# Using Measures of Vowel Space for Autistic Traits Characterization

Chin-Po Chen, *Student Member, IEEE*, Ho-Hsien Pan, Susan Shur-Fen Gau, and Chi-Chun Lee, *Senior Member, IEEE*

***Abstract*—Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder that is prevalent and heterogeneous. Autistic traits describe a wide heterogeneity of behavior symptoms of ASD, and these traits are reflections of core neurodevelopment function deficits. Researchers have predominantly taken a clinical angle to understand autistic traits. They have been developing various clinical-grade instruments with behavioral codes to quantify autistic traits for diagnostic and research purposes. However, the need for highly trained professionals and the inevitable subjectivity limit their usage. Hence, researchers have been developing computational methods to address these issues. Among many efforts, methods based on computing speech have emerged rapidly due to their ability to characterize communicative behaviors and social interactions. Our work addresses one particular under-studied speech aspect: articulation-related acoustics, one of the broad autism spectrum symptoms. In this paper, we examine the articulatory information in a natural spoken interaction through measures of vowel space characteristics (VSCs) to understand autistic traits. Specifically, we approach by modeling statistical relationships of the corner vowel distributions and the interpersonal correlation of these relationships in conversation. Our method is evaluated by deriving VSC features and using them in ASD classification and regression tasks. We found these features predict autism-related communication assessment and add additional information to classification tasks. Furthermore, our analyses show a relationship between VSCs and autism-related communication deficit and also imply differences in VSCs between typical developing people and each ASD subgroup.**

Manuscript received 6 December 2022; revised 19 July 2023; accepted 25 October 2023. Date of publication 6 November 2023; date of current version 7 December 2023. This work was supported in part by the National Health Research Institute under Grants NHRI-EX106-10404PI, NHRI-EX107-10404PI, NHRI-EX108-10404PI, and NHRI-EX110-11002PI, and in part by the Taiwan, and Ministry of Science and Technology under Grants MOST110-2327-B-002-006 and MOST110-2634-F-007-012, Taiwan. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Juan Ignacio Godino-Llorente. *(Chin-Po Chen is co-first author.) (Corresponding authors: Ho-Hsien Pan; Susan Shur-Fen Gau; Chi-Chun Lee.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by NTHU-REC under Application No. 10501HE002 and by RINC under Application No. 201403109, and performed in line with the Declaration of Helsinki.

Chin-Po Chen and Chi-Chun Lee are with the Department of Electrical Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan (e-mail: stu94116@gapp.nthu.edu.tw; cclee@ee.nthu.edu.tw).

Ho-Hsien Pan is with the Department of Foreign Languages and Literatures, National Yang Ming Chiao Tung University, Hsinchu 30013, Taiwan (e-mail: hhpan@nycu.edu.tw).

Susan Shur-Fen Gau is with the Department of Psychiatry, National Taiwan University Hospital and College of Medicine, Taipei 10002, Taiwan (e-mail: gaushufe@gmail.com).

This article has supplementary downloadable material available at https://doi.org/10.1109/TASLP.2023.3330605, provided by the authors.

Digital Object Identifier 10.1109/TASLP.2023.3330605

***Index Terms*—Autism, vowel space characteristics, conversation, severity assessment, diagnosis.**

## I. Introduction

AUTISM Spectrum Disorder (ASD) is a prevalent, and broad heterogeneous spectrum of neurodevelopmental disorders. ASD prevelence rate is approximately 1.5% [1], [2] across countries, and 1% in Taiwan [3]. It is estimated to have a 15 trillion (USD) social cost by 2029, according to Cakir et al. [3]. The cost of medical, therapeutic, or educational expenses has become an issue that cannot be neglected [4]. Symptoms, which vary drastically from one person to another, are often associated with comorbidities [5], and can cause significant health problems for people with ASD.

Autistic traits describe sets of behavioral symptoms of ASD. These traits reflect deficits of core neurodevelopmental functions, featuring social-communicational deficits and repetitive sensory-motor behaviors [6]. For example, ASD often can not have suitable eye contact with other people when having conversations, making their interlocutor feel socially-awkward. Many ASD people are also unwilling to talk to others, despite having the proper speech & language abilities. In severe ASD cases, they can not even convey a complete message with their speech due to their incoherent usage of words or sentences. Furthermore, the inherent heterogeneity of the spectrum of autistic traits makes it difficult to stratify and phenotype ASD easily. For example, the latest edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) can not identify many cases that are diagnosed as ASD in its previous version (DSM-4), which causes controversies [7].

Through decades of studies, researchers have predominantly taken a clinical angle to quantify and understand autistic traits. Various clinical-grade instruments are developed using manual behavioral codes to quantify autistic traits. For example, the Social Responsiveness Scale (SRS) [8] and Social Communication Questionnaire (SCQ) [9] are two commonly used clinical instruments for autistic traits profiling. Advanced ASD assessment instruments such as Autism Diagnostic Observation Schedule (ADOS) [10] and Psychoeducational Profile (PEP-3) [11], measure patient's autistic trait through in-person interviews. These clinical-grade instruments are designed to understand the autistic traits of the patients comprehensively and are used for diagnostic and even research purposes. However, most valid and reliable assessments or diagnostic instruments need highly trained professional practitioners. Moreover, controversial yet probable, a

positional paper stated that human observational assessments would be prone to inevitable subjectivity and most importantly, to non-scalable issues [12], bringing limitations to the current status-quo usage of these existing clinical instruments.

Developing automated methods has been regarded as having the potential to address these issues. Among many efforts, methods based on computing speech and language have emerged rapidly due to their ability to characterize the two most key dimensions of autistic traits: communicative and social behaviors during spoken interactions. Recent speech and language analytics for autistic traits mainly focus on speech acoustics, language, and conversation. A common approach of speech acoustics is to quantify ASD patient's atypical prosody. For example, Bone et al. designed and investigated various acoustic features to quantify abnormal traits in the speech prosody of ASD [13], [14], [15]. Computational methods of word usage has also been used to quantify autistic traits. For example, Li et al. quantified the term frequencies and word attributes of ASD patient's atypical usage of words or phrases [16] and found these measurements reflect the stereotyped idiosyncratic phrases of ASD. As speech acoustics and language are often used to quantify autistic traits, some studies focus on the acoustics within articulation—the acoustic patterns of certain phone units [17], [18]. This approach can measure speech acoustics when the participant is too young to speak meaningfully [17], and can also leverage the characteristic of tonal language (like Mandarin) to study a specific ASD cohort [18]. Lastly, methods for computing conversational dynamics, an angle that focuses on studying interactions, can also quantify autistic traits. For example, our past study used acoustic features with modified BERT embedding derived from interlocutors (where one of them is an ASD patient), and demonstrated that these can classify subtypes of ASD and predict autistic symptom severity [19].

While several related works exist, this work addresses one particular under-studied aspect of speech analysis for ASD: articulation-related acoustics. Speech sound error is one of the autistic traits that continue through adulthood. Few studies quantify this aspect of speech disorder for ASD. Several notable studies include efforts carried out by Bishop et al. and Kissine et al. Bishop et al. found an inverse relationship between autistic severity and vowel intelligibility. To be specific, vowel intelligibility, measured with vowel space area, is significantly correlated with pragmatic communication scores that are inversely correlated with autism severity [20]; Kissine et al. found autistic adults are more rigid in their articulation compared to typical developing adults by measuring their articulatory properties on vowel space [21]. These studies examined articulation acoustics but in highly-controlled experimental settings, such as measuring acoustic values for pre-set words. This study, however, investigates autistic traits of articulatory acoustics when the ASD participants engage in natural spoken interactions by measuring vowel space characteristics (VSCs).

This study contributes one of the first comprehensive and automated studies in developing vowel space characteristics measurements (VSC features) to characterize ASD-related social and communication traits. These VSC features are derived to characterize vowel articulation at different granularity, i.e., at an utterance level for communication function deficit and across an entire interaction episode for social function deficit. Our study shows that by fusing VSC features with the known high predictive power of acoustic-prosodic features improves the results in both ASD classification and severity score regression tasks. Additionally, we observe that formant dependency—a new VSC feature that measures articulator flexibility—and inter-vowel dispersion—indices measuring the discrimination of three corner vowel clusters in vowel space—correlates to the deficit in ASD communication severity. Moreover, simple fusion of VSC feature sets with acoustic-prosodic features achieves a 0.482% Pearson's correlation in regressing ADOS communication score, which is competitive to the prior SOTA that uses complicated and large BERT-based deep learning methods. Our result demonstrates the feasbility of using VSC features to stratify the heterogeneity of ASD.

The roadmap of this paper is in the following. In Section II we summarized prior work and pointed out the novelty of this research. Section III provides details about our experimental material. Section IV elaborates on how to derive those VSC features. Experiments consisting of three classification tasks and a regression task followed by analyses are presented in Section V. Finally, Sections VI and VII are the discussion and conclusion.

## II. Related Works

The following section summarizes relevant prior works that have computationally investigated the two aspects of autistic traits: speech production-related communication impairment and interaction-oriented social reciprocity deficits.

### A. Speech Production-Related Communication Impairment

Many researchers have designed autistic trait-related acoustic parameters to characterize ASD's speech communication impairment. They have shown the effectiveness of using prosodic features in modeling autistic speech [22]. However, the use of articulation-related acoustics parameters to characterize ASD is much less explored. For example, Bishop et al. analyzed the Vowel Space Area (VSA) of adult ASD participants and found that poor pragmatic communication skills cause narrower vowel space expansion and are associated with autistic traits [20]. Another study by Kissine et al. discovered phonetic inflexibility (stability) of autistic adults by analyzing the intra-category vowel dispersion of their vowel production [23]. Talkar et al. supposed that the coordination of articulators characterized by the correlation of the acoustic measurements between several articulators can differentiate ASD from TD. They found that ASD participants have lower precisions in their articulator movements [24].

Despite a few empirical studies that have shown that the articulation properties of ASD differentiate them from typical developing people and also reflect their autistic traits. The role of articulation-related acoustics in autism-related communication deficits remain largely unclear, especially in the context of spontaneous spoken interactions. To better understand the articulatory characteristics of each ASD participant, the setting in which

one performs study to assess an ASD participant's articulatory performance needs to be considered. For example, past research proposed that although speech sound error is one of the symptoms related to ASD, not every standard speech test is suitable for detecting speech sound error (SSE) of ASD testees [25]. McKeever et al. supposed that some of the standard speech test, such as the Photo Articulation Test [26] lacks sufficient articulatory variability. A complicated and naturalistic speech task, such as spontaneous conversation, is more suitable than single-word contexts to characterize the SSE of an individual ASD. The reason is that in normal conversation, a person needs to pay additional attention to the interactional context instead of merely focusing on speaking. Moreover, the challenges of adapting to the ambient environment induce SSE in an ASD patient [25]. In brief, although there have already been abundant studies that found relationships between autistic traits and speech prosody, the relationship between articulation-related acoustics and communication traits of ASD is under-studied. Furthermore, none of these prior research works in a naturalistic spoken interaction context, e.g., spontaneous dialogs.

### B. *Interaction-Oriented Social Reciprocity Deficits*

During face-to-face conversations, people actively adapt their vocal expressions to express their emotions and execute social functions; however, these skills require adaptive skills, which challenges people with ASD. ASD patients demonstrate inflexibility in voice modulation during conversations. For example, several reports from past studies have shown that ASD participants have lesser synchrony and convergence than those without ASD [27], [28], [29]. Owing to the deficit in social functions, the ASD participants present different spoken interaction patterns from the non-autistic participants. Past researchers have developed algorithms to characterize these atypical autistic-related speech behaviors. For example, many studies quantify prosodic attributes to characterize the dyadic interplay in which ASD participants are involved. They found that typical-developing people usually show more speech accommodation in conversations than ASD participants do [27], [28], [29]. Although prior studies have demonstrated repeated evidence that people with ASD have reduced speech accommodation or entrainment to their interlocutor in their speech prosody, the speech accommodation of ASD reflected in articulatory-related attributes seems unclear.

This study further advances the studies of autistic traits from an angle of articulation-related acoustics. This paper advances prior works by measuring vowel space characteristics (VSCs) in conversations, a more realistic phonetic environment, and explores the relationship between VSCs and autism-related communication deficits. We also measure the dyadic interplay of based on these VSC measurements to characterize autistic traits.

## III. Dataset Description

Spoken interaction data is collected through the Autism Diagnostic Observation Schedule (ADOS) interview process cite. ADOS involves a trained investigator conducting a semi-structured interview with the ASD participant. ADOS is organized into different sessions to elicit targeted ASD participant's spontaneous behaviors for the researchers and psychiatrists to assess their social and communicative functions. A couple of sessions are conversations in nature where the investigator and participant engage in dialogs about a certain topic. 'Emotion' session is a part of ADOS where the investigator discusses with the participant about his/her past emotional experiences in life, creating a scenario that simulates natural conversation. This kind of session provides sufficient materials for speech-based algorithm development. In fact, most recent studies on ASD participant's spontaneous speech behaviors utilize this particular session in ADOS as materials, e.g., Bone et al. used these sessions to design features for characterizing ASD-related atypical prosody; Li et al. trained deep neural networks in these sessions for designing diagnosis algorithms.

