INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Multi-Task Learning for Voice Related Recognition Tasks

Ana Montalvo[1], *Jose R. Calvo*[1], *Jean-F. Bonastre*[2]

[1]Advanced Technologies Application Center (CENATAV), Cuba
[2]LIA-CERI Avignon University, France

amontalvo@cenatav.co.cu, jcalvo@cenatav.co.cu, jean-francois.bonastre@univ-avignon.fr

Abstract

Speech is a complex signal conveying numerous information about the message but also various characteristics of the speaker: its sex, age, accent, language. Understanding the use of these features by machine learning (ML) systems has two main advantages. First, it could help prevent bias and discrimination in ML speech applications. Second, joint modeling of this information using multitasking learning approaches (MTL) has great potential for improvement. We explore in this paper the use of MTL in non-linguistic tasks. We compare single- and multi-task models applied to three tasks: (spanish) nativeness, speaker and sex. The effect of training data set size in the performance of both single- and multi-task models is investigated as well as the specific contribution of nativeness and sex information to speaker recognition. Experimental results show that multi-task (MTL) models outperform single task models. We have also found that MTL is beneficial for small training data sets and for low-level acoustic features rather than for pre-trained features such as bottleneck ones. Our results indicate also that more attention should be addressed to the information used by ML approaches in order to prevent biases or discrimination.

Index Terms: Multi-task learning, Convolutional neural networks, Close-set Speaker recognition.

1. Introduction

Speech is a complex signal conveying numerous information about the message but also various characteristics of the speaker like its sex, age, accent or language. The wide deployment of speech technologies in all faces of social life, including banking, health, dating, employment or forensics creates a growing expectation for explainability and transparency. The general public as well as legal systems, see European General Data Protection Regulation for example, are increasingly attentive to potential discrimination or breaches of privacy in AI applications [1, 2]. In the fight against discrimination as well as for the differential or partial preservation of privacy [3, 4, 5], to assess the extent to which a speech-based system captures or exploits information related for instance to sex, age, social level, or accent becomes more and more important.

Following the previous paragraph, a question that arises in modern machine learning systems is what specific information present in the data, during the training and operational phases, leads to a given decision. This question takes on its full value when current supervised artificial intelligence algorithms are almost always able to achieve their objective when applied to data close to training conditions, even if nobody knows what type of information they use for this. Thus, the misuse of a category of information in the incoming data could lead to a specific decision, opening the door to discrimination or loss of performance. Discrimination because the decision could be based on unwanted information, such as the linguistic content of the message in a text-independent speaker verification task. A loss of performance because the learning capacity of the machine could wrongly focus on the information sub-part which allows the best accuracy, even if this information is only a side effect, ignoring more useful data. A good example of this, still in speaker recognition, is to focus training on session-related information (the phone used for example).

In order to better understand this question, we propose in this paper to explore the latent relationship between different voice-related tasks: native-language (for Spanish), speaker sex[1] and speaker recognition. We wish to assess to what extent knowledge of the speaker's sex or whether or not he/she is a native Spanish speaker helps or leads to determining the identity of the speaker and vice versa.

We propose to use Multi-task learning (MTL) [7] versus classical mono-task training in order to evaluate these latent relations between our three tasks.

The basic idea of MTL is to learn related problems simultaneously, using a shared representation. When tasks have a common point and in particular when training data resources are limited, MTL can lead to better performance than a model formed on a single task, allowing the learner to capitalize on the common points between tasks. MTL has been originally proposed as a method that improves the generalization of a classifier by forcing it to learn more than one related task at a time. This has been previously demonstrated in several learning scenarios and areas of machine learning [8, 9, 10]. In this paper, to use MTL in order to improve the performance for our three tasks is only a secondary goal. Here, we use mainly MTL to measure the influence of information related to one of our tasks on the others: a gain in accuracy thanks to multi-task training will make it possible to highlight the links between the two or three categories of information, a loss will tend to show the opposite. In order to implement our machine learning models and our training strategies, we will use a classical convolutional neural networks (CNNs) based approach. The main advantage of CNNs comes from the use of weight sharing, local filters, and pooling since they help to discover robust and invariant representations.

