A. Chowdhury, A. Cozzo, A. Ross, “JukeBox: A Multilingual Singer Recognition Dataset,”
Proc. of Interspeech, (Shanghai, China), October 2020

JukeBox: A Multilingual Singer Recognition Dataset

Anurag Chowdhury, Austin Cozzo, Arun Ross

Michigan State University

{chowdh51, cozzoaus, rossarun}@cse.msu.edu

Abstract

A text-independent speaker recognition system relies on successfully encoding speech factors such as vocal pitch, intensity, and timbre to achieve good performance. A majority of such systems are trained and evaluated using spoken voice or everyday conversational voice data. Spoken voice, however, exhibits a limited range of possible speaker dynamics, thus constraining the utility of the derived speaker recognition models. Singing voice, on the other hand, covers a broader range of vocal and ambient factors and can, therefore, be used to evaluate the robustness of a speaker recognition system. However, a majority of existing speaker recognition datasets only focus on the spoken voice. In comparison, there is a significant shortage of labeled singing voice data suitable for speaker recognition research. To address this issue, we assemble *JukeBox* - a speaker recognition dataset with multilingual singing voice audio annotated with singer identity, gender, and language labels. We use the current state-of-the-art methods to demonstrate the difficulty of performing speaker recognition on singing voice using models trained on spoken voice alone. We also evaluate the effect of gender and language on speaker recognition performance, both in spoken and singing voice data. The complete *JukeBox* dataset can be accessed at http://iprobe.cse.msu.edu/datasets/jukebox.html.

Index Terms: Speaker Recognition, Deep Learning, Singing Voice Dataset

1. Introduction

Speaker recognition entails comparing two audio samples encompassing human voice and determining if the voices pertain to the same individual. A majority of speaker recognition research has focused on modeling the speaker-dependent characteristics from conversational or spoken voice data [1]. However, the spoken voice only exhibits a limited range of possible speaker dynamics [2]. As a result, such speaker recognition systems generalize poorly to a wide variety of speaking styles and vocal effort [3]. The singing voice is one such example of a speaking style [4], where the speaker-dependent voice characteristics depart heavily from the spoken voice of the same speaker. Apart from the perceived differences in intensity, pitch, and timbre, there are also differences in the physiological formation of sung speech [5], especially when considering a trained singer [6]. The different styles of singing further diversify the acoustic differences between spoken and singing speech [7], leading to several challenges for speaker recognition systems. One of the primary challenges of speaker recognition from singing is the increased intra-user variance and decreased inter-user variance due to intentional voice modulation, across a broad acoustic spectrum [2]. In addition, the presence of background music and chorus increases the challenges of the task. Thus, a speaker recognition system's ability to correctly match a singer's voice across multiple songs can be used to assess its robustness.

Table 1: *A list of related music datasets compared to the JukeBox dataset.*

Dataset	Number of Samples	Number of Artists	Label	Raw Audio
UT-Sing [8]	165	33	Singer	Yes
MusiClef [13]	1,355	218	Artist / Group	No
Homburg [14]	1,886	1,463	Artist / Group	Yes
1517-Artists [15]	3,180	1,517	Artist / Group	Yes
Unique [16]	3,115	3,115	Artist / Group	Yes
USPOP [17]	8,752	400	Artist / Group	No
CAL10K [18]	10,271	4,597	Artist / Group	No
MagnaTagATune [19]	16,389	270	Artist / Group	Yes
Codiach [20]	20,849	1,941	Artist / Group	No
FMA [21]	106,574	16,341	Artist / Group	Yes
OMRAS2 [22]	152,410	6,983	Artist / Group	No
MSD [11]	1,000,000	44,745	Artist / Group	No
JukeBox	**7,000**	**936**	**Singer**	**Yes**

However, there appears to be limited amount of work done on this topic. Some of the relevant early literature treat singing voice as a speaking style and cluster it using speaker clustering algorithms [4, 8]. In another work [9], the authors use singing voice to perform speaker recognition; however, no cross-modal experiments were done, i.e. training a model on speaking data and testing on singing data (or vice versa). This work was extended in [10] to evaluate cross-modal speaker recognition; however, poor performance was reported. Notably, the datasets used in [4, 8, 9, 10] were limited to a small set (≤ 50) of speakers.

