INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Cross-Lingual Transfer Learning for Affective Spoken Dialogue Systems

Kristijan Gjoreski, Aleksandar Gjoreski, Ivan Kraljevski, Diane Hirschfeld

voice INTER connect GmbH, Dresden, Germany
{kristijan.gjoreski,aleksandar.gjoreski,ivan.kraljevski,diane.hirschfeld}
@voiceinterconnect.de

Abstract

This paper presents a case study of cross-lingual transfer learning applied for affective computing in the domain of spoken dialogue systems. Prosodic features of correction dialog acts are modeled on a group of languages and compared with languages excluded from the analysis.

Speech from different languages was recorded in carefully staged Wizard-of-Oz experiments, however, without the possibility to ensure balanced distribution of speakers per language. In order to assess the possibility of cross-lingual transfer learning and to ensure reliable classification of corrections independently of language, we employed different machine learning approaches along with relevant acoustic-prosodic features sets.

The results of the experiments with mono-lingual corpora (trained and tested on a single language) and cross-lingual (trained on several languages and tested on the rest) were analyzed and compared in the terms of accuracy and F1 score.

Index Terms: transfer learning, cross-lingual, spoken dialog systems, affective computing

1. Introduction

Affective computing (AC) employs a computational approach to study affect by acquisition and processing bio-signals from different sensor types. The goal is to build an affect model and assign machines the human-like capabilities of observation, interpretation and generation of affect features [1].

Humans express the affects by various facial expressions, body movements, gestures, voice behavior, and other physiological signals, digitally processed [2] and commonly modeled by machine learning [3]. One aspect of the person's affect state is the voice behavior where speech differs regarding the acoustic features as: pitch parameters (F0 level, range, contour and jitter), speaking rate and voice quality. Those acoustic features are used in emotion research along with the pattern recognition methods.

In order to get a representative affect model, it is of great importance to have sufficient amount of annotated and relevant observations. The data could be provided by human experts which could annotate the empirically observable behaviors and actions, relying on inference [4], and the situational context [5]. Also, spontaneous affect response can be elicited by experimental methods, however with no guarantee that the same stimulus will provoke similar response across all participants.

In many real-world applications the data used for training and the data which will be used to infer the model will often differ. The reason, among others, can be insufficient number of participants in respect of specific experimental requirements, like age, gender, language, social background etc. If performed correctly, the knowledge transfer can be employed to overcome necessary acquisition of subjects in experiments or expensive data-labeling efforts. Transfer learning [6] in speech and linguistics generalizes models trained for one setting or task to other settings or tasks [7].

While there are many studies successfully dealing with cross-lingual voice behavior recognition [7, 8], few investigate the possibility of transfer learning applied on cross-lingual classification of correction dialogue acts.

In our previous work [9], cross-lingual differences related to hyperarticulated speech in correction dialogue acts were investigated. It was confirmed that there are distinctive prosodic features across 9 different languages associated with hyperarticulated speech. Although the results showed that differentiation by prosodic features is possible, the influence of the number of observations per language remained for a further research. Since the correction acts distribution is not balanced across the languages, it would be worth investigating if there is distinctive cross-lingual and cultural independent behavior that can be generalized over a group of languages (with most abundant data) equally modeling the under-represented languages in a corpus.

In this study, cross-lingual transfer learning applied in speech feature processing and classification is presented. Prosodic features of correction dialog acts are modeled on a group of languages and deployed on those not included in the training process. The speech recordings in 13 different languages were collected in carefully staged Wizard-of-Oz experiments. However, it was not possible to ensure balanced distribution of speakers per language. Therefore, to assess the possibility of cross-lingual transfer learning and to ensure reliable classification of corrections independently of language, we employed different machine learning methods and relevant features sets. This paper is structured as follows: Section 2 presents the speech database and its organization, as well as the acoustic feature sets employed and statistical analysis of the dataset. Section 3 describes the setup and the methodology of conducting the experiments, while Section 4 and 5 compare the performances of the feature sets on mono-lingual and cross-lingual corpora and concludes the paper.

