrelatedly – it is also due to a lack of readily available data. This is unfortunate because, aside from applications in music retrieval and indexing, sung speech recognition is a challenging problem worthy of study in its own right. For example, when singing the clarity of the speech is often of secondary importance, so sung speech can share many of the characteristics of other poorly intelligible speech signals (e.g., dysarthric speech). It can act as a stress test of state-of-the-art acoustic modelling (AM) techniques.

Until recently, sung speech recognition has relied on small datasets, and typically, datasets where the speech has been mixed with musical accompaniment (where the acoustic modelling problem becomes secondary to the more severe challenge of source separation) [1, 2, 3, 4, 5].

The earliest notable work on sung speech recognition is that of Mesaros and Virtanen [1] as recently as 2010. A monophone Gaussian mixture model (GMM)- hidden Markov model (HMM) model was trained on material from the CMU Arctic[1] speech database. The work employed just 30 minutes of manually annotated monophonic singing recordings divided into 49 fragments (19 males and 30 females) with a length between 20 and 30 seconds each fragment. The musical fragments were used for constrained maximum likelihood linear regression (CMLLR) adaptation obtaining over 87% WER for bigram and over 100% error for tri-gram. More recently, Kruspe [2] trained a fully connected DNN-HMM model with three hidden layers on the DAMP multiple Performance [6] solo-singing data set. This dataset was released without transcriptions or timing information demanding an error-prone process of automatic force alignment against lyric texts obtained from Smule website, reporting phonemes error rates of about 77% on large test set. Other recent work [3] has investigated a state-of-the-art time delay neural network - bidirectional long short-term memory (TDNN-BLSTM) model trained on 110 manually annotated vocal-only singing recordings from YouTube, using 39 Mel frequency cepstral coefficients (MFCC) plus deltas features. To compensate for the small amounts of training data, acoustic models were first pre-trained on 100 hours of spoken speech and then adapted. However, the best WERs achieved were 73.09%. Other attempts at sung speech recognition have all resulted in similarly poor recognition performances [4, 5].

The problems of data availability, encountered in these earlier works, have potentially been solved by the release in 2018 of the Stanford Digital Archive of Mobile Performances annotated database DAMP Sing! 300x30x2 [7]. The database is a collection of 18,670 karaoke performances made by users of the Smule mobile karaoke application. This size of this dataset gives it the potential to transform research in this area. This paper reports work undertaken to build a state-of-the-art ASR benchmark from this dataset, including preprocessing and alignment stages, training and test set definitions, language model construction and baseline system building using state-of-the-art acoustic modelling via the Kaldi toolkit [8].

The remainder of the paper is structured as follows. Section 2 describes how the raw data has been processed to construct a balanced ASR task with separate training, development and evaluation data. Section 3 describe the construction of suitable language models. We present the experimental results of a reference ASR system in Section 4. Finally, Section 5, concludes by discusses remaining challenges and priorities for future work.

2. The Sing! dataset ASR task

We first describe the Damp Sing! dataset and then preprocessing steps that have been taken to prepare the data as a speech recognition task.

2.1. The Sing! dataset

DAMP Sing! 300x30x2 (Sing!)[7] is the third karaoke performances database released by Smule[2] and made available in the DAMP repository. Like the first two releases, the data originates

[1] http://www.festvox.org/cmu_arctic/

[2] http://www.smule.com

http://dx.doi.org/10.21437/Interspeech.2019-2378

from Smule's collaborative karaoke mobile application. However, the new release is more suitable for ASR than the previous two. The first release, DAMP Vocal performance (multiple performance), is a large collection of 34,620 interpretations (i.e., performances) covering 302 different songs. However, there is a big imbalance in the number of performances per song, and the data is made hard to use by a lack of timing information to align lyrics to the performance. The second DAMP Vocal performance (balanced) database [9] is an extension of the previous release containing 24,874 interpretations from 5,429 singers but it covers just 14 song arrangements. Although this dataset is acoustically rich, with only 14 songs, it is does not have enough linguistic variability to train robust speech models. It is better suited to studying singer variability than speech recognition.