TABLE I
Database Description: The Demographics of the Participants in This Database

| Participants | Age: Mean(std) | Gender: (Male/Female) | $ADOS_{comm}$ Mean(std) |
|---|---|---|---|
| ASD total | 16.37 (4.34) | 76/10 | 11.71 (4.49) |
| TD | 13.35 (4.02) | 10/10 | 3.64 (3.92) |

Our ADOS data samples are collected by collaborating with National Taiwan University Hospital.[1] The data contains audio recordings from two lapel microphones attached to the investigator and the participant and video recordings from two fixed cameras facing the front of each person (details can be found in [30]). Table I shows a summary of participant demographics. There are a total of 86 ASD and 20 TD participants recruited in this database. The mean and standard deviation of age of the participants are ASD: 16.37, 4.34 and TD: 13.35, 4.02. The sizes of the cohorts are: ASD: 76 male and 10 female; TD: 10 male and 10 female. The mean and standard deviation of ADOS social and communication scores are: ASD: 11.71, 4.49 and TD: 3.64, 3.92.

This database is one of the largest in scale, and to our knowledge, is one of the few in Mandarin Chinese. This database has already been used in several prior studies. For example, our prior studies derived multi-modal speech and language features for ASD subgroup differentiation [30], [31]. Additionally, our recent study used a conversation-level modeling approach for automatic communication code assessment, and the study achieve 0.567% Pearson's correlation to manual coding [19]. In this work, we utilize the 'emotion' part of ADOS session to measure articulatory properties for autistic trait characterization. The emotion part lasts about 5–10 minutes, starting with positive emotion experiences like happiness and ending with negative emotions like fear. There are 12,010 utterances in the selected dataset totaling 67.35 minutes. Each session contains 111 utterances and 651 phones on average.

[1] Approved by IRB: REC-10501HE002 and RINC-20140319.

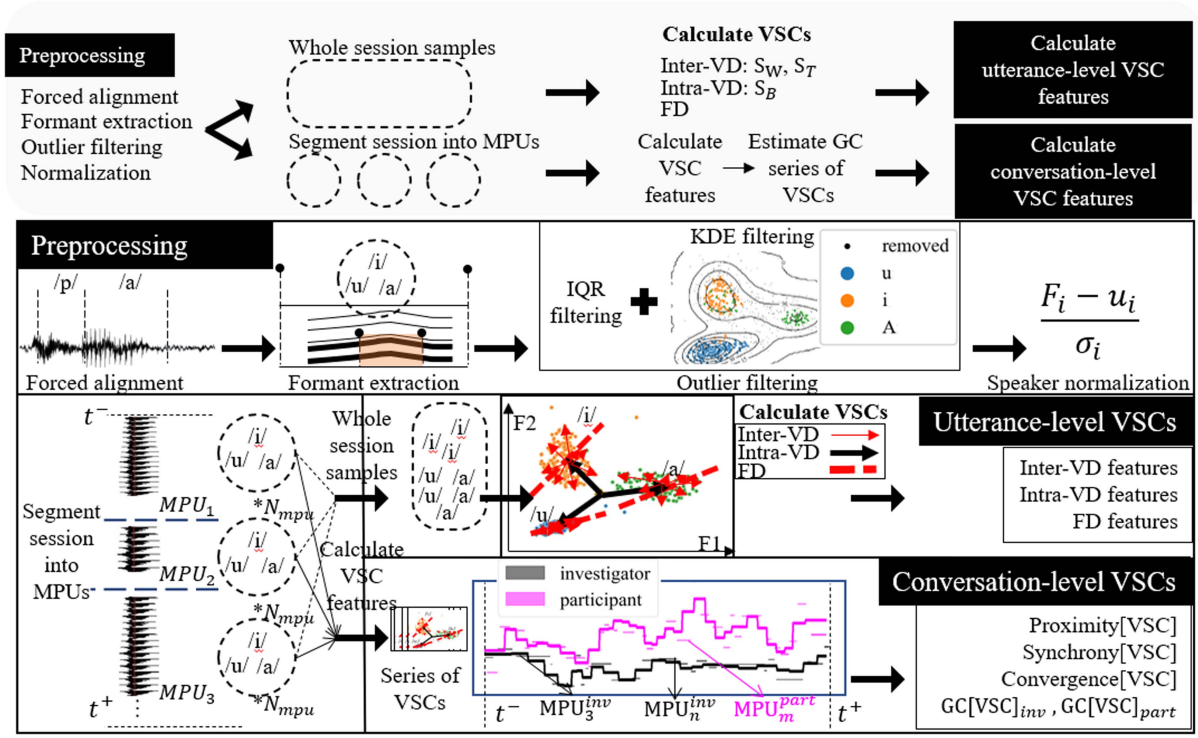


Fig. 1. Overall approach to deriving vowel space characteristic (VSC) features. The top of the figure shows the entire procedure, and the lower part shows the details of each component. The input speech will first undergo a pre-processing procedure, which includes forced alignment, feature extraction, two outlier filtering steps: IQR filtering and KDE filtering, and speaker normalization. Then the process is separated into two streams. The first stream gathers the whole session samples to calculate three vowel space characteristics (VSCs): inter-VD, intra-VD, and FD. Then utterance-level VSC features are calculated from those VSCs. The second stream first segments a session into multiple minimum phone units (MPUs). VSC features are calculated at those units, forming two VSC-based time series. From the two time series, we compute the gradual change (GC) of each and calculate several synchrony measuring metrics on the GC pair, creating conversation-level VSCs.

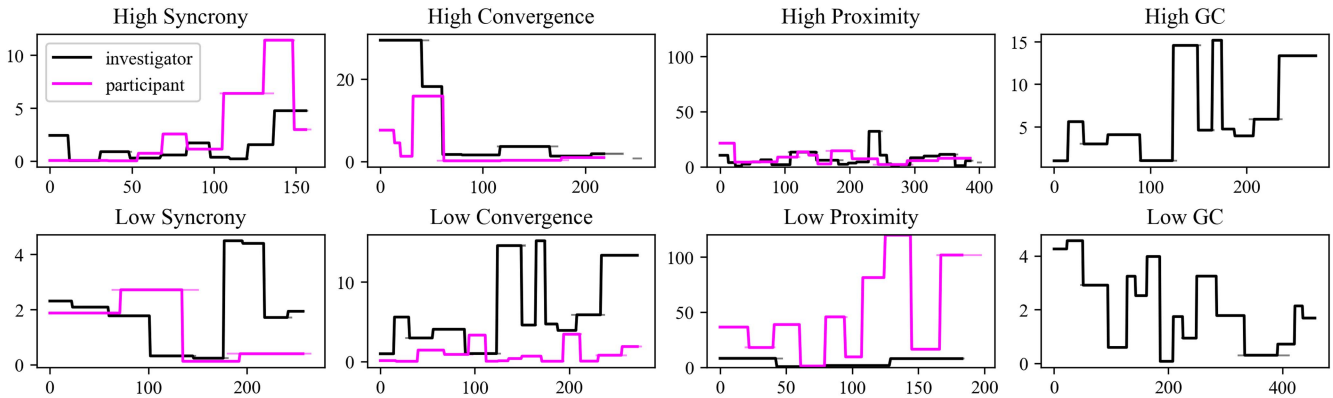


Fig. 2. Illustration of high and low values of conversation-level features.

## IV. Methods

Our goal in this paper is to characterize the communication traits of ASD by measuring the vowel space characteristics (VSCs) of ASD's speech production and the social traits of ASD by measuring the interaction of vowel space characteristics between ASD participants and investigators during dialogs. These measures of VSCs are calculated at two levels of granularities: utterance-level VSCs and conversation-level VSCs. Utterance-level VSCs are acoustic values extracted from the participant's spoken sentences or utterances. In contrast, conversation-level VSCs focus on the interaction between two talkers (an investigator and a participant) in their conversation. The utterance-level VSCs were operationally calculated from the corner vowels collected throughout the session, which are used to characterize ASD communication traits. The conversation-level VSCs features were operationally derived by calculating on two time series from the investigator and participant (as demonstrated in Fig. 2), which characterize the social traits.

For utterance-level VSCs, we derive three feature sets measuring vowel intelligibility, vowel variability, and articulator flexibility, respectively. Vowel intelligibility refers to the clarity or articulation of vowels; vowel variability represents the flexibility or adaptability of a person in articulating vowels, and articulator flexibility refers to the coordination or synchrony between articulators (like tongue and jaw). The vowel intelligibility is characterized as inter-vowel dispersion, while vowel variability is represented as intra-vowel dispersion. These two common feature sets were used in previous research that measured vowel intelligibility with VSA [32] and assessed vowel variability with inter-vowel dispersion [33].

Lastly, articulator flexibility is another important measure. Since speech production is a process that involves the complex coordination of articulators such as the tongue, jaw, and larynx [34], [35], flexible controlling of the articulators is one of the keys to smooth speech production. According to past research, the coordination of fine motor control is atypical in people with ASD [24], [36]. However, there needs to be computational methods for quantifying this coordination perspective of articulatory characteristics. In this work, we propose to compute the dependency between the first and second formant on vowel space to characterize the coordination of articulators.

For conversation-level VSCs, we aim to measure the interaction pattern, so we adopt commonly used interaction metrics [37], [38]: proximity, synchrony, convergence, gradual change of the investigator, and gradual change of participant to represent the interactions of VSCs between the investigators and the participants.

### A. Preprocessing

Since these features are computed in spontaneous conversations (i.e., ADOS sessions), preprocessing is needed to extract robust phone-level estimations. The preprocessing pipeline includes forced alignment, first and second formant extraction, outlier filtering, and speaker normalization. We then compute utterance-level and conversation-level VSC features afterward. The forced alignment is done by our self-trained Taiwan Mandarin force aligner, and the formants are extracted using praat [39]. The details are described below.[2]

*1) The Taiwan Mandarin Force Aligner:* The aligner contains an acoustic model and a language model. The acoustic model is a factorized time-delayed neural network (TDNN-f) [40]. It was first pre-trained on a combination of several Taiwanese Mandarin corpus containing collected from radio broadcast programs [41]. The total duration of the ASR pre-training data is 172 hours. The Mandarin aligner was then fine-tuned on a subset of our ADOS. Given the paired word-level transcript (transcript that is composed of words instead of phones) and audio segments in our ADOS database, we can retrieve the timestamps of the corner phone boundaries with this Taiwan Mandarin aligner. The acoustic values within these phone boundaries are what we are interested.

*2) Formant Extraction:* We extract the first and second formant (F1, F2) values of the corner vowels /a/, /u/, /i/. The formant values are estimated with linear predictive coding (LPC) algorithm using praat. We averaged over the middle position value at 49% to 51% of each vowel's duration. This averaged F1 F2 value represents a data sample on vowel space.

To prevent noisy estimates, the data samples on vowel space underwent subsequent procedures to remove the potential outliers. First, we apply interquartile range (IQR) filtering to each person's data samples. The data samples that fall outside 1.5 times the IQR from the center were removed. After IQR filtering, we apply another outlier filtering by kernel density estimation (KDE) filtering. The procedure segments the F1 F2 space into 100 x 100 linear grids, assigns each gridpoint a density score, and removes the data samples within lower 40% density regions. During KDE filtering, the density score is estimated by a two-dimensional kernel, which fits the data samples on vowel space using density estimation.

To reduce the variation caused by factors irrelevant to autistic traits, we utilize a speaker normalization proposed by Lobanov [42] and tested by flynn [43]. The normalization process is as below,

$$F_i^N = \frac{(F_i - \mu_i)}{\sigma_i} \tag{1}$$

where we subtract each speaker's formant values (F1, F2) from their formant mean and divide the formant values by their formant variances.

The preprocessing step generates robust data samples used to derive the utterance and conversation level VSC features. Note that the Taiwan Mandarin force aligner's accuracy achieve 90.2% (recall) of corner phones with 30 ms phone boundary tolerance. The average formant values are computed, in which /u/ (F1:509.3 $\pm$ 123.6 F2:876.9 $\pm$ 190.1), /i/ (F1:561.9 $\pm$ 249.4 F2:1993.4 $\pm$ 435.1), /a/ (F1:844.9 $\pm$ 262.7); these numbers are similar as previously reported [44].

### B. Utterance-Level VSC Features

The use of acoustic measurements to infer articulatory status has been studied in clinical research [45]. Firstly, the expansion area of the corner vowel can measure articulatory range to assess the severity of speech disorder [32], [46]. Secondly, articulation depends not only on the functions of the articulators (tongue, jaw, larynx) but also on their coordination. We measure the formant-based vowel space characteristics from three aspects: intra-vowel dispersion, inter-vowel dispersion, and formant dependency. Features derived from these aspects are operationally extracted by gathering all phone samples and then processing them with proposed algorithms. The following paragraphs describe the details of this process.