2. Material and methodology

2.1. Speech database

The multilingual speech database was collected by staging a series of Wizard-of-Oz (WOz) experiments. In a preparatory phase, an online questionnaire with a total of 870 participants was carried out in 13 languages [10]. The smart-home control scenarios were thoughtfully designed to elicit spontaneous reactions and to trigger corrections responses from the participants in case of miscommunication. Different correction responses were elicited by presenting specific speech prompts, noted as: *Substitutions* – reaction on wrongly recognized parameters; *Insertions* – reaction to confirmation of a non-uttered sentence; *Deletions* – reaction on request to repeat the last utterance.

http://dx.doi.org/10.21437/Interspeech.2019-2163

2.2. Data organization

Common datasets were compiled for all 13 languages, based on the collected corpus and the time-stamped logs of the dialogue acts. The collected speech for the correction turns were only transcribed by orthography and not evaluated by any other speech characteristic, meaning there are no annotations describing presence of particular speech features. The datasets were transformed by subtracting the features of the *non-correction acts* from the adjacent *corrections acts*, providing quantitative measure how acoustic-prosody features are changed over the both acts. Such *delta features* are considered better suited for analysis and classification, compensating different speakers and environment conditions.

2.3. Acoustic features

We employed 3 different feature sets for the paired acts.

2.3.1. IS09 emotion features

At first, we used the standard feature set designed for emotion recognition: the Interspeech 2009 (IS09) emotion challenge feature set [11]. It contains 384 features extracted from open source feature extraction toolkit openSMILE [12].

2.3.2. GeMAPS and eGeMAPS features

In contrast to fore-mentioned large scale feature set, we also used minimalistic acoustic standard parameter set – the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) and the extended version (eGeMAPS) [13]. The same feature extractor was used as for the IS09 emotion features, providing 3 comparable datasets on the same corpus. The minimalistic GeMAPS has 18 low-level-descriptors (LLD) organized in groups:

- Frequency: pitch, jitter, frequency of (F1-F3), F1 bandwidth;
- Energy/Amplitude parameters: shimmer, loudness, harmonics-to-noise ratio;
- Spectral parameters: alpha ratio, Hammarberg Index, spectral slope, F1-F3 relative energy, harmonic difference H1-H2, harmonic difference H1-A3.

Additional features are derived from the LLDs totalling the number of the features to 62. The extended GeMAPS (eGeMAPS) has additional LLDs: Mel-frequency cepstral coefficients 1-4, spectral flux, F2-F3 bandwidth, arithmetic mean and coefficient of variation applied to LLDs, as well additional features are derived from the LLDs, in total 88 parameters.

2.3.3. Dataset statistical analysis

Shapiro Wilk normality test of the delta features of all 3 datasets, independent of language, showed that they are not represented with a normal distribution ($p<0.05$), as is usually expected for real data. Non-parametric one-sample Wilcoxon test applied to all the delta features, independent of language, showed that there are features with non-zero median values ($p<0.05$): IS09 202, GeMAPS 38 and eGeMAPS 57 features. Better insight about the effects of the factors could be given by employing Linear Mixed Models [14] in R package for statistical computing [15]. The respective features were taken as dependent variables, the speaker as a random factor, the native language and the elicited corrections type as fixed factors, and all their possible interactions were included in the model. The p-values for the fixed effects were calculated by the deviance table analysis (Type II Wald chi-square tests). For the IS09 features the findings could be summarized as: for the language as a factor 98, for correction types, 88 and for the language and correction type interactions, 63 features were not significant. For language, 48 (GeMAPS) and 66 (eGeMAPS) features were significant predictors. For correction type, 43 (GeMAPS) and 57 (eGeMAPS) features were significant. Subsequently, for the interaction correction type and the language, 49 (GeMAPS) and 68 (eGeMAPS) were statistically significant (at least $p<0.05$). The Linear Mixed Model analysis showed that there is a significant influence of the fixed factors language and correction types, as well as their interaction on the delta features.

3. Experiments

3.1. Experiments setup

Our study consists of conducting an experiment where features from user correction utterances in selected languages are used to create a cross-lingual model that could be used when evaluating unknown languages. For this purpose, the three feature sets, GeMAPS, eGeMAPS and IS09, were used to train independent models. Since the language datasets contained relatively small number of observations and their respective classes were unbalanced, the insertion class was omitted as the most inconsistent one, making the problem a binary classification.

To make sure that a fair comparison is made, a separate model was created for every available language, whose performances were then compared with our newly created cross-lingual model from languages proven as most consistent.

All of the experiments were performed using the Python programming language, with scikit-learn [16] as the machine learning library.

3.2. Classification methodology

Three main machine learning approaches – random forest (RF) [17], support-vector classifier (SVC) [18] and multilayer perceptron (MLP) – were used to create an independent *base model* for every feature set using utterances in all available languages, with a purpose of setting a performance baseline to later determine which languages would contribute most to the performance of the new cross-lingual model.

