Multi-level Fusion of Fisher Vector Encoded BERT and Wav2vec 2.0 Embeddings for Native Language Identification

Dani Krebbers[1], Heysem Kaya[1](✉), and Alexey Karpov[2]

[1] Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands
h.kaya@uu.nl

[2] St. Petersburg Federal Research Center of the Russian Academy of Sciences, St. Petersburg, Russia
karpov@iias.spb.su

Abstract. Native Language Identification is a prominent paralinguistic study with applications ranging from biometric analysis to speaker adaptation. Former studies on this task have benefited from alternative acoustic feature representations and pre-trained neural networks. In this work, we explore the Native Language Identification performance of contextual acoustic (wav2vec 2.0) and linguistic (BERT) embeddings as state-of-the-art feature representations and combine them with acoustic features at different levels. We encode acoustic and linguistic features using Fisher Vectors, applying Fisher Vector encoding on BERT word embeddings and wav2vec 2.0 for the first time for a paralinguistic task. We compare this approach with conventional functional summarization. In line with our former study using only acoustic modality, the results indicate the superiority of Fisher Vectors encoding over the traditional techniques. Moreover, we show the efficacy of combining alternative representations now in both acoustic and linguistic modalities. Results indicate a notable contribution of the transformer-based contextual auditory and linguistic feature representations to bimodal Native Language Identification systems.

Keywords: Computational Paralinguistics · Native Language Identification · BERT · Wav2vec 2.0 · Fisher Vector

1 Introduction

Native Language Identification (NLI) is a field related to Natural Language Processing (NLP) that focuses on deriving someone's native language (L1) based on speech or writing in a later learned language (L2). Studies in this field mainly focus on non-native English speakers, where we address the task as a classification problem. The assumption, that motivates this research area, is that a

S. R. M. Prasanna et al. (Eds.): SPECOM 2022, LNAI 13721, pp. 391–403, 2022.
https://doi.org/10.1007/978-3-031-20980-2_34

speaker's L1 influences his/her L2. Therefore, by deriving characteristics from someone's L2, their L1 can be determined.

Motives to explore this domain are manifold. First, we can apply it to computer linguistics, mainly used for authorship profiling [22]. Second, it is helpful for the automated personalization of educational applications. Adapting facets of a system, such as feedback, based on their native language, is beneficial for the learning process [27]. Third, it can be used for spoken language applications, where automatic speech recognition (ASR) systems are customized to specific L1s. Finally, it can help with second language acquisition research [19,20]. Medical spheres that relate to treatment of speech disorders are possible as well.

NLI is closely related to the overarching topic of language identification, and dialect identification [3], given that those topics cover classification tasks based on acoustic and linguistic information as well. Nativeness is another closely related topic, as covered in [28]. It defines the degree of someone's L2 capabilities. This degree is measured on a continuous scale and is determined by expert decisions. This subjective measure is in contrast with NLI, where there is an objective truth for someone's L1. This more objective performance measure for NLI makes it a more sound research domain, less prone to subjective influences.

In this study, we propose a bimodal approach using classical and state-of-the-art acoustic and linguistic feature representations extending our contribution [15] to ComParE INTERSPEECH 2016 Computational Paralinguistics Challenge (ComParE) [29] - Native Language Identification Sub-challenge (NLI SC). This sub-challenge features an 11-class NLI task. The contributions of this work are manifold. First, we apply Fisher Vector (FV) representation on BERT and wav2vec embeddings for the first time, as opposed to applying functionals (e.g., averaging) to these state-of-the-art contextual representations. We propose a bimodal system for the NLI task, comparing alternative fusion strategies at feature and decision levels. We carry out extensive experiments on different acoustic feature representations, FV hyperparameters, and summarize the best performances per feature representation.

2 Background and Related Work

2.1 Transformer-Based Linguistic Features

Given the performance and generalizability of pre-trained transformer-based models, we use BERT to capture the linguistic contents of utterances [5]. While models like BERT are often used in an end-to-end fashion being fine-tuned on specific tasks [9], we extract contextual word embeddings using only the pre-trained model. In [5], the authors found that using a weighted sum of the last four hidden layers to obtain embeddings resulted in the best performance, therefore we adopt this in our system as well.

