



# Prosody and Paralinguistic Features as Meaning-Making Resources in English–Vietnamese Film Dialogue: A Multimodal Framework

Pham Thao Ngoc<sup>1</sup>, Vuong Hong Hanh<sup>\*2</sup>, Nguyen Phuong Mai<sup>3</sup>

<sup>1</sup>Hanoi University of Industry, 298 Cau Dien Street, Hanoi, Vietnam

<sup>\*1, 3</sup> Vietnam Education Strategy and Policy Institute, 101 Tran Hung Dao Street, Hanoi, Vietnam.

## Abstract

*This study proposes an integrated multimodal framework for analysing prosody and paralinguistic features as meaning-making resources in English–Vietnamese film dialogue. The research addresses an important gap in previous studies, which have often examined intonation, lexical tone, or vocal expression separately rather than as coordinated semiotic resources in spoken interaction. The framework combines Systemic Functional Phonology, Van Leeuwen's social-semiotic approach to sound, and multimodal communication research to examine how prosodic, paralinguistic, and bodily-behaviour resources jointly construct interpersonal meaning. A comparative qualitative corpus consisting of two English-language films (*The King's Speech* and *Marriage Story*) and two Vietnamese-language films (*Tro Tàn Rực Rỡ* and *Bên Trong Vò Kén Vàng*) was analysed using ELAN-based multimodal annotation. The study develops a four-layer coding architecture encompassing prosodic, paralinguistic, bodily-behaviour, and meaning layers, and introduces a lexical-tone protection principle to distinguish lexical-tone realisation from discourse-level prosodic interpretation in Vietnamese. The findings suggest that English and Vietnamese frequently realise comparable interpersonal functions through different multimodal configurations. In English, prominence and intonational organisation appear to carry a relatively greater functional load, whereas in Vietnamese, pitch range, tempo, pausing, voice quality, and bodily behaviour interact more extensively within the constraints imposed by lexical tone. The study also distinguishes four analytically independent but interacting domains: natural prosody, poetic rhythm, lexical tone, and paralinguistic meaning. The proposed framework contributes to research on multimodal discourse analysis, English–Vietnamese prosodic comparison, and audiovisual communication, and has implications for EFL pedagogy, intercultural communication, and audiovisual translation. The study argues that a multimodal perspective provides a more comprehensive account of spoken meaning than approaches that treat prosody as an isolated phonological phenomenon.*

**Keywords:** Multimodal communication; prosody; paralinguistic features; English–Vietnamese comparison; interpersonal meaning

Received 25 Sep., 2026; Revised 03 Oct., 2026; Accepted 05 Oct., 2026 © The author(s) 2026. Published with open access at [www.questjournals.org](http://www.questjournals.org)

## I. INTRODUCTION

A recurring pattern in English-language classrooms in Vietnam is that learners often understand the lexical content of an utterance accurately while misinterpreting its interpersonal meaning. A brief acknowledgment such as *okay* or *fine* may be understood as agreement even when it functions as irony, resistance, emotional withdrawal, or restrained opposition (Halliday & Greaves, 2008; Tench, 2011). These interpretations are not determined by words alone but by the coordinated deployment of pitch movement, prominence, loudness, speech tempo, voice quality, facial expression, gaze, and gesture (Van Leeuwen, 1999; Norris, 2004). The

<sup>1</sup>Hanoi University of Industry, 298 Cau Dien Street, Hanoi, Vietnam

<sup>\*2, 3</sup> Vietnam Education Strategy and Policy Institute, 101 Tran Hung Dao Street, Hanoi, Vietnam.  
Corresponding author: [hanhvh@vnies.edu.vn](mailto:hanhvh@vnies.edu.vn)

difficulty therefore lies not simply in linguistic decoding, but in the integration of multiple semiotic resources that jointly construct spoken meaning. From a multimodal perspective, spoken interaction is the simultaneous orchestration of vocal, visual, and bodily resources, each contributing systematically to the construction of meaning (Kress & Van Leeuwen, 2001; Norris, 2004).

