Disentangled Speech Feature Analysis for Identifying Cognitive Decline Through Emotion and Prosody Variations

Logan Bowers
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
logan.bowers@buffalo.edu

Pooja Sami
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
pooja.sami@colostate.edu

Abstract

Speech provides a low-burden window into cognitive-linguistic and affective functioning, yet acoustic markers of cognitive decline are confounded by speaker identity, recording conditions, emotional state, and normal aging. The central problem addressed here is that emotion and prosody vary together in speech, so classifiers trained on raw acoustic descriptors can learn nuisance-driven rather than cognitive-relevant structure. This manuscript develops a methodological framework for disentangled speech feature analysis that separates content, prosodic, emotional, and nuisance factors before estimating cognitive-decline readouts. The framework combines a factorized latent encoder, a reconstruction objective, mutual-information penalties for nuisance leakage, an emotion-prosody interaction model, and an uncertainty-aware screening rule. The analytical contribution is not a reported clinical validation but a technically explicit formulation of how factor separation, invariance constraints, and interpretable readouts interact under realistic confounding. Derived consequences identify three failure modes: residual speaker and channel information inflating apparent prosodic effects, emotion-related variation being misread as cognitive slowing, and unstable decision thresholds when calibration is ignored. The framework clarifies which speech features should be treated as evidence, which should be treated as nuisance, and how uncertainty can be carried into downstream screening. Its significance lies in providing a defensible bridge between representation learning for speech and clinically cautious interpretation of prosodic and emotional variation in cognitive-decline research.

Keywords

cognitive decline; speech prosody; emotional speech; disentangled representation learning; clinical decision support; acoustic biomarkers

1. Introduction

Cognitive decline is a heterogeneous process that ranges from mild cognitive impairment to dementia and involves changes in memory, executive function, language, and social-affective behavior [1]. Diagnostic frameworks increasingly emphasize biomarker-supported characterization and longitudinal change rather than a single cross-sectional label [2], while clinical criteria for mild cognitive impairment recognize the importance of preserved independence alongside measurable impairment [3]. Speech is attractive in this context because it can be collected repeatedly, remotely, and with relatively low participant burden.

Narrative and connected-speech tasks have been associated with lexical, syntactic, and fluency differences in Alzheimer's disease and related conditions [4], and reviews of connected speech in neurodegenerative language disorders show that multiple linguistic levels can be affected [5]. Automatic speech analysis extends this line of work by extracting acoustic, lexical, and interactional descriptors at scale [6], with deep sequential models and disfluency features used to identify mild cognitive impairment and dementia from spontaneous speech [7], [8].

The inferential difficulty is that speech is not a direct readout of cognition. A slowed speaking rate may reflect word-finding difficulty, depression, fatigue, medication effects, hearing loss, or microphone mismatch. Reduced pitch variability may reflect affective flattening, cultural speaking style, respiratory capacity, or the acoustic consequence of a noisy channel. Emotional prosody is especially entangled because it encodes both the speaker's affective state and the linguistic and pragmatic structure of an utterance. A classifier that associates low arousal or flattened intonation with cognitive decline may be learning a depression-related signature, a age-related vocal change, or a task-related artifact. The resulting model can appear accurate in a convenience sample while failing to transport to a new clinic, language, or recording device. This is not merely a statistical nuisance; it is a threat to the construct validity of speech-based cognitive markers.

The gap addressed here is the absence of an explicit framework that treats emotion and prosody as separable sources of variation before cognitive inference. Most feature-based pipelines summarize acoustic signals with low-level descriptors and then fit a classifier. Such pipelines can be effective for prediction, but they provide little control over which variation is being used. End-to-end speech models can learn powerful representations, but without constraints they can encode speaker, channel, and affective nuisance factors that correlate with the target in a particular dataset. Disentangled representation learning offers a language for separating generative factors, but its identifiability is limited and must be stated carefully [12], [14]. Speech decomposition methods based on information bottlenecks provide a constructive precedent for isolating content from style-like variation in audio [15], and self-supervised speech models provide rich phonetic and acoustic representations that can serve as content-side encoders [16], [17]. What remains underdeveloped is a clinically framed account of how such components should be combined, constrained, interpreted, and bounded when the target is cognitive decline rather than speaker conversion or synthesis.

