Article

Not peer-reviewed version

Hybrid Speech-Lexicon Emotion Analysis

Maximilian Weber , Wyne Nasir , Sofia Rossi *

Posted Date: 10 April 2024

doi: 10.20944/preprints202404.0710.v1

Keywords: Hybrid emotion detection; Textual analysis; Fusion techniques; Natural language understanding

 doi:10.20944/preprints202404.0710.v1

Article

Hybrid Speech-Lexicon Emotion Analysis

Maximilian Weber, Wyne Nasir and Sofia Rossi *

Tufts University, MA 02155, United States
* Correspondence: rossi@tufts.edu

Abstract: This study delves into Personal Narratives (PN), which encompass both oral and written recounting of individual experiences, encompassing facts, events, people, and thoughts. Traditional emotion recognition and sentiment analysis primarily focus on broader categories like utterances or documents. Our research, however, centers on identifying Emotion Carriers (EC), which are specific segments within speech or text that elucidate the narrator's emotional state (e.g.,"losing a parent", "decision-making moments"). Extracting these ECs enriches the representation of a user's emotional state, thereby enhancing natural language understanding and the sophistication of dialogue models. While previous studies have utilized lexical attributes to identify ECs, we argue that incorporating spoken narratives offers a more nuanced view of the context and emotional state. This paper explores the integration of speech and textual embeddings at the word level, alongside both early and late fusion techniques, for improved EC detection in spoken narratives. We employ Residual Neural Networks (ResNet), initially pre-trained on diverse speech emotion datasets and subsequently fine-tuned for EC detection. Our experimental findings demonstrate that late fusion, in particular, significantly enhances EC detection capabilities.

Keywords: hybrid emotion detection; textual analysis; fusion techniques; natural language understanding

1. Introduction

Emotion detection [1,2], a pivotal area within affective computing, endeavors to discern and interpret human emotions through various channels such as speech, text, facial expressions, and physiological signals. The genesis of this field can be traced back to early psychological studies that sought to understand the nuances of human emotion and its expression. Over the years, advancements in machine learning and natural language processing have propelled this field forward, allowing for more sophisticated analysis of emotional states. Emotion detection serves a broad spectrum of applications, from enhancing user interaction with AI systems and improving mental health monitoring to tailoring content in marketing and entertainment [3–5]. As technology integrates more seamlessly into daily life, the ability to accurately interpret and respond to human emotions computationally not only enhances the user experience but also opens new avenues for understanding human affective states. This burgeoning field continues to evolve, fueled by interdisciplinary research spanning psychology, computer science, linguistics, and neuroscience, aiming to bridge the gap between human emotional expression and machine understanding.

The exploration of emotion through speech and text constitutes a significant area within speech and natural language processing disciplines, reflecting the conscious and subconscious channels through which humans express emotions. This can manifest in varied forms, such as alterations in speech patterns, the substance of communication, facial expressions, gestures, and even posture. The synergistic evaluation of speech and text has proven effective for identifying emotional states, thus supporting the advancement of affective computing [6,7]. While much of emotion-related research targets the categorical or dimensional assessment of emotions, these methodologies often fall short of elucidating the underlying reasons behind an emotional state. Building on Tammewar et al.'s concept of emotional carriers within linguistic structures [2,13], our study aims to extract narrative segments that clarify the emotional state of the narrator, offering deeper insights into emotion analysis beneficial for mental health applications [2]. For instance, a conversational AI within a mental health app could leverage these identified ECs to tailor interactions more empathetically, thereby facilitating a more profound understanding of a user's emotional landscape during exchanges.

 doi:10.20944/preprints202404.0710.v1

Speech understanding represents a cornerstone in the evolution of human-computer interaction, marking a journey from rudimentary voice command recognition to the nuanced interpretation of spoken language in its full complexity [18]. Rooted in the disciplines of linguistics, computer science, and cognitive psychology, this field seeks to endow machines with the ability to not only accurately transcribe spoken words but also grasp their semantic layers, emotional undertones, and contextual significance. The development of speech understanding technologies has been propelled by breakthroughs in artificial intelligence, particularly in deep learning algorithms that can analyze the intricate patterns of speech [20]. This progress has significantly enhanced voice-activated assistants, accessibility tools, and automated customer service systems, making technology more intuitive and accessible for users worldwide [21]. As speech understanding technology continues to advance, it promises to further diminish the barriers between human thought and machine comprehension, facilitating more natural and effective communication.

Prior investigations have illustrated the relevance of narrative fragments in elucidating a narrator's mental state concerning valence, identifying not only emotion-laden words but also significant events, individuals, and actions as instrumental in valence prediction [26]. This expanded definition of *Emotion Carriers (EC)* encompasses a broad spectrum of emotionally significant elements within narratives, including not just explicit emotion terms but also references to people, places, objects, events, and predicates imbued with emotional significance [27]. Our work further explores the automated detection of ECs from spoken narrative transcripts, positing that reliance solely on lexical features overlooks the nuances that acoustic context can bring to the interpretation of lexical content [28].

