

Expressive Fluency in Natural Speech: A Geometric Expressive Fluency Analysis Model Analysis of an Iranian Azeri–Persian Bilingual Cohort

Taghi Shakouri Youvalri¹ , İnci Zaim Gökbay² 

¹Informatics, PhD Programme, İstanbul University, Institute of Graduate Studies in Sciences, İstanbul, Türkiye

²Department of Artificial Intelligence and Data Engineering, İstanbul University, Faculty of Computer and Information Technologies, İstanbul, Türkiye

Cite this article as: T. S. Youvalri and İ. Z. Gökbay, “Expressive fluency in natural speech: A geometric expressive fluency analysis model analysis of an Iranian Azeri–Persian bilingual cohort,” *Electrica*, 26, 0007, 2026. doi: 10.5152/electrica.2026.26007.

WHAT IS ALREADY KNOWN ON THIS TOPIC?

- Acoustic features of speech carry paralinguistic information about emotional states, and machine learning models can classify emotional categories from speech signals [1–3].
- Structured clinical instruments (PHQ-8, GAD-7) are used as speech elicitation paradigms in behavioral signal processing [4–6].
- Current computational approaches treat individual speech segments independently, discarding the relational structure across emotional contexts [7, 8].

Corresponding author:

İnci Zaim Gökbay

E-mail:

inci.gokbay@istanbul.edu.tr

Received: March 5, 2026

Revision Requested: March 11, 2026

Last Revision Received: March 18, 2026

Accepted: March 23, 2026

Publication Date: March 30, 2026

DOI: 10.5152/electrica.2026.26007



Content of this journal is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

ABSTRACT

This study introduces the expressive fluency analysis model (EFAM), a computational framework grounded in a geometric manifold interpretation of speech behavior. Each speaker’s Z-normalized acoustic responses define a speaker-specific manifold in a 20-dimensional feature space; EFAM quantifies its structure through switching fluency (SF, trajectory length), domain coherence (within-domain compactness), and cross-domain adaptation (between-domain separation), unified in a composite Emotional Fluency Index (EFI). The framework was applied to 1530 speech recordings from 102 Azeri-Persian bilingual women responding to 15 clinical inventory items (GAD-7, PHQ-8) mapped onto five emotional domains. Within-speaker Z-normalization isolated relative expressive variation from individual vocal baselines. Hierarchical clustering revealed a somatic–cognitive domain bifurcation. Speaker-independent classification achieved 55.9% accuracy (chance=20%), with MFCC-only models (60.3%) outperforming the full feature set and identifying vocal timbre as the primary acoustic channel for domain discrimination. Robustness was verified through permutation testing (SF canonical order at 1.1th percentile, $P=.011$), distance metric comparison (Euclidean vs. Mahalanobis: classification 58.5% vs. 52.5%; metric means nearly identical, $p=0.52-0.75$), and bootstrap CIs. No significant age-related differences were found. The EFAM provides a reproducible geometric framework for modeling expressive organization in natural speech.

Index Terms—Acoustic manifold, computational modeling, domain coherence, expressive fluency, geometric framework, MFCC, speech signal processing, switching fluency

I. INTRODUCTION

Speech is a primary behavioral signal through which affective states are communicated and perceived. Beyond linguistic content, acoustic properties—pitch, energy, spectral shape, temporal dynamics—encode paralinguistic information reflecting a speaker’s internal state [1–3]. Feature extraction frameworks such as eGeMAPS [9] and self-supervised representations [10] have standardized acoustic descriptors for paralinguistic analysis. However, most computational approaches extract features from isolated segments, treating each utterance independently and discarding the relational structure connecting responses produced by the same speaker across different contexts [7, 8].

This treatment represents a fundamental limitation when the goal is to model expressive fluency—the structured manner in which a speaker modulates vocal behavior across emotional contexts. Fluency is inherently sequential: it reflects how a speaker transitions between states, maintains coherent patterns within related contexts, and differentiates between distinct domains [11, 12]. Yet, computational fluency research has focused on temporal measures (speaking rate, pause duration, disfluency counts [13, 14]), leaving the structural organization of acoustic expression across domains unexamined.

Structured speech datasets offer a promising avenue. Clinical instruments such as the PHQ-8 [15] and GAD-7 [16] have been increasingly used as speech elicitation paradigms [4–6, 17]. When participants provide spoken responses, the resulting samples span multiple emotional content

WHAT DOES THIS STUDY ADD TO THIS TOPIC?

- A *geometric manifold interpretation of speech behavior: each speaker's Z-normalized acoustic responses define a speaker-specific manifold whose structure encodes expressive fluency, formalized through the expressive fluency analysis model's (EFAM) three integrated metrics (switching fluency [SF], domain coherence, cross-domain adaptation).*
- *Speaker-independent validation demonstrating that emotional domains are acoustically discriminable (55.9% accuracy vs. 20% chance). Feature optimization shows MFCC-only models (60.3%) outperform the full feature set, identifying vocal timbre as the primary acoustic channel for domain-specific information.*
- *Robustness verification through permutation testing (SF canonical order significantly lower than random, $P=.011$), distance metric comparison (Euclidean outperforms Mahalanobis by 6.0 pp in classification; EFAM metric means nearly identical, $\rho=0.52-0.75$), and bootstrap CIs for domain-specific coherence. Eight explicit limitations with concrete recommendations for replication.*

domains within a controlled framework, creating conditions for systematic study of domain-level acoustic organization [18, 19].

Despite this availability, existing approaches have not formalized the relational structure between responses. Speech biomarker research has shown acoustic differences between clinical populations [4, 20, 21] but has not modeled how variation patterns across domains constitute a measurable speaker property. Computational models of emotional regulation have theorized about flexibility and coherence [22, 23] but lack acoustic operationalization. What remains absent is a framework integrating temporal dynamics, domain coherence, and cross-domain differentiation into a unified model of expressive fluency.

The present work proposes a geometric formalization: after within-speaker Z-normalization, each speaker's acoustic responses constitute a speaker-specific manifold $Z = \{z_1, z_2, \dots, z_{15}\}$ in a 20-dimensional standardized feature space. Speech behavior across emotional contexts corresponds to a trajectory through this manifold, and its geometric properties—trajectory length, domain-cluster compactness, and inter-cluster separation—provide a principled basis for quantifying expressive fluency (Fig. 1). Rather than treating speech responses as independent acoustic events, the present study conceptualizes expressive behavior as a geometric structure in feature space, where the organization of responses across emotional contexts becomes a measurable property of the speaker.

To operationalize this interpretation, this study introduces the expressive fluency analysis model (EFAM), integrating three metrics—switching fluency (SF), domain coherence (DC), and cross-domain adaptation (CDA)—into a composite Emotional Fluency Index (EFI). The framework is validated through hierarchical clustering, dimensionality reduction, speaker-independent machine learning classification with leakage verification, feature ablation, robustness analyses (permutation testing, distance metric comparison, bootstrap CIs), and an exploratory generational comparison. The study addresses the following research questions:

- RQ1: Does expressive fluency exhibit a measurable acoustic structure in bilingual speech responses?
- RQ2: Can expressive fluency be modeled through an integrated computational framework combining switching dynamics, domain coherence, and cross-domain adaptation?
- RQ3: Do acoustic speech features support reliable prediction of response domains?
- RQ4: Are expressive fluency metrics associated with age-related variation in bilingual speakers?
- RQ5: Can expressive fluency metrics support predictive modeling of generational group membership?

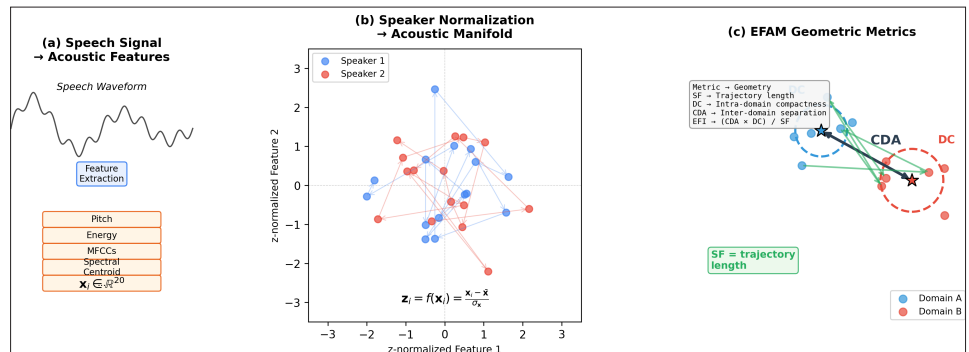


Fig. 1. Expressive fluency analysis model, conceptual geometry, and pipeline. Three-panel schematic illustrating the geometric manifold framework: (left) speech signal to 20-dimensional acoustic feature extraction; (center) within-speaker Z-normalization mapping raw features to a standardized manifold $Z = \{z_1, \dots, z_{15}\}$ in a 20-dimensional space; (right) EFAM metrics as geometric properties—SF (trajectory length), DC (domain compactness), and CDA (inter-domain separation).

