"OK, Siri" or "Hey, Google": Evaluating Voiceprint Distinctiveness via Content-based PROLE Score

Ruiwen He, Xiaoyu Ji, and Xinfeng Li, *Zhejiang University;*
Yushi Cheng, *Tsinghua University;* Wenyuan Xu, *Zhejiang University*

https://www.usenix.org/conference/usenixsecurity22/presentation/he-ruiwen

This paper is included in the Proceedings of the
31st USENIX Security Symposium.

August 10–12, 2022 • Boston, MA, USA

978-1-939133-31-1

Open access to the Proceedings of the
31st USENIX Security Symposium is
sponsored by USENIX.

"OK, Siri" or "Hey, Google": Evaluating Voiceprint Distinctiveness via Content-based PROLE Score

Ruiwen He[1], Xiaoyu Ji[1]*, Xinfeng Li[1], Yushi Cheng[2], and Wenyuan Xu[1]
[1]*Zhejiang University,* [2]*Tsinghua University*
{rwhe97, xji, xinfengli, yushicheng, wyxu}@zju.edu.cn

Abstract

A voiceprint is the distinctive pattern of human voices that is spectrographically produced and has been widely used for authentication in the voice assistants. This paper investigates the impact of speech contents on the distinctiveness of voiceprint, and has obtained answers to three questions by studying 2457 speakers and 14,600,000 test samples: 1) What are the influential factors that determine the distinctiveness of voiceprints? 2) How to quantify the distinctiveness of voiceprints for given words, e.g., wake-up words in commercial voice assistants? 3) How to construct wake-up words whose voiceprints have high distinctiveness levels. To answer those questions, we break down voiceprint into phones, and experimentally obtain the correlation between the false recognition rates and the richness of the phone types, the order, the length, and the elements of the phones. Then, we define PROLE Score that can be easily calculated based on speech content yet can reflect the voice distinctiveness. Under the guidance of PROLE Score, we tested 30 wake-up words of 19 commercial voice assistant products, e.g., "Hey, Siri", "OK, Google" and "Nihao, Xiaona" in both English and Chinese. Finally, we provide recommendations for both users and manufacturers, on selecting secure voiceprint words.

1 Introduction

A voiceprint is a measurable characteristic of human voices that can uniquely identify a person. It is the key biometric in automatic speaker verification (ASV) systems and is used for personal asset access, financial transactions, and even criminal investigation [1–4], e.g., utilizing the voiceprint embedded in wake-up words to activate an Amazon Echo [5], as a password to access a TD bank account [6], or for caller identification in telecommunication fraud cases [7]. Given that voiceprints are increasingly used in sensitive scenarios, it is critical to understand how much we can trust state-of-the-art voiceprint technologies, which commercial ASV system outperforms others in terms of speech contents, and last but not least, what

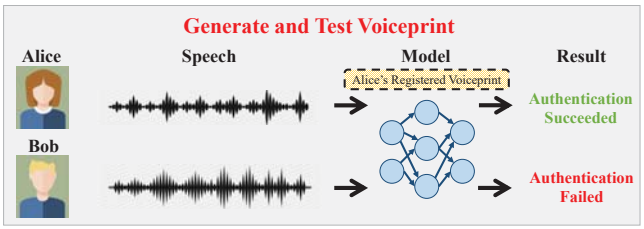


Figure 1: The workflow of a voiceprint ASV system. The distinctiveness of voiceprint is affected by the speaker, the speech contents, and the verification model.

words to say to improve the accuracy of the ASV. The key to answering the aforementioned questions is to define a metric that can quantify the distinctiveness of voiceprints.

A voiceprint is essentially a spectrographically produced mathematical representation of human voices, and its distinctiveness has four contributors: (1) intrinsic physiological features of a speaker, e.g., the vocal tract shape, (2) speech contents, e.g., words with various phonetic alphabets, (3) ASV models that are used to derive and compare voiceprints, and (4) the environment and equipment factors, e.g., environmental noises, acoustic channel responses, or variance caused by device properties, as shown in Figure 1. Instead of studying all influential factors, we study the distinctiveness metric by analyzing the impact of speech contents, complementary to existing research on voiceprint security that focused on performance optimization of models [8–14], attack strategy and defense countermeasures against spoofing, synthesizing, voice conversion, and adversarial samples[1, 15–18].

We aim to define a PROLE Score[†] such that the voiceprint distinctiveness can be directly calculated from the speech contents, yet it reflects the distinctiveness of the voiceprints associated with ASV models[‖]. From a linguistics perspective, a speech is a sequence of phones, i.e., the basic unit of phonetic speech analysis [19], and can be modeled by four phonetic factors: the richness (the number of phone types), the order, the length, and the elements of phones. An ideal PROLE Score definition shall reflect the underlying relation-

*Xiaoyu Ji is the corresponding author.