The optimal values for the hyperparameters were achieved by a grid search. For the RF classifier, we explored different values for the number of trees (ranged from 50 to 400) and number of features to consider when looking for the best split (square root or a binary logarithm of the total number of features). It was noted that increasing the number of trees further did not result in performance increase. To train our MLP classifier, we used rectified linear unit (ReLU) and logistic sigmoid as an activation function for the hidden layers, stochastic gradient descent (SGD), Adam and limited-memory BFGS (LBFGS) were used as solvers for weight optimization with a penalty parameter alpha ranging from 10^{-1} to 10^{-4}. As for the SVC, a kernel was selected between radial basis function (RBF) and linear with a coefficient gamma (ranged from 10^{0} to 10^{-3}), as well as a penalty parameter C (ranged from 10^{1} to 10^{-3}).

The hyperparameter performances were measured with means of accuracy, precision, recall and F1 score over a 5-fold cross-validation, while maintaining the class distribution in every fold. The approach with a most accurate performance for the baseline model was used throughout the remaining parts of the experiment which included individual language models and the creation of a new, cross-lingual model with languages that surpass the baseline. In other words, languages with individual model where at least two out of the three feature sets result to F1 scores below the lowest F1 score were omitted.

4. Results and discussion

Similar performances for the baseline models were achieved by all three methods (Table 1) with standard deviations between 0.01 and 0.03. However, the random forest classifier was proven to give most accurate results across all three feature sets and therefore chosen as the main classifier for the rest of the experiment with the best hyperparameter values of the feature sets.

A highest accuracy of 0.71 was achieved using the IS09 feature set and random forest classifier, which performed best over the eGeMAPS (0.69) and GeMAPS (0.68) as well. All three classifying methods were representative as a highest F1 score classifier. Using SVC, eGeMAPS achieved a F1 score of 0.66, GeMAPS with MLP yielded a 0.69 F1 score and IS09 performed a 0.67 F1 score using RF. From the F1 scores obtained with the random forest classifier a baseline was set for later feature set performance comparison.

Table 1: *Comparison between SVC, MLP and RF performance across the three feature sets.*

	SVC		MLP		RF	
	ACC	F1	ACC	F1	ACC	F1
eGeMAPS	0.68	**0.66**	0.67	0.64	**0.69**	0.64
GeMAPS	0.66	0.63	0.64	0.62	**0.68**	**0.64**
IS09	0.70	0.64	0.70	**0.69**	**0.71**	0.67

To investigate the possibility of transfer learning, an omission of some languages from the base model was needed for evaluating purposes i.e. if a model trained on selected languages can achieve similar (or better) performance on an unknown language.

Figure 1 presents the accuracy and F1 scores of the three feature sets across all 13 individual language models. While the GeMAPS and eGeMAPS feature sets performed fairly similar, the features used from the IS09 set had a slight edge over the accuracy and the F1 scores.

It is evident that the standard deviation differs drastically across the languages, but not across the feature sets. Due to the noticeable deficit in the number of utterances in comparison to the remaining languages, the Danish (DK), Swedish (SE) and Norwegian (NO) have the highest standard deviation and relatively modest performances, however, the model trained on Finnish utterances, although with a similar size and standard deviation, had a greater performance.

The difficulty for a separate model to perform well due to its insufficient number of utterances is one of the main reasons why creation of such model should be avoided and require a language-independent model.

However, by omitting certain languages we most likely lower the performance of the model. Therefore, a careful approach was implemented where separate model for every language was trained using a similar method as the base model and a language was removed only if for that language model the F1 scores of at least two out of the three feature sets were below the lowest F1 score (0.64 with GeMAPS and eGeMAPS, presented with a dashed line on Figure 1). Following this rule, we are assured that same languages are being tested across every feature set. Such languages were Chinese (CN), Danish (DK), Dutch (NL), Norwegian (NO), Swedish (SE) and Turkish (TR). Languages that surpassed the baseline and whose utterances were included in new cross-lingual model were German (DE), English (EN), Spanish (ES), French (FR), Italian (IT) and Russian (RU). The utterances from the Finnish (FI) language have been excluded as well because of their model's high standard deviation, even though similar results were achieved with their inclusion. Note that languages from a same language family do not have to be present in both the training and the test set and therefore, a classification of an unknown language family is made with features extracted from different ones.

On Figures 2 and 3 is shown the difference in performance between the new model, trained on different languages with individual performance over the baseline (F1>0.64), and a separate model, trained on utterances from a single language.