The 2016 ComParE NLI Sub-challenge corpus did not include any textual data. Therefore, we used Google Cloud's speech-to-text[1] services to obtain transcripts of the audio. Valuable linguistic content can be captured using ASR services even though their performance is not perfect. The ASR errors can even be

[1] https://cloud.google.com/speech-to-text.

used to distinguish different speaker classes according to Shivakumar et al. [31], since "an ASR transcript will contain consistent errors based on consistent mispronunciations resulting from L1 specific phonemic confusability".

2.2 Transformer-Based Acoustic Features

In recent years, transformer-based models have been successfully used on various NLP tasks. While they are often used on linguistic content, acoustic input-based models have been implemented as well. The most well-known audio input-based system in this category is wav2vec 2.0 [2]. As seen in Fig. 1, it consists of two main components: a Convolutional Neural Network (CNN) based feature encoder, and a transformer to contextualize the representations. While wav2vec 2.0 is often used in an end-to-end classification fashion, it can also produce vector representations, as the name suggests. Wav2vec 2.0 has been pretrained in a self-supervised manner on large amounts of data. It is often fine-tuned for specific tasks such as speech recognition [4], emotion recognition [21], speaker verification and language identification [7]. Here, we use the wav2vec2-base-960h pretrained model from[2] without further self-supervised training or supervised fine-tuning, extracting embeddings as the low-level descriptors (LLDs) to form the basis for more elaborate feature representations. In [26], the authors showed that aggregating wav2vec 2.0 embeddings outperforms supervised counterparts, and they show aggregation is suitable for extracting phonotactic constraints. In [21], the authors showed the effectiveness of using different layers from the pretrained wav2vec model on emotion recognition tasks. They also showed that wav2vec embeddings improved results in combination with features from other modalities, in their case prosodic. This suggests that using wav2vec embeddings is complementary to other feature representations.

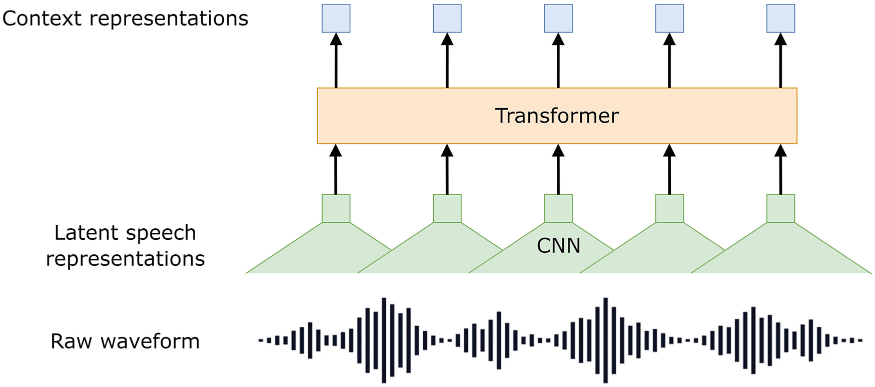


Fig. 1. Framework for wav2vec 2.0, adapted from [2].

[2] https://huggingface.co/facebook/wav2vec2-base-960h.

2.3 Fisher Vector Encoding

With Fisher Vector encoding [23] we can create fixed-length utterance-level features over an arbitrarily long sequence of LLDs. Our proposed system applies FV to the acoustic LLDs as well as the linguistic representations. While FV encoding has been applied to Word2vec word embeddings in the past [24], to our knowledge, we are the first to use FV encoding on contextualized BERT and wav2vec embeddings.

FV encoding takes a background probability model, usually a Gaussian Mixture Model (GMM) with diagonal covariances, and quantifies the change in the GMM parameters needed to fit the new incoming data (e.g., the LLDs of an utterance). We measure both first (mean) and second-order (variance) statistics for the combination of each mixture component and each descriptor. This results in a $2*d*K$ dimensional supervector, where d is the number of dimensions in the data, and K is the number of GMM components. To efficiently learn the diagonal covariance GMM, the data needs to be decorrelated. We apply Principal Component Analysis (PCA) for this purpose, as well as for dimensionality reduction. To reduce computational costs, we use every k-th frame for the acoustic LLDs before learning PCA and the GMM, where initially k is set to 4. However, we apply the fitted PCA model and FV encoding to all frames from an utterance.