[†]PROLE: phonetic richness, order, length, elements.

[‖]Online PROLE Score: https://github.com/USSLab/PROLE-Score.

ship between the phonetic factors and voiceprints, yet capturing it requires a comprehensive dataset with an unbiased combination of phonetic factors. Such a dataset is not only unavailable but also difficult to generate due to the limitation of languages: both the number of words and the combination of phones are limited. We overcame the challenges by segmenting 1000-hour utterances of 2457 speakers in two English speech datasets into phone clips and constructing a dataset of 14,600,000 audio samples. To capture the relationship, we selected three representative ASV models, i.e., i-vector [20], x-vector [21], and the end-to-end U-LEVEL model [22], and evaluated the false recognition rates using the constructed dataset. In total, we executed verification tests more than 107 billion times, and the main observations are listed below:

- **O1:** Among all phonetic factors, the richness, the element, and the length of phones can affect the ASV performance, and their influence levels decrease in order. Interestingly, the order of phones has almost no impact on the voiceprint verification.
- **O2:** The increase of phone richness and length will reduce the false recognition rate, yet the improvement saturates once a speech includes more than 15 phones and 4 types of phones.
- **O3:** The element of phones matters. For instance, phone [**i**] is better than phone [**o**], and the false recognition rate of 20 [**i**]s is 2.43 times less than that of 20 [**o**]s.

We define `PROLE Score` to be linearly inversely proportional to the false recognition rate of a given word, ranging from 0 to 10, and we envision that `PROLE Score` is applicable to both content dependent and independent voiceprint systems. For instance, existing voice assistants (e.g., Google Assistant) utilize content-dependent ASV, and extract users' voiceprints by analyzing the pre-defined wake-up words, e.g., "OK, Google". A `PROLE Score` can help to evaluate such wake-up words and answer the following questions:

- **Q1: Interjection**. Is adding meaningless interjections before wake-up words helpful? For example, is "OK Google" better than "Google'?
- **Q2: Repetition**. Will a wake-up word repetition improve voiceprint distinctiveness? Is "Alexa Alexa" better than "Alexa"?
- **Q3: Comparison**. How to evaluate wake-up words of existing commercial ASV systems? Is "Hey Cortana" more distinct than "Hey Google"?

Without loss of generality, this paper evaluated content-independent voiceprint and analyzed 30 wake-up words from 19 main-stream commercial voice assistant products, in both English and Chinese. Our study confirms the choice of existing commercial wake-up words and answers the above three questions: **A1:** An interjection is useful, especially when the richness and the length of a wake-up word are below the thresholds. Take i-vector as an example, turning "Google" to "OK Google" can reduce the false recognition rate by 10%. **A2:** Repetition can improve the voiceprint performance if the length of a wake-up word is short. For example, repeating "Alexa" once reduces the false recognition rate by 10%, and twice by 14% under the i-vector model. **A3:** Words matter. For instance, "Hey Cortana" outperforms "Hey Google" for all the three voiceprint verification models. To further validate the effectiveness of the `PROLE Score`, we conducted a user study with 40 volunteers, and tested the 30 wake-up words using a third-party ASV system. The measured false recognition rates for each wake-up word match the calculated `PROLE Score` well. Thus, we believe `PROLE Score` can help both manufactures and users to refine their choices of words to be used for extracting voiceprint. In summary, the contributions of the paper are as follows:

- We analyzed the correlation between speech contents and false recognition rates using three representative ASV models, with two datasets of 2457 speakers and 14,600,000 test samples.
- We defined the `PROLE Score` that can quantify the distinctiveness of voiceprints for any given speech content.
- We recommended words that can produce good voiceprint distinctiveness to both manufacturers and users, e.g., good wake-up word candidates.

2 Background and Threat Model

2.1 ASV and Voiceprint

Automatic speaker verification (ASV) systems utilize the voiceprint for user authentication. The workflow of an ASV system is shown in Fig.1. A typical ASV system involves three stages: training, enrollment, and testing. In the training stage, a large number of speech audios with speaker labels are fed into the verification model for parameter optimization. In the enrollment stage, a speaker, e.g., Alice, speaks a phrase several times to register her voiceprint in the verification model. Finally, in the test stage, the ASV system should deny all other unregistered users while accepting Alice by authenticating Alice with her new audio inputs. As a result, the verification accuracy of a given ASV system is affected by three factors: the speaker, the speech content, and the verification model.

