Investigating the role of L1 in automatic pronunciation evaluation of L2 speech

Ming Tu[1], Anna Grabek[2], Julie Liss[1], Visar Berisha[1,2]

[1]Speech and Hearing Science Department
[2]School of Electrical, Computer, and Energy Engineering
Arizona State University
{mingtu, agrabek, jmliss, visar}@asu.edu

Abstract

Automatic pronunciation evaluation plays an important role in pronunciation training and second language education. This field draws heavily on concepts from automatic speech recognition (ASR) to quantify how close the pronunciation of non-native speech is to native-like pronunciation. However, it is known that the formation of accent is related to pronunciation patterns of both the target language (L2) and the speaker's first language (L1). In this paper, we propose to use two native speech acoustic models, one trained on L2 speech and the other trained on L1 speech. We develop two sets of measurements that can be extracted from two acoustic models given accented speech. A new utterance-level feature extraction scheme is used to convert these measurements into a fixed-dimension vector which is used as an input to a statistical model to predict the accentedness of a speaker. On a data set consisting of speakers from 4 different L1 backgrounds, we show that the proposed system yields improved correlation with human evaluators compared to systems only using the L2 acoustic model.

Index Terms: Accentedness, pronunciation evaluation, automatic speech recognition,

1. Introduction

With the development of speech technologies, pronunciation training in second language education can be replaced by computer-based systems. Automatic pronunciation evaluation has always been an important part of Computer Assisted Pronunciation Training (CAPT). The goal of automatic pronunciation evaluation is to build an automatic system which can measure the quality of pronunciation given input speech. Automatic speech recognition (ASR) models play an important role in this area. The acoustic model in an ASR system trained on native speech provides a baseline distribution for each phoneme/word; new speech samples can be projected on this distribution to determine how statistically close the pronunciation is to a native pronunciation.

From a speech learning perspective, accented speech is the result of second language (L2) speech being produced by a sensorimotor control system that has overlearned first language (L1) sound contrasts and rhythmic composition. The Speech Learning Model (SLM) plays an important role in explaining L2 speech learning, which is based on the idea that phonetic systems respond to L2 sounds by adding new phonetic categories, or by modifying existing L1 phonetic categories [1]. The SLM emphasizes the interplay between L1 and L2 in forming the target language phonetic systems of language learners. Based on the SLP hypotheses, an equivalence classification is applied to an L2 phone similar to a previously experienced L1 category, thereby degrading the accuracy of L2 sound production. Since certain phonetic and phonological patterns can be transferred from L1 to the learned L2, English spoken by people from different L1 backgrounds show acoustic characteristics similar to the speakers' mother language [2][3].

However, almost all existing pronunciation evaluation systems only use acoustic models trained on native L2 (i.e. the target language) speech to extract useful measurements to quantify how close non-native speech is to the native pronunciation of the target language. For example, the study in [4] proposed measurements based on both phoneme-level log-likelihood and posterior probabilities calculated from an ASR system trained on native French speech to evaluate the pronunciation of French learners. They showed that the posterior based measurements provided the highest correlation to human scores. Goodness of pronunciation (GOP), which is the log-posterior probability of aligned phonemes normalized by phoneme duration, was used in [5] to detect mispronunciation in non-native speech. The log posteriors in the GOP are also derived from an ASR acoustic model trained on native speech. New advancement in Deep Neural Networks (DNN)-based acoustic models boost the performance of ASR systems, and at the same time these acoustic models have been applied to pronunciation assessment or mispronunciation detection [6][7]. Although some studies also train another ASR system with accented speech in order to generate better recognition/alignment results (thus better fluency and rhythm based features), those measurements related to pronunciation are still extracted from acoustic models trained on native speech [7][8].