The paper is organized in 6 sections: Section 2 describes the background and related work in MTL for speech-related tasks. Details of the dataset, features, and models used are presented in section 3. Our experiments are detailed in section 4. We analyzed the results and their implications in section 5. Section 6 concludes the work and explains possible future research directions.

[1]Throughout this paper, the term *sex* refers to the biological differences between female and male [6]

http://dx.doi.org/10.21437/Interspeech.2020-1857

2. MTL related work

MTL has been widely used in the speech field.[11] proposed a deep learning approach using multimodal features in order to simultaneously recognize speakers and emotions. The design of a secondary task for speaker recognition is presented in [12]. The task is called pseudo task as its target labels are obtained from an unsupervised Gaussian Mixture Model algorithm.

The use of adversarial MTL for learning invariant features is studied in recent papers. [13] explored the potential of adversarial MTL for learning invariant features. It proposes a noise-robust speaker embedding. In [14], the aim is to learn speaker-invariant multilingual bottleneck features for language recognition purposes.

Some papers show how language and sex are speaker related tasks that can be employed as auxiliary tasks [15, 16]. However, besides language and sex, some more complicated speech content features have been employed. A multi-task recurrent neural network model is presented in [17], for joint learning of automatic speech and speaker recognition. It showed improved performance on both automatic speech and speaker recognition tasks regarding single-task systems.

In [18], phoneme recognition is used as a secondary task to improve the performance of speaker recognition, using CNN as shared layers. The main advantage of CNNs comes from the use of local filters, weight sharing and max pooling. Thanks to these characteristics, CNNs provide some degree of invariance to small shifts of speech features along the frequency axis. Each one of these properties has the potential to improve speech recognition performance, and have proven to be important to deal with speaker and environment variations. [19] investigated a CNNs end-to-end trained for speaker verification purposes. It employs them as feature extractors to distinguish between the speaker and non-speaker information. In [20], bottleneck features from a CNN are used to build an ivector system for language identification. The CNN bottleneck features report complementary information to the conventional acoustic features.

CNNs have been widely applied to acoustic modeling for speech recognition, notably by [21], in which convolution was applied to learn more stable acoustic features for classes such as phone, speaker, and gender. [22] described how to apply CNNs to speech recognition, such that the CNNs structure directly accommodates some types of speech variability.

3. Methods

Our framework is based on two models: a multi-task (MTL) and a single-task model (STL). Both of them are based on 2D-CNN. The MTL model has three branches: native-non native Spanish speaker classification task (N_NN task), close-set speaker recognition task (S task), and sex classification task (G task), like illustrated in Figure 1. The MTL model is designed to learn jointly to classify the before mentioned related tasks. On the contrary, the STL model is a single-task CNN which should learn independently to classify each task. In order to help the comparisons between the two models, they share the same architecture. For the STL model, we just consider one given task while ignoring the others. All the models are built using Keras [23].

3.1. Corpora

In this work, we are using West Point Heroico Spanish Speech database [24]. It is composed of digital recordings of spoken Spanish. The corpus consists of two subcorpora, one collected at the Mexican Military Academy in Mexico City, and the other at the United States Military Academy (USMA). We select the USMA subcorpus as it includes data from non-native speakers. The USMA subcorpus includes $1\frac{1}{2}$ hours of speech from non-native speakers and 1 hour of speech from native speakers. All the speech in the USMA corpus was read speech, around 205 utterances per speaker, totaling 3675 utterances. It contains voice recordings from 18 speakers, 8 native Spanish speakers, and 10 non-native Spanish speakers, also labeled by sex divided among 6 female speakers and 12 male speakers.

