Does Language Identification Improve Zero-Shot Cross-Lingual Dementia Detection?

Joseph Kung and Michelle Pan
Department of Electrical and Computer Engineering
University of California, Los Angeles
Email: josephfkung@gmail.com, ypan27@ucla.edu

***Abstract*—Speech-based dementia detection often degrades under cross-lingual transfer. We test whether explicit source information improves zero-shot Alzheimer's disease (AD) versus healthy control (HC) classification across English, Greek, Mandarin, and Spanish, using frozen XLS-R and W2v-BERT 2.0 representations. We compare a pooled baseline, uniform expert ensemble, identity hard routing, combined-distance soft routing, and a post-hoc best-source oracle. Despite near-perfect source identification, routing does not consistently outperform pooling. The best-source oracle shows that source-specific experts retain transferable diagnostic information. For XLS-R on Greek, AUROC rises from 0.441 with pooling to 0.671 with the best source. The failure is selection, not specialization.**

***Index Terms*—Alzheimer's disease, cross-lingual transfer, language identification, mixture of experts, speech classification, zero-shot learning**

I. Introduction

Alzheimer's disease (AD) is a neuro-degenerative disorder that affects memory, executive function, and language. Traditional screening and diagnosis often require trained clinicians, neuro-psychological testing, and in-person evaluations, which can be inefficient for large-scale clinics. An alternative is to use spontaneous speech, which is non-invasive and highly scalable. AD detection on English datasets has been well explored over the past decade, with state-of-the-art speech foundation models (SFMs) and encoders such as HuBERT [1] and wav2vec 2.0 [2] pretrained on massive and diverse datasets, providing transferable representations to downstream AD detection tasks. Benchmarks such as ADReSS [3] and DementiaBank [4] have also supported rapid improvements in AD classification. However, most studies are performed by training and testing on one language or corpus. These advances in English may not yet be generalized to low-resource languages due to the bottleneck from limited labeled data. Performance can deteriorate when the target language differs from the actual training data. Consequently, zero-shot cross-lingual transfer has emerged as a promising research area to address this gap. This paper explores whether it is possible to use language identification to improve zero-shot cross-lingual dementia detection. The literature so far has shown mixed results on the importance of language identification.

Girish et al. [5] proposed that reliable transfer hinges on learning language-invariant multimodal representations, while adversarial learning would prevent the fusion from encoding a specific language identity. By contrast, Wang et al. [6] suggest that features behave inconsistently across languages. They cite a classifier trained on English and zero-shot tested on Mandarin that performed little better than chance, attributing the failure to "cross-lingual polarity flip", a phenomenon in which a feature predicting AD in one language would lean toward predicting HC in another. Wang et al. note that correlation alignment for deep domain adaptation (CORAL) and adversarial learning face challenges in cross-lingual AD detection due to the demand for large amounts of unlabeled data, and the assumption that cross-domain feature shifts are purely numerical, which was not true for the English to Mandarin case. We tested CORAL to see if we should choose the source whose embedding distribution was close to the target, but our findings support the authors' conclusion that cross-domain feature shifts are not consistent across languages.

The limitations of existing domain-adaptation methods have prompted us to seek novel ones. XLS-R [7] and W2v-BERT 2.0 [8], [9] recover source-corpus identity accurately, so we compare the pooled baseline with a uniform expert ensemble, identity hard routing, combined-distance soft routing, and a post-hoc best-source oracle. The key asymmetry is that deployable routing is inconsistent while the experts themselves can transfer useful AD signals: on Greek with XLS-R, the oracle reaches 0.671 AUROC versus 0.441 for the pooled baseline (+0.230).

II. Experiments

A. Datasets and Corpora

Table I reports the AD/HC participants used from four corpora: Mandarin NCMMSC2021 [13], Greek DemCare [14], Spanish Ivanova [15], and English DementiaBank Pitt [4]. Each corpus is held out once as the target. MCI data is excluded before training and evaluation. The Ivanova row (270 participants: 74 AD, 196 HC) is the filtered binary subset used in this study, not the full published cohort, which also includes MCI. Repeated recordings from one participant remain on the same side of every split to prevent leakage. We evaluate five seeds (42–46), extracting frozen representations once. The primary metrics are balanced accuracy (BA), macro-F1, and AUROC.

