Gender representation in French broadcast corpora and its impact on ASR performance

Mahault Garnerin, Solange Rossato, Laurent Besacier
– Université Grenoble Alpes –

AI4TV Workshop – ACMM 2019 – 21st October 2019

Context of the research

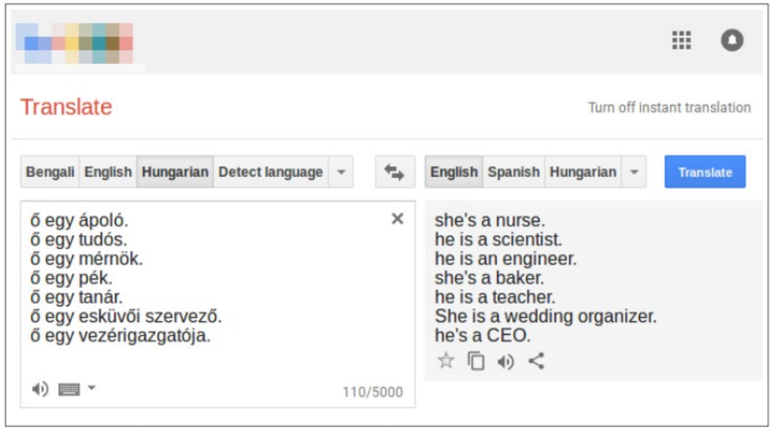
Gender has become a hot topic within the political, social and research spheres

« Gender refers to the socially constructed characteristics of women and men – such as norms, roles and relationships of and between groups of women and men. It varies from society to society and can be changed. » (definition given by WHO)

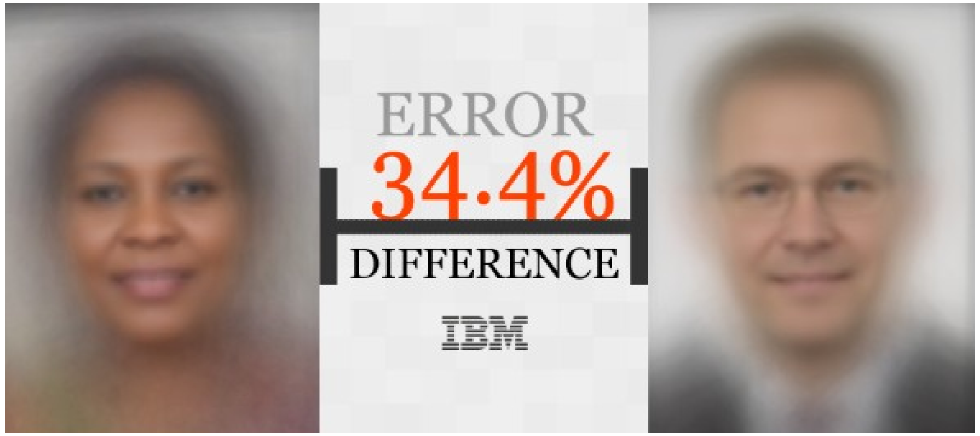
Although the definition of gender can now also englobe people identifying outside of the binary categories, we will only consider the ‘men’ and ‘women’ gender labels in this work

Context of the research

- Gender bias discovered in many AI technologies :
  - Machine translation (Vanmassenhove et al. 2018 ; Prates et al., 2019)
  - Word-embeddings (Bolukbasi et al., 2016 ; Caliskan et al., 2017)
  - Facial recognition (Buolamwini & Gebru, 2018)



From Prates et al., 2019



From GenderShades (Buolamwini & Gebru, 2018)

Context of the research

Why so ?

→ **Data representativity/exhaustivity**

« A data set may have many millions of pieces of data, but this does not mean it is random or representative. To make statistical claims about a data set, we need to know where data is coming from ; it is similarly important to know and account for the weaknesses in that data. » (Boyd & Crawford, 2012)

« A lot of people are saying this is showing that AI is prejudiced. No. This is showing we're prejudiced and that AI is learning it. » (J. Bryson during an interview for The Guardian)

Gender bias in the media

As we are working on **TV and radio broadcast**, we know our data is **unbalanced regarding to gender**

- Global Media Monitoring Project (Macharia et al., 2015)
- CSA report on French media : under-representation of women on TV (CSA , 2018)
- Automatic gender monitoring of French audiovisual streams (Doukhan et al. 2018)