Inspired by the SLM, in this paper we propose to use pronunciation measurements derived from both L1 and L2 acoustic models. We anticipate those features extracted from the L1 acoustic model can provide extra information about the speaker's L2 pronunciation quality. Specifically, two sets of phoneme-level acoustic model confidence scores are implemented: the first is based on the L2 acoustic model (as in [5]); the second one utilizes the forced alignment information derived from the L2 acoustic model and extracts a confidence score of the most likely phoneme from the L1 acoustic model phone sets. These confidence scores represent a projection of the speaker's acoustics on the L1 acoustic model and the L2 acoustic model. One set of features estimates the distance of a phoneme pronounced by non-native speaker from native-like pronunciation and the other one estimates the distance of a phoneme pronounced nonnatively in L2 to the closest phoneme in the speaker's L1. Furthermore, we designed an utterance-level feature extraction scheme based on phoneme-level measurements, which can be concatenated and used as an input to a statistical model to predict a pronunciation score. Both implementations are open-sourced.

To the authors' knowledge, there is only one study that uses both L1 and L2 acoustic models to extract measurements for automatic pronunciation evaluation. The authors in [9] used

utterance-level confidence scores extracted from both L1 and L2 acoustic models, calculated frame-wise and averaged over the utterance. However, our proposed system has an important difference: our confidence scores are calculated on phoneme segments and provide more specific information regarding the accentedness of different phonemic categories. Furthermore, in [9], the authors assume the human evaluator can speak both L1 and L2 and experiments were conducted on only one L1. In our study, we want to investigate if the L1 acoustic model can help improve prediction even if the human evaluators have no knowledge of the underlying L1; we carry out experiments with 4 L1s, including Mandarin, Spanish, German and French. The target language is always American English.

We evaluate the proposed system on an accented speech dataset based on a subset of the GMU speech accent archive [10]. Accentedness scores are collected on Amazon Mechanical Turk (AMT) with judgements from 13 human evaluators for each speaker. Utterance-level features are extracted and sent to a linear regression model. Leave-one-speaker-out cross validation is used to measure the consistency between model predictions and human scores. We show that the proposed system has better consistency with human evaluators compared to systems that only use the target language acoustic model.

2. Datasets and Methods

2.1. Datasets and accentedness annotaion

Native speech corpus: To build the target language acoustic model (English for this study), we use the LibriSpeech corpus [11] and the corresponding training scripts[1] in the Kaldi toolkit [12]. The final acoustic model is a triphone model trained with Gaussian Mixture Model-Hidden Markov Model on 960 hours of speech data. The DNN based model was not used because in our experiments we observed that the DNN acoustic model tended to overestimate the pronunciation score. The feature input is a 39-dimensional second order Mel-Frequency Cepstral Coefficient (MFCC) with utterance-level cepstral mean variance normalization and Linear Discriminant Analysis transformation.

For Mandarin, the publicly accessible AIShell Mandarin Speech corpus (approximately 150 hours training data) [13] and the corresponding Kaldi scripts[2] are used. A pronunciation dictionary is included with the dataset. For the remaining three languages (Spanish, French and German), there are no well organized publicly available data. We use data from the Voxforge project and download the speech corpora for French ($\approx$ 30 hours), German ($\approx$ 50 hours) and Spanish ($\approx$ 50 hours). Kaldi scripts[3] for the Voxforge. The dictionary for these three languages are from the CMU Sphinx (Download available[4]). Feature types and structures of acoustic models for the four languages are the same as those used in the English acoustic model. **Non-native speech corpus and accentedness annotation:** The non-native speech corpus used in this study is a subset of the GMU speech accent archive [10] consisting of speakers whose L1s are the aforementioned four languages and native American English. The speakers are chosen carefully to reduce the accent variability and gender imbalance, and to avoid recordings with high background noise. There are 30 speakers for
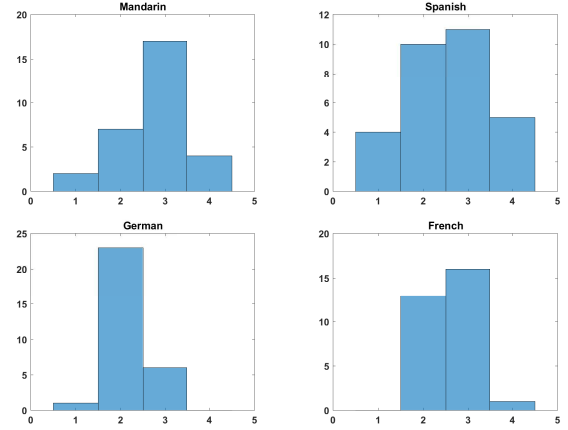
---

