Unravelling Emotional Nuances: A Cross-Linguistic Analysis of Sentiment Differences in Multilingual Movie Versions

Adam Wynn[1[0000−0002−1631−2151]], Jingyun Wang[1[0000−0001−9325−1789]], and Xiaoyan Li[2[0000−0002−1209−7722]]

[1] Durham University, Durham, United Kingdom
{adam.t.wynn,jingyun.wang}@durham.ac.uk
[2] Kyushu University, Fukuoka, Japan

Abstract. To facilitate automatic knowledge creation for the learners of intercultural communication, we propose a pipeline that combines AI analytics with human-in-the-loop evaluation. A case study was conducted on the English and Japanese versions of the film 'Spirited Away'. Speech segments from both versions were extracted and processed using Valence-Arousal-Dominance (VAD) scores and discrete emotion classification, generated by pre-trained Wav2Vec 2.0 models. The segments with the highest discrepancies in emotions were identified for further analysis by a group of human raters. This approach not only improves the understanding of emotional differences in dubbed media but also benefits researchers studying cross-cultural communications by assisting knowledge discovery and the creation of knowledge pools. These insights may also contribute to the design of intelligent tutoring systems (ITS) for cross-cultural communication education, enhancing the overall learning experience for learners.

Keywords: Cross-linguistic analysis, Speech Emotion Recognition, Cultural Learning, Emotion Perception, Wav2Vec 2.0

1 Introduction

Cross-cultural communication education focuses on developing the students' knowledge of how to operate in a culture other than in their own [3]. However, it is time-consuming for instructors to prepare suitable materials manually. This challenge is particularly significant when it comes to addressing emotional differences across cultures, as emotion is deeply tied to an individual's cultural background. Whilst some researchers consider emotion to be a universal construct and that a large part of emotional experience is biologically based, emotion is also heavily influenced by the environment an individual grew up in [4]. Therefore, the way humans express emotions can vary across different languages due to different cultural backgrounds [7].

However, in the field of cross-linguistic sentiment analysis, existing tools tend to focus on text rather than speech, leaving gaps in the comparison of emotional nuances across languages [18]. Developing tools that account for these cultural differences, especially in speech, could significantly enhance educational resources in cross-cultural communication, enabling students to better understand how emotion varies across cultural contexts.

On the other hand, the emotional discrepancies between the same word or phrase under different cultural settings could, when translated directly, often cause confusion, and even lead to conflicts [9]. Analysing emotional content across languages manually is a complex and time-consuming process, especially when dealing with spoken audio. Human annotators often face challenges in consistently identifying and categorizing emotions because emotional perception is subjective where different raters might interpret the same audio in different ways based on their cultural background. Identifying subtle emotional cues requires in-depth knowledge of both languages and culture [17]. Moreover, emotional experiences are often dulled when listening in a second language compared to a listener's native language [2]. Intelligent tutoring systems (ITS) could address these challenges by integrating AI-driven emotion analysis into language learning environments, providing students with real-time feedback on how emotional expression varies across languages.

Therefore, this paper aims to design a AI-based pipeline to support the exploration of the subtle emotional differences between multilingual versions of movies and facilitate the automatic creation of knowledge pools for learners of intercultural communications. To achieve this, a series of deep learning models automatically identify and extract segments of audio, eliminating the need for manual inspection of each sentence. Subsequently, for effectiveness evaluation we conducted a case study on the movie 'Spirited Away' by examining the sentiment conveyed in the audio of the English and Japanese versions to discover the segments that are significantly different in emotion.

In particular, this paper compares how native English and Japanese speakers, as well as non-native bilingual speakers, perceive emotional content in these identified segments. The focus group discussions highlight the importance of a "human-in-the-loop" approach and the importance of human input in understanding emotions that may be missed by AI models alone, and the data analysis is intended to explore how these differences can inform the design of intelligent tutoring systems (ITS) for intercultural communication education, enabling more accurate and culturally aware learning experiences for students.

By automating the process of detecting emotional disparities in subtitles, we aim to facilitate cross-cultural communication analysis and foster a deeper understanding of how cultural contexts shape linguistic expression. This project has the potential to streamline research efforts and contribute to the broader field of intercultural studies. Through the development process of the pipeline, we aim to answer the following research questions:

- **RQ1:** Are the Valence-Arousal-Dominance (VAD) model and discrete emotion classification models appropriate for analysing cross-linguistic emotional content in audio data?