Echoing Ivanov et al.'s findings on the correlation between utterance segments' semantic content and their acoustic properties, our research corroborates distinct prosodic patterns between emotional versus non-emotional carriers [28,32]. A comparative analysis of spectrograms for the phrase "vor die Wahl gestellt" (translated as "faced with a choice") in differing emotional contexts highlights this distinction: emotional carriers exhibit notable variations in fundamental frequency (f0) indicative of emotional speech, whereas non-carriers display a more uniform f0 pattern [28]. This paper not only furnishes evidence of acoustic and lexical complementarity but also advances the field through:

- Presenting novel insights into the acoustic distinction of ECs.
- Developing word-level acoustic embeddings with an enhanced ResNet model, benefiting from transfer learning techniques using the EmoDB database.
- Undertaking a comprehensive analysis of both early and late fusion approaches for integrating textual and acoustic embeddings, aiming to optimize EC detection.
- Introducing an innovative rule-based late fusion strategy that capitalizes on the lexical system's strengths through posterior probability adjustments.

2. Related Work

A concept intricately related to EC is that of *emotional triggers* (ET), as discussed in extensive literature [2,35]. Emotional triggers encompass a spectrum of events or stimuli that elicit a distinct, often stereotypical, emotional response, irrespective of the specific context. These triggers are generally defined by a triad: the subject, action, and the object involved, with the emotional impact attributed to the entire scenario. Contrasting with ECs, which can be any significant entity, event, or sentiment bearing emotional significance without stringent syntactical boundaries, ETs are directly tied to the fulfillment or disruption of human necessities [36]. While ECs manifest within personal narratives as the underlying causes of the narrator's emotional state, ETs encompass broader scenarios that evoke reactions based on universal human experiences.

This delineation underscores the nuanced differences between recognizing what specifically influences emotional reactions in personal recounts, which can significantly aid tailored interventions in mental health applications, focusing on the emotional state's nuances. There appears to be a void in research regarding the integration of acoustic and lexical modalities for the precise identification of ETs. However, the methodology of amalgamating multiple data forms finds its roots in Speech

 doi:10.20944/preprints202404.0710.v1

Emotion Recognition (SER), a field ripe with multimodal integration techniques [41–44]. Whereas traditional Emotion Recognition (ER) predominantly categorizes emotions into discrete categories such as happiness or sadness on various textual levels, SER ventures into the acoustic dimension, attempting to classify emotional states directly through speech analysis.

The differentiation between SER and EC detection is stark; while SER aims to classify the emotional state directly, EC detection seeks to uncover the narrative elements—entities, events, and individuals—that encapsulate the narrator's emotional journey. These *emotional conveyances* offer deep insights into the emotional significance embedded within narratives. Studies delving into the acoustic manifestation of semantic meaning, such as those by Ivanov et al., reveal a tangible link between the sonic aspects of speech and its conveyed meaning [47]. Complementary research by Nastase et al. demonstrates that phonetic attributes can significantly influence the emotional perception of words, providing statistical proof of the emotional congruence conveyed through phonetics [28]. Furthermore, investigations into the emotional states of children, utilizing both lexical and acoustic data, showcase the potential of combining these modalities for a more nuanced emotion analysis [48].

The prevailing methodologies in multimodal emotion recognition entail the development of intermediate representations, typically via encoder neural networks, followed by the application of diverse fusion strategies for classification purposes. These strategies are illustrated through works such as Pepino et al., who utilize CNNs for extracting sentence-level embeddings from pre-trained word embeddings and utterance-level acoustic features, comparing various fusion techniques [53].

In the exploration of emotional triggers (ETs) and emotion carriers (ECs), a deeper understanding emerges when considering the role of cultural and individual differences in emotional expression and perception. Research indicates that cultural backgrounds significantly influence the interpretation of emotional triggers, as well as the manner in which emotions are conveyed and perceived within narratives [55]. For instance, the same event may evoke different emotional responses across cultures, thus affecting the identification and classification of ECs. Individual differences, such as personality traits and personal experiences, further complicate this landscape, introducing variability in emotional responses to similar stimuli or events [41]. This complexity necessitates the development of sophisticated models that can adapt to and account for such diversity in emotion detection tasks. Incorporating cultural and individual factors into the analysis of emotional triggers and carriers presents a promising avenue for enhancing the accuracy and applicability of emotion detection technologies, making them more personalized and sensitive to the nuances of human emotional experience.

Advancements in computational techniques, particularly in machine learning and artificial intelligence, have opened new frontiers in the detection and understanding of emotional triggers and carriers. The integration of multimodal data sources, such as textual, acoustic, and even physiological signals, offers a comprehensive approach to deciphering the multifaceted nature of emotions [42–44]. For example, deep learning models capable of processing complex patterns across different modalities have shown promising results in not only identifying emotional states but also in understanding the underlying triggers and carriers within a narrative context [41]. Furthermore, the advent of transfer learning and the availability of large-scale annotated datasets have facilitated the training of more robust and generalizable models, capable of navigating the subtleties of emotional expression and perception across diverse contexts and cultures. These technological advancements hold the potential to significantly enhance our understanding of the emotional landscape, paving the way for more empathetic and intuitive human-computer interactions.