[1]https://github.com/kaldi-asr/kaldi/tree/master/egs/librispeech/s5
[2]https://github.com/kaldi-asr/kaldi/tree/master/egs/aishell/s5
[3]https://github.com/kaldi-asr/kaldi/tree/master/egs/voxforge/s5
[4]https://sourceforge.net/projects/cmusphinx/files/Acoustic%20
and%20Language%20Models/

Figure 1: *Histograms of accentedness scores of different L1s.*

each language, and each speaker reads the same paragraph in English. This results in a dataset with 150 speech recordings. We recruit 13 human evaluators on AMT to rate the accentedness of the 150 speakers with random order and unknown speakers' L1s. The annotators are all native American English speakers and have no or little experience with the four foreign languages. We use a four point annotation scale: 1 = no accent/negligible accent, 2 = mild accent, 3 = strong accent, and 4 = very strong accent. The average duration of the annotation task is $\approx$ 45 minutes and each annotator received \$1.50 (twice the reward in [10] on similar listening tasks) for their participation in the study.

We take the average of all 13 evaluators as the final accentedness rating for each speaker; other studies have used the average of 10 AMT non-expert annotations in other natural language tasks [14]. The average inter-rater correlation coefficients (calculated as the average of all annotators' correlation with other annotators) is 0.73. In Fig. 1, we show the histograms of the collected ratings across four different foreign languages. Results show that Mandarin speakers have the strongest accent while German speakers have the mildest accent. This is consistent with expectations considering the phonological similarity between German and English as opposed to other 3 languages. For comparison, the average accentedness rating of native English speaker in our dataset is 1.07. The low mean and lack of strongly-accented speakers in the German and French database also means that the variances of the accentedness ratings for these language are relatively low. This poses a challenge in the statistical modeling and will be addressed in section 3.

2.2. Feature extraction and system building

Features based on the L2 acoustic model: Motivated by the work in [5], we measure the goodness of pronunciation for each phoneme in the accented speech. To do this, the accented speech is first force-aligned at the phoneme-level using the L2 acoustic model to provide the start and end frame indices of each phoneme. We define the pronunciation score (PS_{L2}) of the target phoneme p after alignment as

$$PS_{\text{L2}}(p) = \log(P(p|\mathbf{O}^p)) / |\mathbf{O}^p|$$

$$= \log \left[\frac{P(\mathbf{O}^p|p)P(p)}{\sum_{q \in Q} P(\mathbf{O}^q|q)P(q)} \right] / |\mathbf{O}^p|, \quad (1)$$

where $\mathbf{O}^p$ is the feature matrix of phoneme p, $|\mathbf{O}^p|$ is the num-

ber of frames of phoneme p after alignment, and Q is the set of all phonemes. If we assume equal priors for all phonemes, we approximate the denominator in Eq. 1 with max operator,

$$PS_{\text{L2}}(p) = \log \left[\frac{P(\mathbf{O}^p|p)}{\max_{q \in Q} P(\mathbf{O}^q|q)} \right] / |\mathbf{O}^p|. \qquad (2)$$

The conditional likelihood of each phoneme (given the speech frames of the corresponding aligned segment) can be calculated by decoding the sequence of speech features using the L2 acoustic model. It is clear that if the most likely phoneme returned by the acoustic model is the same as the target phoneme p, then $PS_{\text{L2}}(p) = 0$; otherwise, this value will be negative. The interpretation is that the closer $PS_{\text{L2}}(p)$ is to zero, the closer the pronunciation of phoneme p is to that of native speakers.

