

Development and Usability Evaluation of an AI-Based Vocal Emotion Assessment Prototype for Vocational Communication Training

Bayu Sulistiyanto Ipung Sutejo^{1✉}, Muhammad Revin Putra Fahlefi², Habibullah Akbar³,
Andriyanti Asianto⁴, Diah Aryani⁵, Ega Shela Marsiani⁶

(1,2,3,4,5) Universitas Esa Unggul, Jakarta, Indonesia

(6) Universitas Indraprasta PGRI, Jakarta, Indonesia

✉ Corresponding author
bayu.sulistiyanto@esaunggul.ac.id

Received: 08-06-2026 | Revised: 12-08-2026 | Accepted: 05-09-2026

ABSTRACT

Vocational communication training requires opportunities for learners to practise vocal delivery and receive timely feedback, yet conventional assessment often depends on direct instructor observation. This study developed and evaluated a web-based AI-assisted vocal emotion assessment prototype for vocational communication practice. Using evolutionary prototyping, the system integrated browser-based audio recording and file upload, signal preprocessing, *emotion2vec+* inference, probability visualization, and automated feedback. Functional performance was examined through eight black-box testing scenarios, while perceived usability was evaluated using the System Usability Scale with 15 participants comprising 12 vocational students and three instructors. All eight functional scenarios operated successfully, with client-server response times of approximately 0.45–0.65 seconds. The prototype achieved a SUS score of 76.98, indicating favorable perceived usability, particularly for ease of use and learnability. These findings support the prototype's preliminary feasibility as a formative practice tool, although classification validity and educational effectiveness require further evaluation using larger samples, expert-validated emotion labels, and authentic vocational communication settings.

Pelatihan komunikasi vokasional memerlukan kesempatan bagi peserta didik untuk mempraktikkan penyampaian vokal dan memperoleh umpan balik secara tepat waktu, sedangkan asesmen konvensional masih banyak bergantung pada pengamatan langsung instruktur. Penelitian ini mengembangkan dan mengevaluasi prototipe asesmen emosi vokal berbasis web dengan dukungan kecerdasan buatan untuk latihan komunikasi vokasional. Pengembangan menggunakan evolutionary prototyping dengan mengintegrasikan perekaman dan pengunggahan audio, prapemrosesan sinyal, inferensi emotion2vec+, visualisasi probabilitas, serta umpan balik otomatis. Kinerja fungsional diuji melalui delapan skenario black-box testing, sedangkan persepsi usability dievaluasi menggunakan System Usability Scale terhadap 15 partisipan yang terdiri atas 12 peserta didik vokasi dan tiga instruktur. Seluruh skenario pengujian fungsional berhasil dijalankan dengan waktu respons client-server sekitar 0,45–0,65 detik. Prototipe memperoleh skor SUS sebesar 76,98 yang menunjukkan persepsi usability yang positif, terutama pada kemudahan penggunaan dan pembelajaran sistem. Temuan ini mendukung kelayakan awal prototipe sebagai alat latihan formatif, tetapi validitas klasifikasi dan efektivitas pendidikan masih memerlukan pengujian lebih lanjut dengan sampel yang lebih luas, label emosi tervalidasi pakar, dan konteks komunikasi vokasional autentik.

Keywords: artificial intelligence, prototype development, usability evaluation, vocational communication training

Introduction

Vocational education is fundamentally concerned with preparing learners to perform competently within authentic occupational environments (Genito, 2026; Kebede et al., 2024; Qi et al., 2026; Sidorova et al., 2024). This preparation extends beyond mastery of technical procedures because many occupations require individuals to communicate effectively, regulate their responses, and interact professionally with customers, colleagues, and other stakeholders (Haigler, 2021; Laguna et al., 2025; Munir, 2022; Rivas & Husein, 2022). These competencies are particularly important in service-oriented fields such as hospitality, tourism, banking, and customer support, where spoken interaction frequently becomes the primary medium through which professional competence is demonstrated. In such settings, successful communication depends not only on the verbal content of a message but also on how that message is delivered. Vocal characteristics such as pitch, intensity, speaking rate, pauses, and variations in prosody can shape how listeners interpret confidence, calmness, empathy, enthusiasm, or emotional tension. From the perspective of affective computing, these observable characteristics provide meaningful information that computational systems can process to recognize aspects of human emotional expression (Picard, 1997; Schuller & Schuller, 2021).

The importance of vocal expression becomes more apparent when vocational learners encounter communication situations that demand emotional control. Customer complaints, service failures, negotiations, and other high-pressure interactions require learners to maintain an appropriate tone even when responding to difficult situations. Acoustic properties of speech can provide measurable indications of emotional states, and previous research has demonstrated the possibility of using speech characteristics to identify emotional expression and evaluate customer-related interactions in real-world acoustic environments (El Ayadi et al., 2011; Parra-Gallego & Orozco-Arroyave, 2022). For vocational education, this creates an opportunity to treat vocal communication not merely as an interpersonal attribute assessed through general impressions but as a competency that can be repeatedly practiced and supported through observable feedback. Such an approach is especially relevant when learners need opportunities to recognize undesirable vocal tendencies and refine their delivery before participating in authentic workplace interactions.

Despite this need, the assessment of vocal communication in vocational learning environments continues to depend substantially on instructor observation (Abraham & Singaram, 2019). Although instructor judgment remains important for understanding complex interpersonal communication, exclusive reliance on manual assessment presents several practical limitations. Evaluation may vary across instructors, particularly when judgments involve subtle aspects of vocal expression that are difficult to quantify consistently (Waskito et al., n.d.). Large student-to-instructor ratios can further restrict the frequency with which individualized assessment and feedback are provided. Consequently, learners may have limited opportunities to monitor changes in their vocal performance across repeated practice sessions. These constraints are particularly relevant to deliberate practice, in which learners benefit from repeated performance, timely information about their current performance, and opportunities to make targeted adjustments. An automated formative assessment mechanism could therefore complement instructor evaluation by allowing learners to conduct repeated vocal exercises and receive immediate feedback while preserving the instructor's role in interpreting more complex dimensions of professional communication.

