International Journal of Language, Linguistics, Literature and Culture
Vol. 02, No. 04; 2023
ISSN: 2583-6560

RESEARCH ON VOICE QUALITY IN CALL CENTER DISCOURSE: VALIDATING RELIABILITY THROUGH INTERRATER AGREEMENT AND PRAAT

Dr. Yau-ni WAN
Assistant Professor of Department of English Language and Literature at Hong Kong Shue Yan University in Hong Kong, China. Systemic Functional Linguistics.

https://doi.org/10.59009/ijlllc.2023.0029

ABSTRACT

Voice quality is an important component of effective communication in call center conversations. A qualitative research methodology was used in the present study to examine how Filipino customer service representatives (CSRs) and American customers use voice quality features in call center discourse. A precise evaluation of voice quality features can aid in the development of effective communication strategies. Through interrater agreement, the present study aimed to validate the reliability of voice quality feature assessment. In this study, how Filipino CSRs and American customers' call center telephone conversations differed in terms of voice quality characteristics such as volume (Loud/Soft), pitch (High/Low), tension (Tense/Lax), and rhythm (Fast/Slow) was investigated. The conversation transcripts were examined using the aforementioned voice quality features. The transcriptions were then evaluated by three independent raters to determine interrater agreement. The findings revealed a high level of interrater agreement between the three raters of all voice quality features, with up to 0.8 in identifying voice quality changes in generic stages. In particular, a higher agreement was found in the assessment of specific voice quality features such as loudness, high pitch, and tension. Praat software was also used to aid in the analysis of some voice recordings to validate the interrater agreement reliability. The approach taken in the present study for assessing voice quality using interrater agreement and Praat software can improve service quality in the call center industry and provides a reliable and valid framework for future qualitative research in applied linguistics.

Keywords: Interrater Agreement, Test Validity, Qualitative Research, Voice Quality Features, Reliability, Research Instruments.

1. INTRODUCTION

Using elements of social semiotic theories (see Halliday, 1978; van Leeuwen, 1999; 2005), the interpersonal meaning of call center discourse is examined in the present study. In a sample of complex complaint calls, the resource of voice quality is specifically examined to track how the caller and customer service representative (CSR) interpret interpersonal meanings. Special attention is placed on prosodic realization attitude. Understanding the attitudinal and acoustic characteristics of speech makes applied linguistics research on voice quality features an interesting and significant field of study. A qualitative research methodology is used in the present study to analyze attitudes shaped by the CSR and/or the customer as the negotiation and alignment processes are examined. Unfortunately, subjectivity is a problem that qualitative researchers often encounter. The consistency with which raters rate various voice quality attributes can have a significant impact on speech research. The reliability of interrater agreement is critical when validating the accuracy of such studies. As a result, the purpose of

Vol. 02, No. 04; 2023
ISSN: 2583-6560

the present study is to examine the readability of interrater agreement when evaluating voice quality features using both attitudinal and acoustic approaches. The findings can aid in improving the reliability of voice quality assessments and the accuracy of qualitative speech research. This paper begins by introducing the research paradigm, the justification for the methodology, and the research instruments used and then proceeds to describe the validity procedures for qualitative research, the validation of voice quality in the literature, and the development of interrater exercises used in the current study. Finally, the present study's impact on the sound software Praat is demonstrated. The role of voice quality and the relative changes in voice quality features in conversations are two important linguistic resources for comprehending negotiation in call center conversations. The conversational data is carefully examined in the present study to understand the interpersonal features of the text to comprehend how this relationship is realized and how meanings are construed.

Naturalistic Paradigm And Voice Quality Features

A naturalistic paradigm was used in the present study to examine social interactions in call center conversations. Because Systemic Functional Linguistics (SFL) "incorporates the idea that language is a social phenomenon and when dealing with language it works at the level of the text as a unit of meaning," it was chosen as the framework for analyzing spoken conversations in call centers (Forey, 2004, p. 449). Researchers in SFL examine how readers understand texts. According to SFL, language is a social phenomenon. CSRs and customers can interact in a way that resembles social behavior over the phone. By using naturalistic inquiry as a paradigm and SFL as a framework, this linguistics study attempts to address the identified language issue. This knowable negotiation process can be realized through text analysis. The use of text analysis enables the identification of genuine calls. Call center transcripts, which document actual interactions between a customer service representative (CSR) and a customer, were the texts used in this study. These conversations were recorded to reflect what actually happens during calls. This helped to paint a complete picture of the call center sector. According to Scott and Usher (1996, p. 10), researchers should make an effort to systematically look into a few predefined issues. The systematic collection of texts is the knowable result of the SFL approach. Furthermore, the relationship between voice quality features and the naturalistic paradigm is significant because, in contrast to controlled laboratory settings, the naturalistic paradigm offers a more ecologically valid context for studying naturally occurring voice quality features. The naturalistic paradigm in voice quality research involves the analysis of voice characteristics in naturalistic or real-life contexts. In addition, the intricate interactions between various factors, such as the speaker's emotions, social context, and physical surroundings, that influence voice production are explored. This approach aims to capture the natural variation in voice quality that occurs in call center conversations. Therefore, the development of interventions to enhance voice quality in real-world circumstances, such as workplace communication or social interactions, can be informed by the naturalistic paradigm.

Qualitative Research in Applied Linguistics

The present study favored a qualitative research approach over a quantitative one. To focus on locating Maxwellian insights that emerge from the data, researchers can use a qualitative research approach to create their own interpretations of events and phenomena (2005). A sample of particular texts' interpersonal meanings are interpreted in the present study. Determination of the "truth" is not sought by extrapolating from the interpretations since taking a purely quantitative approach is not intended. The interpersonal meanings in complex calls

Vol. 02, No. 04; 2023
ISSN: 2583-6560

are focused on, something that has not yet been discussed in the literature, by concentrating on particular prosodic features and closely examining particular lexicogrammatical choices. Text analysis is a primary human-to-human method for gathering data in naturalistic evaluation. The information was gathered from insurance call centers in the Philippines. The information was gathered through a text analysis of transcriptions of telephone conversations with a focus on interpersonal characteristics at negotiation points in complex calls. Incorporating SFL theory, a thorough text analysis of key linguistic elements was conducted. To "make explicit the relations between meaning constructed at clause level and meaning at the 'larger' levels (paragraphs and text), which in turn can be systematically related to the specified elements of the context," an SFL-style analysis of texts is necessary (Harvey, 1993, p. 25). The study of interpersonal meaning involved using audio-taped transcriptions of call center transactions. Potter (1996) made the claim that a transcription can convert sound into a written format that enables readers to read it quickly. However, in the present study, the transcript cannot completely replace the audio tape. This is because to analyze lexicogrammatical and voice features, it is imperative to work simultaneously with transcripts and tapes. Any study of spoken discourse should, according to Silverman (2000) and Sacks (1992), examine the actual details of conversations.