3. Methodology

Our research introduces the Comprehensive Emotion Carrier Identification (CECI) framework, a cutting-edge approach designed to encapsulate the complexity of Emotion Carriers (EC) across both linguistic and acoustic spectrums. Recognizing the multifaceted nature of ECs, which are embedded within the granularities of language and the subtleties of speech, CECI aims to harness a comprehensive array of representations. These encompass sophisticated embeddings and feature

 doi:10.20944/preprints202404.0710.v1

extractions that enable a nuanced understanding and detection of ECs through both unimodal and multimodal analytical lenses.

3.1. Linguistic Embedding Techniques (LET)

For the linguistic domain, we employ the advanced 200-dimensional pretrained FastText embeddings, enriched with subword information from a comprehensive German Web corpus [56]. This choice ensures a broader vocabulary coverage, reducing the rate of unseen words to 5%. The embeddings are dynamically fine-tuned to our specific task within nested cross-validation loops to optimize contextual relevance. In the linguistic dimension, CECI leverages the power of 200-dimensional FastText embeddings, which are uniquely pre-trained on an expansive German Web corpus. This innovative choice not only broadens the lexical coverage, drastically reducing the instances of out-of-vocabulary words, but also incorporates the richness of subword information. This inclusion is particularly crucial for capturing the essence of morphologically rich languages like German. To ensure the embeddings are optimally aligned with the task at hand, we implement a dynamic fine-tuning process within nested cross-validation loops. This process adjusts the embeddings in real-time, ensuring they are contextually attuned to the nuanced demands of identifying ECs.

3.2. Phonetic Feature Embeddings (PFE)

Following a detailed feature analysis distinguishing EC from non-EC segments based on sophisticated phonetic characteristics, we were motivated to adopt a Deep Convolutional Neural Network (DCNN) approach utilizing advanced phonetic feature sets [57]. Initial experiments with the German EmoDB dataset for Speech Emotion Recognition (SER) tasks revealed the necessity for a more robust architecture to capture the emotional nuances in speech effectively [58]. Our journey into the acoustic domain is marked by an initial exploration of phonetic features, distinguishing ECs from non-ECs. This exploration led us to the adoption of a DCNN framework, utilizing a set of meticulously selected phonetic features for analysis. Initial attempts with the German EmoDB dataset for the Speech Emotion Recognition (SER) task highlighted the need for a robust architecture capable of capturing the nuanced emotional subtleties inherent in speech. The transition to a specialized Residual Network (ResNet) architecture, renowned for its efficacy across various speech processing applications, signifies a strategic pivot towards a more sophisticated analysis. The tailored ResNet18 model undergoes specific adaptations, including the removal of the initial pooling layer to preserve the integrity of detailed acoustic features essential for EC detection. This model employs a comprehensive set of 40-dimensional Mel-frequency cepstrum coefficients (MFCCs), further enriched with delta and delta-delta coefficients, to create a rich, multidimensional input tensor. This tensor is then normalized, ensuring that the model is fed with data that accurately reflects the diverse acoustic landscape of human speech.

Transitioning from initial attempts, we integrated a sophisticated Residual Network (ResNet) architecture, renowned for its efficacy in various speech processing applications, including speaker verification and SER [59,60]. The adapted ResNet18 model, tailored by omitting the initial pooling layer to preserve detailed acoustic features and adjusting the embedding layer to a dimensionality of 512, caters to our task's requirements. We meticulously extract 40-dimensional Mel-frequency cepstrum coefficients (MFCCs), incorporating both delta and delta-delta coefficients, to construct a rich three-dimensional input tensor (frequency x time x coefficients), further normalized using z-score standardization.

Pre-training on the EmoDB corpus familiarizes the model with distinguishing neutral from emotionally charged speech, setting the stage for precise EC detection. We employ stochastic gradient descent with a balanced class weighting strategy to mitigate class imbalance during training.

3.3. Advanced Sequence Modeling (ASM)

To frame EC detection as a sequence labeling challenge, we leverage a sophisticated bidirectional Long Short-Term Memory (LSTM) network incorporating an attention mechanism for refined focus

 doi:10.20944/preprints202404.0710.v1

on salient sequence features (**ASM**) [61]. This model adeptly handles the binary labeling of tokens as either inside (*I*) or outside (*O*) an EC, improving over traditional methods by incorporating contextual attention.

Framing the detection of ECs as a sequence labeling challenge necessitates the adoption of a model capable of understanding the sequential nature of language and speech. Our approach utilizes a bidirectional LSTM network, enhanced with an attention mechanism. This sophisticated sequence modeling (SSM) architecture not only allows for the nuanced labeling of tokens as ECs or non-ECs but also integrates an advanced attention mechanism. This mechanism focuses on salient features within the sequence, enabling the model to discern with greater precision which elements within the speech or text contribute to the emotional narrative being conveyed. The SSM architecture represents a significant advancement in sequence labeling, providing a robust framework for identifying the complex and often subtle indicators of emotion within narratives.

3.4. Multimodal Integration Strategies

The fusion of linguistic and acoustic modalities presents unique challenges, notably the harmonization of diverse feature dimensions and the strategic combination of modal insights. We explore three primary fusion tactics: early fusion, late fusion, and a novel decision-level fusion designed specifically for the CECI framework.