3.2. Basic layout features

To feed the neural network model we considered two features: Mel-filterbank acoustic features and multilingual bottleneck features. To compute the Mel-filterbank features (M-fbank) the signal goes through a bandpass filter, it then gets sliced into non-overlapping frames and a Hamming window function is applied to each frame. Afterward, a short-time Fourier transformation on each frame is computed to obtain the power spectrum or periodogram, subsequently, the Mel filter banks are applied. In the end we get a 40-dimensional frame representation.

On the other hand, multilingual bottleneck features are obtained using the BUT/Phonexia bottleneck feature extractor [25]. The neural network used by the bottleneck extractor package was BabelMulti, trained on 17 languages from the IARPA Babel program. This network has four hidden layers, the third one is the bottleneck layer, its outputs are the 80-dimensional bottleneck (BN) features we will be using.

Each utterance is processed frame by frame to obtain the samples to feed the network. Each frame of speech is appended temporally with a context of 10 frames (± 5 frames of left and right context), conforming an input matrix of 11×40 for the M-fbank and 11×80 for BN features. From the USMA set we finally got 666760 samples and its corresponding triplet of labels. The set of utterances is divided initially in 80% to train and 20% for testing, ensuring that no frame of a train utterance will be seen in the test set. During training, the samples from the train set were split into 2 internal subsets: one used for actual training and the other for validation after each epoch.

3.3. Single-task model

STL is designed to ignore the connection within tasks. We built individuals 2D-CNN models for each single recognition task and each model is trained separately. In the experimental phase of single-task estimation, the input is a matrix formed by a frame vector and its adjacent frames (± 5). The output layer is the classification label corresponding to the specific task.

In this paper, we study three single-task recognition models, namely the speaker's Spanish nativeness, speaker identity, and sex, respectively. N_NN task and G task, are both binary classification tasks and their single-task CNN architecture will be the same. For the close set speaker classification task, after some experiments we decided to add a convolutional layer and a pair of fully connected layers in its particular branch, since it is a more complex classification task.

3.4. Multi-task model

MTL improves the generalization ability of the system by using a shared representation in the parallel processing of multiple related tasks. When these tasks are relevant, joint task learning should work better than learning each task individually, especially when the number of training examples for each task is

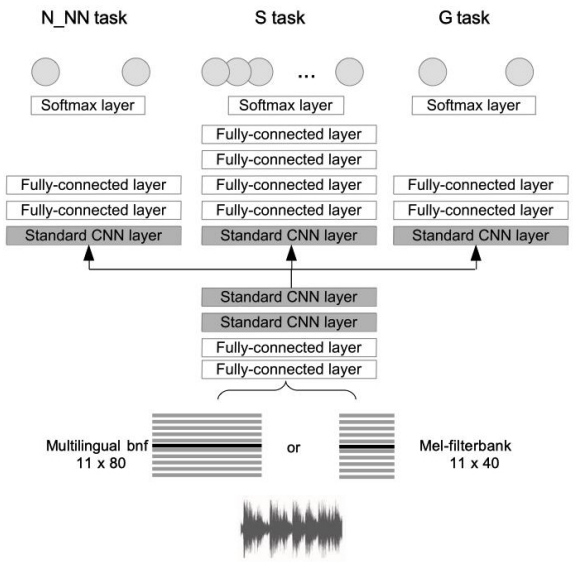

Figure 1: *Schematic diagram of MTL model.*

relatively low.

In MTL, each task contributes to the cost function. The loss value to minimize is the weighted sum of all individual losses:

$$\epsilon_{MTL} = \sum_{n=1}^{N} \lambda_n * \epsilon_{task_n}. \quad (1)$$

Where ϵ_x is the cost function to be minimized, λ_n is a non-negative weight and N the total number of tasks. An λ_n closer to 1 means that all the N tasks have the same impact, while a n near to 0 for a given task means that this task has no influence on the model training.