Recent developments in Speech Emotion Recognition (SER) provide a technological basis for such an assessment mechanism. SER systems use computational approaches to identify

emotional information contained in speech signals, enabling vocal patterns to be processed and classified systematically. Earlier developments employed supervised architectures trained on established emotional speech datasets such as RAVDESS and CREMA-D, while subsequent research introduced CNN-LSTM architectures, attention-based bidirectional models, lightweight convolutional networks, and representation-separation techniques to improve classification performance and robustness under different conditions (Aftab et al., 2022; Cao et al., 2014; Leem et al., 2021; Livingstone & Russo, 2018; Vakariuk, 2022; Zhao et al., 2019). More recently, self-supervised representation learning has expanded the possibilities of SER by reducing dependence on manually engineered acoustic features. The emotion2vec framework represents this development by learning generalized representations of emotional speech through self-supervised pre-training, providing a potentially suitable foundation for applications requiring direct and efficient vocal emotion inference (Ma et al., 2024a).

Nevertheless, advances in emotion-recognition accuracy do not automatically translate into educationally usable systems. A model that performs well computationally must still be integrated into an interaction environment that learners can operate without excessive technical demands. For vocational communication practice, this requirement involves several interconnected processes: capturing or uploading speech through an accessible interface, transforming audio into a format compatible with the inference model, producing emotion classifications within a practical response time, presenting the output in an understandable form, and translating computational predictions into feedback that learners can use during subsequent practice (Budiman, 2025). Usability therefore becomes an important consideration alongside technical functionality. If an assessment tool introduces unnecessary operational complexity, requires extensive technical assistance, or presents results that learners cannot readily interpret, its potential contribution to independent practice becomes limited regardless of the sophistication of the underlying AI model. This relationship places system functionality, user interaction, and formative feedback at the intersection between technological development and vocational learning practice.

Existing studies have contributed substantially to the development of speech emotion datasets, classification architectures, robustness strategies, and increasingly sophisticated representation-learning models. However, the literature incorporated in the present study provides comparatively limited evidence concerning how a self-supervised vocal emotion model can be operationalized as an interactive assessment tool specifically designed for vocational communication practice. This distinction constitutes the central research gap addressed in the present study. Rather than proposing another emotion-classification architecture, the study focuses on translating an existing foundation model into a functional educational prototype that connects browser-based vocal input, standardized audio preprocessing, AI inference, visualized emotion outputs, and automated formative feedback within a single workflow. The novelty therefore lies primarily in the educational operationalization of AI-based vocal emotion recognition and its integration into a user-oriented practice environment, followed by systematic examination of both functional performance and perceived usability. This positioning is deliberately limited to functionality and usability; it does not presume that the prototype has already been demonstrated to improve communication learning outcomes.

Accordingly, this study aims to develop and evaluate an interactive web-based AI vocal emotion assessment prototype intended to support independent communication practice among vocational learners. Two objectives guide the investigation. First, the study examines

the functional performance of the prototype through black-box testing covering browser-based audio capture, signal preprocessing, AI inference, output visualization, and automated feedback generation. Second, it evaluates user acceptance and usability through the System Usability Scale (SUS) involving vocational communication learners and instructors. By addressing these objectives, the study contributes a functionally evaluated and user-oriented implementation of speech emotion recognition for vocational communication practice while providing preliminary empirical evidence regarding its usability. The findings are intended to establish a foundation for subsequent research examining pedagogical effectiveness, broader classroom implementation, and adaptation to diverse Indonesian vocational communication contexts.

Methods

This study employed an evolutionary prototyping approach to develop and evaluate an interactive web-based vocal emotion assessment prototype for vocational communication practice. Evolutionary prototyping enables a functional system to be progressively developed and refined through iterative cycles, making it appropriate when system requirements, technical functionality, and user interaction need to be examined during development rather than determined entirely at the outset (Camburn et al., 2017). The development process comprised five interconnected stages: identifying pedagogical and functional requirements, designing the initial system architecture, constructing the prototype, verifying its functional operation, and evaluating its usability. The prototype was implemented using Python Flask on the server side and HTML5 and JavaScript on the client side, integrating browser-based audio acquisition, signal preprocessing, AI inference, and automated feedback within a single workflow. The study focused specifically on functional performance and perceived usability and therefore did not evaluate changes in learners' communication competence or learning outcomes. The overall architecture and sequence of the prototype are presented in Figure 1.

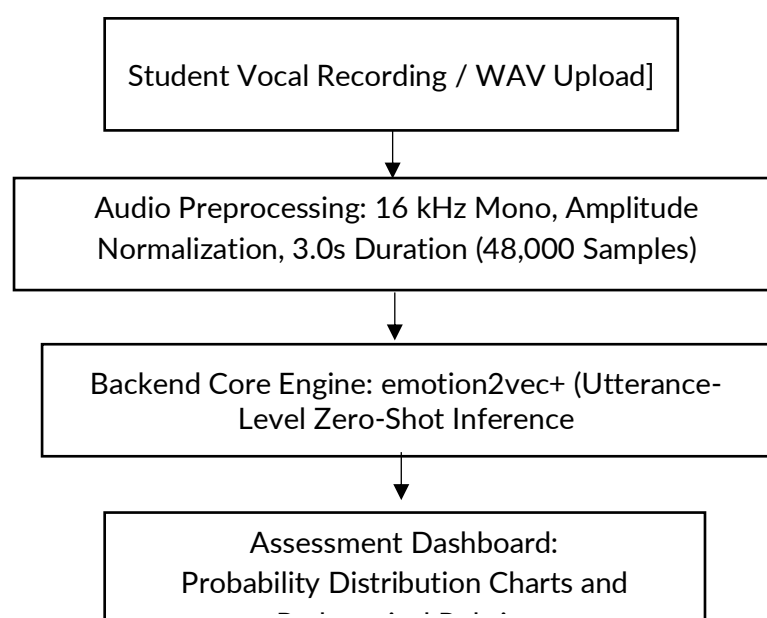


Figure 1. End-to-End Architectural Pipeline of the Vocal Assessment Prototype

The prototype adopted a client-server architecture in which speech could be recorded directly through the browser or submitted as an uncompressed WAV file. Incoming audio was converted into a mono signal, resampled to 16 kHz, amplitude-normalized, and standardized to

a duration of 3.0 seconds or 48,000 samples using Librosa and SoundFile. Recordings exceeding the specified duration were trimmed, whereas shorter recordings were completed through zero-padding. Emotion inference was performed using the pretrained *iic/emotion2vec_plus_base* model through the FunASR and ModelScope environments, and the resulting predictions were mapped into four target categories: anger, sadness, happiness, and neutrality. Functional verification was subsequently conducted through black-box testing by examining eight operational workflows: browser microphone capture, WAV-file upload, audio-duration standardization, AI inference, emotion-label mapping, graphical output visualization, automated feedback generation, and invalid audio handling. For each workflow, the observed system response was compared with its predefined expected behavior, while processing latency was recorded where applicable. This procedure was intended to determine whether the integrated prototype operated according to its functional specifications rather than to independently re-evaluate the classification accuracy of the underlying emotion2vec+ model.