- **RQ2:** How do individuals from different cultural and linguistic backgrounds perceive emotional content in the audio of English and Japanese versions of the same movie, and what implications does this have for education?

In this pipeline, we have, for the first time, devised a subtitle matching algorithm with high accuracy, trained an emotion classification model using high-dimensional audio representations, and performed quantitative analysis to extract discrepancies using the projected VAD states. Through this project, we contribute to the field of cross-linguistic sentiment analysis and provide valuable insights into how cultural differences shape emotional perception for researchers and learners of cross-cultural communication. The findings could inform ITS design, allowing for adaptive learning experiences that help students develop a deeper understanding of how cultural contexts influence emotional expression. By combining the power of AI-based analytics with human expertise, this paper proposes a robust pipeline that extracts the emotional nuances of language and contributes to the advancement of cross-linguistic sentiment analysis.

2 Related Work

2.1 Emotional Theory

In the study of emotions, there are two primary approaches to representing emotion: using Discrete Emotions and the VAD model. Central to the discrete approach is Ekman's identification of six basic emotions: anger, joy, surprise, disgust, fear, and sadness [4]. These emotions, which are fundamental to human experience, form the basis for many psychological and computational models in understanding emotional responses across different cultures and contexts, making them critical to sentiment analysis and cross-cultural emotional studies.

Another approach which quantifies emotions and provides a more detailed view of emotional expression, is the Valence-Arousal-Dominance (VAD) model [12] that represents emotions as a 3-dimensional vector with each dimension ranging from -1 to 1. Valence ranges from negative to positive and expresses the pleasant or unpleasant feeling about something, arousal measures the intensity of the emotion ranging from calm to excitement, and dominance reflects the level of control, from submissive to dominant. This model has been widely adopted in psychological studies and more recently in machine learning for sentiment analysis [16]. By capturing emotions in this multidimensional space, the VAD model offers a more nuanced understanding of emotional expression, particularly in cross-cultural contexts, providing the foundation for further computational approaches, such as audio processing and speech recognition.

2.2 Audio Processing

To prepare continuous speech signals for machine learning models, several preprocessing steps are necessary, including noise reduction and feature extraction.

Noise reduction algorithms reduce unwanted background noise in audio recordings. Traditionally, noise reduction algorithms such as the noisereduce algorithm [13] consist of spatial filtering but if the entire recording has inconsistent noise levels, the de-noised audio outputs will be unsuitable for audio processing. To address this, deep learning approaches, such as Nvidia's CleanUNet model [6], propose to leverage encoder-decoder architectures and self-attention mechanisms to achieve more effective denoising.

Traditional methods for feature extraction involve breaking down raw waveforms into frames, and performing windowing techniques and Fourier transforms to convert signals from the time domain to the frequency domain. Features such as Mel Frequency Cepstral Coefficients [15] have been widely used to capture essential acoustic features, particularly for speech and music processing [14].

2.3 Speech Emotion Recognition

Deep learning enables the use of the rich high-dimensional representations of audio segments, and subsequently allows us to translate this encoding into emotion scores using a decoder model. A prominant example is Wav2vec 2.0 [1] which employs a transformer network to build a contextualised representation, trained similarly on a contrastive task that requires 100 times fewer data. Wav2vec 2.0 proposed several variants, including XLSR-Wav2vec 2.0 (XLSR), which learns cross-lingual speech representations by performing the contrastive training task in more than 56,000 hours in 53 languages and outperforms the best-known results over several benchmarks. However, Wav2vec models are not robust to noisy speeches so speech segments should be de-noised before processing them. Alternatively, HuBERT [5] uses hidden units that are analogous to word tokens in BERT to transform speech data into a more language-like structure.

Pepino et al. (2021) investigated the use of Wav2Vec 2.0 embeddings for speech emotion recognition (SER) by leveraging multiple layers of the pre-trained model with adaptive weighting. Their approach applied transfer learning, utilising a version of Wav2Vec 2.0 fine-tuned for automatic speech recognition (ASR) to enhance SER performance. However, they discovered that ASR fine-tuning can discard features crucial for emotion recognition, such as pitch, which is unnecessary for ASR but plays a key role in SER.

