*INTERSPEECH 2016*
September 8–12, 2016, San Francisco, USA

# Robust multichannel gender classification from speech in movie audio

*Naveen Kumar, Md Nasir, Panayiotis Georgiou, Shrikanth S. Narayanan*

Signal Analysis and Interpretation Lab,
Department of Electrical Engineering,
University of Southern California, Los Angeles.
{komathnk,mdnasir}@usc.edu, {georgiou,shri}@sipi.usc.edu

## Abstract

Speech in the form of scripted dialogues forms an important part of the audio signal in movies. However, it is often masked by background audio signals such as music, ambient noise or background chatter. These background sounds make even otherwise simple tasks, such as gender classification, challenging. Additionally, the variability in this noise across movies renders standard approaches to source separation or enhancement inadequate. Instead, we exploit multichannel information present in different language channels (English, Spanish, French) for each movie to improve the robustness of our gender classification system. We exploit the fact that the speaker labels of interest in this case co-occur in each language channel. We fuse the predictions obtained for each channel using Recognition Output Voting Error Reduction (ROVER) and show that this approach improves the gender accuracy by 7% absolute (11% relative) compared to the best independent prediction on any single channel. In the case of surround movies, we further investigate fusion of mono audio and front center channels which shows 5% and 3% absolute (8% and 4% relative) increase in accuracy compared to only using mono and front center channel, respectively.

**Index Terms**: movie audio processing, gender prediction, rover, multichannel

## 1. Introduction

With the advent of robust multimedia algorithms, content analytics in media is rapidly gaining popularity and a host of new applications are emerging everyday. In addition to classic ones such as indexing and summarization, applications focused more on higher level video understanding [1, 2] have demonstrated significant promise. In the domain of movie content processing, various tasks such as narrative act structure characterization, violent scene detection and saliency prediction [3, 4] for regions of potential greater engagement are some examples of interesting applications. Many of these methods help to analyze movie datasets at scale making it easier for human experts to perform higher level analytics and decision making.

As an application of multimedia content analysis to media informatics and demographics, this paper deals with gender analytics in movie audio for an objective understanding of gender representation in popular media [5, 6, 7]. There is widespread evidence–though largely qualitative–of disparity and differences in gender representations and portrayals in media calling for the development of objective, data-driven methods; the goal of this paper. One such desired metric is what fraction of a movie is shared by female characters. Toward that goal, in this paper, we estimate the total speaking time of characters of each gender. Although the speaking time is not the only measure of presence of characters of a specific gender, important characters in a movie are typically allocated more speaking time. Hence, a systematic computation of speaking time over a large number of movies could reveal any patterns of unconscious bias in these movies [8].

In terms of speech processing techniques, this requires voice activity detection (VAD) followed by gender identification of the speaking characters. A pipeline for such gender specific speaking time measurement from movie audio was proposed earlier by Guha et al. [9]. Background noise such as music or sound effects in movie audio presents the main challenge for robustness in the application of such methods. Moreover, the signal-to-noise ratio for the speech signal is usually variable throughout a movie or across different genres adding to the complexity of the problem.

In this paper, we extend the method discussed in [9] for computing gender-specific speaking time from the audio stream in a movie. The audio system described in [9] analyzed the single audio channel (English, mono) of a movie to estimate the speaking time by gender associated with characters in the movie. The speaking times for each gender are then used to estimate the overall female and male speaking time ratio, a metric desired by social scientists and media scholars [8]. In order to improve the accuracy of the gender prediction system from movie audio, we exploit information from audio channels of multiple languages (often available with most major studio movies, through dubbing) as well as the surround audio channels by fusing predictions from individual channels using Recognition Output Voting Error Reduction (ROVER)[10]. Since the various parallel audio channels are generated independently, the idea is that they differ in their quality, and can be exploited together to provide more robust estimates for diarization. First developed to combine output of multiple Automatic Speech Recognition (ASR) systems, ROVER and similar fusion techniques [11] have been later used for other tasks, typically when the output labels of different systems are not temporally aligned. The applications include emotion recognition [12, 13], speech-to-speech translation systems [14], machine translation [11] *etc.*

We first describe a benchmark dataset for evaluating the accuracy of the gender prediction system with and without fusion. The details of obtaining an annotated dataset for this purpose are discussed in Section 2. Next, we describe the steps involved in our gender prediction system in Section 3. We present the evaluation results as well as results of our proposed system in Section 6. Finally, concluding remarks are made in Section 7.