Existing research addresses this phenomenon from several complementary perspectives. Systemic Functional Phonology, developed by Halliday (1967) and further elaborated by Halliday and Greaves (2008), conceptualises intonation as a grammatical resource that organises information, discourse structure, and interpersonal meaning. This tradition provides a powerful account of prosodic organisation through the systems of tonality, tonicity, and tone (Halliday, 1967; Halliday & Greaves, 2008). However, the framework gives relatively limited analytical attention to the interaction between prosody and co-occurring bodily behaviour. A second tradition, represented by Van Leeuwen's Sound Semiotics (1999) social-semiotic approach to sound, treats sound as a socially regulated semiotic resource whose meanings emerge through culturally recognisable patterns of pitch, loudness, tempo, and voice quality. Although this perspective offers a rich vocabulary for the social interpretation of sound, it is not designed as a systemic description of prosodic organisation (Van Leeuwen, 1999). A third line of research investigates sound symbolism and phonosemantic meaning, focusing primarily on relationships between acoustic form and lexical meaning rather than on the multimodal construction of interactional meaning (Hinton et al., 1994; Dingemanse, 2012). Together these traditions illuminate important aspects of spoken communication, but none of them alone fully explains how prosodic, paralinguistic, and bodily resources jointly construct meaning in English–Vietnamese film dialogue.

This study adopts a multimodal communication perspective that integrates insights from Systemic Functional Phonology and Sound Semiotics within a unified analytical framework (Kress & Van Leeuwen, 2001; Halliday & Greaves, 2008). The framework draws on recent multimodal work that treats spoken interaction as the coordinated operation of prosodic, paralinguistic, and bodily-behaviour channels, interpreted primarily through the experiential and interpersonal metafunctions. Rather than replacing existing theories, the study positions them as complementary analytical resources: Halliday provides the grammatical description of prosodic organisation, Van Leeuwen provides the social-semiotic vocabulary for interpreting sound, and the multimodal framework provides the mechanism through which vocal and visual resources are analysed as integrated meaning-making systems.

The study makes three contributions. First, it proposes a four-channel coding framework that systematically maps prosodic, paralinguistic, and bodily resources onto experiential and interpersonal meanings (Halliday, 1967; Halliday & Greaves, 2008) across English and Vietnamese dialogue. Second, it distinguishes four phenomena that are frequently conflated in research on Vietnamese speech: natural prosody, poetic rhythm, lexical tone, and paralinguistic meaning. Third, it develops a comparative framework for analysing English and Vietnamese film dialogue that has implications for EFL pedagogy, intercultural communication, and audiovisual translation. By considering prosody and paralinguistic behaviour not as separate phonetic features but as integrated semiotic resources (Kress & Van Leeuwen, 2001; Van Leeuwen, 1999; Norris, 2004), the study seeks to offer a more complete picture of the construction of spoken meaning across languages and modalities.

## **II. RESEARCH CONTENTS**

### **Theoretical framework**

#### ***Systemic Functional Phonology***

The analytical framework used in this study is Systemic Functional Phonology (SFP), developed by Halliday (1967) and elaborated by Halliday and Greaves (2008). Within the systemic functional tradition, intonation is treated not as a peripheral expressive feature but as an integral component of the grammar of spoken language. Prosodic choices contribute systematically to the organisation of information, discourse structure, and interpersonal relations, and are therefore analysed as meaning-bearing linguistic resources (Halliday, 1967; Halliday & Greaves, 2008).

SFP describes prosodic organisation through three interrelated systems: tonality, tonicity, and tone. Tonality divides speech into tone groups; tonicity identifies the syllable that carries tonic prominence; and tone specifies the pitch movement associated with that tonic syllable. These systems operate simultaneously within the tone group, which is treated as the basic unit of prosodic analysis. Through their interaction, speakers organise information, signal discourse boundaries, foreground new information, and negotiate interpersonal meanings such as certainty, involvement, reservation, or emotional orientation (Halliday, 1967; Halliday & Greaves, 2008).

A major strength of the Hallidayan framework is that it provides a principled account of how prosodic form is connected to communicative function. Intonation is interpreted in relation to the metafunctional organisation of language rather than as a purely phonetic phenomenon. This functional orientation is particularly relevant for the present study because film dialogue frequently relies on prosodic resources to construct interpersonal meaning beyond the lexical content of the utterance.

Tench (1996, 2011) extends the functional description of intonation through detailed analyses of tone, prominence, pausing, and discourse organisation. Building on this perspective, the present study further incorporates tempo and voice quality as additional analytical resources. These resources interact with the Hallidayan systems in shaping discourse interpretation and interpersonal stance. Accordingly, the present study treats tempo, pausing, loudness, rhythm, and voice quality as analytically distinct but functionally related meaning-making resources. This extension is methodological rather than theoretical: it expands the descriptive toolkit while retaining the Hallidayan conception of prosody as a functional semiotic system. (Halliday & Greaves, 2008; Tench, 2011).