3 Methodology

In this paper, a novel machine learning pipeline is devised to extract instances of emotional misalignment in the movie, as shown in Figure 1. In this section, each step of the pipeline will be illustrated in detail. To preliminarily evaluate our pipeline, we apply it to the analysis of the movie *Spirited Away*, by Hayao Miyazaki. The source audio in both Japanese original and English-dubbed audio is in raw waveform (.wav) format, sampled at 16kHz.

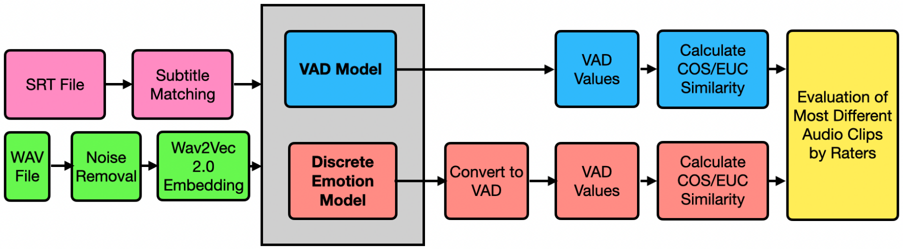

Fig. 1. Overall Pipeline involving AI Emotion Detection and Human Evaluation

3.1 Timestamp matching and de-noising operation

The corresponding subtitles in Japanese and English were obtained from Netflix in SubRip Subtitle (.srt) format. The Japanese subtitle file consists of 1,466 lines, whereas the English subtitle file consists of only 1,119 lines. However, the number of lines in the two subtitle files is different due to the differences in length and grammatical nature of the two languages. This along with accessibility captions gives rise to mismatching timestamps. In order to process texts and audio segments, the subtitle files are pre-processed to find the matching timestamps.

An algorithm to match and merge the subtitles in two languages is designed. We take each pair of timestamps in the Japanese subtitles, and attempt to match a pair of timestamps in the English subtitles where the start and end timestamps are both within a short threshold of the corresponding Japanese start and end timestamps. For unmatched subtitles, the start times were matched using the same threshold, while the end times were matched with an increased threshold in subsequent subtitle entries across both languages. If a match is found, the timestamps and subtitles for all entries between the start and end are merged, resulting in a pair of subtitles with the same duration. Any remaining unmatched entries are discarded. This algorithm maximizes the number of subtitle matches while ensuring that no two matched pairs have overlapping timestamps. As a result, we obtained 1,080 matched pairs of subtitles in both languages.

Given that noisy backgrounds significantly impair model performance [11], a denoising preprocessing step was incorporated into the pipeline to reduce the impact of noise in the subsequent models. The denoising process was applied to the audio segments using the CleanUNet model [6] to remove white noise, and also efficiently remove background music. This denoising operation also reduces the audio to a single channel for subsequent processing.

3.2 Audio Sentiment Analysis

After obtaining the denoised audio, we now predict the sentiment conveyed by the audio by leveraging the XLSR53-Wav2Vec 2.0 model [16] for its robustness to audio perturbations and fair gender accuracy. The extracted speech embeddings are then processed by the VAD model and the Discrete Emotion Model independently, as described below.

VAD Model VAD analysis is performed using the pre-trained speech emotion recognition model as described Wagner et al. [16]. In the implementation, the denoised audio segments were first transformed into the log Mel frequency spectrum using the `audeer` package. The default parameters were used including a sampling rate 16 kHz, 1,024 hidden layer units and default dropout of 0.2. The final linear layer outputs three logit values corresponding to the Valence, Arousal and Domiance scores of the speech segment.

During pre-training, Concordance Correlation Coefficient (CCC) loss was employed alongside the Adam optimiser. During run-time, the logits are passed through a Tanh layer to yield normalised VAD values between 0 and 1. We then linearly interpolated to correspond to the original VAD psychological model.

Discrete Emotion Model Leveraging the representational power of the Wav2Vec 2.0 model, two standalone 6-class emotion recognition models were trained for English and Japanese. The training process is set to run for 10 epochs, using a batch size of 4 with a learning rate of 1e-4 to optimize performance.