The latest Smule data release, Sing! provides 18,676 interpretations from 13,154 singers covering 5,690 songs with a equal number of interpretations per gender separated by country of the singers. Further, it also provides the lyric prompts that were presented to the performer along with the prompt timings, thus it is much easier to align the song lyrics to the signals. The data was collected and processed by Smule during the second half of 2017 and released in early 2018, by selecting the two most popular singers (male and female), from the 300 most popular song arrangements, from 30 countries. The popularity of the song arrangements was determined by counting the number of interpretations, and the interpretation popularity was determined by counting numbers of listens and votes cast by users of the Smule app. It can be assumed that the up-voted interpretations are well sung, good quality recordings.

When using the Smule mobile application, users sing along to a karaoke accompaniment track playing on their device. Users will typically use their headphones so that their voice is captured in isolation of the accompaniment. This means the data can be used to study sung speech recognition in isolation of the challenges of musical source separation. Research can instead focus on the challenges of sung speech recognition itself. To test this assumption a sample of 100 randomly selected recordings were previewed. It was found that for 88% of these recordings users were wearing headphones while for 12% they were not as evident by the absence or presence of the accompaniment audio. Of the accompaniment-free data, about 15% had appreciable levels of noise from the environment (i.e., performers were using the application in a noisy location).

2.2. Preprocessing of the prompt data

The Sing! data provides the text prompts that were shown to performers along with the time that they were displayed. For most songs, these are in a convenient utterance-level format, i.e., one prompt and time-stamp per phrase to be sung (typically a single line from the song lyrics). For some songs though, the prompts are presented in the data as a sequence of words and/or syllables with separate timestamps per unit and with no marker to indicate where utterances start and end. To recover the timings for the start of each utterance, we automatically reconstruct utterance-level prompts from these word/syllable level prompts, by matching the words and syllables to the lines of the song lyrics that can be recovered from the Smule website. This is made possible by a unique Smule song label.

As was mentioned above, Sing! is a collection of multi-language recordings from karaoke vocal tracks. For our work we are currently only interested in songs sung in English. The meta-data does not provide a language identifier, so Non-English songs are identified by using the CLD2 Naive Bayesian Classifier trained on text from web pages. Specifically, any song for which less than 60% of the sentences are classified as non-English were removed from the dataset. Inspection showed this process to be robust and after filtering the 18,676 songs in the original release, 4,460 English songs remained.

2.3. Audio alignment and segmentation

The alignment stage aims to produce a sequence of segmented utterances and their corresponding transcription, using the prompt data (words and prompt timing) and the unsegmented audio performance as input.

There are three main challenges for the alignment process. First, there is often a mismatch between the prompted lyrics and the words actually sung by the performer. This occurs because singers will omit, change or insert entire phrases, either by mistake or to generate a personal interpretation. Second, there can be considerable differences between the prompt timings and the onsets of the corresponding utterances. Generally, prompts appear early to allow the singer time to prepare, but the lead time is not always equal. Further, singers may start utterances considerably late if they are not familiar with the song. Finally, there is not a one-to-one correspondence between utterance-level prompts and sung utterances. A continuously sung utterance may span more than one prompt, i.e., there is not a natural pause at the end of every line of a song. This is especially true for experienced singers who know the song lyrics and do not need to pause to read or prepare.

The alignment process attempts to deal with the above challenges using a rule-based algorithm. The algorithm matches a sequence of utterance-level prompts with a sequence of non-silence signal segments extracted from the recordings. Utterance-level prompts are recovered using the process described in the previous section. An end-time is associated with each prompt taken as the start time of the following prompt. The non-silence signal segments are extracted from the signal using a simple energy-based activity detector using an energy envelope produced using the Pydub[3] implementation. The algorithm uses a 20 ms window and a 1 ms frame step and classifies frames as either silence or non-silence according to whether the windows rms energy is lower or higher than -25 decibels (dB) below the maximum signal amplitude. Silences of less than 20 ms (e.g., within word silences) are converted to non-silence. Then all non-silence segments are located (i.e., sequence of non-silence frames bounded by silence). The start and end time of each segment is noted.