→ **Does an ASR system trained on such data exhibit a gender bias ?**

Our approach

1. How is gender represented in our data?

Hypothesis : *Men are more represented as we work on radio and TV broadcast*

2. How is our system performing on these gender categories? Is some gender bias exhibited by our system?

Hypothesis : *we expect better performances for male speakers, as they are more represented in our dataset*

Our approach

1. How is gender represented in our data?

***Hypothesis :* Males are more represented as we work on radio and TV broadcast**

2. How is our system performing on these gender categories? Is some gender bias exhibited by our system?

Hypothesis : we expect better performances for male speakers, as they are more represented in our dataset

Data presentation

French Radio and TV broadcast from major evaluation campaigns :

- ESTER1 (Galliano et al, 2005)
- ESTER2 (Galliano et al., 2009)
- ETAPE (Gravier et al., 2012)
- REPERE (Giraudel et al., 2012)

~ 300h of manually transcribed recordings (4597 speakers)

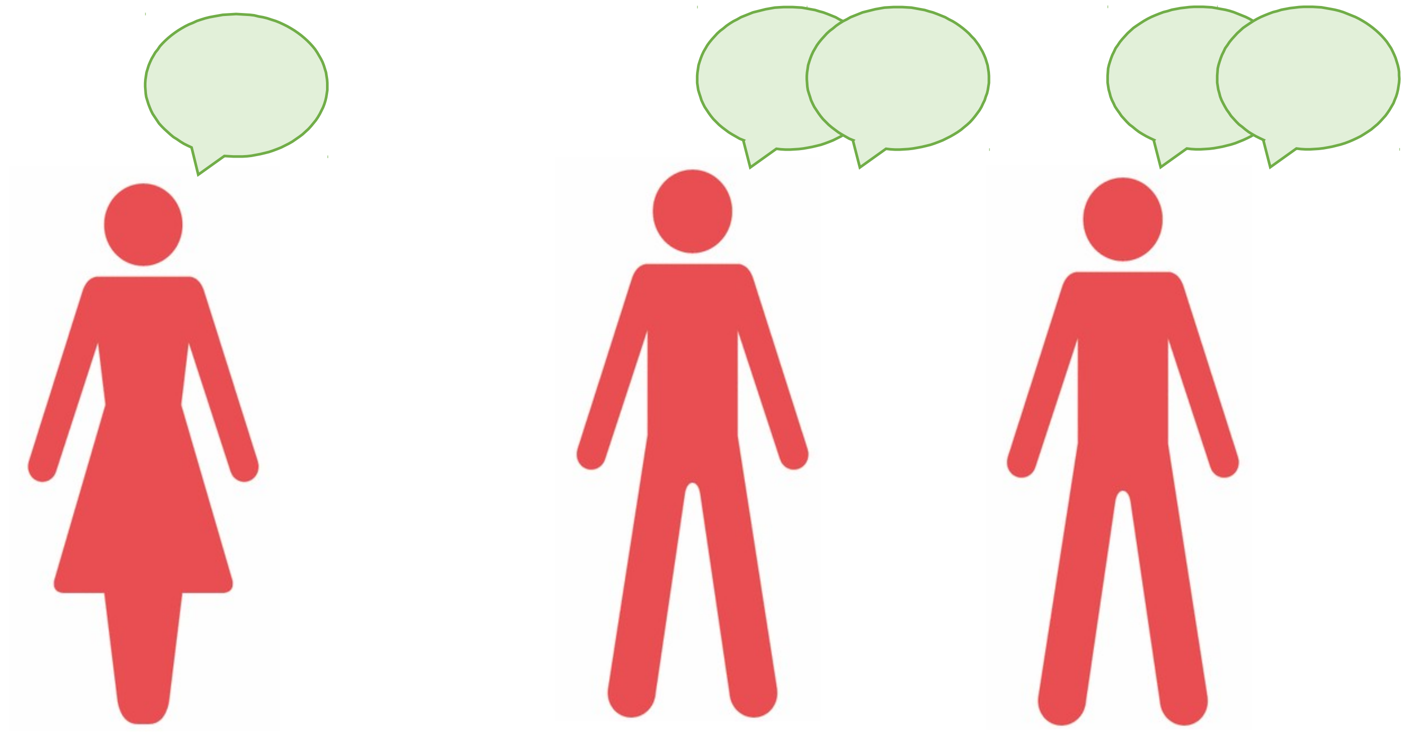
Meta-data available : show name, date, name and gender of speaker (M/F)