Systems for transcription are typically designed to focus on particular aspects of interaction (Potter, 1996). It would be reasonable to only transcribe the information required to address the study's specific research questions (Strauss, 1987). Turns of speakers were recorded in 20 transcripts. "The dynamic and sequence of the talk in which one speaker takes a turn after another speaker, and one sentence leads to another sentence" is how turn-taking organization is described (Eggins, 2004, p. 93). Following consideration, Table 1's transcription key from Eggins and Slade (1997, p. 5) was chosen as the primary transcription code for call center conversations.

Vol. 02, No. 04; 2023
ISSN: 2583-6560

Table 1 Summary of Transcription Key (Eggins & Slade, 1997, p. 5)

Symbol	Meaning
.	certainty, completion (typically falling tone)
no end of turn punctuation	implies nontermination (no final intonation)
,	parceling of talk; breathing time (silent beats in Halliday's 1985/94 system)
?	uncertainty (rising tone, or wh-interrogative)
!	"surprised" intonation (rising-falling tone 5 in Halliday's 1994 system)
WORDS IN CAPITALS	emphatic stress and/or increased volume
" "	change in voice quality in reported speech
( )	untranscribable talk
(words within parentheses)	transcriber's guess
[words in square brackets]	nonverbal information
= =	overlap (contiguity, simultaneity)
…	short hesitation within a turn (less than three seconds)
[pause – 4 secs]	indication of interturn pause length
dash – then talk	false start/restart

Due to the high level of confidentiality surrounding the conversational data used in this study, a transcription code was created to protect policyholder names, dates mentioned, and policy and social security numbers. The transcription code mentioned above called into question the accuracy of the information. However, these adjustments were unavoidable because of research ethics. As paralinguistic and nonverbal information related to the analysis of interpersonal meaning was included in the present study, transcription codes used in previous studies (see Atkinson & Heritage, 1984; Blum-Kulka, Huck-Taglicht, & Avni, 2004; Chafe, 1993; Ehlich, 1993; Johnstone, Andrus, & Danielson, 2006; Lapadat & Lindsay, 1999; Poskiparta, Kettunen, & Liimatainen, 2000; Psathas & Anderson, 1990; Samuelsson, Nettelbladt, & Lofqvist, 2005) were reviewed. According to Bruce (1992), "it is reasonable to believe that a transcription system should be easy to write, easy to read, easy to learn, and easy to search" (p. 145). The present study was innovative in creating and implementing a specific transcription of conversational text that included paralinguistic voice quality features such as Pitch, Volume, Tension, and Rhythm. These transcription conventions were primarily used in the data to indicate changes in voice quality and related features (c.f. Wan, 2017, p. 143 for more details):

Volume (Loud/Soft, dbs)
°word° a passage of talk that was softer than surrounding talk
WORD a passage of talk that was louder than surrounding talk

Pitch Register (Low/High)
↓marked falling shifts in pitch that was lower than surrounding talk
↑marked rising shifts in pitch that was higher than surrounding talk

Rhythm (Slow/Fast)
> word < talk was faster than surrounding talk (fast rhythm)

Vol. 02, No. 04; 2023
ISSN: 2583-6560

< word > talk was slower than surrounding talk (slow rhythm)
:: an extension of a sound or syllable

Tension (Tense/Lax)
~ word ~ talk was laxer than surrounding talk, achieved by relaxing the muscle of the throat
+ word + talk was tenser than surrounding talk, achieved by tensing the muscle of the throat

[pause – x secs] timed intervals in seconds showed length of silence

Denzin and Lincoln (2005) described qualitative research

> as a situated activity that locates the observer in the world. It consists of a set of interpretive, material practices that make the world visible. These practices transform the world. They turn the world into a series of representations, including field notes, interviews, conversations, photographs, recordings, and memos to the self. At this level, qualitative research involves an interpretive, naturalistic approach to the world. Qualitative researchers study things in their natural settings and attempt to make sense of, or interpret, phenomena in terms of the meanings people bring to them. (p. 3)

Qualitative research techniques are not prescriptive or generalizable, and they are made to handle complex data (Lazaraton, 2002; Pennington, 2002; Trappes-Lomax, 2004; see Halliday, 1978). Due to their rarity, complex calls were frequently treated in the current study as single cases or special cases. These examples were incredibly helpful for the investigation of applied linguistics issues, even though the sample size was not very sufficient for generalization. Speaking discourse analysis techniques were used in the current call center study. Discourse research is typically qualitative because it requires interpretation by nature (Creswell, 2007; Denzin & Lincoln, 2000). Its stated objective is to "make sense of or interpret phenomena in terms of the meanings people ascribe to them" (Denzin & Lincoln, 2000, p. 3). The spoken call center discourse was analyzed in this study using transcripts as the primary source of data. Sound files were used to profile the voice quality features that were experienced during calls.

Rating of Voice Quality Features in Existing Research

Features of voice quality can be assessed at various levels using different methodologies. In earlier research, it was thought that interpersonal factors such as attitude, personal involvement, communication patterns, changes in timing, pitch, and rhythm could be used to communicate emotions such as happiness, sadness, fear, and anger (Hill, 1982; Himonides & Welch, 2005; Juslin & Sloboda, 2001). Scales made up of various adjective types were created for the raters, according to the literature (see Kurkul, 2007; La & Davidson, 2005; Orbelo, Testa, & Ross, 2003; O'Sullivan, 2003; Scherer & Ceschi, 2000). Adjectives such as good, bad, happy, sad, and polite/impolite were selected for the raters to evaluate their subjective feelings and judgments. In these previous studies, the raters ranked voice quality by assigning different numerical levels, such as ranking good/bad from *1* (the most likely) to *7* (the most unlikely). However, these Likert-type scales were not suitable for the present study, as such a ranking method only uses adjectives, which when reviewed closely, could be interpreted as undefined, subjective and less systematic for the analysis of a variant in voice quality features. Voice quality analysis above the simple sentence level was absent, and voice quality features were not addressed. The present study used a rating exercise to enhance the validity of the voice

Vol. 02, No. 04; 2023
ISSN: 2583-6560

quality analysis. The profile of the raters is outlined in the next section. Due to the call center company's privacy policy, the original sound tracks of call center conversations cannot be disclosed or reproduced for the sake of disclosure. As pointed out by Short, Semino and Wynne (2002),

> if someone re-says exactly what someone else said we would not expect voice quality features to be the same, simply because we know that the original speaker and the reporter are not the same person. This is because we all automatically make adjustments in accordance with the type/token distinction when we compare one thing with another. (p. 330)

Voice quality features thus were not reproduced and recorded in the present study, as this would change the originality and affect the results.

