A Computational Tool to Study Vocal Participation of Women in UN-ITU Meetings

Rajat Hebbar¹, Krishna Somandepalli¹, Raghuveer Peri¹, Ruchir Travadi¹, Tracy Tuplin², Fernando Rivera², and Shrikanth Narayanan¹

¹Signal Analysis and Interpretation Laboratory, University of Southern California, Los Angeles
²International Telecommunication Union, United Nations, Geneva

Abstract—International organizations such as the United Nations drive policies that impact our everyday lives. Diverse representation of people and ideas in the decision making process of such bodies is critical to ensure that the policies work for everyone. One aspect of the representation is the participants’ expressed gender. In this work, we focus on analyzing meetings at the International Telecommunication Union (ITU). These meetings include a moderator who mediates the proceedings between delegates from across the world speaking in different languages. For the purpose of quantifying the participation of delegates, we propose a scalable, human-in-the-loop system to first identify the moderator’s speech and estimate the speaking time with respect to gender for all the speakers. Our proposed system includes three main audio modules: speech activity detection, gender identification and moderator verification using a human-labelled speech probe. We then estimate percentage of speaking time controlled for the moderator’s speech. We present detailed and multilingual performance evaluation of the component systems using state-of-the-art technologies for these tasks. Finally, we examine the vocal participation of female delegates in the 2018 ITU Plenipotentiary Conference spanning for 18 days and about 108 hours of audio recordings.

Index Terms—Gender representation, United Nations, multilingual meetings, audio systems

I. INTRODUCTION

Over the years, female participation and involvement in social and political spheres has seen a gradual uptrend [1]. However, this is not necessarily reflected in women’s influence in the decision making processes at an institutional level. Meetings, conferences and similar forums of discussion are critical components of this process. A recent study [2] that examined village community level meetings over a period of seven months found that “Women’s lack of participation in important decision making is noted as an obstacle to sustainable development” and “Even when women are present at meetings they are still consistently less likely than men to substantively participate”. Furthermore, multiple studies [3]–[5] also reported under-participation of women in Q&A sessions at academic conferences, despite nearly equal attendance of both men and women. Besides highlighting the inequities in our communities, such studies can also lead to systemic changes in policy with a real impact on our daily lives [6].

A crucial limitation of scaling up such studies in related domains is typically their dependence on manual labeling which is expensive, time-consuming and can be prone to errors. A promising direction to address this limitation is the recent efforts in developing machine learning based tools to automate our understanding of human representation in media content [7]. A study on nearly 600 top grossing Hollywood movies from 2014–2019 found that men speak significantly more than women, even when women appear on screen [7].

Our focus in this work is to measure the participation of women in the policy discussions at the United Nations International Telecommunication Union (UN-ITU)¹. ITU is a specialized agency that deals with setting international standards on issues related to communication technology (internet, TV broadcasts, etc.), managing the radio-spectrum, satellite orbits, and bridging the digital divide in the world. The far-reaching impact of policy decisions made at the ITU is an interesting case to study the diversity of participants. In this context, we wish to quantify female representation beyond simple attendance counts by analyzing the audio recordings of the ITU meetings.

ITU meetings are multilingual (See Fig. 2) and generally presided over by a moderator, who speaks in English. The moderator typically begins the meeting and mediates the representatives (delegates) from different countries throughout the proceedings. As a result, the moderator is likely to account for a substantial fraction of total speech in the meeting. Because our goal is to study the representation of the delegates’ speech in general, we need to reliably estimate the amount of moderator’s speech. In order to address this, we use a speaker verification (SV) module to query speech segments belonging to the moderator and we account for this during the estimation of delegate speaking time.

Acoustic variability poses major challenges to speech processing systems. Different languages, variable or noisy recording conditions and within-individual differences are a few prominent sources of variability. Meetings in particular, are susceptible to cross-talk, channel variations and other noises associated with close-talk microphones such as breath noise [8], [9]. These variables pose challenges to canonical audio systems such as speech activity detection, automatic speech recognition and speaker diarization [10], [11]. In this study, we develop a framework using state-of-the-art methodologies for the different components used to automate this process.