Commonly-used verification models include two types: (1) text-dependent models, and (2) text-independent models. Most existing commercial voice assistant products, e.g, Apple Siri, Amazon Alexa, or Google Assistant, employ the text-dependent model [23] for its superior performance, and authenticate users as they speak pre-defined wake-up words, e.g., "Hey, Siri" or"OK, Google". However, text-dependent models require the enrollment and testing sentences to be the same while text-independent models have no such restriction on the speech contents. Therefore, it is a trend to use text-independent models for speaker recognition [24]. Typical text-independent models include: (1) classic models such as i-vector [25–27], (2) DNN-based ones such as

x-vectors [9, 21, 28], and (3) advanced end-to-end DNN models [10, 24, 29] such as U-LEVEL [22]. We study these representative models i.e., i-vector, x-vectors, and U-LEVEL, in this paper.

2.2 Threat Model

In this paper, we consider the following attack scenario: *An adversary aims to illegally access a voice assistant such as Siri, Amazon Echo, etc., which however is protected by speaker verification system. To break it, the adversary imitates the voice of the victim and repeats the wake-up words or other keywords to spoof the speaker verification system.*

In this case, the goal of the ASV system is to identify the registered speaker while rejecting the illegal ones. Thus, the ASV system shall (1) select distinct wake-up words or keywords, (2) design the appropriate embedding that represents the voiceprint, and (3) improve the accuracy of the classifier. The latter two are usually improved by optimizing the voiceprint model. In this paper, we study how to enhance the security of the ASV system by selecting appropriate speech contents, e.g., wake-up words or keywords.

3 Measurement Methodology

To quantify the distinctiveness of voiceprints in terms of speech contents, we look into the basic unit of speech content, i.e., phones. However, existing public datasets for speaker verification are usually long audio clips without dedicated phonetic information. To overcome this challenge and analyze the correlation between voiceprints and phonetic factors, we design the following measurement methodology with 4 stages, as shown in Fig. 2. First, we dissect a speech into phones and propose four intrinsic phonetic factors to represent any speech content (section 3.1). Second, we derive seven test variables based on the four phonetic factors plus three combinations of two factors (section 3.2). Then, we construct the dedicated test datasets for each test variable by segmenting and reassembling audio clips (section 3.3). Finally, we feed the constructed test datasets into pre-trained models (section 3.4) and measure the impacts of speech contents using two metrics (section 3.5).

3.1 Speech Content Analysis

According to linguistics[30], the basic unit of a given speech is a phone and a speech is a sequence of phones. Different from a phoneme which is a speech sound in a given language, a phone is language-independent and thus more suitable for speech content analysis. There are 107 phones across all the languages and English has 48 of them. To quantitatively model speech contents in terms of phones, we propose four phonetic factors:

- ❶ **Richness:** the number of phone types.
- ❷ **Length:** the number of phones.
- ❸ **Element:** the specific type of phones.
- ❹ **Order:** the sequential relationship among phones.

By varying the test variables derived from these four phonetic factors, we measure the impact of speech contents on voiceprint distinctiveness.

3.2 Test Variable Design

To validate the influence of speech contents, we design seven test variables based on the aforementioned four phonetic factors as well as the combination of two factors (i.e., richness, length and element) as they have mutual influences upon each other in practice:

- ❶ **Richness**. This variable explores the impact of the number of phone types while for each level of richness, the lengths of audio samples are uniformly distributed.
- ❷ **Length.** This variable studies the impact of the number of phones, and for each length, the phone types are uniformly distributed.
- ❸ **Element.** This variable explores the impact of each phone when the length is fixed. We cycle through each phone, and construct audios sample by repeating it.
- ❹ **Order.** This variable explores the impact of the sequence of phones, while the length and the element are fixed.
- ❺ **Synergies of Richness and Length** (❶ + ❷). This setup explores the impact of richness and length, as both vary simultaneously.
- ❻ **Synergies of Length and Element** (❷ + ❸). Similar to ❺, these two factors vary simultaneously, and we construct audio samples by repeating a phone up to the required length.
- ❼ **Synergies of Richness and Element** (❶ + ❸). This combination is meant to investigate the impact of repetition and combination. For instance, will ***A+A*** or will ***A+B*** improve the distinctness, given that ***A*** and ***B*** contain different levels of distinctiveness.

With the 7 test variables, we then construct corresponding test datasets for measurement.

3.3 Test Dataset Construction

To construct test datasets for each variable, we first select appropriate speech datasets to reduce the biases caused by speakers, then prepare the audios by pre-processing, and finally segment and reassemble the processed audios to get the final test datasets.