2.4 Kernel Extreme Learning Machines

In our fusion scheme, we use Kernel Extreme Learning Machine (KELM) [12] to model high-dimensional feature vectors, motivated by its fast and accurate learning capability and state-of-the-art results on paralinguistic challenge corpora [15,16,32]. We obtain kernels from the dataset and use them in KELM, optimizing the hyper-parameters on the development set.

ELM was initially proposed as an alternative to back-propagation: a fast learning method for Single Hidden Layer Feed-forward Networks (SLFN) [13]. In this approach, the input layer weights are randomly generated and then orthogonalized, while the second layer weights are optimized via (regularized) least squares. The Kernel ELM approach, however, does not benefit from random hidden layer generation but from direct use of kernels for regularized least-squares-based learning. Here, a hyper-parameter C is introduced for regularization of the kernel. Given a kernel $\mathbf{K}$ and the label vector[3] $\mathbf{T} \in \mathbb{R}^{N\times 1}$, where N denotes the number of instances, the projection vector β is learned as follows [12]:

$$\beta = (\frac{\mathbf{I}}{C} + \mathbf{K})^{-1}\mathbf{T}. \tag{1}$$

In order to prevent parameter over-fitting, we use the linear kernel $\mathbf{K}(x,y) = x^T y$, where x and y are the (normalized) feature vectors. With this approach, the only parameter of our model is the regularization coefficient C, which we optimize on the development set.

[3] In case of classification, $\mathbf{T}$ represents the one-hot-encoding matrix of the training set class labels.

3 Proposed NLI Framework

The pipeline of the proposed NLI system is illustrated in Fig. 2. In the framework, the upper pipe that combines reduced baseline openSMILE features with FV encoded MFCC+RASTA-PLPC features, extending the best performing approach used in our contribution to the challenge [15]. In our former paper, due to computational limitations, only Mel-Frequency Cepstral Coefficients (MFCC) features were used in FV encoding, while the fused acoustic representations were modeled with both Kernel ELM used here and Kernel Partial Least Squares Regression-based classifier for subsequent weighted decision fusion [33]. The research question tackled in this work investigates the contribution of transformer-based state-of-the-art acoustic and linguistic embeddings and the suitability of the FV encoding of these embeddings. For the final predictions, we use weighted score fusion on the individual classifier scores.

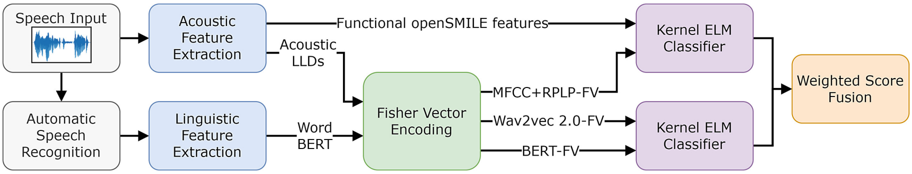


Fig. 2. The pipeline of the proposed bimodal NLI system. RPLP is short for RASTA-PLPC.

3.1 Extracting Conventional Acoustic LLDs

In line with our former work on the same task [15], we extract MFCCs 0-24 and RASTA-style Perceptual Linear Prediction Cepstral Coefficients (RASTA-PLPC) [10,11] using 12th order linear prediction as low-level descriptors. Both of these LLDs are extracted from 25 ms windows with 10 ms steps and are augmented with their first and second-order delta coefficients yielding LLD-vectors of length 75 (3×25) and 39 (3×13) for MFCC and RASTA-PLPC, respectively. For a direct comparison with the openSMILE-based challenge baseline feature set, we also extract LLDs with openSMILE [6] using INTERSPEECH 2013 configuration [30], which has been used in the ComParE series since then.