*1) Intra-Vowel Dispersion (Intra-VD) and Inter-Vowel Dispersion (Inter-VD) Features:* Researchers have used indices of VSCs to predict autistic traits. For example, vowel space area (VSA) and intra-vowel dispersion indices were used to measure

[2]We provide our code for deriving the VSC features on our gitlab website https://biicgitlab.ee.nthu.edu.tw/jack/tbme2021.git. The forced alignment step in preprocessing can be executed using any alternative ASR model.

the vowel intelligibility and stability of ASD participants [20], [21]. The VSA calculates the triangle area expanded by the mean of each corner vowel, and the intra-vowel dispersion proposed by Kissine calculates the euclidean distance of each phone sample to it's corresponding phone center. Both of them capture the mean positional relationship of corner vowels in the F1 F2 formant space. Given our articulation's natural variability, we develop a distributional approach by deriving the intra-vowel dispersion and inter-vowel dispersion feature sets. The intra-vowel dispersion and inter-vowel dispersion are represented by four scatter matrices: the within-class covariance matrix ($S_W$), the between-class covariance matrix($S_B$), the total covariance matrix ($S_T$), and the ratio of $S_B$ to $S_W$ ($S_W^{-1}S_B$). The first three matrices are derived by:

$$S_W = \frac{1}{N}\sum_{i=1}^{C}\sum_{j=1}^{N_i}(x_{ij}-\mu_i)(x_{ij}-\mu_i)^T \tag{2}$$

$$S_B = \frac{1}{N}\sum_{i=1}^{C}N_i(\mu_i-\mu)(\mu_i-\mu)^T \tag{3}$$

$$S_T = \frac{1}{N}\sum_{i=1}^{C}N_i(x_{ij}-\mu)(x_{ij}-\mu)^T \tag{4}$$

where, C and $N_i$ are the number of categories and the total number of samples in the phone category $i$, respectively. N is the total number of samples, and $x_{ij}$ are samples of 2-dimensional vector (F1 and F2). $\mu_i$ is the mean of samples in the phone category $i$, and $\mu$ is the overall mean. These matrices represent generalized variance [47] of these corner phones on F1 F2 vowel space. The within-class covariance matrix measures the intra-vowel dispersion, which reflects vowel stability. The between-class covariance matrix, total covariance matrix, and the ratio of $S_B$ to $S_W$ ($S_W^{-1}S_B$) measure the inter vowel dispersion, reflecting vowel intelligibility but with three different types of measures.

Additionally, all the matrices processed by a determinant or trace that converts each matrix to a single value. The higher determinant or trace of $S_W$, denoted as within-class covariance or variance (WCC or WCV), implies lower vowel stability. The determinant or trace of $S_B$ is denoted as between-class covariance or variance (BCC or BCV), and that of $S_T$ is denoted as total covariance or variance (TC or TV). The higher of these values implies higher vowel intelligibility.

$S_W^{-1}S_B$, a positive semi-definite matrix, represents the ratio of within-class covariance to between-class covariance. First, we calculate the determinant and trace on $S_W^{-1}S_B$, and denote them Det($W^{-1}B$) and Tr($W^{-1}B$). We also calculate four common estimates of within-between class covariance ratio: Pillai's trace, Hotelling-Lawley's trace, Wilk's lambda, and Roy's largest root. The four estimates are derived by: Pillai $=\sum_{i=1}^{q}\frac{\lambda_i}{1+\lambda_i}$, Hotelling $=\sum_{i=1}^{q}\lambda_i$, Wilks $=\prod_{i=1}^{q}\frac{1}{1+\lambda_i}$, Roys $=\max(\lambda_i)$ in which $\lambda_i$ denotes the eigenvalue of $S_W^{-1}S_B$. The higher Pillai, Hotelling, and Roys values represent higher vowel discrimination. On the contrary, the higher index value of Wilks the lower the vowel discrimination.

*2) Formant Dependency (FD) Features:* The intuition of this measurement is to infer the articulatory coordination with acoustic measurements. We calculate four correlation coefficients: Pearson's correlation coefficient, spearman's correlation coefficient, Kendall's tau correlation coefficient, and distance correlation coefficient. The four correlation coefficients are calculated on variables: F1 and F2 from all the retrieved corner vowels. Finally, the FD feature set has 4 features: PearF1F2, SpearF1F2, KendallF1F2, and DCorrF1F2.

In brief, there are totally three feature sets Inter-VD. Intra-VD, and FD composing of 16 features (summarized in Table II). These features, pre-processed and gathered through a session, represent one's vowel space characteristics.

### C. Conversation-Level VSC Features

Measuring interaction between two talkers in dyadic interaction can be operationally designed as computing interrelationship between two time-series of acoustic features. A classic example is measuring the proximity, synchrony, and convergence of acoustic/prosodic between the two feature series to characterize speech entrainment [37]. We follow similar approaches to characterize the dynamic interplay between the investigators and the participants but with VSC features. Furthermore, gradual change of an person's behaviors over time is also important indices to measure interaction as it has been shown to reflect certain types of affectionate behaviors [38]. Motivated by these studies, we derive conversation-level VSC feature sets involving three steps: extracting features from minimum phone units (MPUs), estimating the speaker's gradual change (GC) series, and calculating the phonetic conversation-level VSC features. The purpose is to derive a temporal progression of VSC features from both interlocutors, so that we can calculate metrics to represent the VSCs' relationship from them. To collect enough phones in all temporal segments, we define MPU and calculate VSC features within each of them. The following describes this process.

*1) Extracting VSC Features From MPUs:* To derive a temporal progression, we segment each conversation session into several units and calculate the progression of VSC features. We define the MPU as time intervals, segments of a conversation session that contain at least $N_{mpu}$ amounts of each corner vowel (we set $N_{mpu}$=2 in this study, so there will be at least six vowels contained in an MPU). Then, we calculate inter- and intra-vowel dispersion and vowel formant dependency within each minimum phone unit (MPU). In brief, the VSC features extracted from vowels in each MPU comprise a series of VSC features. The progression and their inter-dependency of the series from two interlocutors is what we aim to estimate.

*2) Estimating the Speakers' Gradual Change (GC) Series of VSCs:* Prior studies derive relationship metrics by first calculating acoustic-prosodic feautres in inter-pausal unit (IPU), and then interpolate the values between the IPUs [37]. We use a similar approach, but define MPU as the basic unit instead. The interpolated time series are denoted as gradual change (GC) series, in which we use k-nearest neighbors (KNN) regression to interpolate values between each MPU. In each

TABLE II
SUMMARY OF THE FEATURES USED IN THE EXPERIMENTS

| Feature sets | | | |
|---|---|---|---|
| Utterance-level Features | VSC features | Inter Vowel Dispersion (Inter-VD) | BCC, BCV, TC, TV, Wilks, Pillai, Hotel, Roys, Det($W^{-1}B$), Tr($W^{-1}B$) |
| | | Intra vowel dispersion (Intra-VD) | WCC, WCV |
| | | Formant dependency (FD) | PearF1F2, SpearF1F2, KendallF1F2, DCorrF1F2 |
| | Acoustic-prosodic features | F0, Intensity, HNR, Jitter, Shimmer | Mean($\overline{int}$),Mean($\overline{F0}$),Mean($\rho$(F0)),Mean(HNR),Mean(Jitter),Mean(Shimmer), Max($\overline{int}$),Max($\overline{F0}$),Max($\rho$(F0)),Max(HNR),Max(Jitter),Max(Shimmer), Std($\overline{int}$),Std($\overline{F0}$),Std($\rho$(F0)),Std(HNR),Std(Jitter),Std(Shimmer) |
| Conversation-level Features | Conversation[P] | Proximity[P] | Proximity of features in feature set: acoustic-prosodic features |
| | | Convergence[P] | Convergence of features in feature set: acoustic-prosodic features |
| | | Syncrony[P] | Syncrony of features in feature set: acoustic-prosodic features |
| | | GC[P]$_{\text{inv}}$ | Investigator's GC of features in feature set: acoustic-prosodic features |
| | | GC[P]$_{\text{part}}$ | Participant's GC of features in feature set: acoustic-prosodic features |
| | Conversation[VSC] | Proximity[VSC] | Proximity of features in feature set: VSC |
| | | Convergence[VSC] | Convergence of features in feature set: VSC |
| | | Syncrony[VSC] | Syncrony of features in feature set: VSC |
| | | GC[VSC]$_{\text{inv}}$ | Investigator's GC of features in feature set: VSC |
| | | GC[VSC]$_{\text{part}}$ | Participant's GC of features in feature set: VSC |

conversation, we estimate the investigator and the participant's GC series, denoting them as $f^{inv}$ and $f^{part}$. Firstly, to prevent the potential outliers before fitting KNN, all phone instances that are three standard deviations away from the mean value in each MPUs are dropped. Secondly, $f^{inv}$ is defined within the interval $[t^{inv}_{\min}, t^{inv}_{\max}]$, where $t^{inv}_{\min}$ and $t^{inv}_{\max}$ represent the initiation and end of investigator's MPUs (same rule is applied to $f^{part}$). Thirdly, the common support: $t^- = \max(t^{inv}_{\min}, t^{part}_{\min})$ $t^+ = \min(t^{inv}_{\max}, t^{part}_{\max})$ are defined to denote the time interval which the two talkers have overlap.

*3) Calculating Conversation-Level VSC Features From the GC Series:* We calculated five types of relationship metrics to derive conversation-level VSC features: 1) Proximity, 2) Convergence, 3) Synchrony, 4) $\text{GC}_{\text{inv}}$ 5) $\text{GC}_{\text{part}}$. The details are in the following.

1) Proximity: the proximity feature represents the closeness of interlocutors across the whole conversation. It is calculated by the mean distance of acoustic features between the two GC series with a negative sign in the front, that is:$-\frac{1}{N}\sum_{t=t^-}^{t^+}|f^{inv}[t] - f^{part}[t]|$. 5

$$-\frac{1}{N}\sum_{t=t^-}^{t^+}\left|f^{inv}[t] - f^{part}[t]\right| \tag{5}$$

Here, N denotes the number of points interpolated by the KNN regressor within the interval $[t^-, t^+]$.

2) Convergence: the convergence feature measures whether the two interlocutors are becoming closer across the whole conversation. It is calculated by the Pearson correlation coefficient between $-|f^{inv}[t] - f^{part}[t]|$ and t:

$$\frac{\sum_{t=t^-}^{t^+}(d[t]-\overline{d})\cdot(t-\overline{t})}{\sqrt{\sum_{t=t^-}^{t^+}(d[t]-\overline{d})^2\cdot\sum_{t=t^-}^{t^+}(t-\overline{t})^2}} \tag{6}$$

where $d[t]{=}-|f^{inv}[t] - f^{part}[t]|$, and $\overline{d} = \frac{1}{N}\sum_{t=t^-}^{t^+} d[t]$.

3) Synchrony: synchrony features measure the leader-follower relationship of two talkers. This is calculated by calculating correlation between $f^{inv}[t+\delta]$ and $f^{inv}[t]$, in which $\delta$ denote a lagging parameter:

$$\frac{\sum_{t=t^-}^{t^+}(f^{inv}[t+\delta]-\overline{f^{inv}})\cdot(f^{part}[t]-\overline{f^{part}})}{\sqrt{\sum_{t=t^-}^{t^+}(f^{inv}[t+\delta]-\overline{f^{inv}})^2\cdot\sum_{t=t^-}^{t^+}(f^{part}[t]-\overline{f^{part}})^2}} \tag{7}$$

we iterate $\delta$ over a range of values: [−15, −10, −5, 0, 5, 10, 15] (the unit of lagging parameter $\delta$ is expressed in seconds). The final synchrony value is determined by the selected $\delta$ where the absolute value of synchrony $|\text{Synchrony}(f^{inv}, f^{part})|$ has the largest value. Positive $\text{Synchrony}(f^{inv}, f^{part})$ indicates one of the two talkers is leading, and the other is following. On the contrary, it means $f^{inv}$ and $f^{part}$ evolve in the opposite direction.

4) GC: The derivation of GC features is similar to that of convergence features. Instead of calculating the distance between the two talkers, we calculate the investigator's or participant's gradual change of VSC features during the conversation, that is:

$$\frac{\sum_{t=t^-}^{t^+}(f^{R}[t]-\overline{f^{R}})\cdot(t-\overline{t})}{\sqrt{\sum_{t=t^-}^{t^+}(f^{R}[t]-\overline{f^{R}})^2\cdot\sum_{t=t^-}^{t^+}(t-\overline{t})^2}} \tag{8}$$

where R $\in$ {inv, part}. Then, the GC features of the investigator and the participant are denoted as $\text{GC}_{\text{inv}}$ and $\text{GC}_{\text{part}}$, respectively.

