



Improving Gender Identification in Movie Audio using Cross-Domain Data

Rajat Hebbar, Krishna Somandepalli, Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory,
Department of Electrical Engineering,
University of Southern California, Los Angeles.
{rajatheb, somandep}@usc.edu, shri@sipi.edu

Abstract

Gender identification from audio is an important task for quantitative gender analysis in multimedia, and to improve tasks like speech recognition. Robust gender identification requires speech segmentation that relies on accurate voice activity detection (VAD). These tasks are challenging in movie audio due to diverse and often noisy acoustic conditions. In this work, we acquire VAD labels for movie audio by aligning it with subtitle text, and train a recurrent neural network model for VAD. Subsequently, we apply transfer learning to predict gender using feature embeddings obtained from a model pre-trained for large-scale audio classification. In order to account for the diverse acoustic conditions in movie audio, we use audio clips from YouTube labeled for gender. We compare the performance of our proposed method with baseline experiments that were setup to assess the importance of feature embeddings and training data used for gender identification task. For systematic evaluation, we extend an existing benchmark dataset for movie VAD, to include precise gender labels. The VAD system shows comparable results to state-of-the-art in movie domain. The proposed gender identification system outperforms existing baselines, achieving an accuracy of 85% for movie audio. We have made the data and related code publicly available¹.

Index Terms: gender identification, voice activity detection, deep neural networks, recurrent neural networks, transfer learning, bi-directional long short-term memory.

1. Introduction and Related Work

In recent years, there has been a growing interest in large-scale automatic analysis of multimedia content such as movies. Audiovisual analysis has enabled us to objectively quantify representations of different characters in a movie along demographics such as gender, age and race [1]. For instance, gender representation in movies has been examined by obtaining the fraction of time female characters are shown on screen (*on-screen time*), and the fraction of time female characters speak in a movie (*speaking time*)².

Precise measurement of on-screen time and speaking time relies on accurate gender identification from video and audio, respectively. Gender classification from face images *in the wild* is possible with an accuracy of over 95% [2]. By comparison, gender identification from audio has also been known to show similar performance [3, 4], although to the best of our knowledge, an accuracy of 71% has been reported on movie audio [5]. Besides obtaining measures such as speaking time, gender identification can also be an important step for several audio tasks

such as speech recognition [6], speaker segmentation [7] and emotion recognition [8].

The task of gender identification from audio in general, has been well studied with published research dating back about three decades [9]. These methods have mostly examined speech acquired in controlled recording conditions (e.g., read speech : WSJ[10] and TIMIT [11]), or telephone speech [12] or radio and outdoor speech [13]. A number of studies have shown that approaches such as Gaussian mixture models (GMM) and support vector machines (SVM; e.g., [14]) with features including pitch and mel-frequency cepstral coefficients (e.g., [3]) are often sufficient to provide accurate gender classification.

Extending the aforementioned approaches to movie audio is non-trivial for two main reasons: 1) voice activity detection (VAD) which is challenging due to the variability in the acoustic conditions, and 2) lack of in-domain data for supervision, i.e., movie audio with precise gender labels. A typical system for gender identification requires VAD, followed by gender classification of the segments identified as speech. VAD is particularly challenging for movie audio because it is heavily *edited* [15]. The final movie audio consists of music tracks, sound effects and ambient noise with speech. Furthermore, the speech sometimes differs from read speech (e.g., shouting, electronic voice). As a result, the acoustic conditions are variable, and possibly non-stationary. Although VAD has been well studied for audio tasks (see [16] for a survey), most off-the-shelf VAD models, including the ones trained for *real-life data* such as [17] fail for movie audio [18].

In order to address the VAD problem for movie audio, we train a recurrent neural network (RNN), with bi-directional long short-term memory (BLSTM) cells [19] using movie audio that are weakly aligned with subtitles (similar to [20]). The primary objective of this work is to identify gender from speech. Although the performance of our model is comparable to that of the state-of-the-art for movie data [18] (see Sec 4), we do not present a detailed error analysis of the VAD aspect of this work.

Obtaining precise³ labels by manually annotating movie audio for gender can be time consuming and arduous. We would also have to sample a large number of movies because on average, a person speaks for only about one-third of the total duration of a movie. Instead, we use human-labeled sound clips of female and male speech, that have been released as part of AudioSet [21]. These clips from YouTube are obtained in varying recording conditions. The variability in the speech quality as well as diverse acoustic environments may be closer to the audio quality in movies compared to read speech or radio data.

