CARMA 2026

8th Int. Conf. on Advanced Research Methods and Analytics
1-3 July 2026, Valencia, Spain
DOI: https://doi.org/10.4995/CARMA2026.2026.21902

Vocal Delivery Predicts Startup Success: A Multimodal Pitch Analysis

Hector Borobia[1] , Elies Seguí-Mas[2] ,Guillermina Tormo-Carbó[2]
[1]VRAIN, Universitat Politècnica de València, Spain, [2]eSMART Research Center, Universitat Politècnica de València, Spain.

Abstract
Investors evaluate startup pitches through multiple channels simultaneously, yet research typically examines these channels in isolation. This study asks which modality—textual content, acoustic delivery, or visual presentation—best predicts subsequent venture outcomes when analysed through machine learning. We combine theory-grounded interpretable features with frozen pretrained embeddings (text-embedding-3-large, Wav2Vec 2.0, VideoMAE) and evaluate 40 pitch videos under 50 repetitions of stratified 5-fold cross-validation. Audio models consistently dominate, with Wav2Vec 2.0 embeddings achieving a median AUC of 0.736. Text and video models hover near chance. Multimodal fusion provides no reliable improvement. SHAP and embedding-probing analyses reveal that discriminative information resides in latent spectrotemporal patterns rather than handcrafted prosodic summaries. These results suggest that how founders speak matters more than what they say for predicting startup success.

Keywords: *startup success prediction; multimodal machine learning; entrepreneurial pitching; audio embeddings; Wav2Vec 2.0; signaling theory.*

1. Introduction

The entrepreneurial pitch constitutes a critical episode in venture development: a compressed window where founders must convey vision, capability, and potential to skeptical investors under extreme uncertainty. Despite decades of research, our understanding remains fragmented. The fundamental reality that investors process pitch information through multiple simultaneous channels—what founders say, how they say it, and how they present themselves visually—has received limited integrated attention. Investors consistently report making intuitive, holistic judgments within minutes (Huang and Pearce, 2015; Tsay, 2021), yet academic research has predominantly relied on single-modality analyses.

This study addresses a fundamental question: Can multimodal machine learning—simultaneously integrating textual, acoustic, and visual signals from pitch videos—predict subsequent startup success, and if so, which modality carries the strongest predictive signal?

Article in Press

This work is licensed under a Creative Commons License CC BY-NC-SA 4.0
Editorial Universitat Politècnica de València

Recent advances in pretrained transformer models, self-supervised audio models, and video transformers now enable researchers to operationalize this multimodal reality. Critically, we use these embeddings in a frozen manner, reducing overfitting risk in small-sample settings.

The practical motivation is compelling. Venture capital decision-making involves substantial uncertainty reduction through pitch evaluation (Gompers et al., 2020). If specific pitch characteristics systematically predict venture outcomes, this knowledge has direct implications for entrepreneur training programs, investor due diligence, and accelerator screening at scale.

This study contributes to computational entrepreneurship (Shrestha and He, 2025) by developing a small-sample-disciplined pipeline combining theory-grounded features with frozen pretrained embeddings. We provide (i) a reproducible pipeline for small-N research, (ii) a controlled comparison of textual, acoustic, and visual signals, and (iii) explainability analyses combining SHAP attributions with embedding probing to link latent representations to interpretable constructs.

2. Conceptual Framework and Empirical Design

2.1. Theoretical Logic

Our framework integrates three theoretical streams. For text, signaling theory (Spence, 1973; Bafera and Kleinert, 2023; Colombo, 2021) provides the logic: content elements—traction, market sizing, team credentials, competitive positioning, funding specificity—function as quality signals reducing investor uncertainty (Gavrilenko et al., 2023; Tarillon et al., 2023). Linguistic style (pronoun usage, sentiment, certainty) further conveys information about founder characteristics (Pennebaker, 2011; Parhankangas and Ehrlich, 2014; Mahmood and Yeganegi, 2023).

For audio, the bidimensional affect model (Russell, 2003) anchors our variables. Pitch variation signals dynamism; speech rate conveys competence; vocal intensity projects authority (Allison et al., 2022). Critically, research reveals inverted-U relationships: moderate levels maximize effectiveness (Miao et al., 2024). Pretrained Wav2Vec 2.0 embeddings capture richer spectrotemporal dynamics—timbre, rhythm, micro-pauses—beyond what summary statistics represent (Goossens et al., 2022; Van Aken et al., 2023).