At the same time, two limitations of SFP should be acknowledged. First, the framework primarily models the prosodic channel of spoken communication. It does not provide a comprehensive account of paralinguistic vocal behaviour (e.g., laughter, sighing, whispering, breathiness, creaky voice, or vocalisations) or bodily behaviour (e.g., facial expression, gesture, gaze, posture, and head movement). In naturally occurring interaction, these resources frequently co-occur with prosody and contribute jointly to the construction of emotion, evaluation, and interpersonal stance. Second, the Hallidayan description was developed primarily on the basis of English. Although its functional principles are widely applicable, the phonological realisation of prosodic systems differs across languages and cannot be transferred mechanically from English to Vietnamese.

Vietnamese presents a particularly important case because it is a lexical-tone language. Lexical tone plays a phonological role in distinguishing word meaning, and this constrains the range of pitch variation available for discourse-level purposes. However, lexical tone does not eliminate discourse intonation. Vietnamese speakers make systematic use of pitch range, register, boundary phenomena, tempo, pausing, loudness, and voice quality to express attitude, stance, emphasis, and interactional meaning (Brunelle, 2009; Kirby, 2011; Michaud, 2004). For this reason, the present study analyses Vietnamese prosody through categories that are functionally comparable to those used for English while recognising the language-specific constraints imposed by lexical tone.

### ***Sound Semiotics***

The second theoretical component is the social-semiotic approach to sound developed by Van Leeuwen (1999). This approach argues that sound functions as a socially organised semiotic resource whose meanings emerge through repeated cultural and communicative practices. In Van Leeuwen's (1999) framework, acoustic properties are not treated as neutral physical phenomena. Rather, Sound Semiotics is concerned with how listeners read patterns of pitch, loudness, tempo and voice quality as socially meaningful signs.

Van Leeuwen (1999) argues that acoustic variation is systematically associated with socially interpretable meanings. High pitch and expanded pitch range may be associated with excitement, urgency, or emotional intensity, whereas low pitch may be associated with calmness, seriousness, or authority (Van Leeuwen, 1999); increased loudness may index anger, confidence, or public orientation; reduced loudness may index intimacy, secrecy, or vulnerability; a faster tempo may index enthusiasm, tension, or urgency; a slower tempo may index reflection, sadness, or solemnity; voice quality can add meanings such as tenderness, emotional strain, distance, or suppression.

Sound Semiotics, for the purposes of the present study, provides an interpretive vocabulary for the analysis of social meanings of vocal behaviour. This is particularly relevant to film dialogue where actors intentionally modulate pitch, loudness, tempo and voice quality to create character identity, emotional states, interpersonal relations and narrative tension (Van Leeuwen, 1999). These acoustic resources often function as culturally recognisable social cues that audiences can interpret as part of character and interactional meaning.

However, Sound Semiotics differs fundamentally from Systemic Functional Phonology in analytical orientation. It is not organised as a systemic network with entry conditions, paradigmatic choices, and realisation rules. Consequently, the present study does not treat Sound Semiotics as a grammatical model of prosodic organisation. Instead, it is used as a social-semiotic interpretive framework that complements the structural description provided by SFP. The Halliday–Tench framework provides the coding categories for prosodic organisation (Halliday, 1967; Halliday & Greaves, 2008; Tench, 2011), while Van Leeuwen (1999) provides the semantic vocabulary for interpreting the social meanings associated with those coded features.

### ***Multimodal communication***

The third theoretical component is a multimodal perspective on spoken communication, which treats meaning as the coordinated product of multiple semiotic resources (Kress & Van Leeuwen, 2001; Norris, 2004). In this perspective, spoken interaction is not simply a sequence of linguistic forms but a temporally organised integration of vocal, visual, and bodily resources.

For analytical purposes, the present study distinguishes three interacting channels: (1) the prosodic channel, including intonation, prominence, rhythm, and discourse-level pitch organisation; (2) the paralinguistic channel, including loudness, tempo, pausing, voice quality, and non-lexical vocalisations; and (3) the bodily-behaviour channel, including gesture, gaze, facial expression, posture, and head movement. These channels are analysed primarily through two metafunctional dimensions: the experiential metafunction, which contributes to

the construction of events, emotional experience, and discourse organisation, and the interpersonal metafunction, which contributes to the negotiation of social relations, evaluation, stance, intimacy, authority, and alignment (Halliday, 1967; Halliday & Greaves, 2008).

A key assumption of the framework is that meaning is often produced through the synchronisation or coordination of channels (Norris, 2004). A prosodic prominence may coincide with a gesture, a gaze shift, and a loudness peak, creating a single integrated emphasis. Conversely, channels may diverge: a smiling face may accompany a tense voice quality, or a calm pitch contour may co-occur with an aggressive gesture. Such cross-channel relationships frequently generate meanings such as irony, mitigation, emotional conflict, or restrained hostility that cannot be adequately explained through prosody alone (Kress & Van Leeuwen, 2001; Norris, 2004).