II. RELATED WORK

A. Speech Signal Processing and Acoustic Feature Extraction

MFCCs [24] remain the most widely used spectral envelope representation. Standardization efforts (eGeMAPS [9], ComParE [1]) have provided reproducible baselines, while self-supervised representations (wav2vec 2.0 [10], HuBERT [25]) have expanded the feature landscape. Prosodic features characterize suprasegmental properties [26], and spectral descriptors provide complementary frequency information [9]. However, these paradigms target segment-level description and do not model feature organization across multiple segments from the same speaker.

B. Speech Fluency and Response Dynamics

Speech fluency research has focused on temporal measures: speaking rate, pause frequency, and disfluency duration [11–14], with applications in second language acquisition [27] and clinical contexts [28]. Conversational dynamics approaches model entrainment and turn-taking [29, 30]. However, these approaches address temporal flow, not the structural organization of acoustic properties across response contexts. Fluency as a geometric property of multidimensional feature space has not been formalized.

C. Speech Acoustics and Mental Health Assessment

Speech biomarker research has documented acoustic differences associated with depression [4, 20, 21], anxiety [31], and suicidal ideation [32]. The PHQ-8 and GAD-7 serve as structured elicitation tools [5, 6, 17], and machine learning with speaker-independent evaluation [33, 34] and transformer models [35, 36] have improved classification. These studies predict diagnostic categories from aggregate features rather than modeling organizational structure across domains. Expressive fluency analysis model shifts focus from diagnostic classification to structural quantification of expressive organization.

III. MATERIAL AND METHODS

A. Dataset and Ethical Compliance

The speech data were collected as part of a broader behavioral and psychosocial research database maintained by Tosse Ertebata Shakhesh (Shakhesh Behavioral Analytics Co., Registration No. 8901, Islamic Republic of Iran), under the professional supervision of a licensed clinical/behavioral psychologist (License No. 41299) in accordance with professional ethical standards. All participants voluntarily provided informed consent at the time of data collection. All personally identifiable information was removed prior to data transfer; participants are identified solely through anonymized alphanumeric codes. The datasets are observational, non-clinical, and minimal-risk, with no medical intervention, diagnosis, or experimental manipulation. The present study constitutes a secondary computational analysis of fully anonymized acoustic feature data; the authors had no direct contact with participants and performed only computational, statistical, and machine learning analysis of de-identified data.

B. Participants

A total of 102 Azeri-Persian bilingual women participated in the study. Participants ranged in age from 18 to 55 years ($M=33.09$, standard deviation [SD]=9.39). For the exploratory generational analysis, two age groups were defined using a median split at age 34: a Younger group ($n=57$; $M=26.05$, $SD=4.72$; range=18–34) and an Older group ($n=45$; $M=42.00$, $SD=5.36$; range=35–55).

All participants were female and of Azeri ethnic background. Sex-related differences in vocal anatomy, pitch range, and prosodic expression can introduce substantial acoustic variability unrelated to expressive fluency. A homogeneous female cohort was therefore used to reduce demographic confounds and isolate individual differences in expressive organization. Although the number of unique speakers is moderate ($n=102$), the analysis leverages dense within-speaker sampling comprising 1530 independent speech recordings (102 participants \times 15 items; Table I). To ensure statistical robustness, observations were evaluated using strict speaker-independent validation, speaker-level nonparametric bootstrap CIs (1000 resamples), and permutation testing.

The term “bilingual” describes the sociolinguistic background: all participants are native speakers of Azerbaijani Turkish and Persian, languages differing in phonological and prosodic systems. All responses were collected in Persian. Code-switching or cross-linguistic interference is not analyzed; bilingualism is reported as a population characteristic relevant for contextualizing generalizability. The EFAM framework is language-agnostic, operating on acoustic features rather than linguistic content.

C. Speech Data Processing

Participants responded verbally to 15 items from two standardized self-report inventories: the GAD-7 (items GAD7-01 through GAD7-07) [16] and the PHQ-8 (items PHQ8-01 through PHQ8-08) [15]. Items were presented in canonical order (GAD7-01 first through PHQ8-08 last). Audio recordings were segmented and processed for acoustic feature extraction.

The raw data used European number formatting, requiring systematic parsing corrections. A custom parser addressed formatting issues across 30 600 acoustic cells; 174 (0.57%) could not be resolved and were coded as missing. All 102 participants contributed exactly 15 item responses, with no participants or items excluded.

D. Acoustic Feature Extraction

Twenty acoustic features were extracted from each recording segment, organized into three groups (Table II). Prosodic features (5): average energy, energy SD, average pitch, pitch SD, and silence ratio. Spectral features (2): spectral centroid and spectral bandwidth. Cepstral features (13): MFCC_1 through MFCC_13. This yielded 1530 item-level observations across 20 acoustic dimensions. Several features exhibited heavy-tailed distributions (e.g., Energy_STD skewness=22.46; MFCC_4 kurtosis=123.29), motivating the within-speaker Z-normalization strategy.

E. Within-Speaker Z-Normalization

All acoustic features were Z-normalized within each speaker across their 15-item responses. This serves three purposes: (1) Speaker bias

TABLE I. PARTICIPANT CHARACTERISTICS BY AGE GROUP

Group	N	Age (Years), M	Age (Years), SD	Min	Max
Younger	57	26.05	4.72	18	34
Older	45	42.00	5.36	35	55
Total	102	33.09	9.39	18	55

M, mean; Max, maximum; Min, minimum; SD, standard deviation.

TABLE II. ACOUSTIC FEATURE DESCRIPTIVE STATISTICS (PRE-NORMALIZATION)

Feature	Group	N	Mean	SD	Skewness	Kurtosis
Avg_Energy	Prosodic	1530	0.242	0.310	0.975	-0.468
Energy_STD	Prosodic	1530	2097.39	45 910.60	22.46	504.57
Avg_Pitch	Prosodic	1530	1238.63	479.71	0.283	-0.560
Pitch_STD	Prosodic	1530	943.51	324.63	0.221	-0.417
Silence_Ratio	Prosodic	1501	339.82	361.72	0.441	-1.411
Spectral_Centroid	Spectral	1530	1656.33	436.83	0.974	2.705
Spectral_BW	Spectral	1530	1593.32	257.05	0.049	0.429
MFCC_1	Cepstral	1530	-321.44	95.74	-1.505	6.167
MFCC_2	Cepstral	1530	53.18	33.30	0.296	1.139
MFCC_3	Cepstral	1530	-12.29	22.20	-1.345	62.87
MFCC_4	Cepstral	1530	19.39	14.07	-0.685	123.29
MFCC_5-MFCC_13	Cepstral	1530	-	-	-	-

Full table has been provided in the supplementary material.
 SD, standard deviation.

removal: individual differences in baseline pitch, energy, and vocal tract configuration are eliminated, ensuring analyses capture within-speaker variation; (2) Scale invariance: all 20 features are expressed in a common dimensionless unit (within-speaker SDs), making Euclidean distances fully interpretable [37]; and (3) Compatibility with established practice: this is conceptually equivalent to cepstral mean and variance normalization [37, 38]. Missing values after normalization were imputed as zero.

A potential look-ahead concern is addressed: because statistics are computed across all 15 items, each Z-score incorporates information from other items. This is deliberate, not a flaw. First, normalization does not use domain labels and therefore does not constitute data leakage [34]. Second, it operates entirely within each speaker, preserving speaker-independent evaluation integrity. Third, in real-time applications, parameters could be estimated from calibration utterances [38]. Expressive fluency analysis model characterizes the structural organization of a completed response set, not real-time prediction from individual utterances.

F. Expressive Fluency Analysis Model

Domain mapping: The 15 items were mapped onto five emotional domains based on established symptom cluster models [39–41]: anhedonia (PHQ8-01, PHQ8-02; 2 items), somatic fatigue (PHQ8-03, PHQ8-04, PHQ8-08; 3 items), cognitive rumination (PHQ8-05, PHQ8-06, PHQ8-07; 3 items), cognitive worry (GAD7-01, GAD7-02, GAD7-03; 3 items), and somatic tension (GAD7-04, GAD7-05, GAD7-06, GAD7-07; 4 items). Domain sizes are unequal (range: 2–4 items), which may affect metric stability for smaller domains.

