https://doi.org/10.5664/jcsm.9798

JCSM | Journal of Clinical Sleep Medicine

SCIENTIFIC INVESTIGATIONS

Selection of OSA-specific pronunciations and assessment of disease severity assisted by machine learning

Yiming Ding, MM[1,2,3,*]; Yuechuan Sun, BE[4,*]; Yanru Li, MD[1,2,3]; Huijun Wang, MM[1,2,3]; Qiang Fang, MD[5]; Wen Xu, MD[1,2,3]; Ji Wu, PhD[4,6]; Jiandong Gao, PhD[4,6]; Demin Han, MD, PhD[1,2,3]

*[1]Beijing Tongren Hospital, Capital Medical University, Beijing, China; [2]Obstructive Sleep Apnea-Hypopnea Syndrome Clinical Diagnosis and Therapy and Research Centre, Capital Medical University, Beijing, China; [3]Key Laboratory of Otolaryngology Head and Neck Surgery, Ministry of Education, Capital Medical University, Beijing, China; [4]Department of Electronic Engineering, Tsinghua University, Beijing, China; [5]Institute of Linguistics, Chinese Academy of Social Sciences, Beijing, China; [6]Center for Big Data and Clinical Research, Institute for Precision Medicine, Tsinghua University, Beijing, China; *Contributed equally*

Study Objectives: To screen all of the obstructive sleep apnea (OSA)–characteristic pronunciations, explore the pronunciations which provide a better OSA classification effect than those used previously, and further clarify the correlation between speech signals and OSA.

Methods: A total of 158 adult male Mandarin native speakers who completed polysomnography at the Sleep Medicine Center of Beijing Tongren Hospital from November 15, 2019, to January 19, 2020, were enrolled in this study. All Chinese syllables were collected from each participant in the sitting position. The syllables, vowels, consonants, and tones were screened to identify the pronunciations that were most effective for OSA classification.

Results: The linear prediction coefficients of Chinese syllables were extracted as features and mathematically modeled using a decision tree model to dichotomize participants with apnea-hypopnea index thresholds of 10 and 30 events/h, and the leave-one-out method was used to verify the classification performance of Chinese syllables for OSA. Chinese syllables such as [leng] and [jue], consonant pronunciations such as [zh] and [f], and vowel pronunciations such as [ing] and [ai] were the most suitable pronunciations for classification of OSA. An OSA classification model consisting of several syllable combinations was constructed, with areas under curve of 0.83 (threshold of apnea-hypopnea index = 10) and 0.87 (threshold of apnea-hypopnea index = 30), respectively.

Conclusions: This study is the first comprehensive screening of OSA-characteristic pronunciations and can act as a guideline for the construction of OSA speech corpora in other languages.

Keywords: OSA, speech signals, speech corpus, machine learning

Citation: Ding Y, Sun Y, Li Y, et al. Selection of OSA-specific pronunciations and assessment of disease severity assisted by machine learning. *J Clin Sleep Med*. 2022;18(11):2663–2672.

BRIEF SUMMARY

Current Knowledge/Study Rationale: Patients with obstructive sleep apnea (OSA) have speech abnormalities compared to healthy individuals, and speech signals can be used to assess the severity of OSA. However, comprehensive screening and evaluation of OSA-characteristic pronunciations and speech corpus have not been performed.

Study Impact: This study is the first systematic screening of OSA-characteristic pronunciations and will help to improve the effectiveness of speech signals for OSA assessment.

INTRODUCTION

With advancements in medical knowledge and health care awareness, obstructive sleep apnea (OSA) has received increasing attention as a primary disease that can trigger or aggravate multisystem diseases (eg, hypertension, coronary heart disease, diabetes mellitus, etc).[1,2] OSA is caused by intermittent upper airway collapse during sleep, manifested by snoring and sleep apnea with hypopnea, and accompanied by varying degrees of decreased oxygen saturation and sleep structure disorders.[3] In addition to causing a variety of diseases, OSA can lead to reduced concentration during wakefulness, low work efficiency, and even drowsiness, leading to various accidents in daily life.[4]

OSA is a widely prevalent problem in the general population.[5] There were an estimated 936 million patients with OSA worldwide in 2019, including 176 million in China,[6] based on American Academy of Sleep Medicine (AASM) 2012 diagnostic criteria. However, a large number of patients with OSA are only concerned with snoring symptoms and do not feel that they need to go to hospital, so many patients with OSA go undiagnosed. The number of undiagnosed patients with OSA in the United States is estimated to be 24 million.[7] Even when they do go to hospital, the gold standard for diagnosing OSA, polysomnography (PSG), is difficult to complete quickly because it is time consuming and the waiting time for an appointment can be several weeks. In addition, PSG is difficult to perform in primary care hospitals because it requires complex testing instruments and specialized sleep technicians for data analysis and interpretation. Therefore, an effective out-of-hospital assessment of individuals who may have OSA would help to increase the rate of OSA visits and diagnosis.

The Stop-Bang sleep questionnaire is the tool most commonly used to screen for OSA, but results from the questionnaire are influenced by self-reported factors and have low specificity, especially for patients with severe OSA, in whom the results have specificity less than 50%.[8]

The speech signal is ideal for the evaluation of OSA, as this information can be easily and quickly obtained and contains a large number of individual characteristics. Earlier studies have identified characteristic differences in the upper airway structure of patients with OSA compared to healthy individuals, and these combined with the effects of long-term snoring can result in patients with OSA developing abnormalities in their speech, including articulation, phonation, and resonance abnormalities.[9] Based on this, researchers have studied the characteristics of abnormal speech signals in patients with OSA to evaluate the severity of the disease. In 2009, Pozo et al presented experimental findings regarding the discriminative power of the Gaussian mixture model applied to severe apnea detection and achieved an 81% correct classification rate.[10] In 2012, Benavides et al improved the correct classification rate to 85% by using hidden Markov models.[11] After that, researchers then successively used speech features to screen for OSA and obtained satisfactory results.[12–14]

A *speech corpus* refers to a collection of selected linguistic materials used for some specific purpose of phonological research. However, earlier studies did not screen the speech corpus used to capture speech signals of patients with OSA, and mainly selected the common vowels /a/, /e/, /i/, /o/, /u/ or sentences as the speech corpus. According to results from our earlier study, different pronunciations affect the classification of OSA.[15] Therefore, we considered it desirable to perform a more comprehensive pronunciation screening on patients with OSA to determine the most suitable pronunciations for OSA classification. Our goals were to improve the efficacy of speech signals as an assessment of the severity of OSA, lessen the burden of recordings on individuals, and further elucidate the association between speech signals and OSA.