This multimodal perspective provides the integrative framework for the present study. Systemic Functional Phonology provides the grammatical description of prosodic organisation (Halliday, 1967; Halliday & Greaves, 2008), Van Leeuwen's (1999) social-semiotic approach provides an interpretive vocabulary for the social meanings of vocal resources, and the multimodal perspective explains how prosodic, paralinguistic, and bodily resources operate together as integrated meaning-making systems in English–Vietnamese film dialogue (Kress & Van Leeuwen, 2001; Norris, 2004).

### **Distinguishing prosody, lexical tone, poetic rhythm, and paralinguistic meaning**

A major methodological challenge in the analysis of Vietnamese spoken discourse is that several sound-related phenomena are frequently treated as if they belonged to the same analytical level. In discussions of Vietnamese intonation and prosody, lexical tone, discourse prosody, rhythmic organisation, and paralinguistic vocal behaviour are frequently discussed under the general label of “intonation” or “prosody”, despite belonging to different analytical levels (Brunelle, 2009; Kirby, 2011). Such conflation obscures important differences in linguistic function, analytical status, and methods of description. The present study therefore proposes a four-way distinction that separates these phenomena into analytically independent domains.

The first phenomenon is natural prosody, which refers to the organisation of spoken discourse through tone-group structure, prominence, rhythm, pausing, and discourse-level intonation (Halliday, 1967; Halliday & Greaves, 2008). Prosody operates above the lexical level and contributes to the organisation of information, discourse segmentation, speaker attitude, and interpersonal meaning. In English, these functions are realised through the interaction of tonality, tonicity, and tone (Halliday, 1967; Halliday & Greaves, 2008). In Vietnamese, comparable discourse functions are realised through rhythmic organisation, pausing, tempo, pitch range, register, and boundary phenomena, although these resources interact with the language's lexical-tone system (Brunelle, 2009; Kirby, 2011; Michaud, 2004)..

The second phenomenon is poetic rhythm, particularly the metrical organisation found in Vietnamese verse traditions such as *ca dao* and *lục bát*. Poetic rhythm is governed by relatively stable conventions on syllable count, rhyme, tonal patterning, and parallelism that belong primarily to the domain of literary composition rather than spontaneous spoken interaction (Nguyen Tai Can, 1998; Tran Dinh Su, 2005). A *lục bát* sequence follows metrical principles that remain comparatively stable across performances, whereas conversational prosody is organised dynamically in response to discourse structure, interactional goals, and interpersonal negotiation. Although poetic performance may employ prosodic resources, poetic rhythm should not be treated as equivalent to discourse prosody because the organising principles of literary metre and conversational speech are fundamentally different (Nguyen Tai Can, 1998; Tran Dinh Su, 2005).

The third phenomenon is lexical tone, which functions as a phonological mechanism for distinguishing lexical meaning (Kirby, 2011; Brunelle, 2009). Vietnamese lexical tones form part of the lexical identity of words, and changes in tonal specification may produce changes in lexical meaning. Lexical tone therefore belongs to the phonological organisation of the lexicon rather than to discourse-level prosodic organisation. This distinction is particularly important because surface pitch movement may simultaneously reflect both lexical-tone realisation and discourse-level prosodic modification.

The relationship between lexical tone and prosody requires careful treatment. Vietnamese speakers can express discourse-level meanings through pitch-related resources, but these meanings are realised within the constraints imposed by lexical tone (Brunelle, 2009; Kirby, 2011; Michaud, 2004). Consequently, not every pitch movement should be interpreted as an intonational choice. The present study adopts a lexical-tone protection principle: if a pitch pattern functions primarily to realise the lexical tone of a Vietnamese word, it is analysed at the lexical layer rather than the prosodic layer. Prosodic analysis is reserved for pitch range, register, boundary phenomena, prominence, rhythmic organisation, pausing, and other discourse-level modifications that do not alter the lexical identity of the word. This principle prevents the common analytical error of treating lexical-tone realisation as discourse intonation.

The fourth phenomenon is paralinguistic meaning, which is realised through loudness, speech tempo, pausing, voice quality, and non-lexical vocalisations such as laughter, sighing, whispering, crying, or breathiness. These resources are interpreted in the present study through Van Leeuwen's (1999) social-semiotic framework.

Unlike lexical tone, paralinguistic resources do not distinguish word meaning. Unlike prosody, they are not organised primarily through the systemic phonological structure of the tone group. Instead, they function as socially interpretable vocal behaviours that contribute to emotion, evaluation, interpersonal stance, authority, intimacy, solidarity, and social distance (Van Leeuwen, 1999). The analytical distinction among these four phenomena can be summarised in Table 1.