4. Experimental settings

To compare the performance of the single-task model and multitask model, we perform experiments on USMA corpora. This section provides the experimental settings we used.

4.1. Network setups

As explained in subsection 3.2, each frame of speech is appended temporally with a context of 5 frames. In the shared section, the model uses 2 fully connected layers followed by 2 convolutional layers with 256 kernels and 3×3 filters. The outputs of the filters are summed and processed with the max-pooling operation, which downsamples the 2D representation along the spectral dimension. The output of the max-pooling is processed with a ReLu activation function and dropout as a standard CNN pipeline. A similar process happens in the second and third convolutional layers.

Next, fully-connected layers and softmax are employed for each of the classification tasks. We use 4 hidden layers with 2048 units in the S branch and two hidden layers for the N_NN and G task, before the final output layer.

The networks are trained with the cross-entropy criterion, using the Adam optimizer. The minibatch size is fixed to 32. The learning rate starts from 1×10^{-3} and exponential decay is used to decrease it. The number of training epochs is 50.

4.2. Single and multi-task learning

The first experiment is devoted to verifying the effectiveness of MTL vs STL. Table 1 presents the classification accuracy of both models fed with multilingual BN features.

In this experiment, all the tasks are impacting the same in the learning process, according to equation 1: $\lambda_n = 1$ for $n = 1, 2, 3$.

Table 1: *Single-task vs Multi-task training with BN features*

Model	**Accuracy per task**		
	N_NN	**S**	**G**
Single-task	0.8745	0.6576	0.8231
Multi-task	0.9128	0.7823	0.8549

As shown in tables 1 and 2, with MTL, one task helps the other tasks, for both feature sets. However, the befit seems to be dependent on the feature set used, BN or M-fbank.

The experiments with M-fbank features (table 2) show that these low-level acoustic features are more informative than "high level" BN features.

4.3. Impact of weighting auxiliary tasks loss function

Assuming S task as the main one, and N_NN and G as auxiliary tasks, we intend to show the interdependence between these three tasks in an MTL framework. By varying the weight of the auxiliary tasks (λ_n) in the cost function to be minimized during the learning process, we can evaluate the individual contribution of each auxiliary task to the main one.

Table 3 shows that considering together the three tasks during the training phase increases the performance for the main task (S task), even if the gain remains quite small. It is not clear which auxiliary task brings more than the other, but the S task obtains profit using both: when one of the auxiliary task is dismissed, a decrease in the main task's accuracy is observed.

4.4. Dataset size

We done several experiments while reducing the size of the training dataset. Figure 2 reports the corresponding results. We observe that as the size of the training set decreases, the performance of the multitasking model increases compared to the single task models: using 50 and 25% of the training data results in an average drop of approximately 3.8 and 7.2% respectively for the single-task model for only about 2.8 and 4.1% for the multi-task model.

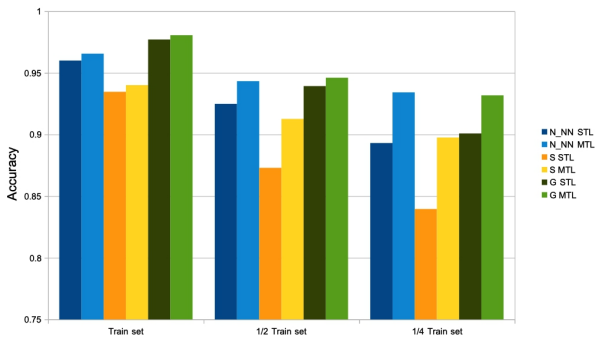

Figure 2: *Models performance as train set size decreases.*

Table 2: *Single-task vs Multi-task training with M-fbank features*

Model	Accuracy per task		
	N_NN	S	G
Single-task	0.9600	0.9147	0.9771
Multi-task	0.9656	0.9401	0.9806

Table 3: *Impact of the auxiliary tasks in the main task, using the multi-task model with M-fbank features*

