Who's speaking? Predicting speaker profession from speech

Yaru Wu, Lihu Chen, Benjamin Elie, Fabian M. Suchanek, Ioana Vasilescu, Lori Lamel

WHO'S SPEAKING? PREDICTING SPEAKER PROFESSION FROM SPEECH

Yaru Wu[1,3,4,*], Lihu Chen[2,*], Benjamin Elie[3,*], Fabian Suchanek[2], Ioana Vasilescu[3], Lori Lamel[3]

[1]CRISCO/EA4255, Université de Caen Normandie, 14000 Caen, France
[2]Télécom Paris, Institut Polytechnique de Paris, France
[3]LISN, Univ. Paris-Saclay, 91405 Orsay cedex, France
[4]Laboratoire de Phonétique et Phonologie (UMR7018, CNRS-Sorbonne Nouvelle), France

yaru.wu@unicaen.fr, lihu.chen@telecom-paris.fr, elie@lisn.fr, fabian.suchanek@telecom-paris.fr,
{ioana.vasilescu, lori.lamel}@lisn.fr

ABSTRACT

Variations in speech can reveal the gender, birth place, age, and socio-economic level of the speaker. In this paper, we show that even the profession of the speaker can be recovered from a recording. For this purpose, we design a method that combines features from both the speech signal and the transcription. For the features from the transcription, we used pre-trained language models. This allows us to train a model that predicts the speaker profession from both signals. Our empirical results show that our model can narrow down the profession of the speakers considerably.

Index Terms: knowledge base, large corpora, multimodal representation, pre-trained language models

1. INTRODUCTION

The recording of someone's speech provides two principal types of signals: the content of the speech (*what* is being said), and the acoustic and prosodic features of the speech (*how* it is being said). The latter features cover both segmental and supra-segmental levels and are reflected in various measures related to the acoustic properties of vowels and consonants, as well as to the prosodic specificity of such units (e.g., segment duration, pitch, speech rate, etc.).

It is known that these features can help predict the background of the speakers, most notably the age, gender, and the region of birth [1, 2, 3]. In this work, we take this analysis one step further: we show that even the profession of the speaker can to some degree be determined from a recording alone. We can show that the content of the speech alone is not sufficient for this purpose. Nor are the speech-based features. However, if both features are combined, we can predict the profession of the speaker with an accuracy of up to 76%. It matters thus not only *what* the speaker says, but also *how* they say it.

Our approach takes as input a speech recording of a speaker and its transcription. It then extracts features from both the speech signal and the transcription, using pre-trained language models for the latter. We then use a machine learning model to predict the speaker profession from these features. For our experiments, we use a corpus of French oral broadcast news from radio and TV channels recorded in the 2000s. As ground truth for training and evaluation, we use profession annotations from the YAGO knowledge base [4] that have been found by previous work [5].

The ability to predict the profession of the speaker can have implications for AI applications such as speaker classification, speaker identification, dialog systems, intelligent assistants etc. It also has ramifications in the domain of privacy: it shows that machines can identify a socio-economic characteristic of speakers, possibly unbeknownst to them or against their will. It is thus important to investigate to what degree such privacy violations are possible with today's means.

2. RELATED WORK

Variation is ubiquitous in human language and is linked to both physiological and psychological properties of the individuals who communicate. According to [6], physiological properties are related to age, gender, health etc., whereas psychological ones are related to their identity as individuals and position or role within a community. Both properties translate into an array of features ranging from very basic (physical) properties of the acoustic speech signal, to high level features (the semantics of the utterances). On the speech processing side, the features cover both acoustic and prosodic levels, ranging from acoustic parameters at the phone level (for instance, the formants of vowels), to prosodic parameters such as duration, fundamental frequency and intensity [7, 8, 9, 10, 11]. Of these features, acoustic properties of consonants and vowels as well as fun-

*These authors contributed equally to this work.

damental frequency related measures appeared to be speaker specific and have allowed to quantify inter-speaker variability [7, 11].