One key reason behind the underrepresented research focus on speaker recognition from singing voice, i.e., singer recognition, is the lack of sufficient development and evaluation data. A review of currently existing music datasets for research (in Table 1) reveals two relevant datasets: the Million Song Dataset (MSD) [11] and the Free Music Archive (FMA) [12]. MSD contains 1,000,000 songs from 44,745 artists/groups. However, the data is available only in the form of audio features and not raw audio, which forces a speaker recognition algorithm to work with a predetermined feature-set. FMA, on the other hand, contains 106,574 songs from 16,341 artists/groups. Here, the ‘artist/group’ label refers to the associated music group/band and not necessarily the individual singer, who might change over time. For example, both Ozzy Osbourne and Ronnie James Dio have sung songs under the artist label of Black Sabbath, thus making group/band labels unsuitable for training or testing a speaker recognition system.

Therefore, in this work, we assemble *JukeBox*, a singing voice dataset annotated with singer, gender, and language labels for the development and evaluation of speaker recognition methods. In the next few sections, we will describe in detail this dataset, the data collection procedure, several experimental protocols, and analyze the performance of state-of-the-art speaker recognition methods on the dataset.

2. *JukeBox* Dataset

The *JukeBox* dataset contains 467 hours of singing audio data sampled at 16 KHz, downloaded from the Internet Archive (IA) [23]. There is a total of 936 different singers in the dataset, of which 533 are male. Figures 1 and 2 summarize the different languages and the distribution of the length of songs in the *JukeBox* dataset. The songs in the *JukeBox* dataset:

- are sung in 18 different languages, as shown in Figure 1, where almost one-fifth of the singers in the dataset sing in non-English languages (i.e., a language other than English).
- are recorded under a wide variety of acoustic environments and recording apparatus, ranging from highly-constrained studio recording setups to completely-unconstrained live concerts.
- contain multiple singers apart from the person-of-interest (POI), for example, vocal duets with overlapped singing and background chorus.
- contain different types of background music (such as drums, piano, or other instrumentation), thus adding to the difficulty of performing speaker recognition.

2.1. Data collection procedure

The *JukeBox* dataset was assembled as follows.

- **Candidate list creation for artists of interest:** We started by compiling a list of artists from Wikipedia, who were tagged as "singer". This yielded a list of 5,046 artists of interest (AOI) from a variety of languages and genres (such as Pop, R&B, Rock, Jazz, Folk, Classical, etc.), with associated metadata such as country of origin (∼ 18 different countries) and years active.
- **Candidate list creation for songs of interest:** The candidate list for AOI was used to query Spotify's song database [24] to generate a list of 162,311 songs. This list was then cross-referenced against IA's repository to generate a list of downloadable songs of interest (SOI). We chose IA as our audio source due to its (a) large collection of audio, (b) public accessibility, (c) nearly unrestricted download access [25], and (d) re-distribution permission for non-commercial purposes.
- **Downloading songs of interest:** The IA repository often contains multiple copies of a song, differing in their audio duration, recording conditions (such as studio versus live versions), and singers (such as original versus cover artists). We specifically avoided cover artists to remove multiple versions of a song and ensure the correctness of artist labels. A large number of the songs on IA were restricted to 30-second duration due to copyright concerns. We preferred the full duration versions of a song, whenever available. Using these criteria, we downloaded a total of 10,063 SOI for 1,341 AOI.
- **SOI pruning for removing non-singing audios:** Voice Activation Detection (VAD) [26] was used on the SOI to remove silent segments. The VAD processed songs were then manually verified to discard audio files that did not contain singing vocals. Note that the human listeners only listened to 5 equally separated 1-second long audio segments in every song to make their decision. This process ensured a practicable manual verification process of 1,500 hours of audio data.
- **Manual verification of language labels in non-English songs:** Nearly one-fifth of the singers in the *JukeBox* dataset are non-English singers. The language labels originally assumed the non-English singers to sing in a non-English language. However, some of the non-English singers were multilingual, and had songs in the English language as well. Therefore, a secondary manual verification of the dataset was conducted to remove English songs for non-English singers. The resulting 7,000 SOI from 936 AOI form the *JukeBox* dataset.

