Krajowa Konferencja Zastosowań Matematyki w Biologii i Medycynie

Regietów, 22nd–26th September 2015

I-VECTORS IN GENDER RECOGNITION FROM TELEPHONE SPEECH

Joanna Grzybowska, Mariusz Ziółko

AGH University of Science and Technology
al. Mickiewicza 30, 30-059 Kraków
{gjoanna, ziolko}@agh.edu.pl

ABSTRACT

This paper addresses gender recognition from telephone speech in emergency situations. A speech corpus of emergency phone calls is characterized by loud background noise, emotional, and often intoxicated speech. For those specific conditions we propose the use of i-vectors to model utterances. We gathered a large multilingual speech database, modified recordings to match those from emergency telephone calls, and for cross-validation we chose utterances which lasted at least 30 seconds. We also examined shorter recordings. In 5-fold cross-validation on the multilingual set of almost 1000 speakers we obtained F-score = 97.5%.

INTRODUCTION

Human voice contains cues to speaker characteristics, like identity, gender, emotional state and age of the speaker. Automatic gender recognition of a speaker is a part of many automatic speech recognition (ASR) and speaker recognition systems for the purpose of choosing gender-specific acoustic models. Systems that contain gender recognition step, achieve better speech or speaker recognition rates.

A supporting system of voice analysis for emergency call centers is being developed at AGH University of Science and Technology in Krakow. It is a human-computer interaction (HCI) system which uses telephone calls to create a caller profile [1]. The system is aimed at facilitating the work of public-safety answering point (PSAP) responders by providing information about the speaker. The information is automatically derived from caller voice. Graphical User Interface (GUI) of this system features several information about the current speaker (e.g. identity, language, gender, age, emotional state, intoxication level). The speech signal is obtained from telecommunication channel – the acoustic band is limited to 300-3400 Hz. The amount of information available in the signal is reduced due to lossy compression.

In the context of emergency phone application gender recognition has specific characteristics that make it different from other applications.

- Very often the incoming speech signal is distorted with loud background noise (e.g. crowd, cars, trains). Therefore a gender identifier has to be robust to acoustic condition changes.

- In those real-life emergency situations, caller speech is often very emotional (mostly intense, negative emotions). This makes CPR database a good material for emotion recognition system, although this characteristic of the CPR database hinders recognition of other human characteristics, e.g. gender.
- Callers are often intoxicated. The degree of their intoxication is reflected in their speech, which also impedes recognition of other caller characteristics (e.g. gender).

This research aims at recognizing speaker gender in those specific conditions.

GENDER RECOGNITION FROM SPEECH

Sex-based differences are visible in laryngeal and vocal tract physiology. Basically, the vocal tract of an adult female is shorter that a man's; a man's pharynx and larynx is larger than a woman's. Those and other physiological and anatomical differences cause differences in the acoustics of male and female voice. Mean fundamental frequency (F0) for men and women differs by about an octave. The average F0 for men is 130 Hz and for women 220 Hz [2]. The differences in the resonances of the vocal tract between men and women can be measured by the formant frequencies which are higher in females. Cues about a speaker's gender may be also found in dynamic features like pitch contours, speaking rate, and details of pronunciation.

A lot of previous studies focused on automatic gender identification from voice. Some researchers use the acoustic level information, some use prosodic information (i.e. pitch, energy, speaking rate), others fuse those both types [3]. We can also find papers which address gender recognition for telephone applications [4, 5]. The features for gender identification most typically found in the literature are pitch, formants, Mel-Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction Coefficients (PLP).

Using i-vectors for speaker recognition [6], as well as for language [7], accent [9] and age [8] recognition, has significantly increased the classification accuracy. I-vector is a low dimensional feature vector used as a model of an utterance. In this paper we propose the use of i-vectors for gender recognition from telephone speech in real-life application.

DATABASES

Seven different databases were used to train and evaluate the performance of gender recognition system. We used four corpuses that are commonly used in speech research: *aGender* [10], *TIMIT* [11], *LUNA* [12], *RSR2015* [13]; we also used three other corpuses: *CPR*, *uzywam*, and *WWW*. In total, we gathered speech samples from 4939 speakers. Number of speakers for each database as well as statistics for utterances duration and gender distribution is presented in Table 1.