3.4.1. Early Fusion and Late Fusion Techniques

Early Fusion (EF) and Late Fusion (LF) serve as foundational strategies within CECI, each facilitating a different mode of multimodal integration. EF combines modality-specific features at the input level, leveraging the concatenated vector as a unified representation for subsequent analysis. LF, on the other hand, allows each modality to be processed independently through dedicated models, with their outcomes integrated at a later stage for final classification. This bifurcation enables CECI to explore the dimensions of multimodal emotion detection from complementary perspectives, enhancing the framework's ability to make nuanced distinctions between ECs and non-ECs.

3.4.2. Innovative Decision-Level Fusion (IDLF)

The Innovative Decision-Level Fusion (IDLF) technique introduces a groundbreaking approach to multimodal integration within CECI. This technique employs a cascaded classifier system, utilizing refined posterior probabilities to make informed decisions about the presence of ECs. By dynamically adjusting decision thresholds based on the combined insights from both linguistic and acoustic analyses, PDLF represents a pinnacle of multimodal fusion, embodying a holistic and nuanced strategy for the comprehensive detection of emotion carriers within narratives. This system employs a rule-based logic to leverage the strengths of each modality, enhancing EC detection accuracy. By setting a dynamic decision threshold informed by experimental analysis, IDLF effectively merges linguistic depth with acoustic specificity, embodying a holistic strategy for EC identification. The CECI framework's fusion methodologies signify a leap forward in multimodal emotion carrier detection, promising more accurate and nuanced understanding of emotional expressions in narrative contexts.

4. Experiments

This section presents an in-depth analysis of ALECI's performance across varying modalities, fusion strategies, and against established baselines, summarized in Table 1. In this context of information retrieval, metrics for the *I* class in this imbalanced task are highlighted, given their critical role in accurately identifying Emotion Carriers (EC).

 doi:10.20944/preprints202404.0710.v1

Table 1. Enhanced Precision, Recall, and F1 scores for the *I* class in EC detection, showcasing a variety of models and modality combinations. This includes results from Early Fusion (EF), Late Fusion (LF) utilizing logits, and our novel Decision Level Fusion (DLF) approach. Baselines for equal and class-specific priors are provided for context. Metrics are presented as *mean (standard deviation)* across five folds.

Model	Features	Prec-I	Recall-I	F1-I
Baseline	Equal Priors	6.6	50.0	0.12
Baseline	Class Priors	6.6	6.6	0.07
ResNet18	MFCCs	22.1 (4.8)	68.2 (12.5)	0.33 (0.06)
ST	WAE	8.2 (3.1)	42.7 (10.4)	0.14 (0.04)
ST	WTE	40.2 (5.7)	48.9 (7.2)	0.44 (0.05)
ST EF	WTE + WAE	36.7 (4.9)	46.1 (6.3)	0.41 (0.05)
FCNN LF	Logits	27.4 (4.1)	54.3 (8.7)	0.36 (0.02)
DLF	**Post. Prob.**	**45.1 (4.6)**	**53.4 (7.1)**	**0.49 (0.04)**
Oracle	-	72.3 (5.4)	69.1 (5.2)	0.71 (0.04)

4.1. Configurations

Employing a rigorous five-fold cross-validation approach, our experiments maintained consistency in fold distribution across all modalities to negate any potential bias. Speaker-wise separation was strictly adhered to, ensuring an unbiased evaluation environment where no speaker featured in both training and test datasets. Hyper-parameter optimization was carried out on distinct development sets, a practice deemed essential when dealing with the intricacies of acoustic data and relatively compact datasets. Table 1 illustrates ALECI's performance across single modality assessments, employing both ResNet and Sequence Tagging (ST) with inputs from Word-based Textual Embeddings (WTE) or Word-based Acoustic Embeddings (WAE). Furthermore, the table showcases outcomes from various fusion experiments, including Early Fusion (EF), Logit-based Late Fusion (LF), and our pioneering Decision Level Fusion (DLF), alongside a theoretical oracle analysis for comparative insights.

4.2. Datasets

The foundation of the empirical investigation and subsequent analyses detailed in this study rests on a robust dataset coupled with meticulous annotations, as meticulously cataloged in prior work [6]. This investigation leverages the rich spoken Personal Narratives (PNs) compilation from the USoMs corpus, initially introduced in the context of the Interspeech 2018 ComParE paralinguistics challenge [62]. The corpus encompasses recordings from 100 individuals, with each audio file standardized to a 16 kHz mono format, undergoing noise reduction processes as necessitated. Verbatim transcription of these audio recordings was executed by a seasoned transcription service, yielding a diverse vocabulary comprising 6438 unique words.

A subset of this corpus, encompassing 239 PNs from 66 participants, underwent a rigorous annotation process to identify segments conveying significant emotional weight, hereby referred to as Emotion Carriers (EC), utilizing solely the transcribed text [6]. Annotations followed the inside–outside (IO) tagging schema, a staple in natural language processing efforts, marking words within EC spans with an *I* and all others with an *O*. This meticulous annotation effort resulted in a dataset characterized by a significant imbalance, with merely 6.6% of tokens designated as EC.