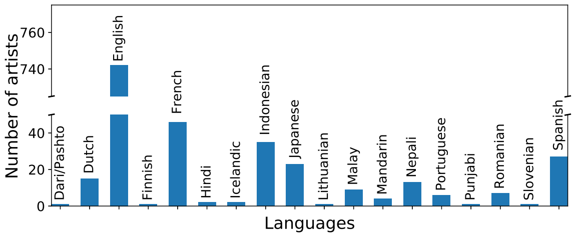


Figure 1: *Distribution of languages in the JukeBox dataset*

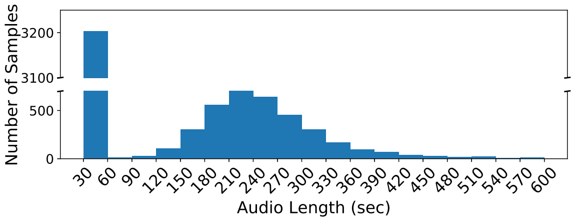


Figure 2: *Distribution of audio length in the JukeBox dataset*

- **Splitting the dataset into the train, test, and auxiliary subsets:** Finally, the set of 936 speakers in the dataset was split into three subsets (shown in Table 2):

– Training set: All speakers with at least three audio samples constitute the training set (670 subjects). This set is reserved for training or fine-tuning speaker recognition models.

– Test set: All speakers with exactly two audio samples constitute the test set (98 subjects). This set is reserved for evaluating trained speaker recognition models on singing voice data.

– Auxiliary set: All speakers with only one audio sample constitute the auxiliary set (168 subjects). This set can be used to augment the training data for speaker recognition models trained in the identification mode. However, the auxiliary set cannot be used to train models in the verification mode, as at least 2 samples per subject are needed to form a genuine pair.

Table 2: *Dataset statistics of the JukeBox dataset*

Dataset	**Train**	**Test**	**Auxillary**
# of Subjects	670	98	168
# of Male Subjects	397	57	79
# of Non-English Subjects	104	21	69
# of Samples	6,636	196	168
# of Hours	385	33	49
Max # of Samples/Speaker	87	2	1
Min # of Samples/Speaker	3	2	1
Avg # of Samples/Speaker	10	2	1

3. Datasets and Experimental Protocols

We propose several experimental protocols for establishing baseline speaker recognition performance on the *JukeBox* dataset. We use state-of-the-art and baseline speaker recognition methods, viz., 1D-Triplet-CNN [27], xVector-PLDA [28], and iVector-PLDA [29] for this purpose. We also evaluate their performance on the *JukeBox* dataset under different conditions based on gender of the artists and language of the songs.

3.1. Datasets

3.1.1. *VoxCeleb2 Dataset*

We use the VoxCeleb2 [30] dataset to perform baseline speaker recognition experiments on spoken voice data (i.e. spoken-to-spoken scenario). We use a subset of the VoxCeleb2 dataset to keep the experiments computationally tractable. A random subset of 5,994 video samples corresponding to the 5,994 celebrities in the VoxCeleb2 dataset forms the training set. Similarly, a random subset of 118 video samples corresponding to 118 celebrities forms the evaluation set. Speech from each video in the dataset is extracted and split into multiple non-overlapping 5-second long audio samples.

3.1.2. *JukeBox Dataset*

Data from *JukeBox* dataset is used to *fine-tune* and evaluate the aforementioned speaker recognition methods on singing voice data (i.e. both spoken-to-singing and singing-to-singing scenarios). Each song in the training set was split into multiple non-overlapping 30-second long segments to increase the number of training samples. In all our experiments, we use the samples from the training set to train the speaker verification algorithms, and the samples from the test set to evaluate the performance of the trained speaker verification models.

3.2. Experimental Protocol

3.2.1. *iVector-PLDA based speaker verification experiments*

We use the MSR Identity Toolkit's [31] implementation of the iVector-PLDA algorithm as our first baseline speaker verification method. A Gaussian-PLDA (gPLDA)-based matcher [31] is used to compare the extracted i-Vector embeddings of a pair of speech samples.

3.2.2. *xVector-PLDA based speaker verification experiments*

We use the PyTorch-based implementation [27] of the xVector algorithm as our second baseline speaker verification method. A gPLDA-based matcher [31] is used to compare the extracted xVector embeddings of a pair of speech samples.