Usability evaluation involved 15 participants comprising 12 vocational students enrolled in customer service and business communication courses and three vocational instructors (Creswell & Creswell, 2018; Suhirman et al., 2026). Participants were recruited through purposive sampling according to their relevance to the intended context of prototype use, with inclusion directed toward vocational learners engaged in customer service or business communication learning and instructors involved in related vocational communication instruction. Following interaction with the prototype, participants completed the ten-item System Usability Scale (SUS) developed by Brooke (1996) using a five-point response scale. For positively worded odd-numbered items, one point was subtracted from the response, whereas negatively worded even-numbered items were scored by subtracting the response from five; the sum of the adjusted scores was then multiplied by 2.5 to generate a composite score ranging from 0 to 100. Descriptive analysis was used to summarize functional status, processing latency, item responses, and overall perceived usability. To minimize unnecessary retention of potentially sensitive voice data, audio was temporarily processed during inference and removed after the assessment output was returned to the user, consistent with the data-handling procedure described in the prototype.

Results

User Interface and Interaction Flow

The development process resulted in a fully operational web-based prototype that integrated audio acquisition, emotion analysis, visualization, and automated feedback within a single user interface. The interface was organized into two principal functional areas to separate the input process from the presentation of assessment results. This arrangement enabled users to complete the assessment sequence within the same browser environment without navigating between multiple pages or external applications. The input area provided two alternative mechanisms for supplying vocal data: direct speech recording through the browser microphone and the selection of a prerecorded WAV audio file. These alternatives accommodated different interaction conditions while directing both forms of input into the same subsequent analytical workflow.

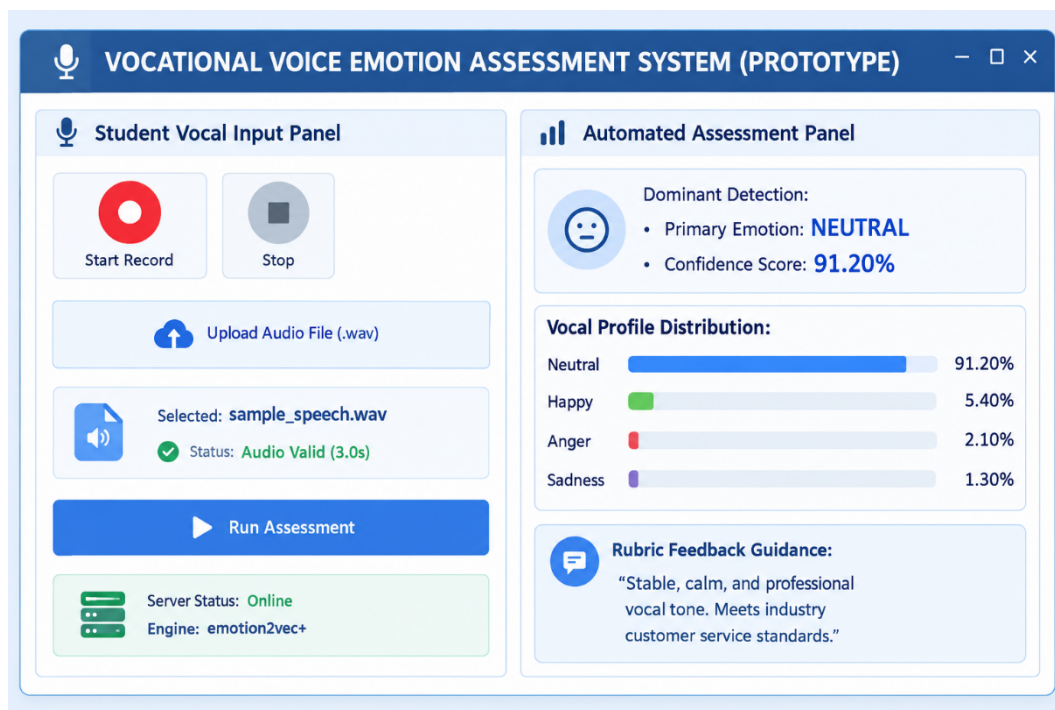


Figure 2. Graphical User Interface of the Interactive Vocal Assessment Prototype

As presented in Figure 2, the left section of the interface functioned as the primary point of interaction for audio acquisition and file management. Once vocal input had been provided, the interface displayed information associated with the recording and prepared the audio for submission to the analytical process. This arrangement established a sequential interaction pattern in which users first supplied vocal data and subsequently initiated the assessment process. The system therefore maintained a visible separation between the stage of preparing vocal input and the stage of receiving analytical output. The interface also provided feedback when the submitted audio satisfied the required processing conditions, allowing the user to determine whether the recording had progressed to the assessment stage before interpreting the resulting emotion classification.

The right section of the interface presented the output generated following AI inference. Rather than displaying only a single categorical prediction, the dashboard organized the results into several complementary forms of information. The dominant emotional state was displayed as the primary classification outcome, while the relative probability distribution across the four target categories—anger, sadness, happiness, and neutrality—was represented through horizontal percentage bars. An automated recommendation was subsequently presented alongside the classification output. Consequently, the complete interaction sequence progressed from vocal input to standardized processing, emotion prediction, visual presentation of probability values, and automated textual feedback within the same interface. The implemented interface therefore demonstrated that the principal components specified during prototype development could be presented as a continuous browser-based interaction rather than as separate technical operations.

Functional Verification (Black Box Testing)

Functional verification was conducted across eight predefined operational scenarios representing the principal functions required for the prototype to complete the vocal assessment workflow. The scenarios covered browser-based microphone recording, WAV-file submission, signal-duration normalization, AI inference, emotion-label mapping, graphical probability rendering, automated rubric generation, and invalid audio handling. Table 1 presents

the expected output, observed system response, measured latency, and final verification status for each test case. Across all scenarios, the observed responses corresponded with the predefined functional requirements, and every test case was classified as **Valid**. Thus, the functional verification produced an operational success rate of 100% across the eight tested workflows.