METHODS

Participants

This study included 158 adult male Mandarin native speakers who visited Beijing Tongren Hospital between November 15, 2019, and January 19, 2020, and who were experiencing sleep snoring. The study excluded individuals with the following conditions: recent upper respiratory tract infection, allergic rhinitis, sinusitis, chronic obstructive pulmonary disease, vocal cord disorders, history of pharyngeal surgery, craniofacial trauma or deformity, history of speech disorders, or history of psychiatric disorders. Central sleep apnea was excluded after PSG had been performed. Basic information such as age, weight, height, and neck circumference were collected from the participants. All individuals underwent PSG at the Sleep Center of Beijing Tongren Hospital and were grouped according to apnea-hypopnea index (AHI) thresholds of 10 events/h and 30 events/h, as shown in **Table 1**. The study was approved by the Ethics Committee of Beijing Tongren Hospital, no. TRECKY2019-049. All participants in this study signed informed consent forms.

Speech signal collection

Before sleep monitoring, the participants were recorded in a sitting position in a quiet environment and in a calm state using a Sony PCM-D10 recorder with a sampling frequency of 44,100 Hz and a 16-bit quantization accuracy. The recorder was placed approximately 50 cm from the participant's head. In total, 374 Chinese syllables were collected from each individual (see **Table S1** in the supplemental materials for details), with a duration of about 1 second and an interval of about 1 second for each syllable. A full audio recording of all syllables has been uploaded in the supplemental material, read by Yiming Ding, one of the authors of this article, a male Mandarin native speaker.

Polysomnography

In this study, PSG was performed on all of the participants using the Philips Respironics G3 sleep diagnostic system, which included 2-channel electroencephalography (C3/M2, C4/M1), 2-channel electro-oculography, anterior tibial electromyogram, electrocardiogram, 2-channel airflow measurement with nasal cannula pressure, recording of respiratory (thoracic and abdominal) movements, and pulse oximetry for oxygen saturation. All of the electrocardiogram and electro-oculography channels were captured at a sampling frequency of 200 Hz and displayed with a 0.3–35-Hz bandpass filter. Anterior tibial electromyogram had a sampling rate of 200 Hz, and the bandpass filter was 10–100 Hz. Three PSG technologists, each with more

Table 1—Participant characteristics.

	Threshold: AHI = 10 events/h			Threshold: AHI = 30 events/h		
	AHI > 10 (n = 117)	AHI ≤ 10 (n = 41)	*P*	AHI > 30 (n = 80)	AHI ≤ 30 (n = 78)	*P*
Age (years)	39.3 ± 8.77	39.1 ± 11.0	.9288	39.3 ± 9.03	39.2 ± 9.77	.9294
Height (cm)	174 ± 5.61	175 ± 5.38	.2034	174 ± 5.79	174 ± 5.36	.9480
Weight (kg)	85.2 ± 12.5	80.1 ± 10.2	.0191	87.8 ± 11.8	79.9 ± 11.1	< .001
BMI (kg/m^2)	28.1 ± 3.48	26.1 ± 3.21	.0013	28.9 ± 3.35	26.3 ± 3.19	< .001
NC (cm)	41.7 ± 3.25	39.8 ± 2.54	.0009	42.3 ± 3.04	40.1 ± 2.95	< .001
AHI (events/h)	46.3 ± 23.6	5.47 ± 2.97	<.0001	59.4 ± 16.1	11.4 ± 7.69	< .001

AHI = apnea-hypopnea index, BMI = body mass index, NC = neck circumference.

than 10 years' experience, scored sleep stages and respiratory events in accordance with the American Association of Sleep Medicine (AASM 2012) guidelines.[16] AHI is defined as the number of apneic and hypopneic events per hour of sleep and is used to indicate the severity of sleep apnea.