B. Speech Representation

All the audio was resampled to 16 kHz. We encoded the datasets with a frozen XLS-R model [7] or W2v-BERT 2.0

TABLE I
AD-VS-HC TARGET PARTICIPANTS

Language	Total	AD	HC	Corpus
Mandarin	70	26	44	NCMMSC2021 [13]
Greek	90	55	35	DemCare [14]
Spanish	270	74	196	Ivanova [15]
English	238	128	110	DementiaBank Pitt [4]

TABLE II
FOUR-WAY SOURCE IDENTIFICATION, PARTICIPANT-LEVEL, 5-FOLD CV × 5 SEEDS

Encoder	Accuracy	Macro-F1
XLS-R	0.984 ± 0.010	0.983 ± 0.010
W2v-BERT 2.0	0.977 ± 0.011	0.973 ± 0.018
Paired fold-level difference (XLS-R − W2v-BERT)		
Accuracy	0.007 ± 0.011; 15 / 5 / 5 win-tie-loss	
Macro-F1	0.010 ± 0.016; 18 / 0 / 7 win-tie-loss	

model [8], [9]. The frame-level hidden states are pooled into 1024 dimensions. The encoders were frozen to prevent full-model fine-tuning and to isolate the effect of explicit language conditioning. A lightweight multilayer perceptron predicts AD probability from the pooled representation. Expert hidden dimensions are reduced so that multi-expert systems match the pooled baseline in trainable capacity, which prevents any parameter-count advantage.

C. Routing Signals

We evaluate two routing signals. The *identity* router is a source classifier producing a posterior $\mathbf{z}(\mathbf{x})$ over the L source corpora. Source examples receive out-of-fold posteriors, and the held-out target is scored by a router trained only on source data, which avoids sample leakage. The *distance* router scores each source by an unsupervised discrepancy between source and target embeddings, with weights given by a softmax over negative distance. The three distances differ in how much distributional structure they capture. The centroid distance uses first-order statistics, CORAL [10] uses second-order statistics through the squared Frobenius norm between feature covariances, and the maximum mean discrepancy (MMD) [11] compares mean embeddings in a reproducing kernel Hilbert space, making it sensitive to all moments under a characteristic kernel. The combined score is the standardized average of the three.

III. SYSTEMS AND ROUTING STRATEGIES

A. Pooled Baseline

The pooled baseline trains one AD classifier on all source corpora:

$$\hat{y} = \sigma(f_{\text{pooled}}(\mathbf{x})) \,. \tag{1}$$

B. Uniform Expert Ensemble

With one expert f_ℓ per source corpus, the uniform ensemble averages expert logits:

$$\hat{y} = \sigma\left(\frac{1}{L}\sum_{\ell=1}^{L} f_\ell(\mathbf{x})\right). \tag{2}$$

C. Identity Hard Routing

The identity router selects $r(\mathbf{x}) = \arg\max_\ell z_\ell(\mathbf{x})$ from the source-corpus posterior $\mathbf{z}(\mathbf{x})$, then predicts

$$\hat{y} = \sigma\big(f_{r(\mathbf{x})}(\mathbf{x})\big) \,. \tag{3}$$

D. Combined-Distance Soft Routing

Centroid, CORAL, and MMD discrepancies are standardized and averaged into d_ℓ. The resulting soft weights combine expert logits:

$$w_\ell = \frac{e^{-d_\ell}}{\sum_j e^{-d_j}}, \qquad \hat{y} = \sigma\left(\sum_{\ell=1}^{L} w_\ell f_\ell(\mathbf{x})\right). \tag{4}$$

E. Best-Source Oracle

The best-source oracle reports the source expert with the highest labeled target AUROC.

IV. RESULTS AND ANALYSIS

A. Cross-Target Patterns

Table II reports four-way source identification using a matched participant-level protocol. Both encoders recover the language corpus identity nearly perfectly: XLS-R reaches 0.984 ± 0.010 accuracy and W2v-BERT 2.0 0.977 ± 0.011. There is a small advantage for XLS-R, 0.007 ± 0.011 accuracy and 0.010 ± 0.016 macro-F1, which are negligible gaps. Both encoders recover source identity well, so this ensures that any downstream router built on them has a clean signal. This eliminates the possibility that the results can be attributed to a weak router.