3.2 Fusion Schemes

Based on our experience with former audio and video signal processing challenges [14,15,32], in this work we propose a stacking framework, where we use feature level fusion followed by a decision level fusion using simple weighted score fusion(SF), where the classifier confidence scores S^A and S^B are fused using the weight $\gamma \in [0, 1]$:

$$S^{fusion} = \gamma * S^A + (1 - \gamma) * S^B. \tag{2}$$

In the pipeline, FV representations of state-of-the-art transformer-based contextual acoustic and linguistic embeddings are fused at feature level on one pipe, and classical acoustic features that are used in our contribution to ComParE 2016 [15] on the other pipe. We further compare the stacking approach with class-based weighted score fusion where we replace γ with a vector Γ containing the weights for each of L classes. This vector contains bounded normal distribution values (mean=0.5 and variance=0.1) where $\Gamma \in [0,1]^L$ and is found by random sampling. A third score fusion method we apply is by using Random Forest (RF) classification applied to the individual classifier scores.

4 Experimental Results

We conducted unimodal and bimodal experiments to showcase the contribution of the proposed transformer-based acoustic and linguistic features with FV encoding. To this end, we also compare the classical functional summarization-based approach with FV representation on acoustic and linguistic representations. In all experiments, we apply a cascaded normalization, composed of feature level z-normalization, value level (signed) power normalization, and feature vector-level L_2 normalization, respectively, as in [15,17]. Below we briefly revisit the dataset and subsequently introduce our results.

4.1 ComParE 2016 Native Language Corpus

As shown in Table 1, in total there are eleven different native languages in the dataset. The complete dataset has around 45 s of speech from each of the 5,132 distinct speakers, all in non-native English. The training partition for all different classes consists of 300 instances, making the class distributions equal. Overall, the classes are distributed fairly equally when we include the development and test set.

Table 1. ETS corpus of non-native spoken English with ComParE 2016 challenge split.

#	Train	Dev	Test	Σ
Arabic	300	86	80	466
Chinese	300	84	74	458
French	300	80	78	458
German	300	85	75	460
Hindi	300	83	82	465
Italian	300	94	68	462
Japanese	300	85	75	460
Korean	300	90	80	470
Spanish	300	100	77	477
Telugu	300	83	88	471
Turkish	300	95	90	485
Σ	3300	965	867	5132

4.2 Comparative Experiments with Unimodal Features

The FV extraction with MFCC+RASTA-PLPC LLD combination, we experimented with P = {90, 100, 110} PCA dimensions and K = {64, 128, 200, 256} GMM components. For a better insight in comparison with the challenge baseline features, we extract the same set of openSMILE (OS13 - openSMILE feature set with INTERSPEECH 2013 configuration) LLDs as the challenge paper [29], apply both functionals, and FV encoding for utterance representation, in addition to the original set of acoustic features. We use 10 functionals that include mean, standard deviation, min (+ its relative location), max (+ its relative location), zero-crossing rate, coefficients of the first (slope, offset) and second-order (curvature) polynomials fit to the LLD contours.

BERT and wav2vec 2.0 output 768 dimensional embeddings and hence we use P = {350, 400, 450, 500} PCA dimensions with K = {64, 128} GMM components. BERT provides word-level embeddings, while wav2vec 2.0 provides an embedding for 25 ms windows of the speech signal with 20 ms steps. To alleviate the computational issues and obtain quasi-phoneme level information, we summarize wav2vec over consecutive 5 frames. The summary of best development set performances in terms of the Unweighted Average Recall (UAR) for each utterance feature representation is given in Table 2. Here, we have multiple observations. The first observation is that FV representation dramatically boosts the performance compared to a simple use of 10 functionals both with proposed MFCC+RASTA-PLPC and with openSMILE LLDs. The second observation is that wav2vec embedding with FV representation outperforms the best conventional LLDs-based FV model, reaching 73.65% UAR on the development set. Third, while a simple mean averaging of BERT outperforms classical acoustic features summarized with 10 functionals, FV encoding of these acoustic features outperforms BERT-FV by 13% to 20% absolute difference.

In line with [15], we apply feature selection using the Canonical Correlation Analysis-based approach [18] and retain the top 5300 features out of the original 6373. This selection improves the Kernel ELM performance of the baseline set to 52.1%. A feature-level combination of the reduced openSMILE set and Ac2 in Table 2 gives a development set UAR score of 72.43%.