The two emotion classification models were trained on noise-free versions of the JVNV dataset [19] for Japanese and the RAVDESS dataset [8] for English. For both datasets, we maintained an equal gender ratio in the speech samples, as well as ensured similar dataset sizes and audio recording lengths. A batch size of 100 was used, with a learning rate starting at $1e^{-4}$. The discrete model achieved an accuracy of 90.5% for the RAVDESS dataset and 98.6% for the JVNV dataset using 5-fold validation. The datasets were balanced for emotion and gender to ensure fairness and reduce potential biases, and across languages, we ensured similar dataset sizes and recording lengths.

Table 1. Emotion Dimension Values for Ekman's six basic emotions [4]

	Valence	Arousal	Dominance
Anger	–0.43	0.67	0.34
Joy	0.76	0.48	0.35
Surprise	0.4	0.67	–0.13
Disgust	–0.6	0.35	0.11
Fear	–0.64	0.6	–0.43
Sadness	–0.63	0.27	–0.33

For each speech segment in both languages, a predicted probability for each of the six emotions is obtained. To compare the emotions quantitatively, we compute a *weighted average VAD score* by applying the VAD coordinates of each of the six basic emotions (see Table 1) weighted by the predicted probability. This method allows for a nuanced analysis of complex emotional states, enabling distance measures between predictions in both languages. For example, a prediction of 50% happiness and 50% surprise can be evaluated continuously between their VAD scores rather than discretely.

3.3 Evaluation Metrics

For the predicted VAD values generated by the VAD and Discrete Emotion models, the audio segments with the highest discrepancies were identified using two measures: cosine similarity and Euclidean distance across each dimension. Cosine similarity provides a similarity score ranging from -1 to 1 and is insensitive to relative magnitude. This is particularly useful for assessing the similarity in the type of emotion conveyed rather than its intensity. However, it is less effective when the same emotion is expressed in both languages but with varying intensities.

Conversely, Euclidean distance is heavily influenced by magnitude, ranging from 0 (indicating identical VAD scores) to infinity. This measure is beneficial for comparing the intensity of emotions; for instance, it can effectively capture the difference in intensity of surprise conveyed between two speech segments. However, Euclidean distance can be affected by the scale of dimensions, as differences in the arousal dimension might disproportionately outweigh differences in the dominance dimension. To evaluate our pipeline, the top 10 audio segments with the highest distances are further analysed.

4 Results

Having identified the sentences with the largest emotional disparities using the machine learning pipeline, we intend to showcase the instances where the emotional interpretations diverged the most. In this study, the audio of the top 10 sentences for each metric (available at https://github.com/adamtwynn/Unravelling-Emotional-Nuances) were manually analysed by four raters: Rater A (native British English speaker, male, aged 20-25), Rater B (native Japanese speaker, female, aged 20-25), and Raters C and D (both proficient in both languages but not native, both female aged 40-50, D is a researcher in Cross-cultural Communications). Many of the sentences were the same across both metrics so any duplicates were removed from the analysis resulting in 15 audio samples for the VAD Model and 11 for the Discrete Model.

4.1 VAD Model Analysis

Each rater was first individually asked to answer the following questions which correspond to valence, arousal and dominance: "Does the English audio sound more positive than the Japanese audio?", "Does the English audio have more energy than the Japanese audio?", and "Does the English audio sound more in control than the Japanese audio?" and could answer on a scale from -2 (significantly less positive/energy/in control) to 2 (significantly more positive/energy/in control) with the option of adding an open comment for each audio clip.

To assess interrater reliability across raters, intraclass correlation coefficients (ICC) were calculated for each VAD dimension. Results indicated poor agreement for valence (ICC = 0.163, 95% CI [0.25, 0.04], p = 0.951) and slight agreement for dominance (ICC = 0.103, 95% CI [0.09, 0.42], p = 0.170), whilst fair agreement was observed for arousal (ICC = 0.287, 95% CI [0.05, 0.60], p = 0.008). Raters were more consistent when evaluating differences in energy levels (arousal) and the lower agreement for valence and dominance likely reflects the greater subjectivity and complexity involved in interpreting emotion across languages.