An *operational* challenge for gender identification in movies is the lack of data with precise labels for evaluation. To address this, we have manually annotated the gender of the

¹<https://github.com/usc-sail/mica-gender-from-audio>

²<https://seejane.org/research-informs-empowers/data>

³time precision of atleast 20ms

speakers in a benchmark VAD dataset (four Hollywood movies; [18]). These labels have been made publicly available, and can be used for a systematic evaluation of VAD and gender in movie audio⁴

Features known to be discriminative for gender identification (such as pitch) cannot be reliably estimated from noisy speech. In this context, noise-robust feature representation of speech can be a useful tool. Such representations or *embeddings* can be obtained from models that are trained with big data for classifying hundreds of classes. These ‘general-purpose’ audio embeddings when employed for a specific task often come under the paradigm of *transfer learning* [22]. For instance, [23, 24] use transfer learning to leverage large-scale audio data to improve the performance of specific tasks such as ASR and speaker recognition. In our work, we use the embeddings that were developed for a large-scale audio classification task using AudioSet [21, 25].

Based on the aforementioned studies, we hypothesize that gender identification in movie audio can be improved by – 1) using embeddings from large-scale audio classification models as features, and 2) supervised gender classification with cross-domain audio data acquired in diverse acoustic conditions. Our proposed model for gender identification incorporates both aspects. Furthermore, our baseline experiments (Sec. 4) are set up to evaluate these two claims.

In summary, the contributions of our work are as follows: 1) we build a robust VAD for movie audio using weakly aligned subtitle information, 2) our proposed model significantly improves gender identification in movie audio, and 3) we have made the gender labels for a subset of four movies publicly available to facilitate future research.

2. Datasets

2.1. Movie dataset for VAD training

We use a subset of 95 Hollywood movies⁵ released in the year 2014. The dataset was partitioned into 82 movies for training and 13 movies for system development. The development split was chosen to have a sample representative of different movie genres [5].

In order to train our VAD, the speech/non-speech labels were obtained as follows: 1) we first aligned movie audio with subtitles using Gentle⁶. The audio segments which were successfully aligned were labeled as speech; 2) audio segments between consecutive subtitle utterances were labeled as non-speech.

2.2. Movie dataset for gender identification evaluation

We manually annotated the gender of the speakers in a benchmark dataset released in [18]. It consists of four Hollywood movies: ‘Kill Bill 1’ (2003), ‘Saving Private Ryan’ (1998), ‘I Am Legend’ (2007), and ‘The Bourne Identity’ (2002).

First, the subtitle utterances for the four movies were segmented (where necessary) to consist of male or female speech only. We obtained utterance-level gender labels from two independent annotators. All the labels were verified to resolve ties by the author. Finally, the gender labels were mapped to the VAD labels at a precision of 20ms [18]. This resulted in 128.5 and 22.4 minutes of speech from male and female speakers re-

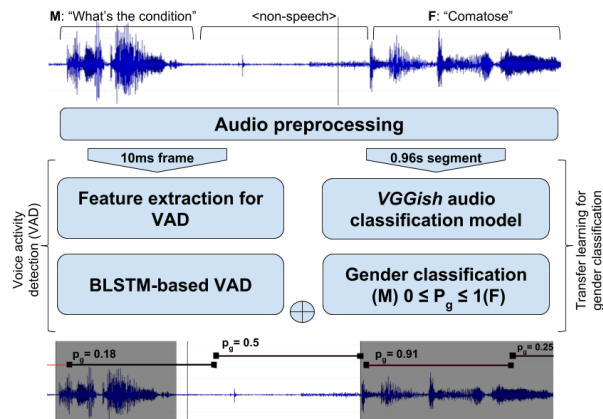


Figure 1: Schematic diagram for our proposed model

spectively from a total audio duration of 8hrs. This allowed us to test the performance of our system at a higher precision than the previous methods which evaluated at the subtitle level [5].

2.3. AudioSet and WSJ corpus

AudioSet [21] is a large corpus of 10s sound clips collected from YouTube for audio event detection (AED). Each clip is manually tagged for at least one of 632 audio events. In our experiments for gender identification, we use those clips tagged either as ‘male speech, man speaking’ or ‘female speech, woman speaking’. Any clips tagged as both male and female speech were excluded. The audio data from these clips was downloaded using *youtube-dl* [26].