Table 1. Databases used for training and evaluation with utterance duration statistics and gender distribution.

		Duration [sec]			Gender		
Corpus name	# of speakers	min	median	max	children	females	males
aGender	770	7	54	110	13.8%	43.5%	42.7%
TIMIT	629	13	21	31	0.0%	30.4%	69.6%
RSR2015	300	40	55	75	0.0%	47.7%	52.3%
LUNA	76	10	20	57	0.0%	65.8%	34.2%
CPR	3058	1	15	73	1.2%	50.2%	48.6%
uzywam	77	4	29	71	0.0%	39.0%	61.0%
WWW	29	10	15	20	0.0%	31.0%	69.0%
Total	**4939**	1	18	110	2.9%	46.4%	50.7%

DATA PREPARATION

Described gender recognition system is aimed to work in real-life emergency situations, where the available signal comes from the telecommunication channel. All of the recordings used in this study (Table 1) were therefore previously processed to match the standard of CPR recordings. All of the recordings were converted to mono signals, they were also downsampled to 8 kHz. Voice activity detection (VAD) is performed using energy variance analysis in different acoustic bands. The total duration of speech after those pre-processing steps (from all databases) is almost 34 hours. Pre-processing steps are illustrated on Figure 1.

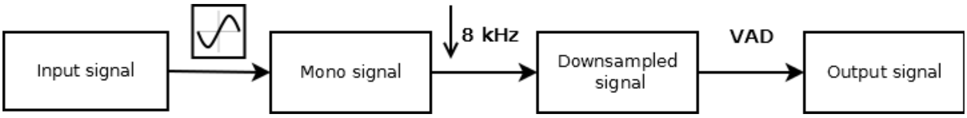

Figure 1. Pre-processing steps.

In the further analysis, we decided not to use children recordings. We performed tests in 2 different configurations: (1) 5-fold cross-validation using only recordings of more than 30 seconds of speech, (2) a single test where recordings of more than 30 seconds were used for creating gender models, and the rest of data was used for testing. Figures 2 and 3 show histograms for two created databases. The database of more than 30 seconds of speech for each speaker contains in total speech from 942 speakers (19.1% of all data). The database of less than 30 seconds of speech for each speaker contains in total speech from 3855 speakers (78.0% of all data).

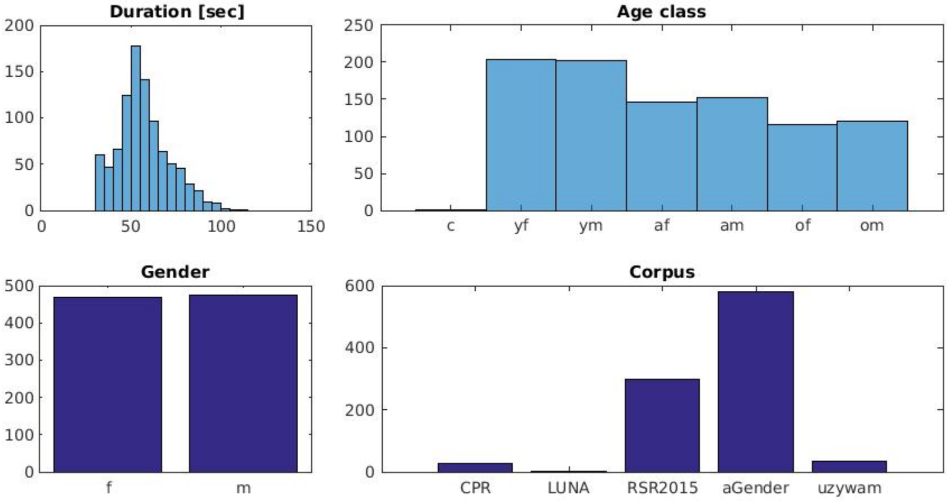

Figure 2. Histograms (duration of recordings for each speaker, age class assignment, gender distribution and source corpus) for a database of more than 30 seconds of speech for each speaker *(f – females, m – males, c – children, yf – young females, ym – young males, af – adult females, am – adult males, of – old females, om – old males).*