3.2.3. *1D-Triplet-CNN based speaker verification experiments*

We also perform speaker verification experiments using the 1D-Triplet-CNN algorithm, due to its demonstrated robustness to audio degradations [27]. The audio samples in the training set are grouped into triplets to train the 1D-Triplet-CNN algorithm. For evaluation, the audio samples are grouped into pairs and processed by the trained model to generate pairs of 1D-Triplet-CNN embeddings. These pairs of embeddings are then matched using the cosine similarity metric.

3.2.4. *Studying the effect of gender on speaker verification*

The fundamental physiological differences between male and female voices [32] have been used to advocate for their separate treatment in the context of speaker recognition [33]. These differences are further pronounced in the singing voice [34]. Male singers, for example, exhibit a larger variation in their falsetto (a method of voice production) [35], potentially making them harder to recognize than their female counterparts. Therefore, in this work, we perform gender-specific speaker verification experiments (Exp. # 10, 11, 14, 15, 18, and 19 in Table 4) to study the effect of gender on speaker verification from singing voice data. We use the following two types of gender-specific trials in our experiments:

Female only trials: In these experiments, the trained models are evaluated on same-gender (female only) trials drawn from 41 female artists in the test set of the *JukeBox* dataset.

Table 3: *Speaker verification results on spoken voice data from the VoxCeleb2 dataset using the 1D-Triplet-CNN **[M1]**, iVector-PLDA **[M2]**, and xVector-PLDA **[M3]** models. The same models are evaluated on the JukeBox dataset to compare the performance on singing voice data. Here, **P1** = VoxCeleb2 , **P2** = JukeBox , and **P3** = Both VoxCeleb2 and JukeBox together.*

Exp. #	Train Set /Test Set	Models	TMR @FMR=1%	minDCF	EER (in %)
1	P1/P1	M1	91.23	1.82	4.09
2		M2	92.79	1.38	3.81
3		M3	65.06	4.15	7.89
4	P1/P2	M1	**24.72**	**8.35**	26.48
5		M2	18	8.99	**24.49**
6		M3	9.9	9.56	31.83
7	P3/P2	M1	29.71	7.91	24.36
8		M2	**30.98**	**7.77**	**23.63**
9		M3	22.82	8.42	26.39

Table 4: *Verification results on the gender and language specific evaluation subsets of the JukeBox dataset using the 1D-Triplet-CNN **[M1]**, iVector-PLDA **[M2]**, and xVector-PLDA **[M3]** methods. All the models were trained on the VoxCeleb2 dataset and fine-tuned using the JukeBox dataset. Here, **C1** = male speakers only, **C2** = female speakers only, **C3** = English speakers only, and **C4** = non-English speakers only.*

Exp. #	Models	Evaluation Condition	TMR @FMR=1%	minDCF	EER (in %)
10	M1	C1	24.6	8.33	24.44
11		C2	37.29	6.4	21.95
12		C3	31.28	7.67	21.7
13		C4	21.91	8.18	33.63
14	M2	C1	30.64	7.87	26.41
15		C2	30.05	7.58	22.43
16		C3	30.51	7.75	23.67
17		C4	23.53	7.67	28.48
18	M3	C1	20.14	8.57	25.09
19		C2	30.59	7.72	29.29
20		C3	22.88	8.41	24.72
21		C4	21.81	8.44	38.96

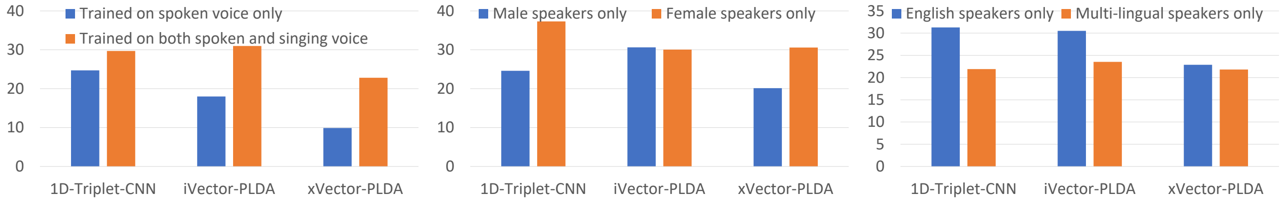
Male only trials: In these experiments, the trained models are evaluated on same-gender (male only) trials drawn from 57 male artists in the test set of the *JukeBox* dataset.