The central contribution is a methodological framework for disentangled speech feature analysis in which cognitive-decline inference is performed on factorized representations of content, prosody, emotion, and nuisance. The framework has four parts. First, an utterance is modeled as a composition of content, prosodic, emotional, speaker-channel, and noise factors. Second, a factorized encoder estimates separate latents for these factors, with reconstruction and mutual-information objectives designed to reduce nuisance leakage while preserving cognitive-relevant variation. Third, a readout model uses interpretable prosodic and emotional summaries, including an emotion-prosody interaction term, rather than an opaque aggregate score. Fourth, an uncertainty-aware decision rule separates calibrated risk from confidence and abstains when evidence is insufficient. The analysis below derives consequences for confounding, identifiability, sensitivity, and failure modes, and it compares the framework with conventional acoustic-feature classifiers and unconstrained end-to-end models.

The significance of the framework is both scientific and practical. Scientifically, it makes explicit which speech variations are treated as markers and which are treated as confounds,

allowing competing hypotheses about emotional flattening, prosodic slowing, and cognitive-linguistic impairment to be tested rather than conflated. Practically, it provides a design logic for screening tools that report uncertainty and avoid overconfident inference from short, noisy, or emotionally ambiguous speech samples. The framework does not replace clinical assessment; it organizes the speech-analysis problem so that subsequent validation can target specific mechanisms and specific error modes.

2. Background and Problem Formulation

2.1 Speech, cognition, and the construct-validity problem

Speech-based cognitive assessment draws on multiple levels of processing. Lexical retrieval, syntactic planning, phonological encoding, articulatory execution, and self-monitoring all contribute to fluency and intelligibility. Connected-speech studies have documented reduced semantic content, increased pauses, and altered syntactic complexity in Alzheimer's disease and related disorders [4], [5]. Automatic analyses have used pause distributions, type-token ratios, repetition, and interactional features to distinguish diagnostic groups [6], [7], [8]. These findings support the general premise that cognitive-linguistic change can leave measurable traces in speech. They do not establish that a given acoustic feature is a specific marker of cognition. The same feature may index arousal, fatigue, respiratory function, or recording quality.

Construct validity therefore requires a separation between the target construct and the measurement channel. Cognitive decline is a latent clinical construct inferred from longitudinal cognitive testing, functional status, and, where available, biomarkers [2], [3]. Speech is an observed behavior that reflects the target construct plus many other influences. A useful speech marker must satisfy three conditions. It must vary with the target construct across individuals or within individuals over time. It must be sufficiently stable under irrelevant changes in channel and context. And it must not be driven primarily by a correlated nuisance variable that differs between clinical groups. The third condition is especially important in cognitive-decline research because age, hearing loss, depression, and medication use are correlated with both cognitive status and vocal behavior.

Emotion and prosody complicate this picture because they are both signal and confound. Emotional prosody can convey affect, attitude, and pragmatic intent, and its acoustic correlates include pitch range, intensity, voice quality, and temporal patterning [10], [11]. Prosody also supports syntax, focus, and turn-taking, and its realization depends on context and communicative function [18], [19]. A speaker with cognitive decline may produce less varied prosody because of reduced affective expressivity, because of motor speech changes, because of depression, or because the elicitation task does not elicit spontaneous affect. Without a framework that separates these possibilities, observed prosodic differences are difficult to interpret.

2.2 Emotion and prosody as entangled factors

Emotional speech and linguistic prosody share acoustic dimensions. Pitch movement can mark a question, signal surprise, or index arousal. Intensity can mark emphasis or anger. Speaking rate can reflect urgency, fatigue, or word-finding difficulty. The overlap is not accidental; emotional and linguistic functions are coordinated in production. This creates a representation problem: if an encoder is trained only to predict cognitive status, it may use whichever acoustic cue is most predictive in the training sample, even if that cue is an emotion-related confound. If an encoder is trained only to reconstruct speech, it may preserve

all factors, including speaker identity and channel, because they are useful for reconstruction. The proposed framework treats this tension explicitly by factorizing the latent space and imposing asymmetric constraints: some information should be preserved, some should be compressed, and some should be predicted.

The clinical literature on emotion in aging provides a further reason for caution. Older adults do not simply show global decline in emotional processing; motivational and attentional changes can alter which emotional information is prioritized [20]. Consequently, differences in emotional speech between clinical groups may reflect normal aging, mood, or task engagement rather than cognitive decline. A disentangled representation should allow the analyst to ask whether a cognitive-decline signal remains after emotion-related variation is modeled rather than ignored. This is different from removing emotion entirely. Emotional flattening may itself be a clinically meaningful feature, but its interpretation depends on whether it is separable from prosodic control, lexical content, and depression-related variation.