In short, we characterize the VSCs interaction status between the investigators and the participants, which can be derived by computing dependency between two VSCs series. We use five functions: Proximity, Convergence, Synchrony, $\text{GC}_{\text{inv}}$, $\text{GC}_{\text{part}}$ to describe the dynamic interplay of these two series. Furthermore, higher values of Proximity, Convergence, Synchrony imply more closeness of the two series. Additionally, $\text{GC}_{\text{inv}}$ and $\text{GC}_{\text{part}}$ represent the global VSCs trend of the investigator and the participant. Positive/negative of these values imply a rising/decreasing trend. Fig. 2 illustrates these features. Since these features are derived at a conversation-level granularity, we denote them as conversation-level features. At last, Fig. 1 summarizes the whole process of deriving utterance- and conversation- level features.

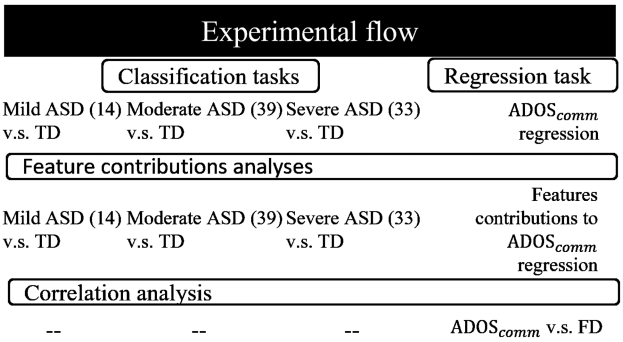


Fig. 3. Experimental flow of this study.

## V. Experiments

This work aims to comprehensively study vowel space characteristics by observing differences in TD's speech and their relationships with ASD-symptom severity. Autism spectrum disorder is known as a heterogeneous disorder, and participants demonstrate a wide variety of functioning skills. Some have less proficient language abilities, making it difficult to collect speech from their sessions. By contrast, some ASD participants show no impairments in communication but show autistic traits on social dimensions. To investigate the differences of VSC for coherent groups of ASD participants as compared to TD, we split our ASD cohort based on ADOS Calibrated Severity Scores (CSS [48])-determined symptom severity into severe, moderate, and mild ASD subgroups. Then, the three subgroups were compared with the TD group in a binary classification task. The ADOS assessment can assist in judging the severity of autistic traits. A higher score on the assessment code in ADOS means that patients have more severe ASD symptoms indicating serious socio-communicative impairments and vice versa. Hence, aside from ASD detection, we also perform a regression task to predict ADOS scores.

In short, several binary classification tasks and a regression task were conducted in this study. The former differentiates coherent cohorts of ASD from TD, and the latter task regresses the ADOS communication score.

### A. Definition of Experimental Parameters

Fig. 3 demonstrates the flow of our experiments, in which two main tasks are followed by analyses. First, classification tasks contain three subtasks: mild ASD vs. TD, moderate ASD vs. TD, and severe ASD vs. TD. Then, regression task contains one task: the ADOS communication code (denoted as $\mathrm{ADOS}_{\mathrm{comm}}$) regression task. For each task, we trained a SVM model with feature sets: VSC features, Conversation[VSC], and Conversation[P]. (Please refer to the Table II) The VSC and Conversation[VSC] feature sets are defined in Section IV. Additionally, we computed the conversation-level acoustic-prosody features according to past studies [27], [37], [49] but with a slight modification; implementation details are in section V-C. In both classification and regression tasks, we present the results of each single feature set (Single feature set prediction) and of the feature fusion (Feature sets fusion prediction). When performing feature fusion, we compared the **best-performing model**—model of highest score including VSC features and the **baseline model**—model of highest score excluding VSC features.

The model parameters and evaluation metrics are defined as follows. In each binary classification subtask, the parameter C of SVC model was tuned within the set: $S = \{0.001, 0.01, 0.1, 1, 5, 10.0, 25, 50, 75, 100\}$, and the classification results are evaluated using unweighted average recall (UAR) and F1-score. In the regression task, the parameter $\epsilon$ was tuned within the set: $\epsilon \in S$. The evaluation metrics of this task include Mean absolute error (MAE), Pearson's correlation coefficient (pear), Spearman's correlation coefficient (spear), and the concordance correlation coefficient (CCC). Lastly, nested cross-validation was implemented in both experiments.

### B. Model Explanation Through SHAPley Analysis

SHAPley analysis [50], an explainable AI approach, can evaluate the contributing factors of each feature in a well-trained model, with each factor explaining the prediction results of each participant. A given SHAPley model can provide measures of contributions from each feature to each testing instance. These measures of contributions denoted as SHAPley values, are used to understand how each feature contributes to the model's final decision. In this study, KernelSHAP was selected as our explanation model.

*1) Explanation of Classification Results:* Given an initial feature combination used to train a classification model, denote as an initial model: **I**, and final model, denoted as **F**, representing the same configuration of classification model trained on another feature combination; for example, a feature combination that contains initial feature sets with some additional feature sets. The change from the initial to the final model is denoted as **I** $\rightarrow$ **F**. There are several key points worth observing. First, the classification decision on certain samples may change. Those initially correct but finally incorrect predictions are denoted as X $\rightarrow$ O. Conversely, the opposite situations are denoted as O $\rightarrow$ X. Second, all of those decisions are determined from decision values (mostly are distances of the sample to the decision boundary). For instance, in the row, Feat set A $\rightarrow$ Feat set $A^+$ in the Table III, participant 21 has a decision value of $-0.55$ and turned to 0.27. A participant with a decision value larger than 0 indicates that he is classified (by prediction model) as TD, and he/she will be classified as ASD if his/her decision value is less than 0. Third, the decision scores can be attributed to the SHAPley values of the feature set combinations and base values from the models. Since each feature has a SHAPley value, we compute the sum of the features' SHAPley values from a feature set to represent that feature set. Then, the SHAPley value of a feature set represents the feature set's contribution to the outcome decision score, and the base value of the model represents the average of the model's decision scores from the training data. These SHAPley values and a base value compose the model's decision value [50]. Hence, a SHAPley value having the same sign with the final

TABLE III
MOCK UP EXAMPLES FOR EXPLANATIONS OF CLASSIFICATION AND REGRESSION TASKS

| Analysis of classification task | | |
|---|---|---|
| | **I → F** | $\text{SHAPley}_{\text{F}}$ |
| Feat set A → Feat set $A^{+}$ | 21: X→O<br>ASD:-0.55→TD:0.27 | **BASEval**:0.2+0.039<br>(k):0.242,(d):0.053<br>(e):-0.139,(c):-0.124 |
| Feat set A → Feat set $A^{-}$ | 21: O→X<br>TD: 0.27→ASD: -0.22 | **BASEval**:0.25-0.011<br>(d):0.024,(e):-0.258<br>(c):-0.233 |

| Analysis of regression task | | | | | |
|---|---|---|---|---|---|
| | | Q1 | Q2 | Q3 | Q4 |
| | N= | 36 | 37 | 6 | 7 |
| BASEval | **F+I** | 6.96 | 6.93 | 6.93 | 6.99 |
| | Δ | -0.05 | -0.05 | -0.04 | -0.04 |
| Inter-VD | **F+I** | -0.01 | -0.03 | -0.61 | 1.06 |
| | Δ | -0.02 | 0.06 | 0.14 | -0.44 |
| FD | **F+I** | -0.12 | 0.10 | -0.18 | 0.37 |
| | Δ | -0.01 | 0.00 | -0.02 | -0.18 |
| Y | | 1.81 | 5.11 | 3.17 | 4.00 |
| Δ MAE | | -0.08 | 0.11 | -0.19 | -0.42 |

N denote number of samples at that group.

decision value means they contribute to the prediction result. Additionally, we present the base value of the final model as that of the initial model plus the difference of base values between the final model and initial model ($\text{BASEval}_{\text{F}} = \text{BASEval}_{\text{I}} + \Delta$ BASEval, where $\Delta$ BASEval = $\text{BASEval}_{\text{F}}$ - $\text{BASEval}_{\text{I}}$). Both feature sets' SHAPley values and the model's base value indicate how the difference between the initial and final model relates to the final prediction outcome.

Take Participant 21 in Table III as an example. Comparing the difference between the initial model—feature set A, and the enhanced initial model—feature set $A^{+}$, the classification result turned from wrong to correct. This participant was initially classified as ASD with a decision value of $-0.55$ and finally classified as TD with a decision value of 0.27. Then, refer to the column: $\text{SHAPley}_{\text{F}}$, the SHAPley values of the model: **F** shows that feature sets (k), (d) are in line with the final decision (both the decision value and SHAPley values are positive), and the feature sets (e), (c) are inconsistent with the final decision value. Additionally, $\Delta$ BASEval (0.039) implies that using the final feature combination increases the model's tendency to predict TD, so Participant 21 was corrected. Similarly, the same method can also explain the changes from the initial to the final models with a reduced feature set compared to the initial model. The purpose is to understand the changes when one or several key feature sets are absent. By referring to the row: Feat set A → Feat set $A^{-}$ in the Table III. Participant 21, in this case, was misclassified as ASD with a decision value of $-0.22$. Feature sets (e) and (c) have the SHAPley values $-0.258$ and $-0.233$, which are the primary factor causing the misclassification. Besides, the difference in the base value ($\Delta$ BASEval = $-0.011$) also attributes to the model's final prediction result. In a nutshell, the impact of including or excluding certain feature sets on the prediction results is what we are interested in. This impact can be observed by adding or removing this feature set when training a classification model. Using decision scores and SHAPley values illustrates the changes from the initial to final models.

*2) Explanation of Regression Results:* The explanation of the Regression result is slightly different from that of the classification. The prediction errors will be measured using Mean Absolute Error (MAE), calculated from the difference between the predicted values from the models and the ground truth. Furthermore, since the predicted values of the models can be expressed by the composition of SHAPley values, the prediction error can be quantified using SHAPley values. Equation (9) to (12) are derived from the relationship between the change in prediction error and the SHAPley values of each feature set. To better discuss the attribution of the prediction errors, the participants were divided into four groups based on their actual and predicted score of $\text{ADOS}_{\text{comm}}$. The groups are defined as follows,

1) Q1: $\hat{Y}_q^F - Y_q > 0, \hat{Y}_q^I - Y_q > 0$
2) Q2: $\hat{Y}_q^F - Y_q < 0, \hat{Y}_q^I - Y_q < 0$
3) Q3: $\hat{Y}_q^F - Y_q > 0, \hat{Y}_q^I - Y_q < 0$
4) Q4: $\hat{Y}_q^F - Y_q < 0, \hat{Y}_q^I - Y_q > 0$

where $\hat{Y}_q^F$ $\hat{Y}_q^I$ represent the predicted $\text{ADOS}_{\text{comm}}$ of the participants by the final model followed by that of the initial model. Additionally, $Y_q$ represents the actual $\text{ADOS}_{\text{comm}}$ of the participants q.

Given an initial and final model, each model has one base value and several SHAPley values from each feature set for each sample. According to the equations (9)– (12), there are two types of values we are interested in: the difference between the initial and final model and the sum of the values from the initial and final model.

For example, in Equations (9) and (10) the differences of base values ($\Delta$ BASEval), and differences of SHAPley values ($\Delta$ $\text{Feat}_{\text{i}}$) are used. In equations (11) and (12), $\text{BASEval}^{\text{F}}+\text{BASEval}^{\text{I}}$ and $\text{Feat}_{\text{i}}{}^{\text{F}}+\text{Feat}_{\text{i}}{}^{\text{I}}$ are used. Hence we present the difference and summation, denoted as $\Delta$ and I+F, just as in Table III. A mockup example is shown in the second part of Table III. The values under the title 'Analysis of regression task' represent each group's mean of SHAPley value. For example, the value 6.96 at column Q1 and the BASEval—**F+I** shows the mean of base values over Q1. First, according to (9), the mean absolute error of the Q1 group can be obtained by the deltas of base values and feature sets. hence, $\overline{\Delta\text{MAE}_{\text{Q1}}} = -0.05 + -0.02 + -0.01 = -0.08$. The calculation on group Q2 is similar to group Q1, but with an additional negative sign. As for Q3, the mean absolute error of this group can be obtained by the summation of values from the initial and final model (I+F). Therefore, $\overline{\Delta\text{MAE}_{\text{Q1}}} = -0.19 = 6.93 + -0.61 + -0.18 - 2 \times 3.17$ (the calculation of group Q4 is similar, but with an additional negative sign). Finally, the average prediction error is the weighted sum of the mean absolute error from each group divided by the total number: $\overline{\Delta\text{MAE}} = -0.08 \times 36 + -0.11 \times 37 + -0.19 \times 6 + -0.42 \times 7 \,/\, 86 = -0.0078$, representing that averagely the model will reduce an error of 0.0078 when assessing a new person.