For video, non-verbal communication constitutes a third information channel encompassing body language, gestures, and presentation quality (Clarke et al., 2019; Tsay, 2021). We employ VideoMAE (Tong et al., 2022) for dense spatiotemporal embeddings. Our late-fusion stacking architecture reflects how investors evaluate modalities semi-independently before forming holistic judgments (Chen et al., 2009).

Article in Press

2.2. Data and Variables

The dataset comprises 40 startup pitch videos (3–6 minutes) from public YouTube competitions, balanced between success (n=20) and failure (n=20) via Crunchbase. Success: active with Series A+ or funding >$10M (Gompers et al., 2020). Failure: closed (n=13) or "zombie" (n=7)—active 8+ years with <$300K in funding (Cantamessa et al., 2018; Ungerer et al., 2021).

Text features: 14 interpretable variables extracted via Whisper transcription and GPT-4o, plus 3,072-dim embeddings (text-embedding-3-large). Audio: 9 prosodic/affective variables plus 768-dim Wav2Vec 2.0 embeddings. Video: 768-dim VideoMAE embeddings only. We evaluate XGBoost (primary) and Random Forest across unimodal, bimodal early fusion, trimodal, and late-fusion stacking configurations. Embeddings are frozen, hyperparameters set a priori, preprocessing within training folds. Evaluation: 50 repetitions of stratified 5-fold CV, reporting median AUC and 95% intervals from out-of-fold predictions.

3. Results

3.1. Predictive Performance

Results are summarized in Table 1 and Figure 1. Three findings stand out. First, the acoustic modality dominates: Audio-Emb (Wav2Vec 2.0) attains the highest median AUC (0.736 [0.607, 0.825]), while Audio-All (manual + embeddings) performs nearly identically (0.735 [0.611, 0.811]). The predictive signal is primarily paralinguistic, captured most effectively by latent acoustic representations, whereas adding handcrafted prosodic summaries provides little incremental benefit.

Second, multimodal fusion does not yield reliable gains. The strongest fusion configuration remains below Audio-Emb (e.g., Audio+Video: 0.657 [0.496, 0.738]). Third, repeated-CV comparisons confirm that late fusion rarely improves upon audio: the median ΔAUC (LateFusion-Text+Audio − Audio-All) is −0.128 with a win rate of only 4% (Figure 2). Text and video models remain near chance (AUCs ≈0.39–0.50). The aggregated out-of-fold confusion matrix for Audio-Emb (Figure 3) shows that the model correctly classifies 71.2% of failures and 64.6% of successes across 2,000 out-of-fold predictions.

Article in Press

Table 1. Predictive performance (XGBoost) under repeated stratified cross-validation (50 rep. × 5-fold, N=40).

Model	Features	XGB AUC
Audio-Emb	Wav2Vec2 (768d)	0.736 [0.607, 0.825]
Audio-All	Manual + embeddings	0.735 [0.611, 0.811]
Audio+Video	Early fusion	0.657 [0.496, 0.738]
LateFusion-All	Stacking (all)	0.575 [0.402, 0.722]
LateFusion-T+A	Stacking (text+audio)	0.593 [0.407, 0.727]
Text+Audio	Early fusion	0.516 [0.331, 0.631]
Video-Emb	VideoMAE (768d)	0.501 [0.414, 0.588]
Text-Manual	Interpretable (14)	0.419 [0.272, 0.488]
Audio-Manual	Prosodic/affective (9)	0.364 [0.277, 0.484]

Note: Values are median [2.5th, 97.5th percentiles]. Classes are balanced (20/20).