This resulted in 47 and 21 hours of speech from male and female speakers respectively. The development set consisted of 60 audio clips for each gender. As mentioned before, since these audio clips were obtained from YouTube, they have diverse recording and acoustic conditions. As such it may account for higher variability (often observed in movie audio) compared to other datasets such as WSJ. We refer to the audio quality of this dataset throughout the paper as ‘diverse’.

We use the WSJ-SI84 dataset, a subset of the WSJ corpus, in order to replicate the experiments performed in [5]. It consists of a total of 15 hours of data with 7184 utterances from 42 male and 41 female speakers. This data consists of read speech was collected in controlled environment. We refer to the audio quality of this dataset throughout the paper as ‘controlled’.

3. Methods

The two main components of our proposed system are: 1) VAD, and 2) transfer learning for gender classification. Fig. 1 shows a schematic of the end-to-end gender identification system. All audio data was resampled to 8kHz, and converted to a single channel. VAD labels were obtained at 10ms frames, and VGGish-embeddings [25] for 0.96s non-overlapping segments, followed by gender classification. Finally, frame-level gender labels were obtained by masking the gender labels at 0.96s scale by the VAD frame-level labels (masking operation is indicated by a \oplus in the Fig. 1)

3.1. Voice activity detection (VAD)

Following empirical observations in [17] we train an RNN-based model. Since it is sufficient to have an offline system, we used both forward and backward context in the form of

⁴<https://goo.gl/forms/tVXE4V52kf7LkCcP2>

⁵<http://www.boxofficemojo.com/yearly/chart/?yr=2014>

⁶<https://github.com/lowerquality/gentle>

Table 1: Performance evaluation of our VAD (ACC: accuracy, PREC: precision, REC: recall, EER: equal error rate)

METHOD	ACC	PREC	REC	EER	F1
openSMILE [27]	0.66	0.41	0.55	0.4	0.45
Lehner et. al.,[18]	0.87	0.75	0.73	0.13	0.74
BLSTM-VAD	0.87	0.80	0.63	0.25	0.70

BLSTM cells in the RNN. Henceforth, we refer to our model as ‘BLSTM-VAD’.

3.1.1. BLSTM-VAD

The VAD model consists of a RNN layer with BLSTM cells followed by a sequence of fully connected (FC) layers with rectified linear unit (ReLU) [28] non-linearity. The final layer has two output nodes for binary classification with softmax activation. We use binary cross-entropy loss and Adam optimizer [29] to train the network. The input to the network is log-Mel features for a 160ms segment with 15 frames as context (=150ms). The output VAD labels are median filtered with a window length of 55 frames.

Network architecture of the VAD model:

INP[23,16] \rightarrow **BLSTM[150]** \rightarrow **FC[256]** \rightarrow **FC[128]** \rightarrow **FC[64]** \rightarrow $\sigma(\mathbf{FC}[2])$,

where **INP** = Input, **FC** = Fully connected layer with relu activation, σ = softmax. Note that in this shorthand notation, $[\cdot]$ indicates the number of nodes in the layer.

3.1.2. Feature extraction for VAD

We use log-Mel filterbank energies (log-Mel features) as input features to the VAD system. Unlike MFCCs which are widely used features for audio tasks, computing log-Mel features does not include the discrete cosine transform (DCT) step. We chose log-Mel features over MFCC in our experiments because the benefit of DCT is mostly compression, and computation power is not an issue. Furthermore, in our preliminary experiments, log-Mel features outperformed MFCC for VAD task.

We extracted 23 log-Mel filterbank coefficients over a window of 25ms with an overlap of 15ms (1 frame=10ms). The features were normalized to have a zero mean and unit variance.

3.2. Transfer learning for gender classification

The objective of the models developed for AED [25] was to classify over six-hundred audio events. The embeddings are features from the penultimate layer of the trained network. They provide noise-robust, low-dimensional representation of an audio segment. Transfer learning deals with adapting these embeddings for a specific task. It is particularly beneficial in the lack of domain-specific training data (which is the case here).

3.2.1. VGGish audio classification model

Convolutional neural networks (CNN) have proven very effective for computer vision tasks. VGG [30] is a widely used network for object recognition in images. Similar network architectures have been shown to perform effectively for audio segment classification [25]. A smaller, modified version trained for the AED task has been made publicly available⁷. We refer to this model as ‘VGGish’. The input to the pretrained VGGish model are the 64 log-Mel filterbank coefficients extracted for a 0.96s audio segment. The output is a 128-dimensional vector.