TABLE IV
THIS TABLE SHOWS THE FIRST EXPERIMENT'S RESULTS: DISCRIMINATION BETWEEN ASD/TD

| Feature sets | Feature subsets | | Mild | | Moderate | | Severe | |
|---|---|---|---|---|---|---|---|---|
| | | | UAR | F1 | UAR | F1 | UAR | F1 |
| Conversation[P] | (a) | Proximity[P] | 0.675 | 0.678 | **0.810** | **0.799** | **0.844** | **0.826** |
| | (b) | Convergence[P] | 0.661 | 0.662 | 0.460 | 0.429 | 0.485 | 0.376 |
| | (c) | Syncrony[P] | 0.732 | 0.729 | 0.597 | 0.598 | 0.529 | 0.524 |
| | (d) | $GC[P]_{inv}$ | **0.782** | **0.785** | 0.635 | 0.636 | 0.634 | 0.637 |
| | (e) | $GC[P]_{part}$ | 0.746 | 0.751 | 0.509 | 0.508 | 0.379 | 0.321 |
| Conversation[VSC] | (f) | Proximity[VSC] | 0.518 | 0.507 | 0.461 | 0.408 | 0.593 | 0.589 |
| | (g) | Convergence[VSC] | 0.539 | 0.539 | 0.610 | 0.614 | 0.464 | 0.450 |
| | (h) | Syncrony[VSC] | 0.489 | 0.489 | 0.586 | 0.584 | 0.559 | 0.552 |
| | (i) | $GC[VSC]_{inv}$ | 0.561 | 0.555 | 0.673 | 0.680 | 0.634 | 0.637 |
| | (j) | $GC[VSC]_{part}$ | 0.550 | 0.549 | 0.460 | 0.443 | 0.494 | 0.476 |
| Vowel space characteristics (VSCs) | (k) | Inter Vowel Dispersion (Inter-VD) | 0.518 | 0.507 | 0.481 | 0.473 | 0.533 | 0.533 |
| | (l) | Intra vowel dispersion (Intra-VD) | 0.504 | 0.502 | 0.497 | 0.492 | 0.704 | 0.702 |
| | (m) | formant dependency (FD) | 0.539 | 0.539 | 0.472 | 0.453 | 0.449 | 0.388 |
| Feature fusion | (n) | Inter-VD+$GC[VSC]_{inv}$+Convergence[VSC]+ Syncrony[VSC]+$GC[P]_{part}$+Proximity[P]+Convergence[P] | **0.904** | **0.908** | - | - | - | - |
| | (o) | $GC[P]_{part}$+Proximity[P]+Convergence[P] | 0.618 | 0.598 | - | - | - | - |
| | (p) | $GC[P]_{part}$+Syncrony[P] | 0.843 | 0.846 | - | - | - | - |
| | (q) | FD+$GC[VSC]_{inv}$+Proximity[P] | - | - | **0.899** | **0.904** | - | - |
| | (r) | Proximity[P] | - | - | 0.81 | 0.799 | 0.844 | 0.826 |
| | (s) | FD+$GC[P]_{inv}$+Proximity[P] | - | - | - | - | **0.845** | **0.855** |
| | (t) | $GC[P]_{inv}$+Proximity[P] | - | - | - | - | 0.759 | 0.759 |

The rows represent the feature sets used in training the Support Vector Machine (SVM) classifier. The columns represent the tasks mild ASD vs TD, moderate ASD vs TD, and severe ASD vs TD, respectively.

### C. Additional Features for This Study

Instead of extracting the acoustic-prosodic features from inter-pausal units as in [37], [51], we extracted features from minimum phone units (MPUs) defined in section IV-C. we derive the conversation-level acoustic-prosodic features similar to VSC features in the following procedures. For each corner vowel, we calculate six acoustic-prosodic features: the mean of intensity (Mean($\overline{int}$)), the mean of F0 (Mean($\overline{F0}$)), the standard deviation of F0 ($\rho$(F0)), Harmonic to Noice Ratio (HNR), Jitter, and Shimmer. Then within each MPU, we calculate three statistical functionals: mean, max, and standard deviation. The same statistical functionals were used in previous studies [37], [51]. Finally, we estimate the GC series and calculate proximity, convergence, synchrony, and GC functions, the same as the conversation-level VSC features.

### D. Experiment1: Classification of ASD/TD

In this section, we compare the ASD/TD classification results under the conditions: with and without VSC features. We observed that including VSC features gains a higher accuracy than not including them in all three binary classification subtasks. Table IV summarizes all the classification results.

*1) Comparison Between Single Feature Sets:* In the Mild ASD vs. TD subtask, single feature sets Synchrony[P] (UAR: 0.732, F1: 0.729), $GC[P]_{\mathrm{inv}}$ (UAR: 0.782, F1: 0.785), $GC[P]_{\mathrm{part}}$ (UAR: 0.746, F1: 0.751) achieve competitive accuracy (of above 0.70), and is the highest among all the feature sets in Table IV. Furthermore, feature set Proximity[P] has the highest performance in both Moderate and Severe ASD vs. TD subtask (UAR: 0.810, F1: 0.799, and UAR: 0.844, F1: 0.826, respectively). Besides, training classifiers on acoustic-prosodic features can achieve high UAR scores (over 0.75) when classifying ASD and TD, which aligns well with past research [27]. Interestingly, the best-performing feature sets in the Mild ASD vs. TD subtask (Synchrony[P], $GC[P]_{\mathrm{inv}}$, $GC[P]_{\mathrm{part}}$) is different from those in Moderate and Severe ASD vs. TD subtask (Proximity[P]). This observation indicates that the acoustic properties differentiating severe ASD from TD and those differentiating mild ASD from TD are different. This difference provides further empirical evidence in the distinctive nature of the autism spectrum.

*2) Comparison Between Feature Fusions With and Without VSC Features:* Among all feature combinations without the VSC features, $GC[P]_{\mathrm{part}}$+Synchrony[P] (UAR: 0.843, F1: 0.846) has the highest performance in the Mild ASD vs. TD subtask. In contrast, among feature combinations with VSC features, the best-performing combination is row (n): Inter-VD+$GC[VSC]_{\mathrm{inv}}$ + Convergence[VSC] + Syncrony[VSC]+$GC[P]_{\mathrm{part}}$ + Proximity[P] + Convergence[P] (UAR: 0.904, F1: 0.908), which is higher than $GC[P]_{\mathrm{part}}$+Synchrony[P]'s (6.1% in terms of UAR and 6.2% in terms of F1). Interestingly, Inter-VD, $GC[VSC]_{\mathrm{inv}}$, Convergence[VSC], Syncrony[VSC], $GC[P]_{\mathrm{part}}$, Proximity[P], Convergence[P] do not perform well by themselves, but including these feature complements the results. Furthermore, the accuracy decreases when these feature sets were removed from the combination: (n) (resulting in combination: (o)). Secondly, row (r): Proximity[P] has the highest performance of all the feature combinations without the VSC features in both the Moderate and the Severe ASD vs. TD subtasks. Additionally, the best-performing feature combination in the Moderate ASD vs. TD subtask is row (q): FD+$GC[VSC]_{\mathrm{inv}}$+Proximity[P], which is UAR: 8.9%, F1: 10.5% better than Proximity[P] alone. Furthermore, the best-performing feature combination in the Severe ASD vs. TD subtask is row (s): FD+$GC[VSC]_{\mathrm{inv}}$+Proximity[P], whose result is UAR: 0.1%, F1: 2.9% higher than Proximity[P] alone. Moreover, removing FD from FD+$GC[VSC]_{\mathrm{inv}}$+Proximity[P] leave only $GC[VSC]_{\mathrm{inv}}$+Proximity[P] (row: (t)), whose

TABLE V
ANALYSIS OF THE CLASSIFICATION TASK

| | | | Major reason | Participants | Example | |
|---|---|---|---|---|---|---|
| Mild ASD vs TD | Baseline → Best-perform: (p) → (n) | X → O | Inclusion of VSC features | **21**,12 | | |
| | | | Changes of base value | None | | |
| | | O → X | Inclusion of VSC features<br>Changes of base value | None | | |
| | | | | | **I → F** | $\text{SHAPley}_{\text{F}}$ |
| | w/ VSC → w/o VSC: (n) → (o) | X → O | Changes of base value | None | 21: X→O | **base val**:0.28+-0.08 |
| | | | Removal of VSC features | **14**,25 | ASD:-0.55→TD:0.11 | (k):0.262,(i):-0.189 (g):-0.039,(h):-0.082 (e):-0.029,(a):-0.006 (b):-0.005 |
| | | O → X | Changes of base value | **2**,3,6,8,9, 12 | 14: X→O | **base val**:0.16+0.13 |
| | | | Removal of VSC features | **4**,10,13,15 | TD: 0.22→ASD: -0.25 | (e):-0.046,(a):-0.404 (b):-0.100 |
| Moderate ASD vs TD | Baseline → Best-perform: (r) → (q) | X → O | Inclusion of VSC features | 2,24,30,31, 33,39,41,48 | 2: O→X | **base val**:0.2+0.16 |
| | | | Changes of base value | **28** | ASD: -0.34→TD: 0.28 | (e):-0.102,(a):0.113 (b):-0.089 |
| | | O → X | Inclusion of VSC features | None | 4: O→X | **base val**:0.2+0.01 |
| | | | Changes of base value | **23**,27,47 | ASD: -0.07→TD: 0.15 | (e):-0.133,(a):0.311 (b):-0.237 |
| Severe ASD vs TD | Baseline → Best-perform: (r) → (s) | X → O | Inclusion of VSC features | 13,19,23,24, 25 | 28: X→O | **base val**:-0.28+-0.1 |
| | | | Changes of base value | 35 | TD:0.29→ASD:-0.07 | (m):0.060,(i):-0.002 (a):0.255 |
| | | O → X | Inclusion of VSC features | **6**,28 | 23: O→X | **base val**:-0.26+-0.1 |
| | | | Changes of base value | 38 | ASD:-0.42→TD:0.21 | (m):0.084,(i):0.192 (a):0.294 |
| | | | Other reason | 45 | | |
| | w/ VSC → w/o VSC: (s) → (t) | X → O | Changes of base value<br>Removal of VSC features | None | 6: X→O | **base val**:-0.14+-0.1 |
| | | | Other reason | **45** | TD:0.12→ASD:-0.02 | (m):-0.015,(d):0.340 (a):-0.106 |
| | | O → X | Changes of base value | None | 45: O→X | **base val**:-0.24+-0.03 |
| | | | Removal of VSC features | 2,3,6,29,36, 44 | TD:0.62→ASD:-0.00 | (m):-0.076,(d):0.073 (a):0.278 |

This table demonstrates the changes from the baseline models (prosodic features only) to the best-performing models and the changes when the VSC features are removed from the best-performing models of each task. The participants whose prediction results had been changed is summarized in the column: Affected participants. The possible reasons that cause these prediction changes are provided in the column: Major reason. We also provide the an example for each possible reason in the column: Example.

prediction accuracy is even worse than only Proximity[P]. These results again demonstrate that regardless of which ASD symptom group we are distinguishing from TD, including VSC features can improve accuracy.

In short, there are two major observation from our results. First, the best prediction accuracy in each subtask is achieved by combining appropriate feature sets instead of combining all individual feature sets with high prediction scores. Second, although the VSC feature sets by themselves do not achieve the best prediction results, fusing VSC features achieves optimal performance in all three binary classification subtasks. More details are shown in our analyses below.

### *E. Analysis of the Classification Tasks*

The classification results have shown that considering VSC features into feature fusion can improve the classification score in all three tasks. We further investigate the improvement details and the influence of the VSC features. To investigate VSC features influences, we focus on two changes. First is the changes from the baseline models (prosodic features only) to the best-performing models of each task, and denote these analyses as baseline → best-perform. Second is the changes when the VSC features are removed from the best-performing models and denote these analyses as w/ VSC → w/o VSC. The changes are illustrated using model explanation techniques in Section V-B.

By inspecting the model explanation data, we are interested in several points. First is how many participants were corrected and misclassified due to the changes. Since the classification score of the best-performing model should be the highest, the corrected participants should be more than the misclassified ones. Second, we are also interested in whether the correction or misclassification is attributed to VSC features or other reasons. Table V shows the participants whose prediction results had changed when VSC features were added or removed and the main factor that caused these changes. We present key finding in the following and leave the complete results to the supplementary material. Last, all of the feature combinations and single feature sets shown in Table V, such as (m), (i) (a), are abbreviated as indicated in Table IV. Our observations from the analyses are described below:

*1) Analysis of the Mild ASD vs. TD Subtask:* As shown in row (p) → (n) in Table V, two participants (21, 12) were corrected by (n) from (p). The main reason is that the added VSC features contribute to the correction. Take participant 21, for example (please see the example on the right of the Table V),

Inter-VD (denoted as (k)) contributes to the correct classification because its SHAPley value (0.262) is the largest, which has the same sign as the decision value of the final model (0.11). If we exclude VSC features from the best-performing model (n), resulting in model (o), the prediction result of 12 people will change. Ten of them will be misclassified, and 2 of them will be corrected (please refer to row (n) $\rightarrow$ (o)). Among the people being misclassified, misclassification is mainly because of removing the VSC features. The others are mainly because of the change in base value. To be specific, referring to Participant 4's example, the feature Proximity[P] has the highest SHAPley value ((a):0.311). Herefore, the removal of VSC features, leaving only prosodic features, causes misclassification. On the other hand, the example of Participant 2 shows that his/her base value increased from 0.2 to 0.36 (0.2+0.16), which is higher than any of the feature sets' SHAPley values. Hence we consider the change in base value to be the main reason for misclassification when VSC features were removed from the best-performing model (n). However, there are 2 participants whose prediction results turned correct. One example is Participant 14, whose prediction result turned correct that the feature Proximity[P]'s SHAPley value ((a): $-0.404$) is lowest and in line with the decision value ($-0.25$).