Work supported by NSF and Google



http://dx.doi.org/10.21437/Interspeech.2016-540

## 2. Datasets and Annotations

Our evaluation dataset comprises 22 Hollywood movies, comprising what we refer to as set A and set B. Set A has 15 movies released in the year 2014. These movies have only mono audio channels for multiple languages (English/en, French/fr, Spanish/es). The second set, set B consists of 7 movies released in 2010. In addition to multiple languages, the audio streams of these movies also have 5.1 surround channels.

All of these movies were selected from diverse genres and subtitles were downloaded for each movie. We manually checked each subtitle file at random to ensure they are accurate and time-aligned. We use movie subtitles for gender annotations since they provide rough timestamps for dialogues facilitating the annotation process. In this paper, we assume subtitle timestamps to be the ground truth, although in practice the subtitle boundaries might not exactly align with speech boundaries as they are primarily meant to be visual aids.

For each movie we asked annotators to listen to the movie audio within subtitle timestamps. All dialogues within each subtitle were associated with one of the gender labels: *M, F, MF, None*. The label *MF* indicates that both male and female characters were speaking within the time duration; *None* is used if no speech is present. Dialogues belonging to these two labels are then excluded from further analysis, since our current gender prediction system can only deal with speaker homogeneous speech segments. Table 1 below shows the list of movies with the number of dialogues used from each dataset. The average duration of utterances was approximately 3s.

| **Movie Title** | **Properties** |
|---|---|
| Set A (15 movies) | |
| Alexander and the Terrible, Horrible, No Good, Very Bad Day; Annie; Blended; Captain America: The Winter Soldier; Cesar Chavez; If I stay; Maleficent; Million Dollar Arm; Need for speed; Selma; Son of God; Hundred Foot Journey; The Hunger Games: Mockingjay - Part 1; Unbroken; Dumb and Dumber to | mono audio, multiple languages (en/es/fr), 21293 dialogues (15168 M, 6125 F) |
| Set B (7 movies) | |
| Cop out; Due date; Easy A; Takers; The Bounty Hunter; The Tourist; Resident Evil: Afterlife | 5.1 channel surround audio, multiple languages (en/es/fr), 8151 dialogues (5966 M, 2185 F) |

Table 1: List of movies in the evaluation set

## 3. Method

In this section, we elaborate on the baseline system used for gender prediction of dialogues in movie audio as discussed in [9]. The goal of such a system would be to provide accumulative statistics of total speaking time associated with male and female characters in the movie. For this purpose, we first isolate and segment speech regions in the movie audio which are later classified as male or female dialogues. Each of the subsystems in the pipeline are described in detail next.

### 3.1. Voice Activity Detection and segmentation

Differentiating between speech and non-speech regions in movie audio can be a challenging task. In this paper, we use the OpenSMILE Voice Activity Detection (VAD) tool which has been used on movie audio [15]. This VAD module is readily available with the OpenSMILE toolkit in the form of a RNN-LSTM model that was trained on data corrupted with synthetic movie noise. [15] shows this VAD tool to be robust to the different audio background present in a movie and we observe similar results in our experiments. The OpenSMILE VAD takes as input a WAV file and first generates VAD activation which are soft values indicating presence or absence of speech in the signal. This information is then passed through the turn-detector module in OpenSMILE that converts voice activity timestamps. Additionally, the WAV segments corresponding to the speech regions are also saved.

Note that though speech regions have been extracted from the movie audio, they need not be speaker homogeneous. This can pose a problem later, while classifying the gender of an utterance. Hence, we perform a further pass of speaker-homogeneous segmentation of each speech utterance obtained in the last step. We use a Bayes Information Criterion (BIC) based segmentation algorithm [16, 17, 18] implemented in the open source speech-recognition toolkit KALDI [19]. We use 13 dimensional MFCC features for speech segmentation in this work. Naturally, as in any segmentation task, there exists a trade-off between purity of segments and over-segmentation. It is important to note in this context, that segmentation is not the final goal of this work. In fact over-segmentation to a certain extent might be desirable if it guarantees speaker purity in each chunk. Thus at the end of speaker segmentation we end up segmenting each speech utterance into a larger number of chunks. Typically, for a movie audio stream 2 hours in duration around 1500 chunks are obtained, with a total duration of about 30 minutes of speaking time.

