AN OPEN-SOURCE SPEAKER GENDER DETECTION FRAMEWORK FOR MONITORING GENDER EQUALITY

David Doukhan, Jean Carrive
French National Institute of Audiovisual
Paris, France
{ddoukhan, jcarrive}@ina.fr

Félicien Vallet†
CNIL
Paris, France
fvallet@cnil.fr

Anthony Larcher, Sylvain Meignier
LIUM - Université du Maine
Le Mans, France
{firstname.name}@univ-lemans.fr

ABSTRACT

This paper presents an approach based on acoustic analysis to describe gender equality in French audiovisual streams, through the estimation of male and female speaking time. Gender detection systems based on Gaussian Mixture Models, i-vectors and Convolutional Neural Networks (CNN) were trained using an internal database of 2,284 French speakers and evaluated using REPERE challenge corpus. The CNN system obtained the best performance with a frame-level gender detection F-measure of 96.52 and a hourly women speaking time percentage error bellow 0.6%. It was considered reliable enough to realize large-scale gender equality descriptions. The proposed gender detection system has been packaged as an open-source framework.

Index Terms— Digital Humanities, Speech Gender Detection, Convolutional Neural Network, i-vector

1. INTRODUCTION

Gender equality in media is a concern which has been analyzed through various descriptors such as the proportion of male and female employees in media (journalists, top decision-making posts) [1], speaker roles (expert, interviewer), topics covered [2] or correlation between the proportion of male and female speakers and the amount of audience [3]. Among the descriptors of gender equality, male and female speaking-time duration has been used in Global Media Monitoring Project world-scale study [4]. Due to the cost associated to the manual annotation of speech, this monitoring is restricted to the analysis of a single day every five years, and limited to news-related content. Such a limitation induces biases, such as the particular context in which these measures were made, which may influence the topics covered in the news, as well as the amount of male and female speakers. Moreover, it describes a limited set of programs that does not fully reflect audiovisual diversity.

In the present work, we design an automatic speaker gender detection system for large scale monitoring of raw audiovisual streams in order to reduce the bias associated to a particular measure context. Quantitative gender speaking time estimates may help describing the evolution of gender equality across time and could be used to guide qualitative studies. They could also enable TV channels and radio stations to monitor gender equality in their programs. Long term goal of this research is based on the assumption that quantitative description of gender equality in media may increase consciousness on those issues and result in societal changes.

As it is easier than other automatic speech-related tasks, gender detection's difficulty is often under-estimated. Female speech is generally associated to higher pitch, vowel formants located in higher frequencies and considered to be more breathy. Language-dependent acoustic variations between male and female speakers were also reported. Distinction between male and female speech is therefore not only due to anatomical differences, but also to a given socio-cultural context [5]. Consequently, the robustness of speech gender detection systems is still challenged by speakers having marked accents (regional, foreign), extreme pitch ranges, or speaking using non-standard intonation.

This study addresses three major issues: Which methods result in the most accurate speech-time estimates? Does the robustness of automatic systems allows to perform reliable gender equality descriptions based on men and women speech-time percentage? What knowledge can be gathered from audiovisual streams using this information?

2. RELATION TO PRIOR WORK

Speech gender detection systems have been used for a long time as a preprocessing step in Automatic Speech Recognition (ASR) engines, allowing the selection of gender dependent acoustic models [6, 7]. Gender detection was not necessarily considered as an end itself, but rather as a way of selecting models having the closest acoustic properties to the utterance being analyzed. More recent studies have used speaker gender detection as a preprocessing step for selecting gender-

† The views and opinions expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the CNIL.

specific emotion recognition engines [8], as well as a way of enhancing human-machine interactions by defining gender-specific interaction strategies [9].

Gaussian Mixture Models (GMM), trained with MFCCs have been used as a *de facto* standard for speech gender recognition since the 90's and are still used in most of modern systems [6]. Approaches based on i-vectors, compact representations of speaker's utterances, have been successfully used over the last few years for speaker recognition tasks [10]. They have recently shown better performances over GMMs for the detection of speaker gender, in a context where already segmented utterances were assumed to belong to a single speaker [9]. Other approaches based on Convolutional Neural Networks (CNN) were proposed and evaluated on utterances assumed to belong to a single speaker [11]. However, those approaches were trained using a limited amount of data, that does not allow to take advantage of deep architectures.