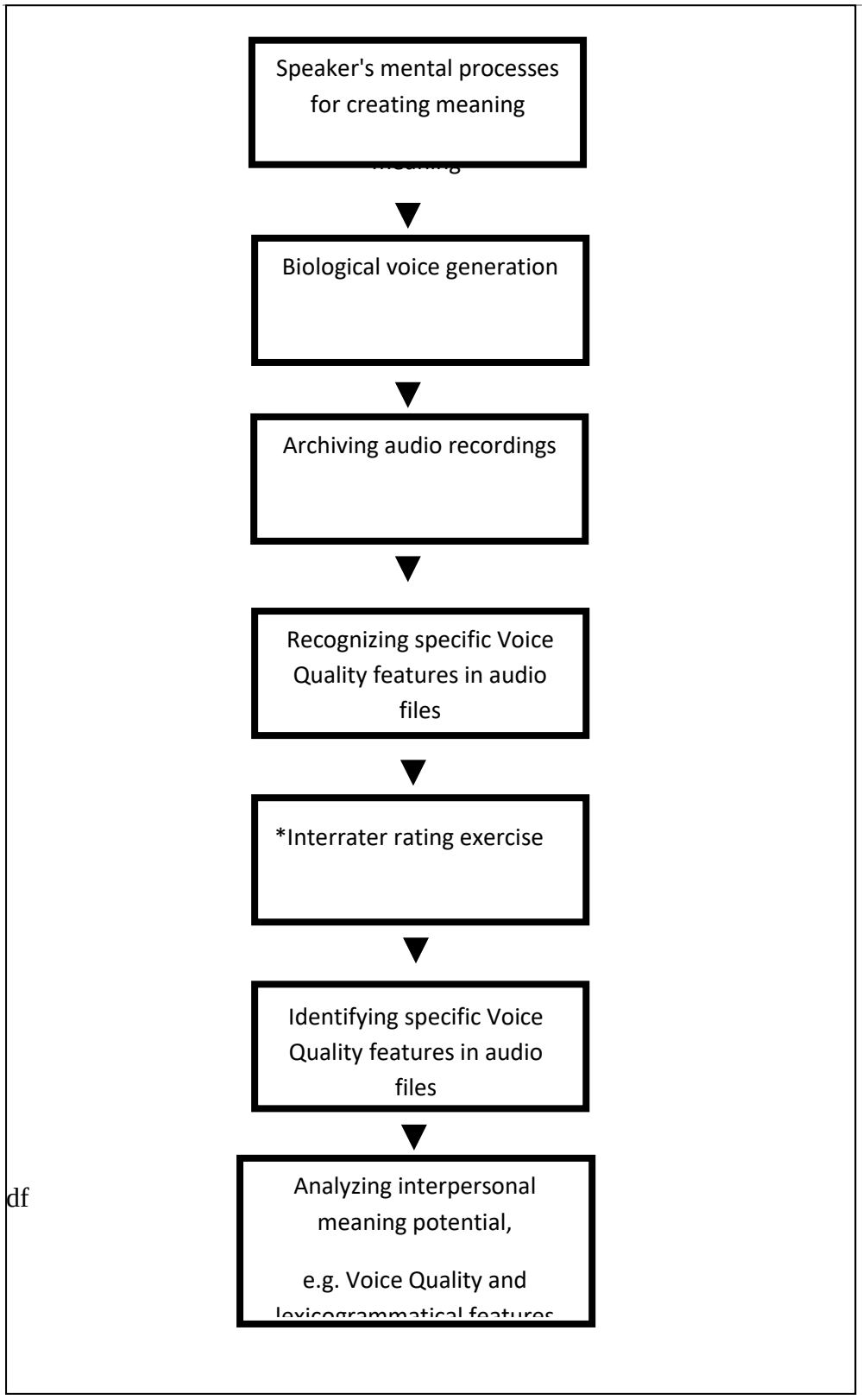
Validity of the Research Method: Interrater Agreement

Hammersley (1992) defined reliability as "the degree of consistency with which instances are assigned to the same category by different observers or by the same observer on different occasions" (p. 67). The most important goal of qualitative research, according to Lincoln and Guba (1985), is credibility. Interrater reliability testing is an important method for determining reliability (Potter, 1996). The present study included a rating exercise for identifying voice quality features. It is crucial to conduct interrater reliability in voice quality research to guarantee the accuracy and dependability of the findings. Voice quality features of the call center data all referred to relative changes in the conversation. A relative shift was an explicit voice change made by one or both speakers that could be recognized in relation to the rest of the conversation. The objective was not to compile a dictionary of specific voice quality features but rather to demonstrate how voice quality features support the exchange of interpersonal meaning in call center discourse. Interrater reliability is used to investigate whether multiple raters agree on a change in voice quality features, and if they do, the results are less likely to be biased. Because interrater reliability increases the consistency of the data obtained, it enables other researchers to replicate the study and obtain similar results. Interrater agreement improves the validity of the research and makes the findings more trustworthy. There are significant implications for validating interrater reliability in voice quality research.

2. METHODOLOGY

Voice quality characteristics are paralinguistic and multimodal resources for conveying meaning (Leijssen, 2006; Martin, 2007). In telephone conversations, identifying changes in voice quality and changes in interpersonal meaning is especially important. Because of the lack of visual cues, communicators must rely heavily on "differences in voice quality to assess emotional impact" (McCoyd & Kerson, 2006, p. 402). Figure 1 depicts the steps taken in the current study to produce, identify, and analyze voice quality.

Vol. 02, No. 04; 2023
ISSN: 2583-6560



df

Figure 1 Steps of Producing, Identifying and Analyzing Voice Quality

A rating exercise session was conducted to identify obvious changes in voice quality that result in a change in interpersonal meaning. The findings from this preliminary investigation into the voice quality validation process can be interpreted as a new challenge in text and multimodal

Vol. 02, No. 04; 2023
ISSN: 2583-6560

analysis validation. These findings are critical for the investigation and development of voice quality features in call center data validation. Furthermore, rigorous validation will aid in the reduction of subjectivity. The present study is anticipated to have significant benefits for voice quality training in the call center sector as well as for methodological considerations in the validation of multimodal text analysis. Four categories, volume (Loud/Soft), pitch (High/Low), tension (Tense/Lax), and rhythm (Fast/Slow), were used by the raters to help them detect changes in voice quality. Finding the level of agreement between two raters' evaluations of the same piece of material is the aim of the interrater agreement measurement. Although 100% agreement is ideal, it is not always feasible or needed. High levels of consistency and agreement are preferred. Setting a goal of total agreement might also be unreasonable, impossible to accomplish, and even discourage raters from being honest in their assessments.