4.3 Proposed Bimodal System and Ablation Studies

The test set performances of the bimodal system in comparison with the works presented in the challenge and the state-of-the-art system presented after the challenge are presented in Table 3. The proposed system reaches a test set UAR performance of 83.89%, outperforming the challenge-winning system (UAR 81.30%), while remaining below the current state-of-the-art work of Qian et al. [25] that employs a pre-trained Time-Delayed Deep Neural Network (TDNN)-based i-vector approach using Probabilistic Linear Discriminant Analysis (PLDA) as a classifier.

Of the constituent models, feature fusion of selected openSMILE features with MFCC+RASTA-PLPC FV gives a test set UAR score of 73.44%, while the

Table 2. Unimodal development set UAR (%) performances of the constituent feature representations of the proposed framework. RPLP: RASTA-PLPC, NFun: Summarization using N functionals. P_{PCA} and K_{GMM} represent the number of PCA dimensions and GMM components, respectively.

SysID	LLD	Utterance rep.	Notes	UAR
Ac1	MFCC+RPLP	10Fun		34.05
Ac2	MFCC+RPLP	FV	$P_{PCA} = 110$, $K_{GMM} = 200$	70.95
Ac3	OS13	10Fun		37.59
Ac4	OS13	54Fun	Baseline set [29]	51.17
Ac5	OS13	FV	$P_{PCA} = 130$, $K_{GMM} = 200$	63.31
Ac6	wav2vec 2.0	FV	$P_{PCA} = 400$, $K_{GMM} = 128$	73.65
Lin1	BERT	Mean		40.20
Lin2	BERT	FV	$P_{PCA} = 400$, $K_{GMM} = 64$	50.67

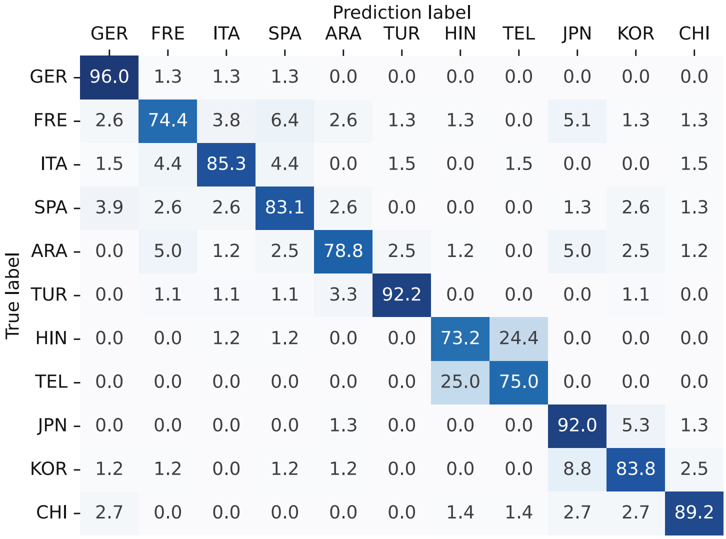
combination of FV representations of wav2vec 2.0 and BERT reaches a test set UAR of 76.52%. We obtain the best performance when we use simple weighted score fusion, resulting in 83.89% UAR on the test set. Removing the wav2vec features from the proposed system decreases the performance to 77.93%. Further removal of the selected openSMILE features (leaving us with BERT FV and MFCC+RASTA-PLPC FV pipes) results in 76.10% UAR on the test set.

Table 3. UAR (%) Performance comparison of the proposed system with literature on ComParE 2016 NLI sub-challenge. The first part reports the performances in the official challenge, the second part reports the performance after the challenge.

Work	Dev.	Test
Baseline paper [29]	45.10	47.50
Gosztolya et al. [8]	70.70	70.10
Kaya and Karpov [15]	67.60	71.50
Shivakumar et al. [31]	78.60	80.13
Abad et al. [1]	84.60	**81.30**
Our system	82.52	83.89
Qian et al. [25]	87.10	**87.20**

Figure 3 illustrates the test set confusion matrix corresponding to the predictions giving 83.89% UAR. Interestingly, results of the current and past research on the dataset have shown that Hindi and Telugu often give the highest confusion. Both languages are used in India, and share similarities in pronunciation, which could be correlated to this higher level of confusion. French, Italian and Spanish share the same language family (Indo-European Romance), while

Japanese, Korean and Turkish are Ural-Altaic languages. Being in the same language families partly explains the relatively higher confusion observed in the former works and the present work for these language subsets. Although Chinese (Sino-Tibetan) is not under the same language family as Japanese and Korean, its tonal nature and geographic proximity to these countries are thought to influence the (mis)classification in this group. Moreover, the confusions within language families (e.g. Indo-European Romance) and among those languages spoken in geographical proximity (such as Hindi & Telugu, as well as Japanese & Korean) can be used to generate groups of languages for a two-stage hierarchical classification in a future work.