Table 1. Functional Verification Results of the Prototype Using Black Box Testing

Test Case ID	Operational Feature / Workflow	Expected Functional Output	Actual System Response	Measured Latency	Test Status
TC-01	Browser microphone streaming	Browser prompts permission and captures live voice stream	Audio captured and buffered in local browser memory	< 200 ms	Valid
TC-02	File upload mechanism (.wav)	Accepts standard WAV audio files via drag-and-drop or file dialog	File transferred successfully to temporary server cache	< 300 ms	Valid
TC-03	Signal duration normalization	Trims audio > 3.0s and applies zero-padding to audio < 3.0s	Array reshaped to exactly 48,000 discrete points	45 ms	Valid
TC-04	AI inference execution	emotion2vec+ processes raw waveform without runtime failure	Generates raw probability vector across emotion classes	45.94–48.78 ms	Valid
TC-05	Output label mapping	Maps model classes into anger, sadness, happy, and neutral	Output cleanly categorized into target four-class domain	< 5 ms	Valid
TC-06	Graphic distribution rendering	Dynamic horizontal progress bars render proportional values	Percentages rendered accurately on the client dashboard	< 100 ms	Valid
TC-07	Rubric logic generation	Displays corresponding constructive text based on primary state	Correct contextual guidance text rendered on interface	< 10 ms	Valid
TC-08	Empty audio input handling	Rejects silent/corrupted files and prompts user to re-record	Displays warning modal without application crash	< 50 ms	Valid

The first group of test cases examined the acquisition and preparation of vocal data before AI inference. In TC-01, the browser successfully requested microphone permission and captured the live voice stream into local browser memory, with a recorded latency below 200 ms. TC-02 confirmed that WAV audio files submitted through the available upload mechanism were successfully transferred into the temporary server cache, with latency remaining below

300 ms. The preprocessing function evaluated through TC-03 also operated according to specification. Audio exceeding the predetermined 3.0-second duration was trimmed, whereas shorter input was processed through zero-padding until the resulting waveform contained exactly 48,000 discrete samples. This normalization procedure required approximately 45 ms. Collectively, these results verified that audio obtained through either of the available input mechanisms could proceed through the required preprocessing sequence without producing a functional failure.

The subsequent test cases examined the analytical and output-generation stages of the workflow. TC-04 demonstrated that the emotion2vec+ inference engine processed the standardized raw waveform and generated a probability vector across the available emotion classes without runtime failure. The measured inference latency ranged from 45.94 to 48.78 ms. TC-05 verified the subsequent mapping procedure, through which the model outputs were categorized into the four target labels of anger, sadness, happiness, and neutrality; this operation required less than 5 ms. TC-06 further confirmed that the resulting probability values could be rendered accurately as dynamic horizontal progress bars on the client dashboard, with a measured latency below 100 ms. These results showed that the sequence from standardized waveform input to emotion classification and graphical output was completed successfully across the tested functional conditions.

The final two test scenarios examined feedback generation and error handling. TC-07 confirmed that the rubric module generated contextual guidance corresponding to the primary detected emotional state, with the text displayed correctly on the interface and processing completed in less than 10 ms. TC-08 examined the system response to empty, silent, or corrupted audio input. Rather than continuing the analytical process or producing an application failure, the prototype rejected the unsuitable input and displayed a warning requesting a new recording. This error-handling procedure was completed in less than 50 ms and did not result in an application crash. The inclusion of this test demonstrated that the functional verification covered not only successful input conditions but also an invalid-input condition that could reasonably occur during user interaction.

When the eight functional scenarios were considered collectively, no critical operational failure was recorded during black-box verification. All test cases achieved Valid status, including functions associated with input acquisition, preprocessing, inference, visualization, feedback generation, and invalid-input handling. At the level of the complete interaction cycle, the client-server round-trip time from assessment initiation to the presentation of the final rubric recommendation ranged from approximately 0.45 to 0.65 seconds. These findings establish the functional performance observed during testing; they should not be interpreted as evidence of the classification accuracy of emotion2vec+ because predictive accuracy was not independently tested using participant recordings in the present study.

Usability Evaluation (System Usability Scale)

Usability evaluation was conducted with 15 participants, consisting of 12 vocational students and three vocational instructors. After interacting with the prototype, participants completed the ten statements included in the System Usability Scale. Table 2 presents the mean Likert response for each item together with its converted score and the resulting composite SUS calculation. The item-level results produced a total average converted score of 30.79. Following the standard SUS conversion procedure, multiplication of this value by 2.5 generated a final composite usability score of 76.98.

Table 2. system usability scale (sus) questionnaire responses and score conversion

Item ID	Standard SUS Questionnaire Statements	Mean Likert Score (1-5)	Conversion Formula	Average Converted Score
Q1	I think that I would like to use this vocal assessment tool frequently	4.20	$X - 1$	3.20
Q2	I found the system unnecessarily complex	1.87	$5 - X$	3.13
Q3	I thought the system was easy to use	4.27	$X - 1$	3.27
Q4	I think that I would need the support of a technical person to be able to use this system	1.73	$5 - X$	3.27
Q5	I found the various functions in this system were well integrated	4.13	$X - 1$	3.13
Q6	I thought there was too much inconsistency in this system	1.80	$5 - X$	3.20
Q7	I would imagine that most students/instructors would learn to use this system very quickly	4.33	$X - 1$	3.33
Q8	I found the system very cumbersome/awkward to use	1.67	$5 - X$	3.33
Q9	I felt very confident using the system	3.93	$X - 1$	2.93
Q10	I needed to learn a lot of things before I could get going with this system	2.00	$5 - X$	3.00
Total	Sum of Average Converted Item Scores (Q1 through Q10)	-	-	30.79
Final	Composite SUS Usability Score (Total Converted Score $\times$ 2.5)	-	-	76.98

A more detailed examination of the positively worded SUS items revealed consistently favorable responses across several dimensions of interaction. The mean response for willingness to use the vocal assessment tool frequently (Q1) reached 4.20, while the perceived ease of use (Q3) recorded a mean of 4.27. The integration of system functions (Q5) received a mean score of 4.13. The highest mean among the positively phrased statements was recorded for Q7, concerning how quickly students or instructors could learn to operate the system, with a value of 4.33. Participant confidence when using the prototype (Q9) produced a somewhat lower but still positive mean of 3.93. Taken together, the positive items therefore ranged from 3.93 to 4.33, with the strongest responses concentrated on perceived learnability and ease of operation.