L1 acoustic model based measurements: In contrast to the PS_{L2} score, there does not exist a transcript in L1 for the accented speech to measure pronunciation of the phonemes in L1. We define a new way to calculate the pronunciation score with the L1 acoustic model which quantifies how close the pronunciation of a phoneme in L2 is to a specific phoneme in L1. The forced-alignment calculated with the L2 acoustic model is used here. We first decode the speech frames with the L1 acoustic model and find the state path with the highest likelihood. In the path, the corresponding phonemes of each HMM state are recorded and the phoneme with the highest occurrence is considered as the most likely L1 phoneme for a given speech segment. Then, the pronunciation score is calculated as

$$PS_{\text{L1}}(p) = \left[\sum_{t \in T_p} \log \frac{\sum_{s \in S_p} P(o_t|s)}{\sum_{s \in S} P(o_t|s)} \right] / |T_p|, \qquad (3)$$

where o_t is the feature vector for frame t and p is the phoneme with the highest occurrences in the best decoding path of the current segment. T_p is the set of frames where each frame corresponds to an HMM state of phoneme p. S_p is the set of HMM states that belong to phoneme p and S is the set of all HMM states. $PS_{\text{L1}}(p)$ essentially quantifies the confidence of the L1 acoustic model that phoneme p was produced for a speech segment. With eq. 3, a pronunciation score based on the L1 acoustic model can be calculated for each phoneme segment in the original alignment. The implementations of both sections are available on Github[5].

Regression-based accentedness prediction: A diagram of the complete system including forced-alignment, phoneme-level pronunciation score calculation, sentence-level feature extraction and accentedness prediction is shown in Fig. 2. After phoneme-level features $PS_{\text{L2}}(p)$ and $PS_{\text{L1}}(p)$, are extracted, we use a sentence-level feature extraction scheme to convert phoneme-level measurements to a feature vector with a fixed dimension for each utterance. We first combine the pronunciation features for vowels, consonants and syllables and then calculate four statistics for each of these three phonemic categories: for both $PS_{\text{L2}}(p)$ and $PS_{\text{L1}}(p)$, we calculate the minimum, mean, standard deviation and mean-normalized standard deviation (standard deviation divided by mean) of phoneme-level pronunciation measurements of vowels, consonants and syllables in each utterance (implementation available[6]). This results in a total of

---
[5]https://github.com/tbright17/kaldi-dnn-ali-gop
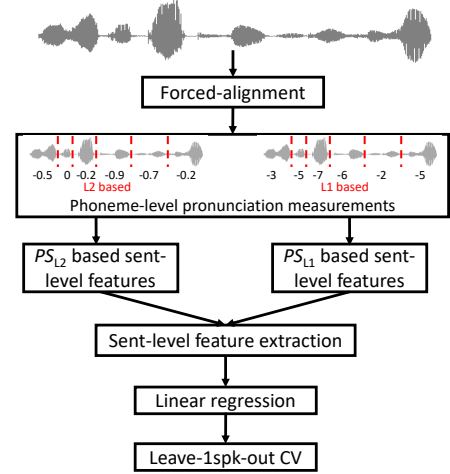[6]https://github.com/tbright17/accent-feat

Figure 2: *System diagram.*

12 utterance-level features for the acoustic model of each language, and a total of 24 utterance-level features combining both pronunciation information from L1 and L2 acoustic models.

To evaluate the predictive ability of this feature set, we build a linear regression model to predict the annotated accentedness from the input feature vector. Since $PS_{\text{L1}}(p)$ is measured with different acoustic models for different languages, we build a different regression model for each L1. We use leave-one-speaker-out cross validation (CV) to estimate the prediction for each test speaker with the remaining speakers used as training data. A simple linear model is used over more complex non-linear models since there are only 30 speakers per language. The system that uses only the 12-dimensional features extracted from only the L2 acoustic model is used as a baseline.

3. Result analysis

Feature Visualization: We first illustrate that the extracted pronunciation features provide relevant information regarding the perceived accentedness ratings. In Fig. 3 we show four scatter plots relating the accentedness ratings and one of the pronunciation features with Pearson correlation coefficients and statistical significance. The two plots in the first row are for Mandarin speakers. The left plot shows the relationship between the human ratings of accentedness (X-axis) and the value of PS_{L2} averaged over all vowels ($PS_{\text{L2}}\_$avgV on Y-axis). The right plot shows the relationship between the human ratings of accentedness (X-axis) and the value of PS_{L1} averaged over all vowels ($PS_{\text{L1}}\_$avgV on Y-axis). The second row shows the same figures for Spanish. It is clear that accentedness and $PS_{\text{L2}}\_$avgV have a negative correlation since larger $PS_{\text{L2}}\_$avgV implies that pronunciation of vowels is closer to native-like pronunciation (and thus a lower accentedness score); accentedness and $PS_{\text{L1}}\_$avgV have a positive correlation since a larger $PS_{\text{L1}}\_$avgV means pronunciation of vowels is closer to L1 pronunciation (and thus higher accentedness score). This provides some confidence that our features exhibit a predictable relationship with accentedness.

