IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE CONSUMER TECHNOLOGY SOCIETY SECTION

Received 26 June 2022, accepted 11 July 2022, date of publication 5 August 2022, date of current version 11 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3196733

APPLIED RESEARCH

Confirming Customer Satisfaction With Tones of Speech

YEN HUEI KO[1], PING-YU HSU[1], YU-CHIN LIU[2], AND PO-CHIAO YANG[1]
[1]Department of Business Administration, National Central University, Taoyuan 320317, Taiwan
[2]Department of Information Management, Shih Hsin University, Taipei 116002, Taiwan

Corresponding author: Yen Huei Ko (yensintong@gmail.com)

ABSTRACT As customer satisfaction explicitly leads to repurchase behavior, the level of customer satisfaction affects both sales performance and enterprise growth. Traditionally, measuring satisfaction requires customers spending extra time to fill out a post-purchase questionnaire survey. Recently, ASR (Automatic Speech Recognition) is utilized to extract spoken words from conversation to measure customer satisfaction. However, as oriental people tend to use vague words to express emotion, the approach has its limitation. To solve the problem, this study strived to complete following tasks: devising a process to collect customer voice expressing satisfaction and corresponding verifiable ground truth; a dataset of 150 customer voices speaking in Mandarin was collected; MFCCs were extracted from the voice data as features; as the size of dataset was limited, Auto Encoder was utilized to further reduce the features of voices; models based on Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNNs) and Support Vector Machine (SVM) were constructed to predict satisfaction. With nested cross validation, the average accuracy of LSTM and SVM could reach 71.95% and 73.97%, respectively.

INDEX TERMS Customer satisfaction, LSTM-RNNs, MFCCs, SVM.

I. INTRODUCTION

Customer satisfaction is crucial for enterprise growth. Satisfaction strongly affects customer repurchase behavior [1], [2]. Oliver [3] proposed the Expectancy Confirmation Theory (ECT), which states that consumers evaluate the performance of products against prior expectations and that the discrepancy between expectations and reality determines the degree of customer satisfaction. Garbarino and Johnson [4] stated that customer decision-making behavior is governed by two mental constructs, namely, perceived service quality and customer satisfaction with the service or product being purchased. Satisfied customers are typically those who consider that their expectations are fulfilled. Therefore, evaluating customer satisfaction with each transaction is essential for enterprises to not only improve their service quality but also gain a clearer understanding of customer expectations [5].

Traditionally, customer satisfaction is measured through post-purchase questionnaire surveys or interviews [4], [6], [7]. In the telemarketing industry, surveys based on pre-prepared questionnaires are performed through telephone interactions between operators and customers [8]. However, several limitations constrain the effective application of telephone surveys. First, most customers are reluctant to participate in a telephone survey because of the time required for participation. Second, the cost of conducting telephone surveys is high because of the additional tasks imposed on call center agents conducting the surveys. Moreover, customers who choose to participate in such surveys tend to have extreme opinions. Customers with mild satisfaction or dissatisfaction usually skip these types of surveys [1]. Therefore, developing a tool for directly evaluating customer satisfaction without the need for post-purchase questionnaire surveys is invaluable.

Several studies have adopted text mining and natural language recognition techniques for measuring customer satisfaction. For example, Kang and Park [9] measured the levels of customer satisfaction with mobile services by using text mining to compile sentiment words from customer reviews and then conducting sentiment analysis to calculate

The associate editor coordinating the review of this manuscript and approving it for publication was Yoonsik Choe.

sentiment scores. Park [10] employed text mining and statistical analysis to calculate sentiment scores for online customer reviews in order to measure customer satisfaction. Automatic Speech Recognition (ASR) is a technique that is primarily used for transcribing speech into spoken words, and the texts corresponding to these spoken words can then be analyzed through various analytical models to measure customer satisfaction [11]–[14].

However, text alone might be unable to reflect the satisfaction level of customers. In particular, for reasons of politeness or avoiding awkwardness, East Asian customers might utter vague words that somewhat suggest satisfaction. For example, Tanaka *et al.* [15] conducted an experiment that involved an emotional Stroop-like task in which participants were required to identify the emotions of facial expressions and ignore the corresponding voice (or vice versa). The experimental results indicated that Japanese people are more conciliatory than Dutch people when expressing their feelings. In general, Japanese people hesitate to express their true thoughts and tend to utter ambiguous answers when pressed. Liu *et al.* [16] confirmed that compared with the comprehension of English, that of Mandarin is more complicated and relies more on tone of voice. Furthermore, a Mandarin speaker is likely to mainly focus on the tone of someone's voice instead of their facial or verbal expression alone to understand their true feeling [17]. Banse and Scherer [18] posited that spoken language is embedded with a variety of nonverbal means to express emotions. As indicated by Bostanov and Kotchoubey [19], tone of voice is the most essential prosodic cue for emotional recognition. Thus, without the recognition of tone of voice, spoken words alone cannot precisely represent customer responses.

To catch the level of customer satisfaction without relying on post-consumption questionnaire surveys and speech recognition, this study strived to collect customer voices with different levels of satisfaction and build a model to distinguish their satisfaction. The aim of the present study was five folds: design an experiment to collect voices corresponding to different levels of satisfaction; verify the ground truth of voice data, of which the satisfaction was measured with questionnaire surveys adapted from [20] and [21]; implement Mel-Frequency Cepstral Coefficients (MFCCs) to extract voice features which were transformed into quantified vectors; propose an Auto-Encoder (AE) to reduce attributes as the number of manually collected voice data was limited; devise two classifiers based on Long Short-Term Memory (LSTM) Recurrent Neural Networks (-RNNs) and Support Vector Machine (SVM) as they have shown good performance at analyzing sequential ratio data. The last but not least, this research also compare their classification performances under different settings, namely, with and without Auto-Encoder, and with different setting of neuro networks and hyper parameters.

The research result contributed to the literature of measuring customer satisfaction with voices in several aspects. A process to collect customer satisfaction voices with verifiable ground truth was designed. A dataset of 150 customer voices speaking in Mandarin was collected. The process along with feasible hyperparameters of MFCCs to extract the features of these voices was shown. As the size of dataset was limited, AE was proved to be beneficial in this study. Different hyperparameters of LSTM and SVM were evaluated and the best result was shown for future references. To balance between data limitation and hyperparameter selections, the accuracy of the proposed model was verified with the approach of nested cross validation. The best average accuracy of LSTM-RNNs and SVM was 71.95% and 73.97%, respectively. SVM outperformed LSTM-RNNs may be due to its simplicity as the data size was limited in this study.

The remainder of this paper is organized as follows: **Section 2** reviews the importance of customer satisfaction and prior research using speech recognition technique to judge if customers were satisfied, and describes the popularity of MFCC process being used to extract features of voices, **Section 3** describes the process to solicit customer voice, the tool to measure the ground truth of the collected data, MFCC process and hyperparameters to extract voice features, the mathematical models and pseudocode of LSTM-RNNs, AE and SVM used in this study, **Section 4** describes the collected voice data, verification of the data validity, process of data clustering into three level of satisfactions, the process to decide AE hyperparameters, the process of the nested cross validation to decide the hyperparameters of LSTM-RNNs and SVM, and compare and contrast the accuracies of models under various settings, and **Section 5** summarizes the contributions of this study, discusses implications of the experimental results, and suggests further research directions.

II. RELATED WORKS

A. SATISFACTION

According to ECT, satisfaction is an emotional attitude or a psychological state that results from the evaluative judgment of the expectation for and performance of a product or service being purchased. Oliver [3] posited that expectation is the baseline or reference level for a customer to evaluate a product or service. Ando *et al.* [5] demonstrated that the degree of customer satisfaction indicates the service quality of enterprises and their performance in customer engagement. Satisfaction is the cognitive assessment of expectation versus actual performance, and satisfaction is achieved when the level of perceived performance is higher than the initial level of expectation.

Several studies have evaluated customer satisfaction to predict repurchase intention [20], [22]–[24]. Moreover, studies have revealed that customer satisfaction is an essential factor for retaining long-term loyal customers and influencing repurchase behavior [1]–[3], [6], [25]. Customer repurchases indicate that customers' expectations are fulfilled. Kincade *et al.* [7] posited that customer demand leads to purchase behavior and that repurchase behavior is driven by customer satisfaction. Furthermore, customer satisfaction

explicitly leads to reduced operating costs and generates both customer referrals and positive word-of-mouth, which increase revenues and profitability [26]–[29]. Therefore, an enterprise must measure the level of customer satisfaction to improve its internal management and maintain valuable relationships with customers [1].