### 3.2. Utterance gender classification

Each chunk output from the segmentation system is then classified as belonging to a male (M)/ female (F) character based on its acoustic features. In this work, we used 13-dimensional MFCC features for this purpose. We avoid other commonly used higher level features such as pitch, because of the difficulty in robustly estimating these from noisy movie audio. We use a classification rule based on Gaussian mixture model (GMM) log-likelihood ratio. Two GMM models are trained using frame level features for each gender. Gender classification for a new utterance with features $X$ can then be done by comparing the log-likelihood according the two models as shown below.

$$\mathcal{L}_{\text{male}}(X) \underset{M}{\overset{F}{\lessgtr}} \mathcal{L}_{\text{female}}(X) \tag{1}$$

We train our gender models on a subset of the Wall Street journal corpus (WSJ-SI84) comprising 7184 utterances from 42 male and 41 female speakers. In our experiments, we use 100 mixtures component GM models, after tuning the parameter on a held out eval set. Feature extraction, GMM training and log-likelihood computation was performed using KALDI [19]. In addition to computing log-likelihood ratios, we also compute a simple confidence measure $c(\cdot)$ for gender classification as shown in Eq.(2)

$$\begin{aligned} p(X) &= \sigma(\mathcal{L}_{\text{male}}(X) - \mathcal{L}_{\text{female}}(X)) \\ c(X) &= 2|p(X) - 0.5| \end{aligned} \tag{2}$$

where $\sigma(.)$ is the sigmoid function. This confidence measure will be useful later when we fuse predictions from multiple audio channels.

# 4. Fusing predictions from multiple audio channels

Nowadays, most Hollywood movies released in the form of DVDs contain multiple audio channels. This often includes audio tracks in multiple languages for movies meant to reach out to a wider audience. Other than the primary language (mostly English) other audio tracks are usually dubbed over by voice artists in the foreign language. While dubbing a movie a voice artist needs to be mindful of recreating each dialogue as close in style and prosody as possible to their original. This is necessary because the dubbed dialogues still need to be synced to the video stream, which remains the same. Additionally, the dialogues also need to have similar time boundaries. This fact can be exploited to obtain gender prediction that leads to optimal consensus on all existing language channels.

In addition to multiple languages each audio track might sometimes also contain spatial audio channels designed to provide viewers with an immersive acoustic experience. One of the most commonly used layouts is the 5.1 surround sound [20] mix that uses 5 full bandwidth channels viz. front left, front center (or, *center* in some literature), front right, surround left and surround right. A sixth channel is usually reserved exclusively for low frequency sound effects. For movie audio in 5.1 mix the surround channels typically never contain any speech. Most of the dialogues occur in the front center channel along with a few other sound effects that are near the person speaking. This fact has been used in the past for designing speech enhancement algorithms for movie audio to assist the hearing impaired [21, 22, 23]. In this paper, we propose a method to exploit the complementary information present in both the surround and multiple language channels.

To obtain a consensus prediction between multiple audio channels we make use of the Recognizer Output Voting Error Reduction (ROVER) scheme [10]. The ROVER scheme was designed to fuse output hypotheses from multiple automatic speech recognizers (ASRs). ROVER takes as input a sequence of symbols from each input hypothesis. It first aligns them using their timestamps and performs a majority voting among the different options for each time segment. Some variants of the ROVER algorithm can also incorporate average confidence information for each symbol by performing a convex combination of symbol frequency and confidence. This can be used via the method *avgconf* in the ROVER tool [24] and requires tuning of the parameter $\alpha$ that assigns weight to the confidence value as opposed to just symbol frequency which would be akin to simple majority voting. Note that it is necessary to first time align each input hypothesis before voting since the number of time segments might differ across each hypothesis. As a result, the final fused prediction also has different timestamps and reflects consensus between different input hypotheses.

# 5. Experiments

Due to different properties of the datasets A and B, we conduct different set of experiments for each of them. For set A we extract single channel (mono) audio tracks corresponding to each of the language tracks in English, Spanish or French. Each mono track is then run through the gender prediction pipeline described in Section 3. Finally, the predictions obtained for each channel are fused together using confidence weighted ROVER.

For the movies in set B we adopt a slightly different approach, by also deriving multiple audio channels from the 5.1 surround sound mix. Since most of the dialogues in a movie occur in the front center channel with a smaller number of background sound effects, we process this channel directly using the gender prediction pipeline. Simultaneously, the 5.1 channel audio is also downmixed to a mono channel and then used for predicting gender. The predictions on the mono and center channels are then fused using confidence weighted ROVER as before. We experimented with different values of $\alpha$ for average confidence ROVER on a held out set and selected $\alpha = 0.4$ as the optimal weight for prediction confidence in ROVER voting. Note that a higher value of $\alpha$ implies that we rely on confidence values more.