*2) Analysis of the Moderate ASD vs. TD Subtask:* According to the row (r) $\rightarrow$ (q) in Table V, 9 participants (2, 24, 28, 30, 31, 33, 39, 41, 48) were originally misclassified by the initial model but corrected by the final model. Among these participants, except Participant 28, the other 8 participants were correctly classified due to the inclusion of GC[VSC]$_{\text{inv}}$ (denoted as (i)). As for Participant 28, the change in base value probably is the reason for him/her to be corrected. As shown in row (r) $\rightarrow$ (q), Participant 28's base value difference ($-0.28 + -0.1 = -0.38$) is most likely to be the reason the model classified this person as ASD ($-0.07$).

In addition, the inclusion of FD+GC[VSC]$_{\text{inv}}$ also induces some classification errors (23, 27, 47). Refer to Participant 23's example; the SHAPley values of FD ((m): 0.084) and GC[VSC]$_{\text{inv}}$ ((i): 0.192) are in line with the final decision value. Hence we consider adding FD+GC[VSC]$_{\text{inv}}$ had made this participant misclassified.

*3) Analysis of the Severe ASD vs. TD Subtask:* As shown in the row (r) $\rightarrow$ (s) in Table V, 6 participants (13, 19, 23, 24, 25, 35) were corrected by the final model (t) from the initial model (s), whereas there are 4 additional misclassified participants (6, 28, 38 and 45). The corrected participants were either corrected because of added VSC features or the change on base values. As for the misclassified people, participants 6 and 28 are mainly because of change in base values. Please refer to participant 6's example, his/her base value had decreased 0.1 that made the decision score lower than 0, hence this participant were incorrectly classified as ASD. In addition, participant 45 was also misclassified. Interestingly, the either the change in base value ($-0.03$) and his/her Shapley values of VSC feature sets ((m):$-0.076$, (d):0.073) seems not high, making us consider them not affective to the result. Furthermore, This's participant's original decision score was a relatively large value (0.62) when having only one feature set Proximity[P] (a). According to these observation, we think the SHAPley value of Proximity[P] in this task was large. However, after the inclusion of VSC features, the SHAPley value of this feature set decreased, indicating that the influence of Proximity[P] had become weaker when being fused with VSC features. Finally, removing VSC features from the best-performing model in this task (as shown in the row (s) $\rightarrow$ (t)) will cause 6 participants being misclassified, but one participant being corrected. Since the misclassified participants are more than corrected participants, the classification score had therefore decreased.

In short, this analysis comprehensively analyzed the differences between the baseline and best-performing models and the changes when we removed the VSC features from the best-performing models. By inspecting the changes in SHAPley values, which represent the contribution of the feature sets, and the changes in decision values, we found that including the VSC features corrects several samples misclassified by the classifiers trained with only conventional acoustic-prosodic features. We observed several possible reasons why the feature fusions with VSC features could improve the prediction scores. The first possibility is that the VSC features increase the best-performing models' decision scores by directly correcting some misclassified participants from the initial models. Another possibility is that the base value, the expected decision score after training on the training set, changes so that the model tends to make a better prediction. The other possibility is that including the VSC features may alter the contribution of each feature set in the models, and this causes some misclassified samples to be corrected (for example, Participant 45). This study further underscores the inherent heterogeneity of the ASD cohort, i.e., not all participants express the same atypicality in the prosodic space. In contrast, some ASD participants show in the articulatory VSC space instead.

### F. Experiment2: Regression of Communication Deficit Score

*1) Result of Experiment2:* As shown in Table VI, Inter-VD (MAE: 1.402, pear: 0.431, spear: 0.440, CCC: 0.263) and FD (MAE: 1.485, pear: 0.406, spear: 0.337, CCC: 0.245) by itself has the best prediction score among all the individual feature sets. The best-performing model trained on the feature combination Inter-VD+FD+GC[VSC]$_{\text{inv}}$+Syncrony[VSC] (MAE: 1.36, pear: 0.487, spear: 0.508, CCC:0.289) are composed of VSC features. The best-performing model outperforms the baseline model Convergence[P] (MAE: 1.530, pear: 0.273, spear: 0.279, CCC:0.156), representing the highest-scoring feature combination without VSC features. Furthermore, we observed that the best-performing model includes the feature sets GC[VSC]$_{\text{inv}}$ and Syncrony[VSC], which do not result in good regression performance. However, removing these two feature sets will degrade the prediction score. Our further analysis is in the Section V-F2 investigates how the feature sets GC[VSC]$_{\text{inv}}$ and Syncrony[VSC] improve the results.

In addition, our previous work develop a deep learning based method for automatic ADOS coding prediction [19]. In that work, we trained an attentional GRU network regressor on converse-level lexical and acoustic embeddings. Although the

TABLE VI
THIS TABLE SHOWS THE RESULTS OF THE SECOND EXPERIMENT: PREDICTION OF $ADOS_{COMM}$

| Feature sets | Feature subsets | MAE | Pear | Spear | CCC |
|---|---|---|---|---|---|
| Conversation[P] | Proximity[P] | 1.576 | 0.130 | 0.125 | 0.060 |
| | Convergence[P] | 1.530 | 0.273 | 0.279 | 0.156 |
| | Syncrony[P] | 1.615 | 0.080 | 0.073 | 0.037 |
| | GC[P]inv | 1.603 | -0.024 | -0.014 | -0.006 |
| | GC[P]part | 1.514 | 0.398 | 0.352 | 0.217 |
| Conversation[VSC] | Proximity[VSC] | 1.639 | 0.117 | 0.108 | 0.061 |
| | Convergence[VSC] | 1.637 | -0.027 | -0.040 | -0.009 |
| | Syncrony[VSC] | 1.721 | -0.184 | -0.179 | -0.089 |
| | GC[VSC]inv | 1.591 | 0.019 | 0.033 | 0.007 |
| | GC[VSC]part | 1.621 | -0.011 | 0.001 | -0.004 |
| Vowel space characteristics (VSC) | Inter-Vowel Dispersion (Inter-VD) | **1.402** | **0.431** | **0.440** | **0.263** |
| | Intra-Vowel Dispersion (Intra-VD) | 1.553 | 0.121 | 0.146 | 0.058 |
| | formant dependency (FD) | 1.485 | 0.406 | 0.337 | 0.245 |
| Feature fusion | Inter-VD+FD+$GC[VSC]_{inv}$+Syncrony[VSC] | **1.357** | **0.487** | **0.508** | **0.289** |
| | Inter-VD+FD+Syncrony[VSC] | 1.381 | 0.467 | 0.486 | 0.317 |
| | Inter-VD+FD+$GC[VSC]_{inv}$ | 1.364 | 0.462 | 0.467 | 0.294 |
| | Inter-VD+FD | 1.387 | 0.456 | 0.443 | 0.308 |
| | Convergence[P] | 1.524 | 0.270 | 0.263 | 0.151 |

The rows represent the feature sets used in training the Support Vector Machine (SVM) regressor. The columns represent the evaluation metrics: mean absolute error, Pearson's coefficient, Spearman's coefficient, and concordance correlation coefficient, respectively.

TABLE VII
RESULTS SHOWS REGRESSION RESULTS COMPARING TO OUR PREVIOUS WORK [19]

| $ADOS_{comm}$ assessment tasks | Model | Pear |
|---|---|---|
| This work | Inter-VD+FD+ $GC[VSC]_{inv}$+Syncrony[VSC] | 0.482 |
| Previous study [19] | $T_A$ | 0.121 |
| | $T_{W2V\ QT}$ | 0.478 |
| | $T_{BERT+A}$ | 0.501 |
| | $C_{BERT+A}$ | 0.567 |

result of the model in our previous study is 8.5% (in terms of Pearson's coefficient) higher than the result in this study (refer to Table VII). This study focuses only on the acoustic part with inclusion of novel measures of the vowel space characteristics. In the next section, we dive into more details about the ASD's characteristics on vowel space, along with the relationships between the VSC features and the ASD communication score.

*2) Analysis of the Feature Contributions to the Regression Task:* As described in the previous section, including $GC[VSC]_{inv}$ and Syncrony[VSC] can improve regression performances; we further analyze the contributing factors with SHAPley values. The needed difference ($\Delta$) and summation (I+F) of SHAPley values from the feature sets Inter-VD, FD, $GC[VSC]_{inv}$, Syncrony[VSC] and base values are shown in Table VIII. These data allow us to attribute the prediction errors of regression task to each feature set as demonstrated in Section V-B. The improvement in model performance is evaluated by the Mean Absolute Error (MAE), which is in lines with the results in Table VI. It is expected that the MAE of the best-performing model (denoted as $MAE^F$) should be lower than that of the baseline mode (denoted as $MAE^I$). In other words, $\Delta$ MAE $= MAE^F$-$MAE^I$ is expected to be a negative value. In addition, we found the $\Delta$ and I+F subcolumns under the feature sets $GC[VSC]_{inv}$ and Syncrony[VSC] are equal. The reason they are equal is because $GC[VSC]_{inv}$ and Syncrony[VSC] are absent in the initial model. The SHAPley value of these two feature sets is zero (they do not affect the model's prediction). Therefore, the difference and summation between the initial and final models are the same.

Table VIII demonstrates the results of the analysis of the differences between the initial model (**I**: Inter-VD+FD) and the final model (**F**: Inter-VD+FD+$GC[VSC]_{inv}$+ synchrony[VSC]). First, among the participants in group Q1, feature sets $\overline{\Delta \mathrm{BASEval}}$, $\overline{\Delta \mathrm{Inter-VD}}$, $\overline{\Delta \mathrm{FD}}$, $\overline{\Delta \mathrm{GC[VSC]_{inv}}}$ and $\overline{\Delta \mathrm{Syncrony[VSC]}}$ are negative. Together with the (9), these values will result in a negative $\Delta$MAE. Hence, it implies that the all these feature sets and base values contribute to correcting the regression errors, making the average MAE of group Q1 decrease (refer to the row Q1 (36) and column $\overline{\Delta \mathrm{MAE}}$). Second, among the participants classified as Q2, the results show $\overline{\Delta \mathrm{BASEval}}$, $\overline{\Delta \mathrm{Inter-VD}}$ and $\overline{\Delta \mathrm{FD}}$ are negative, which will increase the prediction error (notice that there's an additional negative sign in (10)). Hence these feature sets cause the mean absolute error to increase.

Third, the average MAE of group Q3 will decrease by 0.114 ($\overline{\Delta \mathrm{MAE}} = -0.114$) if $GC[VSC]_{inv}$ and Syncrony[VSC] are included. As shown in the Section V-B only $\overline{\mathrm{Y}}$ and the subcolumns of Table VIII denoted by I+F matters. In the subcolumns I+F of Table VIII, it shows that only the feature sets $GC[VSC]_{inv}$ and Syncrony[VSC] are positive. Although $GC[VSC]_{inv}$ and Syncrony[VSC] increased the mean absolute error, the Inter-VD, FD feature sets still reduced the error.