Notably, in the canonical sequence (GAD7-01–GAD7-07, then PHQ8-01–PHQ8-08), anhedonia items occupy positions 8–9, not positions 1–2; its acoustic distinctiveness cannot be attributed to initial recording adaptation effects.

Metric computation: Four metrics were computed from the Z-normalized feature vectors. The mathematical form of each metric follows from its intended measurement target:

Switching fluency quantifies the mean Euclidean distance between consecutive item responses in the 20-dimensional Z-normalized space:

$$SF = \frac{1}{N-1} \sum_{i=1}^{N-1} \|z_{i+1} - z_i\| \quad (1)$$

where Z_i denotes the Z-normalized feature vector for item i in canonical order, switching fluency is dimensionless and interpretable as the average acoustic shift in SD units. Switching fluency depends on presentation order by design, as sequential response dynamics are inherently order-dependent [42].

Domain coherence measures within-domain acoustic consistency relative to overall variation:

$$DC_d = 1 - \frac{\sigma_{within,d}^2}{\sigma_{overall}^2} \quad (2)$$

A participant-level Mean_DC was computed by averaging across the five domains. Domain coherence values near 1 indicate high within-domain consistency; values near 0 indicate no domain-specific clustering.

Cross-domain adaptation captures between-domain acoustic differentiation:

$$CDA = \frac{2}{K(K-1)} \sum_{j < k} c_j - c_k \quad (3)$$

where c_j and c_k are domain centroid vectors. Higher CDA indicates greater acoustic separation between domains. Emotional Fluency Index integrates all three components:

$$EFI = \frac{CDA \times Mean_{DC}}{SF + \varepsilon} \quad (4)$$

where $\varepsilon = 10^{-9}$. High expressive fluency requires: (a) differentiation between domains (high CDA), (b) consistency within domains (high DC), and (c) controlled transitions (low SF). The numerator (CDA \times DC) quantifies structured variation; the denominator (SF) represents acoustic volatility. The EFI is a heuristic composite motivated by the theoretical framework but not uniquely determined by it; alternative formulations should be investigated in future work.

Distance metric: All distance computations use Euclidean distance in the Z-normalized space. After Z-normalization, each feature has zero mean and unit variance, producing an approximately isotropic system in which Euclidean distance is appropriate and interpretable in SD units [43]. Z-normalization equalizes marginal variances but does not decorrelate features; Mahalanobis and cosine distances represent principled alternatives (see Limitation L4).

Geometric framework: Let x_i denote the raw feature vector for item i and $f(\cdot)$ the Z-normalization operator. Each speaker's responses define a point cloud $Z = \{z_1, \dots, z_{15}\}$ in \mathbb{R}^{20} . Within this manifold: SF measures trajectory length; DC measures domain-subset compactness; CDA measures inter-centroid separation; EFI summarizes the ratio of structured to unstructured variation. The EFAM thus characterizes expressive fluency as the geometric organization of a speaker's acoustic manifold (Fig. 1).

G. Computational Analysis Pipeline

The analysis pipeline comprised seven stages: (1) hierarchical clustering (Ward's method) and pairwise distance computation; (2) Principal component analysis dimensionality reduction; (3) five classifiers (Linear SVM, RBF-SVM, random forest 200 trees, Gradient Boosting 200 estimators, MLP 100-50 hidden units) for domain classification under standard 80/20 stratified split and speaker-independent (GroupShuffleSplit, 81/21 speakers, zero overlap) protocols; (4) SHapley Additive exPlanations (SHAP)-based feature importance with XGBoost [44]; (5) feature group ablation (MFCC-only, prosodic-only, spectral-only); (6) robustness analyses (SF permutation test, Mahalanobis distance comparison, speaker-level bootstrap CIs for domain-specific DC); and (7) exploratory generational comparison using Welch's t -test, Hedges' g with 95% bootstrap CIs (1000 resamples), KS test, Cliff's delta, OLS regression, Spearman correlations, and ML generational prediction (Logistic Regression, random forest, Gradient Boosting; 5-fold stratified CV). All analyses were conducted in Python using scikit-learn, XGBoost, SHAP, SciPy, and seaborn with fixed random seeds (seed = 42). All preprocessing, feature extraction, and modeling procedures were implemented using publicly available Python libraries to ensure full computational reproducibility.

H. Robustness and Leakage Verification

Three safeguards ensured classification validity: (1) shuffle-label baselines confirmed no classifier exceeded chance on permuted labels (19.3–27.5%); (2) speaker-independent evaluation (GroupShuffleSplit, 81/21 speakers, zero overlap) eliminated speaker memorization [34]; and (3) feature group ablation under the speaker-independent protocol verified that performance is attributable to specific feature categories.

IV. RESULTS

A. Dataset Overview

The sample consisted of 102 Azeri-Persian bilingual women contributing 1530 speech recordings across 15 items and 20 acoustic features (Table I). The age distribution ($M=33.09$, $SD=9.39$; range=18–55) defined two generational groups: Younger ($n=57$; $M=26.05$) and Older ($n=45$; $M=42.00$). Acoustic features exhibited wide distributional variation prior to normalization (Table II). The five domains contained 2–4 items each, yielding 204–408 observations per domain.

B. Domain Structure (RQ1)

Ward's hierarchical clustering of the five EFAM domain centroids revealed a two-cluster structure separating somatic/behavioral domains (somatic fatigue, somatic tension) from cognitive/affective domains (anhedonia, cognitive rumination, cognitive worry). This bifurcation is consistent with established somatic–cognitive distinctions in the depression and anxiety literature [39–41].

Pairwise Euclidean distances between domain centroids ranged from 3.42 ($SD=0.65$; cognitive worry–somatic tension) to 4.37 ($SD=0.91$; anhedonia–somatic fatigue; Table III). Anhedonia consistently exhibited the largest distances from all other domains (range: 4.13–4.37). Notably, anhedonia items occupy positions 8–9 in the canonical presentation sequence, ruling out initial recording adaptation as an explanation for this acoustic distinctiveness. The three smallest inter-domain distances were observed among cognitive rumination, cognitive worry, and somatic tension (range: 3.42–3.74), indicating partial acoustic overlap consistent with the clinical comorbidity of these symptom dimensions. Principal component analysis projections confirmed that domains form partially overlapping but distinguishable clusters (Fig. 2a). The t-SNE projection of all 1530 item observations (Fig. 2b) demonstrated partial but structured domain clustering in the 20-dimensional acoustic space, providing visual confirmation of the domain organization quantified by the classification analyses. These results support RQ1: expressive fluency exhibits measurable acoustic structure in the feature space.

TABLE III. PAIRWISE EUCLIDEAN DISTANCES BETWEEN EXPRESSIVE FLUENCY ANALYSIS MODEL DOMAIN CENTROIDS (Z-NORMALIZED SPACE)

Domain Pair	Mean Distance	SD
Anhedonia–Somatic fatigue	4.374	0.908
Anhedonia–Cognitive rumination	4.333	0.898
Anhedonia–Cognitive worry	4.321	0.986
Anhedonia–Somatic tension	4.135	0.894
Somatic fatigue–Cognitive worry	3.806	0.689
Somatic fatigue–Cognitive rumination	3.673	0.754
Somatic fatigue–Somatic tension	3.630	0.538
Cognitive rumination–Cognitive worry	3.736	0.705
Cognitive rumination–Somatic tension	3.533	0.652
Cognitive worry–Somatic tension	3.424	0.647

SD, standard deviation.