B. No Method Consistently Wins

Table III reports zero-shot AD detection for the pooled baseline, uniform expert ensemble, identity hard routing, combined-distance soft routing, and best-source oracle. The overall absolute performance is weak: excluding the oracle, the target AUROC ranges from 0.330 (XLS-R, Mandarin, identity hard routing) to 0.634 (XLS-R, Spanish, identity hard routing), and balanced accuracy never exceeds 0.569. Next, the best method changes with every target and encoder. The pooled baseline is the best on English for both encoders, which is expected since English has the biggest speech database. Identity hard routing is the best on Spanish for XLS-R but the worst on Spanish for W2v-BERT 2.0. Another result is that the uniform expert ensemble is best on Spanish for W2v-BERT 2.0 and among the worst on Mandarin for XLS-R. This shows that no ordering of the systems is stable across targets.

Another pattern is that for Mandarin and Greek with XLS-R, and on Greek with W2v-BERT 2.0, AUROC is below 0.5, in several cases by more than three seed-level standard deviations

TABLE III
PARTICIPANT-LEVEL ZERO-SHOT AD DETECTION, MEAN ± STANDARD DEVIATION OVER FIVE SEEDS.

		XLS-R			W2v-BERT 2.0		
Target	Method	BA	Macro-F1	AUROC	BA	Macro-F1	AUROC
English	Pooled baseline	**0.556**±0.021	**0.538**±0.020	**0.566**±0.013	**0.527**±0.026	**0.514**±0.032	**0.537**±0.017
	Uniform expert ensemble	0.530±0.027	0.481±0.073	0.563±0.033	0.509±0.012	0.506±0.014	0.507±0.011
	Identity hard routing	0.517±0.018	0.507±0.015	0.529±0.018	0.517±0.022	0.511±0.027	0.525±0.028
	Combined-distance soft routing	0.531±0.030	0.490±0.069	0.548±0.034	0.503±0.020	0.492±0.030	0.494±0.022
	Best-source oracle	*0.535*±0.022	*0.512*±0.043	*0.581*±0.026	*0.519*±0.008	*0.473*±0.023	*0.539*±0.023
Mandarin	Pooled baseline	**0.527**±0.055	**0.428**±0.040	**0.460**±0.041	0.520±0.033	**0.496**±0.035	0.520±0.062
	Uniform expert ensemble	0.405±0.029	0.384±0.018	0.382±0.020	0.516±0.043	0.465±0.050	0.493±0.086
	Identity hard routing	0.417±0.036	0.394±0.034	0.330±0.025	**0.525**±0.058	0.488±0.071	**0.527**±0.083
	Combined-distance soft routing	0.402±0.047	0.379±0.036	0.355±0.032	0.522±0.037	0.471±0.045	0.505±0.071
	Best-source oracle	*0.490*±0.031	*0.449*±0.049	*0.527*±0.019	*0.549*±0.057	*0.512*±0.059	*0.582*±0.089
Spanish	Pooled baseline	0.534±0.026	0.529±0.033	0.568±0.048	0.501±0.026	0.500±0.028	0.483±0.034
	Uniform expert ensemble	**0.569**±0.040	0.508±0.074	0.601±0.044	**0.541**±0.015	**0.510**±0.020	**0.551**±0.022
	Identity hard routing	0.549±0.022	**0.545**±0.031	**0.634**±0.045	0.480±0.029	0.448±0.039	0.472±0.043
	Combined-distance soft routing	0.537±0.042	0.425±0.121	0.594±0.033	0.526±0.010	0.494±0.037	0.544±0.021
	Best-source oracle	*0.571*±0.048	*0.512*±0.070	*0.645*±0.055	*0.536*±0.026	*0.505*±0.021	*0.566*±0.030
Greek	Pooled baseline	0.470±0.038	0.434±0.015	0.441±0.015	0.390±0.057	0.355±0.037	**0.419**±0.033
	Uniform expert ensemble	**0.491**±0.046	**0.463**±0.035	**0.502**±0.075	**0.433**±0.043	**0.418**±0.047	0.385±0.032
	Identity hard routing	0.483±0.044	0.403±0.026	0.462±0.046	0.422±0.049	0.400±0.053	0.404±0.066
	Combined-distance soft routing	0.480±0.047	0.402±0.063	0.498±0.052	0.426±0.042	0.402±0.062	0.366±0.029
	Best-source oracle	*0.606*±0.048	*0.580*±0.083	*0.671*±0.066	*0.505*±0.043	*0.454*±0.081	*0.551*±0.068

(i.e., XLS-R Mandarin identity routing, 0.330 ± 0.025). Below-chance AUROC means that the signal transferred with its sign flipped, thus ranking target AD participants below the target HC participants. This is the cross-lingual polarity flip described by Wang et al. [6].