3.2.5. *Studying the effect of language on speaker verification*

Speaker recognition performance of both humans and machines degrade when the speech audio being evaluated is in a language unknown or unfamiliar to the listener [36]. This is also known as the language-familiarity effect (LFE) [37]. In this work, we perform additional speaker verification experiments on the *JukeBox* dataset to evaluate the effect of language on speaker verification performance from singing audio. We perform two different types of language-based speaker verification experiments, given by Exp. # 12, 13, 16, 17, 20, and 21 in Table 4 and described below. All the models in this set of experiments were trained and fine-tuned using the multilingual speech data from the VoxCeleb2 and the *JukeBox* datasets, respectively.

Same language, English only trials: In these experiments, the models are evaluated on same-language (English only) trials drawn from 77 English singers in the test set of *JukeBox*.

Multilingual, non-English trials: In these experiments, the models are evaluated on multilingual trials drawn from 21 non-English singers in the test set of *JukeBox*. The songs in the multilingual trials are sung in one of these 9 different non-English languages: Dari/Pashto, Dutch, French, Japanese, Mandarin, Nepali, Punjabi, Romanian, Spanish.



(a) *Effect of training data on singer verification.* (b) *Effect of gender on singer verification.* (c) *Effect of language on singer verification.*

Figure 3: *Summary of verification performance (TMR@FMR=1%) across different evaluation conditions on the JukeBox dataset.*

Table 5: *Effect of prosody modeling for singing-style based speaker recognition. The 1D-Triplet-CNN + GST model performs singing-style based speaker recognition. The numbers represent performance when trained on the VoxCeleb2 dataset only / on both the VoxCeleb2 and the JukeBox datasets*

Models	TMR@FMR=1%	minDCF	EER (in %)
1D-Triplet-CNN	**24.72/29.71**	**8.35/7.91**	**26.48**/24.36
1D-Triplet-CNN + GST	19.42/26.80	8.78/8.24	26.55/**24.27**

3.2.6. Studying the effect of singing style modeling on speaker verification

Finally, we also perform a fusion of Global Style Token (GST) [38] based prosodic speech features with the 1D-Triplet-CNN based speaker embedding to facilitate singing style modeling for speaker verification. In these experiments, we extract the speaker embeddings obtained from the 1D-Triplet-CNN and input it to GST to extract prosodic speech features. These prosodic speech features are further fused with the 1D-Triplet-CNN based speaker embeddings to derive a style-sensitive speaker embedding. This embedding is then used to perform speaker verification experiments, given in Table 5.

4. Results and Analysis

The results of all the experiments described in Section 3.2 are given in Tables 3, 4, and 5, and Figure 3. For all the speaker verification experiments, we report the True Match Rate at a False Match Rate of 1% (TMR@FMR=1%), minimum Detection Cost Function (minDCF) and Equal Error Rate (EER in %). The minimum Detection Cost Function (minDCF) is computed at a prior probability of 0.01 for the specified target speaker (P_{tar}) with a cost of missed detection of 10 (C_{miss}).