We also performed the ROVER fusion of predictions from multiple languages, separately with corresponding mono and center channels for movies in set B. Finally, we obtained the gender predictions for these movies by fusing all available channels used so far: mono and center channels of multiple languages.

In all experiments, gender predictions are evaluated at the subtitle level. For each gender annotated subtitle in our dataset, we match the subtitle timestamps to find all time overlapping predictions. For the overlapping predictions we then compute an average prediction weighted by their confidences. This predicted gender is compared against the annotated ground truth to compute the gender classification accuracy. Note that the classification error in this case will include subtitles for which no overlapping prediction segments were found because of VAD errors. These are used to compute the percentage VAD miss rate. For the evaluation of the gender classification system, we first calculate the accuracy of assigning correct labels to utterances from each gender and then compute their unweighted average (UWA) as the performance metric. This metric is chosen as it is robust to bias in class proportions on the evaluation set.

# 6. Results

We divide the results into 3 subsections; rover using a) multiple language channels b) 5.1 surround channels and c) both language and surround channels.

## 6.1. Multiple language channels

Results obtained on Set A by fusing predictions across multiple language tracks are shown in Table 2. We note that ROVER helps in improving the accuracy of gender prediction in this case. This might suggest that while the underlying gender information is consistent across different language tracks, their prediction errors are complementary in nature. We also observe that using confidence information in ROVER generally improves efficiency of the fusion algorithm, although the parameter $\alpha$ must be chosen with care. It is additionally interesting to note that ROVER-ing across multiple language channels also helps reduce the average VAD miss rate as the final fused out timestamps are obtained through an initial time alignment that tries to achieve consensus between different input hypotheses.

## 6.2. Surround sound channels

Unlike the multiple language tracks, speech in the mono and center channels that we use for ROVER in this case are not com-

| Method | Accuracy (UWA in %) | VAD miss (in %) |
|---|---|---|
| English only | 62.6 | 18.4 |
| rover ($\alpha = 0$) | 64.1 | 16.3 |
| rover ($\alpha = 0.4$) | 69.2 | 10.7 |

Table 2: Performance obtained on the multilingual dataset Set A by fusing prediction over language channels.

pletely independent. In fact, compared to the center channel, speech in the mono channel can be thought of to be a noisier version of the same underlying signal. Hence, it is not surprising to see that gender prediction is more accurate on the center channel as compared to mono (Table 3). However, the information contained in the center channel is still noisy and complementary, which is suggested by the further improvement obtained on fusing the predictions on the mono and center channels. Note that the figures in Tables 2 and 3 cannot be directly compared as they are on different sets of data. However the overall trends are similar as we show next.

| Method | Accuracy (UWA in %) | VAD miss (in %) |
|---|---|---|
| mono | 64.0 | 16.2 |
| center | 66.5 | 12.8 |
| mono + center | 69.1 | 10.9 |

Table 3: Performance improvement obtained on Set B (7 movies) by using 5.1 surround channels only.

### 6.3. Combining both surround and language channels

Finally, we also experiment by combining the multi language tracks in addition to the surround sound tracks in Set B. Gender predictions on different language tracks are first fused individually on the mono and center channels, followed by a joint ROVERing of all channels for each movie. The results are shown in Table 4.

| Method | Accuracy (UWA in %) | VAD miss (in %) |
|---|---|---|
| mono multilang | 68.6 | 10.3 |
| center multilang | 68.9 | 9.1 |
| rover all | 71.1 | 7.5 |

Table 4: Performance improvement obtained on Set B (7 movies) by using both multichannel and multi-language information.

These results clearly show the merit in the joint fusion of surround and multiple language channels. We also notice that the proposed fusion also yields an improvement in the VAD miss rate.

## 7. Conclusion

In this paper we describe a framework for performing gender prediction on speech in movie audio. Such a system could find use in several applications including gender-specific speaking time estimation as discussed in [9]. The problem of gender classification from movie audio is inherently very challenging because of the diverse nature of background noise in movies. This can be seen from the low classification accuracy when the gender system is directly run on the center channel.

At the same time, movie audio is also very structured which allows for redundancy in the information present across different channels- both spatial and linguistic. We seek to exploit this latent structure in this paper, by obtaining consensus of gender predictions on different channels.

Our results suggest that fusing gender predictions over multiple audio channels in movies helps improve its robustness. We obtain a 7% absolute improvement in gender classification accuracy (11% relative) by combining both multiple languages and spatial channels. The results on mono-center fusion are perhaps the most interesting since each channel contains partial but somewhat diverse information about the underlying gender labels.