As a result, complete agreement among the three raters was not sought in the present study when identifying the change in voice quality features. The goal of the rating exercise was to boost the credibility of the current voice quality study. In language and behavioral studies, it is suggested that a rating exercise be conducted on a sample selection of data, ranging from 20% to 30% of the data (Liu et al., 2005; Wiebe, Wilson, & Cardie, 2005). Approximately 30% of the total interactional data were chosen, as shown in Table 2, to be the basis for the rating exercise. A sample size of approximately 30% is considered adequate to produce a trustworthy rating (see Liu et al., 2005; Wiebe et al., 2005). Approximately 41 working hours were used to complete the rating exercise.

Table 2 Selected Calls for Conducting the Rating Exercise

Transcription Code	Duration	Words
T1	10 mins 04 secs	1,474
T2	10 mins 06 secs	1,523
T3	9 mins 23 secs	1,682
T4	14 mins 36 secs	3,015
T5	13 mins 43 secs	1,665
T10	3 mins 44 secs	676
Total	**1 hour 1 min 36 sec**	**10,035**

Three raters participated in the rating process. Table 3 provides the profile of the raters. The accuracy and reliability of interrater agreement in call center conversations voice quality research can be significantly impacted by the linguistic background and call center expertise of raters. The raters had solid educational backgrounds in English and language studies and were linguistic and language experts. For instance, one associate professor and two others had PhDs in applied linguistics. The reliability of the interrater results was strengthened by their solid linguistic and educational backgrounds.

Vol. 02, No. 04; 2023
ISSN: 2583-6560

Table 3 Profile of Raters for Interrater Agreement

Rater	R1	R2	R3
Age	25-30	30-35	35-40
Nationality	Hong Kong Chinese	Hong Kong Chinese	Mainland Chinese
Gender	Female	Female	Female
First language	Cantonese	Cantonese	Putonghua
English study	23 years	20+ years	25+ years
Education	• BA in Japanese and English Studies • MPhil in Japanese Studies	• BA in Translation • Master in Translation	• BA in language • Master in English studies • PhD in applied linguistics
Major	Language	Translation and language	Linguistics
English teaching experience	• 2 years; • workshop and secondary school English teacher	• 2 years; • Tertiary education • Teaching associate	• 10 years; • University Associate Professor
Call center research experience	2 years	2 years	1.5 years
Call center visits	5 call centers in the Philippines and India	10 call centers in Hong Kong, Mainland China and the Philippines	10 call centers in the Philippines and Mainland China
Hour(s) involved	10 hours	14 hours	10 hours
Follow-up interview	27 mins 12 secs	1 hour 49 mins 34 sec	30 mins 08 sec
Transcripts	**T1-5, T10**	**T1-5, T10**	**T1-5, T10**

Call center experience may help raters better recognize specific voice quality characteristics that are important in call center conversations, such as active listening cues, tone and pitch variations. Additionally, they might be more familiar with call center procedures and customer service benchmarks, which can make it easier for them to distinguish between interactions of high and low service quality. The raters had visited five to ten call centers and were knowledgeable about the call center industry as well as the specific language features of call center conversations, which proved useful when conducting the rating exercise. Furthermore, designing rating criteria is a critical component of voice quality research. The current study's rating was created with the help of linguistically diverse groups and call center professionals and aided in the collection of various variables influencing quality conversation in call centers.

Raters from various linguistic backgrounds may have cultural and linguistic biases that influence their interpretation. The impact of linguistic background and call center knowledge, on the other hand, can be mitigated through adequate training and preparation of raters. For example, providing training materials on voice quality knowledge related to rating exercises

Vol. 02, No. 04; 2023
ISSN: 2583-6560

can improve rating accuracy. Raters can be trained to be consistent, reliable, and impartial to reduce bias and increase interrater agreement. The rating exercise was conducted according to the following procedure: Each rater received a copy of Chapter 6 of van Leeuwen's (1999) book Speech music and sound, which served as the main framework for the current voice quality analysis, two weeks before the test. Next, each rater received a one-hour individual briefing. During the testing, the raters were given six blank transcripts as well as the original audio tapes. Instead of giving them excerpts from each call, they were given the entire transcript. This was done because the present study emphasized that the meaning of a call center conversation can only be understood holistically. The raters' job was to note any obvious changes in relative voice quality on the transcripts. The change in voice quality could be identified by a syllable, word, or turn. The raters were asked to categorize noticeable changes in voice quality levels in terms of pitch (High/Low), volume (Loud/Soft), tension (Tense/Lax), and rhythm (Slow/Fast). A voice quality feature could fall under more than one category, such as a High, Tense, or Loud voice, as they are not mutually exclusive. Each rater listened to the audio files three times in total during the test. The goal of the first listen was to comprehend the entire call's content. They completed their first scoring in the transcripts after their second listen. The raters were asked to proofread and double-check their evaluations during the final listen. Each call was handled in between thirty and forty-five minutes. Each rater underwent an individual interview after the test. Before trying to reach a consensus among raters, the potential reasons for disagreements regarding identifications were discussed. The rating exercise had two distinct outcomes, from agreement on I) identifying general stages of voice change to II) identifying voice quality features of specific voice changes. Overall, in call center conversation voice quality research, a number of variables influence the precision and dependability of interrater agreement. The present study attempted to consider these variables during the interrater agreement design, data collection, and analysis to generate accurate and objective insights.

3. FINDINGS AND DISCUSSION: INTERRATER AGREEMENT

It is necessary to discuss voice quality changes in relation to the call center generic call stages that were present to review the rating process. The texts identified the following generic stages of insurance call center conversations: *Opening ^{Identification} ^{Purpose} ^(Clarification) ^ [(Objection) ^ (Legitimization)] ^ {Servicing} ^ (Transfer ^ Transfer-Opening ^ Transfer-Identification ^ Transfer-Purpose ^ Transfer-Closing) ^ (Closing)* (c.f. Wan 2023 for a detailed discussion of generic stage elements). The caret sign represents the stage order, brackets ( ) represent the genre's optional elements, square brackets [ ] represent recursive elements, and brace brackets { }represent recurring stages (cf. Halliday & Hasan, 1980). The voice quality changes noted by Raters 1, 2, and 3 in the corresponding generic stages in Transcriptions 1 through 5 and 10 (T1 to T5 and T10) are compiled for a thorough discussion of the standard stages in call center calls. Agreement occurs when at least two out of the three raters make a change to the same generic stage. The research reveals an 80% agreement rate in the generic analysis between the three raters (32 out of 40 examples). Interrater agreement is influenced by a variety of factors, including the difficulty of the task or material being evaluated, the clarity of the evaluation criteria, and individual differences in human judgment among raters. The degree of agreement among raters can be impacted by these factors, which can lead to differences in perceptions or interpretations. In the present study, some degree of disagreement between raters cannot be avoided. However, a rating of 80% agreement is deemed adequate to reflect the natural diversity and range of opinions among human interpretation, producing accurate and reliable results.