I-VECTORS FOR GENDER RECOGNITION

I-vectors were previously used for speaker [6], language [7], accent [9] and age [8] recognition from voice. We propose to use i-vectors for gender identification. In i-vector approach, models are estimated through eigenvoice adaptation. A total variability subspace is learned; it captures

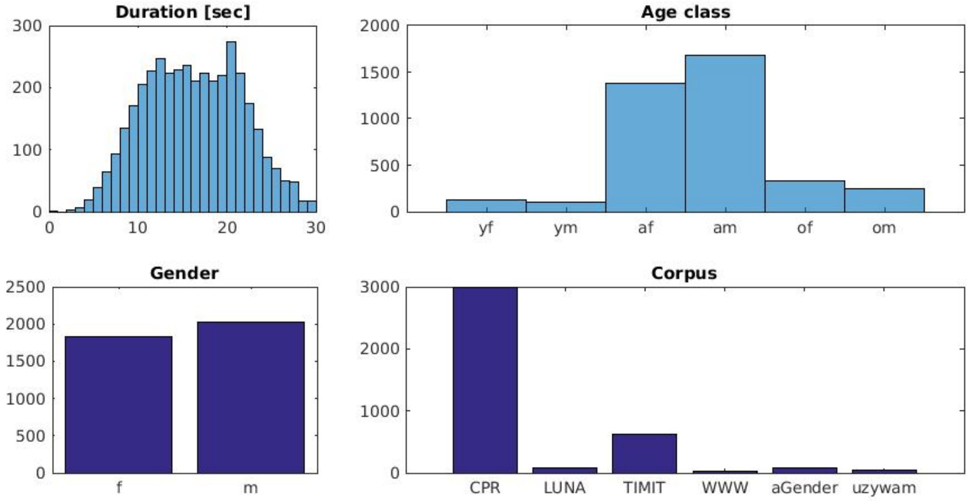

Figure 3. Histograms (duration of recordings for each speaker, age class assignment, gender distribution and source corpus) for a database of less than 30 seconds of speech for each speaker *(f – females, m – males, c – children, yf – young females, ym – young males, af – adult females, am – adult males, of – old females, om – old males).*

both speaker and channel effects [6]. Given a C component GMM UBM model λ with $\lambda_c = \{p_c, \mu_c, \Sigma_c\}, c = 1, ..., C$ and an utterance with a L frame feature sequence $\{y_1, y_2, ..., y_L\}$, the $0^t h$ and $1^s t$ order Baum-Welch statistics on the UBM are calculated as follows:

$$N_c = \sum_{t=1}^{L} P(c|y_t, \lambda) \tag{1}$$

$$F_c = \sum_{t=1}^{L} P(c|y_t, \lambda)(y_t - \mu_t) \tag{2}$$

where $c = 1, ..., C$ is the GMM component index and $P(c|y_t, \lambda)$ is the occupancy probability for y_t on λ_c.

Centered GMM mean supervector $\tilde{F}$ is projected on a low rank factor loading matrix T following the standard factor analysis framework:

$$\tilde{F} \rightarrow Tx, \tag{3}$$

where T is a rectangular total variability matrix of low rank and x is the i-vector [6, 14].

The acoustic features used to obtain i-vectors were MFCC with their first order derivatives. To measure the similarity between two i-vectors we used cosine measure, defined as follows:

$$k(x_1, x_2) = \frac{< x_1, x_2 >}{\|x_1\| \|x_2\|} \tag{4}$$

Model and test utterances are represented as i-vectors of size 300.

RESULTS AND DISCUSSION

As stated before, we performed 5-fold cross-validation on data of more than 30 seconds for each speaker. We examined the influence on the recognition results of number of GMM components used to train UBM. Figure 4 shows those results for F-score. F-score is calculated as follows:

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{5}$$

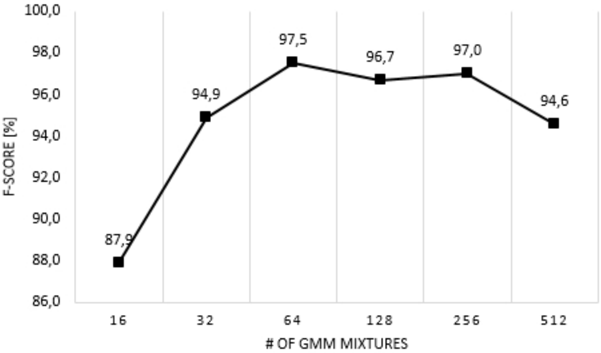

Figure 4. F-score results for different number of GMM components used to train UBM, obtained in 5-fold cross-validation.