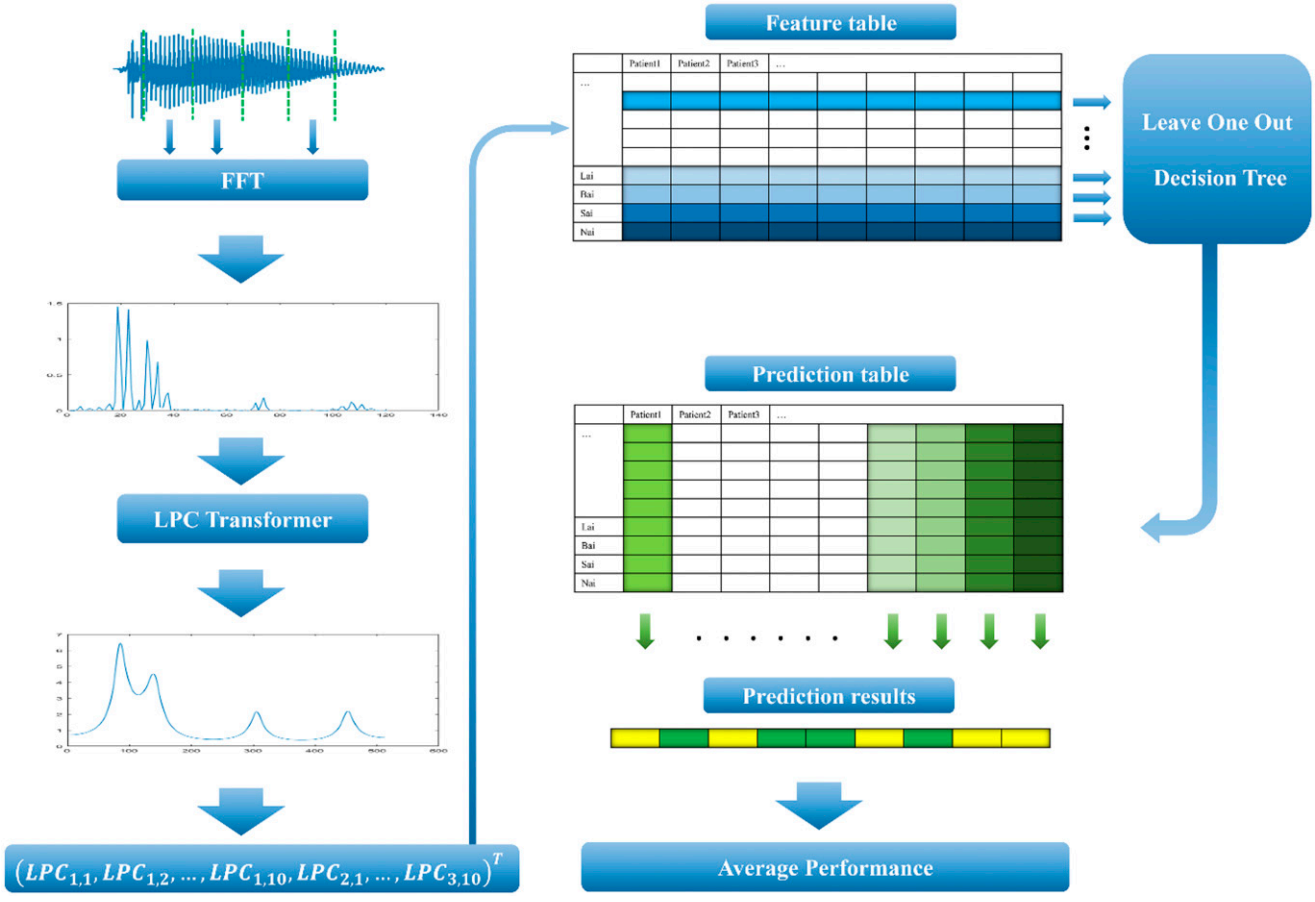
Speech signal processing and modeling

We preprocessed the speech data composed of Chinese single-character syllables, extracted the linear prediction coefficients (LPC) frame-by-frame, and performed mathematical modeling using a decision tree model to obtain the classification performance of single-character syllables for OSA using the leave-one-out validation method. The results were then combined to obtain the average performance of Chinese vowels and consonants in OSA classification. The specific process is shown in **Figure 1**.

Preprocessing and feature extraction

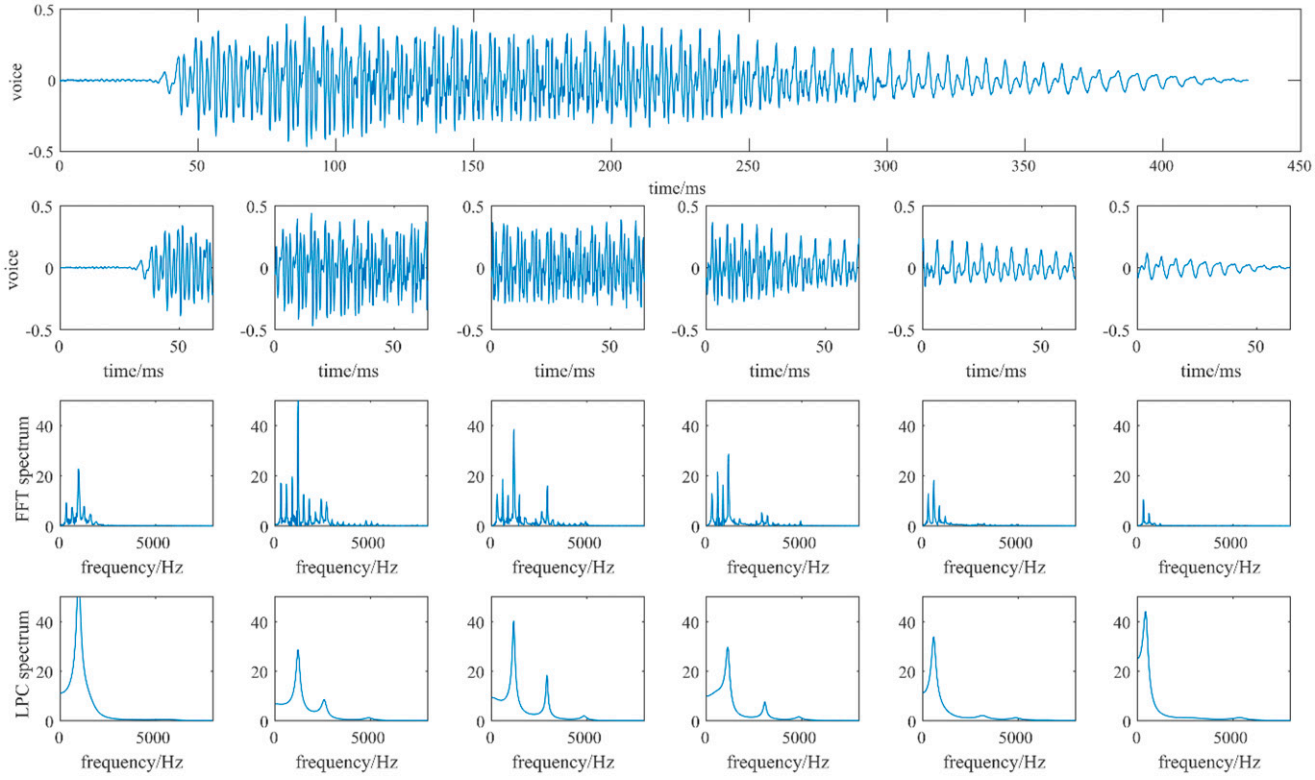
Considering that Chinese single-character syllables include a vowel part and a consonant part, and the vowel part is more variable, by including single vowels (eg, [a], [i]) and compound vowels (eg, [iao], [uai]) in the preprocessing stage, we expected to synthesize the pronunciation characteristics of each part of the single character for problem modeling and prediction. Instead of using the customary ∼20–40-ms speech frame length, we divided the pronunciation into 6 equal frames based on the stages of the single-character syllable, and the frame length floated from 50 ms to 70 ms according to the pronunciation length. The patients' pronunciation was slow and steady, so short-time smoothness of speech was guaranteed. The comparison test determined that segment 2 was the stage at which the consonants were clearer, and segments 3 and 5 were the stages at which the vowels were clearer. Taking the syllable [hua] as an example, there are 3 pronunciation stages: [h], [u], and [a]. The syllable segmentation is shown in **Figure 2**. The first row of **Figure 2** shows the waveform of syllable [hua], the second row shows the waveform of the pronunciation of the syllable cut into 6 segments on average, the third row shows the fast Fourier transform amplitude spectrum of each segment,