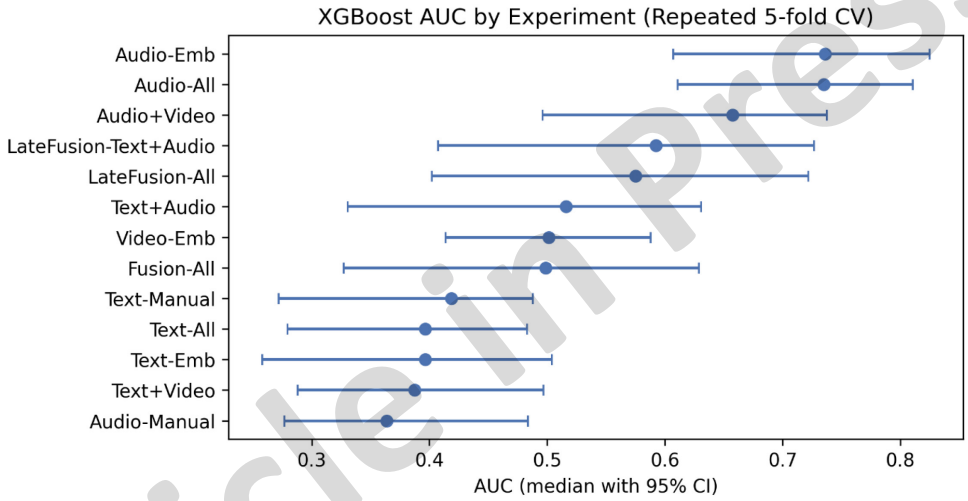


Figure 1. XGBoost AUC across model configurations. Points: median AUC; bars: 95% interval across 50 repetitions of stratified 5-fold CV (N = 40).

Article in Press

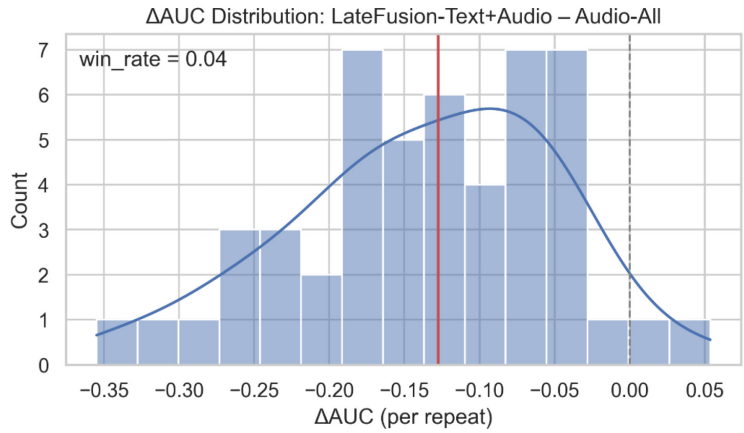


Figure 2. ΔAUC distribution (LateFusion-Text+Audio − Audio-All). The distribution is predominantly negative (win rate = 0.04), confirming fusion does not improve upon the audio baseline.

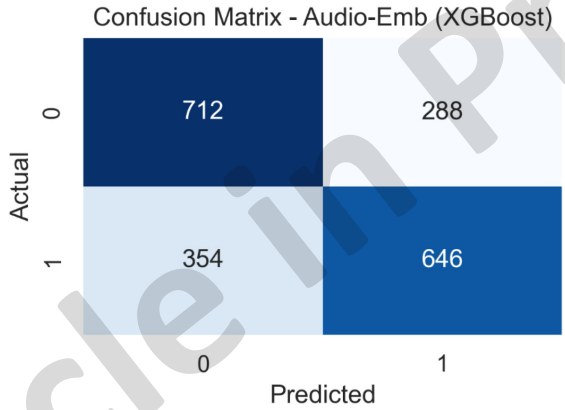


Figure 3. Aggregated out-of-fold confusion matrix for Audio-Emb (XGBoost). Counts aggregate predictions across 50 repetitions (N = 40 × 50 = 2,000 predictions). The model correctly identifies 71.2% of failures and 64.6% of successes.

3.2. Explainability

We compute SHAP values (Lundberg and Lee, 2017) and probe embedding spaces by correlating interpretable variables with embedding dimensions. For manual audio features, SHAP magnitudes are small—only mean loudness shows a noticeable contribution, while other prosodic summaries remain near zero. By contrast, audio embedding dimensions (Figure 4) exhibit substantially larger SHAP effects with clear directional separation between success and failure, indicating reliance on richer spectrotemporal configurations: rhythm, micro-pauses,

Article in Press

emphasis patterns, and timbre. This confirms that simple prosodic averages are too coarse to capture discriminative vocal delivery.

