MANDARIN PROSODIC FOCUS BY SPEAKERS WITH AUTISM SPECTRUM DISORDERS

Ho-hsien Pan¹, Shaoren Lyu¹, Chin-po Chen², Hai-ti Lin³, Susan Shur-fen Gau³

¹National Yang Ming Chiao Tung University, ²National Tsing Hua University, ³National Taiwan University Hospital

hhpan@nycu.edu.tw, shaorenlyu@gmail.com, stu9116@gapp.nthu.edu.tw, htdaisy125@gmail.com, gaushufe@gmail.com

ABSTRACT
The production of prosodic stress is a potential source of communication difficulties experienced by individuals with autism spectrum disorder (ASD). To investigate the acoustic cues used by individuals with ASD to distinguish given information from narrowly focused new information and to enhance narrow focus during post-focus syllables, read and spontaneous speech corpora were analyzed for both typically developing (TD) and ASD speakers of Taiwan Mandarin. The results revealed that both groups utilized duration lengthening and initial creakiness to mark narrow focus. However, duration shortening, which is a post-focus compression (PFC) cue used by TD speakers, was not used by ASD speakers. Additionally, initial glottalization with tense voice quality, as indicated by low H1-A3c values, was found to be significantly correlated with ASD speakers' memory and executive function scores. These findings suggest that acoustic cues could potentially serve as a noninvasive objective index of ASD.

Keywords: Tense voice quality, duration, executive function, memory, narrow focus.

1. INTRODUCTION
Autism Spectrum Disorder (ASD) is a neurodevelopmental condition that often leads to difficulties in communication and interpersonal interactions within a community. The perception and production of prosodic stress on prominent words in utterances are crucial for effective communication. However, individuals with ASD are prone to misperception of stress, particularly when the stress is located early in sentences [1]. Those with ASD who can stress important information are more likely to receive better communication ratings on the Autism Diagnostic Observation Schedule-Generic (ADOS-G) Communication scores [2].

Typically, new and contrasting information that act as the heads of utterances tend to be stressed prosodically. For instance, when asked to produce narrowly focused (NF) new information for words such as "CHOCOLATE" or "ICE CREAM" in contexts like "Did you eat the vanilla ice cream?" "I ate a CHOCOLATE ice cream." or "Did you eat the vanilla cake?" "I ate the vanilla ICE CREAM," individuals with ASD ambiguously distinguish given and NF new information [3, 4, 5].

Previous studies on the production of narrow focus in ASD speech have mainly focused on English. It remains unclear how ASD speakers of native languages with different prosodic typologies, such as Mandarin, express narrow focus. Furthermore, our understanding of the acoustic data used by native Mandarin speakers with ASD to signal NF new information is limited.

Those aims of this study is to investigate the following questions: (1) How do native Mandarin speakers with ASD distinguish NF new information from unfocused given information? (2) How do they enhance NF new information with post-focus compression? (3) Are there significant correlations between acoustic cues for narrow focus and the scores in the Cambridge Neuropsychological Test Automated Battery (CANTAB)?

Works on NF new information in Mandarin were conducted in controlled lab speech produced by typically developing (TD) native Beijing and Taiwan Mandarin speakers. These studies found that duration lengthening and F0 range expansion as the two major acoustic cues for NF syllables [6, 7, 8]. Moreover, Beijing Mandarin speakers lowered and compressed the pitch ranges of post-focus words to enhance the preceding NF new information. However, this post-focus compression (PFC) was not observed among Taiwan Mandarin or Taiwan Min Nan [9].

Although initial vowel glottalization was found to mark prosodic prominence in English [10], the potential use of initial glottalization as a cue for NF in Mandarin never explored. In this study, we aim to investigate this possibility, on top of duration lengthening and F0 range expansion. The two voice quality cues associated with initial glottalization were H1-H2 and H1-A3. Phonation-wise, the lower the H1-H2 values become, the creakier the voice quality is. The lower the H1-A3 values are, the more tense and abrupt the glottal closure becomes.

By comparing the acoustic cues of the same words carrying given versus NF new information or by
5. Phonation and Voice Quality

comparing the acoustic cues associated with post-
given versus post-new information in sequential
sentences produced, we explored the focus and post-
focus markings by both TD and ASD speakers of
Mandarin. Additionally, we reported individual cue
weighting in addition to group results to highlight the
heterogeneous nature of ASD speakers. Finally, we
correlated the effective acoustic measures with
cognition scores from the Cambridge
Neuropsychological Test Automated Battery
(CANTAB).

Based on our hypothesis, we expected that NF new
information would be marked by duration
lengthening, F0 range expansion, and initial
glottalized creaky or tense voice quality. On the other
hand, post-focus syllables were predicted to be
shorter in duration, narrower in F0 range, and more
modal-like in terms of voice quality to enhance
preceding NF information.