C. No Comparison Survives Correction

Table IV shows the prespecified paired participant-and-seed bootstrap for the combined-distance soft routing against the pooled baseline and against identity hard routing. Across all 48 comparisons, no Holm-corrected [12] p-value falls below 0.05. The smallest p-value is 0.128. Only three unadjusted 95% confidence intervals exclude zero, and two of them run against combined-distance soft routing: XLS-R Mandarin balanced accuracy ($\Delta = -0.185$, CI $[-0.300, -0.040]$) and XLS-R Spanish macro-F1 versus identity hard routing ($\Delta = -0.244$, CI $[-0.364, -0.030]$). The single interval in its favor is the W2v-BERT Spanish AUROC versus identity ($\Delta = +0.091$, CI $[0.010, 0.155]$). Since about 2.4 of 48 intervals would exclude zero by chance at the 95% level, observing three is not substantial evidence of an effect.

The unadjusted intervals suggest target-specific effects, but the experiment does not support a general conclusion that combined-distance soft routing beats the pooled baseline.

D. The Router Does Not Know the Transfer

Across 20 seed blocks, the identity router selects the best-transferring expert 6 times for each encoder, a hit rate of 0.300 (95% CI $[0.145, 0.519]$) compared to a chance rate of 0.333 with three candidate sources. The router that recovers the corpus with 98% accuracy therefore selects the clinically best source almost at random.

The two questions are unrelated. The rank correlation between routing similarity and target AUROC is -0.002 (CI $[-0.157, 0.185]$; Pearson 0.080) for XLS-R and -0.162 (CI $[-0.412, 0.110]$; Pearson -0.136) for W2v-BERT 2.0. Both point estimates are non-positive.

The mean AUROC regret, defined here as the AUROC forfeited relative to always selecting the oracle source, is 0.108 (CI $[0.054, 0.170]$) for XLS-R and 0.074 (CI $[0.043, 0.110]$) for W2v-BERT 2.0. Unlike the correlations, these intervals exclude zero, so routing on identity carries a measurable cost.

Table V shows how unevenly that cost falls. The router's dominant choice matches the oracle only for XLS-R on Mandarin, where it routes to English in 4 of 5 seeds and English is indeed the strongest source (regret 0.014). XLS-R on Greek instead favors Mandarin while the oracle is Spanish, which no seed selects. Spanish receives only 0.189 of the routing mass, and the resulting 0.323 AUROC loss accounts for roughly three quarters of total XLS-R regret. Across the eight encoder–target configurations, Spanish is the dominant choice in four, and in three of those it is not the oracle. This is expected of a router responding to corpus size and recording conditions rather than clinical content.

E. Distribution Distance

An objection is that a discriminative identity classifier is the wrong tool, and that an unsupervised measure of distribution shift would do better at tracking transfer. Table VI tests four of these measures. None of them does well. Hard oracle hit rates range from 0.050 to 0.350, and all but one are at or below the

TABLE IV
PRIMARY CONTRASTS FOR COMBINED-DISTANCE SOFT ROUTING AND PAIRED PARTICIPANT-AND-SEED BOOTSTRAP. Δ IS THE MEAN PAIRED DIFFERENCE. p IS HOLM-CORRECTED OVER ALL 48 TESTS.

		vs. pooled baseline		vs. identity hard routing	
Target	Metric	Δ	p	Δ	p
XLS-R					
	BA	-0.033	1.000	-0.022	1.000
English	Macro-F1	-0.018	1.000	-0.021	1.000
	AUROC	-0.030	1.000	$+0.023$	1.000
	BA	-0.185	0.128	-0.061	1.000
Mandarin	Macro-F1	-0.097	1.000	-0.057	1.000
	AUROC	-0.136	0.592	$+0.016$	1.000
	BA	-0.064	1.000	-0.050	1.000
Spanish	Macro-F1	-0.261	0.528	-0.244	0.152
	AUROC	$+0.024$	1.000	-0.048	1.000
	BA	-0.013	1.000	-0.039	1.000
Greek	Macro-F1	-0.069	1.000	-0.036	1.000
	AUROC	$+0.075$	1.000	$+0.046$	1.000
W2v-BERT 2.0					
	BA	-0.035	1.000	-0.046	1.000
English	Macro-F1	-0.023	1.000	-0.045	1.000
	AUROC	-0.067	1.000	-0.045	1.000
	BA	-0.022	1.000	-0.030	1.000
Mandarin	Macro-F1	-0.064	1.000	-0.065	1.000
	AUROC	-0.019	1.000	-0.044	1.000
	BA	$+0.001$	1.000	$+0.048$	1.000
Spanish	Macro-F1	-0.029	1.000	$+0.038$	0.735
	AUROC	$+0.055$	1.000	$+0.091$	0.256
	BA	$+0.005$	1.000	$+0.006$	1.000
Greek	Macro-F1	$+0.017$	1.000	$+0.014$	1.000
	AUROC	-0.037	1.000	-0.007	1.000

TABLE V
PER-TARGET IDENTITY ROUTER AND ORACLE ALIGNMENT.