Figure 1—Flow diagram for extraction and modeling of speech signal features.



First, each single-character syllable from each patient was divided into 6 equal segments, and LPC features were extracted from segments 2, 3, and 5. The top 10 orders of LPC coefficients were taken for each segment, stitched together to obtain 30-dimensional features, and these features were recorded in a table. After that, the features were selected for all patients under each single-character syllable for decision tree modeling and leave-one-out validation, and the prediction results were obtained for each single-character syllable for each patient. Finally, the relevant single-character syllables were selected according to their vowels or consonants, and their prediction results were voted on for each patient to obtain the prediction results for that patient. Then the prediction performance was evaluated as the average performance for that vowel or consonant. FFT = fast Fourier transform, LPC = linear prediction coefficients.

Figure 2—Syllable syncopation diagram.



Taking the syllable [hua] as an example, there are 3 pronunciation stages: [h], [u], and [a]. The first row of the figure shows the waveform for syllable [hua], the second row shows the waveform for the pronunciation of the syllable cut into 6 segments on average, the third row shows the FFT amplitude spectrum for each segment, and the fourth row shows the LPC amplitude spectrum for each segment. It can be seen that the first segment contains a change from silence to the beginning of articulation, which is not in line with the short-term stability; the second segment is mainly pronounced with the consonant [h] and part of the vowel [u]; the third segment is mainly pronounced with the vowel [u], and the frequency point of the second resonance peak of the LPC spectrum is shifted to the right compared with the second segment; the pronunciation of the fourth paragraph is the conversion stage from vowel [u] to vowel [a], and the LPC spectrum also changes synchronously; the main pronunciation of the fifth paragraph is vowel [a], and its LPC spectrum is quite different from that of the second and third paragraphs; the main pronunciation of the sixth paragraph is still vowel [a], but the waveform amplitude is small. Considering them together, we selected the second, third, and fifth pronunciation segments to be spliced together as the training features. FFT = fast Fourier transform, LPC = linear prediction coefficients.

and the fourth row shows the LPC amplitude spectrum of each segment. It can be seen that the first segment contains a change from silence to the beginning of articulation, which is not in line with the short-term stability; the second segment is mainly pronounced with the consonant [h] and part of the vowel [u]; the third segment is mainly pronounced with the vowel [u], and the frequency point of the second resonance peak of the LPC spectrum is shifted to the right compared with the second segment; the fourth segment is the conversion stage from vowel [u] to vowel [a], and the LPC spectrum is also changing synchronously; the main pronunciation of the fifth segment is vowel [a], and its LPC spectrum is quite different from that of the second and third paragraphs. The main pronunciation of the sixth segment is still vowel [a], but the waveform amplitude is small. By consensus, we selected the LPC coefficients of the second, third, and fifth pronunciation segments to be spliced together as the training features of the word for machine learning. We extracted the common speech features including formants (F1–F4), Mel frequency cepstral coefficients and their first- and second-order difference splicing, filter banks (fbanks) and their first- and second-order difference splicing, linear prediction coefficients, and linear prediction cepstral coefficients. Through comparative experiments based on classification performance, we determined the top 10 coefficients of LPC to be the most suitable characteristics for resolving problems related to Chinese single-character syllables. The 3 stages of speech signals provide us with Chinese single-character syllable features with a feature dimension of 30, which achieves a more effective feature-dimension compression.

Problem modeling and prediction

We modeled the problem using a decision tree model. We hoped to obtain the performance of all Chinese single-character syllables for OSA classification, so we used the leave-one-out cross-validation method to collect all available patient data for each Chinese single-character syllable, perform feature extraction, model prediction, and performance evaluation, and finally obtain the classification performance for each Chinese single-character syllable. As the performance of a single-character syllable was unreliable, it was important to obtain a more stable

and reliable performance, and to determine the classification performance for each vowel and consonant, to facilitate the subsequent analysis. To do this, we calculated the classification performance for each vowel and consonant based on the results for single-character syllables containing the same vowel or consonant for each patient. The result treats the majority of multiple classification results as the final result. Specifically, we selected the prediction results for all single-character syllables containing a certain vowel or consonant and voted them by the patient, from which the positive and negative class prediction results for patients were obtained and the prediction performance was evaluated as the average performance for that vowel or consonant.

RESULTS

Participant characteristics

In this study, all participants were dichotomized 2 times with the AHI = 10 events/h classification to initially determine whether participants had OSA, and then using the AHI = 30 events/h classification to determine whether participants had severe OSA. The specific information on participants after grouping is shown in **Table 1**.

Classification results for the Chinese syllables, consonants, vowels, and tones of participants

The participants were dichotomized using different Chinese syllables, consonants, vowels, and tones with thresholds of AHI = 10 and 30 events/h, and the classification effects for each are shown in **Table 2**, **Table 3**, **Table 4**, and **Table 5**, respectively, arranged in descending order for each classification effect index (including area under curve [AUC], accuracy, precision, sensitivity, and specificity). Only the 2 best and worst effects are listed for each item.