Traditionally, customer satisfaction is measured through post-consumption questionnaire surveys [7], [20], [21]. Text mining has been used to analyze sentiment words from customer reviews for measuring the level of customer satisfaction [9] and [10]. Additionally, speech recognition techniques have been adopted to measure customer satisfaction [11]–[14]. Sun *et al.* [13] used information fusion to analyze customer satisfaction. In telemarketing, communication between call agents and customers is conducted through voice calls; a telephone operator administers a questionnaire survey to customers through phone calls. However, a telephone survey has some limitations, such as customer reluctance and cost constraints. Therefore, a tool must be developed for determining customer satisfaction by using acoustic data without conducting a survey.

In the present study, a discriminative model was developed for determining customer satisfaction. The study conducted an experiment to collect voice samples for analysis in the model. This model was developed to verify whether voice data can be used to distinguish the degree of customer satisfaction.

B. FEATURE EXTRACTION TECHNIQUES

Because people's voice tracts have unique physiological structures, the frequency spectrum of speech signals can be used for various speech applications, including speech and speaker recognition [30]–[35]. For effective speech recognition, spectral feature extraction is used to transform raw speech into compressed signals. Effective feature extraction techniques include those that are based on MFCCs, Linear Prediction Coefficients (LPCs), and Linear Prediction Cepstral Coefficients (LPCCs). MFCCs have the advantages of reducing error rates, improving efficiency, and increasing accuracy [30]–[33], [35]. Singh *et al.* [36] presented that MFCC-based features extraction methods achieve superior performance to methods based on prosodic features alone; this is because MFCCs can be used to accurately discriminate signal characteristics. Tiwari [32] demonstrated that combining MFCCs with Vector Quantization (VQ) is suitable for feature extraction for speaker identification. Singh and Rajan [34] concluded that voice samples used for model training and testing should be collected under the same conditions; they also determined that noise is the main factor affecting the accuracy of MFCC-VQ based method for speaker recognition. Martinez *et al.* [33] observed that feature extraction using MFCCs and VQ achieved higher accuracy than that achieved using LPCCs. Currently, MFCC-based feature extraction is the most widely used technique for speaker recognition [36].

Voice features extracted using MFCCs are short-term features that reflect the frequency spectrum of a speech signal. Prosodic features with rhythmic properties are long-term features that reflect a voice's fundamental frequency. Accordingly, extracting specific types of prosodic features, including pitch, intensity, and duration, can enhance acoustic applications [37]–[41]. The extraction of prosodic features is a relatively simple and effective process in speech and speaker recognition tasks. Sönmez *et al.* [42] developed a piecewise linear model for capturing stylized pitch contours to represent an individual's speaking style in order to achieve speaker verification. Pitch features reflect the rising-falling patterns of signals and enable the achievement of robust and accurate recognition performance [38]. Sinith *et al.* [41] fed different combinations of MFCC-based features and prosodic features to an SVM classifier for speech emotion recognition. Their results indicated that the combination of MFCC-based features, pitch, and intensity engendered the highest accuracy in speech emotion recognition.

Accordingly, the present study collected quantified feature vectors that comprised voice features extracted using MFCCs and prosodic features namely voice pitch and voice intensity extracted using Praat and fed them to the constructed LSTM-RNN model for measuring customer satisfaction.

III. METHODOLOGY

A. RESEARCH DESIGN

The experiment conducted in this study was based on those conducted in [43] and [44]. The experiment was designed to capture participants' confirmation of expected utility and utterances conveying their level of satisfaction simultaneously. Two sets of data were collected in the experiment: the first set comprised customer responses to a questionnaire survey adapted from the relevant literature, and the second set comprised audio clips indicating the same participants' satisfaction level in Mandarin. The questionnaire data were validated through statistical analysis and utilized for supervised learning to classify audio data. Voice data collection was conducted in a soundproof laboratory to avoid noise interference. Through a data preprocessing process, features extracted from the voice data were transformed into quantified vectors that comprised MFCCs, logarithmic energy, voice pitch, and voice intensity. These vectors were then fed to the constructed LSTM-RNN model for predicting customer satisfaction. This model was trained and evaluated using a five-fold nested cross-validation method. Fig. 1 shows the architecture of research framework.

B. DATA COLLECTION PROTOCOL

The design of the conducted experimental was based on ECT [3], which states that satisfaction is the discrepancy between a consumer's expectation and the actual utility of a product. Therefore, satisfaction is determined by two factors: expectation (e.g., pre-consumption) and confirmation (e.g., post-consumption or experience). To collect information related

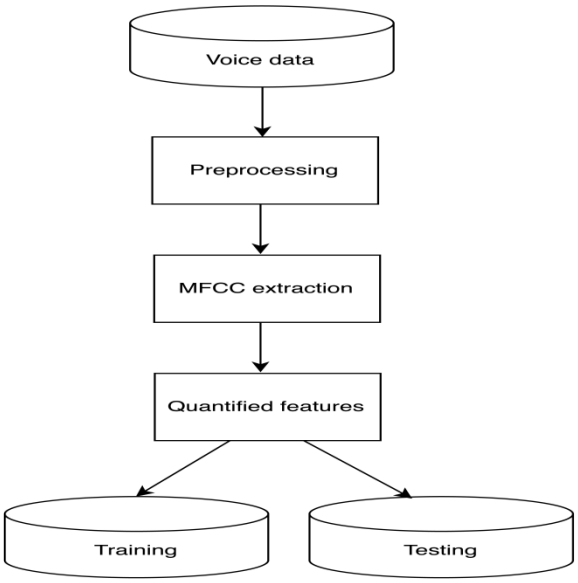


FIGURE 1. Architecture of research framework.

TABLE 1. Satisfaction statement in mandarin.

Please utter the following statement
我對於該產品滿意

to expectation, prior to being served, the participants were requested to indicate their order of preference for the following beverages: coffee, milk tea, green tea, fruit and vegetable juice, and Chinese bitter tea. Subsequently, approximately one-third of the participants were each served their most preferred, moderately preferred, and least preferred beverage; each participant was served one drink. The data collection process involved the following steps:

1) Before drinking a beverage, the participants recorded a voice file by uttering a short statement in Mandarin, which is presented in Table 1. This statement translates to "I am satisfied with this item." This step helped familiarize the participants with the prompted utterance.
2) The participants were then invited into the recording room and served a beverage. The beverage was poured into a dark glass before the participants entered the room to ensure that they had no prior knowledge of what beverage would be served to them.
3) Immediately after drinking the provided beverage in the blinded test, the participants uttered the statement presented in Table 1, which was then recorded. This step was conducted to determine the participants' degree of satisfaction with the provided beverage.

The post-consumption questionnaire used in this study was adapted from those presented in [20] and [21], which were based on the work of Oliver [3]. The level of satisfaction was measured using a 5-point Likert scale ranging from 0 ("Very Dissatisfied") to 4 ("Very Satisfied"). The devised questionnaire is presented in Table 2.

TABLE 2. Scale items related to satisfaction in the post-study questionnaire.

	How do you feel after drinking the beverage?	
S1	I feel	(highly unacceptable, unacceptable, neutral, acceptable, or highly acceptable)
S2	I feel	(very displeased, displeased, neutral, pleased, or very pleased)
S3	I feel	(very dissatisfied, dissatisfied, neutral, satisfied, or very satisfied)
S4	I feel	(absolutely terrible, terrible, neutral, delighted, or absolutely delighted)

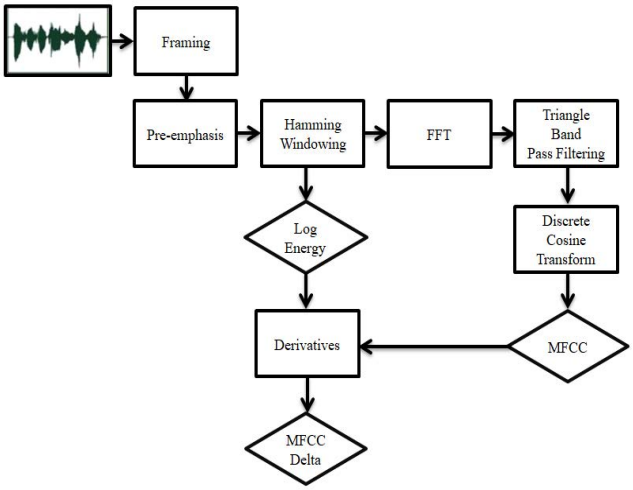


FIGURE 2. The process of voice feature extraction.