	GER	FRE	ITA	SPA	ARA	TUR	HIN	TEL	JPN	KOR	CHI
GER	96.0	1.3	1.3	1.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
FRE	2.6	74.4	3.8	6.4	2.6	1.3	1.3	0.0	5.1	1.3	1.3
ITA	1.5	4.4	85.3	4.4	0.0	1.5	0.0	1.5	0.0	0.0	1.5
SPA	3.9	2.6	2.6	83.1	2.6	0.0	0.0	0.0	1.3	2.6	1.3
ARA	0.0	5.0	1.2	2.5	78.8	2.5	1.2	0.0	5.0	2.5	1.2
TUR	0.0	1.1	1.1	1.1	3.3	92.2	0.0	0.0	0.0	1.1	0.0
HIN	0.0	0.0	1.2	1.2	0.0	0.0	73.2	24.4	0.0	0.0	0.0
TEL	0.0	0.0	0.0	0.0	0.0	0.0	25.0	75.0	0.0	0.0	0.0
JPN	0.0	0.0	0.0	0.0	1.3	0.0	0.0	0.0	92.0	5.3	1.3
KOR	1.2	1.2	0.0	1.2	1.2	0.0	0.0	0.0	8.8	83.8	2.5
CHI	2.7	0.0	0.0	0.0	0.0	0.0	1.4	1.4	2.7	2.7	89.2

Fig. 3. The test set confusion matrix (in %) of the proposed system yielding 83.89% UAR.

4.4 Effect of Design Choices on the Proposed Pipeline

To test the robustness of the proposed pipeline, we test for alternative pipeline components. Simple weighted score fusion is substituted by the individual class-weighted score fusion technique. In this approach, we randomly generate fusion weights for each class independently, and check the fusion UAR performance for a pool of such generated fusion vectors. Using a pool of 50K randomly generated fusion vectors, on the same feature fusion pipelines, this results in 83.50% UAR on the test set, obtaining slightly lower performance compared to the current best pipeline. Stacking the classifier scores to RF results in 80.60% on the test set. This shows that the simple weighted score fusion method is the best option out of these three methods for the task, both regarding performance and simplicity.

The current approach applies L_2 normalization as a final step of the normalization process, followed by calculating a linear kernel as the first step of the classification process. This two-step process essentially results in a cosine kernel. Alternatively, we remove the L_2 normalization and replace the linear kernel with an RBF kernel, since non-linear kernels usually can fit the data better. However, we see a performance drop of around 3% UAR, justifying the use of the simpler cosine kernel approach.

By only retaining the first-order statistics (means) from FV, we reduce the dimensionality by half of the initial size. The resulting feature vector is in line with the Vectors of Locally Aggregated Descriptors (VLAD) representation, even though K-means is more commonly used as a background probability model. Interestingly, as Table 4 shows, we find only minimal performance differences between FV and VLAD.

Table 4. Development and test set UAR (%) performances of the two pipes and the proposed pipeline using alternative clustering-based feature representations and score fusion (SF) methods. Pipe 1 (upper pipe in Fig. 2) uses classical acoustic features with functional and clustering-based utterance representations. Pipe 2 uses transformer-based acoustic and linguistic embeddings.

System	VLAD		FV	
	Dev.	Test	Dev.	Test
Pipe 1	70.55	71.22	72.43	73.44
Pipe 2	76.50	78.45	75.51	76.52
Simple weighted SF of Pipe 1 & 2	83.23	**83.80**	82.52	**83.89**
Class-weighted SF of Pipe 1 & 2	84.37	**83.91**	83.88	83.50

4.5 Further Experiments

Since the FV features consist of large dimensions, we try several feature reduction techniques to further improve the performance by removing redundant information. Instead of single-column feature reduction techniques, we reduce features GMM component-wise. We apply the first method, Permutation Feature Importance (PFI), per section of the whole FV encoding. While results improved marginally during testing (an increase of $< 0.5\%$), we decide the dramatically increased computational costs do not weigh up against this marginal improvement.