Accentedness Prediction: After extracting utterance-level features, each speaker has a feature vector and a corresponding accentedness score (in the range of 1 to 4). For speakers that belong to the same L1 category, a linear regression model with an L2 norm regularizer (or ridge regression) is built with data from
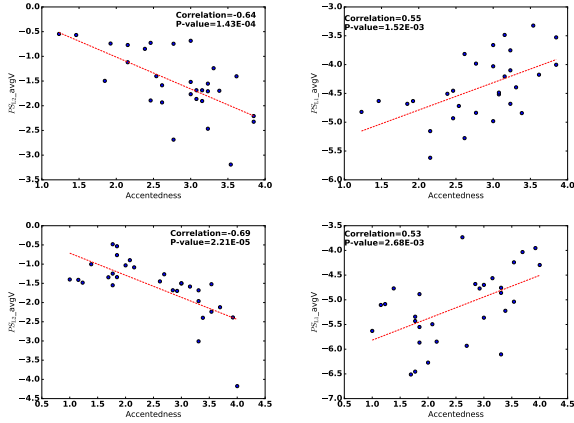
Figure 3: *Scatter plots between accentedness scores and one dimension of features for Mandarin (first row) and Spanish (second row) speakers.*

29 speakers used to train the model and the remaining speaker used to evaluate the model. Feature selection based on the univariate linear regression test [15] was also used to select the most predictable features. The scikit-learn toolkit was used to implement feature selection and ridge regression [16]. To generate the accentedness prediction for all speakers, we perform the evaluation using leave-one-speaker-out CV, which is an unbiased estimate of generalization error [17]; this means that a feature selector and a ridge regression model is trained on all combinations of 29 speakers out of 30 speakers and tested on the 1 remaining. For different input features (12-dimensional utterance-level features or 24-dimensional utterance-level features) we tuned hyperparameters for optimal performance.

As mentioned in section 2.1, the accentedness label distributions for German and French speakers do not span the 1-4 rating scale uniformly. Our initial result revealed that the model performance on German and French speakers was comparatively lower (but there was still improvement over the baseline model). In an attempt to train our model with more uniformly distributed labels, we down-sample the German speakers from 30 to 18 and French speakers from 30 to 22 in an attempt to uniformly sample the labels. For other two languages, there are still 30 speakers in the results. The Pearson correlation coefficient (PCC, higher better) and the mean absolute error (MAE, lower better) are used to measure the relationship between model prediction and human scores.

In table 1, we show both the PCCs and MAEs between model predicted accentedness and human annotated accentedness for 4 groups of speakers. We also show the results of German and French speakers before down-sampling in the parentheses. The results for French are comparatively lower; this could be because of the acoustic model was built with a smaller dataset or because of random sampling. In terms of the main purpose of this study, there is a clear improvement when adding PS_{L1} based features for all 4 L1s. It shows that there is an improvement in model performance consistently and across all languages after adding features from the L1 acoustic model. This is despite the fact that the annotators know little about the acoustic properties of the speakers' L1s.