In the alignment algorithm we use the start and end times to pair utterances prompts to corresponding signal segments. However, in some cases it is necessary to join two or more non-silence segments to match with a single prompt. This occurs when an utterance has been split by the existence of a small silence (e.g. due to aspiration). In other cases it is necessary to prompt texts to match a single utterance, i.e., when the performer sings more than one line without an intervening pause. To achieve this the algorithm proceeds as follows:

1. Prompts that do not intersect with any existing non-silence segment are discarded (the singer failed to sing the lyric).

2. Non-silence segments that do not intersect with any existing prompts are discarded, (i.e., typically extraneous noise such as coughing).

[3] https://github.com/jiaaro/pydub

3. Wherever more the one non-silence segment intersects with the same prompt, the segments are joined.
4. Wherever more than one prompt intersects with the same non-silence segment, the prompts are joined.
5. If every segment does not intersect with only one prompt return to step 4
6. Non-silence segments are now paired to their intersecting prompt.

After running the algorithm, a sample of 100 segments was examined to evaluate the quality of the alignment. It was found that for 60% the segments were correctly aligned to the prompt, i.e., with correct timings and with prompts that provided the correct transcription. A further 32% were only partially correct. In these cases, the segment-to-prompt association was correct, but in the singer there was the addition, removal or substitution of a word or words with respect to the prompted lyric. In 8% of the segments alignment had totally failed. Typically in these cases prompts were being aligned to segments containing only background noise introduced by failure of the earlier voice activity detection stage.

Due to these imperfections our baseline alignment processing is only used for generating the training data. To ensure that accurate recognition performances can be measured, for the test data, a gold standard has been constructed using human annotators to correct the alignment timings and to re-transcribe the speech.

2.4. Training and test set definition

First, taking advantages of the singer country information, we split the data into three datasets, *DSing1*, *DSing3* and *DSing30*, that progressively introduce performances from a broader set of countries. DSing1, is constructed using the subset of recordings from singers registered as users in Great Britain. DSing3 is constructed from the subset of recordings from singers registered in one of the three native English speaking countries, namely, Great Britain, USA and Australia. Finally, the largest data set, DSing30, is constructed using singers from all 30 countries available in the Sing! dataset. Note, in all cases only the English songs are being used, i.e., DSing30 will contain many recordings sung in English by non-native English speakers.

The data in DSing1 is further split into train, test and development sets including 80%, 10% and 10% of the data respectively. Care has been taken to ensure that the sets are disjoint with respect to both singers and arrangements, i.e., no singer or arrangement seen in one set is seen in any other set. This is made complicated by the many-to-many association between singers and arrangements and some data has to be lost to meet this constraint. Any arrangements occurring in the DSing1 development and test set are removed from DSing3 and DSing30 so that these dataset can be used for training. The final size of each DSing training datasets is presented in table 1. For each training dataset the gender balance is roughly equal.

Table 1: *Description of the DSing training sets.*

Set	Singers	Songs	Utterances	Hours
DSing1	352	434	8,794	15.1
DSing3	1,050	1,343	25,526	44.7
DSing30	3,205	4,324	81,092	149.1

For the construction of gold standard evaluation data, roughly 600 utterances were randomly selected from the data assigned to development and test sets. For these utterances the alignments were corrected by humans and the prompts were replaced with human transcriptions of the words that were actually sung. Utterance were discarded if they were found to be contaminated with the background track (i.e., where users were not wearing earphones) and care was taken to keep a maximum of 20 utterance per speaker. This process resulted in 482 utterances covering 40 speakers (27 female and 13 males) for the development set and 480 utterances covering 43 speakers (30 females and 13 males) for the test set. The gold standard evaluation data is summarised in Table 2.