Lastly, the average MAE of the final model has decreased by 0.128 from the initial model in group Q4. The result shows that the I+F of the base value, Inter-VD, and FD's SHAPley values are positive. According to (12), these positive values will reduce the MAE in this group. In contrast, $GC[VSC]_{inv}$ and Syncrony[VSC] increase the MAE because their SHAPley

TABLE VIII
THIS TABLE DEMONSTRATES THE ANALYSIS OF CHANGES IN PREDICTION RESULTS WHEN GC[VSC]$_{\text{INV}}$+SYNCRONY[VSC] IS ADDED TO INTER-VD+FD

| **I**: Inter-VD+FD → **F**: Inter-VD+FD+GC[VSC]$_{\text{inv}}$+ synchrony[VSC] | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\overline{\text{BASEval}}$ | | $\overline{\text{Inter-VD}}$ | | $\overline{\text{FD}}$ | | $\overline{\text{GC[VSC]}_{\text{inv}}}$ | | $\overline{\text{Synchrony[VSC]}}$ | | $\overline{\text{Y}}$ | $\overline{\Delta \text{ MAE}}$ |
| | **F+I** | $\Delta$ | **F+I** | $\Delta$ | **F+I** | $\Delta$ | **F+I** | $\Delta$ | **F+I** | $\Delta$ | | |
| Q1 (36) | 6.973 | -0.006 | -0.141 | -0.018 | -0.092 | -0.003 | -0.008 | -0.008 | -0.001 | -0.001 | 1.811 | -0.036 |
| Q2 (37) | 6.942 | -0.010 | 0.155 | -0.023 | 0.057 | -0.015 | 0.022 | 0.022 | 0.021 | 0.021 | 5.077 | 0.004 |
| Q3 (6) | 6.992 | -0.002 | -0.554 | 0.299 | -0.197 | 0.019 | 0.052 | 0.052 | 0.164 | 0.164 | 3.286 | -0.114 |
| Q4 (7) | 7.011 | 0.050 | 1.320 | -0.430 | 0.738 | -0.061 | -0.225 | -0.225 | -0.050 | -0.050 | 4.333 | -0.128 |

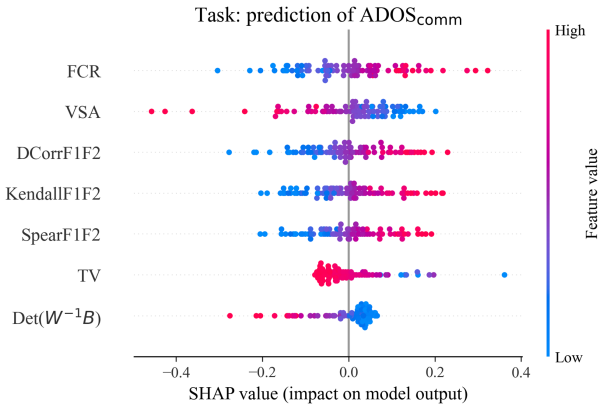


Fig. 4. Top seven features according to SHAPley values. This figure shows the correlation between features' values and their corresponding SHAPley value, in which the features are from feature sets: formant dependency and Inter-VD.

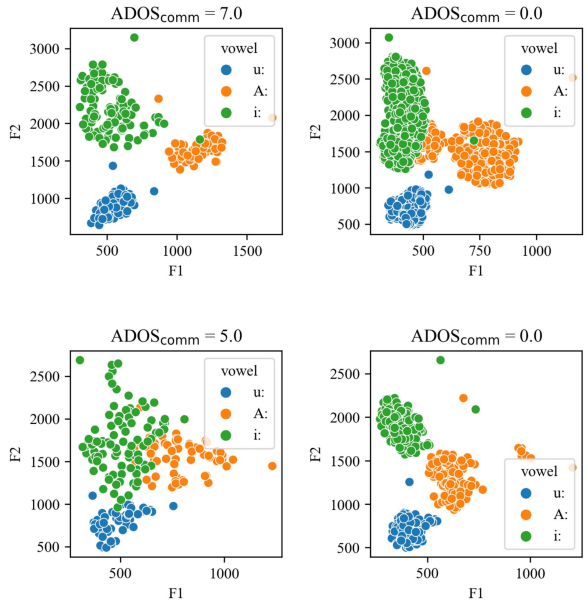


Fig. 5. Four examples of the high-score (the figure to the left) and low-score (the figure to the right) ASD participants' vowel space characteristics.

values are negative. The result implies that feature sets Inter-VD and FD are the major feature sets that help improve prediction accuracy.

In a nutshell, we observed that the inclusion of feature sets GC[VSC]$_{\text{inv}}$ and Syncrony[VSC] have decreased the prediction error by 0.031 $((-0.036 \times 36 + 0.004 \times 37 + -0.114 \times 6 + -0.128 \times 7) / 86)$ in average. Furthermore, by performing feature fusion with GC[VSC]$_{\text{inv}}$ and Syncrony[VSC], the prediction accuracies of the Q3 and Q4 groups, have increased the most ($\overline{\Delta\text{MAE}_{\text{Q3}}} = -0.114$, $\overline{\Delta\text{MAE}_{\text{Q4}}} = -0.128$). In other words, the participants initially predicted lower and higher but ultimately predicted higher and lower than the real assessment are the groups that benefit more.

*3) Analysis of the Relationship Between ADOS Communication Deficit Score and Feature Sets: FD and Inter-VD:* Two utterance-level VSCs feature sets, Inter-VD and FD, are analyzed for having competitive single feature prediction results in this task. According to the SHAPley analysis demonstrated in Fig. 4, all features in FD are positively correlated to their SHAPley values. Furthermore, features representing the dispersion of three corner vowels (BCC, BCV, ...) are negatively correlated to their SHAPley values, and formant centralization ratio (FCR, which has the opposite meaning to inter vowel dispersion) is positively correlated to their SHAPley values. These results indicate that participants with higher ADOS$_{\text{comm}}$ correspond to larger feature values of features in FD and lower feature values of that in Inter-VD. This result implies that ASD participants with more severe communication symptoms have less flexibility in their articulators and lower vowel intelligibility. Fig. 5 demonstrates the distribution of corner vowels on vowel space from two different severity-level ASD. According to this figure, the higher ADOS$_{\text{comm}}$ ASD participant might have a higher correlation between F1 and F2 or a lower distinction between vowel clusters. Notice that participants with low severity scores might still have overlap in vowel space, and participants with high severity scores might have distinct vowel clusters. The patterns of distinct vowel clusters and dependant formant values can only be considered as factors in identifying ASD.

## VI. DISCUSSION

This study uses formants of corner phones to model vowel intelligibility, vowel variability, and articulator flexibility, which are common measurements for autistic traits in past research. Our experimental results demonstrate that these measurements can effectively distinguish ASD and TD participants. In addition, both vowel intelligibility and articulator flexibility are associated with ASD's communication deficit. Furthermore, the articulation of severe ASD tends to have less vowel intelligibility and articulator flexibility, which aligns with prior studies [20], [24], [25]. However, merely vowel intelligibility and articulator flexibility can not be diagnostic standards; instead, they should be considered as factors that characterize the speech production

of ASD participants. Due to the heterogeneity of ASD, the observation that severe ASD is associated with less vowel intelligibility and articulator flexibility did not apply to all participants but a significant proportion of them.

Although the proximity, synchrony, convergence, and gradual changes of VSC features did not achieve competitive results by themselves, we still observed interesting trends. For example, we found less severe ASD participants may have closer vowel intelligibility to their investigators, indicated by the proximity metrics on vowel intelligibility. However, only a few features from vowel-intelligibility proximity features, the proximity of within-class variance, between-class variance, and total variance, correlate to ASD communication deficit score. The other features, like proximity of within-class covariance in the same category, did not show a significant correlation, and thus the feature set Proximity[VSC] did not perform well on the regression task. Perhaps there are factors unrelated to ASD that affects the robustness of features related to vowel-intelligibility proximity, making the features in this set inconsistent.

The limitation of this study is that we only investigate the acoustics of corner phones in conversations, whereas the articulatory dynamics in conversations might be complex. For example, each phone may be influenced by its contextual phones, and hence the phone sequences such as diphones and triphones need to be considered to understand better the dynamical properties of articulatory movement; identifying these phones in conversations will require advanced methodology. In addition, Mandarin is a tonal language. The interaction between tones and formants should also be considered when deriving VSC features. However, past research has found that young people with autism exhibit distinctive ways of the third tone, distinguishing them from typically developing people [18]. Perhaps the vowel space characteristics, tone, and the interaction of them can better characterize autistic traits.

Acoustics and language should also be considered jointly when modeling speech for ASD characterization. The deficit in communication not only pertains to speech acoustics and terminology usage but also encompasses the acoustics within each spoken word. Many prior works have established acoustic and natural language processing algorithms for modeling atypical patterns to characterize ASD. Furthermore, fusing acoustic and lexical features have shown competitive prediction results in our past study [19]. This study further shows that articulatory acoustics is important in understanding autistic trait. Therefore, a possible future extension is to explore more acoustic-linguistic measurements. For example, the acoustics of advanced phonetic sequences such as triphone ordiphone or the interaction of speaking tone and corresponding acoustic values may be relevant to modeling the social interaction of ASD. While past research has successfully modeled communication through lexical and acoustic inputs, the acoustic measured conditioned on certain linguistic tokens is yet to be studied.

## VII. Conclusion

This research characterizes autistic traits by measuring vowel space characteristics from two aspects: utterance-level for speech production and conversation-level for interaction dependency. These VSC measurements are used in training a classifier to predict whether the unseen testing instance is ASD or TD (the binary-classification subtasks) and a regressor to predict the ASD-related communication score (the $\text{ADOS}_{\text{comm}}$ regression task). The experiment results demonstrate that the conversation-level acoustic-prosodic features help classify ASD and TD. Then, although VSC feature sets by themselves do not have high prediction scores in the binary-classification subtasks, encompassing these feature sets can correct the prediction of several hard-to-classify participants who are originally mispredicted. This observation implies that a distinctive autistic profile of participants can not be correctly classified by simply looking at the prosody characteristics (it would require the vowel space characteristics). Furthermore, according to the SHAPley analysis, we observe that our VSC features improve the prediction scores by either directly contributing to the models' decision values or indirectly changing the influences of the other features on the models' decision values. Next, the regression experiment shows that VSC features related to formant dependency and inter-vowel dispersion are positively and inversely related to ASD communication severity. Additionally, similar to the binary classification task, we also observe that including appropriate VSC-related features in feature fusion can improve the prediction score; despite some of these features having low prediction scores by themselves. Moreover, through SHAPley analysis, we observe that these types of features correct mostly participants from defined Q3 and Q4 groups in Section V-B by reforming the contributing weight of features in the original model.

Lastly, autistic traits is a large umbrella term covering heterogeneous behavior symptoms. Speech provides a naturally-rich information signal that includes multi-faceted manifestations of autistic traits. For this purpose, this paper provides one of the few pieces of research in understanding vowel space characteristics for quantifying ASD participant's traits during spontaneous dialogue.

## References

[1] G. Xu, L. Strathearn, B. Liu, and W. Bao, "Prevalence of autism spectrum disorder among US children and adolescents, 2014-2016," *JAMA*, vol. 319, no. 1, pp. 81–82, 2018.

[2] S. Qiu et al., "Prevalence of autism spectrum disorder in Asia: A systematic review and meta-analysis," *Psychiatry Res.*, vol. 284, 2020, Art. no. 112679.

[3] J. Cakir, R. E. Frye, and S. J. Walker, "The lifetime social cost of autism: 1990–2029," *Res. Autism Spectr. Disord.*, vol. 72, 2020, Art. no. 101502.

[4] N. Rogge and J. Janssen, "The economic costs of autism spectrum disorder: A literature review," *J. Autism Develop. Disord.*, vol. 49, no. 7, pp. 2873–2900, 2019.

[5] M.-H. Chen et al., "Autistic spectrum disorder, attention deficit hyperactivity disorder, and psychiatric comorbidities: A nationwide study," *Res. Autism Spectr. Disord.*, vol. 10, pp. 1–6, 2015.

[6] American Psychiatric Association, D. S. M. T. F., and American Psychiatric Association, Diagnostic and statistical manual of mental disorders: DSM-5, vol. 5, no. 5, Washington, DC: American psychiatric association, 2013.

[7] J. C. Wakefield, "Diagnostic issues and controversies in DSM-5: Return of the false positives problem," *Annu. Rev. Clin. Psychol.*, vol. 12, pp. 105–132, 2016.

[8] J. N. Constantino and C. P. Gruber, *Social Responsiveness Scale: SRS-2*. Torrance, CA, USA: Western Psychol. Serv., 2012.

[9] L. C. Eaves, H. D. Wingert, H. H. Ho, and E. C. Mickelson, "Screening for autism spectrum disorders with the social communication questionnaire," *J. Develop. Behav. Pediatr.*, vol. 27, no. 2, pp. S95–S103, 2006.