2.3 Why disentanglement is appropriate and what it cannot guarantee

Disentangled representation learning aims to recover factors that generate observed data and to organize them so that changes in one factor do not produce uncontrolled changes in others. Variational autoencoders with constrained priors can encourage factorized representations [12], but identifiability is not automatic. Theoretical work has shown that unsupervised disentanglement is impossible without inductive biases or additional information, and that many objectives can produce similarly scoring but semantically different representations [14]. Speech decomposition via information bottlenecks offers a more targeted approach by separating content from speaker or style information using mutual-information constraints [15]. The lesson for cognitive-decline research is that disentanglement should be treated as a modeling discipline rather than a guarantee. The framework must specify which factors are assumed separable, which observations identify them, and which residual confounds remain.

This is why the present framework does not claim that emotional and prosodic factors can be perfectly isolated from a single utterance. Instead, it defines a set of latent variables, an objective that penalizes unwanted dependence, and a readout that is interpretable under stated assumptions. The goal is to make the inferential pathway from speech to cognitive-decline probability explicit, not to assert that the pathway has been empirically validated. In this sense, the framework is closer to a measurement model than to a black-box classifier. It specifies what would count as evidence for a prosodic or emotional effect and what would count as evidence of nuisance leakage.

3. Methodological Framework

3.1 Signal model and factor definitions

Consider an utterance indexed by u, with observed acoustic waveform or feature sequence x_u. The framework assumes that x_u is generated by a composition of content, prosody, emotion, speaker-channel, and noise factors. Content c_u includes lexical, phonological, and syntactic information. Prosody p_u includes pitch, timing, intensity, and voice-quality patterns that organize speech and convey structure. Emotion e_u includes affective state and emotional expression as realized in voice. Speaker-channel s_u includes vocal tract characteristics, age-related voice change, microphone response, and background noise. The residual n_u captures stochastic variation not assigned to the other factors. The generative relationship is written as a conceptual composition rather than a known physical equation:

$$x_u = G(c_u, p_u, e_u, s_u, n_u)$$

Here G denotes a possibly nonlinear generation process, and the factors are latent variables rather than directly observed labels. The equation is not a claim that speech is generated by a simple additive mechanism. It states the factorization assumption that organizes the analysis: variation in x_u is attributed to distinct sources, and the inference problem is to estimate representations that respect those sources. The composition is useful because it separates two questions that are often conflated. The first is whether cognitive status is associated with prosodic or emotional variation. The second is whether that variation can be estimated independently of speaker, channel, and content. A model that answers the first question without addressing the second is vulnerable to confounding.

The factor definitions also impose boundary conditions. Content and prosody are not independent in natural speech; syntactic structure influences prosodic phrasing, and emotional state influences content selection. The framework therefore treats factorization as a soft constraint rather than a hard causal claim. It aims to reduce unwanted dependence, not to deny the coordination among factors. This distinction matters for interpretation. A residual correlation between prosody and content may reflect genuine linguistic coupling or imperfect separation, and the analytical assessment below distinguishes these possibilities.

3.2 Factorized encoder and latent structure

The encoder estimates a posterior over four latent variables: z_c for content, z_p for prosody, z_e for emotion, and z_n for nuisance. The factorization assumption is conditional independence given the utterance:

$$q(z_c, z_p, z_e, z_n | x_u) \;=\; \prod_\{k \in \{c, p, e, n\}\}\, q(z_k | x_u)$$

In this expression, q denotes the encoder distribution, k indexes the latent factors, and the product runs over the four factors. The notation does not imply that the true posterior factorizes; it specifies an inductive bias that makes the latent space interpretable. Each factor can be implemented with a different encoder inductive bias. The content encoder can use self-supervised speech representations that preserve phonetic and lexical information [16], [17]. The prosody encoder can use contour-based features such as fundamental frequency, energy, duration, and voice quality, which are standard in affective computing and voice research [9]. The emotion encoder can use similar acoustic inputs but be trained with affective supervision or weak emotional labels when available. The nuisance encoder can use speaker identity, channel metadata, or augmentation labels to capture variation that should not drive the cognitive readout.

The separation between z_p and z_e is the most delicate part of the design. Both factors can be informed by pitch and energy. If the emotion encoder is allowed to absorb all prosodic variation, then z_p becomes uninformative. If the prosody encoder is allowed to absorb all emotion-related variation, then z_e becomes redundant. The framework resolves this by using two different readout targets. The prosody factor is encouraged to preserve variation that predicts cognitive status but not speaker or channel. The emotion factor is encouraged to preserve variation that predicts affective state but not cognitive status, except through the explicit interaction term in the readout. This asymmetry is a modeling choice, not a discovered fact. It reflects the hypothesis that emotional state is more likely to be a confound or a mediator than a direct diagnostic criterion, while prosodic control is a plausible pathway through which cognitive-linguistic changes become observable.