Presented on Figure 2, the accuracy of the individually trained models has been generally predominant with an emphasis of the Danish, Swedish and Turkish language. Here, both the GeMAPS and eGeMAPS have worked best when tested on some languages (up to 13% and 8% respectively on Norwegian), but poorly on other (down to 11% on Turkish and 13% on Swedish respectively).

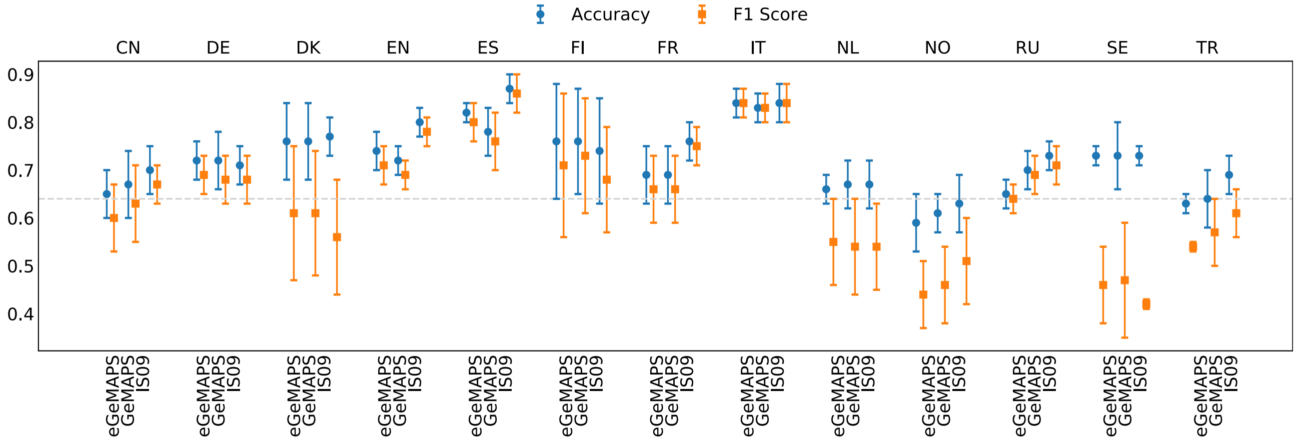

Figure 1: *Accuracy and F1 score values with their respective standard deviation for all languages as individually trained models using the eGeMAPS, GeMAPS and IS09 feature sets. The F1 baseline is presented with a dashed line.*

However, since the difference between accuracy and F1 score on some separate models such as Danish and Swedish is apparent (Fig. 1), a considerably better performance representation of the new model would be through the F1 score comparison.

On Figure 3 the F1 scores of the model trained on languages over the baseline have been evidently dominant over almost every test language.

A highest improvement of 27% has been registered for the Norwegian language using the GeMAPS feature set and, similar as eGeMAPS, in only two (CN, TR) out of the seven sub-baseline language models has occurred some performance decrease, in contrast to the three languages (CN, DK, FI) when using the IS09 feature set.

An remarkable difference range from -6% (TR) to 27% (NO) has been achieved using the GeMAPS feature set, -5% (TR) to 22% (NO) when the eGeMAPS feature set was used and IS09 feature set yielded to a difference range from -5% (DK, FI) to 16% (NL).

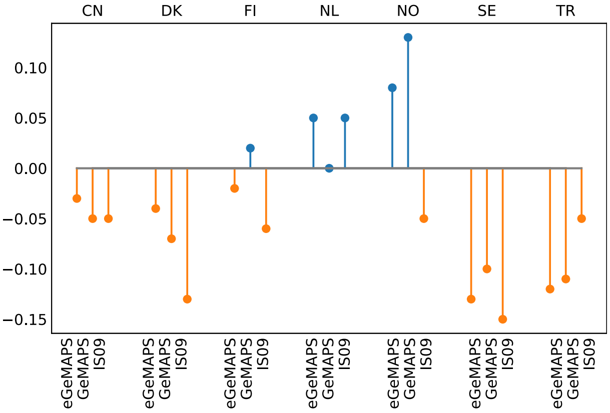

Figure 2: *Accuracy difference between the model trained without the sub-baseline language models and the individual language models.*