For syllables, the classification effect for [leng] and [gua] is better when the threshold of AHI = 10 events/h is used for classification, and the classification effect for [jue] and [qia] is better when the threshold of AHI = 30 events/h is used (**Table 2**). The complete Chinese syllable classification results are shown in **Table S2** in the supplemental material.

For consonants, the average classification effect for all Chinese syllables containing a certain consonant was used as the classification effect for that consonant. The classification effect for [zh] and [d] is better when the threshold of AHI = 10 events/h is used for classification, and the classification effect for [f] and [q] is better when the threshold of AHI = 30 events/h is used (**Table 3**). The complete consonant classification results are shown in **Table S3** in the supplemental material.

For vowels, the average classification effect for all Chinese syllables containing a certain vowel was used as the classification effect for that vowel. The classification effect for [ing] and [in] is better when the threshold of AHI = 10 events/h is used for classification, and the classification effect for [ai] and [ia] is better when the threshold of AHI = 30 events/h is used (**Table 4**). The complete vowel classification results are shown in **Table S4** in the supplemental material.

For tones, the average classification effect for all Chinese syllables containing a certain tone was used as the classification effect for that tone. There are 4 tones in Chinese syllables, which are represented by "1", "2", "3", and "4". The classification of Chinese syllables with tones 1 and 3 is better when the threshold of AHI = 10 events/h is used for classification, and the classification of Chinese syllables with tones 2 and 1 is better when the threshold of AHI = 30 events/h is used (**Table 5**).

Table 2—Classification results for Chinese syllables of participants.

Threshold: AHI = 10 events/h				
AUC	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
leng (0.74)	xi (77.9%)	bin (86.2%)	xi (87.8%)	bin (69.2%)
gua (0.73)	qiu (76.7%)	qu (84.4%)	ni (87.0%)	ban (61.0%)
……	……	……	……	……
cuo (0.47)	ke (46.4%)	sen (64.8%)	suan (57.3%)	hu (5%)
sa (0.46)	bang (43.5%)	bang (44.4%)	bang (51.3%)	tong (5%)
Threshold: AHI = 30 events/h				
AUC	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
jue (0.70)	yin (73.9%)	hong (74.3%)	yin (76.9%)	cha (72.5%)
qia (0.70)	cha (71.4%)	cha (72.5%)	dui (74.7%)	cui (72.0%)
……	……	……	……	……
se (0.48)	zhuai (39.0%)	kuan (37.5%)	kuan (30.9%)	ruo (33.8%)
guai (0.47)	kuan (37.4%)	zi (37.3%)	shei (30.9%)	luo (29.9%)

The classification effects are arranged in descending order. Only the 2 best and worst effects are listed for each item. The complete Chinese syllable classification results are shown in **Table S2**. AHI = apnea-hypopnea index, AUC = area under curve.

Table 3—Classification results for consonants of participants.

Threshold: AHI = 10 events/h				
AUC	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
zh (0.71)	sh (67.5%)	sh (85.9%)	x (87.8%)	sh (67.5%)
d (0.70)	h (66.2%)	h (84.8%)	n (87.0%)	q (65.0%)
……	……	……	……	……
r (0.49)	b (46.4%)	r (74.5%)	s (57.3%)	f (36.6%)
k (0.44)	s (43.5%)	s (70.2%)	b (51.3%)	r (32.5%)
Threshold: AHI = 30 events/h				
AUC	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
f (0.68)	n (69.8%)	x (72.1%)	n (71.4%)	h (72.2%)
q (0.68)	x (69.6%)	h (71.0%)	b (69.6%)	x (71.2%)
……	……	……	……	……
zh (0.51)	f (50.3%)	k (51.4%)	sh (44.3%)	zh (50.0%)
p (0.49)	k (50.0%)	f (51.2%)	m (44.2%)	f (45.9%)

The classification effects are arranged in descending order. Only the 2 best and worst effects are listed for each item. The complete consonant classification results are shown in **Table S3**. AHI = apnea-hypopnea index, AUC = area under curve.

Construction of OSA classification model

To construct a robust and effective OSA classification model, 10 syllables with highest AUC values are combined for OSA classification. For the group taking AHI = 10 as the threshold, the 10 syllables with highest AUC values are [leng], [gua], [ding], [ba], [lue], [ting], [du], [pu], [qiong], and [xue]. The receiver operating characteristic curve is shown in **Figure 3A**, with an AUC of 0.83. For the group using AHI = 30 as the threshold, the 10 syllables with highest AUC values include [jue], [qia], [zhui], [yu], [fu], [li], [wei], [ta], [jiu], and [xing]. The receiver operating characteristic curve is shown in **Figure 3B**, with an AUC of 0.87. The complete classification results of the model are shown in **Table 6**.

DISCUSSION

To the authors' knowledge, this is the first time that a comprehensive screening of pronunciations has been conducted to assess OSA. It was found that the classification effect for different pronunciations used to assess OSA had obvious differences. Chinese syllables such as [leng] and [jue], consonant pronunciations such

Table 4—Classification results for vowels of participants.

Threshold: AHI = 10 events/h				
AUC	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
ing (0.73)	ian (68.5%)	iu (86.1%)	ian (75.3%)	ü (75.0%)
in (0.72)	ei (66.7%)	ü (85.5%)	a (70.1%)	iu (73.2%)
……	……	……	……	……
o (0.43)	ia (49.0%)	uan (71.4%)	ue (49.1%)	ou (39.0%)
uai (0.35)	uan (47.1%)	uai (71.4%)	uan (47.4%)	un (34.1%)
Threshold: AHI = 30 events/h				
AUC	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
ai (0.72)	ang (67.2%)	ang (70.5%)	ü (71.8%)	ang (73.1%)
ia (0.67)	in (67.1%)	in (67.1%)	iu (70.2%)	eng (70.8%)
……	……	……	……	……
ou (0.47)	ian (52.0%)	o (51.1%)	u (47.3%)	iang (46.7%)
uai (0.40)	ei (50.4%)	ei (50.6%)	o (43.3%)	ei (40.0%)

The classification effects are arranged in descending order. Only the 2 best and worst effects are listed for each item. The complete vowel classification results are shown in **Table S4**. AHI = apnea-hypopnea index, AUC = area under curve.