⁷<https://github.com/tensorflow/models>

Table 2: Data and methods used for baseline experiments

Data/Methods	MFCC-GMM	Transfer learning
WSJ	Baseline 1 (B1)	Baseline 2 (B2)
AudioSet	Baseline 3 (B3)	proposed model

VGGish network architecture:

INP[96,64] \rightarrow **Conv[64]** + **MP** \rightarrow **Conv[128]** + **MP** \rightarrow (**Conv[256]**)x2 + **MP** \rightarrow (**Conv[512]**)x2 + **MP** \rightarrow **Flatten** \rightarrow **FC[4096]**x2 \rightarrow **FC[128]**,

where **Conv** = 2D Convolutional layer (kernel: 3x3, stride=1x1), **MP** = 2D Max Pooling layer (kernel: 3x3, stride=2x2).

3.2.2. Gender Classification

The input to the model is the VGGish-embeddings extracted for 0.96s audio segments. We replace the top layers after VGGish network shown above with FC network which are tuned for gender classification task. The network consists of three FC layers with ReLU activations. Batch normalization (BN) [31] is applied after each layer.

Network architecture for gender classification:

INP[128] \rightarrow **FC[256]** \rightarrow **BN** \rightarrow **FC[64]** \rightarrow **BN** \rightarrow **FC[16]** \rightarrow **BN** \rightarrow $\sigma(\mathbf{FC}[2])$

For all FC networks, the number of nodes and layers were tuned to minimize the loss on the development set. We also tested a logistic regression model. Experiments were also conducted excluding BN.

4. Experiments and Results

We performed a comprehensive evaluation of our system by first evaluating the BLSTM-VAD that was trained with movie audio forced-aligned with the subtitles using Gentle. We compared the performance of our VAD to that of the state-of-the-art for movie data, and also against openSMILE VAD, which is used in [5]. We then evaluated all gender identification models with, a) ground-truth (or oracle) VAD [18], and b) the proposed BLSTM-VAD (see Sec 4). Evaluation using the oracle VAD provides the performance of our gender model when perfect VAD is available. Meanwhile, evaluation with BLSTM-VAD gives the performance of the end-to-end system. This may also indicate if the same audio segments were misclassified in both VAD and gender identification systems.

All evaluations are performed on frame-level ground-truth labels for the four Hollywood movies (see Sec 2.2). We use unweighted accuracy (UWA) as the performance measure for evaluation. UWA is a more suitable measure compared to precision or recall since it assigns equal weights to each class, and hence can be used to evaluate on datasets with class imbalance (as in our evaluation set). We also calculate the area under ROC curve (A') [32], a complementary measure of the predictive accuracy of a model. $A' > 0.9$ indicates an ‘excellent’ model.

4.1. Performance of BLSTM-VAD

We evaluated our VAD model on the benchmark data released in [18]. Performance measures averaged across the four movies are shown in the Table 1. BLSTM-VAD performs significantly better than an off-the-shelf RNN-based VAD model trained for clean speech [27]. This suggests that training with in-domain data certainly improves VAD for movie audio.

The performance of our model (f1-score=0.70) is comparable to that of the state-of-the-art VAD for movie data (f1-score=0.74, [18]). However we achieve a lower system recall

Table 3: Performance evaluation of our proposed system; unweighted accuracy, UWA (mean \pm stdev), and area under ROC, A' for the movie dataset

VAD			B1		B2		B3		Proposed	
Method	Miss rate	FAR	UWA	A'	UWA	A'	UWA	A'	UWA	A'
BLSTM-VAD	0.40	0.03	0.87 \pm 0.02	0.90	0.86 \pm 0.04	0.92	0.88 \pm 0.02	0.90	0.93 \pm 0.02	0.97
Oracle-VAD	-	-	0.74 \pm 0.20	0.87	0.76 \pm 0.05	0.84	0.81 \pm 0.03	0.88	0.85 \pm 0.03	0.93

that results in a higher fraction of missed speech. During the performance evaluation of gender identification systems, we report miss-rate ($1 - \text{recall}$), along with false alarm rate (FAR) for speech. This highlights the importance of a robust VAD for the gender identification task.