**Table 1. Four analytically distinct sound-related phenomena**

| Phenomenon             | Primary function                                                     | Analytical framework                                                        |
|------------------------|----------------------------------------------------------------------|-----------------------------------------------------------------------------|
| Natural prosody        | Information structure, discourse organisation, interpersonal meaning | Systemic Functional Phonology (Halliday, 1967; Halliday & Greaves, 2008)    |
| Poetic rhythm          | Literary metre and compositional structure                           | Literary and rhetorical analysis (Nguyen Tai Can, 1998; Tran Dinh Su, 2005) |
| Lexical tone           | Lexical contrast and phonological identity                           | Vietnamese phonology (Kirby, 2011; Brunelle, 2009)                          |
| Paralinguistic meaning | Emotion, stance, evaluation, and social meaning                      | Sound Semiotics of sound (Van Leeuwen, 1999)                                |

This classification has important methodological consequences for the coding procedure developed later in the paper. Each observed sound-related feature is assigned to one analytical layer rather than being treated as a general prosodic phenomenon. For example, a rise in pitch at the end of a Vietnamese utterance may involve (i) lexical-tone realisation, (ii) discourse-level boundary marking, or (iii) paralinguistic emotional expression (Brunelle, 2009; Kirby, 2011; Michaud, 2004). These possibilities are distinguished through the combined analysis of lexical identity, tone-group organisation (Halliday, 1967; Halliday & Greaves, 2008), and accompanying paralinguistic and bodily-behaviour evidence.

The four-way distinction is not intended to deny interaction among these phenomena. In naturally occurring speech they frequently co-occur and influence one another (Norris, 2004). Rather, the purpose of the distinction is analytical clarity. By separating natural prosody, poetic rhythm, lexical tone, and paralinguistic meaning into independent but interacting domains, the present study establishes a more reliable foundation for the cross-linguistic comparison of English and Vietnamese film dialogue and for the multimodal coding procedure developed in the following sections.

### III. Methodology

#### *Corpus selection and scene matching*

The study adopts a comparative qualitative corpus design based on English and Vietnamese feature films. Film dialogue is selected because it provides relatively naturalistic spoken interaction while allowing repeated observation of prosodic, paralinguistic, and bodily-behaviour resources (Baldry & Thibault, 2006; Norris, 2004). Unlike laboratory speech, film dialogue contains emotionally rich interpersonal exchanges in which prosodic and multimodal resources contribute to character construction and audience interpretation (Baldry & Thibault, 2006; Van Leeuwen, 1999).

The corpus consists of two English-language films (*The King's Speech* and *Marriage Story*) and two Vietnamese-language films (*Tro Tàn Rực Rỡ* and *Bên Trong Vỏ Kén Vàng*). These films were chosen because they contain a high density of interpersonal interaction, emotional negotiation, and evaluative dialogue. They also represent contemporary cinematic speech rather than highly stylised poetic or theatrical language.

Rather than comparing films as complete texts, the study analyses matched interactional scenes. Scenes are paired across the two languages according to four criteria: (1) comparable interpersonal relationships (e.g., family members, spouses, or authority figures); (2) similar communicative purposes (e.g., confrontation, reconciliation, confession, or emotional withdrawal); (3) comparable levels of emotional intensity; and (4) conversational rather than monologic discourse. This scene-matching procedure is intended to improve functional comparability across the English and Vietnamese datasets.

#### *Annotation framework*

The data are annotated in ELAN (EUDICO Linguistic Annotator), a tool widely used for multimodal annotation and time-aligned analysis (Wittenburg et al., 2006), using a four-layer analytical architecture proposed in the present study. The tone group is treated as the basic alignment unit for annotation (Halliday, 1967; Halliday & Greaves, 2008), allowing prosodic, paralinguistic, and bodily resources to be analysed within the same temporal framework.

The annotation architecture consists of the following layers:

- Prosodic layer (P): tone-group segmentation, tonic prominence, pitch movement, rhythmic organisation, and pausing.
- Paralinguistic layer (Pa): loudness, speech tempo, voice quality, and non-lexical vocalisations.
- Bodily-behaviour layer (B): gaze, facial expression, gesture, posture, and head movement.

- Meaning layer (M): experiential interpretation and interpersonal interpretation (Halliday, 1967; Halliday & Greaves, 2008), including stance, evaluation, emotion, authority, solidarity, intimacy, and social distance.