The negatively worded items displayed the inverse response pattern expected in a favorable usability assessment. Perceived unnecessary complexity (Q2) recorded a mean of 1.87, while the perceived need for technical assistance (Q4) produced a mean of 1.73. The

statement concerning excessive inconsistency within the system (Q6) recorded a mean of 1.80, and the perception that the system was cumbersome or awkward to operate (Q8) obtained the lowest mean among the negative items at 1.67. Q10, which assessed whether substantial prior learning was required before participants could begin using the system, recorded a mean of 2.00. After application of the standard SUS conversion procedure, these negatively phrased responses contributed positively to the overall usability score because lower raw responses represented lower perceived complexity, inconsistency, dependence on technical assistance, and operational burden.

At the item-conversion level, the average adjusted scores ranged from 2.93 to 3.33. The highest converted values were obtained for Q7 and Q8, both at 3.33, followed by Q3 and Q4 at 3.27. Q1 and Q6 each produced converted scores of 3.20, whereas Q2 and Q5 each contributed 3.13. Q10 yielded a converted value of 3.00, and Q9 produced the lowest converted contribution at 2.93. The sum of these converted item values was 30.79, from which the final composite SUS score of 76.98 was derived. The item-level distribution is important because it indicates that the final composite score was not generated by a single exceptionally high item; rather, the responses displayed a broadly consistent pattern across the ten dimensions represented in the instrument.

Using the interpretive categories already adopted in the original manuscript, the composite SUS score of 76.98 was placed within the Acceptable acceptability range, associated with Grade B, and described using the adjective rating Good. Among the individual responses, perceived rapid learnability (Q7 = 4.33) and ease of use (Q3 = 4.27) constituted the two highest positive ratings, whereas low scores for cumbersome operation (Q8 = 1.67) and dependence on technical assistance (Q4 = 1.73) provided complementary results from the negatively phrased items. Accordingly, the usability findings consistently reflected favorable participant responses across ease of operation, learnability, functional integration, complexity, and technical-support requirements. These results describe users' perceptions of the prototype after use and do not, by themselves, establish improvement in vocal communication competence or other educational outcomes.

Discussion

The findings of this study demonstrate that the web-based vocal assessment prototype successfully integrated audio recording or file submission, signal preprocessing, AI-based emotion inference, probability visualization, and automated feedback within a continuous interaction sequence. In the context of vocational education, this integration is particularly relevant because communication competence requires more than conceptual understanding; learners also need opportunities to perform, obtain information about their performance, and reflect on how they communicate in practice-oriented situations. The prototype supported this process through an interface that clearly separated the input and output areas while maintaining both within the same operational environment. Users could provide a speech sample, allow the system to process the input, and subsequently examine the predicted emotional category, probability distribution, and corresponding recommendation without moving between different applications. This sequence established a direct connection between vocal performance and the information generated from that performance, making the analytical process more visible to the user. Rather than positioning AI inference as an isolated computational procedure, the prototype embedded it within an interaction that progressed from action to analysis and from analysis to feedback. The principal contribution observed at

this stage therefore lies in translating vocal emotion processing into an integrated educational interaction that can provide a basis for subsequent reflection on vocal performance.

The successful integration of this workflow is closely associated with the use of *emotion2vec+* as the principal speech-analysis component. Ma et al. (2024) demonstrated that the *emotion2vec* approach employs self-supervised pre-training to construct speech representations containing emotion-related information and can perform competitively across multiple emotion-recognition tasks. This capability has potential relevance to vocational fields characterized by intensive interpersonal communication, including customer service, hospitality, tourism, banking, and other service occupations, because professional performance in these settings depends not only on the semantic content of a message but also on how it is vocally delivered. Nevertheless, the performance of a pretrained model reported in previous studies cannot automatically be transferred to every population or context of use. Previous speech emotion recognition research has shown that classification performance can be affected by differences in datasets, language, speaker characteristics, acoustic conditions, recording devices, and environmental noise (George & Odgers, 2015; Tripathi et al., 2021; Zhang et al., 2021). Consequently, the anger, sadness, happiness, and neutrality outputs produced by the present prototype should be interpreted as model-generated analytical information that can support reflection on vocal delivery rather than as definitive determinations of a learner's emotional state. This distinction is particularly important because the present study did not independently compare *emotion2vec+* predictions from participant recordings with expert-established ground-truth labels. The contribution of the study therefore lies primarily in operationalizing an existing speech emotion model within a vocational educational workflow rather than establishing new evidence of its classification accuracy for Indonesian vocational learners.

Functional verification demonstrated that all eight predefined black-box testing scenarios achieved Valid status, including browser-based microphone recording, WAV-file submission, signal-duration normalization, AI inference, emotion-label mapping, probability visualization, automated recommendation generation, and invalid-input handling. This finding indicates that the sequence required to transform vocal input into analytical output operated according to the specified functional requirements under the conditions examined. Comparable applications of black-box testing in educational software have emphasized the importance of verifying core system functions before usability or broader user evaluation is undertaken (Saputra & Parhusip, 2026). In the present study, the absence of critical functional failure across the eight scenarios provides an important technical foundation for repeated learner-system interaction. The client-server round-trip time of approximately 0.45–0.65 seconds further indicates that analytical output could be returned shortly after an assessment was initiated, thereby maintaining continuity between vocal performance and the resulting feedback. Such responsiveness is relevant to practice-oriented learning because prolonged delays may interrupt the connection between an action, its evaluation, and the learner's subsequent attempt. Nevertheless, the 100% result must remain within the boundaries of the black-box procedure: it represents successful completion of the eight tested functions, not 100% emotion-recognition accuracy, evidence of improved communication competence, or assurance that identical performance would be obtained across all devices, network conditions, languages, accents, and recording environments. Functional reliability, predictive validity, and educational effectiveness therefore remain distinct dimensions of evaluation.