Vol. 02, No. 04; 2023
ISSN: 2583-6560

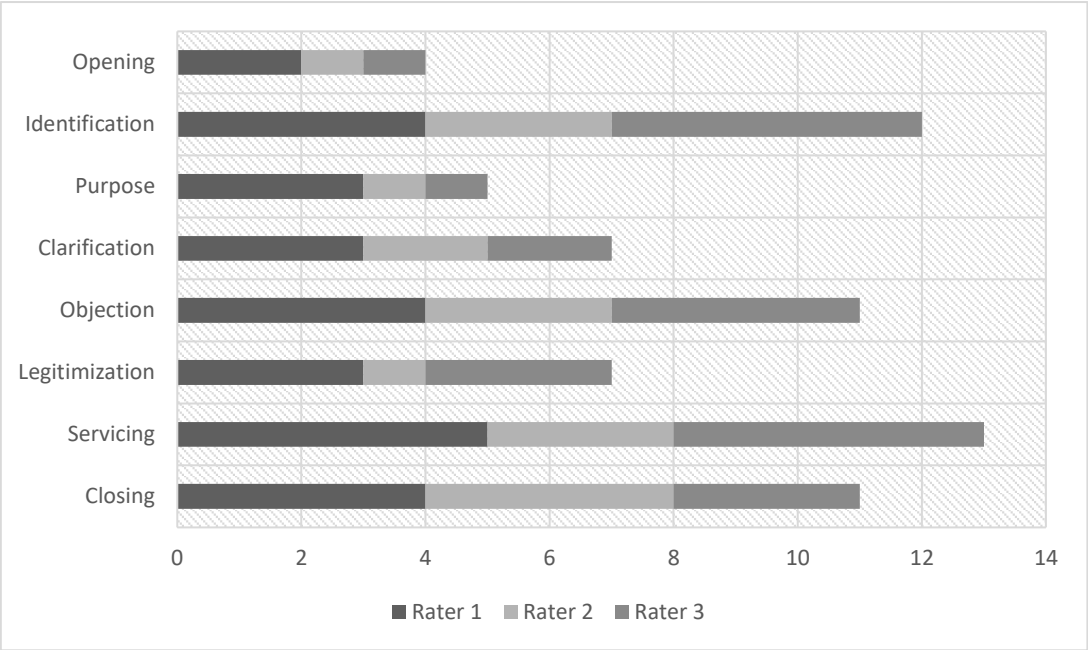


Figure 2 Rater Agreement on Changes in Voice Quality Features in Different Generic Stages

Forey and Lockwood (2007) state that "certain features, such as a breakdown in communication, occurred during specific stages," and that "the problems tended to occur during the purpose and service stages" (p. 318). Explicit changes in voice quality occurred during the generic stages of Servicing (top category with *n=13*), Identification (*n=12*), Objection (*n=11*) and Closing (*n=11*), as shown in Figure 2. The raters concurred that explicit voice quality changes occur frequently in these stages:

The generic stage of *Identification* establishes and verifies the callers' identities. Names, addresses, telephone numbers, social security numbers, and policy numbers are among the personal data that are frequently checked in this stage. The customer and CSR must concentrate to ensure that the information is accurate and correct. This increased focus can lead to changes in voice quality and emotional intensity. In addition, both the customer and the representative may feel anxious at this point because they are discussing personal information with a stranger. The exchange of personal information could be sensitive and dangerous. The CSR must carefully verify the customer's identity to avoid fraud or impersonation. Therefore, the CSR may use a formal tone to convey the operator's confidence and a sense of authority. During the initial stages of the conversation, their voice quality may be impacted by this seriousness, resulting in frequent alterations in voice quality features.

The customer may be angry, frustrated, or irritated as a result of their problem during the *Servicing* stage, whereas the CSR may be sympathetic or concerned. These emotions can be expressed through changes in voice quality. In addition, the servicing stage may include complex information or instructions pertaining to the customer's problem resolution. The CSR may be required to explain technical or procedural details in a clear and concise manner, which may affect the quality of their voice and rate of delivery. Compliance issues, such as adhering to internal process guidelines or meeting regulatory requirements, may be addressed during the *Servicing* stage. The CSR might need to use a more formal tone of voice to ensure compliance

Vol. 02, No. 04; 2023
ISSN: 2583-6560

and project their authority. In general, a detailed, frequently complex, and emotionally charged information exchange occurs during the *Serving* stage.

In the *Objection* stage, the customer may propose an objection by rejecting the CSR's solution and providing additional information to build up the complaints. Their pitch may rise and fall depending on the seriousness of the issue. There may also be more frequent pauses or hesitations in the speaker's rhythm pattern as they process and organize their thoughts around the new information or approach or as they recall a memory. Emotions are significant at this stage, and discussing complaints can elicit a variety of emotional responses. The customer's voice quality can easily reflect their level of anxiety and uncertainty. Because objections and complaints frequently involve the customer's negative emotions, such as frustration, anger, or disappointment, the CSR may react defensively (using a harsher tone and faster pace) or apologetically depending on the situation. To acknowledge the customer's concerns, show empathy, and effectively resolve the issue, the CSR must demonstrate active listening skills. They may alter their voice quality with empathetic tones or apologies to establish rapport. To convey trust, speakers may lower their voices. Using a more warm and friendly tone aims to set a more positive tone for the call. The CSR may employ different voice quality techniques to persuade the customer and reach an agreement. As a result, during the *Objection* stage, when conflicts may arise, calmness and control through voice quality can help deliver effective resolutions.

At the *Closing* stage, the CSR may formalize the resolution by summarizing the agreed-upon actions, next steps, or follow-ups. This stage frequently requires the CSR to speak in a clear and concise manner. The closing stage is also an important opportunity for the CSR to make a long-term positive impression on the customer. They may use polite and friendly language and appropriate voice quality to finish the conversation on a positive note, leaving the customer satisfied with the service provided. The customer's satisfaction or dissatisfaction with the resolution can be easily conveyed through voice quality. If the customer is satisfied, they may express relief, happiness, or gratitude; if they are dissatisfied, they may express disappointment, frustration, or anger in their tone. These emotional changes are clearly reflected in changes in voice quality features located in different generic stages, as evidenced by the interrater agreement.