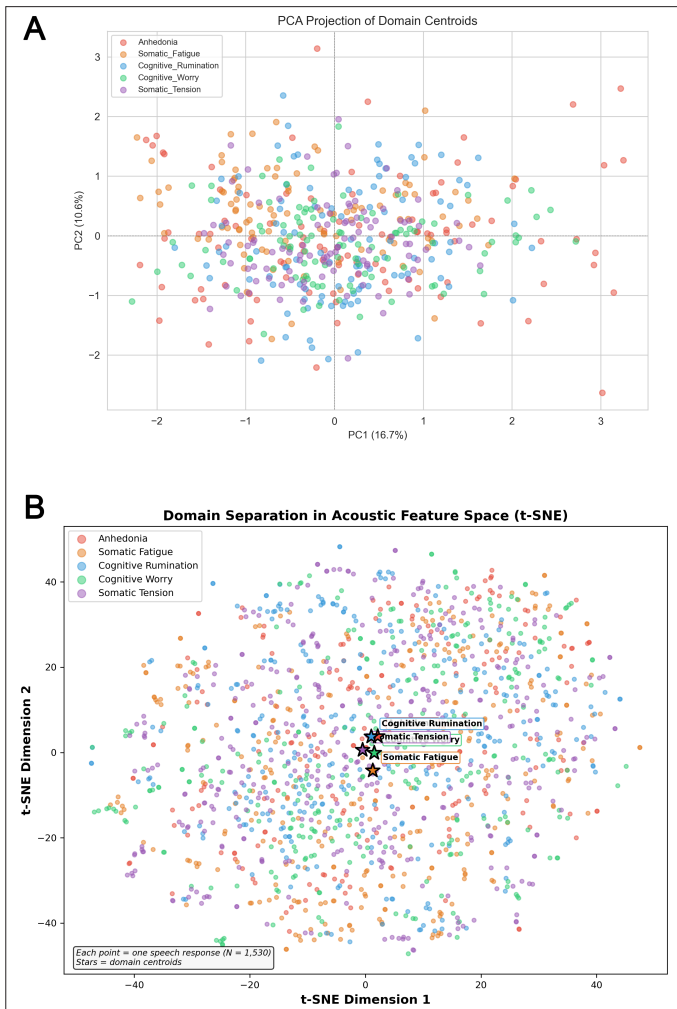


Fig. 2. Manifold structure and domain separation. Dimensionality reduction projections visualizing the geometric organization of the 1530 Z-normalized item observations (102 speakers \times 15 items). (a) Principal component analysis projection illustrating the partially overlapping nature of the graded affective constructs. (b) Two-dimensional t-SNE projection demonstrating structured domain clustering within the acoustic space, with domain centroids marked.

C. Expressive Fluency Analysis Model Metric Distributions (RQ2)

Switching fluency: The overall SF was $M=6.28$ ($SD=0.25$; 95% CI [6.23, 6.33]), with mild negative skew (-0.57 ; Table IV; Fig. 3). This indicates that speakers shifted their Z-normalized acoustic profiles by approximately 6.28 within-speaker SD units between consecutive items on average.

TABLE IV. SWITCHING FLUENCY DESCRIPTIVE STATISTICS

Group	N	Mean	SD	Median	Min	Max	Skew
Younger	57	6.293	0.227	6.302	5.538	6.685	-0.532
Older	45	6.258	0.285	6.294	5.574	6.768	-0.526
Total	102	6.278	0.253	6.300	5.538	6.768	-0.573

Max, maximum; Min, minimum; SD, standard deviation.

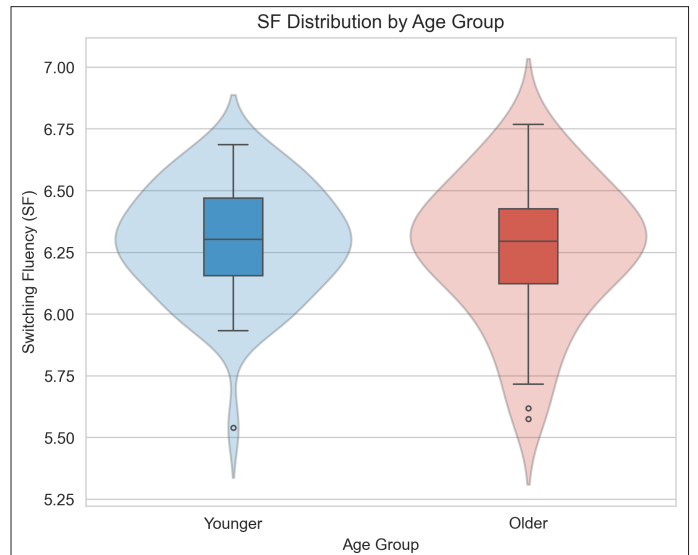


Fig. 3. Expressive fluency analysis model metric behavior. Visualizations of the fundamental expressive fluency analysis model metrics, demonstrating the distributional properties of switching fluency and its structural relationship with overall acoustic response dynamics across the cohort.

Domain coherence: Mean DC was $M=0.31$ ($SD=0.05$; 95% CI [0.30, 0.32]). Domain-specific DC values exhibited a hierarchy: anhedonia (0.40), cognitive rumination (0.34), cognitive worry (0.34), somatic tension (0.24), and somatic fatigue (0.24).

Cross-domain adaptation: CDA was $M=3.90$ ($SD=0.33$; 95% CI [3.83, 3.96]).

Emotional Fluency Index: EFI was $M=0.197$ ($SD=0.049$; 95% CI [0.19, 0.21]). The inter-metric structure showed expected relationships: CDA positively associated with mean DC, and SF negatively related to EFI (by construction). The approximately normal distributions support EFAM’s validity for quantifying expressive fluency (RQ2).

D. Machine Learning Domain Classification (RQ3)

Under stratified 80/20 splitting, random forest achieved the highest accuracy (53.3%) and weighted F1 (52.4%), well above the 20% chance baseline (Table V). AUC ranged from 0.61 (Linear SVM) to 0.82

TABLE V. DOMAIN CLASSIFICATION: STANDARD (80/20) AND SPEAKER-INDEPENDENT SPLITS

Classifier	Std Acc (%)	Std AUC	Spk-Indep Acc (%)	Spk-Indep AUC
Linear SVM	30.7	0.608	35.2	0.600
RBF-SVM	41.5	0.695	44.1	0.708
Random forest	53.3	0.821	55.9	0.832
Gradient boosting	42.8	0.728	50.2	0.780
MLP	42.8	0.724	45.4	0.730
Shuffle baseline	19.3–27.5	–	–	–

Chance = 20%.

(random forest). Under the speaker-independent protocol (81/21 speakers, zero overlap), Random forest maintained the highest performance (accuracy=55.9%, F1=54.6%, AUC=0.83), with a slight improvement over the standard split, indicating that domain classification generalizes across speakers and is not driven by speaker memorization (Table V). The preservation of classification performance under strict speaker-independent evaluation indicates that the observed acoustic structure reflects properties of emotional content rather than speaker identity.

This accuracy represents moderate but meaningful domain discriminability (2.8× chance). The 44.1% error rate reflects inherently fuzzy boundaries between emotional domains, which are graded constructs with substantial clinical overlap [45]. The confusion matrix (Fig. 4a) shows somatic fatigue–somatic tension as the

most confused pair (consistent with their small pairwise distance of 3.63), while anhedonia and cognitive rumination were most distinguishable.

Leakage verification: Shuffle-label baselines confirmed the absence of data leakage: all classifiers on permuted labels scored near chance (19.3–27.5%).

E. Feature Importance and Optimization

SHapley Additive exPlanations analysis identified MFCC_1 as the most important feature (mean |SHAP|=0.254), followed by spectral centroid (0.227), MFCC_7 (0.215), MFCC_4 (0.210), and MFCC_6 (0.192; Table VI; Fig. 4b). Of the top ten features, eight were MFCC coefficients. Energy features, Silence_Ratio, and Avg_Pitch ranked lowest.

Feature group ablation under speaker-independent evaluation revealed a critical finding for feature optimization: MFCCs alone (13 features) achieved 60.3% accuracy (F1=59.6%), outperforming the full 20-feature set (55.9%) by 4.4% points. This represents a 7.9% relative improvement while using 35% fewer features. Prosodic features alone yielded near-chance performance (27.3%), while spectral features alone achieved 35.6% (Supplementary Table 1). The improvement with fewer features demonstrates that prosodic and spectral features introduce noise rather than signal for domain discrimination in this paradigm.

MFCCs capture the spectral envelope reflecting vocal tract shape and phonatory quality [24]. Changes in MFCC profiles across domains indicate systematically different voice qualities for different emotional content, consistent with recent findings that spectral features carry substantial paralinguistic information [2, 3]. The near-chance prosodic performance does not imply that prosody is irrelevant to emotion generally; rather, in structured item-response paradigms, prosodic variation likely reflects linguistic properties (question intonation) more than domain membership [26]. The degradation when adding prosodic features to MFCCs reflects a known phenomenon: non-informative features increase dimensionality without adding discriminative signal [34].