We apply PCA after obtaining the FV encodings in a similar fashion as PFI. Different approaches for retaining a considerable amount of explained variance in the data 85%, 95%, 99% did not yield improved results. Furthermore, Linear Discriminant Analysis (LDA) for feature reduction, applied in the same GMM component-wise style, did not improve results either.

5 Conclusions and Future Work

In this paper, we employ transformer-based acoustic and linguistic embeddings as LLDs and model them via Fisher Vector over the utterance for NLI tasks.

Without any further self-supervised pre-training or task-dependent fine-tuning, the transformer-based acoustic and linguistic embeddings modeled with FV provide a marked contribution to the traditional acoustic features for the NLI task. The ablation studies show the overall robustness and preference for simpler versions of pipeline components. Thus, future works may benefit from a simplistic approach with clustering-based modeling using GMM, probably comparing the performances of FV and VLAD earlier, in a preliminary set of experiments. Other future works include task-based fine-tuning of the transformer networks and further prosodic modeling for NLI tasks. Furthermore, a two-level hierarchical classification approach can be exploited to minimize the confusion among highly confused L1 classes.

Acknowledgments. Work of A. Karpov is supported by the RSF (project No. 22-11-00321). We also thank the data donors and challenge organizers for making this research possible.

References

1. Abad, A., Ribeiro, E., Kepler, F., Astudillo, R., Trancoso, I.: Exploiting phone log-likelihood ratio features for the detection of the native language of non-native English speakers. In: Proceedings of Interspeech 2016, pp. 2413–2417 (2016). https://doi.org/10.21437/Interspeech.2016-1491
2. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: a framework for self-supervised learning of speech representations. Adv. Neural Inf. Process. Syst. **33**, 12449–12460 (2020)
3. Chowdhury, S.A., Ali, A., Shon, S., Glass, J.: What does an end-to-end dialect identification model learn about non-dialectal information? In: Proceedings of Interspeech 2020, pp. 462–466 (2020). https://doi.org/10.21437/Interspeech.2020-2235
4. Conneau, A., Baevski, A., Collobert, R., Mohamed, A., Auli, M.: Unsupervised cross-lingual representation learning for speech recognition. arXiv preprint arXiv:2006.13979 (2020)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2019)
6. Eyben, F., Weninger, F., Groß, F., Schuller, B.: Recent developments in opensmile, the Munich open-source multimedia feature extractor. In: Proceedings of the 21st ACM International Conference on Multimedia, pp. 835–838. ACM (2013)
7. Fan, Z., Li, M., Zhou, S., Xu, B.: Exploring wav2vec 2.0 on speaker verification and language identification. In: Proceedings of Interspeech 2021, pp. 1509–1513 (2021). https://doi.org/10.21437/Interspeech.2021-1280
8. Gosztolya, G., Grósz, T., Busa-Fekete, R., Tóth, L.: Determining native language and deception using phonetic features and classifier combination. In: Proceedings of Interspeech 2016, pp. 2418–2422 (2016). https://doi.org/10.21437/Interspeech.2016-962
9. Hao, Y., Dong, L., Wei, F., Xu, K.: Visualizing and understanding the effectiveness of bert. arXiv preprint arXiv:1908.05620 (2019)
10. Hermansky, H.: Perceptual linear predictive (PLP) analysis of speech. the J. Acoust. Soc. Am. **87**(4), 1738–1752 (1990)