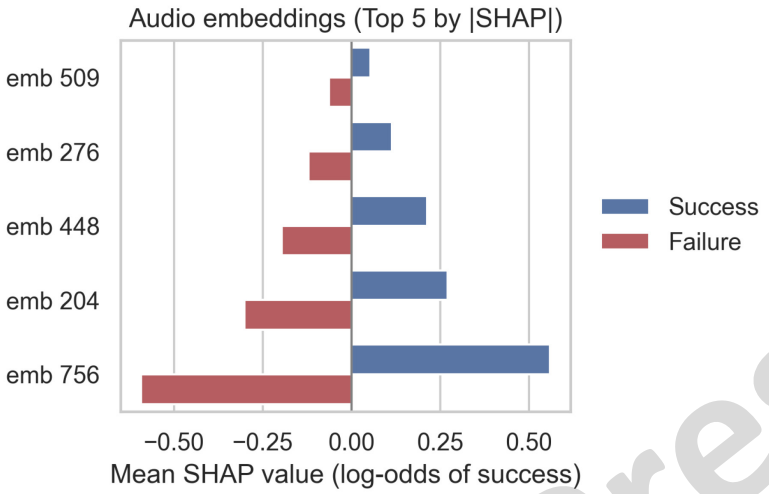


Figure 4. *SHAP summary for Wav2Vec 2.0 embedding dimensions (Top 5). Effects are substantially larger than for handcrafted prosodic summaries, with clear directional separation between success and failure.*

Probing analysis (Figure 5) reveals that several manual variables align strongly with embedding dimensions: mean pitch, pitch variability, speech rate, and loudness variability reach max $|r| \approx 0.6–0.75$. This confirms that Wav2Vec 2.0 encodes core prosodic components while capturing additional discriminative information. Key pitch evaluations reflect holistic impressions—preparedness, credibility, fluency—arising from subtle combinations of timing, emphasis, and consistency that pretrained embeddings capture effectively, explaining why embedding-based features dominate contribution even when theory-grounded manual variables are included.

Article in Press

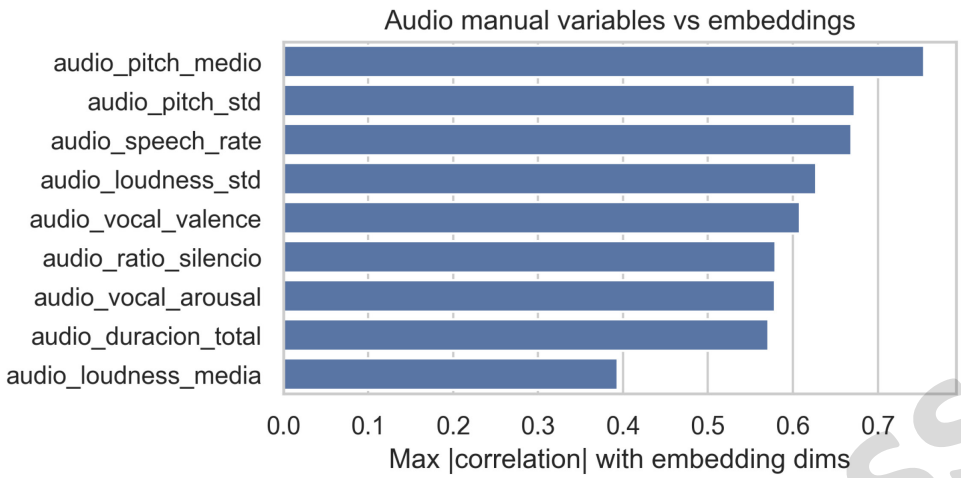


Figure 5. *Probing audio embeddings with manual variables. Max absolute correlation between each handcrafted variable and any Wav2Vec 2.0 dimension. Higher values indicate strong encoding of that construct.*

4. Discussion and Conclusion

This study examined whether multimodal machine learning can predict startup success from pitch videos. Using frozen pretrained representations and 50 repetitions of stratified 5-fold CV (N = 40), audio-based models yield the strongest discrimination (median AUC = 0.736). These findings extend prior research on vocal delivery (Allison et al., 2022; Goossens et al., 2022) by showing that latent acoustic embeddings, not coarse prosodic averages, carry most predictive information.

Our results nuance claims that visual information dominates pitch evaluation (Tsay, 2021): under strict out-of-fold validation, speech embeddings generalize more reliably than video embeddings derived from sparsely sampled frames. From a signaling perspective (Spence, 1973; Colombo, 2021), canonical textual elements (traction, market size, team) do not translate into stable out-of-sample performance, suggesting pitches converge on similar content templates, limiting discriminative power. Vocal delivery may function as a higher-cost, harder-to-script signal reflecting fluency, emotional regulation, and confidence—qualities consistent with holistic investor judgment under uncertainty (Gompers et al., 2020).