The separation of these layers allows multiple semiotic resources to be coded independently while preserving their temporal coordination during analysis (Norris, 2004; Baldry & Thibault, 2006).

### ***Coding procedure***

The analysis follows a four-stage coding procedure. First, each scene is segmented into tone groups on the basis of discourse organisation rather than orthographic sentence boundaries. Second, prosodic features are coded using the functional categories derived from Systemic Functional Phonology (Halliday, 1967; Halliday & Greaves, 2008; Tench, 2011), including tone-group structure, prominence, pitch movement, rhythm, and pausing. Third, paralinguistic features are coded using a descriptive inventory of loudness, tempo, voice quality, and vocal behaviour (Van Leeuwen, 1999). Fourth, bodily-behaviour features are coded from the video data (Norris, 2004), after which the three channels are interpreted jointly at the meaning layer.

For the Vietnamese data, coding is guided by the lexical-tone protection principle introduced in Section 3. Pitch events that primarily realise lexical tone (Kirby, 2011; Brunelle, 2009) are analysed at the lexical level, whereas discourse-level pitch range, boundary phenomena, prominence, tempo, pausing, and voice quality are coded as prosodic or paralinguistic resources. This procedure is intended to reduce the risk of conflating lexical-tone realisation with discourse-level prosodic organisation.

### ***Analytical strategy and reliability***

The analysis is primarily qualitative and comparative. The aim is not statistical generalisation but the identification of recurrent multimodal configurations that contribute to interpersonal meaning in English and Vietnamese film dialogue. Comparative interpretation focuses on how similar communicative functions—such as authority, emotional withdrawal, solidarity, or conflict—are realised through different combinations of prosodic, paralinguistic, and bodily resources across the two languages.

To enhance analytical reliability, a subset of the data is independently re-annotated after an interval to assess coding consistency, following established practices in multimodal annotation (Norris, 2004; Baldry & Thibault, 2006). Ambiguous cases, particularly those involving the interaction of lexical tone and discourse-level pitch movement in Vietnamese, are examined through cross-layer comparison rather than isolated acoustic interpretation. Coding decisions are documented with annotation notes that specify the relevant channel and the rationale for the interpretation. This transparent coding procedure is intended to improve the replicability and analytical transparency of the multimodal annotation framework and to provide a systematic basis for the cross-linguistic comparison developed in the following section (Baldry & Thibault, 2006; Wittenburg et al., 2006).

## **Cross-linguistic comparison of English and Vietnamese film dialogue**

### ***Functional comparability across typologically different languages***

The comparison between English and Vietnamese is motivated by a functional rather than purely structural perspective. The two languages differ substantially in their phonological organisation, particularly in the role of lexical tone, stress, and intonation (Halliday & Greaves, 2008; Brunelle, 2009; Kirby, 2011). English relies heavily on prominence, rhythm, and intonational organisation (Halliday, 1967; Halliday & Greaves, 2008; Tench, 2011) to structure discourse and negotiate interpersonal meaning, whereas Vietnamese operates within a lexical-tone system (Kirby, 2011; Brunelle, 2009) in which pitch simultaneously serves lexical and discourse-related functions. These typological differences make direct one-to-one mapping between English intonation categories and Vietnamese pitch patterns problematic.

However, despite these structural differences, speakers of both languages must accomplish comparable communicative tasks. They need to foreground information, express emotion, evaluate situations, negotiate social relations, and position themselves toward other participants in interaction. The present study therefore compares multimodal functional configurations rather than isolated acoustic features (Kress & Van Leeuwen, 2001; Norris, 2004). The central question is not whether English and Vietnamese use identical prosodic forms, but how each language mobilises prosodic, paralinguistic, and bodily resources to realise similar interpersonal functions.

### ***Shared multimodal resources***

The analysis suggests that English and Vietnamese share a substantial repertoire of multimodal meaning-making resources. In both languages, speakers can systematically exploit rhythm, pausing, speech tempo, loudness, voice quality, gaze, facial expression, and gesture to construct interpersonal meanings (Norris, 2004; Van Leeuwen, 1999). Emotional intensity, interpersonal distance, solidarity, authority, and hesitation are rarely conveyed through pitch alone; instead, they emerge through coordinated patterns across multiple channels.

One recurrent pattern involves prosodic prominence synchronised with bodily behaviour (Halliday, 1967; Halliday & Greaves, 2008; Norris, 2004). In English film dialogue, an emphatic lexical item is frequently accompanied by increased loudness, expanded pitch range, direct gaze, and a reinforcing gesture. Vietnamese dialogue also displays coordinated multimodal emphasis, although the acoustic realisation may differ because pitch interacts with lexical tone (Brunelle, 2009; Kirby, 2011). In both languages, interpersonal meanings such as insistence, accusation, or emotional commitment are often realised through the convergence of vocal and bodily resources rather than through a single prosodic feature (Norris, 2004; Van Leeuwen, 1999).