After discussing the findings of voice quality changes in the generic stages, individual and specific voice quality features in specific phrases or words should be considered. Call center negotiation is distinguished by a shift in lexicogrammatical choices as well as a change in voice quality features. The second level of the rating exercise is to identify specific voice quality features: There are eight distinct features of volume (Loud/Soft), pitch (High/Low), tension (Tense/Lax), and rhythm (Fast/Slow). This was accomplished by selecting 74 representative examples from six sample transcripts while keeping the raters anonymous. Two rules are established prior to the tests. The first rule is that contradictions are not permitted (for instance, neither Loud and Lax nor Low and High can coexist). Another rule is that the criteria for agreement are met if two of the three raters choose the same voice quality. For example, the results of Example 2 are an example of agreement; the results of Example 63 are an example of disagreement between three raters.

Vol. 02, No. 04; 2023
ISSN: 2583-6560

E.g. 2 Transcript 1 Turn 31	Rater	Loud	Soft	High	Low	Tense	Lax	Fast	Slow
	R1	✓							
✓	R2			✓					
	R3	✓		✓					

E.g. 63 Transcript 5 Turn 56	Rater	Loud	Soft	High	Low	Tense	Lax	Fast	Slow
	R1	✓						✓	
×	R2	✓		✓					
	R3			✓					

The three raters correctly identified specific voice quality features in 48 of the 74 examples in the six transcripts with a 65 percent accuracy rate. According to the interrater response, there was 80% agreement at the Level I Generic stage level. Given that the interrater agreement at the Level II Leixcogrammatical level was only 65%, it was less consistent than at Level I. This observation can be explained using Matthiessen's (2007) process of constructing interpersonal meaning, as shown in Figure 3.

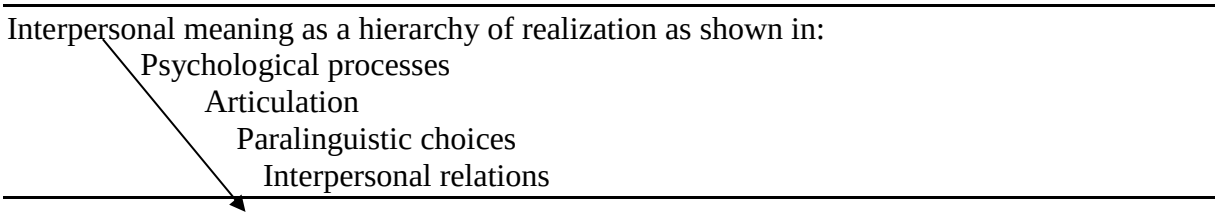


Figure 3 Construction of Interpersonal Meaning (Matthiessen, 2007)

Changes in voice quality affect articulation and alter the paralinguistic choices used to create different interpersonal meanings (Matthiessen, 2007). According to Martin, language is divided into three strata: semantics, lexicogrammar, and phonology (2007). Phonology and sound quality, including voice quality, are examples of concurrent expression forms. Generic analysis in lexicogrammar is a broader discourse category than extremely specific voice quality features. As a result, the rate of agreement between raters for the generic stage was 80%, while the rate of agreement for the voice quality feature was 65%. Furthermore, the data showed that the agreement strengthened as raters gained more experience. Since Transcript 10 was the shortest text at the outset, it was prepared for the first test. However, out of the six transcripts used to represent agreement, the three raters' overall level of agreement was the lowest for Transcript 10. Five examples of preset voice quality came with Transcript 10. The three raters agreed on only 25% (1/4) of the content of Transcript 10. The agreement rate rose and was higher in Transcripts 1 to 5 after the raters repeated the test and invested more time and effort. For Transcripts 1 through 5, the average level of agreement was 67%. For instance, the agreement rates were 75% (6/8) in T1, 69% (9/13) in T2, 78% (7/9) in T3, 65% (20/31) in T4, and 56% (5/9) in T5. To recognize specific voice quality features, it is therefore predicted that if the raters or the CSR had more training, they would reach a higher rate of agreement. The CSR must develop their ability to distinguish between different voice quality features and comprehend the interpersonal connotations that are elicited by hearing such features.

Vol. 02, No. 04; 2023
ISSN: 2583-6560

Additionally, out of the eight voice quality features, it appears that the three raters recognize the change of the High (n=22), Loud (n=11), and Fast (n=11) voice quality features in the total of 74 examples with "100% agreement". These three voice quality features may be easier for raters to recognize in the interrater exercise because they are more salient and noticeable to associate with emotional changes, particularly during communication breakdown or crucial complaint moments. Pitch variations in particular were the simplest for raters to assess in comparison to other parameters such as tension (Tense/Lax). Changes in tension, particularly changes in throat tension, are the most challenging categories for raters to identify in interrater exercises. This is because tension is less overtly present and audible than other aspects of voice quality. Additionally, the speaker's traits, accent, and speech practice can all affect the level of tension. For raters unfamiliar with the background and vocal characteristics of the speaker, this parameter (Tense/Lax) might be challenging to identify and evaluate, which could lead to lower interrater agreement.

The management of the call center was very interested in finding solutions to the problems with call center conversations and researching techniques for handling challenging negotiations, according to the results of the follow-up conversation with call center professionals. As mentioned above in the agreement result of Level I, Generic stages, the majority of changes in voice quality features are discovered during points of negotiation in the stages of *Identification*, *Objection*, *Servicing*, and *Closing*. Among the eight voice quality features used in this rating exercise, High, Loud, and Fast are the easiest to distinguish. If the call center provided more voice quality training and assessment practices to CSRs, they would be able to develop the skills required for identifying changes in voice quality features and identifying negotiation opportunities in complex calls. Understanding the problem and determining its root cause will help CSRs perform better and successfully handle more complex calls in the long run.