Gender representation in our data

65% male et 35% female speakers corroborating previous studies (CSA, 2018 ; Macharia et al., 2015)

Women tend to speak only half as much as men when counting number of speech turns (as in Doukhan et al., 2018)

No significant difference on speech turns length between gender (but observable difference between type of media)



Gender and roles

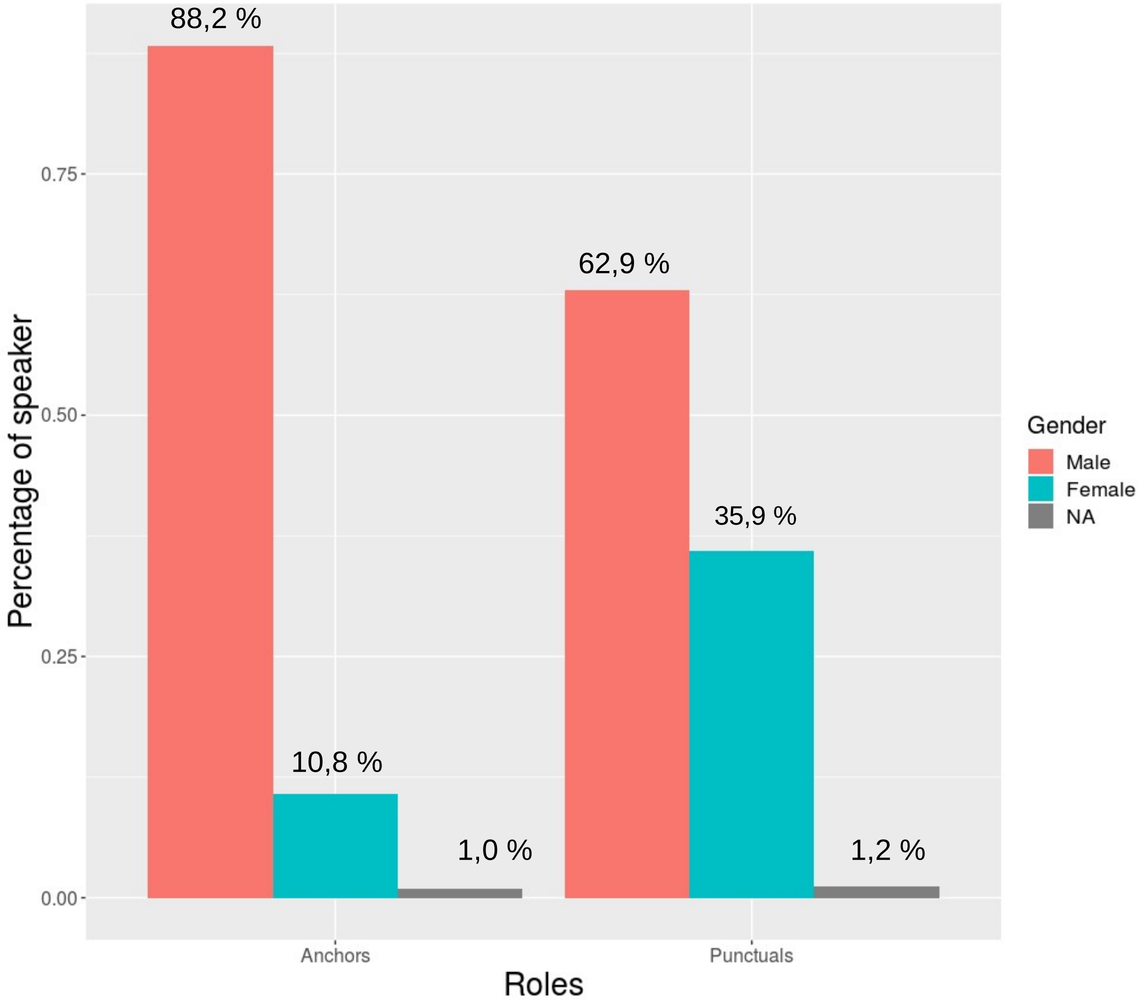
Speaker roles as a proxy for data availability

- <u>Anchor speakers:</u> above number of turn and duration thresholds
- <u>Punctual speakers</u>: below nomber of turn and duration thresholds

We focused on these two extreme categories as they were few speakers outside of them and there is no real-world interpretation of their status

Gender gap is even bigger as media exposure increased.