Practically, entrepreneur training programs should prioritize vocal delivery—speech dynamics, confidence signaling, vocal variety—alongside content development. For investors and accelerators, audio-based assessment tools could efficiently screen large applicant pools as complementary aids. Limitations include pilot scale (N = 40), wide uncertainty intervals, and contextual heterogeneity across public competitions. Future research should scale the dataset,

Article in Press

incorporate temporal representations, and transition toward domain-adapted multimodal LLMs capable of producing structured evaluative reports, while prioritizing transparency and fairness (Shrestha and He, 2025).

References

Allison, T.H., Warnick, B.J., Davis, B.C., Cardon, M.S. (2022). Can you hear me now? Engendering passion and preparedness perceptions with vocal expressions in crowdfunding pitches. Journal of Business Venturing, 37(3), 106198.

Bafera, J., Kleinert, S. (2023). Signaling theory in entrepreneurship research: A systematic review and research agenda. Entrepreneurship Theory and Practice, 47(6), 2165–2199.

Cantamessa, M., Gatteschi, V., Perboli, G., Rosano, M. (2018). Startups' roads to failure. Sustainability, 10(7), 2346.

Chen, X., Yao, X., Kotha, S. (2009). Entrepreneur passion and preparedness in business plan presentations. Academy of Management Journal, 52(1), 199–214.

Clarke, J.S., Cornelissen, J.P., Healey, M.P. (2019). Actions speak louder than words. Academy of Management Journal, 62(2), 335–360.

Colombo, O. (2021). The use of signals in new-venture financing. Journal of Management, 47(1), 237–259.

Gavrilenko, A., et al. (2023). Machine learning prediction of startup success. Proc. AAAI, 37(4), 4123–4131.

Gompers, P.A., Gornall, W., Kaplan, S.N., Strebulaev, I.A. (2020). How do venture capitalists make decisions? Journal of Financial Economics, 135(1), 169–190.

Goossens, I., Jung, M.M., Liebregts, W., Postma, E. (2022). To invest or not to invest: Using vocal behavior to predict investor decisions. Int. Conf. Pattern Recognition, 12–25.

Huang, L., Pearce, J.L. (2015). Managing the unknowable. Administrative Science Quarterly, 60(4), 634–670.

Lundberg, S.M., Lee, S.I. (2017). A unified approach to interpreting model predictions. Advances in NeurIPS, 30.

Mahmood, A., Yeganegi, S. (2023). Linguistic certainty and entrepreneurial success. Entrepreneurship Theory and Practice, 47(2), 561–589.

Miao, C., et al. (2024). The sound of success: Speech rate and entrepreneurial financing. J. Business Venturing Insights, 21, e00420.

Parhankangas, A., Ehrlich, M. (2014). How entrepreneurs seduce business angels. Journal of Business Venturing, 29(4), 543–564.

Pennebaker, J.W. (2011). The secret life of pronouns. Bloomsbury Press.

Russell, J.A. (2003). Core affect and the psychological construction of emotion. Psychological Review, 110(1), 145–172.

Shrestha, Y.R., He, V.F. (2025). Integrating multimodal data and ML for entrepreneurship research. Strategic Entrepreneurship Journal, forthcoming.

Spence, M. (1973). Job market signaling. Quarterly Journal of Economics, 87(3), 355–374.

Article in Press

Tarillon, C., Grazzini, F., Boissin, J.P. (2023). Team signals in entrepreneurial pitches. Entrepreneurship & Regional Development, 35(3–4), 267–289.

Tong, Z., Song, Y., Wang, J., Wang, L. (2022). VideoMAE: Masked autoencoders are data-efficient learners. Advances in NeurIPS, 35, 10078–10093.

Tsay, C.J. (2021). Visuals dominate investor decisions about entrepreneurial pitches. Academy of Management Discoveries, 7(3), 343–366.

Ungerer, C., Reuther, K., Baltes, G. (2021). The lingering living dead phenomenon. J. Business Venturing Insights, 16, e00250.

Van Aken, P., Jung, M.M., Liebregts, W., Onal Ertugrul, I. (2023). Deciphering entrepreneurial pitches: A multimodal deep learning approach. Proc. ICMI, 144–152.

Wang, Y., et al. (2021). Audio mining and vocal tone in persuasion. Journal of Consumer Research, 48(2), 269–288.

Article in Press