4.2. Baseline experiments

We set up three baseline experiments (Table 2) to evaluate the two claims that drive our proposed model (see Sec. 1). These experiments were set up to understand the independent effects of 1) training with ‘controlled’ vs. ‘diverse’ audio, and 2) gender classification with VGGish-embeddings vs. MFCC features.

As shown in Table 2, **B1** is a GMM model trained on MFCC features extracted from WSJ dataset. Per our claims, we expect **B1** to perform poorer than the other models. **B2** is the transfer learning model, except that the embeddings for training are obtained for WSJ. We then train a MFCC based GMM model on the cross-domain AudioSet dataset (**B3**). To show the effect of the choice of training data, we compare **B1** and **B3**. Comparing **B1** and **B2** would show the importance of VGGish-embeddings.

Similarly, by comparing **B3** and **B2**, we can understand the relative gain in performance between ‘transfer learning over MFCC-GMM’ and ‘controlled vs diverse data for training’ (C2). In simpler terms, if B3 performs better than B2, it would suggest that the effect of training data on improving gender identification performance is *larger* than that of features used and the classification methods. Finally, we analyze the performance of our proposed model with respect to all the baselines.

4.2.1. B1: WSJ/MFCC-GMM

We performed feature extraction and trained a system similar to [5], but with BLSTM-VAD instead of openSMILE VAD. Speaker-homogeneous segmentation was performed using Bayesian information criterion similar to that in [5]. We then trained a GMM with 100 mixture components for 200 iterations of expectation-maximization (EM) [33] till the model converged. Accuracy on the development set was 78.8%.

In Table 3, we present the results for all experiments using both oracle VAD and BLSTM-VAD. In the latter case, we computed UWA only in regions where the predicted VAD labels match with ground-truth. We also present VAD miss rate and FAR for completeness.

As shown in Table 3, across the board, UWA is lower for oracle VAD compared to BLSTM-VAD. This supports our hypothesis that the VAD and gender models likely fail for the same audio segments.

4.2.2. B2: WSJ/Transfer Learning

We identified speech segments using the BLSTM-VAD and split them into 0.96s segments to extract VGGish embeddings. These embeddings were used as input to train the FC network to classify gender. The network architecture was chosen by tuning the the number of nodes and layers. The average accuracy

on the development set was 87%. The results suggest that the VGGish-embeddings from ‘diverse’ audio data, may not be effective in representing WSJ (Table 3 *c.f.* UWA: **B2** < **B1**).

4.2.3. B3: AudioSet data/MFCC-GMM

Here, the feature extraction was similar to **B2**. Additional experiments were conducted to tune the system parameters, which gave an accuracy of 71.3% on the development set. The results show that ‘diverse’ data for training improves gender identification performance (Table 3: *c.f.* UWA: **B3** > **B1** > **B2**).

4.2.4. Proposed model

As described in Sec 3.2, audio gender is classified with a FC network using the VGGish-embeddings. Results in Table 3 (*c.f.* UWA: **Proposed** > **B3** > **B1** > **B2**), show that our proposed model significantly⁸ outperforms baseline models with an improvement of about 4% over the best baseline model. This observation is consistent when evaluating with oracle VAD as well. It suggests that transfer learning from VGGish-embeddings is beneficial to improve gender classification for challenging audio such as in movies.

5. Discussion and Future Work

As noted earlier, the ratio of male to female speech in the evaluation dataset is 5:1. To ensure that our gender system does not perform adversely for one of the classes, we used a class-balanced development set and UWA to measure performance. This is important, especially for a confident estimation of measures such as *female speaking time* in movies.

A caveat to our proposed system is that gender labels are predicted for 0.96s segments. The duration of speech for the correctly classified segments is significantly higher than that for the misclassified segments with median values of 1.06s and 0.88s respectively⁹. This indicates the need for *VGGish-ish* embeddings for audio segments of shorter durations which would be part of our future work.

The BLSTM-VAD model, although effective for movie audio is prone to higher miss-rates. A detailed error-analysis is the subject of our future work.

6. Conclusions

In this work, we develop a model for VAD in movies using subtitle information which performs similar to that of the state-of-the-art for this domain. We have released a movie benchmark dataset with precise gender labels, to enable a joint system evaluation of VAD and gender identification in movies. Finally, our experiments show that using embeddings from large-scale audio classification models from cross-domain speech data acquired in diverse acoustic conditions significantly improves audio gender prediction in movies.