- In the experiments 1 to 3 given in Table 3, baseline speaker verification performance is established for all the models on spoken voice data from the VoxCeleb2 dataset. The relatively lower performance of the xVector-PLDA model is attributed to the limited training data being insufficient for learning xVector-PLDA model's considerably larger parameter space.
- Further, in experiments 1 to 6, a large performance drop is noted across all models when they are evaluated on the *JukeBox* dataset when compared to the VoxCeleb2 dataset. This indicates the difficulty of performing singer recognition using models that are pre-trained on spoken voices.
- Fine-tuning the models pre-trained on the VoxCeleb2 dataset, using the training set of *JukeBox* (in experiments 7 to 9) improved the average performance (TMR@FMR=1%) of all the models by $\sim$ 10.29%. This indicates the benefit of using *JukeBox* for fine-tuning pre-trained speaker recognition models for the task of singer recognition.
- We also performed speaker identification experiments corresponding to the experimental protocol given in Table 3. The identification results follow the trend seen in verification. Best performance is observed when the models are trained and tested on spoken voice. Worst performance is observed when the models are trained on spoken voice and tested on singing voice. Fine-tuning the models trained on spoken voice with singing voice improves the performance on singing voice.
- In the gender-based speaker verification experiments (10, 11, 14, 15, 18, and 19) given in Table 4, majority of the models perform better on female subjects. This is an interesting result because (a) both the VoxCeleb2 and *JukeBox* datasets have a higher proportion of male subjects in the training data, and (b) gender-based speaker recognition experiments on spoken speech data usually perform better for males [32, 33]. This demonstrates the effect of gender-specific voice range profiles of the singing voice [34] in the context of speaker recognition.
- In the language-based speaker verification experiments (12, 13, 16, 17, 20, and 21) given in Table 4, majority of the models perform better on English-only trials. This indicates the presence of the LFE even in singing audios, where the speaker models trained on English-majority speech data performs better on English-only speech data compared to non-English speech.
- The inclusion of prosody modeling for encoding the singing style in the speaker embeddings degrades the speaker verification performance (see Table 5). This can be attributed to the large intra-speaker variance due to different singing styles used in different songs. This indicates that the singing-style of the singer estimated from a fixed set of songs does not generalize well across other songs, leading to a drop in performance.

5. Summary

We assembled a multilingual singer recognition dataset called *JukeBox*. The evaluation of state-of-the-art speaker recognition methods trained only on spoken voice data, on the *JukeBox* dataset, revealed the challenges posed by singing voice data to speaker recognition. The *JukeBox* dataset can be used to address these challenges by facilitating speaker recognition research on singing voice data. Additionally, the dataset is annotated for language and gender labels, which can be used to investigate their effects on singer recognition performance. In the future, we plan to extend this dataset to include spoken voice audios for the singers in the current dataset. This will help us study the relationship between the spoken voice and the singing voice of a subject, in the context of speaker recognition.

6. Acknowledgements

We thank Dr. J. P. Campbell from MIT Lincoln Laboratory for his valuable suggestions. We also thank Jefferson Bailey and Alexis Rossi from the Internet Archive for assisting with data collection for the *JukeBox* dataset.

A. Chowdhury, A. Cozzo, A. Ross, “JukeBox: A Multilingual Singer Recognition Dataset,”
Proc. of Interspeech, (Shanghai, China), October 2020

7. References

[1] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors,” *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

[2] J. Sundberg, “The acoustics of the singing voice,” *Scientific American*, 1977.

[3] E. Shriberg, M. Graciarena, H. Bratt, A. Kathol, S. S. Kajarekar, H. Jameel, C. Richey, and F. Goodman, “Effects of vocal effort and speaking style on text-independent speaker verification,” in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[4] M. Mehrabani and J. H. Hansen, “Singing speaker clustering based on subspace learning in the GMM mean supervector space,” *Speech Communication*, 2013.

[5] R. Daniloff, K. Wolf, G. Larsen, and L. Evans, “Allophonic variation in spoken and sung speech,” *The Journal of the Acoustical Society of America*, 1994.

[6] W. Brown Jr, E. Hunt, and W. N. Williams, “Physiological differences between the trained and untrained speaking and singing voice,” *Journal of Voice*, 1988.

[7] R. Stone, T. Cleveland, J. Sundberg, and J. Prokop, “Aerodynamic and acoustical measures of speech, operatic, and broadway vocal styles in a professional female singer,” *Journal of Voice*, 2003.

[8] M. Mehrabani and J. Hansen, “Speaker clustering for a mixture of singing and reading,” *INTERSPEECH*, 2012.

[9] H. A. Patil, M. C. Madhavi, and N. H. Chhayani, “Person recognition using humming, singing and speech,” in *International Conference on Asian Language Processing*, 2012.

[10] N. H. Chhayani and H. A. Patil, “Development of corpora for person recognition using humming, singing and speech,” in *International Conference Oriental held jointly with Conference on Asian Spoken Language Research and Evaluation*, 2013.

[11] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, “The million song dataset,” *International Society for Music Information Retrieval Conference*, 2011.

[12] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, “FMA: A dataset for music analysis,” in *International Society for Music Information Retrieval Conference*, 2017.