In the future, we would like to investigate if the benefits of ROVER fusion scale with accuracy of the baseline gender prediction system. Further experiments also need to be performed for more efficient fusion schemes by incorporating domain knowledge. Finally, it would be interesting to see how the cross-channel information could be useful in other speech related tasks such as diarization in movie audio.

## 8. References


[1] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko, "Translating videos to natural language using deep recurrent neural networks," *arXiv preprint arXiv:1412.4729*, 2014.

[2] T. Guha, N. Kumar, S. Narayanan, and S. Smith, "Computationally deconstructing movie narratives: an informatics approach," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014.

[3] G. Evangelopoulos, K. Rapantzikos, P. Maragos, Y. Avrithis, and A. Potamianos, "Audiovisual attention modeling and salient event detection," in *Multimodal Processing and Interaction*. Springer, 2008, pp. 1–21.

[4] 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," *Multimedia, IEEE Transactions on*, vol. 15, no. 7, pp. 1553–1568, 2013.

[5] S. L. Smith, M. Choueiti, K. Pieper, T. Gillig, C. Lee, and D. DeLuca, "Inequality in 700 popular films: Examining portrayals of gender, race, & LGBT status from 2007 to 2014," *Media, Diversity, & Social Change Initiative*.

[6] S. L. Smith and M. Choueiti, "Gender disparity on screen and behind the camera in family films; the executive report," *Geena Davis Institute on Gender in Media*, 2011, http://seejane.org/wp-content/uploads/full-study-gender-disparity-in-family-films-v2.pdf.

[7] S. Smith and C. A. Cook, "Gender stereotypes: An analysis of popular films and TV," *Geena Davis Institute on Gender in Media*, vol. 208, pp. 12–23, 2008, http://seejane.org/wp-content/uploads/GDIGM_Gender_Stereotypes.pdf.

[8] J. Glascock, "Gender roles on prime-time network television: Demographics and behaviors," *Journal of Broadcasting & Electronic Media*, vol. 45, no. 4, pp. 656–669, 2001.

[9] T. Guha, C. Huang, N. Kumar, Z. Yan, and S. Narayanan, "Gender representation in cinematic content: A multimodal approach," in *Seventeenth ACM International Conference on Multimodal Interaction*, 2015.

[10] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*. IEEE, 1997, pp. 347–354.

[11] S. Bangalore, G. Bordel, and G. Riccardi, "Computing consensus translation from multiple machine translation systems," in *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*. IEEE, 2001, pp. 351–354.

[12] A. Batliner, S. Steidl, B. Schuller, D. Seppi, K. Laskowski, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous *et al.*, "Combining efforts for improving automatic classification of emotional user states," *Proc. IS-LTC*, pp. 240–245, 2006.

[13] J. Pittermann and A. Pittermann, "A post-processing approach to improve emotion recognition rates," in *Signal Processing, 2006 8th International Conference on*, vol. 1. IEEE, 2006.

[14] D. Mostefa, O. Hamon, and K. Choukri, "Evaluation of automatic speech recognition and speech language translation within tc-star: Results from the first evaluation campaign," *Proceedings of LREC'06*, 2006.

[15] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 483–487.

[16] E. C. Ozan, S. Tankiz, B. O. Acar, and T. Ciloglu, "An unsupervised audio segmentation method using bayesian information criterion," in *Communications, Control and Signal Processing (ISCCSP), 2014 6th International Symposium on*. IEEE, 2014, pp. 640–643.

[17] A. Tritschler and R. A. Gopinath, "Improved speaker segmentation and segments clustering using the bayesian information criterion." in *Eurospeech*, vol. 99, 1999, pp. 679–682.

[18] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1557–1565, 2006.

[19] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[20] T. Lund, "Enhanced localization in 5.1 production," in *Audio Engineering Society Convention 109*. Audio Engineering Society, 2000.

[21] J. T. Geiger, P. Grosche, and Y. L. Parodi, "Dialogue enhancement of stereo sound," in *Signal Processing Conference (EUSIPCO), 2015 23rd European*. IEEE, 2015, pp. 869–873.

[22] C. Uhle, O. Hellmuth, and J. Weigel, "Speech enhancement of movie sound," in *Audio Engineering Society Convention 125*. Audio Engineering Society, 2008.

[23] E. Vickers, "Frequency-domain two-to three-channel upmix for center channel derivation and speech enhancement," in *Audio Engineering Society Convention 127*. Audio Engineering Society, 2009.

[24] "ROVER: Recognition output voting error reduction," http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/rover.htm.