A related engineering analogue appears in hierarchical style transfer for singing voice synthesis, where content and style are modeled separately to permit out-of-domain style control [13]. That work is not a clinical study and offers no evidence about cognitive decline, but it illustrates a design principle: when a signal contains both content-like and style-like variation, explicit separation can improve control and interpretation. The present framework borrows that principle without assuming that emotional expression in older adults is equivalent to singing style. The analogy is useful because it clarifies why a factorized representation may be more interpretable than a single pooled embedding, not because the underlying tasks are the same.

3.3 Objective function and invariance constraints

The encoder and generator are trained with a composite objective that balances reconstruction, nuisance compression, and cognitive-relevant information. A compact form is:

$$L = L_rec + \alpha I(z_p; z_n) + \beta I(z_e; z_n) - \delta I\left((z_p, z_e); y\right) + \lambda \sum_k KL_k$$

Here L_rec is a reconstruction or predictive loss for the observed speech representation, I(·;·) denotes mutual information, y is the cognitive-status or cognitive-severity target, and α, β, δ, and λ are nonnegative weights. The term I(z_p;z_n) penalizes dependence between the prosody latent and the nuisance latent. The term I(z_e;z_n) penalizes dependence between the emotion latent and the nuisance latent. The negative term −δ I((z_p,z_e);y) rewards preservation of joint prosody-emotion information that is predictive of cognitive status. The final term KL_k is the Kullback-Leibler divergence between the encoder posterior for factor k and its prior, which regularizes the latent distributions. The objective is a proposed design; the mutual-information terms must be estimated with mini-batch estimators or variational bounds, and the weights must be selected through validation rather than assumed.

The objective exposes a trade-off that is often hidden in end-to-end models. Increasing δ encourages the representation to retain cognitive-relevant prosodic and emotional variation, but it can also retain nuisance variation that happens to correlate with y in the training sample. Increasing α and β reduces nuisance leakage but can discard clinically meaningful variation if the nuisance and target are entangled. The appropriate balance depends on the intended use. For a screening tool that will be deployed across recording devices, stronger nuisance penalties are justified. For a mechanistic study of emotional flattening in a single clinical center, weaker penalties may be preferable because they preserve more affective detail. The framework therefore treats the weights as design parameters that encode an explicit measurement policy.

The negative mutual-information term requires care because y may be correlated with age, education, or depression. If those variables are not modeled, maximizing I((z_p,z_e);y) can simply encode the confound. The framework therefore includes a separate invariance constraint on nuisance variables and recommends that covariates be included in the readout model as calibration terms rather than hidden in the latent space. This separation does not eliminate confounding, but it makes the assumptions visible. A model that predicts y from z_p and z_e while also adjusting for measured covariates is more interpretable than a model that predicts y from a pooled representation with no explicit covariate structure.

3.4 Emotion-prosody interaction and cognitive readout

The readout model operates on interpretable summaries of the latent factors rather than on the full latent vector. Let φ_p be a vector of prosodic summaries, such as pitch range, pitch variability, articulation rate, pause proportion, and voice-quality stability. Let φ_e be a vector of emotional summaries, such as valence-related pitch contour, arousal-related intensity, and affective dynamic range. Let x_cov include measured covariates such as age, sex, education, hearing status, and depression score when available. A linear readout with an interaction term is:

$$\eta_u = \beta_0 + \beta_p \cdot \varphi_p + \beta_e \cdot \varphi_e + \beta_ep \cdot (\varphi_p \odot \varphi_e) + \beta_x \cdot x_cov$$

In this equation, η_u is the linear predictor for utterance u, β_0 is an intercept, β_p, β_e, β_ep, and β_x are coefficient vectors, and ⊙ denotes elementwise multiplication. The interaction term is included because emotional state can change how prosodic variation is expressed. For example, a reduction in pitch range may be more informative when it occurs during positive-affect elicitation than during a neutral reading task. The interaction term allows the model to represent such context dependence without creating a separate model for every emotion category. The coefficients are not assumed to be known; they are estimated in a validation study. The equation is a specification of the readout, not a report of fitted values.