Satisfaction Labeling: After verifying the validity of the satisfaction construct, this study used the factor loading and score obtained for each item to calculate the satisfaction level of each participant. The scores were discretized into three bins through the Visual Binning (VB) process in IBM®SPSS®. The three bins were marked as satisfaction, neutrality, and dissatisfaction. The participants and their corresponding audio files were labeled appropriately into the three bins.

C. MFCC PROCESSING

MFCC-based techniques have been widely used to extract crucial voice features [45]. The process of extracting voice features is shown in Fig. 2.

MFCC processing typically involves the following step:

1) *Processing of the input signal:* In the time domain, a signal waveform expresses the relationship between signal intensity and time. The dynamic signal $X(t)$ is a continuous function of time. However, for conducting digital sampling and numerical calculations, $X(t)$ must be converted into a discrete function $S(n)$. The most common approaches for converting continuous analog signals into discrete analog signals are Pulse Amplitude Modulation (PAM) and Pulse Code Modulation (PCM). PAM and PCM are used to sample the amplitude of an analog signal at equal intervals.

2) *Framing:* Because a voice signal changes with time, it must be partitioned into short frames to capture the acoustic characteristics. The size of a frame may be set to 20 to 40 ms (milliseconds) to avoid excessive changes between adjacent frames. According to the procedure described by [46], this study set the frame size to 25 ms. To avoid considerable changes between frames, consecutive frames can be set to overlap with each other. As in most relevant studies [30], [31], [33], [41], the overlap interval in this study was set to 10 ms.
3) *Pre-emphasis:* In the vocalization step, a high-pass filter can be used to compensate for the energy of high frequencies suppressed by a pronunciation system. The pre-emphasis formula can be expressed as follows:

$$S'[n] = S[n] - \beta S[n-1], \quad n = 1, \ldots, N \tag{1}$$

where $S[n]$ is the voice signal of the n^{th} sampling point in a frame, $S'[n]$ is the output of the voice signal obtained after pre-emphasis, and β is the coefficient of the high-pass filter. In this study, β, which is usually a constant, was set to 0.97 in accordance with the suggestion of [46].
4) *Hamming Windowing:* After the framing step, a Hamming window function can be applied to each frame to reduce signal discontinuity for minimizing the distortion in the spectra at the beginning and end of a frame.

$$\tilde{S}[n] = S'[n]\, W[n] \tag{2}$$

$$W[n] = \begin{cases} (1-\alpha) - \alpha\cos\left(\dfrac{2\pi n}{N}\right); & 1 \le n \le N \\ 0; & \text{otherwise} \end{cases} \tag{3}$$

where $\tilde{S}[n]$ is the output signal obtained after Hamming windowing, $W[n]$ is the Hamming window function, and $\alpha \le 0.5$ is the adjustment parameter for reducing the signal intensity at the beginning and end of a frame. The parameter α is usually set to approximately 0.46 [31] and [47].
5) *Fast Fourier Transform:* This step utilizes Discrete Fourier Transform (DFT) to transform $\tilde{S}[n]$ from the time domain to the frequency domain for effectively describing signal characteristics. The DFT of a frame, for frequency $k\left(\hat{S}[k]\right)$ among a total of K discrete frequencies can be expressed as follows:

$$\hat{S}[k] = \sum_{n=1}^{N} \tilde{S}[n]\, e^{-\frac{j2\pi kn}{N}}, \quad 1 \le k \le K \tag{4}$$

where j is an imaginary number.
The periodogram-based power spectrum of each frame ($X[k]$) at frequency k can be expressed as follows:

$$X[k] = \frac{1}{N}\left|\hat{S}[k]\right|^2 \tag{5}$$

The characteristics of a signal can be obtained by observing the energy distribution of a spectrum at different frequencies.

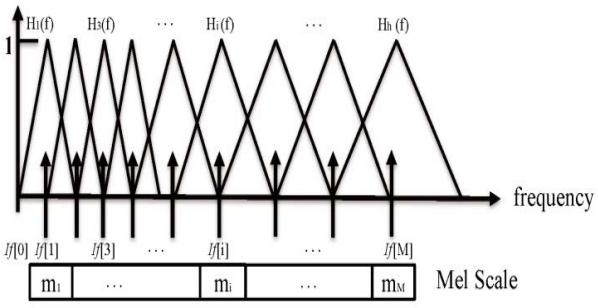


FIGURE 3. Mel filter banks.

6) *Mel Filtering:* Although $X[k]$ can be derived through calculations, not all $X[k]$ values affect human hearing in the same manner. Moreover, frequencies related to human hearing, should be emphasized. Therefore, Mel filter banks can be used to enhance and reduce the weights of $X[k]$ in particular frequency intervals. A Mel filter bank contains a set of band-pass filters defined on a constant Mel scale interval. As displayed in Fig. 3, Mel frequencies overlap with adjacent frequencies. Consequently, if the number of intervals is M, M+2 border points exist in the frequency domain.

- *Definition 1*: Let f be a frequency (in Hz). The Mel frequency can be computed as follows:

$$\text{Mel}(\text{f}) = 1127ln\left(1 + \frac{\text{f}}{700}\right)$$

Given M intervals and $\text{f}_{\text{lb}} \ge 0$ and $\text{f}_{\text{ub}} \le K$ be the lower and upper bounds of the frequencies of interest, respectively. The Mel frequency of the m^{th} border point can be expressed as follows:

$$\text{Mf}_{\text{m}} = \text{Mel}(\text{f}_{\text{lb}}) + \frac{\text{m}-1}{\text{M}+1}\left(\text{Mel}(\text{f}_{\text{ub}}) - \text{Mel}(\text{f}_{\text{lb}})\right),$$

where $\text{m} = 1, \ldots, \text{M}+2$.
- *Definition 2:* Let $If[\text{m}]$ be the inverse Mel frequency of the m^{th} border point.

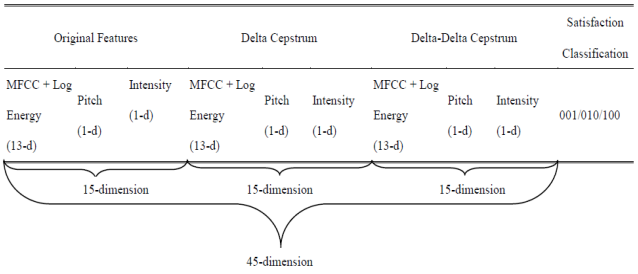
$$If[\text{m}] = 700\left(\text{Exp}\left(\frac{\text{Mf}_{\text{m}}}{1127}\right) - 1\right),$$

where $\text{m} = 1, \ldots, \text{M}+2$.
- *Definition 3:* Given the Mel frequency filter operated in the h^{th} interval;
$\text{H}_{\text{h}}(\text{f})$ is the weight of the h^{th} Mel scale, where $\text{h} = 1, \ldots, \text{M}$. The relevant formula can be expressed as follows:

$$\text{H}_{\text{h}}(\text{f}) = \begin{cases} 0 & \text{f} < If[\text{h}-1] \\ \dfrac{\text{f} - If[\text{h}-1]}{If[\text{h}] - If[\text{h}-1]} & If[\text{h}-1] \le f \le If[\text{h}] \\ \dfrac{If[\text{h}+1] - \text{f}}{If[\text{h}+1] - If[\text{h}]} & If[\text{h}] \le f \le If[\text{h}+1] \\ 0 & \text{f} > If[\text{h}+1] \end{cases}$$

where $\text{H}_{\text{h}}(\text{f}) = 1$ when $\text{f} = If[\text{h}]$

TABLE 3. Information of feature.



When the entire signal of each frame passes through a Mel filter bank, this operation is used to imitate the human ear to achieve precise frequency identification, smooth the spectrum, and reduce the number of calculations.