Similarly, reduced loudness, slower tempo, gaze withdrawal, and limited gestural movement frequently co-occur in scenes of emotional withdrawal or relational distance. These shared multimodal configurations suggest that important aspects of interpersonal communication operate across languages even when their phonological systems differ (Norris, 2004; Van Leeuwen, 1999).

### ***Language-specific configurations***

Although the two languages share multimodal resources, they differ in the distribution of functional load across those resources. In English, prominence placement and intonational contour play a particularly important role in signalling information structure, contrast, and interpersonal orientation (Halliday, 1967; Halliday & Greaves, 2008; Trench, 2011). Speakers often concentrate evaluative force on the tonic syllable through the interaction of stress, pitch movement, loudness, and temporal lengthening (Halliday & Greaves, 2008; Trench, 2011).

Vietnamese appears to distribute comparable interpersonal meanings across a broader combination of pitch range, register, tempo, pausing, loudness, and voice quality (Brunelle, 2009; Kirby, 2011; Michaud, 2004). Because lexical tone constrains pitch realisation at the word level, discourse-level meanings are frequently reinforced by non-tonal vocal resources and by bodily behaviour (Brunelle, 2009; Kirby, 2011). Authority, emotional suppression, tenderness, resignation, or relational negotiation may be realised through tempo reduction, increased pausing, breathy or tense voice quality, and coordinated gaze or gesture patterns (Van Leeuwen, 1999; Norris, 2004).

This does not imply that Vietnamese lacks discourse-level intonational meaning. Rather, discourse intonation operates within a different phonological ecology. The interaction between lexical tone and discourse prosody requires analysts to distinguish lexical-tone realisation from discourse-level prosodic modification, as discussed in Section 3. The comparative analysis therefore focuses on multimodal configurations rather than assuming that English intonational categories can be directly transferred to Vietnamese.

### ***Constrained correspondences***

A particularly important finding concerns surface similarities that mask different functional organisations. A rising pitch movement in English may function primarily as a discourse-level intonational choice (Halliday & Greaves, 2008), whereas a similar acoustic rise in Vietnamese may reflect lexical-tone realisation, discourse-level boundary marking, or an interaction between lexical and discourse factors (Brunelle, 2009; Kirby, 2011). Surface acoustic similarity should therefore not be interpreted as evidence of functional equivalence.

For this reason, the study proposes the notion of constrained correspondence. English and Vietnamese often realise comparable interpersonal meanings, but the correspondence is constrained by the different organisational principles of their phonological systems. The relationship is therefore functional rather than formally identical. A translation, dubbing, or interpretation that preserves lexical content while ignoring multimodal coordination may reproduce propositional meaning but fail to preserve interpersonal meaning (Kress & Van Leeuwen, 2001; Norris, 2004).

### ***Implications for pedagogy and interpretation***

The comparative findings have implications for both English language teaching and audiovisual translation. In many EFL contexts, pronunciation instruction has traditionally focused on segmental accuracy and isolated intonation patterns (Halliday & Greaves, 2008; Trench, 2011). The present analysis suggests that learners also need explicit awareness of tempo, pausing, loudness, voice quality, gaze, and gesture as functional resources for interpersonal communication (Norris, 2004; Van Leeuwen, 1999). Misunderstanding these resources may lead to pragmatic failure even when grammatical and lexical comprehension is accurate.

For translation and dubbing, the findings suggest that interpersonal equivalence requires more than lexical or grammatical correspondence. Translators and voice actors need to reproduce the multimodal configuration of meaning, including vocal dynamics and bodily timing, rather than merely the propositional content of the dialogue (Kress & Van Leeuwen, 2001; Norris, 2004). This is particularly important in emotionally intense scenes where interpersonal meaning is distributed across multiple semiotic channels.

Overall, the comparison supports the central claim of the paper: English and Vietnamese differ in their phonological organisation, but both languages construct interpersonal meaning through coordinated multimodal

resources. A comparative framework that integrates prosody, paralinguistic behaviour, and bodily action therefore provides a more adequate account of cross-linguistic meaning-making than analyses restricted to intonation alone (Kress & Van Leeuwen, 2001; Norris, 2004).