2. METHODS

2.1. Speakers

Spontaneous speech was collected from 81
participants, and all of whom were of Chinese descent.
The participants were divided into eight groups: five
TD boys, seven TD adult males, three TD girls, eight
TD adult females, seven ASD boys, 44 ASD adult
males, two ASD girls and five ASD females. The pre-
pubescent ASD and TD speakers were under 13 years
of age. The TD participants did not have any history
of speech or hearing problems. The ASD participants
were clinically diagnosed and confirmed by the
Autism Diagnostic Interview-Revised at the National
Taiwan University Hospital in Taipei, Taiwan.

2.2. Corpus

Spontaneous speech was collected from both ASD
speakers and TD children during the administration
of Module 3 of the Mandarin-ADOS-G, a semi-
structured clinical instrument designed to assess
ASD-related deficits. This module is intended for use
with children and adolescents who exhibit fluent
spontaneous speech and require the ability to talk
about objects or events that are not immediately
present. The participants’ responses to social-
emotional questions on topics such as emotions,
friends, loneliness, marriage, social difficulties, and
annoyance were recorded using a lapel microphone
attached to their collars. During the assessment, the
examiners, who were professionally trained therapists
or psychiatrists, observed and skipped inappropriate
questions that the participants had difficulty
answering. They also rated the subjects according to
socio-communication codes. The Mandarin-ADOS-
G has demonstrated good inter-rater reliability (0.91),
test-retest reliability (intraclass correlations 0.55-
0.73), and low to high good internal consistency
(Cronbach’s alpha 0.27-0.86) [11].

The linguistic institute of National Taiwan
University recorded the TD adults’ spontaneous
monologues. Each recording lasted approximately 30
minutes.

2.3. Data transcription

The sound files were transcribed in Praat [12] on
multiple tiers, including Chinese orthography, words,
syllables, tones, and surface segments. A dual-
language forced aligner was utilized to generate
segmental boundaries for both Taiwan Mandarin and
Min Nan. This aligner was developed through a
collaboration with the team that developed Munich
Automatic Segmentation System (MAUS) and the
Common Language Resources and Technology
Infrastructure (CLARIN) [16]. The aligner is
available through BAS Web Services [13, 14, 15].

A program developed in the lab was utilized to
detect repeated content words in consecutive
sentences. The first occurrence of a content word was
labelled as new information whereas subsequent
repetitions were marked as given information.
Similarly, in consecutive sentences, the words
following new and given information were labeled as
post-new and post-given, respectively.

Next, VoiceSauce [15] was employed to extract
acoustic parameters at every 10% vowel intervals
during syllables conveying NF new, given, post-new,
and post-given information.

2.4. Data analysis

Linear mixed-effect regression models were used to
analyze the acoustic measures of syllables carrying
NF new information compared to those carrying
given information. The models included speakers and
syllables as random effects, with given information as
the baseline data. Normalized duration, F0 range, and
normalized H1-H2c and H1-A3c were compared at
initial 10% vowel points. Similarly, the acoustic
measures of syllables following NF new information
were compared to those following given information
using linear mixed-effect regression models, with
post-given syllables as the baseline. These models
also included speakers and syllables as random
effects.

The H1-H2c measures the difference in amplitude
between the first and second harmonics, taking into
account the resonance effect of the vocal tract. The
H1-A3c corrected measures the magnitude difference
between the first harmonic and the strongest
harmonic of the third formant. The spectral tilt of H1-
H2c reflects the open phase of glottal vibration, with high values indicating breathy voice quality and low values indicating creaky voice quality. H1-A3c reflects the degree of abruptness in glottal closure, with lower values indicating more abrupt glottal closures and tense voices [16, 17].

3. RESULTS

3.1. New vs. Given information

The group results of linear mixed effect regression models comparing given vs. new information showed that TD adult males ($\beta = 0.26^{***}$), TD adult females ($\beta = 0.28^{***}$), ASD girls ($\beta = 0.46^{**}$) and ASD adult males ($\beta = 0.15^{**}$) produced NF new information with significantly longer normalized duration than given information.

Among all subject groups, only TD adult females produced a more expanded F0 range for NF new information ($\beta = 0.12^{***}$).

Regarding voice quality cues of NF syllables, TD adult males ($\beta = -0.08^{***}$), TD adult females ($\beta = -0.06^{*}$), and ASD adult males ($\beta = -0.07^{*}$) produced significantly lower H1-A3c values, indicating a more tense voice quality during the initial portion of NF syllables. However, the group results for H1-H2c did not show initial creaky voice for NF syllables.

Individual results revealed that five TD adult males, seven TD adult females, eight ASD adult males, and one ASD girl produced narrow focus with significantly longer duration.

Additionally, three TD adult females produced a more expanded F0 range for narrow focus. Furthermore, four ASD adult males produced narrow focus with creakier voice quality.

In summary, the group results revealed that both TD and ASD speakers produced narrow focus with longer duration and a more tense voice quality at the beginning of the syllable. The TD speakers also showed F0 range expansion for narrow focus. However, there were discrepancies between the group and individual results. Individual results showed that narrow focus was marked with longer duration by most speakers, F0 range expansion by a few TD speakers, and initial creaky voice only by some ASD speakers.