[13] M. Schedl, N. Orio, C. C. Liem, and G. Peeters, “A professionally annotated and enriched multimodal data set on popular music,” in *ACM Multimedia Systems Conference*, 2013.

[14] H. Homburg, I. Mierswa, B. Möller, K. Morik, and M. Wurst, “A benchmark dataset for audio classification and clustering.” in *International Society for Music Information Retrieval Conference*, 2005.

[15] K. Seyerlehner, G. Widmer, and P. Knees, “Frame level audio similarity-a codebook approach,” in *International Conference on Digital Audio Effects*, 2008.

[16] K. Seyerlehner, G. Widmer, and T. Pohle, “Fusing block-level features for music similarity estimation,” *International Conference on Digital Audio Effects*, 2010.

[17] A. Berenzweig, B. Logan, D. Ellis, and B. Whitman, “A large-scale evaluation of acoustic and subjective music similarity measures,” *Computer Music Journal*, 2003.

[18] D. Tingle, Y. E. Kim, and D. Turnbull, “Exploring automatic music annotation with “acoustically-objective” tags,” in *International Conference on Multimedia Information Retrieval*, 2010.

[19] E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie, “Evaluation of algorithms using games: The case of music tagging.” in *International Society for Music Information Retrieval Conference*, 2009.

[20] C. Mckay, “A large publicly accessible prototype audio database for music research,” in *International Society for Music Information Retrieval Conference*, 2006.

[21] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, “FMA: a dataset for music analysis,” in *International Society for Music Information Retrieval Conference*, 2017.

[22] M. Mauch, C. Cannam, M. Davies, S. Dixon, C. Harte, S. Kolozali, D. Tidhar, and M. Sandler, “OMRAS2 metadata project 2009,” in *International Society for Music Information Retrieval Conference*, 2009.

[23] “Internet archive,” https://archive.org, accessed: 2020-03-03.

[24] “Spotify API,” https://developer.spotify.com/documentation/web-api, accessed: 2020-03-04.

[25] “Internet archive API,” https://archive.org/services/docs/api, accessed: 2020-03-04.

[26] “Google WebRTC voice activity detection,” https://webrtc.org, accessed: 2020-03-04.

[27] A. Chowdhury and A. Ross, “Fusing MFCC and LPC features using 1D Triplet CNN for speaker recognition in severely degraded audio signals,” *IEEE Transactions on Information Forensics and Security*, 2020.

[28] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-Vectors: robust DNN embeddings for speaker recognition,” in *International Conference on Acoustics, Speech, and Signal Processing*, 2018.

[29] D. Garcia-Romero and C. Espy-Wilson, “Analysis of i-vector length normalization in speaker recognition systems.” in *INTERSPEECH*, 2011.

[30] J. S. Chung, A. Nagrani, and A. Zisserman, “Voxceleb2: Deep speaker recognition,” in *INTERSPEECH*, 2018.

[31] S. Sadjadi, M. Slaney, and L. Heck, “MSR identity toolbox v1.0: A MATLAB toolbox for speaker-recognition research,” *Speech and Language Processing Technical Committee Newsletter*, 2013.

[32] J. Mason and J. Thompson, “Gender effects in speaker recognition,” *International Conference on Signal Processing*, 1993.

[33] L. Li and T. F. Zheng, “Gender-dependent feature extraction for speaker recognition,” in *China Summit and International Conference on Signal and Information Processing*, 2015.

[34] A. Sulter, H. Schutte, and D. Miller, “Differences in phonetogram features between male and female subjects with and without vocal training,” *Journal of Voice*, 1996.

[35] G. Welch, D. Sergeant, and F. MACCURTAIN, “Some physical characteristics of the male falsetto voice,” *Journal of Voice*, vol. 2, pp. 151–163, 12 1988.

[36] L. Lu, Y. Dong, X. Zhao, J. Liu, and H. Wang, “The effect of language factors for robust speaker recognition,” in *International Conference on Acoustics, Speech, and Signal Processing*, 2009.

[37] D. Fleming, B. L. Giordano, R. Caldara, and P. Belin, “A language-familiarity effect for speaker discrimination without comprehension,” *Proceedings of the National Academy of Sciences*, 2014.

[38] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, “Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis,” in *International Conference on Machine Learning*, 2018, pp. 5180–5189.