#### IV. CONCLUSION

This study has proposed an integrated multimodal framework for analysing prosody and paralinguistic features as meaning-making resources in English–Vietnamese film dialogue. By integrating insights from Systemic Functional Phonology (Halliday, 1967; Halliday & Greaves, 2008), Van Leeuwen's (1999) social-semiotic framework and Leeuwen's (1999) social-semiotic approach to sound, and multimodal communication research (Kress & Van Leeuwen, 2001; Norris, 2004), the paper argues that interpersonal meaning is constructed through the coordinated operation of prosodic, paralinguistic, and bodily-behaviour resources rather than through intonation alone.

The study makes three principal contributions. First, it develops a four-channel analytical framework that integrates prosodic organisation, paralinguistic vocal behaviour, bodily action, and metafunctional interpretation within a unified coding architecture. Second, it establishes a clear distinction among natural prosody, poetic rhythm, lexical tone, and paralinguistic meaning, thereby addressing a recurrent source of conceptual ambiguity in the analysis of Vietnamese spoken discourse. The lexical-tone protection principle proposed in the study provides a methodological mechanism for separating lexical-tone realisation from discourse-level prosodic interpretation. Third, the comparative analysis demonstrates that English and Vietnamese often realise similar interpersonal functions through different multimodal configurations, suggesting that cross-linguistic comparison should focus on functional meaning-making patterns rather than on isolated acoustic features.

The findings also have broader implications for EFL pedagogy, intercultural communication, and audiovisual translation. They suggest that learners need explicit awareness of tempo, pausing, loudness, voice quality, gaze, and gesture as interpersonal resources, and that translation and dubbing require attention to multimodal interpersonal equivalence rather than lexical correspondence alone.

Several limitations should be acknowledged. The corpus is relatively small and limited to contemporary feature films and the analysis is largely qualitative. Future work might extend the corpus to include television drama, naturally occurring conversation, interpreted interaction and longitudinal audiovisual data. Quantitative multimodal annotation and acoustic analysis may also offer more robust empirical evidence for the functional patterns identified here.

The paper argues that a multimodal perspective allows for a more complete account of spoken meaning than approaches that treat prosody as an isolated phonological phenomenon (Kress & Van Leeuwen, 2001; Norris, 2004). The proposed framework combines grammatical, social-semiotic and multimodal perspectives in order to provide a systematic basis for comparing spoken discourse in English and Vietnamese, and for future research into multimodal meaning-making across languages.

#### REFERENCES

- [1] Baldry, A., & Thibault, P. J. (2006). *Multimodal transcription and text analysis: A multimedia toolkit and coursebook*. Equinox.
- [2] Brunelle, M. (2009). Tone perception in Northern and Southern Vietnamese. *Journal of Phonetics*, 37(1), 79–96. <https://doi.org/10.1016/j.wocn.2008.09.003>
- [3] Dingemanse, M. (2012). Advances in the cross-linguistic study of ideophones. *Language and Linguistics Compass*, 6(10), 654–672. <https://doi.org/10.1002/lnc3.361>
- [4] Halliday, M. A. K. (1967). *Intonation and grammar in British English*. Mouton.
- [5] Halliday, M. A. K., & Greaves, W. S. (2008). *Intonation in the grammar of English*. Equinox.
- [6] Hinton, L., Nichols, J., & Ohala, J. J. (Eds.). (1994). *Sound symbolism*. Cambridge University Press.
- [7] Kirby, J. (2011). Vietnamese (Hanoi Vietnamese). *Journal of the International Phonetic Association*, 41(3), 381–392. <https://doi.org/10.1017/S0025100311000181>
- [8] & Van Leeuwen, T. (2001). *Multimodal discourse: The modes and media of contemporary communication*. Arnold.
- [9] (2004). Final consonants and glottalization: New perspectives from Hanoi Vietnamese. *Phonetica*, 61(2–3), 119–146. <https://doi.org/10.1159/000082092>
- [10] Nguyễn Tài Căn. (1998). *Ngữ pháp tiếng Việt: Tiếng – từ ghép – đoản ngữ*. Nhà xuất bản Đại học Quốc gia Hà Nội.
- [11] Norris, S. (2004). *Analyzing multimodal interaction: A methodological framework*. Routledge.
- [12] Tench, P. (1996). *The intonation systems of English*. Cassell.
- [13] Tench, P. (2011). *Transcribing the sound of English: A phonetics and phonology workbook for words and discourse* (2nd ed.). Cambridge University Press.
- [14] Trần Đình Sử. (2005). *Thi pháp văn học trung đại Việt Nam*. Nhà xuất bản Đại học Quốc gia Hà Nội.
- [15] Van Leeuwen, T. (1999). *Speech, music, sound*. Macmillan.
- [16] Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)* (pp. 1556–1559). European Language Resources Association.