After each rater rated the top 10 audio clips for each metric, a focus group discussion was held to identify sentences where the raters had disagreements in their evaluations and to further discuss their ratings. Overall, out of the 15 sentences Rater A and B identified three sentences, and Rater D identified four sentences as having the same valence, arousal and dominance in both English and Japanese whereas rater C found no sentences with matching emotional expressions. This suggests that at least 11 out of the 15 sentences identified by the VAD Model contain differences in sentiment and are meaningful for cross-linguistic sentiment analysis.

Individual rating result analysis Rater A, the native English speaker, found that the valence and dominance of English and Japanese were similar on average (Valence: M=0.07, SD=0.8; Dominance: M=-0.13, SD=0.83). The mean value for arousal (M =- 0.53, SD = 1.06) suggest a slight tendency to perceive the Japanese audio as more arousing and having more energy for the selected sentences compared to the English audio. Rater A also commented that the English audio often conveys more exaggerated or heightened emotions and hears sarcasm when the Japanese version sounds more angry. The Japanese audio is commented as sounding more in control and more accepting of situations particularly when the English version sounds worried. Rater B, the native Japanese speaker showed that Japanese valence (M=-0.4, SD=1.12) and arousal (M=-0.4, SD=1.24) were stronger than English indicating that the Japanese audio is slightly more negative with less energy. Similar to Rater A, the dominance mean (M=0.00, SD=0.84) suggests the rater on average thought dominance was similar in both language versions.

Raters C and D are proficient in both English and Japanese. Rater C rated the English audio, with a higher mean valence score (M=0.86, SD=0.86), while the arousal mean (M=-0.21, SD=1.31) indicates a slightly calmer interpretation of the Japanese audio. The dominance mean (M=0.00, SD=1.15) suggests that this rater thought dominance was similar in both language versions. Rater C identified the Japanese audio as more worried and intense, often conveying feelings of anger or frustration, whereas the English audio was perceived as more relaxed and more straightforward in delivery, and lacked energy compared to the Japanese version. In contrast, Rater D's scores for valence (M=-0.13, SD=1.18), arousal (M=-0.40, SD=1.18), and dominance (M=-0.33, SD=1.29) indicate a slight tendency to perceive the Japanese audio as more negative, arousing, and dominant. According to Rater D's comments, the Japanese version was often described as more direct and commanding while the English version appeared softer and more friendly, sometimes altering the meaning of the sentence, as seen in examples like 「イヤっ行かないでここにいてお願い」 (Please don't go)" in Japanese versus "No, don't leave me, I don't want to be alone" in English.

In summary, across the four raters, the Japanese audio is often described as more direct and intense, with stronger expressions of anger and worry, and the English audio, is typically seen as more calm and less emotional. The overall mean values across all raters: valence (0.13), arousal (-0.39), and dominance (-0.10), suggest that while English was generally perceived as slightly more positive in valence, Japanese tended to be more arousing and controlling. However, both Rater B and Rater C reported a mean dominance score of 0 which could suggest that the concept of dominance was hard to hear in the audio especially as differences in valence and arousal were found and emotional differences were mentioned in the open comments.

Focus Group Discussion for VAD Model In the focus group, sentences with rater disagreement were discussed. The first pair of sentence where raters disagreed (A: 0,1,0; B: 0,2,1; C: 0,-1,0; D: 1, 0 -1) was "Welcome. Always nice to see you." and 「いらっしゃいませお早いお着きで」. Rater A commented that the English sounded exaggerated, possibly because British English speakers don't use the phrase "Welcome" as frequently as in Japanese. Rater B, the native Japanese speaker felt that the Japanese version lacked the depth of feeling conveyed and the emotion wasn't as strong. Rater C (proficient in both languages) noted that the English version sounded less passionate, and rater D (also proficient in both) commented that the English version felt friendlier, while the Japanese sounded more in control. Moreover, the Japanese audio translated directly to "Welcome, please quickly get dressed," which may have affected the perception of dominance.