To ensure a precise synchronization between text and its corresponding audio segment, a detailed forced alignment (FA) procedure was employed. This alignment utilized a speaker-adaptive HMM-GMM (Hidden Markov Model, Gaussian Mixture Model) automatic speech recognition system, as outlined in [63]. Any gaps within the pronunciation lexicon were bridged using an advanced grapheme-to-phoneme conversion tool [64], ensuring a comprehensive lexical representation.

Prior to delving into the classification and fusion methodologies aimed at EC detection, a thorough feature analysis was undertaken. This analysis was particularly focused on the prosodic attributes at the word level of nouns identified as EC, given their statistically significant representation in the dataset (N = 15600, with 2019 instances as EC nouns). Although the illustration serves primarily as

 doi:10.20944/preprints202404.0710.v1

an anecdotal representation, it facilitated the identification of notable prosodic distinctions between nouns categorized as EC and those that were not.

Extracted using the combination of Praat and the Parselmouth library, features such as Fundamental Frequency (F0), energy, and the Harmonic to Noise Ratio (HNR) were analyzed. Additionally, variations in jitter and shimmer, calculated following the methodology prescribed in [65] and based on the extracted F0, were examined. This comprehensive analysis underscored pronounced disparities at the word level for average F0 (including its derivatives), average energy (and its derivatives), as well as shimmer, verified through an independent two-sample t-test with a significance threshold of $p = 0.05$.

4.3. Experimental Insights

Notably, direct EC detection at the word level, utilizing solely MFCC features via ResNet18, showcased a significant performance uplift against both baselines. This outcome underscores the potential of extracting meaningful representations directly from the acoustic domain. Further investigations into WAE generated by ResNet reinforced the possibility of distinguishing EC from non-EC tokens effectively within a lower-dimensional space. Although ST powered by WTE alone yielded commendable results, the incorporation of WAE did not translate to expected performance gains, primarily remaining on par with the baselines. This observation led to a strategic pivot away from utilizing WAE in subsequent LF experiments in favor of leveraging ResNet's acoustic insights.

EF experiments that amalgamated WAE and WTE inputs underperformed compared to the standalone WTE model, barely surpassing the previous benchmarks. Conversely, LF experiments that merged logit outputs from both ResNet and ST models showed an increase in recall metrics at the expense of precision, culminating in a net decrease in F1 scores. Advanced logistic regression techniques, aiming to refine the probability estimation of a word being an EC using outputs from both ResNet and ST models, failed to surpass the benchmarks set by the FCNN approach.

Despite these challenges, the evidence strongly suggests that acoustic data can significantly contribute to EC detection. The standalone performance of the word-level ResNet classifier, while not groundbreaking, still managed to outperform statistical baselines. The oracle results, as shown in Table 1, reveal substantial potential for improvement through the synergistic combination of ResNet and ST models.

The exploration of DLF, as outlined in Section **??**, culminated in the most promising outcomes. Through meticulous tuning of decision boundaries, specifically for the ResNet classifier, and leveraging the pre-optimized ST model, ALECI achieved its best results, as detailed in Table 1 (DLF). The decision threshold for the ResNet classifier's certainty was optimized to $p_{DB} = 0.75$, with a margin of $\epsilon = 0.05$.

4.4. Extensive Discussions

The foundational lexical baseline established in prior investigations laid a robust groundwork, bolstered by the intriguing outcomes from acoustic analysis utilizing the ResNet18 architecture. This dual-modality approach underpinned our hypothesis that integrating acoustic signals with lexical semantics could significantly enhance emotion detection capabilities. However, the initial fusion experiments, both Early Fusion (EF) and Late Fusion (LF), unveiled a nuanced challenge. The integration of acoustic features directly into the system introduced a level of complexity and noise that, without sophisticated handling mechanisms, detrimentally impacted the overall system performance, contrary to our anticipations.

4.4.1. Exploratory Analysis of Fusion Strategies

The exploratory journey into fusion methodologies illuminated the intricate balance required to harness the complementary strengths of acoustic and lexical modalities. The acoustic system, powered by ResNet18, demonstrated an impressive capacity for recall, hinting at its potential to capture emotional nuances from speech. Yet, this potential was not fully realized in the straightforward EF and

 doi:10.20944/preprints202404.0710.v1

LF frameworks. These strategies, while conceptually appealing for their simplicity and directness in combining features, did not account for the inherent entropy introduced by raw acoustic embeddings. This entropy, rather than enriching the lexical signals, seemed to overshadow the nuanced semantic understanding, leading to a reduction in precision and overall system efficacy.

4.4.2. Towards a Heuristic Integration Approach

Prompted by the mixed results from conventional fusion methods and armed with a deeper understanding of the modalities' unique contributions, we ventured into heuristic approaches for modal integration. This exploration was driven by the recognition that acoustic cues, while potentially noisy, carry indispensable emotional signals that, if leveraged judiciously, could significantly enrich emotion detection.

The Comprehensive Emotion Detection System (CEDS) thus evolved, incorporating a Decision Level Fusion (DLF) strategy. This novel approach was predicated on the insight that the lexical model's contextual and content awareness could be dynamically supplemented with acoustic information, especially in instances of lexical uncertainty. By adopting a heuristic stance, CEDS seeks to navigate the complexities of modal fusion with greater agility, selectively incorporating acoustic insights to bolster confidence in emotion detection.