Enc.	Target	Routed	Oracle	Hits	Mass	Regret
XLS-R	English	Spanish	Greek	1/5	0.204	0.029
	Mandarin	English	English	4/5	0.474	0.014
	Spanish	English	Greek	1/5	0.251	0.067
	Greek	Mandarin	Spanish	0/5	0.189	0.323
W2v-B	English	Spanish	Mandarin	1/5	0.352	0.054
	Mandarin	Spanish	English	1/5	0.246	0.118
	Spanish	Mandarin	Greek	1/5	0.252	0.049
	Greek	Spanish	Spanish	3/5	0.420	0.074

0.333 chance level, and the one exception (CORAL with W2v-BERT 2.0) sits within sampling noise of it. Rank correlations between similarity and downstream AUROC are negative for six out of the eight encoder–distance pairs. For XLS-R, greater embedding similarity predicts worse transfer under three of four distances. Soft weighting places only 0.225 to 0.366 of its mass on the oracle source. CORAL is the only measure with positive correlations under both encoders and may merit further study, but distribution distance is not reliable.

F. Specialization versus Selection

The best-source oracle in Table III is the strongest positive result of the study. It sits near the top of every target column, showing that source-specific experts retain clinically transferable information that pooling dilutes. For XLS-R on Greek, the oracle reaches 0.671 AUROC compared to 0.441 for the pooled baseline, with balanced accuracy rising from 0.470 to 0.606. On Spanish, AUROC increases from 0.568 to 0.645.

The central asymmetry of the investigation is that *expert specialization helps even where expert selection fails*. Identity hard routing, centroid distance, CORAL, MMD, and combined-distance soft routing all fail to recover the best source reliably. The oracle narrows the problem rather than fixing it.

TABLE VI
SOURCE–TARGET DISTANCE AS A TRANSFERABILITY SIGNAL. ρ IS THE SPEARMAN CORRELATION BETWEEN SIMILARITY AND TARGET AUROC

Encoder	Distance	ρ	Hit rate	Regret	Oracle mass
XLS-R	Centroid	-0.225	0.200	0.136	0.248
	CORAL	$+0.039$	0.200	0.144	0.348
	MMD	-0.296	0.050	0.190	0.244
	Combined	-0.315	0.050	0.190	0.225
W2v-B	Centroid	-0.334	0.150	0.116	0.303
	CORAL	$+0.251$	0.350	0.038	0.366
	MMD	-0.334	0.150	0.116	0.234
	Combined	-0.073	0.150	0.108	0.305

V. DISCUSSION

The router and the clinical system answer two different questions. The router asks which source corpus a recording resembles, while the classifier needs to know which source expert transfers AD-related information best. Channel and recording conditions carry corpus identity, but they reveal nothing about whether pause structure, prosody, fluency, or lexical retrieval continue to transfer.

Soft routing does not solve this mismatch. Averaging a well-transferring expert with weaker ones will dilute its predictions toward the decision boundary, and our soft router assigns at most 0.366 of its weight to the oracle source even in the best case, which leaves most of the mass on the suboptimal experts.

The below-chance AUROCs on Mandarin and Greek point to a bigger problem. Since source identity remains highly decodable from the same representations, the embedding space carries substantial corpus-specific information. The classifiers may be exploiting features that correlate with diagnosis in a source corpus but relate to it differently in the target, leaving the transferred decision function misaligned or inversely aligned with target labels. This matches the cross-lingual polarity flip reported by Wang et al. [6], and no routing rule can repair it.

VI. LIMITATIONS AND FUTURE WORK

This investigation has several limitations. Language and corpus identity are largely confounded, since each language is represented by a different dataset. The high source-identification accuracy may be due to differences in recording channel rather than linguistic information. We interpret our results in terms of source identity instead of language identity.