The logarithmic filter bank energy ($E[h]$) is the output of the weighted sum of the power spectrum passing through the hth Mel filter bank.

- *Definition 4:* The following equation can be obtained when the number of Mel filters is M and the number of discrete frequencies is K:

$$E[h] = ln\left[\sum_{k=1}^{K} |X[k]|^2 H_h(k)\right], \quad h = 1, 2, \ldots, M$$

Through the Discrete Cosine Transform (DCT), M log energy values can be obtained in the frequency domain, and l^{th} feature can be extracted.

$$C[l] = \sum_{h=1}^{M} E[h] \cos\left((h - 0.5)\frac{l\pi}{M}\right), \quad l = 1, 2, \ldots, L$$

where $C[l]$ is the l^{th} feature extracted using MFCCs and L is the total number of extracted features for each examined frame. In practice, L < M. According to the procedures described in [33], [41], [46], and [48], the present study set L and M to 12 and 26, respectively.

7) *Delta Cepstrum Coefficient:* A voice signal changes over time, similar to the slope of a formant at its transitions. Therefore, the trajectory of MFCCs over time (i.e., the Delta cepstral coefficient) must be determined. The term $\partial C[l]/\partial t$ is the first-order derivative of the l^{th} feature vector of an MFCC in an time window. The addition of acceleration features such as Delta-Delta cepstral coefficients, which are obtained through double partial differentiation with respect to time, generally leads to superior classification performance [41].

8) *Logarithmic Energy:* The energy of each frame is a crucial feature that represents the variation in amplitude and provides acoustic information [33], [41], and [48].

- *Definition 5:* Given N sampling points in a Hamming window, the logarithmic energy of each frame (LE) can be expressed as follows:

$$LE = ln\left[1 + \sum_{n=1}^{N} \left(\tilde{S}[n]\right)^2\right]$$

9) *Quantified Feature Vector:* This study combined 13 features derived through MFCC processing with prosodic features of voice signals which are widely used for speaker recognition [32], [36], [38], [48], and [49]. The prosodic features, namely voice pitch contours and intensity, were extracted using Praat (version 6.0). Thus, the quantified feature vector for each frame contained 15 dimensions. When the first- and second-order derivatives of time windows were considered, the vector had a total of 45 dimensions. The vectors were fed to LSTM-RNNs for satisfaction classification. Because the outputs were the categorical variables, the present study adopted one-hot encoding to denote which were satisfaction, neutrality, and dissatisfaction, labeled with "001", "010", and "100",

TABLE 4. Notations for LSTM-RNN formulation.

Notation	Description
W^{ci}, W^{co} and W^{cf}	the weights from input cepstrum to the input gate, output gate and forget gate, respectively
W^{hi}, W^{ho} and W^{hf}	the weights for the transmission of the value in the hidden units h to the input gate, output gate, and forget gate, respectively
b_i, b_o, b_f and b_e	the bias values of the input gate, output gate, forget gate, and hidden layer, respectively
T_S	the output values of hidden units for frame s
I_s, O_s and F_s	the values of the input gate, output gate, and forget gate, respectively, for frame s
$\tilde{C}_s$ and C_S	the short-term memory and long-term memory, respectively, for frame s
W^{ch} and W^{hh}	the weights from the input cepstrums and previous output values to the hidden unit, respectively
$b_{\tilde{C}}$	the bias value of the short term memory, $\tilde{C}_S$
W^{oq}	the weights from output values to the output nodes q which is a categorical variable
M_h	Number of hidden layer
E^e	Values of e^{th} hidden layer
Y_q	the ratio of the output value of q.

where

$$I_S = \sigma(W^{ci}C_S^E + W^{hi}T_{S-1} + b_i)$$
$$O_S = \sigma(W^{co}C_S^E + W^{ho}T_{S-1} + b_o)$$
$$F_S = \sigma(W^{cf}C_S^E + W^{hf}T_{S-1} + b_f)$$

where

$$\tilde{C}_S = tanh(W^{ch}C_S^E + W^{hh}T_{S-1} + b_{\tilde{C}})$$
$$C_S = F_S C_{(S-1)} + I_S \tilde{C}_S$$

The output value, $T_S = O_S tanh(C_S)$

Let $T = [T_1, \cdots, T_S]$

$$E^1 = \sigma_1(W^{T,1}T + b_1)$$
$$E^e = \sigma_e(W^{e-1,e}E^{e-1} + b_e), \text{ as } 2 \le e \le M_h$$
$$Y_q = \frac{e^{E^{M_h} \cdot W^{oq}}}{\sum_{q=1}^{Q} e^{E^{M_h} \cdot W^{oq}}}$$
$$O_q = arg\_\max_{q \in Q} Y_q$$

Algorithm: Pseudocodes of MFCC and LSTM-RNNs

Pseudocode of the discriminative model

A_{pc}// the audio clip of input of participant, p

Input: $A_{pc}, F_s, F_i, N, S_r, \beta, \alpha, \mathrm{M}, \mathrm{L}$(MFCC parameters),
$C^E_{p,s}, H_l, H_u, L_r, E_p, Q$(LSTM parameters)

Output: W (LSTM minimizes loss function by tuning W)

// Construction of quantified feature vectors

1. For $p = 1$ to P# number of participants
2. $C_{p,s} = \mathrm{MFCC}(A_{pc}, F_s, F_i, N, S_r, \beta, \alpha, \mathrm{M}, \mathrm{L})$
3. End For
4. $C^E_{p,s} = \langle C_{p,s}, \Delta C_{p,s}, \Delta\Delta C_{p,s} \rangle$

// Training the discriminative model

5. For $p = 1$ to P
6. $\mathrm{W} = \langle W^{ci}, W^{co}, W^{cf}, W^{hi}, W^{ho}, W^{hf}, W^{ch}, W^{hh}, W^{oq} \rangle$

// Train LSTM to derive W which minimizes the loss function

7. $\mathrm{W} = \mathrm{TrainLSTM}(C^E_{p,s}, H_l, H_u, L_r, E_p, Q)$
8. End for
9. Return

FIGURE 4. Pseudocode of the discriminative model.

respectively. Table 3 showed the information of feature vector and labeling.

D. LSTM MODEL CONSTRUCTION

The 45-diminsional vectors along with corresponding labels were fed into LSTM-RNNs to develop a discriminative model for determining the level of customer satisfaction.

Table 4 shows the notations for LSTM-RNN formulation. $C_{p,s}$ is the array of cepstrum of frame s for participant p and $C^E_{p,s}$ is the extended $C_{p,s}$ with Delta and Delta-Delta, time derivative.

$C^E_{p,s} = \langle C_{p,s}, \Delta C_{p,s}, \Delta\Delta C_{p,s} \rangle$. To simplify the notation p is omitted in LSTM-RNN formulation.

Fig. 4 shows the complete pseudo code.

E. SVM MODEL CONSTRUCTION

SVM is a robust machine learning algorithm which can reach good performance with the relatively limited training data [41]. SVM is mostly used in pattern classification by finding a best hyperplane with maximum margin to separate data of one class from those of the other. Typically, by using a non-linear kernel function, the original feature vector can be transformed into a hyper plane where the data can be linearly classified or non-linearly separable. Given a training dataset $\left\{\left(C^E_p, y_p\right), p \in P\right\}$, where $y_p \in \{-1, 1\}$, kernel function, $\Phi(*)$, the goal is to adjust the vector of separating hyperplane, $(\vec{w}_h, b)$ to maximize the margin between supporting vectors under constraints.

$$\frac{1}{2}\|\vec{w}_h\|^2 + \delta \sum\nolimits_{p \in P} \xi_p$$

with constraints that

$$y_p\left(< \vec{w}_h, \Phi(C^E_p) > +b\right) \geq 1 - \xi_p$$

where $\xi_p \geq 0$, is the error distance, δ is the cost of error inflicted on the classifier. To classify a sample, C^E_p, the decision function of prediction is as follows:

$$f(x) = sign\left(< \vec{w}, \Phi(C^E_p) > +b\right)$$

where *sign* is the function to justify the positive or negative argument by returning $+1$ and -1, respectively.

F. AUTO-ENCODER NEURAL NETWORKS

AE neural networks comprise three major components: an encoder is used to reduce the dimensionality of input data and develop a compressed representation of input data; a latent space shows the compressed representation; a decoder is to reconstruct the input data from the compressed representation. The encoder and decoder consist of a series of neural layers containing the decreasing and increasing nodes, respectively.