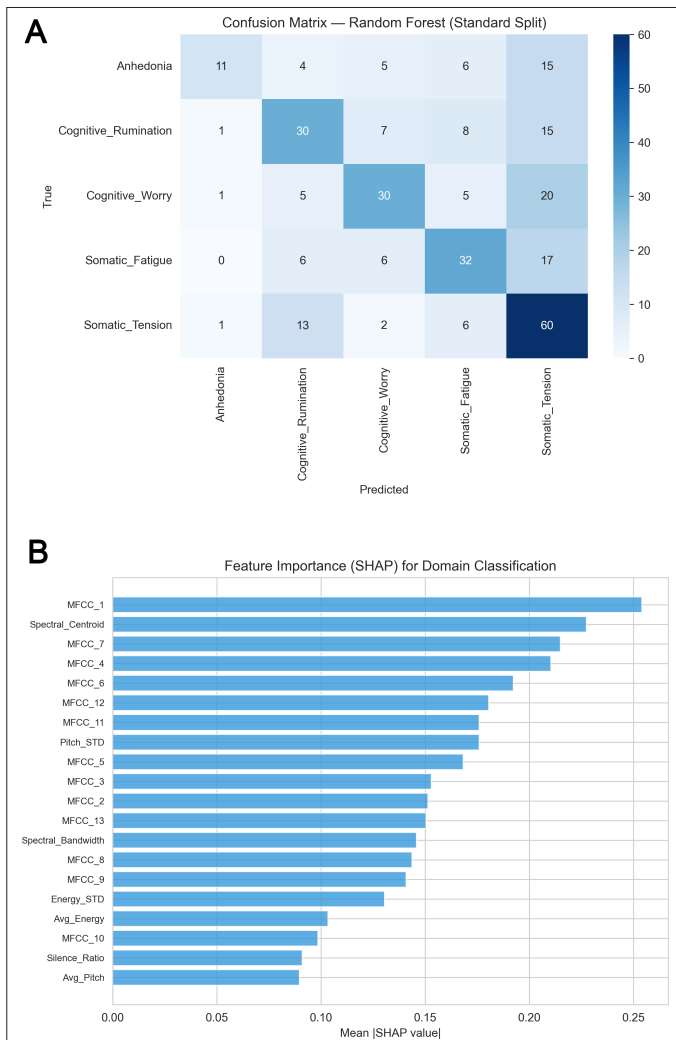


Fig. 4. Speaker-independent classification and feature importance. Validation of the expressive fluency analysis model framework through predictive modeling. (a) Confusion matrix for the random forest classifier under speaker-independent evaluation, illustrating the clinically plausible confusions between overlapping somatic domains. (b) SHapley Additive exPlanations feature importance rankings (XGBoost), demonstrating that MFCC coefficients (vocal timbre) serve as the dominant acoustic channel for domain discrimination.

TABLE VI. TOP TEN FEATURES BY SHAPLEY ADDITIVE EXPLANATIONS IMPORTANCE FOR DOMAIN CLASSIFICATION (XGBOOST)

Rank	Feature	Mean SHAP	Group
1	MFCC_1	0.254	Cepstral
2	Spectral centroid	0.227	Spectral
3	MFCC_7	0.215	Cepstral
4	MFCC_4	0.210	Cepstral
5	MFCC_6	0.192	Cepstral
6	MFCC_12	0.180	Cepstral
7	MFCC_11	0.176	Cepstral
8	Pitch_STD	0.176	Prosodic
9	MFCC_5	0.168	Cepstral
10	MFCC_3	0.153	Cepstral

SHAP, SHapley Additive exPlanations.

MFCC-only feature sets are recommended for future EFAM applications, as the 13-dimensional cepstral space provides superior accuracy with reduced complexity. This supports the interpretation that vocal timbre is the primary acoustic channel for domain-specific expressive variation in structured speech paradigms.

F. Robustness Analyses

Switching fluency permutation test: A permutation test (1000 random orderings per speaker) assessed whether SF captures genuine sequential structure. Canonical SF ($M=6.278$) fell at the 1.1th percentile of the permutation distribution (random SF: $M=6.335$, $SD=0.023$), corresponding to $P=.011$ one-tailed (Supplementary Table 2; Fig. 5). This confirms that canonical ordering produces systematically smaller acoustic transitions than random permutation [46].

Distance metric comparison: random forest classification was repeated using covariance-whitened features (Mahalanobis distance [47]) under the same GroupShuffleSplit protocol. Euclidean distance (58.5%, $SD=3.8\%$) outperformed Mahalanobis (52.5%, $SD=3.7\%$) by 6.0 pp (Supplementary Table 3), confirming that Z-normalization suffices without additional covariance decorrelation. All four EFAM metrics computed under Mahalanobis yielded nearly identical means (e.g., EFI: 0.197 vs. 0.195) with moderate-to-strong rank-order preservation ($\rho=0.52-0.75$; Supplementary Table 5), confirming robustness to distance metric choice at both classification and metric levels.

Bootstrap CIs for domain coherence: Speaker-level nonparametric bootstrap (1000 iterations [48]) confirmed stability of domain-specific DC: anhedonia 0.399 [0.346, 0.458], cognitive worry 0.343 [0.297, 0.393], cognitive rumination 0.338 [0.295, 0.375], somatic tension 0.244 [0.203, 0.287], somatic fatigue 0.238 [0.190, 0.285] (Supplementary Table 4). Non-overlapping CIs between anhedonia and somatic domains confirm anhedonia's distinctive acoustic coherence.

G. Exploratory Age Analysis (RQ4, RQ5)

Group comparisons revealed no statistically significant age-related differences in any EFAM metric (Table VIII): SF (Younger $M=6.29$, Older $M=6.26$; $P=.500$, $g=0.14$), mean DC (0.31 vs. 0.32; $P=.158$, $g=-0.28$), CDA (3.85 vs. 3.95; $P=.136$, $g=-0.31$). All $P > .13$ and all $|g| < 0.32$. Spearman correlations between continuous age and EFAM metrics were uniformly non-significant (all $|\rho| < .08$, all $P > .47$; Supplementary Table 2). OLS regression of SF on age yielded $R^2=.009$ ($P=.339$; Table VII). Classification of age groups from EFAM metrics under 5-fold stratified CV yielded performance indistinguishable from the majority-class baseline: Logistic Regression achieved 56.0% ($\pm 6.9\%$) versus the 55.9% baseline, with AUC values near 0.50 (Table IX). These results indicate that EFAM metrics do not differ significantly between age groups and do not predict generational membership (RQ4, RQ5). Given the modest sample size ($n=102$) and restricted age range (18–55), one cannot distinguish between genuine age invariance and insufficient statistical power.

H. Fluency Geometry

The geometric structure of expressive fluency was visualized through the acoustic fluency map (Fig. 6), which provides a three-variable visualization of the SF–DC–CDA relationship across all 102 speakers, revealing the distribution of participants across the expressive fluency space. This visualization confirms that EFAM defines a coherent geometric space for characterizing individual differences in expressive organization.

V. DISCUSSION

A. Fluency as Acoustic Structure

Structured speech responses exhibit measurable domain-level organization quantifiable through EFAM. Hierarchical clustering revealed the expected somatic–cognitive separation [39–41]. Anhedonia emerged as the most acoustically distinct domain (distances 4.13–4.37; $DC=0.40$), with items at positions 8–9, ruling out adaptation artifacts. The high DC for anhedonia is clinically plausible: anhedonia