We obtained the best results for 64 GMM components. The confusion matrix for 64 GMM components is shown in Table 2.

Table 2. Confusion matrix of true and predicted labels of gender classes obtained in 5-fold cross-validation (*F – females, M – males*).

		Predicted	
		F	M
True	F	**97%**	3%
	M	4%	**96%**

We also performed a single test where recordings of more than 30 seconds were used for creating gender models and the rest of recordings were used for testing. In this case scenario, depending on the number of GMM components used for UBM training, the F-score varied from 63.4% for 16 components to 86.1% for 64 components.

Although for shorter recordings used for testing the F-score results were much lower, the best result for the second test was also obtained for 64 GMM components.

CONCLUSIONS

In this paper we used i-vectors to address gender recognition in emergency situations. We merged several different databases to create a multilingual corpus of almost 5000 speakers. We divided collected data into two parts (recordings lasting more and less than 30 seconds) and we performed tests to evaluate our approach. Our results confirm the effectiveness of the proposed scheme.

ACKNOWLEDGEMENTS

The project was supported by the National Research and Development Center granted by decision 0072/R/ID1/2013/03.

REFERENCES

[1] J. Gałka, J. Grzybowska, M. Igras, P. Jaciów, K. Wajda, M. Witkowski, and M. Ziółko: *System supporting speaker identification in emergency call center*, Proceedings of the INTERSPEECH 2015 (to appear).

[2] J. M. Hillenbrand and J. M. Clark: *The role of F0 and formant frequencies in distinguishing the voices of men and women*, Attention, Perception & Psychophysics **71** (2009), 1150–1166.

[3] M. Li, K. J. Han, and S. Narayanan: *Automatic speaker age and gender recognition using acoustic and prosodic level information fusion*, Computer Speech and Language **27** (2013), 151–167.

[4] T. Bocklet, A. Maier, J. G. Bauer, F. Burkhardt, and E. Noth: *Age and gender recognition for telephone applications based on GMM supervectors and Support Vector Machines*, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2008, pp. 1605–1608.

[5] F. Metze, J. Ajmera, R. Englert, U. Bub, F. Burkhardt, J. Stegmann, C. Muller, R. Huber, B. Andrassy, J. G. Bauer, and B. Littel: *Comparison of four approaches to age and gender recognition for telephone applications*, Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2007, pp. 1520–6149.

[6] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet: *Front End Factor Analysis for Speaker Verification*, IEEE Transactions on Audio, Speech and Language Processing (2010).

[7] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak: *Language Recognition via Ivectors and Dimensionality Reduction* (INTERSPEECH 2011).

[8] M. H. Bahari, M. McLaren, H. van Hamme, and D. A. van Leeuwen: *Speaker age estimation using i-vectors*, Eng. Appl. of AI **34** (2014), 99–108.

[9] M. H. Bahari, R. Saeidi, H. van Hamme, and D. A. van Leeuwen: *Accent recognition using i-vector, Gaussian Mean Supervector and Gaussian posterior probability supervector for spontaneous telephone speech*, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 7344–7348.

[10] F. Burkhardt, M. Eckert, W. Johannsen, and J. Stegmann: *A Database of Age and Gender Annotated Telephone Speech*, Proceedings of the Language and Resources Conference (LREC), 2010.

[11] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren: *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1*, Web Download. Philadelphia: Linguistic Data Consortium, 1993.

[12] *LUNA – multi-domain multilingual dialogue corpus* (2007).

[13] A. Larcher, K. A. Lee, B. Ma, and H. Li: *The RSR2015: database for text-dependent speaker verification using multiple pass-phrases*, Annual Conference of the International Speech Communication Association (Interspeech), 2012, pp. 1580–1583.

[14] M. Li and S. Narayanan: *Simplified Supervised I-vector Modeling with Application to Robust and Efficient Language Identification and Speaker Verification*, Computer, Speech, and Language (2014).