On the textual side, Natural language processing (NLP) methods have been used for a variety of prediction tasks about the speaker from discourse, such as political profiling [12], ideology classification [13], candidate to job matchings [14], or stance detection [15]. Closer to our approach, recent approaches combine speech and NLP features for more accurate prediction of speaker characteristics. For instance, in [16] the speaker stances in Mandarin ideological debate competitions are predicted by combining n-grams and acoustic prosodic features. In a similar direction, another approach [17] combines text and acoustics for ideology detection in Youtube data. However, the prediction of a socio-economic indicator as precise as the speaker profession has so far been out of reach.

Closest to our approach is the work of [5], whose dataset we use here. While that work found correlations between speech features and professions, it was unable to *predict* the profession from the speech features. This is because the speech features alone are not sufficient for that task, as our work shows. Only the combination with the speech content that we propose here allows a narrowing-down of the profession.

3. DATA AND METHODOLOGY

Given as input a recording of speech by a speaker with its transcription, and given a list of professions, our goal is to predict the profession of the speaker. Our approach proceeds in three steps: (1) feature extraction from the speech, (2) feature extraction from the transcription of the speech, and (3) classification of the feature vector to one of the professions. To train the model of step (3), and to evaluate our method, we need a ground truth, i.e., an annotation of speakers with professions for part of the data.

3.1. Feature extraction from speech

The first step of our method extracts features from speech signal. We focus on global features, computed at the speech chunk level. This level corresponds roughly to an utterance. This choice is made in order to avoid considering speaker-specific features that could be irrelevant for a speaker profession recognition. Computing global features at a larger time-scale also enables us to extract psychological variations of speech. The following features were extracted from the speech signal : speech rate, mean pitch, pitch span, pitch declination, normalized pitch peak extent, local peak dynamics of pitch, mean for each of the 4 formants, formant span for each of the 4 formants, voicing ratio and devoicing ratio. Speech rate is calculated on speech utterances between pauses that are equal to or greater than 200ms. The pitch and formant span are defined as the difference between the average distance between the maximums and the local mean contour and the average distance between the minimums and the local mean contour [18]. We extracted and computed the characteristics of the 4 first formant trajectories – hence the 4 mean formants and 4 formant spans in the table. Voiced and unvoiced ratio relate to non-canonical surface forms, corresponding to stops that changed their voicing category. Voicing refers to canonical /ptk/ pronounced [bdg], while unvoicing refers to canonical /bdg/ pronounced [ptk]. The estimation of non-canonical realizations is obtained with the method described in [19]. We then compute the ratio of produced non-canonical forms in each category (voicing and devoicing) by the speaker as marks for consonant reduction (voicing) or strengthening (devoicing) that may be speaker specific. Except for the voiced and unvoiced ratio, which are computed over all speech chunks of a particular speaker, features are computed at the speech chunk level, and then averaged over each speaker. Each speaker is then assigned the mean value and the standard deviation of each features. This results in a feature vector $\mathbf{s}$ of length $N = 30$ for a given speech recording.

3.2. Feature extraction from the transcription

Formally, the transcription of a speech recording is a word sequence of length l: $W = \{w_1, w_2, ..., w_l\}$. We extract a textual feature vector $\mathbf{t} \in \mathbb{R}^n$ from the sequence by using a encoder function $\phi(\cdot)$, $\mathbf{t} = \phi(W)$. The resulting vector $\mathbf{t}$ can be regarded as a document-level representation for the input transcription text. We adopt Pre-trained Language Models (PLMs) [20] as an encoder to extract textual features from transcriptions for speakers because PLMs have achieved the state-of-the-art across a series of Natural Language Processing tasks. More specifically, we use CamenBERT [21, 22], which targets French texts. The input transcription is truncated into word sequences with a maximum length of 500, and CamenBERT is applied to the input. Afterward, a pooling layer is used to the whole sequence of the last-layer hidden states to obtain the textual feature $\mathbf{t}$. We use a CamenBERT-large model[1] of dimension $n = 1048$. When a transcription is missing, we use a zero padding vector.

Profession	Politician	Journalist	Artist	Movie Person	Scientist	Lawyer	Business Person	Sports Person	All
Number of speakers	189	56	53	70	23	12	24	21	448
Utterances per speaker	505	1869	881	519	600	506	241	132	695

Table 1: Statistics of the top-8 frequent professions.