Loss weight per task			Accuracy per task		
N_NN	S	G	N_NN	S	G
0	1	1	0.5221	0.9203	0.9787
0.2	1	0.7	0.9424	0.9351	0.9707
0.5	1	0.5	0.9619	0.9378	0.9740
0.7	1	0.2	0.9636	0.9391	0.9784
1	1	0	0.9638	0.9212	0.2911

5. Results and discussion

Models lower performance of BN features compared with M-fbank features are partially explained because the BN features are optimized for automatic speech recognition senone classification, so they lack from linguistically-irrelevant information, e.g., speaker change. MTL approach seems to undermine useful information on lower level traits better (table 1 and 2).

An average improvement on MTL's performance in comparison with STL is observed for most task. Although this is not surprising, there was not reported researches about using speaker nativeness as an auxiliary task for speaker recognition, and that is a remarkable contribution of this paper.

The best MTL performance is obtained with an equally weighted loss function per task $\lambda_n = 1$ (table 2). Assuming different weights for the auxiliary tasks (table 3) it is shown that considering both auxiliary tasks is beneficial for the main task, even when we are not able to say that one task is contributing more or less than the other.

As we can see in figure 2 it is corroborated that, in general, the performance of the models is affected by the reduction of the train data. However, it is for smaller sets that the impact of MTL becomes more remarkable compared with STL.

6. Conclusions

In this paper we investigated the links between different paralinguistic information in the view of three voice-related detection tasks, speaker nativeness, sex and identity. We used MTL as a measure system by comparing independent or joint modelling of the three tasks. We observed that the multi-task models produced clearly better results for one the task, speaker recognition, and slightly better performance for the two others, when applied on low-level features like M-fbank. On higher level features, like BN features coming from a discriminatively trained neural network, the three tasks took clearly advantage of the shared representation proposed by MTL. This finding could help explain why state-of-the-art speaker recognition systems often combine BN and low-level characteristics: BN characteristics offer several advantages, such as better modeling in long acoustic units (senone), but are too task-oriented and lose other source information still present at a lower level. We have also found that MTL becomes more beneficial when the training set is smaller.

In addition, the selection of the secondary task appears to be crucial and affects the performance of the main task. Through our different configurations, it appears that nativeness and speaker-sex detection are auxiliary tasks which benefit the task of recognition of the speaker. This is a promising result which strongly encourages us to go further with MTL.

Although preliminary, the results presented in this document also showed that more attention should be paid to the information used by ML approaches to prevent bias or discrimination. This is particularly important regarding the sensitive nature of paralinguistic information regarding privacy issues.

7. Acknowledgements

The authors would like to thank LIA and CENATAV IT colleagues whose help was essential to carry out the experiments with an ocean in the middle.

8. References

[1] L. Costa and Y. Poullet, "Privacy and the regulation of 2012," *Computer Law & Security Review*, vol. 28, no. 3, pp. 254 – 262, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0267364912000672

[2] A. Nautsch, A. Jiménez, A. Treiber, J. Kolberg, C. Jasserand, E. Kindt, H. Delgado, M. Todisco, M. A. Hmani, A. Mtibaa, M. A. Abdelraheem, A. Abad, F. Teixeira, D. Matrouf, M. Gomez Barrero, D. Petrovska Delacrétaz, G. Chollet, N. Evans, J.-F. Bonastre, B. Raj, I. Trancoso, and C. Busch, "Preserving privacy in speaker and speech characterisation," *Computer Speech and Language, June 2019*, 06 2019. [Online]. Available: http://www.eurecom.fr/publication/5910

[3] C. Dwork, "Differential privacy: A survey of results," in *Theory and Applications of Models of Computation*, M. Agrawal, D. Du, Z. Duan, and A. Li, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 1–19.

[4] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3?4, p. 211?407, Aug. 2014. [Online]. Available: https://doi.org/10.1561/0400000042

[5] C. Glackin, G. Chollet, N. Dugan, N. Cannings, J. Wall, S. Tahir, I. G. Ray, and M. Rajarajan, "Privacy preserving encrypted phonetic search of speech data," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 6414–6418.