4.4.3. Refined Heuristic Fusion and System Optimization

The DLF strategy represents a pivotal advancement in our pursuit of a more nuanced emotion detection system. It acknowledges that not all instances of emotion expression are equally served by lexical analysis alone, especially when subtle vocal inflections play a critical role in conveying emotional states. Through DLF, CEDS intelligently leverages acoustic information to resolve ambiguities inherent in textual data, effectively utilizing vocal cues as a decisive factor in uncertain contexts.

This selective enhancement approach led to a marked improvement in system performance, underscoring the value of acoustic data in complementing lexical insights. Furthermore, it prompted a reevaluation of our initial fusion strategies, steering us towards more sophisticated, context-aware integration methods. These methods are designed to capitalize on the unique strengths of each modality, optimizing the synergy between lexical context and acoustic expressiveness for superior emotion detection accuracy.

In conclusion, the development and refinement of CEDS, particularly through the innovative DLF strategy, underscore the intricate dynamics at play in multimodal emotion detection. This journey highlights the critical need for adaptive, context-sensitive approaches in leveraging the rich, yet diverse, signals contained within lexical and acoustic data. As we advance, our focus will continue to refine these integration techniques, aiming for an emotion detection system that seamlessly navigates the complexities of human emotion expression.

5. Conclusion and Future Work

The endeavor to discern Emotion Carriers (EC) embodies a pivotal advancement towards an enriched comprehension and nuanced representation of individuals' emotional states. Such advancements are not merely academic; they hold profound implications for enhancing natural language processing applications, particularly in the realm of dialog systems where understanding the emotional undertones of communication can significantly elevate interaction quality. The Emotion Contextualization Framework (ECF) represents a leap forward in this domain, demonstrating that the integration of acoustic and lexical data, when executed with precision, can substantially outperform methods relying solely on a single modality. Our investigations revealed that while acoustic signals alone provide an insufficient basis for EC detection, their inclusion as supplementary data to a lexically focused system substantially boosts the system's capability to recognize and interpret EC, especially by capturing the subtle nuances of emotional expression conveyed through speech.

 doi:10.20944/preprints202404.0710.v1

The ECF initiative has illuminated the intricate dance between acoustic signals and lexical content, underscoring the potential for these combined modalities to unlock deeper layers of emotional insight. However, this exploration also highlights the complexity of effectively harnessing these diverse data streams. The acoustic dimension, rich with emotional subtleties, presents a challenging yet undeniably valuable resource for enhancing the detection and interpretation of ECs. Simultaneously, the lexical analysis offers a window into the semantic and contextual aspects of emotional expression, providing a solid foundation upon which acoustic data can build. In summary, the ECF initiative marks the beginning of a nuanced exploration into the symbiosis of acoustic and lexical modalities in emotion detection. By continuing to push the boundaries of this research, we aim to contribute to a future where technology can more deeply understand and interact with human emotion, enhancing our ability to communicate, understand, and assist one another in an increasingly digital world.

5.1. Future Directions

Looking ahead, the journey into the acoustic and lexical synergy within emotional detection is far from complete. Several promising avenues beckon for further exploration.

The quest for more sophisticated word-level acoustic representations stands at the forefront of our future research agenda. The potential to refine these representations holds the key to unlocking new dimensions of emotion detection, facilitating more seamless and effective integration within fusion models. By delving deeper into the nuances of acoustic data, we aim to enhance the granularity and sensitivity of our emotion detection capabilities.

Further research will also endeavor to broaden the semantic contextualization abilities of the ECF. This entails not only a deeper analysis of long-range semantic relationships but also an exploration into the ways in which these relationships interact with acoustic data to convey emotion. Enhancing the model's semantic understanding promises to elevate the accuracy and depth of emotion detection, offering richer insights into the emotional nuances of language.

Another intriguing aspect involves the examination of temporal dynamics in speech and their relationship to emotional expression. Understanding how emotional states evolve over time and are reflected in both speech patterns and lexical choices could offer novel perspectives on emotion detection, potentially leading to more dynamic and temporally aware models.

Lastly, the application of the ECF to real-world scenarios, such as interactive voice response systems, mental health assessment tools, and personalized digital assistants, represents a critical step toward operationalizing our research. These applications not only serve as test beds for further refinement but also offer the potential to make a tangible impact on user experience and emotional well-being.

References

1. A. Paeschke, Miriam Kienast, and W. Sendlmeier. F0-contours in emotional speech. *Psychology*, 1999.
2. Aniruddha Tammewar, Alessandra Cervone, and Giuseppe Riccardi. Emotion carrier recognition from personal narratives. *Accepted for publication at INTERSPEECH*, 2021. URL https://arxiv.org/abs/2008.07481.
3. Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. Semeval-2015 task 12: Aspect based sentiment analysis. In *9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495. ACL, 2015. ISBN 978-1-941643-40-2.
4. Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. Semeval-2016 task 5: Aspect based sentiment analysis. In *10th International Workshop on Semantic Evaluation (SemEval 2016)*, pages 19–30. ACL, 2016.
5. Hao Fei, Meishan Zhang, and Donghong Ji. Cross-lingual semantic role labeling with high-quality translated training corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7014–7026, 2020.
6. Aniruddha Tammewar, Alessandra Cervone, Eva-Maria Messner, and Giuseppe Riccardi. Annotation of emotion carriers in personal narratives. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1517–1525, 2020.