Table 5—Classification results for tones of participants.

Threshold: AHI = 10 events/h				
AUC	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
1 (0.62)	2 (69.6%)	3 (84.9%)	2 (71.8%)	3 (65.9%)
3 (0.61)	3 (67.9%)	2 (84.8%)	4 (69.2%)	2 (63.4%)
2 (0.56)	4 (65.0%)	1 (81.6%)	3 (68.7%)	1 (53.8%)
4 (0.55)	1 (64.7%)	4 (81.0%)	1 (68.4%)	4 (52.5%)
Threshold: AHI = 30 events/h				
AUC	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**
2 (0.68)	1 (68.8%)	1 (70.3%)	2 (68.5%)	1 (71.1%)
1 (0.67)	2 (67.1%)	2 (66.7%)	1 (66.7%)	2 (65.8%)
4 (0.65)	3 (62.0%)	3 (60.8%)	3 (64.9%)	3 (59.2%)
3 (0.60)	4 (60.1%)	4 (59.0%)	4 (64.5%)	4 (55.8%)

There are 4 tones in Chinese syllables, which are represented by "1", "2", "3", and "4". AHI = apnea-hypopnea index, AUC = area under curve.

as [zh] and [f], and vowel pronunciations such as [ing] and [ai] had the best classification effects. An OSA classification model consisting of several syllable combinations was constructed, with AUCs of 0.83 (threshold of AHI = 10) and 0.87 (threshold of AHI = 30), respectively, confirming that the speech signal had desirable efficacy when used for OSA assessment.

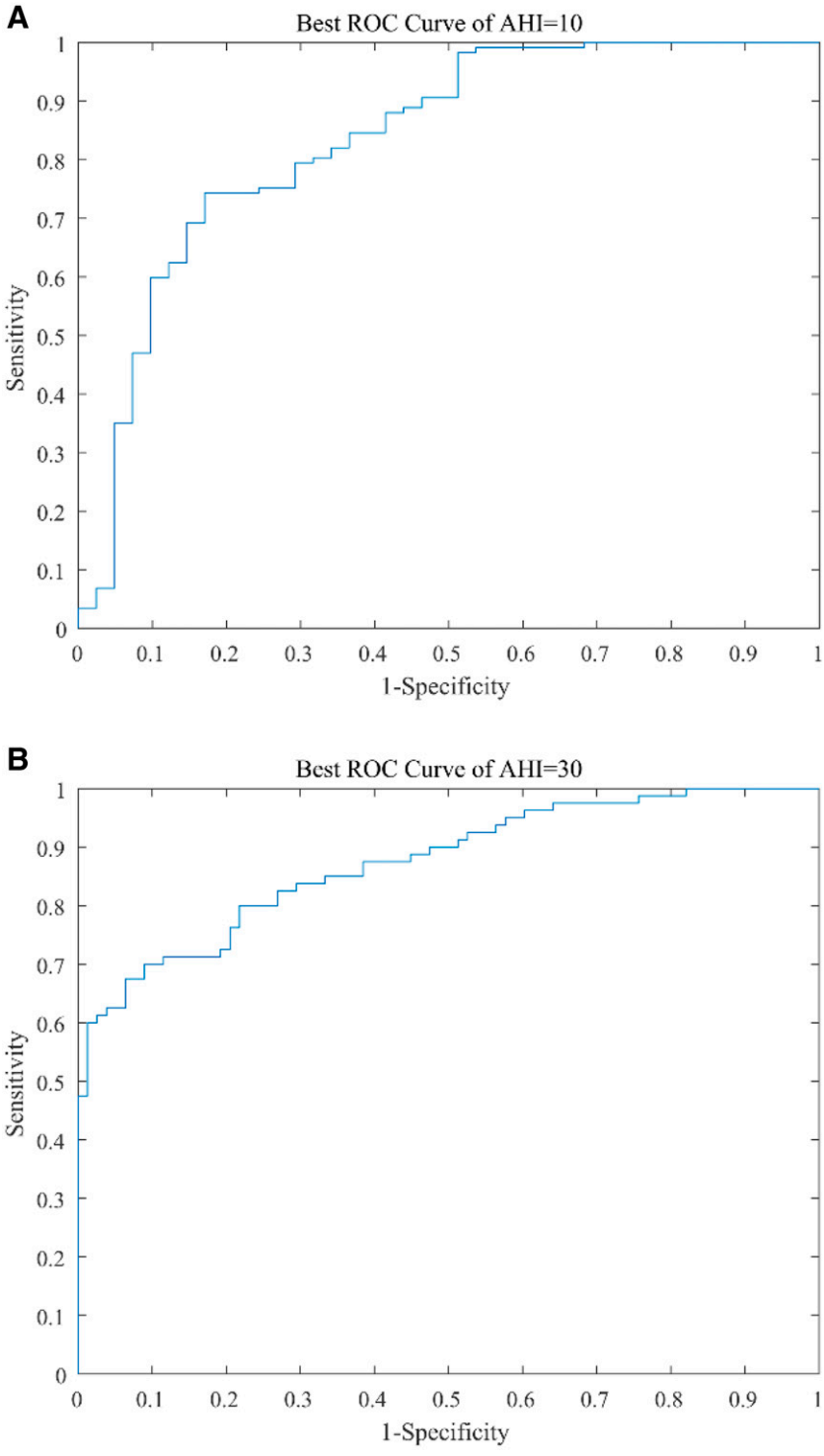
In 1989, Fox et al confirmed that the speech signal in patients with OSA was different from that in healthy individuals,[8] and since then, further research has been carried out. However, many of the earlier studies selected 2 groups of participants with large differences in AHI for comparative analysis. Although producing better results than previous samples, these groups could not be used for comprehensive assessment of OSA as they did not include cases with AHI between 10 and 30 events/h.[17–19] In our study, all participants were dichotomized 2 times with the AHI = 10 events/h classification to initially determine whether participants had OSA, and then again with the AHI = 30 events/h classification to determine whether participants had severe OSA, thus allowing the model to assess the severity of disease more comprehensively in patients with OSA.