The use of interpretable summaries is a deliberate trade-off. A fully end-to-end model might achieve better predictive performance in a large sample, but it would provide weaker control over which variation is used. The proposed readout sacrifices some flexibility for transparency. It allows an analyst to inspect whether the cognitive signal is carried by prosodic variation, emotional variation, or their interaction. It also allows sensitivity analysis: if the interaction term is removed, does the model rely more heavily on main effects that may be confounded? If the emotion summaries are residualized against depression scores, does the prosodic effect remain? These questions are central to construct validity, and they are easier to answer with a structured readout than with a pooled embedding.

The linear predictor is converted to a probability through a link function, such as a logistic function, and the resulting probability is calibrated before use. Calibration is not a cosmetic step. Cognitive-decline screening typically involves asymmetric costs: a missed case may delay evaluation, while a false positive may cause anxiety and unnecessary testing. A probability that is not calibrated cannot be translated into a defensible threshold. The framework therefore separates three quantities: the latent factor estimates, the linear predictor, and the calibrated probability. Each can be audited independently.

3.5 Uncertainty, abstention, and decision logic

Speech samples vary in duration, noise, and emotional ambiguity. A screening system that produces a probability for every sample, regardless of data quality, will be overconfident on short or noisy recordings. The framework uses predictive entropy as a simple uncertainty summary:

$$H(y|x_u) = -\sum_\{y \in \{0,1\}\} p(y|x_u) \log p(y|x_u)$$

Here H(y|x_u) is the entropy of the predicted cognitive-status distribution for utterance u, and p(y|x_u) is the calibrated probability for class y. Entropy is not a complete measure of uncertainty, but it is a useful signal for abstention. A decision rule can require both a calibrated probability above a threshold and entropy below a ceiling before returning a positive screen. If entropy is high, the system can return an indeterminate result or request a longer sample. This is preferable to forcing a binary decision from ambiguous evidence.

The threshold and entropy ceiling should be chosen with explicit cost considerations. Let τ be the probability threshold and h_0 the entropy ceiling. A positive screen is issued only when $p(y=1|x_u) \geq \tau$ and $H(y|x_u) \leq h_0$. Otherwise, the sample is classified as not positive or as indeterminate, depending on the clinical workflow. The framework does not prescribe a single threshold because the appropriate trade-off depends on prevalence, follow-up capacity, and the consequences of false positives and false negatives. It does prescribe that the threshold be selected on calibration data and reported alongside the model, rather than tuned on the evaluation sample.

Uncertainty also has a representation-learning interpretation. If the encoder is uncertain about whether a pitch change reflects emotion or prosody, the readout should reflect that ambiguity. One way to operationalize this is to use an ensemble of factorized encoders and measure disagreement in φ_p and φ_e. High disagreement across encoders indicates that the factor separation is unstable, which is a warning sign for interpretation. The entropy of the final prediction captures only part of this instability. A full treatment would propagate factor uncertainty into the readout and report both predictive uncertainty and factor uncertainty. The framework can accommodate this extension, but the simpler entropy rule is sufficient to show why abstention is a necessary component of a clinically cautious system.

4. Analytical Assessment

4.1 Identifiability and the limits of factor separation

The factorization in the encoder is not guaranteed to recover the true factors. A local identifiability condition can be stated using the Jacobian of the generative map. If G is differentiable and the factors p, e, and s are locally separable, then the Jacobian with respect to those factors should have full column rank:

$$rank(\partial G/\partial(p,e,s)) = d_p + d_e + d_s$$

Here $\partial G/\partial(p,e,s)$ is the Jacobian of the generative map with respect to prosody, emotion, and speaker-channel factors, and d_p, d_e, and d_s are their respective dimensions. The condition states that small changes in different factors should produce linearly independent changes in the observed speech representation. If the rank is deficient, then different combinations of prosody, emotion, and speaker characteristics produce the same observation. In that case, no encoder can distinguish the factors from a single utterance without additional assumptions or auxiliary data. This is a structural limitation, not a failure of optimization.

The condition is useful because it identifies where auxiliary information is needed. If prosody and emotion are collinear for a given task, then additional elicitation conditions may be required, such as neutral and affective speech samples from the same speaker. If speaker and channel factors are collinear with prosody, then channel augmentation or speaker normalization may be necessary. If content and prosody are strongly coupled, then a factored model may need to condition on syntactic or lexical context. The framework does not assume that these conditions are satisfied; it provides a language for diagnosing when they are not.