4. FINDINGS AND DISCUSSION: PRAAT

In addition to human interpretation, to facilitate the identification of the features of voice quality, a computer speech software program was used in the present study. Praat (meaning "talk" in Dutch) is the name of the free computer program created to detect a change in the relative pitch of the sounds used. Researchers working in the fields of phonetics and speech analysis use this program to visualize, quantify, and analyze different aspects of voice quality. The user-friendly interface allows researchers to manipulate audio files and perform acoustic analyses of sound characteristics. Linguists and phoneticians use this software to examine speech sounds frequently (Halliday & Greaves, 2008). It is possible to use Praat to support specific sound file examples in the present study. Decibel (dB), Hertz, the length of the utterance, and the range of energy in a spectrogram are a few of the crucial functions that Praat can facilitate with the validation of the voice quality features in the call center conversation. Praat allows researchers to analyze voice intensity, which is related to loudness. The decibel level (dB) of a call center conversation is measured. The absence of loudness is 0 dB. In this study, intensity analysis is used to quantify the level of voice variability and investigate how it relates to vocal quality. A Loud voice can be projected at up to 81.17 dB, while a Low voice can be projected at up to 65.47 dB in call center audio files. In addition, the pitch analysis of the software enables researchers to determine the fundamental frequency of the voice. In the call center conversation file, the Hertz ranges from 50 to 500 Hz. The speaker projects the High voice up to 284.6 Hz and the Low voice down to 158. Using this computer software, changes in pitch can be easily detected. Third, Praat includes a spectrogram analysis tool for visualizing

Vol. 02, No. 04; 2023
ISSN: 2583-6560

the spectral content of the voice signal. Using a spectrogram to identify different voice qualities such as Tense, i.e., with more energies shown, and Lax, i.e., with less energy input and breathiness in the sound files. By observing the energy input of the conversations visualized in the spectrogram, Tense and Lax voice quality features can be distinguished. Finally, by calculating how many words or syllables can be uttered per second, Praat also calculates the length of the utterance. Obtaining accurate measurements of speech acoustic properties is effective in validating the voice quality features of rhythm (Fast/Slow). The validated measures can then be used to examine the relationships between vocal quality and communicative purpose, providing insight into how vocal quality affects speech perception.

However, it appears that Praat cannot be selected as the primary analytical tool in the interrater agreement due to the following considerations. Analyzing voice quality is a conceptual analysis (van Leeuwen, 1999). In the present study, the main goal is to investigate how interpersonal meaning potential is created in conversation. The study advocates examining the overall information flow of the text as well as the potential for interpersonal meaning. This can only be partially aided by software. All of the call center conversations that were recorded for this study happened at authentic workplaces. Background noise annoyances such as typing and other external sounds were frequently found in the recording because the audio files were not intended to be used for experimental purposes. The results of using Praat could be unfavorably impacted by this. On occasion, speakers stepped away from the microphone. The recording quality was adversely affected by this. Some fundamental technical problems also existed: Praat would be unable to recognize or multicode a variant because voice quality features can be categorized into different groups simultaneously. There are times when a voice quality phase began with high, tense features and ended with a breathy voice. Praat could not completely replace human interpretation from raters. Praat only partially supported and aided in the validation of the rating exercises in the present study. Although it is acknowledged that there is still room for improvement in the validation of the results, the validation in the present study marks a significant advancement in call center training and voice quality analysis. The results of this interrater agreement and voice quality analysis were deemed innovative and beneficial in a follow-up conversation with call center experts.

5. CONCLUSION

The goal of naturalistic researchers is to observe and document human communication and behavior in their unrestricted natural settings. Studying voice quality features can provide insights into the dynamics of real-life interactions because they are influenced by situational or emotional factors that naturally occur in everyday conversation. The application of linguistic theory to real-world contexts is the focus of applied linguistics. It investigates how language is used in various contexts and how it can be used to achieve specific goals. One significant area of applied linguistics is voice quality research, which deals with the study of voice-related features in speech such as volume, rhythm, pitch, and tension. Voice quality features contribute to signify various social and cultural roles, express emotion, and convey meaning. In applied linguistic research, the validation of voice quality feature analysis is closely related to the development of research methodologies. For measuring and analyzing voice features, accurate methods are needed. A well-validated method ensures that the research data are reliable and trustworthy. Voice research in applied linguistics must include the validation of voice quality features. Voice quality features are acoustic characteristics of speech that reveal information about the psychological and emotional state of the speaker. These characteristics can have a significant impact on human communication. In the present study, the objective is to increase

Vol. 02, No. 04; 2023
ISSN: 2583-6560

the validity of voice quality features. A multimodal analysis was conducted on the voice quality features and interpersonal meaning found in the transcribed calls in accordance with the goal and plan of the present study. The interrater exercises were carried out by raters to identify features of voice quality, such as High/Low, Loud/Soft, Fast/Slow and Tense/Lax. The findings demonstrate the validity of identifying and measuring voice quality features in call center interactions using both interrater agreement and the speech analysis software Praat. The findings strongly suggested that voice quality features are not subjective and can be objectively assessed by raters and measured by Praat in a variety of generic stages and lexical phrases. Validated and reliable measurements can enable researchers to investigate the relationship between voice quality characteristics and communicative goals such as understanding, persuasion, or social influence. The present study outlines the voice quality and rating exercise validation methods created to support ongoing social science research. The results of the present study will be useful to researchers, call center managers and agents, and other speech and communication analysis specialists. By identifying important voice quality features and validating the accuracy of their measurement, the present study advances knowledge of the qualitative research validation method in the field of applied linguistics as well as knowledge of call center discourse and its role in fostering effective communication in professional contexts.

REFERENCES

Atkinson, J. M., & Heritage, J. (1984). *Structures of social action: Studies in Conversation Analysis*. Cambridge: Cambridge University Press.

Blum-Kulka, S., Huck-Taglicht, D., & Avni, H. (2004). The social and discursive spectrum of peer talk. *Discourse Studies, 6*(3), 307-328.

Bruce, G. (1992). Comments. In J. Svartvik (Ed.), *Directions in corpus linguistics: Proceedings of the nobel symposium 82, Stockholm August 4-8, 1991* (pp. 145-147). New York: Mouton de Gruyter.

Chafe, W. L. (1993). Prosodic and functional units of language. In J. A. Edwards & M. D. Lampert (Eds.), *Talking data: Transcription and coding in discourse research* (pp. 33-43). Hillsdale, NJ: Lawrence Erlbaum.

Creswell, J. W. (2007). *Qualitative inquiry & research design: Choosing among five approaches* (2 ed.). London: Sage.

Denzin, N. K., & Lincoln, Y. S. (2005). *The Sage handbook of qualitative research* (3 ed.). Thousand Oaks, CA: Sage.

Eggins, S. (2004). *An introduction to Systemic Functional Linguistics* (2 ed.). New York, London: Continuum.

Eggins, S., & Slade, D. (1997). *Analysing causal conversation*. London: Cassell.

Ehlich, K. (1993). HIAT: A transcription system for discourse data. In J. A. Edwards & M. D. Lampert (Eds.), *Talking data: Transcription and coding in discourse research* (pp. 123-148). Hillsdale, NJ: Lawrence Erlbaum.

Forey, G. (2004). Workplace texts: Do they mean the same for teachers and business people? *English for Specific Purposes, 23*(4), 447-469.