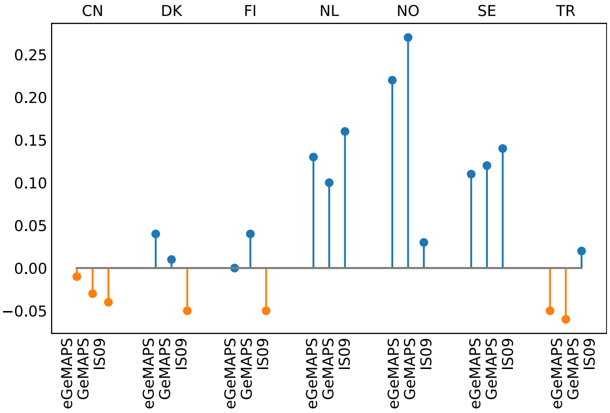

Figure 3: *F1 score difference between the model trained without the sub-baseline language models and the individual language models.*

It is worth noting that the model created from the most consistent correction utterances across several languages, tested on utterances of different languages, achieved insignificantly inferior results in few, but far superior in most comparisons to individual test language models and therefore, confirmed the possibility of cross-lingual transfer learning for correction act classification.

5. Conclusions

In this paper we implemented a case of cross-lingual transfer learning applied for affective computing. Prosodic features were extracted from correction utterances in selected languages which were then used to create a cross-lingual model for evaluating unknown languages.

Using a carefully staged Wizard-of-Oz experiments, speech recordings were collected in 13 different languages, with an emphasis on a natural user correction reaction. To evaluate the possibility of cross-lingual transfer learning and to ensure reliable classification of corrections independently of language, different machine learning methods along with relevant acoustic-prosodic features sets were tested.

With a purpose of setting a performance baseline, random forest, support vector and multilayer perceptron classifiers were used to create an independent base model for the GeMAPS, eGeMAPS and IS09 feature sets using utterances in all available languages.

To maintain a fair comparison, a separate model was created for every language, whose performances were then compared with our newly created cross-lingual model from languages proven to have the most consistent correction utterances.

The models created with the proposed approach, trained across the most consistent languages, when tested on utterances of different languages, achieved insignificantly inferior results in few, but far superior in most comparisons to individual test language models, tackling the difficulty for a separate language model to perform well when an insufficient number of utterances are present and confirming the possibility of cross-lingual transfer learning for correction act classification.

For our future work, our objective is to re-evaluate the presented approach after improving the target class annotations which is expected to further increase the classification performance.

6. References

[1] J. Tao and T. Tan, “Affective computing: A review,” in *International Conference on Affective computing and intelligent interaction*. Springer, 2005, pp. 981–995.

[2] A. Vinciarelli, M. Pantic, and H. Bourlard, “Social signal processing: Survey of an emerging domain,” *Image and vision computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[3] P. M. Domingos, “A few useful things to know about machine learning.” *Commun. acm*, vol. 55, no. 10, pp. 78–87, 2012.

[4] M. Mehu and K. R. Scherer, “A psycho-ethological approach to social signal processing,” *Cognitive processing*, vol. 13, no. 2, pp. 397–414, 2012.

[5] A. Kappas, U. Hess, and K. R. Scherer, “Voice and emotion,” *Fundamentals of nonverbal behavior*, vol. 200, 1991.

[6] K. Weiss, T. M. Khoshgoftaar, and D. Wang, “A survey of transfer learning,” *Journal of Big Data*, vol. 3, no. 1, p. 9, 2016.

[7] D. Wang and T. F. Zheng, “Transfer learning for speech and language processing,” in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2015, pp. 1225–1237.

[8] S. M. Feraru, D. Schuller *et al.*, “Cross-language acoustic emotion recognition: An overview and some tendencies,” in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 125–131.

[9] I. Kraljevski and D. Hirschfeld, “Hyperarticulation of corrections in multilingual dialogue systems.” in *INTERSPEECH*, 2017, pp. 2531–2535.

[10] I. Wendler, A. Jatho, I. Kraljevski, and M. Wenzel, “Nutzerzentrierter Entwurf von Multimodalen Bedienkonzepten,” in *28. Konferenz Elektronische Sprachsignalverarbeitung 2017*, 2017.

[11] B. Schuller, S. Steidl, and A. Batliner, “The interspeech 2009 emotion challenge,” in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[12] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor,” in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[13] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, “The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing,” *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[14] D. Bates, M. Mächler, B. Bolker, and S. Walker, “Fitting linear mixed-effects models using lme4,” *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[15] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014. [Online]. Available: http://www.R-project.org/

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[17] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[18] C.-C. Chang and C.-J. Lin, “Libsvm: a library for support vector machines,” *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, p. 27, 2011.