The literature on disentanglement supports a cautious interpretation. Unsupervised disentanglement is not identifiable without inductive biases, and many objectives can produce different factorizations with similar reconstruction quality [14]. Speech decomposition methods use information bottlenecks to impose structure, but they still rely on assumptions about what content and style mean in the data [15]. The present framework therefore treats factor labels as operational constructs. The prosody latent is not "the" prosody factor in an

absolute sense; it is the representation that the encoder learns under the specified objective and data. This operational stance is important for clinical interpretation because it prevents overclaiming. A statistically separated factor is not necessarily a biological mechanism.

4.2 Confounding and bias propagation

The most important analytical consequence of entanglement is bias propagation. Suppose the observed prosody representation is contaminated by a speaker or channel factor s according to z_p^obs = z_p + Γs + δ_p, where Γ is a loading matrix and δ_p is residual noise. If the readout omits s, the observed linear predictor contains a term that depends on the nuisance factor:

$$\eta_obs = \beta_p \cdot z_p + \beta_e \cdot z_e + \beta_p \cdot \Gamma s + \varepsilon_obs$$

In this expression, η_obs is the observed linear predictor, β_p and β_e are the readout coefficients, Γ is the nuisance loading, s is the nuisance factor, and ε_obs collects residual terms. The term β_p·Γs is the bias introduced by nuisance leakage. If s differs between clinical groups, the model can attribute a group difference to prosody even when the true effect is due to speaker age, recording device, or vocal pathology. The equation also shows why post hoc correction is difficult: the bias depends on the product of the readout coefficient and the loading matrix, neither of which is directly observed. Disentanglement aims to reduce Γ during representation learning, while covariate adjustment aims to reduce the residual effect of s in the readout. Both are needed when the nuisance factor is only partially observed.

Emotion introduces a parallel bias. If z_e contains both affective state and cognitive-related slowing, then the coefficient β_e conflates two pathways. The interaction term in the readout can partially separate them, but only if the emotion summaries are measured under conditions that vary affect independently of cognitive load. This is a design requirement, not a statistical fix. The framework therefore recommends elicitation tasks that include neutral, positive, and negative affect conditions, as well as tasks that vary cognitive load. Without such variation, the interaction term is weakly identified, and the model may default to a main effect that is confounded.

4.3 Sensitivity and failure modes

The framework has several characteristic failure modes. The first is over-compression. If the nuisance penalty is too strong, the encoder may discard prosodic variation that is genuinely associated with cognitive decline because that variation is correlated with speaker age. The result is a conservative model that misses true effects. The second is under-compression. If the nuisance penalty is too weak, the encoder may retain channel and speaker information, producing unstable predictions across devices. The third is emotion absorption. If the emotion encoder is too expressive, it may absorb prosodic variation that should be attributed to linguistic or cognitive processing. The fourth is task dependence. If the speech sample does not elicit enough emotional or prosodic variation, the latent factors are weakly identified, and the readout becomes sensitive to incidental correlations.

These failure modes can be assessed with sensitivity analyses. Varying the weights α, β, and δ and observing changes in the readout coefficients tests the stability of the factor separation. Replacing the emotion summaries with residualized versions that remove depression scores tests whether the emotional effect is distinct from mood. Training on augmented channel conditions and evaluating on held-out devices tests nuisance robustness. Comparing the structured readout with a pooled end-to-end classifier tests whether

interpretability costs predictive power. These analyses do not prove construct validity, but they make the model's dependencies visible.

The framework also has boundary conditions. It is designed for research and screening support, not for diagnosis. It assumes that cognitive status can be measured with reasonable accuracy through clinical assessment and, where available, biomarkers. It assumes that speech samples are long enough to estimate prosodic and emotional summaries reliably. It assumes that the recording protocol is consistent enough to permit channel invariance to be learned. When these assumptions fail, the model's output should be treated as low-confidence regardless of its predicted probability. The entropy rule provides one safeguard, but it cannot compensate for a fundamentally invalid measurement design.

4.4 Comparison with conventional pipelines

Conventional acoustic-feature pipelines extract low-level descriptors such as pitch, energy, and spectral statistics, then fit a classifier. These pipelines are transparent at the feature level and can be effective when the training and test distributions are similar. Their weakness is that the features are not organized around causal or measurement roles. A pitch statistic can reflect emotion, syntax, speaker identity, or microphone response. A classifier can exploit any of these associations. The proposed framework differs by assigning features to factors and by constraining the dependence between factors. It does not eliminate the need for feature engineering, but it gives the engineering a measurement interpretation.