The usability findings provide complementary evidence concerning the preliminary feasibility of the prototype from the perspective of its intended users. The composite SUS score of 76.98 indicates a generally favorable perception of usability and, according to the interpretive framework applied in this study, places the prototype within the Acceptable range with a Good adjective rating. Lewis (2018) describes SUS as a widely used instrument for assessing perceived usability, while Hyzy et al. (2022) note the widespread use of approximately 68 as a general benchmark for interpreting SUS scores, although meaningful variation can occur across systems and domains. The score obtained in the present study exceeded this commonly referenced benchmark. More importantly, the item-level pattern was consistent with the composite result. Participants gave relatively high ratings to rapid learnability (Q7 = 4.33) and ease of use (Q3 = 4.27), while reporting relatively low perceptions that the system was cumbersome (Q8 = 1.67) or required technical assistance (Q4 = 1.73). These findings have practical relevance because a technology intended for repeated vocational practice should not require learners to devote excessive cognitive or instructional resources merely to understand its operation. At the same time, SUS measures perceived usability rather than learning effectiveness, and the current evaluation involved only 15 participants comprising 12 vocational students and three instructors. The result should therefore be regarded as preliminary evidence of favorable usability within the pilot group rather than as a basis for generalizing acceptance to the wider vocational education population.

One plausible explanation for the favorable usability pattern is the relatively simple interaction sequence and the integration of the principal functions within a single environment. Users did not need to move between separate applications to record speech, initiate analysis, inspect results, and receive recommendations, while the visual separation between input and output areas provided a recognizable sequence of actions. From a human-centred artificial intelligence perspective, such characteristics are important because the educational value of AI is not determined exclusively by algorithmic sophistication. Alfredo et al. (2024) emphasize the importance of comprehensibility, stakeholder involvement, human control, reliability, safety, and trustworthiness when learning analytics and AI are designed around human needs. The use of probability bars in the present prototype is relevant in this respect because it provides users with information beyond a single dominant emotion label, allowing them to observe how model outputs are distributed across the available categories. This design is also consistent with developments in learning analytics in which visualization is frequently employed to communicate analytical information to learners and educators (Banihashem et al., 2022) while educational dashboards have increasingly moved from simple data presentation toward supporting learning-oriented interpretation and action (Paulsen & Lindsay, 2024). Nevertheless, Matcha et al. (2020) and de Vreugd et al. (2024) caution, in different learning-analytics contexts, that displaying information does not automatically guarantee that users will interpret it appropriately or modify their learning behavior. Probability visualization should therefore be treated as a reflective aid that makes model output more inspectable, rather than as an objective and self-sufficient measurement of a learner's emotional condition.

The automated textual recommendations extend the prototype beyond the presentation of analytical output toward a mechanism with potential relevance to formative feedback and deliberate practice. Deliberate practice involves more than repeated performance; improvement depends on obtaining sufficiently specific information about current performance and using that information to guide subsequent attempts. The prototype establishes the technical conditions for such a cycle because learners can produce a vocal sample, receive an

analysis within a short interval, inspect the classification and recommendation, and subsequently modify their vocal delivery during another attempt. Banihashem et al. (2022) similarly indicate that learning analytics can contribute to feedback processes when analytical information is communicated in forms that are meaningful for learners and educators. Evidence from adjacent speech technologies further illustrates the importance of feedback design. Ngo-Thi-Kim et al (2026) , in a meta-analysis of automatic speech recognition in ESL/EFL pronunciation learning, reported an overall positive effect on pronunciation development and found that outcomes were influenced by feedback and instructional characteristics. These findings cannot be transferred directly to speech emotion recognition because pronunciation and vocal emotion represent different constructs and involve different analytical processes. They nevertheless illuminate a relevant pedagogical mechanism: automated speech analysis becomes educationally meaningful when its output is returned to learners in a form that can inform reflection and subsequent performance. The automated recommendations in the present prototype can therefore be conceptualized as formative prompts that may support reflection on vocal delivery. However, because the study did not measure communication competence before and after system use, this function represents a pedagogical potential of the prototype rather than demonstrated evidence of learning effectiveness.

At a theoretical level, the findings position AI-supported vocal assessment at the intersection of affective computing, speech emotion recognition, learning analytics, formative assessment, and human-centred educational technology. Research on speech emotion recognition demonstrates that deep-learning and self-supervised approaches can computationally represent affective information embedded in acoustic signals (Abbaschian et al., 2021; Ma et al., 2024) , whereas research on human-centred AI emphasizes that analytical capability alone does not determine educational value (Alfredo et al., 2024) . Algorithmic outputs acquire educational significance when they are translated into forms that users can observe, interpret, question, and, where appropriate, use to inform subsequent action. The present study provides an implementation-level illustration of this relationship because emotion inference was not deployed as an isolated computational component but embedded within a sequence of vocal input, processing, visualization, and feedback that received generally favorable usability responses. More importantly, the findings highlight the need to distinguish four levels of evidence: functional validity, perceived usability, predictive validity, and educational effectiveness. Black-box testing provides evidence that specified functions operate as intended; SUS captures users' perceptions of interaction with the system; predictive validity requires comparison between model classifications and defensible ground-truth labels; and educational effectiveness requires evidence of meaningful changes in communication performance or learning outcomes following system use. Maintaining these distinctions prevents technical performance from being interpreted prematurely as pedagogical effectiveness and provides a clearer framework for subsequent evaluation of AI-based educational systems.

Taken together, the successful functional verification, relatively short response time, integrated interaction sequence, and favorable preliminary usability results provide a basis for further development of the prototype as a formative support tool for vocational communication training. Potential applications include independent practice and instructor-supported simulations involving customer service encounters, complaint handling, presentations, front-desk interactions, and other occupational situations in which vocal delivery contributes to professional performance. In such contexts, AI-generated classifications and recommendations

should function as additional sources of information for reflection rather than as autonomous judgments of learner competence or emotional condition. This orientation is consistent with human-centred AI principles that preserve meaningful human involvement in interpretation and educational decision-making (Alfredo et al., 2024). It is also necessary because voice recordings may contain sensitive information and because pretrained emotion models may respond differently across linguistic, cultural, and acoustic conditions. Future research should therefore involve larger and more diverse samples of students and instructors, multi-institutional testing, authentic Indonesian vocational speech, expert-established ground truth, comparisons across recording devices and environmental conditions, and longitudinal or quasi-experimental designs capable of measuring changes in communication performance. Such investigations would allow the evidence base to progress systematically from demonstrating that the prototype functions and is perceived as usable toward determining whether its classifications are valid, whether its feedback produces measurable learning benefits, and under what pedagogical, linguistic, and sociocultural conditions AI-supported vocal emotion assessment can be implemented responsibly in vocational education.