[6] V. Prince, "Sex vs. gender," *International Journal of Transgenderism*, vol. 8, no. 4, pp. 29–32, 2005. [Online]. Available: https://doi.org/10.1300/J485v08n04_05

[7] R. Caruana, "Multitask learning: A knowledge-based source of inductive bias," in *Proceedings of the Tenth International Conference on Machine Learning*. Morgan Kaufmann, 1993, pp. 41–48.

[8] R. K. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data," *Journal of Machine Learning Research*, vol. 6, no. Nov, pp. 1817–1853, 2005.

[9] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4460–4464.

[10] A. Maurer, M. Pontil, and B. Romera-Paredes, "The benefit of multitask representation learning," *J. Mach. Learn. Res.*, vol. 17, no. 1, p. 2853?2884, Jan. 2016.

[11] S. Novitasari, Q. T. Do, S. Sakti, D. P. Lestari, and S. Nakamura, "Multi-modal multi-task deep learning for speaker and emotion

recognition of tv-series data," *2018 Oriental COCOSDA - International Conference on Speech Database and Assessments*, pp. 37–42, 2018.

[12] X. Lu, P. Shen, Y. Tsao, and H. Kawai, "A pseudo-task design in multi-task learning deep neural network for speaker recognition." in *ISCSLP*. IEEE, 2016, pp. 1–5. [Online]. Available: http://dblp.uni-trier.de/db/conf/iscslp/iscslp2016.html

[13] J. Zhou, T. Jiang, L. Li, Q. Hong, Z. Wang, and B. Xia, "Training multi-task adversarial network for extracting noise-robust speaker embedding," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2019.8683828

[14] Z. Peng, S. Feng, and T. Lee, "Adversarial multi-task deep features and unsupervised back-end adaptation for language recognition," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5961–5965.

[15] H. Chen, L. Xu, and Z. Yang, "Multi-dimensional speaker information recognition with multi-task neural network," in *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, 2018, pp. 2064–2068.

[16] F. Richardson, D. Reynolds, and N. Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1671–1675, 2015.

[17] Z. Tang, L. Li, D. Wang, and R. Vipperla, "Collaborative joint training with multitask recurrent model for speech and speaker recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 3, pp. 493–504, 2017.

[18] Y. Pan and W.-Q. Zhang, "Multi-task learning based end-to-end speaker recognition," in *Proceedings of the 2019 2nd International Conference on Signal Processing and Machine Learning*, ser. SPML 19. New York, NY, USA: Association for Computing Machinery, 2019, p. 5661. [Online]. Available: https://doi.org/10.1145/3372806.3372818

[19] H. Salehghaffari, "Speaker verification using convolutional neural networks," *ArXiv*, vol. abs/1803.05427, 2018.

[20] S. Ganapathy, K. Han, S. Thomas, M. Omar, M. Van Segbroeck, and S. Narayanan, "Robust language identification using convolutional neural network features," in *Proceedings of the 2014 Annual Conference of the International Speech Communication Association, INTERSPEECH*, 09 2014.

[21] D. Hau and K. Chen, "Exploring hierarchical speech representations with a deep convolutional neural network," in *Proceedings of UKCI'11*, 9 2011.

[22] O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[23] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[24] "West Point Heroico Spanish Speech, Philadelphia: Linguistic Data Consortium," WWW page, 2006. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2006S37

[25] A. Silnova, P. Matějka, O. Glembek, O. Plchot, O. Novotný, F. Grézl, P. Schwarz, and J. Černocký, "But/phonexia bottleneck feature extractor," in *Proceedings of Odyssey 2018*, vol. 2018, no. 6. International Speech Communication Association, 2018, pp. 283–287. [Online]. Available: https://www.fit.vut.cz/research/publication/11789