Unconstrained end-to-end models offer greater representational flexibility and can exploit large self-supervised speech models [16], [17]. They are attractive when the goal is pure prediction and the evaluation distribution matches the training distribution. Their weakness is interpretability and transportability. A pooled representation can encode nuisance factors in ways that are difficult to audit. The proposed framework can use self-supervised representations as content encoders while imposing explicit constraints on prosody, emotion, and nuisance. This hybrid design retains the benefits of large speech models without treating the entire representation as a single predictive object.

A third alternative is to remove emotion and prosody entirely and rely on lexical or fluency features. This reduces confounding but may discard clinically meaningful variation. Emotional flattening and prosodic changes can be early signs of altered communication, and ignoring them limits the scope of the analysis. The proposed framework keeps these factors but separates them, allowing the analyst to test whether they contribute to cognitive inference after adjustment. This is a more informative scientific design than either ignoring emotion or pooling it with cognition.

5. Technical Discussion

5.1 Interpretability of prosodic and emotional factors

The interpretability of the framework depends on the mapping between latent factors and observable summaries. The prosody factor should be accompanied by summaries that are familiar to speech scientists: pitch range, pitch variability, articulation rate, pause duration, and voice-quality stability. The emotion factor should be accompanied by summaries that are grounded in affective prosody research: pitch contour direction, intensity dynamics, and valence-arousal correlates [10], [11]. These summaries are not perfect measures of the latent factors, but they provide a bridge between the representation and clinical interpretation. A

reader can ask whether a model's cognitive signal is carried by reduced pitch variability, increased pause duration, or an interaction between positive affect and prosodic range.

The interaction term is especially important for interpretation. A main effect of reduced pitch variability may reflect motor speech changes, depression, or cultural speaking style. An interaction between positive affect and pitch variability is more specific: it suggests that the speaker's prosodic response to positive affect is altered. Such a pattern could be consistent with affective flattening, but it could also reflect task engagement or medication. The framework does not resolve these possibilities automatically. It provides a structured way to test them by comparing models with and without the interaction, and by examining whether the interaction persists after adjustment for depression and hearing status.

5.2 Clinical translation and decision support

Clinical translation requires more than a predictive model. It requires a clear statement of intended use, a validated threshold, a plan for handling indeterminate results, and a pathway for follow-up. The framework supports these requirements by separating the cognitive readout from the uncertainty rule. A positive screen is not a diagnosis; it is a signal that further assessment is warranted. An indeterminate result is not a negative result; it is an invitation to collect a better sample. This distinction is important because speech-based screening may be used in settings where follow-up capacity is limited. A system that produces too many indeterminate results will not be useful, while a system that suppresses uncertainty will produce false reassurance.

The calibration step is central to clinical translation. A probability of 0.7 should mean that approximately 70 percent of similar cases are positive, not merely that the model's score is above a threshold. Calibration can be assessed with reliability diagrams, Brier scores, and decision-curve analysis in a validation study. The framework does not specify a calibration method because the choice depends on sample size and the number of covariates. It does specify that calibration must be performed before threshold selection and that the threshold must be reported with the model. Without this discipline, the uncertainty rule is not meaningful.

5.3 Limitations and research implications

The framework has several limitations. It assumes that the factor structure is stable across tasks and populations, which may not hold in multilingual or cross-cultural settings. Prosodic norms vary across languages and dialects, and emotional expression is culturally patterned. A model trained in one context may not transport to another without adaptation. The framework also assumes that nuisance factors can be measured or approximated. Speaker identity and channel metadata are often available, but depression, hearing loss, and medication use may not be. Unmeasured nuisance can still bias the readout. Finally, the framework does not provide empirical evidence of clinical validity. It provides a design and an analytical rationale that must be tested in prospective studies with appropriate clinical endpoints.

These limitations suggest a research agenda rather than a finished tool. The first step is to establish whether the factorized representation is stable across elicitation tasks. The second is to test whether the prosody-emotion interaction improves construct validity beyond main effects. The third is to evaluate transportability across devices and languages. The fourth is to compare the structured readout with pooled end-to-end models under distribution shift. These questions are answerable with the framework, but they are not answered by it. The contribution is to make the questions precise.

6. Conclusion

Speech-based identification of cognitive decline is constrained by the entanglement of content, prosody, emotion, speaker identity, and channel. A classifier that ignores this entanglement can produce predictions that are accurate in one sample but uninterpretable and unstable in another. This manuscript developed a disentangled speech feature analysis framework that separates these factors, constrains nuisance leakage, models emotion-prosody interaction, and carries uncertainty into a screening decision. The framework is grounded in existing work on connected speech in cognitive decline, acoustic prosody, affective speech, disentangled representation learning, and self-supervised speech models. Its analytical value lies in making the inferential pathway explicit: which variation is treated as cognitive evidence, which is treated as emotion or prosody, and which is treated as nuisance.