1) Processing of Encoder:

$$h^u_e = \sigma^u_e(W^u_e h^{u-1}_e + b^u_e)$$

where σ^u_e, b^u_e, and W^u_e is the activation function, bias vector and input weight vectors of the u^{th} encoding layer, respectively, and $h^0_e = C^E_P$.

2) Processing of Decoder:

$$h^u_d = \sigma^u_d(W^u_d h^{u-1}_d + b^u_d)$$

where σ^u_d, b^u_d, and W^u_d is the activation function, bias vector and input weight vectors of the u^{th} decoding layer, respectively, and $h^0_d = h^\Gamma_e$. Γ is the layers of encoding.

The lost function is $\left|\left|C^E_P - h^\Gamma_d\right|\right|$.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. DATA COLLECTION

This study comprised two major tasks: determining the participants' degree of satisfaction through a post-consumption questionnaire survey; and developing a discriminative model of satisfaction by feeding quantified feature vectors to LSTM-RNNs. Because a dataset containing acoustic files related to satisfaction is not publicly available, this study collected an acoustic dataset by using the protocol described in Section III.

The recording equipment used in the experiment was an external microphone (E-books S60) that can collect similar signals to those collected by mobile phones. The participants were instructed to make the statement "我對於該產品滿意" (it means, I am satisfied with this item) after drinking the provided beverage. The voices of the participants were recorded as acoustic files in .wav format.

TABLE 5. Demographics of the participants (Total = 150).

	Categories	Numbers of participants	Percentage
Age (years)	19 and below	18	12%
	20-39	66	44%
	40-64	45	30%
	65 and above	21	14%
Gender	Female	88	59%
	Male	62	41%
Education	High school	24	16%
	College graduate	114	76%
	Postgraduate	12	8%

TABLE 6. Satisfaction-related information for the provided beverages.

	Satisfaction	Neutrality	Dissatisfaction	Total
Coffee	11	13	10	34
Milk tea	9	12	9	30
Green tea	15	12	4	31
Fruit and Vegetable juice	11	10	10	31
Chinese bitter tea	5	5	14	24
Total	51	52	47	150

TABLE 7. Descriptive statistics of the questionnaire data.

	Item 1	Item 2	Item 3	Item 4
Numbers of responses	150	150	150	150
Minimum	0	0	0	0
Maximum	4	4	4	4
Mean	2.31	2.11	2.03	2.08
Median	2.50	2.00	2.00	2.00
Mode	3.00	2.00	2.00	2.00
S.D.	1.12	0.95	1.01	1.05

B. EXPERIMENTAL RESULTS

1) Statistical Analysis: The demographic information of the 150 participants is presented in Table 5. The research sample had a diverse age profile, with 44% of the participants being aged 20-39 years. Moreover, 59% and 76% of the sample comprised women and college graduates, respectively.

Table 6 presents the relationship between the level of satisfaction and the drink served. Green tea and Chinese bitter tea were the most and least preferred beverages, respectively. The number of different drinks served was balanced among the participants. Moreover, the numbers of satisfied, neutral, and dissatisfied participants were similar.

After consuming the beverage, the participants were requested to fill out the adopted questionnaire surveys to indicate their level of satisfaction with the beverage served. The descriptive statistics of the questionnaire data are presented in Table 7. Four items related to satisfaction were evaluated on a 5-point Likert scale ranging from 0 (''very dissatisfied'') to 4 (''very satisfied''). The total number of responses for each item was 150. The mean scores for the four items were 2.31, 2.11, 2.03, and 2.08, demonstrating the tendency of neutrality in the conducted survey.

The factor loadings of the four items are presented in Table 8, indicating that they are above the threshold of 0.6.

TABLE 8. Validity analysis.

Construct	Items	Factor loadings	AVE(0.5)	CR(0.7)
Satisfaction	S01	0.822	0.600	0.857
	S02	0.805		
	S03	0.764		
	S04	0.701		

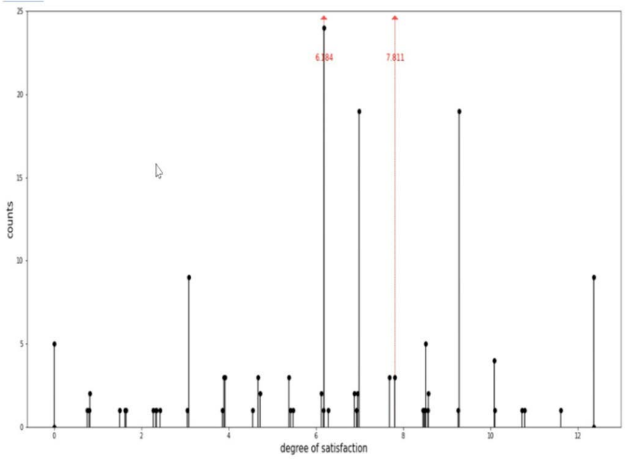


FIGURE 5. Clusters of participant satisfaction.

The Cronbach's Alpha (CA), Average Variance Extracted (AVE), and Composite Reliability (CR) of the satisfaction construct were 0.950, 0.600, and 0.857, respectively, which are above the relevant thresholds of 0.7, 0.5, and 0.7, respectively [50] and [51]. These results confirm the reliability and validity of the adopted questionnaire; thus, the survey results could be used for accurately identifying the level of satisfaction in the collected acoustic files.

The satisfaction vector obtained from the survey data for each participant comprised four features. The inner product of this vector with the factor loading represented each participant's level of satisfaction. The VB process in SPSS (version 18) was used to automatically cluster the participants into three bins, and the corresponding result is displayed in Fig. 5, where the X-axis represents the degree of satisfaction and the Y-axis represents the number of participants. The cut points were set to 6.184 and 7.811. The numbers of participants clustered into satisfaction, neutrality, and dissatisfaction were 51, 52, and 47, respectively.

2) Acoustic Extraction: The voice of each participant was recorded for approximately 2.5 s and saved in .wav format. The recorded voice file was then cleaned using Audacity 2.3 which is an open-source software program. The noise reduction function of this software was used to eliminate environmental noise. Voices were divided into multiple 25-ms frames, and 12 features were extracted. By combining the logarithmic energy as well as the voice intensity and voice pitch determined using Praat with the aforementioned 12 features, this study derived a 15-dimensional vector

TABLE 9. Parameters used for feature extraction.

Items	Parameter setting
Frame size, F_S	25 ms
Frames overlap interval, F_i	10 ms
Sampling rate, S_r	44.1 kHz
Bit depth	16
Sampling points, N	1100
Pre-emphasis coefficient, β	0.97
Hamming window parameter, α	0.46
Number of Mel scale filter bank, M	26
Number of extracted features, L	12

TABLE 10. The ranges of hyperparameters used for auto-encoder neural networks.

Hyperparameters	Range
Hidden layer in Encoder	1, 2, 3, 4, 5
Dimensions of hidden layer	43, 41, 39, 37, 35
Number of Epochs	1 to 500
Batch size	10
Activation function	Sigmoid, ReLU, ELU
Optimizer	Adam
Loss function	Mean Square Error
Objective function	Minimize (loss)

TABLE 11. The summaries of AE configuration.

Input dimensions	layers deployed in AE	Output dimensions	Activation function	Optimizer	Number of parameters
45	43	45	ELU	Adam	3958
45	43,41,43	45	ELU	Adam	7568
45	43,41,39,41,43	45	ELU	Adam	10846
45	43,41,39,37,39,41,43	45	ELU	Adam	13808
45	43,41,39,37,35,37,39,41,43	45	ELU	Adam	16470

for each frame. Finally, by calculating the Delta and Delta-Delta cepstral coefficients between frames, the study derived 45-dimensional feature vectors. These vectors were fed to the developed discriminative model for detecting the level of satisfaction. The parameter values adopted for feature extraction are presented in Table 9.