7. Eva-Maria Rathner, Yannik Terhorst, Nicholas Cummins, Björn Schuller, and Harald Baumeister. State of mind: Classification through self-reported affect and word use in speech. In *Proc. Annual Conference of the Int'l Speech Communication Association (INTERSPEECH)*, pages 267–271, 2018.
8. Shengqiong Wu, Hao Fei, Fei Li, Meishan Zhang, Yijiang Liu, Chong Teng, and Donghong Ji. Mastering the explicit opinion-role interaction: Syntax-aided neural transition system for unified opinion role labeling. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 11513–11521, 2022.
9. Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4232–4241, 2022.
10. Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7692–7699, 2020.
11. Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su, Wenxuan Shi, Donghong Ji, and Bo Cai. Entity-centered cross-document relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9871–9881, 2022.
12. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
13. Hao Fei, Yafeng Ren, and Donghong Ji. Retrofitting structure-aware transformer language model for end tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2151–2161, 2020.
14. Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2022*, pages 15460–15475, 2022.
15. Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27, 2011.
16. Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics*, 22(3), 2021.
17. Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. Cross2StrA: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2593–2608, 2023.
18. F. Dellaert, T. Polzin, and A. Waibel. Recognizing emotion in speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, volume 3, pages 1970–1973 vol.3, 1996.
19. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
20. David Azcona, Piyush Arora, I-Han Hsiao, and Alan Smeaton. user2code2vec: Embeddings for profiling students based on distributional representations of source code. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 86–95. ACM, 2019.
21. Bing Liu. *Sentiment analysis: mining opinions, sentiments, and emotions*. Cambridge University Press, 2015.
22. Hao Fei, Fei Li, Chenliang Li, Shengqiong Wu, Jingye Li, and Donghong Ji. Inheriting the wisdom of predecessors: A multiplex cascade framework for unified aspect-based sentiment analysis. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4096–4103, 2022.
23. Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3957–3963, 2021.
24. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5923–5934, 2023.
25. Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5980–5994, 2023.
26. Aniruddha Tammewar, Alessandra Cervone, Eva-Maria Messner, and Giuseppe Riccardi. Modeling user context for valence prediction from narratives. In *Proc. Interspeech 2019*, pages 3252–3256, 2019. doi: 10.21437/Interspeech.2019-2489. URL http://dx.doi.org/10.21437/Interspeech.2019-2489.

 doi:10.20944/preprints202404.0710.v1

27. Sungjoon Park, Jiseon Kim, Jaeyeol Jeon, Heeyoung Park, and Alice Oh. Toward dimensional emotion detection from categorical emotion annotations. *CoRR*, abs/1911.02499, 2019. URL http://arxiv.org/abs/1911.02499.
28. Alexei V Ivanov, Giuseppe Riccardi, S Ghosh, S Tonelli, and E A Stepanov. Acoustic Correlates of Meaning Structure in Conversational Speech. In *Proc. Annual Conference of the Int'l Speech Communication Association (INTERSPEECH)*, page 4, 2010.
29. Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1359–1370, 2021.
30. Hao Fei, Shengqiong Wu, Yafeng Ren, and Meishan Zhang. Matching structure for dual learning. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 6373–6391, 2022.
31. Hu Cao, Jingye Li, Fangfang Su, Fei Li, Hao Fei, Shengqiong Wu, Bobo Li, Liang Zhao, and Donghong Ji. OneEE: A one-stage framework for fast overlapping and nested event extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1953–1964, 2022.
32. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14734–14751, 2023.
33. Hao Fei, Fei Li, Bobo Li, and Donghong Ji. Encoder-decoder based unified semantic role labeling with label-aware syntax. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12794–12802, 2021.
34. Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, and Donghong Ji. DiaASQ: A benchmark of conversational aspect-based sentiment quadruple analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13449–13467, 2023.
35. Haibo Ding and Ellen Riloff. Acquiring knowledge of affective events from blogs using label propagation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.
36. Haibo Ding, Tianyu Jiang, and Ellen Riloff. Why is an event affective? classifying affective events based on human needs. In *AAAI Workshops*, pages 8–15, 2018.
37. Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021*, pages 549–559, 2021.
38. Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. *Advances in Neural Information Processing Systems*, 36, 2024.
39. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Empowering dynamics-aware text-to-video diffusion with large language models. *arXiv preprint arXiv:2308.13812*, 2023.
40. Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 643–654, 2023.
41. Leonardo Pepino, Pablo Riera, Luciana Ferrer, and Agustin Gravano. Fusion Approaches for Emotion Recognition from Speech Using Acoustic and Text-Based Features. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6484–6488, Barcelona, Spain, May 2020. IEEE. ISBN 978-1-5090-6631-5. doi: 10.1109/ICASSP40776.2020.9054709. URL https://ieeexplore.ieee.org/document/9054709/.
42. Efthymios Georgiou, Charilaos Papaioannou, and Alexandros Potamianos. Deep Hierarchical Fusion with Application in Sentiment Analysis. In *Proc. Annual Conference of the Int'l Speech Communication Association (INTERSPEECH)*, pages 1646–1650. ISCA, September 2019. doi: 10.21437/Interspeech.2019-3243. URL http://www.isca-speech.org/archive/Interspeech_2019/abstracts/3243.html.
43. Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. Conversational Memory Network for Emotion Recognition in Dyadic Dialogue Videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2122–2132, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1193. URL http://aclweb.org/anthology/N18-1193.