¹https://www.itu.int/en/Pages/default.aspx
II. SPEAKING TIME ESTIMATION

In this section, we describe the various audio systems implemented in the speaking time estimation pipeline. As shown in the Fig. 1, the overall pipeline includes a human-in-the-loop to annotate a single sample (about 1min long) of moderator speech (used in speaker verification module) for each session. We refer to this annotated sample as the moderator probe.

Speech Activity Detection: Speech activity detection (SAD) is the task of segmenting an audio clip into speech and non-speech segments. We implement a segment-level (0.64s) convolutional neural network (CNN) model trained with movie audio [12]. We use the time-distributed CNN flavor described in [12], which we refer to as movie-SAD henceforth. Following SAD, we perform speaker-homogeneous segmentation using Bayesian information criterion (BIC) [13] based segmentation.

Gender Identification: Assigning a gender ID to the resulting segments is the next step in our pipeline. We use a 2-layer fully connected neural network model [14] trained on YouTube audio to predict gender at 0.96s intervals. Majority voting is used to determine gender ID for a speaker homogeneous segment longer than 1s. At this stage of the pipeline, we can estimate the overall percentage female speaking time as the fraction of total speech that belongs to a female speaker.

Moderator Detection: We pose the task of detecting moderator’s speech in these meetings as a speaker verification task given an exemplar of the speech. Speaker verification is the task of accepting or rejecting a test speech segment to belong to the same speaker identity as a query segment (moderator-probe in Fig. 1).

Low-dimensional representations obtained using total variability modeling, referred to as i-vectors [15] are a popular method of capturing speaker related information [16]. In our work, we use i-vector models trained on the NIST SRE-2008 database [17]. We use two variants of i-vector models, a) gender-dependent (GD) - trained separately for each gender, and b) gender-neutral (GN) - trained independent of gender. Recently, neural networks have shown improved results for speaker recognition tasks [11], [18]. We evaluate two variants of x-vectors [11], a) Callhome x-vector, and b) Voxceleb x-vector. We use the PLDA scoring [19] as similarity metric.

Speaking time estimation pipeline: The above described modules are used in combination to estimate the speaking time (Fig. 1). Given a meeting audio file, SAD and gender ID are applied to detect speech segments and assign gender. Speaker segmentation between these two is crucial to ensure that gender ID posteriors are aggregated only over a single speaker. Speaker-embeddings are then extracted for each speech segment, and the moderator-probe segment identified by human annotation. PLDA scores are computed for pairs of embeddings of the speech segments and the moderator-probe. Speech segments that correspond to the top 75th percentile of scores are classified as moderator segments, used to estimate the amount of moderator speech. Delegate speaking time is then estimated by subtracting the moderator speaking time from the overall speech duration. We can now readily estimate female delegate speaking time as a percentage of the overall speaking time after accounting for the moderator.

III. ITU MEETING DATA

In this section, we describe the data used to evaluate the constituent audio systems. We also provide details on the ITU Plenipotentiary meetings, on which we deploy our proposed system and provide analysis on female vocal participation.

A. Development Set

To evaluate the systems described in the previous section, we use two ITU sessions. The sessions were chosen so as to have different moderator gender.