Profession	**Keywords**
Politician	France, politique, gouvernement, ministre, pays
Journalist	France, heures, président, Inter, soir
Artist	temps, film, films, cinéma, France
Movie Person	France, films, français, personnage, film
Scientist	France, monde, temps, Irak, Israel
Lawyer	vérité, société, France
Bussiness Person	travailler, monde, année
Sports Person	France, l'équipe, défense

Table 2: Top-k keywords extraction from all transcripts using YAKE [23]: Top-5 keywords for professions with sufficiently large samples; top-3 for those with fewer samples.

Profession	**Speech signal**	**Text**		**Both**	
	F1	F1	M	F1	Acc
Politician	70.9	65.9	41%	**77.2**	76.6
Journalist	22.6	38.4	29%	**44.0**	42.3
Artist	15.6	16.3	30%	**19.1**	13.6
Movie Person	49.6	8.1	74%	**60.1**	60.6
Scientist	6.7	**12.5**	26%	6.1	5.7
Lawyer	0.0	0.0	33%	**12.5**	15.4
Business Person	0.0	18.2	33%	**27.8**	22.9
Sports Person	14.8	**38.9**	14%	35.3	34.3
All	22.6	24.5	41%	**35.3**	33.9

Table 3: Performances (F1 score) of profession prediction with speech signal and textual features. Macro F1 and Accuracy for **Both** column. Column "M" gives the rate of speakers without transcriptions.

	Hit@1	Hit@2	Hit@3
Politician	66.3	92.9	96.3
Journalist	48.6	67.7	77.1
Artist	19.0	26.6	38.4
Movie Person	54.1	65.5	72.5
Scientist	22.2	30.1	30.3
Lawyer	0.0	10.2	20.1
Business Person	30.3	32.3	33.4
Sports Person	50.0	68.2	68.2
All	55.8	68.5	74.3

Table 4: Hit@K performance with combined speech and transcription features.

3.3. Model prediction

We now have, for the given speech recording, a feature vector $\mathbf{s} \in \mathbb{R}^m$ of speech features, and a feature vector $\mathbf{t} \in \mathbb{R}^n$ of text features. We combine both vectors by concatenation, and obtain a multimodal representation $\mathbf{u} \in \mathbb{R}^{m+n}$.

We then classify this vector using a Multi-Layer Perceptron (MLP) with one hidden layer [24]. We have tested different architectures and sets of input features by Leave-one-out cross-validation across speakers. The best results were obtained with a hidden layer size of 100, a RELU activation function, and a L2 penalty parameter of 1 for regularization. The number of finally selected features is reduced to 26, as some speech features are non-discriminant for this task. The rejected features are the standard deviation of the 3rd formant span, the mean value of the averaged 3rd formant, and the mean values of the averaged pitch and the pitch span. This can be explained by the fact that averaged values of the fundamental frequency and higher formants are expected to be more dependent on speaker characteristics (e.g. gender) than on speaker profession.

4. EXPERIMENTS

4.1. Dataset

We use the annotated dataset from [5]. It consists of the ESTER corpus [25] (80 hours of French formal journalistic speech, mainly from radio broadcast news) and the ETAPE corpus [26] (about 40 hours of less formal journalistic French speech data, mainly from debates and interviews). The two corpora were transcribed manually, and 66% of the utterances have an associated manual transcription. The transcription was force-aligned with the speech signal using the LISN (former LIMSI) speech transcription system [27, 28]. The speech transcription system was used to segment the speech data automatically to word and phone levels. Pauses, hesitations or breaths were detected automatically by the system. The minimum segment duration is 30ms, corresponding to 3 frames [29]. The dataset contains only utterances whose speaker could be identified. We combine all utterances per speaker, so that we arrive at one (pseudo-) recording per speaker. The dataset contains 342 distinct named speakers. For each speaker, the dataset also contains the ground truth in the form of a profession from the YAGO knowledge base [4]. Table 1 shows the statistics of the professions in our dataset.