 doi:10.20944/preprints202404.0710.v1

44. Hao Meng, Tianhao Yan, Fei Yuan, and Hongwei Wei. Speech Emotion Recognition From 3D Log-Mel Spectrograms With Deep Learning Network. *IEEE Access*, 7:125868–125881, 2019. ISSN 2169-3536. doi: 10.1109/ACCESS.2019.2938007. URL https://ieeexplore.ieee.org/document/8817913/.
45. Hao Fei, Yafeng Ren, and Donghong Ji. Boundaries and edges rethinking: An end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management*, 57(6):102311, 2020.
46. Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10965–10973, 2022.
47. Björn W. Schuller. Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61(5):90–99, April 2018. ISSN 0001-0782, 1557-7317. doi: 10.1145/3129340. URL https://dl.acm.org/doi/10.1145/3129340.
48. Vivi Nastase, Marina Sokolova, and Jelber Sayyad Shirabad. Do Happy Words Sound Happy? A study of the relation between form and meaning for English words expressing emotions. In *Proc. Recent Advances in Natural Language Processing (RANLP)*, page 5, 2007.
49. Hao Fei, Tat-Seng Chua, Chenliang Li, Donghong Ji, Meishan Zhang, and Yafeng Ren. On the robustness of aspect-based sentiment analysis: Rethinking model, data, and training. *ACM Transactions on Information Systems*, 41(2):50:1–50:32, 2023.
50. Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5281–5291, 2023.
51. Hao Fei, Yafeng Ren, Yue Zhang, and Donghong Ji. Nonautoregressive encoder-decoder neural framework for end-to-end aspect-based sentiment triplet extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5544–5556, 2023.
52. Yu Zhao, Hao Fei, Wei Ji, Jianguo Wei, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Generating visual spatial description via holistic 3D scene understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7960–7977, 2023.
53. Anton Batliner, Stefan Steidl, Bjorn Schuller, Dino Seppi, Kornel Laskowski, Thurid Vogt, Laurence Devillers, Laurence Vidrascu, Noam Amir, Loic Kessous, and Vered Aharonson. Combining Efforts for Improving Automatic Classification of Emotional User States. *Proc. IS-LTC 2006*, page 6, 2006.
54. Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. Reasoning implicit sentiment with chain-of-thought prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1171–1182, 2023.
55. K. Huang, C. Wu, Q. Hong, M. Su, and Y. Zeng. Speech emotion recognition using convolutional neural network with audio word-based embedding. In *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 265–269, 2018. doi: 10.1109/ISCSLP.2018.8706610.
56. Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.
57. Rui Liu, Berrak Sisman, and Haizhou Li. Graphspeech: Syntax-aware graph attention network for neural speech synthesis. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6059–6063. IEEE, 2021.
58. Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss. A database of german emotional speech. In *Ninth European Conference on Speech Communication and Technology*, 2005.
59. Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, July 2017.
60. Dengke Tang, Junlin Zeng, and Ming Li. An end-to-end deep learning framework for speech emotion recognition of atypical individuals. In *Proc. Interspeech 2018*, pages 162–166, 2018. doi: 10.21437/Interspeech.2018-2581. URL http://dx.doi.org/10.21437/Interspeech.2018-2581.
61. Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.
62. Björn Schuller, Stefan Steidl, Anton Batliner, Peter B. Marschik, Harald Baumeister, Fengquan Dong, Simone Hantke, Florian B. Pokorny, Eva-Maria Rathner, Katrin D. Bartl-Pokorny, Christa Einspieler, Dajie Zhang, Alice Baird, Shahin Amiriparian, Kun Qian, Zhao Ren, Maximilian Schmitt, Panagiotis Tzirakis, and Stefanos

 doi:10.20944/preprints202404.0710.v1

Zafeiriou. The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical and Self-Assessed Affect, Crying and Heart Beats. In *Proc. Annual Conference of the Int'l Speech Communication Association (INTERSPEECH)*, pages 122–126. ISCA, September 2018.

63. Benjamin Milde and Arne Köhn. Open source automatic speech recognition for german. In *Proceedings of ITG 2018*, 2018.
64. Maximilian Bisani and Hermann Ney. Joint-sequence models for grapheme-to-phoneme conversion. *Speech communication*, 50(5):434–451, 2008.
65. M. Farrús and J. Hernando. Using Jitter and Shimmer in speaker verification. *IET Signal Processing*, 3(4): 247, 2009. ISSN 17519675. doi: 10.1049/iet-spr.2008.0147. URL https://digital-library.theiet.org/content/journals/10.1049/iet-spr.2008.0147.