The study presented in this paper aims at designing a speaker gender detection system for large-scale analysis of raw audio streams. For this purpose, we compare the performances of GMMs, i-vector and CNN based systems. Audiovisual streams contain speech recorded in various conditions (telephone of variable quality, studio, spontaneous and prepared speech, speech-over-music, speech-over-noise) which should be handled robustly. While GMM and CNN approaches may be directly used on speech streams, i-vector based systems require a preliminary speaker segmentation step, in charge of splitting speech excerpts into homogeneous segments assumed to belong to a single speaker. The reliability of this pre-segmentation stage may affect the discriminative power of the resulting i-vectors.

3. TRAINING DATASET: INA'S SPEAKER DICTIONARY

Gender detection models were trained using INA's Speaker Dictionary [12, 13], which is to our knowledge the largest manually annotated speaker database issued from broadcast material reported in the literature, providing a definite technological advantage regardless of the algorithmic strategy used (see [14] for a recent review of speaker databases). This audiovisual corpus, constituted using a semi-automatic annotation protocol, contains about 32000 speech excerpts, corresponding to 1780 male (94 hours) and 494 female speakers (27h).

A first set of excerpts was collected from TV news broadcasted from 2007 to 2014. Speech streams were first segmented in speaker turns using LIUM's open-source software SpkDiarization, based on HAC-BIC clustering and Viterbi decoding [15]. An Optical Character Recognition system was used to match personality name appearing on screen to referenced people [16]. The corresponding excerpts were thus presented to annotators in charge of validating speech excerpts. A second set of archives was obtained from INA's collections, for a period of time ranging from 1957 to 2012. Queries based on famous personalities names were performed on `Ina.fr` website. Speaker diarization procedures were used on resulting archives and segments found in largest speaker clusters were presented to annotators. The last set of excerpts was obtained from 3 radio stations collected from 2012 to 2014. Extracted speech turns were associated to i-vectors, and matched to previously known speakers, before being manually validated.

4. GENDER DETECTION SYSTEMS

Gender detection processing pipeline is composed of several modules. A Speech/Music segmenter based on CNN is used to discard music and empty segments [17]. Features corresponding to speech segments are then computed using a common extraction framework. A simple energy threshold is used to discard frames associated with low energy. GMM, i-vector and CNN systems are then used to classify the remaining speech segments into male and female excerpts.

4.1. Feature Extraction

Feature extraction was realized on audio excerpts sampled at 16KHz using SIDEKIT [18]. 24 Mel-scaled filter-banks coefficients were computed on 25ms sliding windows with a 10ms shift. Those coefficients were fed directly to CNN models while a second set of acoustic features (MFCC) was obtained by applying Discrete Cosine Transform (DCT) to those filter-bank coefficients to extract 19 Mel-Frequency Cepstrums coefficients (MFCC), which were augmented with log energy. A feature warping method was used to gaussianize MFCC distributions using windows of 3 seconds [19]. Time derivatives (Δ) and acceleration ($\Delta\Delta$) coefficients were concatenated to MFCC frames, resulting in 60-dimensional input features used by GMM and i-vector models.

4.2. Gaussian Mixture Models

GMMs are the most common approach used for modeling speaker gender, and were used as a baseline for evaluating more complex models. Female and male GMM's were trained using 1024 gaussians and diagonal covariance matrices. Each GMM was trained using 2000 randomly selected frames for each speaker found in the training set. Resulting models were used in a 2 state Hidden Markov Model (HMM) to segment speech into male and female excerpts.

4.3. I-vector gender segmentation system

Our i-vector-based gender segmentation system is composed of three main modules. A speech diarization module is firstly used to segment speech into homogeneous speaker turns. An i-vector extractor is then used in a second stage to model each

speaker turn as a compact vector, assumed to contain discriminative speaker information. Lastly, a classifier trained with i-vectors is used to predict the gender associated to a given speaker turn.

The diarization procedure was realized using S4D [20]. Speech segments were pre-segmented using Gaussian Divergence criterion, with a 2.5 second-window. Segments assumed to belong to the same speaker were merged through two clustering procedures based on Bayesian Information Criterion: a first procedure of Linear clustering, merging adjacent candidates, followed by a global Hierarchical Agglomerative Clustering procedure. Viterbi re-segmentation procedure was carried on the resulting clusters. Clustering algorithms and Viterbi re-segmentation were used with default thresholds (2, 3, -250). I-vector extraction was realized using SIDEKIT [18]. A 1024 gaussians Universal Background Model (UBM) with diagonal covariance matrices was trained using 2500 randomly selected frames per female and 700 frames per male speaker, in order to deal with the gender balance of the training dataset. The total variability matrix was defined with rank 200, and trained using 10 randomly selected excerpts per speaker. Training excerpts used for the estimation of the total variability matrix were used to extract i-vectors, which were fed into a linear SVM in charge of predicting speaker gender. Training examples were weighted according to the gender distribution found in the training examples.