4.2. Using only the speech signal

The performance results for profession prediction by the speech signals alone are shown in Table 3, in the column "Speech Signal". The best scores are obtained for *Politician* (F1=70.9) and *Movie Person* (F1=49.6). Both professions and in particu-

lar *Politician* are well represented in the two corpora. *Politicians* exhibit idiosyncratic patterns such as slow speech rate and high rate of disfluencies. They also have a tendency to consonant reduction through increased rate of non-canonical voiced consonants. This is in line with previous findings [5]. Beyond that, the median value of the speech rate across *Politicians* is the lowest among the professions. These trends can be associated with the profession. Reduction can be linked to a rather spontaneous and thus relaxed pronunciation and one can speculate that *Politicians* are used to speaking publicly and their speech is rather relaxed. *Movie Persons*, too, can be assimilated to trained public speakers and exhibit recognizable speaking traits resulting in a more spontaneous, relaxed speech. The scores for the other professions are much lower, indicating that speech signals alone cannot predict the profession.

4.3. Using only the speech transcriptions

The performance results for profession prediction using only the transcriptions are shown in the third column ("Transcription") of Table 3. We first observe that the textual features are more helpful than the speech features. The macro F1 has a lead of 1.9 absolute percentage points. To better understand what triggers this performance, we extracted the top keywords of each profession (Table 2). The results are consistent with the trends we saw for speech features: *Politician* and *Movie Persons* have rather profession-specific and frequent keywords, which explains the good performance of the textual model. For instance, *Politicians*, often refer to *"politique (politics)"*, *"gouvernement (government)"* and *"ministre (minister)"* in their speeches. We also notice that the performances for the *Movie Person* and *Lawyer* are extremely low. The former is caused by the high missing rate of transcriptions (74%) and the latter is a consequence of the small number of data samples (12 samples). *Artists* as broader category are also hard to distinguish from *Movie Persons*, as they use overlapping topical keywords (e.g., "film", "cinema").

4.4. Combining speech signal and transcription

We now combine the features from the speech signal and the text. The results are shown in the last two columns of the Table 3. The combination consistently performs the best, and the macro F1 across all professions is at least 10 percentage points higher than the other two (shown in the last row). This proves the interest of combining low and high level linguistic features: the integration of features extracted from both the speech signal and the text can better characterize the speaker and consequently narrow its profession. The result also suggests the reasons why human strategies are more effective: the human perception relies on a complex integration of both low and high level information and take advantage of the communicative context in a broad sense to decode speaker turns properties including the characteristics of the speaker that utter it.

The accuracy of the model is not perfect. We hypothesize that the small number of training samples constrains the performance of machine learning. Furthermore, some parts of the audio did not have manual transcripts (column M). Nevertheless, our model can correctly predict the professions for more than half of the speakers. Table 4 shows the Hit@k measure, i.e., the percentage of people for whom the correct profession is among the top k predictions. We see that, while the approach cannot outright guess the profession, the speech nevertheless gives away enough information to narrow down the set of professions to 2 or 3.

5. CONCLUSION

We have shown that it is possible to narrow down, or even predict, the profession of a speaker from a recording of their speech. This is an interesting finding that can be further deepened, e.g., by using more features from speech, using more training data, or using extrinsic features. The main finding underlines the complementarity of low and high level language features and the need to combine them in order to improve performance.

The results point to interesting avenues of research, where intrinsic characteristics (e.g. age, gender), meta-cognitive characteristics (e.g. current state of mind or subconscious attitude towards a topic), or socio-economic characteristics (e.g. social status) can be predicted from multi-level acoustic and linguistic features. Such work poses obvious challenges for the privacy of individuals. It is thus important to investigate to which degree such privacy violations are possible, and whether they require regulation. However, such work also brings opportunities: It could help dialog systems adapt their discourse to the user. For example, the explanation of an insurance case would be different for a lawyer and for a scientist. The work could also help identify fake recordings or fake videos of public personalities, if the extracted characteristics do not correspond to the known characteristics of the speaker. Finally, such analyses carry educational and scientific value, as they reveal speech characteristics that give away personal details.

6. REFERENCES

[1] L. Lamel, J. Gauvain, and L. Canseco-Rodriguez, "Speaker diarization from speech transcripts," in *INTERSPEECH 2004 - ICSLP, 8th International Conference on Spoken Language Processing, Jeju Island, Korea, October 4-8, 2004*. ISCA, 2004. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2004/i04_0601.html

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

[3] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: Recent advances with deep learning," 2021.