The sentence pair "Shoot, this is clearly harassment" (ちぇっ！見えすいたイビリしやがって) also had varying emotional interpretations among the raters (A: 0,-1,1; B: -1,-1,0; C: 0,1,0; D: 1, -1, 1). Rater A explained that the English audio sounded more sarcastic than angry, a nuance that the other raters did not pick up on. Rater B commented that English sounded more like complaining and the Japanese version was less strong and professional. Rater C noted that the Japanese version felt more emotional, slightly angry and "kuyashii", which translates to a feeling of frustration or regret over something unfair, and the English audio was more straightforwardly angry. Finally, Rater D commented that the Japanese version expressed stronger personal feelings, and similar to Rater B, the English version felt more professional and in control.

Another pair of sentence where raters disagreed (A: 1,-2,-1; B: -2,-1,0; C: 1,-2,Not Clear; D: -1, -1, 2) was "He's destroying everything. It's costing us a fortune. " (何をグズグズしてたんだいこのままじゃ大損だ). Rater A commented that Japanese sounded angrier and Rater C noted that the English version conveyed more disgust, whereas the Japanese version felt angrier, hence more negative. On the other hand, Rater B noted that the Japanese version sounded more acted and the anger didn't sound genuine. Rater D observed that the Japanese version criticized someone's behaviour personally, with a sense of threat for future consequences, while the English version emphasised that the damage had already been done, focusing on the fact that everything was already ruined.

Overall, the focus group discussion revealed how factors such as cultural nuances, as well as the raters' language background could cause differences in how the raters interpreted the emotion of English and Japanese audio clips. The discussion also highlighted the complexity of emotions, with raters often struggling to classify a single emotion in a single utterance and it was also noted that raters' opinions could change after listening again to the same audio.

4.2 Discrete Emotion Model Analysis

To evaluate the discrete model, at first each rater was individually asked to assign one of six possible emotions (angry, disgust, fear, happy, sad and surprised) to both the English and Japanese audio and also provided open-ended comments; then a focus group discussion was conducted. Interrater reliability was assessed using Fleiss' Kappa ($\kappa = 0.34$), indicating fair agreement among raters. Overall, Rater B found that all 11 sentences conveyed the same emotion in both languages, whilst Rater A identified 2 sentences with different emotions, Rater C identified 3, and Rater D identified 8. It is worth mentioning that rater D (a researcher in Cross-cultural Communications) highlighted that 6 out of 8 sentences are related to fear and the fear emotion in Japanese was weakened or absent in the English version.

Individual rating result analysis Rater A frequently identified fear as the dominant emotion. In the English audio, fear was selected five times, along with disgust and surprise twice, and sadness appeared once. For the Japanese audio, fear was also the most common emotion, identified five times, with angry, disgust, surprise, and sad each selected once. In rater A's open-ended responses they often felt that fear was more intense in the Japanese version, while anger, although present in both versions, was portrayed with a more restrained tone in the English audio.

In Rater B's evaluation of the audio, the same emotion was portrayed for both English and Japanese. Fear was identified in 6 out of 10 clips, anger appeared twice, and happy and disgust were each identified once for both the English and Japanese audio. Despite the ratings being the same, Rater C frequently described emotions in the Japanese audio as more powerful or expressive.

Rater C also identified fear as the dominant emotion in the English audio (5 times),and sadness and surprise appeared twice. For the Japanese audio, fear was again the dominant emotion, appearing six times, with sadness twice, and angry and surprise chosen once. In their open-ended responses, Rater C often commented on the increased emotional intensity in the Japanese audio. They remarked that the Japanese characters expressed stronger feelings, noting the louder, more direct expressions of emotions such as anger and sadness. For example, the Japanese version of a scene conveyed a greater sense of frustration, whereas the English counterpart seemed subtler or less emotionally charged.

Rater D's evaluation presented more variation in the emotions selected. For the English audio, they identified fear four times, anger three times, and surprise twice, disgust and sadness once each. In the Japanese audio, fear appeared three times, angry three times, sadness twice, and surprise twice. Rater D discussed that fear and anger, were more subdued in the English audio and was more urging than angry.

In summary, across the four raters, fear was the most consistently identified emotion in both the English and Japanese versions but its intensity differed across languages, with the Japanese audio frequently perceived as more emotionally charged. Anger and sadness were also commonly noted, particularly in the Japanese version, where several raters highlighted how emotions like anger and sadness were conveyed with greater emotion intensity.