⁸McNemar test with continuity correction, $p \ll 0.01$

⁹Mann-Whitney U test, $p \ll 0.01$

7. References

- [1] T. Guha, C. Huang, N. Kumar, Y. Zhu, and S. S. Narayanan, "Gender representation in cinematic content: A multimodal approach," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, November 09 - 13, 2015*, pp. 31–34. [Online]. Available: <http://doi.acm.org/10.1145/2818346.2820778>
- [2] J. Mansanet, A. Albiol, and R. Paredes, "Local deep neural networks for gender recognition," *Pattern Recognition Letters*, vol. 70, pp. 80–86, 2016.
- [3] A. DeMarco and S. J. Cox, "An accurate and robust gender identification algorithm," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.
- [4] E. Yux0308cesoy and V. V. Nabiyev, "Gender identification of a speaker using mfcc and gmm," *2013 8th International Conference on Electrical and Electronics Engineering (ELECO)*, pp. 626–629, 2013.
- [5] N. Kumar, M. Nasir, P. G. Georgiou, and S. S. Narayanan, "Robust multichannel gender classification from speech in movie audio," in *INTERSPEECH*, 2016.
- [6] W. H. Abdulla and N. K. Kasabov, "Improving speech recognition performance through gender separation," 1988.
- [7] B. M. Ore, R. E. Slyh, and E. G. Hansen, "Speaker segmentation and clustering using gender information," *2006 IEEE Odyssey - The Speaker and Language Recognition Workshop*, pp. 1–8, 2006.
- [8] I. Bisio, A. Delfino, F. Lavagetto, M. Marchese, and A. Sciarone, "Gender-driven emotion recognition through speech signals for ambient intelligence applications," *IEEE Transactions on Emerging Topics in Computing*, vol. 1, no. 2, pp. 244–257, Dec 2013.
- [9] K. Wu and D. G. Childers, "Gender recognition from speech. part i: Coarse analysis," *The Journal of the Acoustical Society of America*, vol. 90 4 Pt 1, pp. 1828–40, 1991.
- [10] M. S. Phillips, J. R. Glass, J. Polifroni, and V. Zue, "Collection and analyses of wsj-csr data at mit," in *HLT*, 1992.
- [11] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.
- [12] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "The interspeech 2010 paralinguistic challenge," in *Proc. INTERSPEECH 2010, Makuhari, Japan, 2010*, pp. 2794–2797.
- [13] H. Harb and L. Chen, "Gender identification using a general audio classifier," in *ICME*, 2003.
- [14] E. Ramdinmawii and V. K. Mittal, "Gender identification from speech signal by examining the speech production characteristics," *2016 International Conference on Signal Processing and Communication (ICSC)*, pp. 244–249, 2016.
- [15] E. Weis and B. John, *Film Sound - Theory and Practice*. Columbia University Press, 1985.
- [16] M.-W. Mak and H.-B. Yu, "A study of voice activity detection techniques for nist speaker recognition evaluations," *Comput. Speech Lang.*, vol. 28, no. 1, pp. 295–313, Jan. 2014. [Online]. Available: <http://dx.doi.org/10.1016/j.csl.2013.07.003>
- [17] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 483–487.
- [18] B. Lehner, G. Widmer, and R. Sonnleitner, "Improving voice activity detection in movies," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
- [19] C. Ding, P. Zhu, and L. Xie, "Blstm neural networks for speech driven head motion synthesis," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
- [20] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis, "Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention," *IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1553–1568, 2013.
- [21] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.
- [22] D. Wang and T. F. Zheng, "Transfer learning for speech and language processing," *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 1225–1237, 2015.
- [23] J. Kunze, L. Kirsch, I. Kurenkov, A. Krug, J. Johannsmeier, and S. Stober, "Transfer learning for speech recognition on a budget," in *Rep4NLP@ACL*, 2017.
- [24] A. Diment and T. Virtanen, "Transfer learning of weakly labelled audio," *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 6–10, 2017.
- [25] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017. [Online]. Available: <https://arxiv.org/abs/1609.09430>
- [26] R. G. Gonzalez, "Youtube-dl: download videos from youtube.com," 2006.
- [27] F. Eyben, M. Wöllmer, and B. W. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *ACM Multimedia*, 2010.
- [28] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.
- [29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.
- [30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.
- [31] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *ICML*, 2015.
- [32] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.
- [33] A. P. Dempster, N. Laird, and D. D. B. D. Rubin, "Maximum likelihood from incomplete data via the em algorithm," 1977.