3.2 Post-focus vs. post-given syllables

To investigate how PFC enhances NF new information, post-new and post-given syllables were compared. It was hypothesized that post-focus syllables would exhibit significantly shorter duration, narrower F0 range, and less initial glottalization, as indicated by less creaky (high H1-H2c) or less tense (high H1-A3c) voice quality.

Group results showed that both TD adult males ($\beta = -0.10^{*}$) and females ($\beta = -0.11^{**}$) produced significantly shorter post-new syllables compared to post-given positions. However, individual analysis revealed that only one TD male adult and two TD female adults produced shorter post-focus syllables. No other PFC cues were observed among the eight subject groups.

3.3. Individual cue weighting

The acoustic differences between syllables carrying given and NF new information were plotted against acoustic differences between post-given and post-new to explore individual cue weightings.

As shown in Figure 1, TD speakers’ duration data points were clustered in the second quadrant. In other words, TD speakers produced longer durations during narrow focus and shorter durations during post-focus syllables.

For voice quality, most TD speakers produced more tense (H1-A3c in the third quadrant) but not creakier voice quality (H1-H2c in the first and second quadrants) to mark narrow focus. Post-focus syllables were produced with tense voice qualities to enhance the preceding narrow focus. Post-focus creaky voice was found only among some speakers.

For ASD speakers, the distribution of H1-H2c and H1-A3c data among the four quadrants was heterogeneous. However, it was observed that most ASD speakers produced syllables carrying NF new information with longer duration compared to those carrying given information.
5. Phonation and Voice Quality

Figure 1: Individual cue weighting of duration, H1-H2c, and H1-A3c differences between new vs. given information and between post-given vs. post-new syllables in ASD and TD speakers.

3.4. Cognitive functions and ASD acoustic cues

In the study of ASD individuals, linear regression models were employed to investigate the correlations between acoustic cues and various measures of memory or executive functions in the cognitive domain, which were assessed using the CANTAB tasks [18].

The correlation analysis revealed that the mean subsequent thinking time (three moves) of the Stocking of Cambridge (SOCst T3) was significantly correlated with the differences in H1-A3c values between syllables carrying new and given information (r=0.66**).

The Stocking of Cambridge (SOC) task is designed to assess visuospatial planning and problem-solving skills as part of the executive functions. During the task, subjects match a set of stockings containing three balls of different colors at the bottom of the screen to a pattern displayed at the top. The SOCstT3 is a measure of the subject's speed of movement, calculated as the difference in time between selecting the first ball and completing the problem in three moves (https://www.cambridgecognition.com/cantab/executive-function/stockings-of-cambridge-soc/). Higher SOCstT3 values indicate poorer executive function performance.

The positive correlation between SOCstT3 and H1-A3c during the initial portion of NF syllables suggests that ASD individuals who took longer times to complete the task also produced higher H1-A3c values, indicating a less tense voice quality, during narrow focus. Therefore, it can be concluded that ASD speakers with better executive function tend to produce more tense voice quality during the initial portion of narrow focus. The use of tense voice quality as a NF marking was also observed among the TD speakers.

Apart from the SOC task, the correlation analysis also showed that the H1-A3c values during the initial portion of post-focus syllables were significantly correlated with the total and adjusted total number of presentations required to locate the correct patterns in the Paired Associate Learning task (PALtT & PALtTA: r= -0.45*).

During the Paired Associate Learning (PAL) task, subjects were presented with a sequence of boxes that were opened randomly. The majority of the boxes were empty, except for one or two boxes containing patterns. The subjects were then asked to match the patterns of trials to the correct boxes (https://www.cambridgecognition.com/cantab/executive-tests/memory/paired-associates-learning-pal/).

The negative correlations showed that ASD subjects who required fewer trials to complete the PAL task tended to produce higher H1-A3c values, suggesting a less tense voice quality, at the initial portion for post-focus syllables. In other words, subjects who required fewer trials tended to produce a more modal-like voice quality during the initial portion of post-focus syllables to enhance the preceding narrow focus, which was produced with tense voice quality.

In summary, tense voice quality production correlated significantly with the ASD speakers’ executive function and visuospatial memory. Those ASD speakers who scored lower in PALtT & PALtTA but higher in executive function tasks (SOCst T3) take less time to memorize the patterns and complete the tasks faster. These ASD individuals also signalled narrow focus with initial tense voice quality and enhanced narrow focus with less tense voice quality during post-focus syllables.

4. DISCUSSION

Both TD and ASD speakers utilized duration lengthening and initial tense voice to mark narrow-focus syllables. However, only TD speakers exhibited post-focus duration shortening following narrow focus, indicating PFC enhancement. ASD speakers showed more variability, however, there was a positive correlation between the executive and visuospatial memory function of ASD speakers and the use of tense voice quality. Overall, these findings reveal the use of duration and tense voice quality mark narrow focus, with some differences in cue weighting between TD and ASD speakers.
5. REFERENCES