Previous OSA-related speech studies usually selected continuous vowels and sentences as the speech corpus, and the vowels selected, /i/, /e/, /a/, /o/, /u/, were considered able to cover comprehensively the speech signals of participants.[20,21] However, using 5 vowels is far from sufficient, and a variety of combinations can arise between the vowels, in addition to a large number of consonant pronunciations that were not included. Regarding the selection of sentences, Pozo et al first selected 4 Spanish sentences,[10] which were widely used in subsequent studies, and their choice of utterances contained consecutive vowels /a/ and /i/, nasals and nonnasals, and a certain number of voiced sounds, speech features that were considered to differ between OSA and non-OSA individuals. Even so, that study did not implement a comprehensive screening and discussion of OSA-specific pronunciations, and it is impossible to be sure that the selected corpus contained all OSA-specific speech signals. Our study was based on Chinese pronunciation, and syllables containing all vowel-consonant combinations were used as the speech corpus. We found that different pronunciations had very different effects on OSA classification, and that a selection of syllables, vowels, and consonants gave better results for OSA classification, and thus they would be useful for the construction of OSA-specific corpora in other languages.

Interestingly, we found that comparing the results for AHI = 10 and AHI = 30 events/h, the syllables, vowels, consonants, and tones offering the best classification effect were different. For example, when AHI = 10 events/h, the syllable with the best classification effect is [leng], while when AHI = 30 events/h, the syllable with the best classification effect was [jue], and even the top 10 syllables with the best classification effect were not the same. It is speculated that the degree of speech abnormalities between the 2 groups would be unequal when different AHI thresholds are used. Most of the participants with AHI < 10 events/h were healthy people, and the degree of abnormal speech was low. However, there were many patients with OSA among the participants with AHI > 10 events/h. The degree of abnormal speech was very different from that in participants with AHI < 10 events/h. Nevertheless, there were a certain number of patients with OSA among the participants with AHI < 30 events/h, and the difference in the degree of speech abnormality was relatively small compared with participants with AHI > 30 events/h. This led to differences with those syllables, vowels, and consonants having good classification results when the thresholds AHI = 10 and AHI = 30 events/h were considered. Therefore, we believe that the best speech corpus may be different for different levels of OSA classification, and it may be difficult to classify OSA in different levels using a unified speech corpus. Therefore, the best corpus to evaluate the degree of OSA should be selected according to the setting.

All Chinese syllables are composed of consonants and vowels. For example, the syllable [ding] is composed of consonant [d] and vowel [ing]. In our study, comparing the classification results for syllables with consonants and vowels, we found that the syllables with the best classification effect were not composed of the consonants and vowels with the best classification effect. For example, when AHI = 10 events/h was used for

Figure 3—The ROC curves for syllables combination of participants.



Ten syllables with highest AUC values were combined for OSA classification. For the group taking AHI = 10 as the threshold, the ROC curve is shown in (**A**), with an AUC of 0.83. For the group taking AHI = 30 as the threshold, the ROC curve is shown in (**B**), with an AUC of 0.87. AHI = apnea-hypopnea index, AUC = area under curve, OSA = obstructive sleep apnea, ROC = receiver operating characteristic.

classification, the syllable with the best classification effect was [xi], while the consonants and vowels with the best classification effect were [sh] and [ian], respectively, which did not match. Indeed, this was the case for the top 10 syllables, consonants, and vowels. The reason for this may be that, in the process of syllable pronunciation, the structure of the upper airway is constantly changing due to the actions of the lips, tongue, and surrounding muscles. In this process, speech signals other than the consonants

Table 6—Classification results for syllables combination of participants.

Classification Effect Index	Threshold: AHI = 10 events/h	Threshold: AHI = 30 events/h
AUC	0.83	0.87
Accuracy	81.7%	80.3%
Sensitivity	81.8%	78.1%
Specificity	81.3%	82.6%
Precision	93.1%	82.6%

AHI = apnea-hypopnea index, AUC = area under curve.

and vowels in the syllable may be produced. Due to the complexity of the speech formation and pronunciation processes, how to match different speech signals to their corresponding upper airway states requires further study.

Limitations

Since this study used all of the Chinese syllables as the corpus to record the participants, the process was time consuming, and thus single-sitting recordings were made in the seated position, because the participants had difficulty in cooperating with multiple recordings and different positions. In future studies, the corpus will be filtered and streamlined to allow for multiple recordings in multiple positions. Due to the small number of female and pediatric participants, it was difficult to construct a stable prediction model, so only adult male participants were included in this study. In future studies, an expanded sample size is expected to be included in the analysis to explore female- and pediatric-specific speech in patients with OSA. Due to the limited sample size and the small number of participants with AHI < 5 events/h, we selected the same AHI = 10 and 30 events/h cut-offs for classification as in most earlier studies to facilitate comparison of results. The next step is to expand the sample size to allow for dichotomous and multicategory studies with multiple AHI cut-offs. This study is based only on Chinese pronunciation, so its generalizability is limited, but it can act as a reference for other languages with similar pronunciations.

CONCLUSIONS

In this study, a comprehensive screening of pronunciation was conducted in patients with OSA for the first time. Several characteristic pronunciations were identified as more effective for OSA classification than those used previously, which will be helpful in improving the evaluation of speech signals in patients with OSA. In order to obtain more accurate classification results, a speech corpus should be selected to contain speech signals that can classify different degrees of OSA.