→ Gender representation is **dependant on the speaker role**



Gender in data : a summary

- **Gender unbalance observed in our data** in terms of speakers and speech time (men more represented in our data)

- This gender gap is even bigger when looking at Anchors speakers

→ This directly leads to a **gap in data available** for both gender categories. **How does it impact our model ?**

Our approach

1. How is gender represented in our data?

Hypothesis : Males are more represented as we work on radio and TV broadcast

2. How is our system performing on these gender categories? Is some gender bias exhibited by our system?

Hypothesis :* *we expect better performances for male speakers, as they are more represented in our dataset

System description

- ASR system developped at the LIG (Elloumi et al., 2018) using a subset of the four corpora
- State of the art, Kaldi-based system (Povey et al., 2011)
- Hybrid HMM-DNN acoustic model
- 5-gram language model

Training data

Show	Duration	Medium	Type
BFM Story	25h 36min	TV	P
France Info Infos	11h 23min	Radio	P
France Inter Infos	42h 45min	Radio	P
LCP Infos	10h 6min	TV	P
RFI Infos	1h 49min	Radio	P
Top Questions	7h 59min	TV	P
Total	**99h 38min**	**-**	**P**

Evaluation data

Show	Duration	Medium	Type
Africa1	1h 21min	Radio	P
Comme On Nous Parle	2h 14min	Radio	S
Culture et Vous	1h 16min	TV	S
La Place du Village	1h 24min	TV	S
Le Masque et la Plume	4h 12min	Radio	S
Pile et Face	7h 52min	TV	P
Planète Showbiz	1h 12min	TV	S
RFI Infos	24h 14min	Radio	P
RTM Infos	22h 00min	Radio	P
Service Public	2h 30min	Radio	S
TVME Infos	57min	Radio	P
Un Temps de Pauchon	1h 31min	Radio	S
Total	**70h 43min**	**-**	**-**

Evaluation metric

Word Error Rate (WER)

$$\text{WER} = \frac{\text{insertions} + \text{substitutions} + \text{deletions}}{\text{number of words in the reference transcription}}$$

REF	mais	je	connaissais	absolument	rien	au	milieu	de-la	recherche	donc	c'	est	c'	est	quelque-chose	mais	c'	est	vraiment	***	parce-que	la	thématique	m'	a	intéressé
HYP	mais	je	connais	absolument	rien	au	lieu	de-la	recherche	donc	c'	est	c'	est	quelque-chose	mais	c'	est	vraiment	à	parce-que	la	thématique	m'	***	intéressait
OP	C	C	S	C	C	C	S	C	C	C	C	C	C	C	C	C	C	C	C	I	C	C	C	C	D	S