The main defensible insights are three. First, cognitive-decline inference should not be performed on pooled acoustic representations without an explicit nuisance model, because speaker and channel leakage can mimic prosodic effects. Second, emotion and prosody should be modeled jointly rather than collapsed, because their interaction can be clinically informative and because emotional variation can otherwise be misattributed to cognitive decline. Third, uncertainty and calibration are not optional add-ons; they determine whether a screening threshold is meaningful. The framework does not claim that factor separation is perfect or that the proposed readout has been clinically validated. It claims that the measurement problem can be stated more precisely and that a factorized design provides a stronger basis for validation than an unconstrained classifier.

The broader implication is that speech markers of cognitive decline should be treated as measurement constructs rather than raw predictions. A model that reports a probability without stating which speech variation supports it is difficult to falsify, compare, or improve. A factorized model with interpretable summaries and explicit nuisance constraints is more demanding to build, but it supports the kind of mechanistic and transportability questions that clinical translation requires. Future work can use the framework to design elicitation protocols, estimate factor stability, and test whether prosodic and emotional markers remain informative after adjustment for mood, hearing, and recording conditions.

References

1. Livingston, G., Huntley, J., Sommerlad, A., Ames, D., Ballard, C., Banerjee, S., ... & Mukadam, N. (2020). Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. The Lancet, 396(10248), 413-446.

2. Jack, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., ... & Sperling, R. (2018). NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. Alzheimer's & Dementia, 14(4), 535-562.

3. Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., ... & Phelps, C. H. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & Dementia, 7(3), 270-279.

4. Fraser, K. C., Meltzer, J. A., & Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. Journal of Alzheimer's Disease, 49(2), 407-422.

5. Boschi, V., Catricalà, E., Consonni, M., Chesi, C., Moro, A., & Cappa, S. F. (2017). Connected speech in neurodegenerative language disorders: A review. Frontiers in Psychology, 8, 269.

6. Pulido, M. L. B., Hernández, J. B. A., Ballester, M. Á. F., González, C. M. T., & Jiménez, M. D. C. P. (2020). Alzheimer's disease and automatic speech analysis: A review. Expert Systems with Applications, 150, 113213.

7. Themistocleous, C., Eckerström, M., & Kokkinakis, D. (2020). Identification of mild cognitive impairment from speech in Swedish using deep sequential neural networks. Frontiers in Neurology, 11, 578696.

8. Nasreen, S., Rohanian, M., Hough, J., & Purver, M. (2021). Alzheimer's dementia recognition from spontaneous speech using disfluency and interactional features. Frontiers in Computer Science, 3, 640669.

9. Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., ... & Truong, K. P. (2016). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing, 7(2), 190-202.

10. Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code? Psychological Bulletin, 129(5), 770-814.

11. Bänziger, T., & Scherer, K. R. (2005). The role of intonation in emotional expressions. Speech Communication, 46(3-4), 252-267.

12. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., ... & Lerchner, A. (2017). beta-VAE: Learning basic visual concepts with a constrained variational framework. In International Conference on Learning Representations.

13. Zhang, Y., Huang, R., Li, R., He, J., Xia, Y., Chen, F., ... & Zhao, Z. (2024, March). Stylesinger: Style transfer for out-of-domain singing voice synthesis. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 17, pp. 19597-19605).

14. Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., & Bachem, O. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. In Proceedings of the 36th International Conference on Machine Learning (PMLR 97, pp. 4114-4124).

15. Qian, K., Zhang, Y., Chang, S., Yang, X., & Hasegawa-Johnson, M. (2020). Unsupervised speech decomposition via triple information bottleneck. In Proceedings of the 37th International Conference on Machine Learning (PMLR 119, pp. 7836-7846).

16. Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In Advances in Neural Information Processing Systems (Vol. 33, pp. 12449-12460).

17. Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 3451-3460.

18. Wagner, M., & Watson, D. G. (2010). Experimental and theoretical advances in prosody: A review. Language and Cognitive Processes, 25(7-9), 905-945.

19. Cole, J. (2015). Prosody in context: a review. Language, Cognition and Neuroscience, 30(1-2), 1-31.

20. Mather, M., & Carstensen, L. L. (2005). Aging and motivated cognition: The positivity effect in attention and memory. Trends in Cognitive Sciences, 9(10), 496-502.