4. Discussion

The results in table 1 reveal that the improvement in performance varies across different L1s. There are several possible

Table 1: *PCCs and MAEs between predicted accentedness and human scores for speakers of 4 different L1s.*

	PS_{L2} only		PS_{L2} and PS_{L1}	
	PCC	MAE	PCC	MAE
Mandarin	0.707	0.343	**0.727**	**0.329**
Spanish	0.681	0.535	**0.730**	**0.464**
German	0.751 (0.082)	0.192 (0.301)	**0.833** **(0.144)**	**0.163** **(0.287)**
French	0.371 (0.254)	0.373 (0.406)	**0.556** **(0.411)**	**0.315** **(0.370)**

reasons for this including the different modeling quality of the L1s' ASR systems, the accentedness annotation quality, or the contribution of articulation features to perceived impressions of accentedness for different languages. Another interesting aspect that is worthy of additional investigation is that although there is knowledge transfer from L1 to L2 during L2 acquisition, this influence can vary across different L1s and even different speakers. For example, recent research suggests that there exist some universal effects in L2 learning process that are independent of a speaker's L1 [2]. Our approach may provide a means of comparing L1-specific and L1-agnostic pronunciation errors to computationally identify some of the universal effects.

We have shown that our proposed feature sets can boost the performance of accentedness prediction. However, there is still room for improvement. First, as mentioned previously, the GMU speech accent archive dataset has a limited number of speakers and small variation of accentedness for some languages. The recording environment also varies by speaker. A cleaner dataset with uniform accentedness ratings is better suited for our application. Second, the amount and quality of training data for L1 acoustic models can be improved since it is quite limited for some of the languages (Spanish, German and French in this study). More accurate L1 acoustic models may result in an improvement of algorithm performance. Third, it is well known that accentedness is related to both pronunciation and rhythmic features. It is natural to extend the same framework for pronunciation scoring to rhythm features.

5. Conclusions

In this paper, we used both the L1 and L2 acoustic models to extract features for automatic pronunciation evaluation of accented speech. Two sets of phoneme-level pronunciation measurements are developed to quantify both the deviation of native L2 pronunciation and the similarity with speaker's L1 pronunciation. By combining these two sets of features, we develop a new scheme for extracting sentence-level features to predict human-perceived accentedness scores of accented speech. Experiments on accented speakers from 4 different L1s show that there is an improvement in the model's ability to predict accentedness when pronunciation features from both L1 and L2 are included in the model. Future work includes improving the quality of the L1 models we use in the feature extraction and expanding the model to suprasegmental prosodic features in an attempt to model language rhythm.

6. Acknowledgements

7. References

[1] J. E. Flege, "Second language speech learning: Theory, findings, and problems," *Speech perception and linguistic experience: Issues in cross-language research*, pp. 233–277, 1995.

[2] C. B. Chang, *First language phonetic drift during second language acquisition*. University of California, Berkeley, 2010.

[3] Y. Jiao, M. Tu, V. Berisha, and J. M. Liss, "Accent identification by combining deep neural networks and recurrent neural networks trained on long and short term features." in *Interspeech*, 2016, pp. 2388–2392.

[4] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, "Automatic pronunciation scoring for language instruction," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 2. IEEE, 1997, pp. 1471–1474.

[5] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[6] W. Hu, Y. Qian, and F. K. Soong, "An improved dnn-based approach to mispronunciation detection and diagnosis of l2 learners' speech." in *SLaTE*, 2015, pp. 71–76.

[7] J. Tao, S. Ghaffarzadegan, L. Chen, and K. Zechner, "Exploring deep learning architectures for automatically grading non-native spontaneous speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 6140–6144.

[8] Y. Qian, K. Evanini, X. Wang, C. M. Lee, and M. Mulholland, "Bidirectional lstm-rnn for improving automated assessment of non-native childrens speech," in *Proceedings of Interspeech*, 2017, pp. 1417–1421.

[9] N. Moustroufas and V. Digalakis, "Automatic pronunciation evaluation of foreign speakers using unknown text," *Computer Speech & Language*, vol. 21, no. 1, pp. 219–230, 2007.

[10] S. Weinberger, "Speech accent archive," *George Mason University*, 2013.

[11] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 5206–5210.

[12] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[13] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," *arXiv preprint arXiv:1709.05522*, 2017.

[14] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks," in *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2008, pp. 254–263.

[15] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[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] A. Elisseeff, M. Pontil *et al.*, "Leave-one-out error and stability of learning algorithms with applications," *NATO science series sub series iii computer and systems sciences*, vol. 190, pp. 111–130, 2003.