Conclusion

This study developed and evaluated a web-based AI-assisted vocal emotion assessment prototype designed to support communication practice in vocational education. The prototype successfully combined browser-based audio recording or file submission, signal preprocessing, emotion2vec+ inference, probability visualization, and automated feedback within a unified interaction flow. Functional verification confirmed that all eight predefined black-box scenarios operated according to the specified requirements, while the relatively short client-server response time supported a continuous transition from vocal input to analytical feedback. The usability evaluation produced a SUS score of 76.98, accompanied by favorable responses particularly in relation to ease of use and learnability. These findings establish preliminary evidence that AI-based vocal emotion analysis can be implemented in a technically functional and generally usable form for vocational communication practice. In practical application, the prototype can be incorporated into independent exercises or instructor-facilitated simulations involving customer service, complaint handling, presentations, front-desk communication, and other occupational interactions in which vocal delivery is relevant. Its outputs should, however, be used as formative information that encourages learners to examine and adjust their vocal performance rather than as an autonomous judgment of communication competence or emotional state. Instructors therefore remain essential for interpreting AI-generated results within the professional, situational, and sociocultural context of each communication task.

The findings also define the next stage required for the development and evaluation of this approach. Although the prototype demonstrated functional reliability and favorable perceived usability, the present evidence does not establish the predictive validity of its emotion classifications or demonstrate that its use improves vocational communication competence. Further research should therefore validate model outputs against expert-established ground truth using authentic Indonesian vocational speech and examine performance across different speakers, accents, recording devices, acoustic environments, and occupational scenarios. Studies involving larger and more diverse samples across multiple institutions are also required to determine whether the usability findings remain consistent beyond the present pilot setting. More importantly, longitudinal, experimental, or quasi-experimental designs should investigate whether repeated use of AI-generated vocal feedback

contributes to measurable changes in emotional regulation, vocal delivery, and professional communication performance. Future development should simultaneously strengthen data protection, transparency, and human oversight so that voice-based AI remains a supportive component of formative assessment rather than a substitute for professional judgment. These directions would enable subsequent research to progress from demonstrating that the prototype functions and is usable toward establishing when its classifications are valid, whether its feedback generates meaningful educational benefits, and how it can be responsibly integrated into vocational communication training.

References

- Abbaschian, B. J., Sierra-Sosa, D., & Elmaghraby, A. (2021). Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models. *Sensors*, 21(4), 1249. <https://doi.org/10.3390/s21041249>
- Abraham, R. M., & Singaram, V. S. (2019). Using deliberate practice framework to assess the quality of feedback in undergraduate clinical skills training. *BMC Medical Education*, 19(1), 105. <https://doi.org/10.1186/s12909-019-1547-5>
- Aftab, A., Morsali, A., Ghaemmaghami, S., & Champagne, B. (2022). LIGHT-SERNET: A lightweight fully convolutional neural network for speech emotion recognition. *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6912–6916. <https://doi.org/10.1109/ICASSP43922.2022.9746405>
- Alfredo, R., Echeverria, V., Jin, Y., Yan, L., Swiecki, Z., Gašević, D., & Martinez-Maldonado, R. (2024). Human-centred learning analytics and AI in education: A systematic literature review. *Computers and Education: Artificial Intelligence*, 6, 100215. <https://doi.org/10.1016/j.caeai.2024.100215>
- Banihashem, S. K., Noroozi, O., van Ginkel, S., Macfadyen, L. P., & Biemans, H. J. A. (2022). A systematic review of the role of learning analytics in enhancing feedback practices in higher education. *Educational Research Review*, 37, 100489. <https://doi.org/10.1016/j.edurev.2022.100489>
- Brooke, J. (1996). SUS: A “Quick and Dirty” Usability Scale. In P. W. Jordan, B. Thomas, I. L. McClelland, & B. Weerdmeester (Eds.), *Usability Evaluation in Industry* (pp. 189–194). Taylor & Francis. <https://doi.org/10.1201/9781498710411-35>
- Budiman, R. D. A. (2025). Effectiveness of AI-Driven Assessments in Enhancing Learning Evaluation through Predictive Technology in Vocational Secondary School. *International Journal of Information and Education Technology*, 15(7), 1410–1417. <https://doi.org/10.18178/ijiet.2025.15.7.2342>
- Camburn, B., Viswanathan, V., Linsey, J., Anderson, D., Jensen, D., Crawford, R., Otto, K., & Wood, K. (2017). Design prototyping methods: state of the art in strategies, techniques, and guidelines. *Design Science*, 3, e13. <https://doi.org/DOI: 10.1017/dsj.2017.10>
- Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., & Verma, R. (2014). CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4), 377–390. <https://doi.org/10.1109/TAFFC.2014.2315692>
- Creswell, J. W., & Creswell, J. D. (2018). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Sage Publications.
- De Vreugd, L., Van Leeuwen, A., Jansen, R., & Van der Schaaf, M. (2024). Learning Analytics Dashboard Design and Evaluation to Support Student Self-Regulation of Study Behaviour. *Journal of Learning Analytics*, 1–14. <https://doi.org/10.18608/jla.2024.8529>
- El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3), 572–587. <https://doi.org/10.1016/j.patcog.2010.09.020>
- Genito, P. (2026). The impact of AI on reading and literacy practices: implications for vocational training in the Adriatic-Ionian Region. *Economia Marche-Journal of Applied Economics*, 45(1).