11. Hermansky, H., Morgan, N.: Rasta processing of speech. IEEE Trans. Speech Audio Process. **2**(4), 578–589 (1994)
12. Huang, G.B., Zhou, H., Ding, X., Zhang, R.: Extreme learning machine for regression and multiclass classification. IEEE Trans. Syst. Man Cybern. Part B Cybern. **42**(2), 513–529 (2012)
13. Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme learning machine: theory and applications. Neurocomputing **70**(1), 489–501 (2006)
14. Kaya, H., Gurpinar, F., Ali Salah, A.: Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video CVS. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 1–9, July 2017
15. Kaya, H., Karpov, A.A.: Fusing acoustic feature representations for computational paralinguistics tasks. In: Proceedings of Interspeech 2016, pp. 2046–2050 (2016). https://doi.org/10.21437/Interspeech.2016-995
16. Kaya, H., Karpov, A.A.: Introducing weighted kernel classifiers for handling imbalanced paralinguistic corpora: snoring, addressee and cold. In: Proceedings of Interspeech 2017, pp. 3527–3531 (2017). https://doi.org/10.21437/Interspeech.2017-653
17. Kaya, H., Karpov, A.A., Salah, A.A.: Robust acoustic emotion recognition based on cascaded normalization and extreme learning machines. In: Cheng, L., Liu, Q., Ronzhin, A. (eds.) ISNN 2016. LNCS, vol. 9719, pp. 115–123. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-40663-3_14
18. Kaya, H., Özkaptan, T., Salah, A.A., Gürgen, F.: Random discriminative projection based feature selection with application to conflict recognition. IEEE Sig. Process. Lett. **22**(6), 671–675 (2015). https://doi.org/10.1109/LSP.2014.2365393
19. Malmasi, S., Dras, M.: Language transfer hypotheses with linear SVM weights. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1385–1390 (2014)
20. Malmasi, S., et al.: A report on the 2017 native language identification shared task. In: Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 62–75 (2017)
21. Pepino, L., Riera, P., Ferrer, L.: Emotion recognition from speech using wav2vec 2.0 embeddings. arXiv preprint arXiv:2104.03502 (2021)
22. Perkins, R.: Native language identification (NLID) for forensic authorship analysis of weblogs. In: New threats and Countermeasures in Digital Crime and Cyber Terrorism, pp. 213–234. IGI Global (2015)
23. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2007)
24. Plummer, B.A., Kordas, P., Kiapour, M.H., Zheng, S., Piramuthu, R., Lazebnik, S.: Conditional image-text embedding networks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 249–264 (2018)
25. Qian, Y., et al.: Improving sub-phone modeling for better native language identification with non-native English speech. In: Proceedings of Interspeech 2017, pp. 2586–2590 (2017). https://doi.org/10.21437/Interspeech.2017-245
26. Ramesh, G., Kumar, C.S., Murty, K.S.R.: Self-supervised phonotactic representations for language identification. In: Proceedings of Interspeech 2021, pp. 1514–1518 (2021). https://doi.org/10.21437/Interspeech.2021-1310
27. Rozovskaya, A., Roth, D.: Algorithm selection and model adaptation for ESL correction tasks. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 924–933 (2011)

28. Schuller, B., et al.: The interspeech 2015 computational paralinguistics challenge: Nativeness, parkinson's & eating condition. In: Proceedings of Interspeech 2015, pp. 478–482 (2015). https://doi.org/10.21437/Interspeech.2015-179
29. Schuller, B., et al.: The interspeech 2016 computational paralinguistics challenge: deception, sincerity & native language. In: Proceedings of Interspeech 2016, pp. 2001–2005 (2016). https://doi.org/10.21437/Interspeech.2016-129
30. Schuller, B., et al.: The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In: Proceedings of Interspeech 2013, pp. 148–152 (2013). https://doi.org/10.21437/Interspeech.2013-56
31. Shivakumar, P.G., Chakravarthula, S.N., Georgiou, P.: Multimodal fusion of multirate acoustic, prosodic, and lexical speaker characteristics for native language identification. In: Proceedings of Interspeech 2016, pp. 2408–2412 (2016). https://doi.org/10.21437/Interspeech.2016-1312
32. Soğancıoğlu, G., et al.: Is everything fine, grandma? acoustic and linguistic modeling for robust elderly speech emotion recognition. In: Proceedings of Interspeech 2020, pp. 2097–2101 (2020). https://doi.org/10.21437/Interspeech.2020-3160
33. Wold, H.: Partial least squares. In: Kotz, S., Johnson, N.L. (eds.) Encyclopedia of Statistical Sciences, pp. 581–591. Wiley, New York (1985)