ABBREVIATIONS

AHI, apnea-hypopnea index
AUC, area under curve
LPC, linear prediction coefficients
OSA, obstructive sleep apnea
PSG, polysomnography

REFERENCES

1. Somers VK, White DP, Amin R, et al. Sleep apnea and cardiovascular disease: an American Heart Association/American College of Cardiology Foundation Scientific Statement from the American Heart Association Council for High Blood Pressure Research Professional Education Committee, Council on Clinical Cardiology, Stroke Council, and Council on Cardiovascular Nursing. *J Am Coll Cardiol.* 2008; 52(8):686–717.
2. Tasali E, Ip MSM. Obstructive sleep apnea and metabolic syndrome: alterations in glucose metabolism and inflammation. *Proc Am Thorac Soc.* 2008;5(2):207–217.
3. Veasey SC, Rosen IM. Obstructive sleep apnea in adults. *N Engl J Med.* 2019; 380(15):1442–1449.
4. George CFP. Reduction in motor vehicle collisions following treatment of sleep apnoea with nasal CPAP. *Thorax.* 2001;56(7):508–512.
5. Morsy NE, Farrag NS, Zaki NFW, et al. Obstructive sleep apnea: personal, societal, public health, and legal implications. *Rev Environ Health.* 2019;34(2): 153–169.
6. Benjafield AV, Ayas NT, Eastwood PR, et al. Estimation of the global prevalence and burden of obstructive sleep apnoea: a literature-based analysis. *Lancet Respir Med.* 2019;7(8):687–698.
7. Young T, Palta M, Dempsey J, Peppard PE, Nieto FJ, Hla KM. Burden of sleep apnea: rationale, design, and major findings of the Wisconsin Sleep Cohort study. *WMJ.* 2009;108(5):246–249.
8. Chung F, Yegneswaran B, Liao P, et al. STOP questionnaire: a tool to screen patients for obstructive sleep apnea. *Anesthesiology.* 2008;108(5): 812–821.
9. Fox AW, Monoson PK, Morgan CD. Speech dysfunction of obstructive sleep apnea. A discriminant analysis of its descriptors. *Chest.* 1989;96(3): 589–595.
10. Pozo RF, Murillo JLB, Gómez LH, Gonzalo EL, Ramirez JA, Toledano DT. Assessment of severe apnoea through voice analysis, automatic speech, and speaker recognition techniques. *J Adv Signal Process.* 2009:982531.
11. Benavides AM, Blanco JL, Fernández A, et al. Using HMM to detect speakers with severe obstructive sleep apnoea syndrome. In: Torre Toledano D et al, eds. *Advances in Speech and Language Technologies for Iberian Languages.* Berlin: Springer; 2012:121–128.
12. Espinoza-Cuadros F, Fernández-Pozo R, Toledano DT, Alcázar-Ramírez JD, López-Gonzalo E, Hernández-Gómez LA. Reviewing the connection between speech and obstructive sleep apnea. *Biomed Eng Online.* 2016; 15(1):20.
13. Kriboy M, Tarasiuk A, Zigel Y. Obstructive sleep apnea detection using speech signals. *Proceedings of the Annual Conference of the Afeka-AVIOS in Speech Processing.* 2013 https://events.eventact.com/afeka/aclp2012/Obstructive%20Sleep%20Apnea_Kriboy%20et%20al.pdf; accessed December 12, 2021.
14. Pang KG, Hsung TC, Law AKW, et al. Optimal vowels measurements for obstructive sleep apnea detection using speech signals. *2020 IEEE 3rd International Conference on Information Communication and Signal Processing* (ICICSP). 2020:143–147.
15. Ding Y, Wang J, Gao J, et al. Severity evaluation of obstructive sleep apnea based on speech features [published online ahead of print, 2020 Oct 27]. *Sleep Breath.*
16. Berry RB, Budhiraja R, Gottlieb DJ, et al; Deliberations of the Sleep Apnea Definitions Task Force of the American Academy of Sleep Medicine. Rules for scoring respiratory events in sleep: update of the 2007 AASM Manual for the Scoring of Sleep and Associated Events. *J Clin Sleep Med.* 2012;8(5):597–619.
17. Perero-Codosero JM, Espinoza-Cuadros F, Antón-Martín J, et al. Modeling obstructive sleep apnea voices using deep neural network embeddings and domain-adversarial training. *IEEE J Sel Top Signal Process.* 2020;14(2): 240–250.

18. Montero Benavides AM, Pozo RF, Toledano DT, et al. Analysis of voice features related to obstructive sleep apnoea and their application in diagnosis support. *Comput Speech Lang.* 2014;28(2):434–452.
19. Blanco JL, Hernández LA, Fernández R, et al. Improving automatic detection of obstructive sleep apnea through nonlinear analysis of sustained speech. *Cognit Comput.* 2013;5(4):458–472.
20. Solé-Casals J, Munteanu C, Martín OC, et al. Detection of severe obstructive sleep apnea through voice analysis. *Appl Soft Comput.* 2014; 23:346–354.
21. Montero Benavides A, Blanco Murillo JL, Fernández Pozo R, et al. Formant frequencies and bandwidths in relation to clinical variables in an obstructive sleep apnea population. *J Voice.* 2016;30(1):21–29.

ACKNOWLEDGMENTS

The authors are grateful for the contributions of the study participants and otolaryngologists and technologists at the Department of Otolaryngology Head and Neck Surgery, Beijing Tongren Hospital.

SUBMISSION & CORRESPONDENCE INFORMATION

Submitted for publication June 6, 2021
Submitted in final revised form November 17, 2021
Accepted for publication November 18, 2021
Address correspondence to: Demin Han, MD, PhD, Beijing Tongren Hospital, Capital Medical University, 1 Dongjiaominxiang, Dongcheng District, Beijing, People's Republic of China; Email: deminhan_ent@hotmail.com; and Jiandong Gao, PhD, Room 8301, Luomu Building, Tsinghua University, Haidian District, Beijing, People's Republic of China; Email: jdgao@tsinghua.edu.cn

DISCLOSURE STATEMENT

All authors have seen and approved this manuscript. This study was funded by the National Key Research & Development Program of China (2018YFC0116800), the National Natural Science Foundation of China (81970866), Beijing Municipal Administration of Hospitals' Youth Programme (QMS20190202), and Beijing Municipal Natural Science Foundation (L192026). The authors report no conflicts of interest.