3) Feature Reconstruction with Auto-Encoder: As the amount of voice clips was 150 and the feature vectors were of 45 dimensions, this study utilized Auto-Encoder (AE) to further reduce dimensionality and reconstruct the feature vectors. To decide the optimal number of hidden layers deployed in the AE neural networks, a bi-node encoding and decoding approach was adopted: the encoding side started with a layer of 43 hidden nodes, followed by a series of layers, each reduced two nodes from the previous layer until the desired number of features was reached. The decoding side mirrored the encoding with each layer adding two nodes. The loss function adopted was mean square error (MSE).
The ranges of hyperparameters being considered are presented in Table 10. Five-fold cross validation was adopted to decide what hyperparameters were the most feasible for each AE configuration. The hyperparameters with the best average performance were selected to reconstruct the original feature vectors. Table 11 summarizes the AE configurations with the best performance.

TABLE 12. The ranges of hyperparameters used for model optimization.

Hyperparameters	Range
Hidden layer, H_l	1, 2, 3
Number of cells, H_u	25, 30, 35, 40
Learning rate, L_r	0.001, 0.01, 0.1
Number of Epochs, E_p	1 to 50
Activation function	Sigmoid, ReLU, Tanh
Optimizer	Adam, SDG, RMSProp

TABLE 13. LSTM-RNN parameters for 45 dimensions.

Parameter Setting	
Hidden layer, H_l	1
Number of cells, H_u	30
Learning rate, L_r	0.001
Input variable, $C_{p,s}^{E}$	45
Output variable, Q	3
Batch	All input data
Epoch, E_p	10
Activation function	Sigmoid
Output activation function	Softmax
Optimizer	Adam
Loss function	Cross-entropy
Objective function	Minimize (loss)

The feature vectors extracted from the most dense layer of the AE were fed to both SVM and LSTM-RNN classifiers to develop the discriminative model.

4) Classification: To compare and contrast the performance, the original 45-dimensional and the reconstructed feature vectors were fed into LSTM-RNNs and SVM to develop the discriminative model independently. Each model was established and verified by five-fold nested cross-validation to tune hyperparameters, train model parameters, and evaluate the accuracy of the developed models.
For LSTM-RNNs, categorical cross-entropy and Softmax were selected as the loss and activation functions of the output layer, respectively. A Grid Search was conducted to determine values of hyperparameters among given ranges, which are presented in Table 12. Table 13 shows the determined hyperparameters for the model with the original 45-dimensional feature vectors

TABLE 14. The results overview of LSTM-RNNs with different AE configurations.

Classifier	Input Dimensionality	Number of Cells	Activation Function	Output Activation Function	Optimizer	Highest Predictive Accuracy	Average Predictive Accuracy
LSTM without flatten layer	45	30	Sigmoid	Softmax	Adam	75.86%	70.68%
LSTM with flatten layer	45	30	Sigmoid	Softmax	Adam	65.63%	55.13%
LSTM with flatten layer + AE	43	40	Tanh	Softmax	RMSProp	65.52%	57.86%
	41	35	Tanh	Softmax	RMSProp	62.50%	56.57%
	39	35	Tanh	Softmax	RMSProp	72.41%	64.87%
	37	30	Sigmoid	Softmax	RMSProp	75.86%	71.95%
	35	30	Tanh	Softmax	SGD	62.07%	53.28%

TABLE 15. LSTM-RNN hyperparameters of the best performance model.

Parameter Setting	
Hidden layer, H_l	1
Number of cells, H_u	30
Learning rate, L_r	0.1
Input variable, $C_{p,s}^{E}$	37
Output variable, Q	3
Batch	All input data
Epoch, E_p	10
Activation function	Sigmoid
Output activation function	Softmax
Optimizer	RMSProp
Loss function	Cross-entropy
Objective function	Minimize (loss)

TABLE 16. Confusion matrix Of LSTM-RNNs with the best performance.

	Satisfaction (001)	Neutrality (010)	Dissatisfaction (100)
Predicted satisfaction	7	2	1
Predicted neutrality	2	7	1
Predicted dissatisfaction	1	0	8

as the input. The optimized networks contained only 1 level of hidden layer with 30 cells.

The input data are sequentially into the LSTM-RNN model. The output data are accumulated in two ways. In the first one, input is combined with the hidden value of the previous step to calculate the current state. The hidden values of the last step which contains the summarized information are delivered for classification. In the second way, to preserve the information of all states, a flatten layer is utilized to collect the hidden value of each step for classification.

In this study, a hidden layer was affixed to the flatten layer for further feature reduction. The experimental results are shown in Table 14.

At first, the model without the flatten layer seemed to outperform the model with the flatten layer. However, with the enhancement of AE, the models with flatten layers gradually improved and reached the best with 37-dimensional latten features as the input, of which the hightest and average prediction accuracy were 75.86% and 71.95%, respectively. The

TABLE 17. Ranges of the hyperparameters of SVM used for model optimization.

Hyperparameters	Range
Kernel	Polynomial, RBF, Sigmoid
Cost	1 to 10
Gamma	Scale, 0 to 1

TABLE 18. The evaluation results of the SVM with different AE configurations.

Classifier	Input Dimensionality	Kernel	Cost	Gamma	Highest Predictive Accuracy	Average Predictive Accuracy
SVM	45	RBF	1	Scale	65.52%	57.95%
SVM+AE	43	Poly	1	0.2413	62.07%	53.79%
	41	RBF	3	0.1975	67.74%	61.84%
	39	RBF	6	0.0132	75.86%	73.35%
	37	RBF	7	0.0991	79.31%	73.97%
	35	RBF	7	0.1274	75.00%	69.24%

TABLE 19. Confusion matrix of SVM with the best performance.

	Satisfaction (001)	Neutrality (010)	Dissatisfaction (100)
Predicted satisfaction	5	3	2
Predicted neutrality	0	9	1
Predicted dissatisfaction	0	0	9

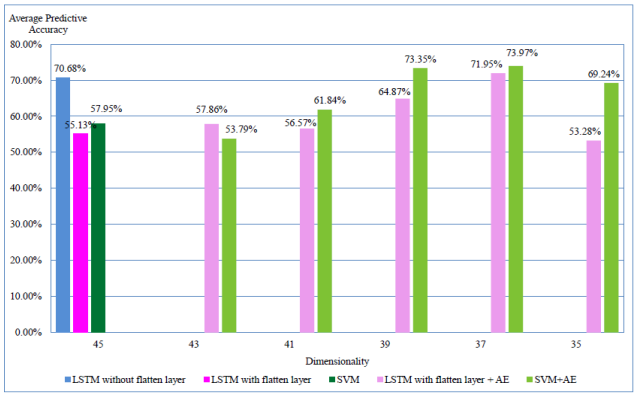


FIGURE 6. The performance comparison of LSTM-RNNs and SVM with different AE configurations.

hyperparameters of the best model is shown in Table 15 and the confusion matrix is illustrated in Table 16.

An SVM, which is a supervised machine learning classifier, is among the most commonly used models in emotion classification [41], [48] and [52]. Accordingly, this study compared the classification performance of the SVM models with that of the LSTM-RNN models. Three SVM hyperparameters were tuned through nested cross validation, namely, Kernel, Cost, and Gamma. The ranges of these hyperparameters are presented in Table 17. The evaluation results of the SVM with the original and the reconstructed feature vectors as input are summarized in Table 18.

Suprisingly, with 37-dimensional latten features, SVM model with Radial Basis Function (RBF) Kernel, Cost and

Gamma being set at 7 and 0.0991, reached the best and average accuary of 79.31% and 73.97%, repsectively, which were higher than the corresponding LSTM-RNN model's. The average accuracy of the SVM models with the reconstructed feature vectors were significantly improved from 53.79% to 73.97%. Table 18 and 19 show the performace of various parameters and the confusion matrix of the best model.

AE neural networks reduced the high dimensionality of the feature vectors and improved the accuracy of both LSTM-RNN and SVM models. Interestingly, in this study, SVM models consistenly outperformed LSTM-RNN models while applying AE as shown in Fig. 6. The result revealed that when the number of data was limited, SVM may outperform LSTM-RNNs due to its simplicity.

V. CONCLUSION

Customer satisfaction drives customer decisions and highly affects customer repurchase behavior. Therefore, determining customer satisfaction is essential for enterprises to gain a clearer understanding of customer expectations and improve their customer service quality. Traditionally, a post-purchase questionnaire survey is conducted to collect the information. However, such surveys require considerable effort from both customers and enterprises. Therefore, a tool that can conveniently collect customer satisfaction should be in high order. However, before the effectiveness of a tool can be verified, a sound data set has to be collected.