Focus Group Discussion for Discrete Model As with the VAD model, sentences were identified where the raters disagreed on their evaluations. One sentence was 「血！わかる？血！」 ("That's blood! Get it! It's blood"). In English this sentence was localised as "I got germs, see". As the sentences were not literally translated, this may have influenced the raters' scores. Rater A (native English) heard disgust in the English audio and fear in the Japanese. Rater B (native Japanese) perceived anger in both versions, explaining that it sounded like the speaker was angrily showing something. Raters C and D (proficient in both languages) felt surprise in the English audio and fear in the Japanese. Only Rater A detected disgust in English, but everyone agreed it was not heard in the Japanese version whilst C and B noted that the Japanese speaker's voice conveyed surprise, which was less apparent in English.

The next pair of sentences was "It's water" (水だ). Rater A heard surprise in the English audio and fear in Japanese, commenting that the English sounded more shocked, whereas Rater B perceived fear in both versions. Rater C detected a mix of surprise and happiness in both, noting that emotions are often more complex than a single label which highlights a limitation of our AI model, which can only classify one emotion per utterance. Rater D heard surprise in the English version and happiness in the Japanese, suggesting they interpreted the discovery of water positively, possibly influenced by the content of the sentence. All raters agreed that the English audio sounded more surprised overall.

Another pair of sentences identified was "Sen, I am sorry I called you a dope before!" (せーん！お前のことどんくさいって言ったけど). Rater A heard disgust in both versions, noting that the Japanese audio sounded more like shouting, edging toward anger but still expressing disgust. They commented that the speaker didn't seem to want to apologize, especially in the English version. Rater B perceived happiness in both. It was discussed that in general, the raters selected the "happy" emotion when they perceived the audio as generally positive, even when "happy" may not have been the most accurate emotion as it was the closest emotion out of happiness, sadness, fear, anger, surprise and disgust. Rater C detected sadness, interpreting both speakers as feeling sorry, though the Japanese version was louder. Rater D heard sadness in the English audio but anger in the Japanese, attributing this to the use of the 'kedo' particle, which made the Japanese version sound more confrontational.

Moreover, in the focus group it was discussed that the native English rater (Rater A) deliberately avoided focusing on the content of the dialogue and relied only on audio cues, unlike the other raters who, being bilingual, considered both the content and the emotional tone of the audio. Whilst the raters tried to not listen to the content, there was still a slight bias as it was difficult to completely disregard the content. This suggests a natural bias when humans label audio compared to the AI models which only have audio as input, as humans can't fully separate cultural knowledge from their evaluations. Whilst AI analyses audio without this bias, the group recognised that the raters' cultural knowledge was valuable, enhancing their interpretations and providing insights that AI alone might miss.

5 Conclusion, Limitations and Future Work

To facilitate automatic knowledge pool creation for the learners of intercultural communication, we developed a novel pipeline combining AI analytics and human-in-the-loop methods for the automated extraction of emotional discrepancies in film dubbing and conducted a case study on the English and Japanese versions of Spirited Away. To answer the first research question, our analysis suggest that whilst both the VAD and discrete models have their strengths, the discrete classification approach proved more intuitive for both human raters and AI when distinguishing between singular emotions. The VAD model was more effective for capturing the emotional depth of more complex audio segments, offering a more nuanced understanding of subtle differences in valence and arousal. However, the human annotators found the VAD approach hard to interpret when comparing it to their own emotional perceptions.

The focus group discussions highlighted the impact of cultural and linguistic backgrounds on emotional perception. To answer the second research question, the raters from different backgrounds often had different interpretations of the same audio clips, revealing how cultural context shapes emotional understanding. The AI models identified the utterances with the highest emotional discrepancies, but human raters offered insights into the emotional complexity that AI struggled to capture, such as detecting multiple emotions in a single utterance. This finding emphasises the importance of integrating human expertise into the analysis process to account for subjective emotional interpretations. Understanding these nuances can support the development of more effective teaching strategies for cross-cultural learners, particularly within ITS, by providing examples of emotional variation in speech and enhancing learners' ability to navigate cross-cultural communication challenges.