Both the sessions were annotated for SAD, gender labels and language spoken at the granularity of 1 second. The distribution of UN-official languages in this data is shown in Fig. 2. Furthermore, speech corresponding to the moderator was also annotated for the purpose of validating the speaker verification module. Details regarding distribution of speech across delegates within each session are shown in Table I.
TABLE I
DISTRIBUTION OF SPEECH IN DEVELOPMENT SET (TIME IN MIN)

<table>
<thead>
<tr>
<th></th>
<th>Total duration</th>
<th>Total speech</th>
<th>Avg. delegate floor time</th>
<th>Moderator Gender</th>
<th>Gender Time</th>
<th>% female speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>Session 1</td>
<td>173</td>
<td>141</td>
<td>1.1 (± 1.12)</td>
<td>M</td>
<td>33</td>
<td>15.4</td>
</tr>
<tr>
<td>Session 2</td>
<td>193</td>
<td>150</td>
<td>1.65 (± 1.58)</td>
<td>F</td>
<td>12</td>
<td>28.6</td>
</tr>
</tbody>
</table>

B. Plenipotentiary Meetings

We deploy our system to analyze meetings from the 2018 ITU Plenipotentiary (PP-18) conference\(^3\), which is the top policy making body of the ITU. These meetings are held every 4 years and are critical for setting a roadmap for the subsequent period. The dataset consists of 54 sessions that were held over a period of 3 weeks, each varying in duration - ranging from 26 minutes up to 14 hours. These meetings are also multi-lingual in accordance with UN standards. The multi-lingual nature poses an additional challenge to the robustness of our audio systems to language changes during a session.

IV. EXPERIMENTS

In order to baseline speech segmentation, we use openSMILE [20], which is a recurrent neural network model developed for SAD. We also use a CNN model (INA) [21] developed for the purpose of speaking time estimation in French TV broadcasts. Since speaking time estimates are adversely affected by both missed-speech and false alarms, we report precision, recall and F1 score as performance metrics for SAD. For the gender ID system, we use the INA model [21] as a baseline. To evaluate gender ID, we use oracle-SAD labels in order to decouple the errors in gender identification from that of speech detection. We report unweighted-average recall as the primary metric for gender ID.

For the verification task, we use one minute of annotated moderator speech as query segments for each of the sessions. We report verification results based on oracle SAD and gender ID. We test both gender dependent and gender-independent i-vectors and the two variants of x-vectors for extracting speaker-discriminative embeddings. We report equal error rate (EER), which is the standard metric for verification task [15].

We then apply the best performing systems in the sequence described in Sec. II to estimate percentage female speaking time estimates controlling for the moderator’s speech. We compare these estimates with the ground truth measures on development set and further deploy the system for PP meetings. We also conduct performance evaluation of the SAD system across languages to understand robustness of our proposed system to different languages.

V. DISCUSSION

A. System Development

As shown in Table II, the movie-SAD outperforms the openSMILE system, missing only 5% of speech. INA results in similar performance, with a trade-off in precision and recall.

\(^3\)https://www.itu.int/web/pp-18/en/

---

**TABLE II**

**SPEECH ACTIVITY (SAD) AND GENDER ID (GID) PERFORMANCE ON ITU DEVELOPMENT SET. METRICS ARE AVERAGED OVER TWO SESSIONS**

<table>
<thead>
<tr>
<th>Model</th>
<th>SAD Precision</th>
<th>SAD Recall</th>
<th>SAD F1</th>
<th>GID UAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Movie SAD / AudioSet</td>
<td>0.97</td>
<td>0.95</td>
<td>0.96</td>
<td>1</td>
</tr>
<tr>
<td>OpenSMILE [20]</td>
<td>0.98</td>
<td>0.88</td>
<td>0.93</td>
<td>-</td>
</tr>
<tr>
<td>INA [21]</td>
<td>0.93</td>
<td>1</td>
<td>0.96</td>
<td>0.96</td>
</tr>
</tbody>
</table>

**TABLE III**

**SPEAKER VERIFICATION PERFORMANCE (EER in %) ON ITU DEVELOPMENT SET. GN: GENDER NEUTRAL, GD: GENDER DEPENDENT**

<table>
<thead>
<tr>
<th>Session</th>
<th>Moderator</th>
<th>GN i-vector</th>
<th>GD i-vector</th>
<th>Callhome x-vector</th>
<th>Voxceleb x-vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>Session 1</td>
<td>Male</td>
<td>1.33</td>
<td>2.87</td>
<td>2.74</td>
<td>0.0</td>
</tr>
<tr>
<td>Session 2</td>
<td>Female</td>
<td>7.22</td>
<td>21.7</td>
<td>3.9</td>
<td>12.99</td>
</tr>
</tbody>
</table>

Gender ID performance is shown in Table II over 338 oracle speaker homogeneous segments. The model developed on AudioSet shows perfect gender classification on both the sessions, while the INA model shows slightly poorer performance.