This study developed a discriminative model for detecting the level of customer satisfaction by analyzing the implicit information embedded in the tone of customers' voices. The study methodology involved three steps: feature extraction, feature reduction, and model construction. Features extracted through an MFCC-based technique were combined with prosodic features, namely voice pitch and voice intensity, to obtain 45-dimensional feature vectors. AE was used to reduce features for model construction. The reconstructed feature vectors were then fed to SVM and LSTM-RNNs to devise the discriminative model.

The level of satisfaction was divided into three categories: satisfied, neutral, and dissatisfied. The developed SVM model combined with AE achieved the highest accuracy of 79.31% in identifying participants' satisfaction level; it outperformed the LSTM-RNN model with AE, which reached accuracy of 75.86%.

A. IMPLICATIONS

During the recording of voice tones, this study instructed all participants to read the same statement: ''I am satisfied with this item.'' To extend the findings of this study, future studies could request participants to read different sentences to verify whether satisfaction levels can still be accurately identified. Furthermore, researchers could use real-world acoustic data to verify the feasibility of the approach used in this study.

The results of this study indicated that MFCC-based technique can extract voice features for satisfaction recognition. Future research could perform experiments by using other feature extraction methodologies, such as LPC- and LPCC-based techniques, and compare the results obtained.

This study confirmed that SVM and LSTM-RNNs can be used to construct a model that can detect the level of customer satisfaction. AE neural networks used to reconstruct features fed into SVM and LSTM-RNNs can significantly improve the predictive accuracy. However, because machine learning technology is advancing rapidly, new and improved methodologies for evaluating customer satisfaction can be explored in future research.

Telemarketing has the advantages of saving time and effort in the execution of marketing campaigns and provision of customer support. The identification of customer satisfaction through acoustic analysis can facilitate the implementation of customer relationship management. According to the results of this study, instead of measuring customer satisfaction through a post-purchase survey, customer satisfaction may be determined using a voice recognition model during interactions between agents and customers. The use of such a model can enhance customer relationships, provide managerial insights to improve internal management, and even support new marketing campaigns if it can be integrated into the business process of enterprises. The findings of this study pave the way for understanding customer preference and attitude through voice analysis.

B. LIMITATIONS AND SUGGESTIONS FOR FURTHER RESEARCH

This study has some limitations. First, the voice recordings used in this study were in Mandarin. Future studies can collect acoustic data in other languages to verify the feasibility of the proposed method. Second, because self-collected voice files were used in this study, the performance of the classification model might have been affected by the quality of the recorded voices, environmental noise, and variations in the participants' voices. Future studies can execute voice recordings in different environmental settings and examine whether the developed model is still accurate. Third, beverages were served free of charge in this study. Customers' satisfaction level might change when they must pay for products or services. Therefore, cautions should be exercised when generalizing the results of this research to environments where services or products are provided for a fee.

REFERENCES

[1] N. Kamaruddin, A. W. A. Rahman, and A. N. R. Shah, ''Measuring customer satisfaction through speech using valence-arousal approach,'' in *Proc. 6th Int. Conf. Inf. Commun. Technol. Muslim World*, Nov. 2016, pp. 298–303.

[2] Q. Llimona, J. Luque, X. Anguera, Z. Hidalgo, S. Park, and N. Oliver, ''Effect of gender and call duration on customer satisfaction in call center big data,'' in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, Sep. 2015, pp. 1825–1829.

[3] R. L. Oliver, ''A cognitive model of the antecedents and consequences of satisfaction decisions,'' *J. Marketing Res.*, vol. 17, no. 4, pp. 460–469, Nov. 1980.

[4] E. Garbarino and M. S. Johnson, ''The different roles of satisfaction, trust, and commitment in customer relationships,'' *J. Marketing*, vol. 63, no. 2, pp. 70–87, Apr. 1999.

[5] A. Ando, R. Masumura, H. Kamiyama, S. Kobashikawa, Y. Aono, and T. Toda, "Customer satisfaction estimation in contact center calls based on a hierarchical multi-task model," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 715–728, 2020.

[6] J. J. Cronin and S. A. Taylor, "Measuring service quality: A reexamination and extension," *J. Marketing*, vol. 56, no. 3, pp. 55–68, Jul. 1992.

[7] D. H. Kincade, V. L. Giddings, and H. J. Chen-Yu, "Impact of product-specific variables on consumers' post-consumption behaviour for apparel products: USA," *J. Consum. Stud. Home Econ.*, vol. 22, no. 2, pp. 81–90, Jun. 1998.

[8] M. Brennan, S. Benson, and Z. Kearns, "The effect of introductions on telephone survey participation rates," *Int. J. Market Res.*, vol. 47, no. 1, pp. 65–74, Jan. 2005.

[9] D. Kang and Y. Park, "Review-based measurement of customer satisfaction in mobile service: Sentiment analysis and VIKOR approach," *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1041–1050, Mar. 2014.

[10] J. Park, "Framework for sentiment-driven evaluation of customer satisfaction with cosmetics brands," *IEEE Access*, vol. 8, pp. 98526–98538, 2020.

[11] S. Godbole and S. Roy, "Text classification, business intelligence, and interactivity: Automating C-Sat analysis for services industry," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2008, pp. 911–919.

[12] Y. Park and S. C. Gates, "Towards real-time measurement of customer satisfaction using automatically generated call transcripts," in *Proc. 18th ACM Conf. Inf. Knowl. Manage. (CIKM)*, Nov. 2009, pp. 1387–1396.

[13] J. Sun, W. Xu, Y. Yan, C. Wang, Z. Ren, P. Cong, H. Wang, and J. Feng, "Information fusion in automatic user satisfaction analysis in call center," in *Proc. 8th Int. Conf. Intell. Hum.-Mach. Syst. Cybern. (IHMSC)*, Aug. 2016, pp. 425–428.

[14] M. Plaza and L. Pawlik, "Influence of the contact center systems development on key performance indicators," *IEEE Access*, vol. 9, pp. 44580–44591, 2021.

[15] A. Tanaka, A. Koizumi, H. Imai, S. Hiramatsu, E. Hiramoto, and B. de Gelder, "I feel your voice: Cultural differences in the multisensory perception of emotion," *Psychol. Sci.*, vol. 21, no. 9, pp. 1259–1262, Aug. 2010.

[16] P. Liu, S. Rigoulot, and M. D. Pell, "Culture modulates the brain response to human expressions of emotion: Electrophysiological evidence," *Neuropsychologia*, vol. 67, pp. 1–13, Jan. 2015.

[17] ScienceDaily. (Feb. 24, 2015). *Cross-Cultural Communication: Much More Than Just a Linguistic Stretch*. Accessed: Apr. 22, 2020. [Online]. Available: https://www.sciencedaily.com/releases/2015/02/150224102843.htm

[18] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *J. Personality Social Psychol.*, vol. 70, no. 3, pp. 614–636, Apr. 1996.

[19] V. Bostanov and B. Kotchoubey, "Recognition of affective prosody: Continuous wavelet measures of event-related brain potentials to emotional exclamations," *Psychophysiology*, vol. 41, no. 2, pp. 259–268, Mar. 2004.

[20] A. Bhattacherjee, "Understanding information systems continuance: An expectation-confirmation model," *MIS Quart.*, vol. 25, no. 3, pp. 351–370, 2001.

[21] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[22] E. W. Anderson and M. W. Sullivan, "The antecedents and consequences of customer satisfaction for firms," *Marketing Sci.*, vol. 12, no. 2, pp. 125–143, 1993.

[23] P. G. Patterson and L. W. Johnson, "Disconfirmation of expectations and the gap model of service quality: An integrated paradigm," *J. Consum. Satisfaction, Dissatisfaction Complaining Behav.*, vol. 6, no. 1, pp. 90–99, 1993.

[24] V. A. Zeithaml, L. L. Berry, and A. Parasuraman, "The behavioral consequences of service quality," *J. Marketing*, vol. 60, no. 2, pp. 31–46, Apr. 1996.

[25] E. Karahanna, D. Straub, and N. L. Chervany, "Information technology adoption across time: A cross-sectional comparison of pre-adoption and post-adoption beliefs," *MIS Quart.*, vol. 23, no. 2, pp. 183–213, 1999.