In summary, the analysis results indicate that our pipeline effectively identifies audio segments with emotional divergences, which are crucial for understanding how cultural backgrounds influence emotional nuances. In addition, the comparison of discrepancies under various metrics has provided insights into the subtleties of cross-linguistic emotional expression. However, the pipeline lacks the ability to capture contextual factors from audio utterances alone critical for accurate emotional interpretation. Furthermore, the English version of Spirited Away is dubbed, meaning the voice actors are not the same as in the original Japanese version which can create differences in how well emotions are expressed compared to the original actors [10], potentially affecting both AI and human emotional assessments. Additionally, the small sample size of movies and focus groups limits the generalisability of the findings. Future work will involve expanding the dataset to include a broader range of movies with multimedia data. We will also conduct evaluations with larger and more diverse focus groups to further explore how cultural differences influence emotional perception.

Acknowledgments. We would like to thank Peihang Li for their support with coding the pipeline during their undergraduate project at Durham University. We would also like to thank all of the raters for their participation in the focus group discussions.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems **33**, 12449–12460 (2020)
2. Bellini, C., Del Maschio, N., Gentile, M., Del Mauro, G., Franceschini, R., Abutalebi, J.: Original language versus dubbed movies: Effects on our brain and emotions. Brain and Language **253**, 105424 (2024). https://doi.org/https://doi.org/10.1016/j.bandl.2024.105424, https://www.sciencedirect.com/science/article/pii/S0093934X24000476
3. Chiper, S.: Teaching intercultural communication: Ict resources and best practices. Procedia - Social and Behavioral Sciences **93**, 1641–1645 (2013). https://doi.org/https://doi.org/10.1016/j.sbspro.2013.10.094, 3rd World Conference on Learning, Teaching and Educational Leadership
4. Ekman, P., et al.: Basic emotions. Handbook of cognition and emotion **98**(45-60), 16 (1999)
5. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 3451–3460 (2021)
6. Kong, Z., Ping, W., Dantrey, A., Catanzaro, B.: Speech denoising in the waveform domain with self-attention. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7867–7871. IEEE (2022)
7. Lim, N.: Cultural differences in emotion: differences in emotional arousal level between the east and the west. Integrative Medicine Research **5**(2), 105–109 (2016). https://doi.org/https://doi.org/10.1016/j.imr.2016.03.004
8. Livingstone, S.R., Russo, F.A.: The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. PloS one **13**(5), e0196391 (2018)

9. Monaghan, L.: Perspectives on intercultural discourse and communication (Jan 2012). https://doi.org/10.1002/9781118247273.ch2
10. Naranjo, B.: The role of emotions in the perception of natural vs. play-acted dubbing: An approach to angry and sad vocal performances. Meta **66**(3), 580–600 (2021). https://doi.org/https://doi.org/10.7202/1088351ar
11. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning. pp. 28492–28518. PMLR (2023)
12. Russell, J.A., Mehrabian, A.: Evidence for a three-factor theory of emotions. Journal of research in Personality **11**(3), 273–294 (1977)
13. Sainburg, T., Gentner, T.Q.: Toward a computational neuroethology of vocal communication: from bioacoustics to neurophysiology, emerging tools and future directions. Frontiers in Behavioral Neuroscience **15**, 811737 (2021)
14. Spanias, A., Painter, T., Atti, V.: Audio signal processing and coding. John Wiley & Sons (2006)
15. Tiwari, V.: Mfcc and its applications in speaker recognition. International journal on emerging technologies **1**(1), 19–22 (2010)
16. Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., Schuller, B.W.: Dawn of the transformer era in speech emotion recognition: closing the valence gap (2022). https://doi.org/10.48550/ARXIV.2203.07378
17. Wierzbicka, A.: Emotions across languages and cultures: Diversity and universals (11 1999). https://doi.org/10.1017/CBO9780511521256
18. Xia, Y., Shin, S.Y., Kim, J.C.: Cross-cultural intelligent language learning system (cils): Leveraging ai to facilitate language learning strategies in cross-cultural communication. Applied Sciences **14**(13), 5651 (Jun 2024). https://doi.org/10.3390/app14135651
19. Xin, D., Jiang, J., Takamichi, S., Saito, Y., Aizawa, A., Saruwatari, H.: Jvnv: A corpus of japanese emotional speech with verbal content and nonverbal expressions. IEEE Access (2024)