Forey, G., & Lockwood, J. (2007). 'I'd love to put someone in jail for this': An initial investigation of English in the Business Processing Outsourcing (BPO) Industry. *English for Specific Purposes, 26*(3), 308-326.

Halliday, M. A. K. (1978). *Language as social semiotic: The social interpretation of language and meaning*. London: Edward Arnold.

Vol. 02, No. 04; 2023
ISSN: 2583-6560

Halliday, M. A. K., & Greaves, W. (2008). *Intonation in the grammar of English*. London: Equinox.

Halliday, M. A. K., & Hasan, R. (1980). Text and context: Aspects of language in a social-semiotic perspective. *Sophia Linguistica (Working Papers in Linguistics), 6*, 4-91.

Hammersley, M. (1992). *What's wrong with ethnography? Methodological explorations*. London: Routledge.

Harvey, N. (1993). Text analysis for specific purposes. *Prospect*, *8*(3), 24-41.

Hill, C. E. (1982). Counselling process research: Philosophical and methodological dilemmas. *The Counselling Psychologist, 10*(4), 7-19.

Himonides, E. T., & Welch, G. F. (2005). Building a bridge between aesthetics and acoustics with new technology: A proposed framework for recording emotional response to sung performance quality. *Research Studies in Music Education, 24*(1), 58-74.

Johnstone, B., Andrus, J., & Danielson, A. E. (2006). Mobility, indexicality, and the enregisterment of 'Pittsburghese'. *Journal of English Linguistics, 34*(2), 77-104.

Juslin, P. N., & Sloboda, J. A. (2001). *Music and emotion: Theory and research*. New York: Oxford University Press.

Kurkul, W. W. (2007). Nonverbal communication in one-to-one music performance instruction. *Psychology of Music*, *35*(2), 327-362.

La, F., & Davidson, J. W. (2005). Investigating the relationship between sexual hormones and female western classical singing. *Research Studies in Music Education, 24*(1), 75-87.

Lapadat, J. C., & Lindsay, A. C. (1999). Transcription in research and practice: From standardization of technique to interpretive positionings. *Qualitative Inquiry, 5*(1), 64-86.

Lazaraton, A. (2002). Quantitative and qualitative approaches to discourse analysis. In M. McGroarty (Ed.), *Discourse and dialogue. Annual review of applied linguistics 22* (pp. 32-51). Cambridge: Cambridge University Press.

Leijssen, M. (2006). Validation of the body in psychotherapy. *Journal of Humanistic Psychology*, *46*(2), 126-146.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.

Liu, M., Chen, X., Rubin, K. H., Zheng, S., Cui, L., Li, D., et al. (2005). Autonomy vs. connectedness-oriented parenting behaviours in Chinese and Canadian mothers. *International Journal of Behavioral Development, 29*(6), 489-495.

Martin, J. R. (2007, December, 10-12). *Multimodality - some issues.* Paper presented at the Semiotic Margins: Reclaiming meaning, Department of Linguistics, University of Sydney, Australia.

Matthiessen, C. M. I. M. (2007b, 17 December). *The interpersonal lexicogrammar of assessment.* Paper presented at The Hong Kong Polytechnic University, Department of English Seminar, Hong Kong, China.

Maxwell, J. A. (2005). *Qualitative research design: An interactive approach* (2 ed. Vol. 41). California: Thousand Oaks.

McCoyd, J. L. M., & Kerson, T. S. (2006). Conducting intensive interviews using email: A serendipitous comparative opportunity. *Qualitative Social Work, 5*(3), 389-406.

O'Sullivan, M. (2003). The fundamental attribution error in detecting deception: The boy-who-cried-wolf effect. *Personality and Social Psychology Bulletin, 29*(10), 1316-1327.

Orbelo, D. M., Testa, J. A., & Ross, E. D. (2003). Age-related impairments in comprehending affective prosody with comparison to brain-damaged subjects. *Journal of Geriatric Psychiatry and Neurology, 16*(1), 44-52.

Vol. 02, No. 04; 2023
ISSN: 2583-6560

Pennington, M. (2002). Examining classroom discourse frames: An approach to raising language teachers' awareness of and planning for language use. In H. Trappes-Lomax & G. Ferguson (Eds.), *Language in language teacher education* (pp. 149-172). Amsterdam: John Benjamins.

Poskiparta, M., Kettunen, T., & Liimatainen, L. (2000). Questioning and advising in health counselling: Results from a study of Finnish nurse counsellors. *Health Education Journal, 59*(1), 69-89.

Potter, J. (1996). Discourse analysis and constructionist approaches: Theoretical background. In J. T. E. Richardson (Ed.), *Handbook of qualitative research methods for psychology and the social sciences* (pp. 125-140). UK: The British Psychological Society.

Psathas, G., & Anderson, T. (1990). The 'practices' of transcription in Conversation Analysis. *Semiotica, 78*(1/2), 75-99.

Sacks, H. (1992). Lectures on conversation. In G. Jefferson with introduction by E. A. Schegloff (Ed.), (Vol. 1 and 2). Oxford: Blackwell.

Samuelsson, C., Nettelbladt, U., & Lofqvist, A. (2005). On the relationship between prosody and pragmatic ability in Swedish children with language impairment. *Child Language Teaching and Therapy, 21*(3), 279-304.

Scherer, K. R., & Ceschi, G. (2000). Criteria for emotion recognition from verbal and nonverbal expression: Studying baggage loss in the airport. *Personality and Social Psychology Bulletin, 26*(3), 327-339.

Scott, D., & Usher, R. (1996). *Understanding educational research*. New York and London: Routledge.

Short, M. H., Semino, E., & Wynne, M. (2002). Revisiting the notion of faithfulness in discourse presentation using a corpus approach. *Language and Literature, 11*(4), 325-355.

Silverman, D. (2000). *Doing qualitative research: A practical handbooks*. London: SAGE.

Strauss, A. L. (1987). *Qualitative analysis for social scientists*. Cambridge: Cambridge University Press.

Trappes-Lomax, H. (2004). Discourse analysis. In A. Davies & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 133-164). Malden, Mass: Blackwell.

van Leeuwen, T. (1999). *Speech, music, sound*. London: Palgrave Macmillan.

van Leeuwen, T. (2005). *Introducing social semiotics*. New York: Routledge.

Wan, Y. N. (2017). Construing negotiation: The role of voice quality features in American- Filipino business telephone conversations. *Language and Dialogue*, 7(2), pp. 137-163.

Wan, J. Y. N. (2023). Structuring logical relations in workplace English telephone negotiation. *International Journal of Language Studies, 17*(1), *71-96.*

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation, 39*(2-3), 165-210.