[4] T. Pellissier-Tanon, G. Weikum, and F. M. Suchanek, "YAGO 4: A Reason-able Knowledge Base," in *ESWC*, 2020.

[5] Y. Wu, F. Suchanek, I. Vasilescu, L. Lamel, and M. Adda-Decker, "Using a knowledge base to automatically annotate speech corpora and to identify sociolinguistic variation," in *LREC*, 2022.

[6] T. Schultz, "Speaker characteristics," in *Speaker classification I*. Springer, 2007.

[7] M. Sambur, "Selection of acoustic features for speaker identification," *Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 2, 1975.

[8] J. Lindh and A. Eriksson, "Robustness of long time measures of fundamental frequency," in *Interspeech*, 2007.

[9] M. Ajili, J.-F. Bonastre, K. W. BEN, S. Rossato, and J. Kahn, "Comparaison des voix dans le cadre judiciaire: influence du contenu phonétique," *Journées d'Études sur la Parole*, 2018.

[10] O. Niebuhr and R. Skarnitzl, "Measuring a speaker's acoustic correlates of pitch–but which? a contrastive analysis based on perceived speaker charisma," in *ICPhS*, 2019.

[11] G. Chignoli, C. Gendrot, and E. Ferragne, "Caractérisation du locuteur par cnn à l'aide des contours d'intensité et d'intonation: comparaison avec le spectrogramme," in *Journées d'Études sur la Parole*, 2020.

[12] C. Mallavarapu, R. Mandava, S. Kc, and G. m Holt, "Political profiling using feature engineering and nlp," in *SMU Data Science Review*, 2018.

[13] A. A. Simoes and M. Castanos, "Fine-tuned bert for the detection of political ideology," 2020.

[14] T. Van Huynh, K. Van Nguyen, N. L.-T. Nguyen, and A. G.-T. Nguyen, "Job prediction: From deep neural network models to applications," in *RIVF*, 2020.

[15] D. Küçük and F. Can, "Stance detection: A survey," in *ACL*, 2020.

[16] L. Li, Z. Wu, M. Xu, H. M. Meng, and L. Cai, "Combining cnn and blstm to extract textual and acoustic features for recognizing stances in mandarin ideological debate competition," in *Interspeech*, 2016.

[17] Y. Dinkov, A. Ali, I. Koychev, and P. Nakov, "Predicting the leading political ideology of youtube channels using acoustic, textual, and metadata information," in *arXiv*, 2019.

[18] D. Patterson and D. R. Ladd, "Pitch range modelling: linguistic dimensions of variation," in *ICPhS*, 1999.

[19] I. Vasilescu, Y. Wu, A. Jatteau, M. Adda-Decker, and L. Lamel, "Alternances de voisement et processus de lénition et de fortition: une étude automatisée de grands corpus en cinq langues romanes," *Traitement Automatique des Langues*, vol. 61(1), 2020.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv*, 2018.

[21] L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, É. V. De La Clergerie, D. Seddah, and B. Sagot, "Camembert: a tasty french language model," in *ACL*, 2020.

[22] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," in *EMNLP demos*, 2020.

[23] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, "Yake! keyword extraction from single documents using multiple local features," *Information Sciences*, vol. 509, 2020.

[24] C. Bishop, *Neural networks for pattern recognition*. Oxford university press, 1995.

[25] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ESTER phase II evaluation campaign for the rich transcription of French broadcast news," in *Interspeech*, 2005.

[26] G. Gravier, G. Adda, N. Paulsson, M. Carré, A. Giraudel, and O. Galibert, "The ETAPE corpus for the evaluation of speech-based TV content processing in the french language," in *LREC*, 2012.

[27] J.-L. Gauvain, L. Lamel, and G. Adda, "The limsi broadcast news transcription system," *Speech communication*, vol. 37, no. 1, 2002.

[28] J.-L. Gauvain, G. Adda, M. Adda-Decker, A. Allauzen, V. Gendner, L. Lamel, and H. Schwenk, "Where are we in transcribing french broadcast news?" in *Interspeech*, 2005.

[29] M. Adda-Decker and L. Lamel, "The use of lexica in automatic speech recognition," in *Lexicon Development for Speech and Language Processing*. Springer, 2000.