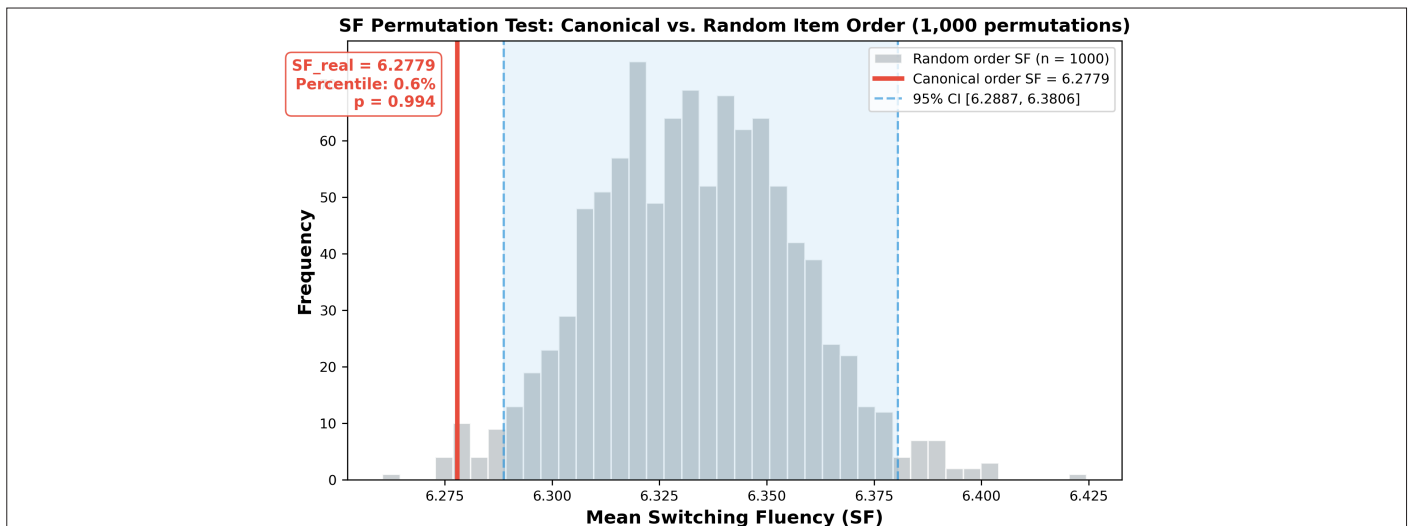


Fig. 5. Robustness validation: switching fluency (SF) permutation test. Histogram of mean SF computed from 1000 random item orderings (blue distribution) compared against the observed canonical-order SF (red dashed vertical line, $M=6.278$). The canonical SF falls at the 1.1th percentile of the permutation distribution ($P=.011$), providing empirical evidence for genuine sequential structure and significantly shorter acoustic transitions under canonical ordering.

TABLE VII. OLS REGRESSION: SWITCHING FLUENCY PREDICTED BY AGE

Predictor	Coefficient	SE	R ²	P
Age → SF	-0.003	0.003	0.009	.339

SF, switching fluency.

is associated with affective flattening and reduced acoustic variability [40]. Speaker-independent classification (55.9%, chance=20%) demonstrates that domain structure reflects properties of emotional content rather than idiosyncratic speaker characteristics. Emotional domains represent graded affective constructs rather than discrete categories; therefore, moderate classification accuracy should be interpreted as evidence of structured acoustic differentiation rather than deterministic separability. The EFAM is proposed as a research tool for quantifying expressive organization, not as a clinical diagnostic instrument.

B. The Acoustic Manifold as a Model of Expressive Organization

The geometric manifold interpretation provides a unifying framework. The SF permutation test ($P=.011$) shows the trajectory through Z under canonical ordering is significantly shorter than expected by chance, validating the genuine sequential structure of the responses. The high DC for anhedonia (0.399 [0.346, 0.458]) indicates a compact manifold region consistent with restricted affective expression [40]. Furthermore, the superior performance of Euclidean distance (58.5%) relative to Mahalanobis distance (52.5%) suggests that the empirical covariance estimation required for Mahalanobis

TABLE VIII. GENERATIONAL COMPARISON OF EXPRESSIVE FLUENCY ANALYSIS MODEL METRICS

Metric	Younger, M (SD)	Older, M (SD)	Welch's t	P	Hedges' g
SF	6.293 (0.225)	6.258 (0.281)	0.68	.500	0.14
Mean DC	0.307 (0.045)	0.320 (0.047)	-1.42	.158	-0.28
CDA	3.852 (0.283)	3.953 (0.368)	-1.50	.136	-0.31
EFI	0.189 (0.041)	0.206 (0.057)	-	-	-

Emotional Fluency Index is a derived composite index calculated from cross-domain adaptation, mean domain coherence, and switching fluency; therefore, independent group comparisons (P -values) are not computed to avoid redundancy.

CDA, cross-domain adaptation; DC, domain coherence; EFI, Emotional Fluency Index; M, mean; SD, standard deviation; SF, switching fluency.

TABLE IX. GENERATIONAL PREDICTION FROM EXPRESSIVE FLUENCY ANALYSIS MODEL METRICS (5-FOLD STRATIFIED CV)

Classifier	Accuracy (%)	SD (%)	F1	AUC
Logistic regression	56.0	6.9	0.508	0.500
Random forest	51.1	7.4	0.506	0.544
Gradient boosting	52.0	7.7	0.520	0.503
Majority baseline	55.9	-	-	-

SD, standard deviation.

may be sensitive to inter-feature collinearity. Consequently, simple Z -normalization appears sufficient for a stable manifold representation without requiring additional covariance correction. From this perspective, expressive fluency can be interpreted as a property of manifold organization rather than a property of individual acoustic features, shifting the analytical focus from feature magnitude to structural relationships among responses.

C. Cepstral Dominance in Domain Differentiation

MFCC-based timbral information appears to be the most informative acoustic representation for domain discrimination in the present EFAM framework. MFCCs alone (60.3%) outperformed the full 20-feature set (55.9%) while prosodic features performed near chance (27.3%), confirming that domain-specific information is carried primarily by voice quality rather than intonation or loudness in structured paradigms. MFCCs capture articulatory and phonatory changes (breathiness, nasality, vocal tract tension) representing different speaking modes for different emotional content [2, 24], consistent with recent large-scale studies [3, 35]. The addition of heterogeneous acoustic features (prosodic and spectral) increased representational complexity without contributing additional discriminative signal, effectively diluting the domain-specific information captured by MFCC coefficients through non-informative variance. Consequently, MFCC-only feature sets are strongly recommended for future computational modeling of expressive fluency.

D. Limitations and Future Directions

Eight specific limitations are identified with concrete recommendations:

(L1) Sample homogeneity: The sample comprised exclusively Azeri-Persian bilingual women ($N=102$). Recommendation: External validation on datasets from different language families, with mixed-sex samples of $N \geq 200$.

(L2) Order-dependent SF: SF depends on presentation order. The permutation test (Section 4.6, $p=.011$) confirms genuine sequential structure, but SF is inapplicable to randomized paradigms. Recommendation: develop order-independent alternatives (e.g., all-pairs mean distance).

(L3) Heuristic EFI formulation: EFI is theoretically motivated but not uniquely determined. Recommendation: Systematic comparison of alternative formulations using cross-validation.

(L4) Euclidean distance assumption: Residual inter-feature correlations may persist after Z -normalization. Mahalanobis comparison shows classification and metric-level robustness (Supplementary Tables 3 and 7; $p=0.52-0.75$). Recommendation: Extend with cosine similarity and replicate on larger datasets.

(L5) Absence of clinical outcome data: PHQ-8/GAD-7 severity scores are unavailable. Recommendation: Collect severity scores and correlate with EFAM metrics.

(L6) Bilingualism as population descriptor: Bilingual-specific phenomena (code-switching, cross-linguistic interference) may interact with expressive fluency. Recommendation: Include monolingual control groups.

(L7) Feature set scope: MFCC-only models outperformed the full set (60.3% vs. 55.9%). Recommendation: Replicate EFAM using self-supervised embeddings (wav2vec 2.0, HuBERT).