Table 2: *Description of the hand-corrected 'gold standard' development and test datasets.*

Set	Singers	Songs	Utterances	Hours
dev	40	66	482	0.7
test	43	70	480	0.8

3. Language models

To construct an in-domain language model, we select lyrics from LyricsWikia, a free wiki website that stores the lyrics of about 2 millions songs[4]). The selection is made to match the style of music found on the Smule karaoke app. We first select the lyrics of all songs by all artists featured in the DSing3 training set (LMSmule). We then add lyrics from all artists from the Billboard 'The Hot 100' for the 31st December of the years 2015 to 2018 (LMSmule+).

To avoid the inclusion of the lyrics of the songs in the Sing! test sets we discard arrangements that share more than half of their sentences with one of the test set songs. This filtering is based on content rather than song title because song title in the Smule data and LyricsWikia are not always easily comparable. For example, arrangements in Smule can have a suffix describing some characteristic of the arrangement (e.g., *bohemian rhapsody short version*) which will not match with the official song title. We also took care when selecting songs to avoid songs being included multiple times, as can happen when songs appear on multiple albums covered by different artists.

A series of text normalisation steps was applied to the raw lyrics: numbers are converted to text; non-lyric text is removed (e.g., labels such as 'verse' and 'chorus'); non-ASCII characters are replaced using NFKD unicode normalisation. Some lyrics contain words with atypical spellings where letters have been repeated to indicate that the singer should sustain the sound (e.g. in Celine Dion's 'Love Can Move Mountains' the word *LOOOOOOVE* and in Coldplay's 'How You See The World' the word *YOUOOOOOOOOOOH*). These letter repetitions can be detected and corrected with the use of a dictionary.

The language model lyrics selection process results in a total of 44,287 song lyrics from a list of 456 artists with 125 of those artists featuring in the DSing3 training set. In total there are 1,747,731 lyric lines consisting of 11.5 million tokens and 91,654 unique words. A lexicon of size 28K was defined by selecting the most frequently occurring words. This was found to encompass 92% of the DSing1 training dataset vocabulary and 97% of the development dataset. Pronunciations were obtained

[4] http://lyrics.wikia.com/wiki/Special:Statistics

from the CMU pronunciation dictionary which covered 80% of the words. For the remaining words, pronunciations were automatically generated using the Phonetisaurus G2P [10] toolkit.

A 3-gram and 4-gram MaxEnt LM were built using the SRILM [11] toolkit with dev set perplexities of 103 and 100 respectively, when trained on LMSmule. Retraining on LMSmule+ resulted in lower perplexities of 73 and 60 respectively proving the benefits of expanding the corpus with the Billboard song lyric data. For comparison, the 3-gram full and 4-gram full LMs from LibriSpeech [12] has perplexities of 206 and 196 respectively on Smule data.

4. Experimental results

In this section we presents the results obtained using a GMM-HMM and a factorised TDNN [13] state-of-art acoustic model.

The acoustic features used are 13 MFCC plus delta, delta-delta and energy, with 25 milliseconds frame length and 15 milliseconds of overlapping. We firstly trained a GMM-HMM triphone speaker adapted GMM on top of fMLLR. This model was used to apply a cleanup process (standard in Kaldi[5]) on the acoustic training set to remove bad utterances from the training data, (e.g., filtering out those with an incorrect transcription). This process removed about 10% of the training utterances. Using the 'clean' taining data we train a factorised TDNN with lattice-free MMI [14]. For all experiments, the LM employed is a 3-gram MaxEnt model trained on the LMSmule+ data using a 28K word vocabulary(see Section 3).

In an initial experiment the acoustic model was trained on *clean-100* LibriSpeech training acoustic material (100h of annotated audiobooks) [12] to test the performance of the recogniser when trained on well-labeled, but out-of-domain spoken speech data. We compare this mismatched baseline with results achieved when training on the three Smule Sing! karaoke DSing training datasets described previously, i.e., DSing1, DSing3 and DSing30. For each system, performance is measured using the gold standard dev and test sets described in Section 2.4. Performances in terms of WER are summarised in Table 3.