4.4. Convolutional Neural Network system (CNN)

Figure 1 describes the deep CNN architecture which was implemented using Keras to discriminate between male and female speech [21]. It is inspired from speech recognition models [22], and differs by using filter banks frames having lower dimensionality, associated to a larger time context of 68 frames (695 ms, 24 dimensions)[1]. Two 3x3 convolutional layers of 64 neurons are first used, followed by a 1x2 max-pooling layer, aimed at providing invariance in the frequency domain. Two additional 3x3 layers of 128 neurons are then followed by a 2x2 max-pooling layer and a last 3x3 convolutional layer of 256 neurons. A maximal temporal pooling layer, in charge of selecting the most discriminating pattern found in this relatively large time interval is finally used, and followed by 4 dense layers of 512 neurons, associated to increasing dropout rates, before the last softmax layer outputting 1 probability per gender. Each layer of the network is followed by batch normalizations and RELU activations.

A data selection strategy aimed at using all available training data with limited over-fitting for a particular gender or speaker was used, consisting in generating 1000 examples batches, composed of 500 randomly examples of distinct speakers of each gender. Epochs were defined as an amount of 1,000 batches.

Model selection consisted in splitting INA's speaker dictionary into training and validation subsets containing respectively 80% and 20% of disjoint speakers. Models' performances were assessed after each epoch on the validation set, using speaker sampling strategies. F-measure was used as performance criterion, and training iterations were stopped after 5 epochs not resulting in performance improvement. Best models were obtained using less than 10 epochs. CNN's output was then used to feed a 2 states HMM.

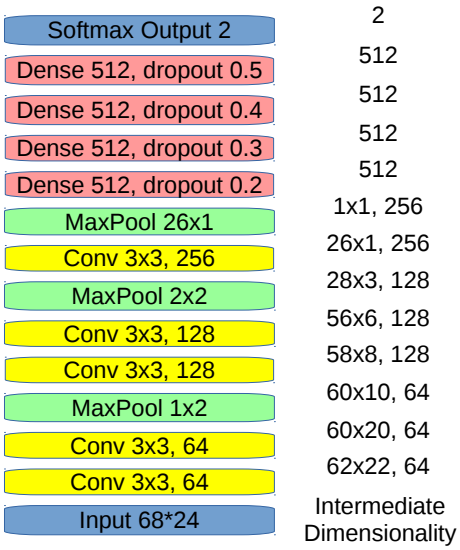

Fig. 1. CNN gender detection architecture, using input features of 68 concatenated 24-dimension filter banks frames.

5. EVALUATION

5.1. Evaluation Material

REPERE challenge corpus contains annotated TV streams gathered from two French channels from 2011 to 2013 [23]. This includes street interviews, debates and news corresponding to 46 hours of speech attributed to 1129 male speakers, and 12 hours attributed to 557 female speakers[2]. 457 speakers found in REPERE corpus were also found in INA's speaker dictionary used for training, and 606 speakers were associated to a gender label, without identity annotation.

5.2. Frame-Level Evaluation

Table 1 presents speaker gender recognition frame-level performances obtained on REPERE corpus. I-vector speech-turn estimates were mapped to frame-level with same sampling frequency than GMM and CNNs. Since small pauses are not systematically annotated in ground-truth, this evaluation

[1] Additional experiments were realized using 40 dimension filter banks, as well as compressed log-spectrograms [11], resulting in lower performances.

[2] 92 speech segments corresponding to 22 minutes of speech were found to be associated to wrong gender annotations and were corrected manually.

Model	Men recall	Women recall	F-score	Accuracy
GMM	97.25	94.28	95.74	96.63
I-vector	98.19	92.98	95.51	97.11
CNN	98.04	95.05	**96.52**	**97.42**

Table 1. Gender Detection frame-level Performances obtained on REPERE challenge corpus

was carried on the subset of frames annotated as speech in REPERE and predicted as speech by the speech/music segmenter, resulting in a binary classification problem. Best results were obtained with the CNN model, which was associated to a F-measure of 96.52 and accuracy percentage of 97.42. All models were associated to better recall rates for male speakers, which may lead to systems underestimating female speaking time. This may be explained by the largest diversity of male speakers found in the training database.