From Table III, we see that SV performs significantly worse with a female query segment than with male query segment for both i-vector models. The Voxceleb x-vector model also performs significantly worse on female segments, while giving perfect performance on male segments. This could be attributed to the gender imbalance in the training data. However, the disparity in performance across gender is significantly reduced with the Callhome x-vector model. One reason why the Callhome model (trained on telephonic speech) outperforms the Voxceleb model (trained on celebrity interviews) could be that the recording conditions in the former are more closely matched to our application.

Using the best performing systems for each module we estimate 17.2% and 26.1% female speech (controlled for moderator) in sessions 1 and 2 respectively. We can see from Table 1 that these are both within 2.5% absolute and 12% relative of the ground truth percentages.

Since we are interested in how our audio system performs on different languages, we evaluate the SAD module on each of the languages individually. From Table IV, we can see that the movie-SAD model performs well (<5% MR) on all languages except Russian, where it has a MR of 14%. This is consistent with previous work [22]. Movie-SAD performs significantly better than OpenSMILE across all the languages.

B. Plenipotentiary Meetings

As in the development set, we use annotated samples of moderator speech for each session for verification. Out of the 54 sessions, 11 had a female moderator. The estimated female delegate speaking time percentages for each of the PP sessions has been shown in Fig. 3. The estimates indicate that female delegates accounted for an average of 25.12% (±12% std) of the total speech controlled for the moderator. Conversely, the female delegate attendance in the PP-18 sessions stood at around 27%. The standard-deviation of 12% shows that vocal participation varies to a large degree across the different...
sessions - with 21 (~ 40%) sessions recording a female speaking-time of less than 20%, showing the need for our proposed system to distinguish from simple attendance counts.

Furthermore, the percentage delegate speech detected is significantly higher in sessions moderated by a female chair - 34% (±5% std), compared to those moderated by a male chair - 23% (±13% std). This indicates the need for our system to identify potential moderator-dependent disparity in meetings.

VI. CONCLUSION

In this work, we developed a human-in-the-loop system to quantify how much female delegates speak in ITU meetings. We performed a systematic evaluation of the constituent systems in our pipeline to provide robust estimates of the target measures. We then deployed this system for the 2018 ITU Plenipotentiary conference to analyze the participation at this event. Our analysis shows an overall percentage of 25.12% of female delegate speaking time compared to an attendance rate of 27%. While this is only a first step towards understanding representation and inclusion in policy discussions, our framework can also applied to other forums such as debates and news panels. Future efforts include automating the moderator detection system and using natural language understanding technologies to model which delegate and how they influence the discussions in policy meetings.

REFERENCES


<table>
<thead>
<tr>
<th>Movie SAD</th>
<th>Arabic</th>
<th>Chinese</th>
<th>English</th>
<th>French</th>
<th>Russian</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAD MISS-RATE (MR) FOR DIFFERENT LANGUAGES</td>
<td>5.15</td>
<td>1.35</td>
<td>4.67</td>
<td>4.74</td>
<td>13.77</td>
<td>4.31</td>
</tr>
<tr>
<td>OpenSMILE</td>
<td>12.83</td>
<td>8.625</td>
<td>10.175</td>
<td>10.6</td>
<td>22.945</td>
<td>9.35</td>
</tr>
</tbody>
</table>

Fig. 3. Speaking time estimates for PP-18 sessions

Authorized licensed use limited to: University of Southern California. Downloaded on July 25,2021 at 22:00:26 UTC from IEEE Xplore. Restrictions apply.