Table 3: *Dev and test WERs when training on LibriSpeech (LS) or the three DSing datasets for both the GMM-HMM and TDNN-F acoustic models and the 3-gram or 4-gram lyrics LM.*

Train Set	AM	LM	dev	test
LS	GMM	3-gram	87.98	85.09
	TDNN-F	3-gram	71.00	65.27
DSing1	GMM	3-gram	64.65	62.60
	TDNN-F	3-gram	45.90	42.28
	TDNN-F	4-gram	41.24	37.63
DSing3	GMM	3-gram	57.45	54.19
	TDNN-F	3-gram	33.00	28.67
	TDNN-F	4-gram	29.60	24.27
DSing30	GMM	3-gram	52.95	49.50
	TDNN-F	3-gram	26.24	22.32
	TDNN-F	4-gram	**23.33**	**19.60**

The highest WER was obtained when training on LibriSpeech data, this was expected due to the different acoustical nature of the data. With the smallest dataset DSing1 we obtained a 42.3% WER which is comparable to the best results reported for karaoke data [4] but without the need for speech-to-singing adaptation of models trained non-singing material. When expanding the training data to include Australian and US recordings (DSing3) recognition performance improved significantly, falling from 42.3% to 28.7%. Rescoring with the 4-gram LM reduced the WER further to 24.3%. Increasing the training data set size from 15.4 hours to 44.7 hours has proved extremely beneficial, despite the additional 29.3 hours being US and Australian English with potential mismatch to the UK test data. This is possibly because, for native English singers, accent variation is neutralised when singing [15] and there is a tendency to move towards US pronunciation [16].

Expanding the training data further by adding in the extra 150 hours from DSing30 led to a further increase in performance with WERs decreasing from 24.3% to 19.6%. The WER decreases despite the introduction of the greater variability of pronunciations that DSing30 possesses. This might be explained by the tendency of non-native English speakers to neutralise their accent during singing [17, 18].

5. Conclusions and Future Work

This paper has provided a new state-of-the-art open baseline for sung speech recognition by building an easily replicable ASR task from the recently released DAMP Sing! 300x30x2 Karaoke performances vocal tracks database collected and distributed by Smule. We have presented techniques for pre-processing the Smule recordings including song-language classification, energy-based utterance segmentation and segment-to-prompt alignment. Further, we have defined training datasets (DSing1, DSing3 and DSing30) starting with GB English, then adding Australian and US English, and finally adding singers from non English speaking countries. We have also described the steps taken to build an in-domain LM collecting a set of lyrics from LyricsWikia website, obtaining perplexities of 73 and 60 for 3-gram and 4-gram LMs respectively.

We have built a Kaldi-based benchmark system using a state-of-the-art TDNN-F acoustic model trained with LF-MMI. This system has produced WERs of 42.3%, 28.7% and 22.3% when trained with DSing1, DSing3 and DSing30 respectively and using a 3-gram LM. When rescoring with 4-gram LM WERs fall to 37.6%, 24.3% and 19.6% respectively. To the best of our knowledge, the lowest WER 19.6% represents a new state of the art performance for unaccompanied singing recognition.

There is still plenty of scope for further research. WERs of around 20% are high compared with spoken speech tasks, e.g., for LibriSpeech audiobooks [12] lowest WERs are 3.19% [19]; for the TED talks corpus Tedlium [20] the lowest WER is 6.5% [19], and for the WSJ corpus [21] WERs have reached 2.9% [22]. The challenge in sung speech recognition remains the lower intelligibility compared with normal speech [23] (e.g., formants can be hard to determine when F0 is high; consonants may be poorly articulated). New approaches to acoustic modelling may be needed, e.g., conditioning phone discrimination on pitch information, or on musical constraints such as beat or tempo. Other progress may be achieved by making better use of the weakly labelled training data, i.e., approaches that use official song lyrics as a guide rather than as a ground truth.

We are making our system fully open with published Kaldi recipes and scripts for reconstructing the segmentation and datasets in the hope that this groundwork will stimulate the community to make new progress in this challenging area (visit https://github.com/groadabike/Kaldi-Dsing-task).

[5] https://github.com/kaldi-asr/kaldi/blob/master/egs/wsj/s5/steps/cleanup/clean_and_segment_data.sh

6. References

[1] A. Mesaros and T. Virtanen, "Recognition of phonemes and words in singing," in *35th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, pp. 2146–2149.