The corpora also differ in size and class balance, with the Mandarin and Greek cohorts being small. Participant-level evaluation, class-balanced metrics, multiple seeds, and hierarchical bootstrap analysis mitigate this, but uncertainty is still higher for those targets. Absolute zero-shot performance is close to chance throughout.

Because the source–target similarity is estimated from the full unlabeled target cohort, our routing is transductive rather than inductive, and routing a single unseen participant remains unexplored. The post-hoc oracle additionally uses target labels, so it is an upper bound rather than an attainable result.

Finally, the centroid, CORAL, and MMD capture global representation similarity and may miss the dimensions that carry impairment-related information. With two frozen encoders, four corpora, and a binary AD-versus-HC classification task excluding MCI, establishing generality will require broader languages and encoders.

These limitations point to the clearest direction for future work, which is to reduce corpus–label confounding during representation learning [5]. Where source-specific experts are retained, the router should be optimized for transferability rather than source identity. A transfer-aware router could be learned from a source-to-source transfer matrix, treating each source corpus as a pseudo-target and scoring the remaining experts on it to supervise selection. Another exploration is having the system fall back to pooled training when no source assignment is confident, with expert disagreement serving as a label-free uncertainty signal.

VII. Conclusion

In conclusion, source identity is almost perfectly recoverable from frozen multilingual representations, at 98.4% for XLS-R and 97.7% for W2v-BERT 2.0. This provides almost no benefit for zero-shot classification with hard or combined-distance soft routing, but shows that matching a best-source oracle to the target corpus may achieve higher accuracy for certain languages.

Neither identity hard routing nor combined-distance soft routing consistently outperforms the pooled baseline. None of the comparisons is statistically significant after Holm correction, source–target similarity is uncorrelated with target AUROC, and the router identifies the best-transferring source at the chance rate. The cost of routing is noticeable. The mean AUROC regret reaches 0.108 for XLS-R and 0.074 for W2v-BERT 2.0.

A post-hoc oracle over the same experts recovers up to 0.230 AUROC, so the gap lies in expert selection rather than the experts. The below-chance AUROCs on Mandarin and Greek further suggest that corpus-specific information can invert the transferred decision function, which no routing method explored can repair. As a result, routers should be trained on transferability, not similarity.

References

[1] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.

[2] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, 2020.

[3] S. Luz, F. Haider, S. de la Fuente, D. Fromm, and B. MacWhinney, "Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge," in *Proc. Interspeech*, 2020, pp. 2172–2176.

[4] J. T. Becker, F. Boller, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis," *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, 1994.

[5] Girish, M. M. Akhtar, F. Sheth, M. Singh, J. Gerard, P. McClean, and K. Wong-Lin, "Synergizing zero-shot cross-lingual Alzheimer detection with language-invariant multimodal bi-geometric adversarial learning," *arXiv preprint arXiv:2606.17254*, 2026.

[6] Q. Wang and M. Wu, "Cross-lingual Alzheimer's disease speech detection: Polarity inversion and few-shot calibration strategies," *Bioengineering*, vol. 13, no. 6, p. 629, 2026.

[7] A. Babu, C. Wang, A. Tjandra *et al.*, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," in *Proc. Interspeech*, 2022, pp. 2278–2282.

[8] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "w2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in *Proc. IEEE ASRU*, 2021, pp. 244–250.

[9] Seamless Communication, L. Barrault *et al.*, "SeamlessM4T: Massively multilingual and multimodal machine translation," *arXiv preprint arXiv:2308.11596*, 2023.

[10] B. Sun and K. Saenko, "Deep CORAL: Correlation alignment for deep domain adaptation," in *Proc. ECCV Workshops*, 2016, pp. 443–450.

[11] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, pp. 723–773, 2012.

[12] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, no. 2, pp. 65–70, 1979.

[13] THU Speech and Audio Technology Laboratory, "AD2021: Alzheimer's Disease Recognition Evaluation 2021," NCMMSC2021 Alzheimer's Disease Recognition Challenge resource, 2021.

[14] A. Karakostas, A. Briassouli, K. Avgerinakis, I. Kompatsiaris, and M. Tsolaki, "The Dem@Care experiments and datasets: A technical report," *arXiv preprint arXiv:1701.01142*, 2017.

[15] O. Ivanova, J. G. Meilán, F. Martínez-Sánchez, I. Martínez-Nicolás, T. E. Llorente, and N. C. González, "Discriminating speech traits of Alzheimer's disease assessed through a corpus of reading task for Spanish language," *Computer Speech & Language*, vol. 73, p. 101341, 2022.