- George, M. J., & Odgers, C. L. (2015). Seven fears and the science of how mobile technologies may be influencing adolescents in the digital age. *Perspectives on Psychological Science*, 10(6), 832–851.
- Haigler, K. (2021). The role of occupational research in complex assessments: Aligning educational practices with workplace requirements. *The Journal of Writing Analytics*, 5(1), 259–291.
- Hyzy, M., Bond, R., Mulvenna, M., Bai, L., Dix, A., Leigh, S., & Hunt, S. (2022). System Usability Scale Benchmarking for Digital Health Apps: Meta-analysis. *JMIR MHealth and UHealth*, 10(8), e37290. <https://doi.org/10.2196/37290>
- Kebede, A., Asgedom, A., & Asfaw, A. (2024). Is linking technical and vocational training with industry a bridge to employability? Lessons from a literature review for Ethiopia. *Cogent Education*, 11(1), 2406721.
- Laguna, R. F. D., Aguinaga, D. L., Torres, D. E. C., & Cueto, B. A. L. (2025). Soft skills and the use of industry 4.0 as determinants of professional development in engineering graduates: A SEM approach. *Sustainable Futures*, 10, 100742.
- Leem, S.-G., Fulford, D., Onnela, J.-P., Gard, D., & Busso, C. (2021). Separation of emotional and reconstruction embeddings on ladder network to improve speech emotion recognition robustness in noisy conditions. *Proceedings of Interspeech 2021*, 2871–2875. <https://doi.org/10.21437/Interspeech.2021-1200>
- Lewis, J. R. (2018). The System Usability Scale: Past, Present, and Future. *International Journal of Human-Computer Interaction*, 34(7), 577–590. <https://doi.org/10.1080/10447318.2018.1455307>
- Livingstone, S. R., & Russo, F. A. (2018). 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. <https://doi.org/10.1371/journal.pone.0196391>
- Ma, Z., Zheng, Z., Ye, J., Li, J., Gao, Z., Zhang, S., & Chen, X. (2024a). emotion2vec: Self-supervised pre-training for speech emotion representation. *Findings of the Association for Computational Linguistics: ACL 2024*, 15747–15760. <https://doi.org/10.18653/v1/2024.findings-acl.933>
- Ma, Z., Zheng, Z., Ye, J., Li, J., Gao, Z., Zhang, S., & Chen, X. (2024b). emotion2vec: Self-supervised pre-training for speech emotion representation. *Findings of the Association for Computational Linguistics: ACL 2024*, 15747–15760. <https://doi.org/10.18653/v1/2024.findings-acl.933>
- Matcha, W., Uzir, N. A., Gašević, D., & Pardo, A. (2020). A Systematic Review of Empirical Studies on Learning Analytics Dashboards: A Self-Regulated Learning Perspective. *IEEE Transactions on Learning Technologies*, 13(2), 226–245.
- Munir, F. (2022). More than technical experts: Engineering professionals' perspectives on the role of soft skills in their practice. *Industry and Higher Education*, 36(3), 294–305.
- Ngo-Thi-Kim, H., Pham, Q. V., Phan, T. H. T., Dang, U. P., & Minh, B. N. H. (2026). Artificial intelligence and teacher education: A bibliometric analysis between 1991 and 2023. *Journal of Educational and Social Research*, 16(2), 631–642. <https://doi.org/10.36941/jesr-2026-0314>
- Parra-Gallego, L. F., & Orozco-Arroyave, J. R. (2022). Classification of emotions and evaluation of customer satisfaction from speech in real world acoustic environments. *Digital Signal Processing*, 120, 103286. <https://doi.org/10.1016/j.dsp.2021.103286>
- Paulsen, L., & Lindsay, E. (2024). Learning analytics dashboards are increasingly becoming about learning and not just analytics - A systematic review. *Education and Information Technologies*, 29(11), 14279–14308. <https://doi.org/10.1007/s10639-023-12401-4>
- Picard, R. W. (1997). *Affective computing*. The MIT Press. <https://doi.org/10.7551/mitpress/1140.001.0001>
- Qi, Z., Wang, J., Du, Y., & Zhang, X. (2026). The effect of blended learning approaches on vocational school students' practical skills performance and learning behaviors: virtual

- simulations or hands-on activities while accessing instructional videos. *Instructional Science*, 54(1), 39.
- Rivas, D. F., & Husein, S. (2022). Empathy, persuasiveness and knowledge promote innovative engineering and entrepreneurial skills. *Education for Chemical Engineers*, 40, 45–55.
- Saputra, S. D., & Parhusip, J. (2026). PENGUKURAN USABILITY APLIKASI WEB MENGGUNAKAN SUS (SYSTEM USABILITY SCALE) DAN PENGUJIAN BLACK BOX PADA WEBSITE E-LEARNING (GEN-IT) KABUPATEN KATINGAN. *Jurnal Informatika Dan Teknik Elektro Terapan*, 14(1). <https://doi.org/10.23960/jitet.v14i1.8390>
- Schuller, D., & Schuller, B. W. (2021). A review on five recent and near-future developments in computational processing of emotion in the human voice. *Emotion Review*, 13(1), 44–50. <https://doi.org/10.1177/1754073919898526>
- Sidorova, M., Kopus, T., & Yurasova, I. (2024). Digital transformation for teaching management accounting: training with a simulation in an authentic professional environment. *Accounting Education*, 33(5), 621–648.
- Suhirman, L., Nugroho, A. Y., Dhyanaputri, I. G. A. S., Adawiyah, R., Soraya, A., Dzaky, M. I., Alunaza, H., Akbar, S. A., Hafizi, M. Z., & Junida, D. S. (2026). *Dasar Metodologi Penelitian*. CV. Askara Sastra Media.
- Tripathi, J. P., Bugalia, S., Burdak, K., & Abbas, S. (2021). Dynamical analysis and effects of law enforcement in a social interaction model. *Physica A: Statistical Mechanics and Its Applications*, 567, 125725. <https://doi.org/10.1016/j.physa.2020.125725>
- Vakariuk, L. V. (2022). To the issue of the efficiency of legal regulation of labor relations and optimization of the labor-law regime. *Naukovij Visnik Užgorods'kogo Nacional'nogo Universitetu*, 2(72), 9–14. <https://doi.org/10.24144/2307-3322.2022.72.34>
- Waskito, Irzal, Wulansari, R. E., & Kyaw, Z. Y. (n.d.). The Adventure of Formative Assessment with Active Feedback in The Vocational Learning: The Empirical Effect for Increasing Students' Achievement. *Journal of Technical Education and Training*, 14(1), 54–62. Retrieved <https://journal.uthm.edu.my/index.php/JTET/article/view/9793>
- Zhang, S., Liu, R., Tao, X., & Zhao, X. (2021). Deep Cross-Corpus Speech Emotion Recognition: Recent Advances and Perspectives. *Frontiers in Neurorobotics*, 15, 784514. <https://doi.org/10.3389/fnbot.2021.784514>
- Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control*, 47, 312–323. <https://doi.org/10.1016/j.bspc.2018.08.035>