[26] B. Edvardsson, M. D. Johnson, A. Gustafsson, and T. Strandvik, "The effects of satisfaction and loyalty on profits and growth: Products versus services," *Total Qual. Manage.*, vol. 11, no. 7, pp. 917–927, Sep. 2000.

[27] Y.-K. Lee, C.-K. Lee, J. Choi, S.-M. Yoon, and R. J. Hart, "Tourism's role in urban regeneration: Examining the impact of environmental cues on emotion, satisfaction, loyalty, and support for Seoul's revitalized Cheonggyecheon stream district," *J. Sustain. Tourism*, vol. 22, no. 5, pp. 726–749, Jan. 2014.

[28] C. Ranaweera and J. Prabhu, "The influence of satisfaction, trust and switching barriers on customer retention in a continuous purchasing setting," *Int. J. Service Ind. Manage.*, vol. 14, no. 4, pp. 374–395, Oct. 2003.

[29] F. F. Reichheld, "The one number you need to grow," *Harvard Bus. Rev.*, vol. 81, no. 12, pp. 46–55, Dec. 2003.

[30] O. Abdel-Hamid, A. R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 10, pp. 1533–1545, Oct. 2014.

[31] P. M. Chauhan and N. P. Desai, "Mel frequency cepstral coefficients (MFCC) based speaker identification in noisy environment using Wiener filter," in *Proc. Int. Conf. Green Comput. Commun. Electr. Eng. (ICGCCEE)*, Mar. 2014, pp. 1–5.

[32] V. Tiwari, "MFCC and its applications in speaker recognition," *Int. J. Emerg. Technol.*, vol. 1, no. 1, pp. 19–22, Feb. 2010.

[33] J. Martinez, H. Perez, E. Escamilla, and M. M. Suzuki, "Speaker recognition using Mel frequency cepstral coefficients (MFCC) and vector quantization (VQ) techniques," in *Proc. 22nd Int. Conf. Electr. Commun. Comput.*, Feb. 2012, pp. 248–251.

[34] S. Singh and E. G. Rajan, "MFCC VQ based speaker recognition and its accuracy affecting factors," *Int. J. Comput. Appl.*, vol. 21, no. 6, pp. 1–6, May 2011.

[35] L. Wang, K. Minami, K. Yamamoto, and S. Nakagawa, "Speaker identification by combining MFCC and phase information in noisy environments," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2010, pp. 4502–4505.

[36] N. Singh, R. A. Khan, and R. Shree, "MFCC and prosodic feature extraction techniques: A comparative study," *Int. J. Comput. Appl.*, vol. 54, no. 1, pp. 9–13, Sep. 2012.

[37] P. P. Dahake, K. Shaw, and P. Malathi, "Speaker dependent speech emotion recognition using MFCC and support vector machine," in *Proc. Int. Conf. Autom. Control Dyn. Optim. Techn. (ICACDOT)*, Sep. 2016, pp. 1080–1084.

[38] G. Friedland, O. Vinyals, Y. Huang, and C. Müller, "Prosodic and other long-term features for speaker diarization," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 5, pp. 985–993, Jul. 2009.

[39] I. Luengo, E. Navas, I. Hernáez, and J. Sánchez, "Automatic emotion recognition using prosodic parameters," in *Proc. 9th Eur. Conf. Speech Commun. Technol.*, Sep. 2005, pp. 493–496.

[40] R. W. M. Ng, T. Lee, C.-C. Leung, B. Ma, and H. Li, "Analysis and selection of prosodic features for language identification," in *Proc. Int. Conf. Asian Lang. Process.*, Dec. 2009, pp. 123–128.

[41] M. S. Sinith, E. Aswathi, T. M. Deepa, C. P. Shameema, and S. Rajan, "Emotion recognition from audio signals using support vector machine," in *Proc. IEEE Recent Adv. Intell. Comput. Syst. (RAICS)*, Dec. 2015, pp. 139–144.

[42] K. Sönmez, E. Shriberg, L. Heck, and M. Weintraub, "Modeling dynamic prosodic variation for speaker verification," in *Proc. 5th Int. Conf. Spoken Lang. Process. (ICSLP)*, Nov. 1998, pp. 3189–3192.

[43] P. Rosso, L.-F. Hurtado, E. Segarra, and E. Sanchis, "On the voice-activated question answering," *IEEE Trans. Syst., Man, Cybern., C, Appl. Rev.*, vol. 42, no. 1, pp. 75–85, Jan. 2012.

[44] X. Sun, Z. Pei, C. Zhang, G. Li, and J. Tao, "Design and analysis of a human–machine interaction system for researching human's dynamic emotion," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 10, pp. 6111–6121, Oct. 2021.

[45] *Welcome to Python_Speech_Features's Documentation!*. Accessed: May 20, 2017. [Online]. Available: https://python-speech-features.readthedocs.io/en/latest/

[46] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proc. IEEE*, vol. 66, no. 1, pp. 51–83, Jan. 1978.

[47] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," *Proc. Inst. Phonetic Sci.*, vol. 17, no. 1193, pp. 97–110, Mar. 1993.

[48] M. Pakyurek, M. Atmis, S. Kulac, and U. Uludag, "Extraction of novel features based on histograms of MFCCs used in emotion classification from generated original speech dataset," *Elektronika Elektrotechnika*, vol. 26, no. 1, pp. 46–51, Feb. 2020.

[49] P. G. Patterson and R. A. Spreng, "Modelling the relationship between perceived value, satisfaction and repurchase intentions in a business-to-business, services context: An empirical examination," *Int. J. Service Ind. Manage.*, vol. 8, no. 5, pp. 414–434, Dec. 1997.
[50] C. Fornell and D. F. Larcker, "Evaluating structural equation models with unobservable variables and measurement error," *J. Marketing Res.*, vol. 18, no. 1, pp. 39–50, Feb. 1981.
[51] J. F. Hair, R. E. Anderson, R. L. Tatham, and W. C. Black, *Multivariate Data Analysis*. Hoboken, NJ, USA: Prentice-Hall, 1998.
[52] A. Shirani and A. R. N. Nilchi, "Speech emotion recognition based on SVM as both feature selector and classifier," *Int. J. Image, Graph. Signal Process.*, vol. 8, no. 4, pp. 39–45, Apr. 2016.

YEN HUEI KO received the bachelor's degree from the Department of Chemical Engineering, National Cheng Kung University, in 1987, the master's degree from the Department of Chemical Engineering, University of California at Davis, in 1990, and the master's degree in business administration from the University of Southern California, in 1993. He is currently pursuing the Ph.D. degree with the Department of Business Administration, National Central University, Jhongli, Taoyuan, Taiwan. His current research interests include utilizing machine learning for data mining, business intelligence, and data modeling in business domains.

PING-YU HSU received the bachelor's degree from the Department of CSIE, National Taiwan University, in 1987, the master's degree from the Department of Computer Science, New York University, in 1991, and the Ph.D. degree from the Department of Computer Science, UCLA, in 1995. He is currently a Distinguished Professor with the Business Administration Department, National Central University, Taiwan, and the Secretary-in-Chief of the Chinese ERP Association. He is also the Dean of the School of Management, National Central University. He has published more than 100 journals and conference papers. His papers have been published in *Decision Support Systems*, *European Journal of Information Systems*, IEEE TRANSACTIONS, *Information Systems*, *Information Sciences*, and various other journals. His research interest include the business data related applications, including business analytics, data mining, business intelligence, and adoption issues of enterprise systems.

YU-CHIN LIU received the bachelor's and master's degrees from the Institute of Information Management, National Chiao Tung University, in 1988 and 1994, respectively, and the Ph.D. degree from the Department of Business Administration, National Central University, in 2006. She is currently an Associate Professor with the Department of Information Management, Shih Hsin University, Taiwan. Her research interest include business analytics and data mining applications in business domains. She has published several journals and conference papers. Her articles have been published in *Information Sciences*, *Expert System with Applications*, and various other journals.

PO-CHIAO YANG received the B.S. degree in information management from Tamkang University, in 2017, and the M.B.A. degree from the National Central University, in 2019. He is currently working as an Infrastructure Consultant and a Solution Architecture. His research interests include data mining, business intelligence, artificial intelligent, signal analytic, and adoption issue of enterprise systems. From 2017 to 2019, he focus on the research of building business model with artificial intelligence, mainly with neural network.