$$\text{WER} = \frac{1+3+1}{25} = \frac{5}{25} = 0{,}2 = 20\,\%$$

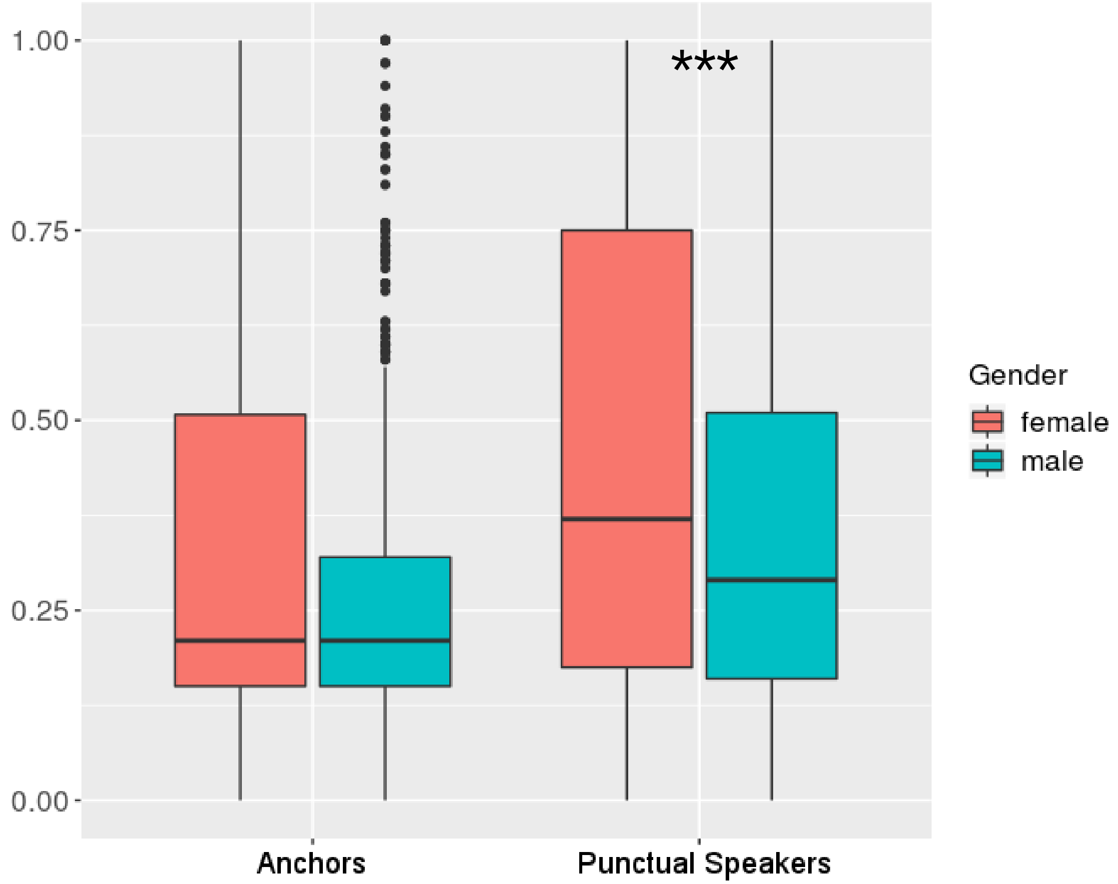
Evaluation procedure

Traditionnally, we compute the WER at the corpus level, meaning we sum all the errors made by the system divided by the total number of words

As we are working on gender, and **gender is a characteristic of an individual**, we decided to compute the **WER at the speaker level**, in a given episode of a show.

We then analyzed our results depending on our defined gender and roles categories

Results (1/2)



42,9 % WER for female speakers (N = 831)
34,3 % WER for male speakers (N = 1637)

When crossing with role factor, **better performance for male speaker within the Punctual** speaker category

But when looking at the Anchor category, the difference between gender is not significant anymore (role of speaker adaptation?)

→ When the **system works poorly it is even worse on women voice**

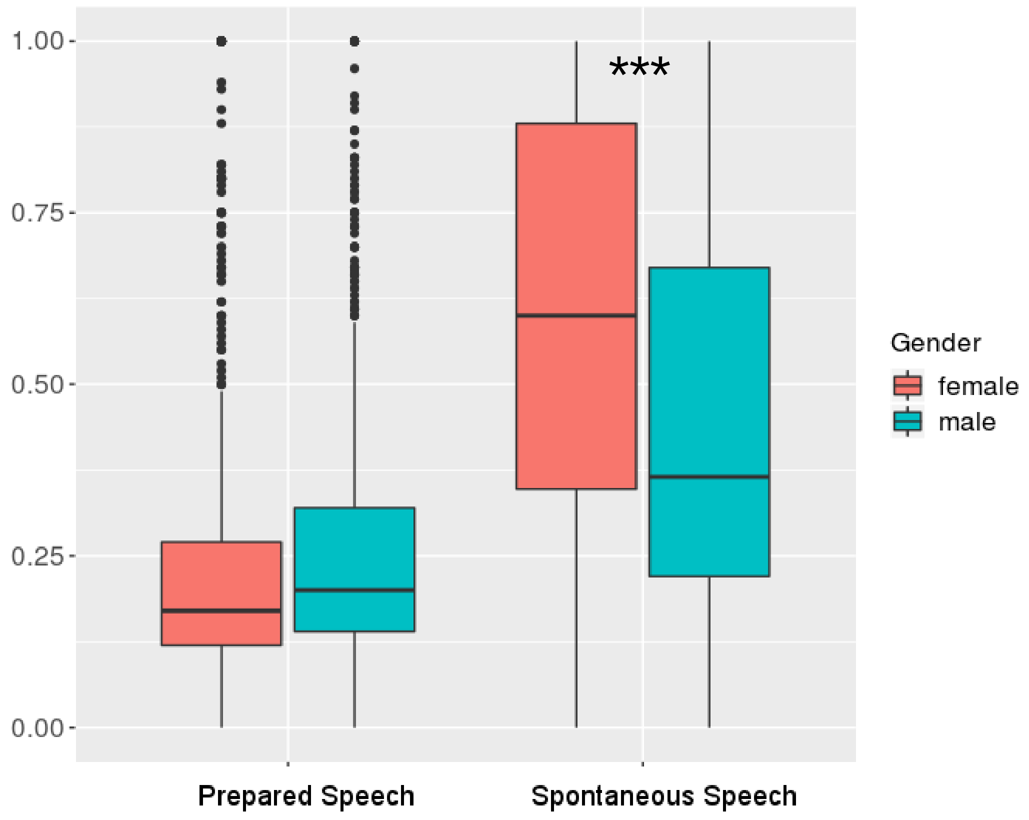
Results (2/2)