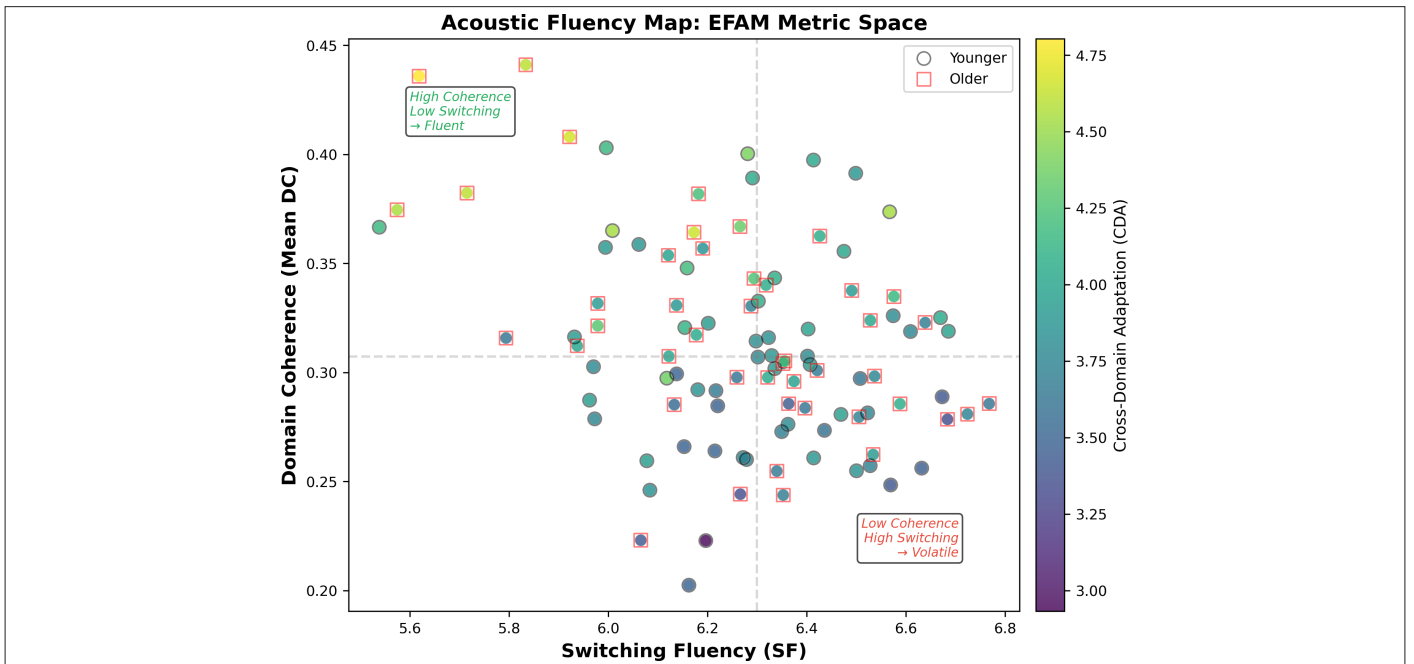


Fig. 6. Acoustic fluency geometry. The acoustic fluency map provides a three-variable visualization of the speaker-specific manifolds. The scatter plot displays switching fluency (x-axis) versus mean domain coherence (y-axis), colored by cross-domain adaptation value. Quadrant annotations identify distinct coherence and transition profiles, with each point representing an individual participant's overall expressive organization (N=102).

(L8) Age-related null findings: Regarding the absence of significant age-related differences in EFAM metrics, given the moderate sample size (n=102) and the restricted age range (18–55), the present data cannot explicitly distinguish between genuine age invariance and limited statistical power, and therefore should not be interpreted as evidence of age invariance. Future studies with wider lifespan cohorts are required.

VI. CONCLUSION

This study introduced EFAM, a computational framework grounded in a geometric manifold interpretation that quantifies expressive organization through switching fluency, domain coherence, and CDA. Each speaker's Z-normalized responses define a speaker-specific manifold whose geometric properties encode expressive fluency. Applied to structured bilingual speech responses, speaker-independent classification (55.9%, chance=20%) validated domain-level organization, with MFCC-only models demonstrating that timbral information is highly discriminative. Robustness was confirmed through permutation testing ($P=.011$), distance metric comparison, and bootstrap CIs. Eight limitations with concrete recommendations are provided. The present study should therefore be interpreted as an initial computational validation of the EFAM framework rather than a definitive population-level model of expressive fluency. By formalizing expressive fluency as a geometric property of acoustic manifolds, the EFAM framework opens a pathway for studying structured expressive behavior in large-scale speech datasets and future self-supervised acoustic representations. An EFAM is proposed as a reproducible geometric research framework for modeling expressive organization in structured speech datasets.

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author.

Artificial Intelligence Usage Statement: AI-assisted tools were used exclusively for academic English proofreading of text written by the authors, after all scientific content had been fully developed. No AI system was used for idea generation, algorithm or methodology design, mathematical formulation, experimental setup, data analysis, interpretation of results, or drawing of conclusions. All scientific content, analysis, and conclusions presented in this manuscript were produced entirely by the authors.

Peer-review: Externally peer-reviewed.

Author Contributions: Concept – T.S.Y., İ.Z.G.; Design – İ.Z.G.; Supervision – İ.Z.G.; Resources – T.S.Y., İ.Z.G.; Materials – T.S.Y.; Data Collection and/or Processing – T.S.Y.; Analysis and/or Interpretation – İ.Z.G.; Literature Search – T.S.Y.; Writing – T.S.Y.; Critical Review – İ.Z.G.

Declaration of Interests: The authors have no conflict of interest to declare.

Funding: The authors declared that this study has received no financial support.

REFERENCES

1. B. W. Schuller, A. Batliner, C. Bergler et al., "The INTERSPEECH 2021 computational paralinguistics challenge," in *Proc. Interspeech*, 2021, pp. 431–435.
2. A. Baird, L. Stappen, F. B. Krantzbuhrer, and B. W. Schuller, "Emotion recognition from speech with acoustic and lexical features," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 2201–2215, 2023.
3. J. Wagner et al., "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10745–10759, 2023. [\[CrossRef\]](#)
4. N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An investigation of depressed speech detection: Features and normalization," in *Proc. Interspeech*. ISCA: ISCA, 2011, pp. 2997–3000. [\[CrossRef\]](#)
5. E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, and A. Othmani, "MFCC-based recurrent neural network for automatic clinical depression