5.3. End-to-End Evaluation

This evaluation describes the ability of models to optimize Women Speaking-Time Percentage (WSTP) estimation on variable length raw audiovisual excerpts. WSTP is the main descriptor retained in our ongoing gender equality studies, and was computed for each annotated stream found in REPERE challenge dataset using the following formula:

$$\text{WSTP} = 100 \times \frac{\text{female speech time}}{\text{male speech time} + \text{female speech time}} \quad (1)$$

Frame-level male and female gender detection errors may counter-balance and provide robust WSTP estimators for a reasonably long time-interval. Based on this assumption, REPERE shows were split into four categories according to their duration, resulting in 133 excepts of less than 2 minutes, 82 excerpts of duration varying between 2 and 10 minutes excerpts, 104 10-30 minutes excerpts and 25 30-60 minutes excerpts. Figure 2 shows Root Mean Square Error (RMSE) between estimated and reference WSTP, obtained using variable length recordings. As expected, longer excerpt durations were associated to lower WSTP estimation errors. Best results were obtained with the CNN model, associated to a RMSE of 0.68 for excerpts longer than 30 minutes together with WSTP estimation mean and worst-case error rates as low as 0.59% and 1.8%.

Maximal error-rates were found to be always lower for CNN than for i-vector model. This phenomenon may be explained by the architecture of the i-vector model, relying on a baseline diarization step, which may miss small speaker turns, resulting in i-vectors corresponding to several speakers.

6. CONCLUSION AND FUTURE WORK

A deep CNN model has been realized to segment speech streams into male and female excerpts and compared to GMM and i-vector models on REPERE corpus, showing superior performances. The proposed CNN model is accessible from INA's GitHub repository: `http://github.com/ina-foss`.

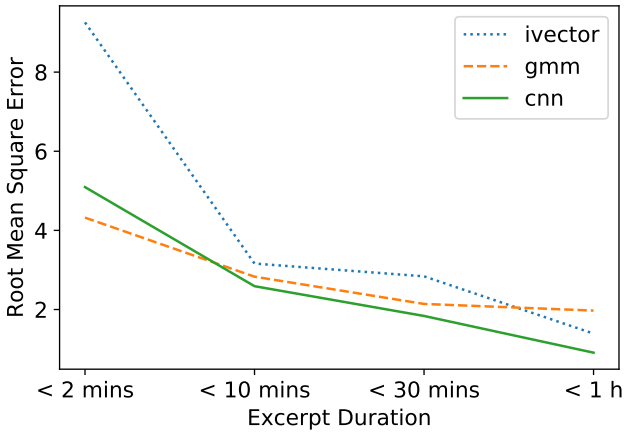

Fig. 2. Root Mean Square Error associated to the estimation of women speaking time percentage, obtained on variable-length shows found in REPERE corpus.

This positive outcome allows considering the reliability of this gender detection system sufficient to perform large-scale gender equality descriptions, offering concrete perspectives for digital humanities. Detailed analyses realized using this framework goes beyond the scope of this paper and are addressed in other studies [24]. Results obtained using 580.000 hours of French TV and radio streams broadcasted from 2001 to 2017 highlighted several interesting tendencies: Speaking time was mainly attributed to male speaker, especially in radio material. Women speaking time was lower during peak viewing time. Percentage of women speaking time has increased of about 0.5% per year between 2001 and 2017. Proportion of male speakers was higher in sport, cultural and educational TV channels. Longer term studies may take advantage of INA's available meta-data (detailed audience, type of program, speaker identities) to propose deeper analyses.

The impact of our evaluation is limited to the properties of REPERE corpus, containing only broadcast news archives. It does not include child speakers, nor very specific speech utterances such as found in movies, comedy sketches or speaker imitations, which should be addressed in future studies. Improvements of the proposed i-vector system could be investigated using more sophisticated diarization methods [25]. The use of recurrent neural network units in CNN's last layers, replacing the final temporal max-pooling layer proposed will also be investigated, together with training example balancing strategies based on pitch estimation and data augmentation.

Since speaking time is not sufficient to describe fully gender equality, longer-term research may include using ASR to describe gender roles and interactions.

7. REFERENCES

[1] Gender equality commission, "Handbook on the implementation of recommendation cm/rec(2013)1 of the committee of ministers of the council of europe on gender equality and media," *Council of Europe*, 2015.