[2] A. M. Kruspe, "Bootstrapping a system for phoneme recognition and keyword spotting in unaccompanied singing," in *17th International Society for Music Information Retrieval Conference (ISMIR)*, New York, NY, USA, 2016, pp. 358–364.

[3] C.-P. Tsai, Y.-L. Tuan, and L.-S. Lee, "Transcribing Lyrics from Commercial Song Audio: The First Step Towards Singing Content Processing," in *43th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada, 2018, pp. 5749–5753.

[4] D. Kawai, K. Yamamoto, and S. Nakagawa, "Speech analysis of sung-speech and lyric recognition in monophonic singing," in *43rd International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 271–275.

[5] G. Roa, "Automatic Speech Recognition in Music," Unpublished MSc. Dissertation, University of Sheffield, UK, 2016.

[6] Smule Vocal Performances (multiple songs) Dataset, "https://ccrma.stanford.edu/damp/," in *accessed July 2018*.

[7] Smule Sing! 300x30x2 Dataset, "https://ccrma.stanford.edu/damp/," in *accessed September 2018*.

[8] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[9] Smule Vocal Performances (balanced) Dataset, "https://ccrma.stanford.edu/damp/," in *accessed July 2018*.

[10] J. R. Novak, N. Minematsu, and K. Hirose, "Phonetisaurus : Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework," *Natural Language Engineering*, vol. 22, no. 6, pp. 907–938, 2015.

[11] A. Stolcke, "SRILM — An Extensible Language Modeling Toolkit," in *7th International Conference on Spoken Language Processing (ICSLP2002 – INTERSPEECH 2002)*, Denver, Colorado, USA, 2002, pp. 901–904.

[12] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *40th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brisbane, Australia, 2015, pp. 5206–5210.

[13] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks," in *19th conference of the International Speech Communication Association (ISCA – Interspeech 2018)*, Hyderabad, India, 2018, pp. 3743–3747.

[14] D. Povey, V. Peddinti, D. Galvez, P. Ghahrmani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *17th conference of the International Speech Communication Association (ISCA –Interspeech 2016)*, San Francisco, California, USA, 2016, pp. 2751–2755.

[15] A. Gibson, "Production and perception of vowels in New Zealand popular music," MPhil Thesis, Auckland University, New Zealand, 2010.

[16] M. Konert-Panek, "Overshooting americanisation. accent stylisation in pop singing acoustic properties of the bath and trap vowels in focus," *Research in Language*, vol. 15, pp. 371–384, 12 2017.

[17] M. Hagen, J. Kerkhoff, and C. Gussenhoven, "Singing your accent away, and why it works," in *17th International Congress of Phonetic Sciences (ICPhS XVII)*, Hong Kong, 2011, pp. 799–802.

[18] M. Mageau, "Foreign Accents in Song and Speech," MPhil Thesis, Carleton University, Canada, 2016.

[19] K. J. Han, A. Chandrashekaran, J. Kim, and I. R. Lane, "The CAPIO 2017 conversational speech recognition system," *CoRR*, vol. abs/1801.00059, 2017.

[20] A. Rousseau, P. Delglise, and Y. Estve, "TED-LIUM: an Automatic Speech Recognition dedicated corpus," in *8th International Conference on Language Resources and Evaluation (LREC 2012)*, 2012, pp. 125–129.

[21] D. B. Paul and J. M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," in *5th DARPA Speech and Natural Language Workshop*, Harriman, NY, USA, 1992, pp. 357–362.

[22] H. Hadian, H. Sameti, D. Povey, and S. Khudanpur, "End-to-end Speech Recognition Using Lattice-free MMI," in *19th conference of the International Speech Communication Association (ISCA – Interspeech 2018)*, Hyderabad, India, 2018, pp. 12–16.

[23] L. Collister and D. Huron, "Comparison of word intelligibility in spoken and sung phrases," *Empirical Musicology Review*, vol. 3, pp. 109–125, 07 2008.