[10] C. Lord et al., "The autism diagnostic observation schedule–generic: A. standard measure of social and communication deficits associated with the spectrum of autism," *J. Autism Develop. Disord.*, vol. 30, no. 3, pp. 205–223, 2000.
[11] E. Schopler, M. D. Lansing, R. J. Reichler, and L. M. Marcus, *PEP-3, Psychoeducational Profile.* Austin, TX, USA: Pro-ed, 2005.
[12] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proc. IEEE*, vol. 101, no. 5, pp. 1203–1233, May 2013.
[13] D. Bone et al., "The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody," *J. Speech, Lang., Hear. Res.*, vol. 57, no. 4, pp. 1162–1177, 2014.
[14] D. Bone, J. Mertens, E. Zane, S. Lee, S. S. Narayanan, and R. B. Grossman, "Acoustic-prosodic and physiological response to stressful interactions in children with autism spectrum disorder," in *Proc. InterSpeech*, 2017, pp. 147–151.
[15] D. Bone, S. Bishop, R. Gupta, S. Lee, and S. S. Narayanan, "Acoustic-prosodic and turn-taking features in interactions with children with neurodevelopmental disorders," in *Proc. InterSpeech*, 2016, pp. 1185–1189.
[16] M. Li et al., "An automated assessment framework for atypical prosody and stereotyped idiosyncratic phrases related to autism spectrum disorder," *Comput. Speech Lang.*, vol. 56, pp. 80–94, 2019.
[17] J. A. Richards, D. Xu, and J. Gilkerson, "Development and performance of the lena automatic autism screen," *Lena Found.*, 2010.
[18] C. Guo, F. Chen, Y. Chang, and J. Yan, "Applying random forest classification to diagnose autism using acoustical voice-quality parameters during lexical tone production," *Biomed. Signal Process. Control*, vol. 77, 2022, Art. no. 103811.
[19] C.-P. Chen, S. S.-F. Gau, and C.-C. Lee, "Learning converse-level multimodal embedding to assess social deficit severity for autism spectrum disorder," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2020, pp. 1–6.
[20] J. Bishop et al., "Brief report: Autistic traits predict spectral correlates of vowel intelligibility for female speakers," *J. Autism Develop. Disord.*, vol. 52, pp. 2344–2349, 2021.
[21] M. Kissine, P. Geelhand, M. Philippart De Foy, B. Harmegnies, and G. Deliens, "Phonetic inflexibility in autistic adults," *Autism Res.*, vol. 14, no. 6, pp. 1186–1196, 2021.
[22] R. Fusaroli, A. Lambrechts, D. Bang, D. M. Bowler, and S. B. Gaigg, "Is voice a marker for autism spectrum disorder? A systematic review and meta-analysis," *Autism Res.*, vol. 10, no. 3, pp. 384–407, 2017.
[23] M. Kissine and P. Geelhand, "Brief report: Acoustic evidence for increased articulatory stability in the speech of adults with autism spectrum disorder," *J. Autism Develop. Disord.*, vol. 49, no. 6, pp. 2572–2580, 2019.
[24] T. Talkar et al., "Assessment of speech and fine motor coordination in children with autism spectrum disorder," *IEEE Access*, vol. 8, pp. 127535–127545, 2020.
[25] L. McKeever, J. Cleland, and J. Delafield-Butt, "Aetiology of speech sound errors in autism," in *Speech Production and Perception: Learning and Memory*. Peter Lang GmbH, 2019, pp. 109–138.
[26] B. A. Lippke, S. E. Dickey, J. W. Selmar, and A. L. Soder, *PAT-3: Photo Articulation Test.* Austin, TX, USA, Pro-Ed, 1997.
[27] K. Ochi et al., "Quantification of speech and synchrony in the conversation of adults with autism spectrum disorder," *PLoS One*, vol. 14, no. 12, 2019, Art. no. e0225377.
[28] H. Lehnert-LeHouillier, S. Terrazas, and S. Sandoval, "Prosodic entrainment in conversations of verbal children and teens on the autism spectrum," *Front. Psychol.*, vol. 11, 2020, Art. no. 2718.
[29] J. Kruyt and Ş. Beňuš, "Prosodic entrainment in individuals with autism spectrum disorder," *Topics Linguistics*, vol. 22, no. 2, pp. 47–61, 2021.
[30] C.-P. Chen, S. S.-F. Gau, and C.-C. Lee, "Toward differential diagnosis of autism spectrum disorder using multimodal behavior descriptors and executive functions," *Comput. Speech Lang.*, vol. 56, pp. 17–35, 2019.
[31] C.-P. Chen, X.-H. Tseng, S. S.-F. Gau, and C.-C. Lee, "Computing multimodal dyadic behaviors during spontaneous diagnosis interviews toward automatic categorization of autism spectrum disorder," in *Proc. InterSpeech*, 2017, pp. 2361–2365.
[32] S. Sapir, C. Fox, J. Spielman, and L. Ramig, "Acoustic metrics of vowel articulation in parkinson's disease: Vowel space area (VSA) vs. vowel articulation index (VAI)," in *Proc. Int. Workshop Models Anal. Vocal Emissions Biomed. Appl.*, 2011, pp. 173–175.
[33] C. DiCanio, H. Nam, J. D. Amith, R. C. García, and D. H. Whalen, "Vowel variability in elicited versus spontaneous speech: Evidence from mixtec," *J. Phonetics*, vol. 48, pp. 45–59, 2015.
[34] M. K. Belmonte, T. Saxena-Chandhok, R. Cherian, R. Muneer, L. George, and P. Karanth, "Oral motor deficits in speech-impaired children with autism," *Front. Integrative Neurosci.*, vol. 7, 2013, Art. no. 47.
[35] J. P. McCleery, N. A. Elliott, D. S. Sampanis, and C. A. Stefanidou, "Motor development and motor resonance difficulties in autism: Relevance to early intervention for language and communication skills," *Front. Integrative Neurosci.*, vol. 7, 2013, Art. no. 30.
[36] C. J. Wynn, E. R. Josephson, and S. A. Borrie, "An examination of articulatory precision in autistic children and adults," *J. Speech, Lang., Hear. Res.*, vol. 65, pp. 1416–1425, 2022.
[37] R. H. Gálvez, L. Gauder, J. Luque, and A. Gravano, "A unifying framework for modeling acoustic/prosodic entrainment: Definition and evaluation on two large corpora," in *Proc. SIGDIAL*, 2020, pp. 215–224.
[38] J. V. Quiros, O. Kapcak, H. Hung, and L. Cabrera-Quiros, "Individual and joint body movement assessed by wearable sensing as a predictor of attraction in speed dates," *Trans. Affect. Comput.*, vol. 14, no. 3, pp. 2168–2181, Jul.–Sep. 2023.
[39] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot. Int.*, vol. 5, no. 9, pp. 341–345, 2001.
[40] F. Wu, L. P. García-Perera, D. Povey, and S. Khudanpur, "Advances in automatic speech recognition for child speech using factored time delay neural network," in *Proc. INTERSPEECH*, 2019, pp. 1–5.
[41] Y.-F. Liao, Y.-H. S. Chang, Y.-C. Lin, W.-H. Hsu, M. Pleva, and J. Juhar, "Formosa speech in the wild corpus for improving Taiwanese Mandarin speech-enabled human-computer interaction," *J. Signal Process. Syst.*, vol. 92, no. 8, pp. 853–873, 2020.
[42] B. M. Lobanov, "Classification of Russian vowels spoken by different speakers," *J. Acoust. Soc. Amer.*, vol. 49, no. 2B, pp. 606–608, 1971.
[43] N. Flynn, "Comparing vowel formant normalisation procedures," *York Papers Linguistics Ser.*, vol. 2, no. 11, pp. 1–28, 2011.
[44] Z. Gu, H. Mori, and H. Kasuya, "Analysis of vowel formant frequency variations between focus and neutral speech in Mandarin Chinese," *Acoust. Sci. Technol.*, vol. 24, no. 4, pp. 192–193, 2003.
[45] R. D. Kent and H. K. Vorperian, "Static measurements of vowel formant frequencies and bandwidths: A review," *J. Commun. Disord.*, vol. 74, pp. 74–97, 2018.
[46] N. Roy, S. L. Nissen, C. Dromey, and S. Sapir, "Articulatory changes in muscle tension dysphonia: Evidence of vowel space expansion following manual circumlaryngeal therapy," *J. Commun. Disord.*, vol. 42, no. 2, pp. 124–135, 2009.
[47] S. S. Wilks, "Multidimensional statistical scatter," in *Contributions to Probability and Statistics*, vol. 28, I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, and H. B. Mann, Eds. Stanford, California: Stanford University Press, pp. 486–503, 1960.
[48] K. Gotham, A. Pickles, and C. Lord, "Standardizing ADOS scores for a measure of severity in autism spectrum disorders," *J. Autism Develop. Disord.*, vol. 39, no. 5, pp. 693–705, 2009.
[49] C. J. Wynn and S. A. Borrie, "Classifying conversational entrainment of speech behavior: An expanded framework and review," *J. Phonetics*, vol. 94, 2022, Art. no. 101173.
[50] M. Sundararajan and A. Najmi, "The many Shapley values for model explanation," in *Proc. IEEE Int. Conf. Mach. Learn.*, 2020, vol. 119, pp. 9269–9278.
[51] S. Weidman, M. Breen, and K. C. Haydon, "Prosodic speech entrainment in romantic relationships," in *Proc. Speech Prosody*, 2016, pp. 508–512.

**Chin-Po Chen** (Student Member, IEEE) is currently working toward the Ph.D. degree in electrical engineering from National Tsing Hua University, Hsinchu, Taiwan. He is also an AI Research Intern with Inventec Corp, Taipei, Taiwan. He is also working on fine-tuning private large language model for question answering tasks. His research interests include audio signal processing (ASR, speaker diarization), generative AI, and explainable AI for health care.

**Ho-Hsien Pan** received the B.A. degree from Tamkang University, Taipei, Taiwan, in 1988, and the M.A. and Ph.D. degrees from Ohio State University, Columbus, OH, USA, in 1995. She is currently a Professor of foreign languages, literatures, and linguistic with National Yang Ming Chiao Tung University, Hsinchu, Taiwan. She is also the Director of speech and hearing science curriculum (NYCU) and was the Chair of foreign languages and literatures (during 2016–2017 and 2002–2003). Since 2018, she has been on the Editorial Board of the *Journal of International Phonetic Association*. She is the first and corresponding Author for 17 EI/SCI/SSCI papers. Her main research interests include pathological speech, Taiwan min nan speech corpus, laboratory phonology, speech perception, phonation, speech prosody. She was the recipient of the three outstanding teaching awards from the National Yang Ming Chiao Tung University, Taiwan.

**Susan Shur-Fen Gau** received the M.D. and Ph.D. degrees from Yale University, New Haven, CT, USA, in 2001. She is currently the Vice President of the National Taiwan University (NTU) Hospital, Taipei, Taiwan, and a Distinguished Professor of psychiatry, psychology, epidemiology, brain and mind sciences, clinical medicine, and occupational therapy with NTU, Taiwan. She was the Director of Departments of Psychiatry NTU Hospital and College of Medicine during 2009–2015, Director of Department of Medical Genetics with National Taiwan University Hospital during 2015–2018, President of the Taiwanese Society of Child and Adolescent Psychiatry during 2014–2018, and Vice-President of International Association of Child and Adolescent Psychiatry, and Allied Professionals (IACAPAP, during 2014–2018. She has conducted several studies on pharmacotherapy for ADHD, and been conducting follow-up, family, neuropsychological, neuroimaging, neurophysiological, microbiomes, metabolomics, genetic, and artificial intelligence studies on attention-deficit hyperactivity disorder (ADHD), and autism spectrum disorders (ASD). Her collaborative research also covers animal (mice & flies) and cellular (iPSC) models. She has authored more than 300 SCI/SSCI articles since 2001, of which she is the first author/corresponding author for more than 200 papers. She was the recipient of the outstanding research awards from the National Science Council in 2012, National Taiwan University in 2013, National Health Research Institute in 2014, and NTUH in 2016, Lifetime Academic Achievement Award from the Taiwanese Society of Psychiatry in 2019, Physician Model Award from Taiwan Medical Association (Taiwan MA, in 2020), Advancing Curiosity Award by Micron Foundation in 2021, USA, 31st WANG MING-NING Award in 2021, and 30th Hsin-Lin Award from Taipei Medical Association (Taipei MA). She is the keynote/state-of-the-art/plenary speaker at more than 20 international congresses. She and her team have presented their work in peer-reviewed Congress on more than 800 occasions.

**Chi-Chun Lee** (Senior Member, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from the University of Southern California, Los Angeles, CA, USA, in 2007 and 2012, respectively. He is currently a Professor with the Department of Electrical Engineering, National Tsing Hua University (NTHU), Hsinchu, Taiwan. His research interests include the speech and language, affective computing, health analytics, and behavioral signal processing. He has been an Associate Editor for the IEEE TRANSACTION ON AFFECTIVE COMPUTING since 2020, the *Journal of Computer Speech and Language* since 2021, the *APSIPA Transactions on Signal and Information Processing* and a TPC Member for APSIPA IVM and MLDA committee. He was an Associate Editor for IEEE TRANSACTION ON MULTIMEDIA during 2019–2020, He is the General Chair for ASRU 2023, Area Chair for Interspeech 2016, 2018, 2019, senior program committee for ACII 2017, 2019, publicity Chair for ACM ICMI 2018, late breaking result Chair for ACM ICMI 2023, sponsorship and special session Chair for ISCSLP 2018, 2020. He was the recipient of the Foundation of Outstanding Scholar's Young Innovator Award in 2020, CIEE Outstanding Young Electrical Engineer Award in 2020, IICM K. T. Li Young Researcher Award in 2020, NTHU Industry Collaboration Excellence Award in 2021 and 2023, and NSTC Futuretek Breakthrough Award in 2018 and 2019. He led a team to the 1st place in Emotion Challenge in Interspeech 2009, and with his students won the 1st place in Styrian Dialect and Baby Sound subchallenge in Interspeech 2019. He is a coauthor on the best paper award/finalist in Interspeech 2008, Interspeech 2010, IEEE EMBC 2018, Interspeech 2018, IEEE EMBC 2019, APSIPA ASC 2019, IEEE EMBC 2020, and the most cited paper published in 2013 in *Journal of Speech Communication*. He is also an ACM and ISCA Member.