Prepared VS Spontaneous speech

Performance are better for prepared speech (data type aligned with training set)

Better performance on women for prepared speech (not stat. sig.) but trend is reversed in spontaneous setting

We can assume that **our canonical system is made for men uttering prepared speech**, if we step away from these two characteristics, performance decrease drastically



Limits and future work

- High standard deviation when looking at female Anchors speaker, maybe our definition of roles needs to be refined?

- How to **compensate this gender gap** in data? Speaker adaptation, data augmentation, adversarial learning, etc.?

- What is the **impact of such difference** in performance?
  → poorer transcription for women speaking in the media
    → less indexation of content starring women → **invisibilisation of women in media**
    → poorer quality when subtitling → **misrepresentation of women speech in media**

Thank you for your attention!

Gender representation in French broadcast corpora and its impact on ASR performance

Mahault Garnerin, Solange Rossato, Laurent Besacier
– Université Grenoble Alpes –

AI4TV Workshop – ACMM 2019 – 21st October 2019

References

T. Bolukbasi, K. Chang, J.Y. Zou, V. Saligrama and A.T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the 30 th Conference on Neural Information Processing Systems (NIPS 2016)*. 4349–4357.

Danah Boyd and Kate Crawford. 2012. Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Information, communication & Society* 15, 5 (2012), 662–679.

J. Buolamwini and T. Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the Conference on Fairness, Accountability and Transparency (ACM FAT 2018). 77–91.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

Conseil Supérieur de l'Audiovisuel (CSA). 2018. La Représentation des Femmes à la Télévision et à la Radio. Rapport d'Exercice 2017. Retrieved July 8, 2019 from https://en.calameo.com/read/00453987548c2c813939e?page=1

D. Doukhan, J. Carrive, F. Vallet, A. Larcher, and S. Meignier. 2018. An open-source speaker gender detection framework for monitoring gender equality. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018). 5214–5218.

S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J-F. Bonastre, and G. Gravier. 2005. The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. In Proceedings of the 9th European Conference on Speech Communication and Technology (INTERSPEECH 2005). 1149–1152.

S. Galliano, G. Gravier, and L. Chaubard. 2009. The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. In Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH 2009). 2583–2586.

References

A. Giraudel, M. Carré, V. Mapelli, J. Kahn, O. Galibert et L. Quintard. 2012. The REPERE Corpus: a multimodal corpus for person recognition. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012). 1102–1107.

G. Gravier, G. Adda, N. Paulson, M. Carré, A. Giraudel and O. Galibert. 2012. The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012). 114–118.

Z. Elloumi, L. Besacier, O. Galibert, J. Kahn, and B. Lecouteux. 2018. ASR Performance Prediction on Unseen Broadcast Programs using Convolutional Neural Networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018). 5894–5898.

S. Macharia, L. Ndangam, M. Saboor, E. Franke, S. Parr, and E. Opoku. 2015. Who Makes the News. Global Media Monitoring Project 2015. Retrieved July 8, 2019 from http://cdn.agilitycms.com/who-makes-the-news/Imported/reports_2015/global/gmmp_global_report_en.pdf

D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer and K. Vesely. 2011. The Kaldi Speech Recognition Toolkit. (2011).

Marcelo O. R. Prates, Pedro H. C. Avelar, and Luís C Lamb. 2018. Assessing Gender Bias in Machine Translation - A Case Study with Google Translate. (2018). arXiv:1809.02208 http://arxiv.org/abs/1809.02208

E. Vanmassenhove, C. Hardmeier, and A. Way. 2018. Getting Gender Right in Neural Machine Translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2018). 3003–3008.