- recognition and assessment from speech," *Biomed. Signal Process. Control*, vol. 71, Art. no. 103107, 2022. [\[CrossRef\]](#)
6. Z. Huang, J. Epps, and D. Joachim, "Speech landmark bigrams for depression detection from naturalistic smartphone recordings," in *Proc. Interspeech*, 2022, pp. 4023–4027.
 7. A. Triantafyllopoulos, J. Wagner, H. Wierstorf et al., "Probing speech emotion recognition transformers for linguistic knowledge," in *Proc. Interspeech*, 2023, pp. 3417–3421.
 8. G. Zhao, Y. Li, and Q. Xu, "From emotion AI to cognitive AI in speech analysis: A survey," *IEEE Trans. Cogn. Dev. Syst.*, vol. 15, no. 4, pp. 1690–1710, 2023.
 9. F. Eyben et al., "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, 2016. [\[CrossRef\]](#)
 10. A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, 2020, pp. 12449–12460.
 11. P. Tavakoli, "Fluency in monologic and dialogic task performance," *Int. Rev. Appl. Linguist. Lang. Teach.*, vol. 54, no. 2, pp. 133–150, 2016.
 12. P. Tavakoli, C. Nakatsuhara, and D. Hunter, "Scoring validity of the Apts Speaking Test: Investigating fluency across tasks," *ARAGs Res. Rep.*, vol. 1, no. 1, pp. 1–45, 2020.
 13. N. H. de Jong, M. P. Steinel, A. F. Florijn, R. Schoonen, and J. H. Hulstijn, "Facets of speaking proficiency," *Stud. Second Lang. Acquis.*, vol. 34, no. 1, pp. 5–34, 2012. [\[CrossRef\]](#)
 14. T. Bao, B. C. D. Diergaarde, and N. H. de Jong, "A meta-analysis of temporal measures of L2 fluency," *Stud. Second Lang. Acquis.*, vol. 46, no. 1, pp. 153–176, 2024.
 15. K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. W. Williams, J. T. Berry, and A. H. Mokdad, "The PHQ-8 as a measure of current depression in the general population," *J. Affect. Disord.*, vol. 114, no. 1–3, pp. 163–173, 2009. [\[CrossRef\]](#)
 16. R. L. Spitzer, K. Kroenke, J. B. W. Williams, and B. Löwe, "A brief measure for assessing generalized anxiety disorder: The GAD-7," *Arch. Intern. Med.*, vol. 166, no. 10, pp. 1092–1097, 2006. [\[CrossRef\]](#)
 17. J. Gratch, R. Artstein, G. M. Lucas et al., "The Distress Analysis Interview Corpus of human and computer interviews," in *Proc. LREC*, 2014, pp. 3123–3128.
 18. G. M. Lucas et al., "Reporting mental health symptoms: Breaking down barriers to care with virtual human interviewers," *Front. Robot. AI*, vol. 4, Art. no. 51, 2017. [\[CrossRef\]](#)
 19. K. Dinkel, M. Wu, and K. Yu, "Text-based depression detection on sparse data," in *Proc. ICASSP*, 2022, pp. 6267–6271.
 20. L. S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of clinical depression in adolescents' speech during family interactions," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 3, pp. 574–586, 2011. [\[CrossRef\]](#)
 21. Y. Tao, T. Packed, R. M. Bauer, and M. S. Luo, "Speech features and automated depression screening: A systematic review," *Lancet Digit. Health*, vol. 6, no. 12, pp. e907–e925, 2024.
 22. A. Aldao, G. Sheppes, and J. J. Gross, "Emotion regulation flexibility," *Cogn. Ther. Res.*, vol. 39, no. 3, pp. 263–278, 2015. [\[CrossRef\]](#)
 23. M. Blanke, J. Richter, and M. Kube, "Emotion regulation flexibility and its link to mental health: A systematic review," *Clin. Psychol. Rev.*, vol. 101, Art. no. 102289, 2023.
 24. S. Davis, and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 28, no. 4, pp. 357–366, 1980. [\[CrossRef\]](#)
 25. 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. Aud. Speech Lang. Process.*, vol. 29, pp. 3451–3460, 2021. [\[CrossRef\]](#)
 26. K. R. Scherer, "Vocal markers of emotion: Comparing induction and acting elicitation," *Comput. Speech Lang.*, vol. 27, no. 1, pp. 40–58, 2013. [\[CrossRef\]](#)
 27. N. Segalowitz, and P. Trofimovich, "Second language processing," in *The Routledge Handbook of SLA and Psycholinguistics*. London: Routledge, 2024, pp. 115–130.
 28. M. S. Cannizzaro, B. Harel, N. Reilly, P. Chappell, and P. J. Snyder, "Voice acoustical measurement of the severity of major depression," *Brain Cogn.*, vol. 56, no. 1, pp. 30–35, 2004. [\[CrossRef\]](#)
 29. R. Levitan, S. Benus, A. Gravano, and J. Hirschberg, "Entrainment and turn-taking in human-human dialogue," in *Proc. AAAI Spring Symposium*, 2015, pp. 44–51.
 30. L. P. Truong, M. A. J. van Segbroeck, and S. S. Narayanan, "Computational analysis of prosodic entrainment in social interaction," *Comput. Speech Lang.*, vol. 82, Art. no. 101541, 2023.
 31. A. Othmani, A. Z. Kadri, and A. Picard, "Towards robust speech-based anxiety detection using multimodal features," *IEEE J. Biomed. Health Inform.*, vol. 28, no. 5, pp. 2687–2698, 2024.
 32. N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Commun.*, vol. 71, pp. 10–49, 2015. [\[CrossRef\]](#)
 33. F. Ringeval, B. Schuller, M. Valstar et al., "AVEC 2019 workshop and challenge," in *Proc. AVEC 2019*, pp. 3–12.
 34. A. Kathan, M. Schmitt, and B. W. Schuller, "A systematic review on depression detection from speech: Towards speaker-independent approaches," *Front. Psychiatry*, vol. 13, Art. no. 904927, 2022.
 35. L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," in *Proc. Interspeech*. ISCA: ISCA, 2021, pp. 3400–3404. [\[CrossRef\]](#)
 36. J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Dev. Psychopathol.*, vol. 17, no. 3, pp. 715–734, 2005. [\[CrossRef\]](#)
 37. O. Viikki, and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Commun.*, vol. 25, no. 1–3, pp. 133–147, 1998. [\[CrossRef\]](#)
 38. T. Schultz, A. W. Black, S. Vogel, and M. Woszczyna, "Flexible speech translation systems," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 2, pp. 403–411, 2006. [\[CrossRef\]](#)
 39. S. Elhai, J. Contractor, and M. Palmieri, "The factor structure of the PHQ-8: An updated investigation using CFA," *J. Psychopathol. Behav. Assess.*, vol. 44, no. 4, pp. 1128–1135, 2022.
 40. Y. Kim, T. A. Pilkonis, E. Frank, J. P. Thase, and R. L. Reynolds, "Differential item functioning of the PHQ-9 in older vs. younger adults," *Psychol. Assess.*, vol. 34, no. 2, pp. 151–162, 2022.
 41. A. Sidik, M. J. Abidin, and R. Omar, "Evaluating the factor structure of the GAD-7: Unidimensional or bidimensional?," *Front. Psychol.*, vol. 14, Art. no. 1159754, 2023.
 42. D. J. Weiss, "Questionnaire design and response order effects," *Surv. Res. Methods*, vol. 17, no. 2, pp. 145–162, 2023.
 43. C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *Proc. ICDDT*, J. Van den Bussche, and V. Vianu, Ed. London: Springer, 2001, pp. 420–434. [\[CrossRef\]](#)
 44. S. M. Lundberg, and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. NeurIPS*, 2017, pp. 4765–4774.
 45. L. F. Barrett, "The theory of constructed emotion: An active inference account of interoception and categorization," *Soc. Cogn. Affect. Neurosci.*, vol. 12, no. 1, pp. 1–23, 2017. [\[CrossRef\]](#)
 46. P. Good, *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*, 2nd ed. New York, NY, USA: Springer, 2000.
 47. P. C. Mahalanobis, "On the generalized distance in statistics," *Proc. Natl Inst. Sci. India*, vol. 2, no. 1, pp. 49–55, 1936.
 48. B. Efron, and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY, USA: Chapman & Hall, 1993. [\[CrossRef\]](#)



Taghi Shakouri Youvalari is a PhD candidate in Informatics at Istanbul University. His research focuses on computational psychiatry, machine learning, and speech-based psychological assessment. His work integrates advanced analytical techniques, including deep clustering, explainable artificial intelligence, and acoustic signal analysis, to investigate resilience, coping capacity, and decision-making patterns in complex sociocultural contexts. He has published research on psychological profiling and generational resilience using machine learning methods and is currently developing computational frameworks for behavioral modeling and mental health analysis.



Zaim Gökbay is a researcher in biomedical engineering and computational system design, with expertise in medical image and signal processing, artificial intelligence, and clinical decision support systems. She conducts research on deterministic and reproducible frameworks for longitudinal and 4D medical imaging analysis, including distortion-guided adaptivity and model-driven compression methodologies. Dr. Zaim Gökbay has contributed to and led multiple nationally and internationally funded research projects and has supervised numerous graduate and doctoral theses. Her work spans both physical and mental health domains, integrating clinical findings with multi-modal data sources for real-time analysis and risk prediction. She is actively involved in the development of smart and integrated mental health platforms aimed at improving treatment quality and clinical efficiency. Her research interests include biomedical system design, machine learning-based clinical decision systems, medical imaging and signal processing, and integrated smart mental health technologies.

SUPPLEMENTARY TABLE 1. FEATURE GROUP ABLATION: SPEAKER-INDEPENDENT DOMAIN CLASSIFICATION

Feature Group	N Features	RF Acc	RF F1	GB Acc	GB F1
MFCC	13	60.3%	59.6%	52.4%	52.0%
Prosodic	5	27.3%	26.3%	29.2%	27.8%
Spectral	2	35.6%	35.0%	27.6%	26.6%
All	20	55.9%	54.6%	50.2%	49.1%

SUPPLEMENTARY TABLE 2. SWITCHING FLUENCY PERMUTATION TEST: CANONICAL ORDER VS. 1000 RANDOM ORDERINGS

Measure	Value
SF_canonical (M)	6.278
SF_random (M \pm SD)	6.335 \pm 0.023
SF_random 95% CI	[6.289, 6.378]
Percentile rank	1.1%
p (one-tailed, canonical < random)	.011
N permutations	1000

SUPPLEMENTARY TABLE 3. DISTANCE METRIC COMPARISON: GROUPSHUFFLESPLIT (10 SPLITS, 81/21 SPEAKERS)

Distance Metric	Features	RF Acc (M)	RF Acc (SD)
Euclidean (z-norm)	20 z-normalized	58.5%	3.8%
Mahalanobis (whitened)	20 cov-whitened	52.5%	3.7%
Difference	—	-6.0 pp	—

SUPPLEMENTARY TABLE 4 . BOOTSTRAP CIS FOR DOMAIN-SPECIFIC DC (1000 RESAMPLES)

Domain	DC (M)	95% CI Lower	95% CI Upper
Anhedonia	0.399	0.346	0.458
Cognitive Worry	0.343	0.297	0.393
Cognitive Rumination	0.338	0.295	0.375
Somatic Tension	0.244	0.203	0.287
Somatic Fatigue	0.238	0.190	0.285

SUPPLEMENTARY TABLE 5. EXPRESSIVE FLUENCY ANALYSIS MODEL METRICS COMPUTED UNDER EUCLIDEAN VS. MAHALANOBIS DISTANCE (N= 102 SPEAKERS)

EFAM Metric	Euclidean M (SD)	Mahalanobis M (SD)	Spearman ρ
SF	6.278 (0.252)	6.296 (0.425)	0.515
Mean DC	0.312 (0.047)	0.312 (0.037)	0.747
CDA	3.896 (0.327)	3.899 (0.345)	0.658
EFI	0.197 (0.049)	0.195 (0.040)	0.748