[2] Michèle Reiser and Brigitte Gresy, "L'image des femmes dans les médias," 2008.

[3] Conseil supérieur de l'audiovisuel (CSA), "La représentation des femmes à la télévision et à la radio - rapport sur l'exercice 2016," 2017.

[4] Sarah et al. Macharia, *Who Makes the News?: Global Media Monitoring Project 2015*, World Association for Christian Communication, 2015.

[5] Erwan Pépiot, "Voice, speech and gender:. male-female acoustic differences and cross-language variation in english and french speakers," *Corela. Cognition, représentation, langage*, , no. HS-16, 2015.

[6] Lori F Lamel and Jean-Luc Gauvain, "A phone-based approach to non-linguistic speech feature identification," *Computer Speech & Language*, vol. 9, no. 1, pp. 87–103, 1995.

[7] Tobias Bocklet, Andreas Maier, Josef G Bauer, Felix Burkhardt, and Elmar Noth, "Age and gender recognition for telephone applications based on gmm supervectors and support vector machines," in *Acoustics, Speech and Signal Processing, ICASSP*. IEEE, 2008, pp. 1605–1608.

[8] Rui Xia, Jun Deng, Bjorn Schuller, and Yang Liu, "Modeling gender information for emotion recognition using denoising autoencoder," in *Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 990–994.

[9] Laurent El Shafey, Elie Khoury, and Sébastien Marcel, "Audio-visual gender recognition in uncontrolled environment using variability modeling techniques," in *International Joint Conference on Biometrics (IJCB)*. IEEE, 2014, pp. 1–8.

[10] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[11] Jimena Royo-Letelier, Romain Hennequin, and Manuel Moussallam, "Detection and characterization of singing voice using deep neural netwo rks," *UPMC-Paris*, 2015.

[12] François Salmon and Félicien Vallet, "An effortless way to create large-scale datasets for famous speakers.," in *LREC*, 2014, pp. 348–352.

[13] Félicien Vallet, Jim Uro, Jérémy Andriamakaoly, Hakim Nabi, Mathieu Derval, and Jean Carrive, "Speech trax: A bottom to the top approach for speaker tracking and indexing in an archiving context.," in *LREC*, 2016.

[14] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: A large-scale speaker identification dataset," *Proc. Interspeech 2017*, pp. 2616–2620, 2017.

[15] Sylvain Meignier and Teva Merlin, "Lium spkdiarization: an open source toolkit for diarization," in *CMU SPUD Workshop*, 2010, vol. 2010.

[16] Johann Poignant, Laurent Besacier, Georges Quénot, and Franck Thollard, "From text detection in videos to person identification," in *Multimedia and Expo (ICME)*. IEEE, 2012, pp. 854–859.

[17] David Doukhan and Jean Carrive, "Investigating the Use of Semi-Supervised Convolutional Neural Network Models for Speech/Music Classification and Segmentation," in *The Ninth International Conferences on Advances in Multimedia (MMEDIA 2017)*, Apr. 2017.

[18] Anthony Larcher, Kong Aik Lee, and Sylvain Meignier, "An extensible speaker identification sidekit in python," in *Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5095–5099.

[19] Jason Pelecanos and Sridha Sridharan, "Feature warping for robust speaker verification," 2001.

[20] Sylvain Meignier and Anthony Larcher, "S4d: Sidekit for speaker diarization," `http://lium.univ-lemans.fr/sidekit/s4d/`, 2015.

[21] François Chollet et al., "Keras," `https://github.com/fchollet/keras`, 2015.

[22] Yanmin Qian, Mengxiao Bi, Tian Tan, and Kai Yu, "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2263–2276, 2016.

[23] Aude Giraudel, Matthieu Carré, Valérie Mapelli, Juliette Kahn, Olivier Galibert, and Ludovic Quintard, "The repere corpus: a multimodal corpus for person recognition.," in *LREC*, 2012, pp. 1102–1107.

[24] David Doukhan, Géraldine Poels, and Jean Carrive, "Describing gender equality in french audiovisual streams with a deep learning approach (accepted)," *Journal of European Television History and Culture (VIEW)*, 2018.

[25] Gaël Le Lan, Delphine Charlet, Anthony Larcher, and Sylvain Meignier, "A triplet ranking-based neural network for speaker diarization and linking," *Proc. Interspeech 2017*, pp. 3572–3576, 2017.