Dataset Selection. As mentioned, the speaker is a subjective factor that may affect the voiceprint security. To reduce its impact, we choose two popular English speech datasets, i.e., VCTK [31] and LibriSpeech [32], which contain audios from 2,457 speakers (45 from VCTK and 2,412 from LibriSpeech) with various genders, speech rates, channel noises, etc. The average speech duration of each speaker is more

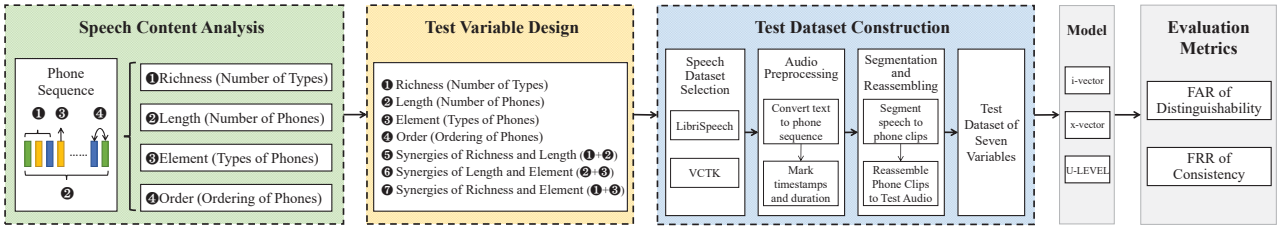


Figure 2: Measurement methodology overview. First, we regarded the phone as the atomic unit of speech content and propose four phonetic factors, i.e., richness, length, element and order. Then, we used seven test variables from the four factors considering the synergies of multiple factors. Under each variable, we constructed its dataset from two popular open speech datasets. Finally, each test variable was validated by calculating the corresponding false recognition rate, i.e., the sum of false acceptance rate and false rejection rate under a specific ASV model.

than 30 minutes in VCTK and more than 20 minutes in LibriSpeech respectively, providing sufficient test samples. We hope these two speech datasets can reduce the biases caused by unbalanced speakers or insufficient test samples.

Audio Pre-processing. With the selected datasets, we preprocess the speech audios. First, we extract the texts corresponding to the speech contents from the datasets. According to the International Phonetic Alphabet (IPA), there are 48 types of phones in English, including 16 vowels, 1 rhoticity vowel, 5 diphthongs, 1 triple vowel, 22 consonants, 1 coarticulated consonant, and 2 affricates. Among these phones, diphthongs and triple vowel are composed of several vowels, and affricate is composed of several consonants. Then, in line with the IPA, we employ the Phonemizer tool [33] and the G2P tool [34] to convert the texts into phone sequences, and employ the Montreal-Forced-Aligner (MFA) tool [35] to calculate the timestamp and the duration of each phone in the speech audios. By statistical analysis, the accuracy of the MFA aligner is 92.4%, and we will discuss the impact of MFA errors on the phonetic factors in Sec. 7.

Audio Segmentation and Reassembling. Based on the results of MFA, we segment the original audios in the open datasets into audios in the unit of phone. Then, We randomly select phone audios to avoid bias and reassemble them to form test datasets designed for the seven variables.

Datasets for ❶ ❷ ❺: For these three datasets, we vary the value of *length*, *richness*, or the $(length, richness)$ pair respectively while keeping the other phonetic factors evenly-distributed to construct the test datasets. Since the diphthong and the affricate consist of 2 single phones and the triple vowel consists of 3 single phones, we count their richness as 2, 2, and 3, respectively. Thus, in these three datasets, we have a total of 40 types of phones.

Datasets for ❸ ❻: For these two datasets, we repeat each phone to reach a specific length to construct various test samples. The difference lies in that the length used in the Dataset for ❸ is a fixed large one, while in the Dataset for ❻, it is a variable one in the range of [1, 20]. Since the triple vowel [aɪʊ] is infrequent in speakers' speech contents, and the vowels [a] and [e] do not appear alone in English, we do not consider these phones in these two datasets.

Datasets for ❹ ❼: For these two datasets, we first construct a group of high-distinctiveness phones (denoted as ***A***) and a group of low-distinctiveness phones (denoted as ***B***) based on the experimental results of ❸ and ❻. Then, to construct the Dataset for ❹, we randomly select an ***A*** and a ***B***, repeat both of them for several times, and change their orders to construct various test samples. For the Dataset of ❼, we construct three types of test samples, i.e., ***AAAA***, ***BBBB*** and ***ABAB***. The used phones ***A*** and ***B*** are also randomly selected from the aforementioned groups.

3.4 Model Selection

The constructed test datasets are then fed into voiceprint verification models to evaluate their impacts. Similar to the speaker, the model is another factor that affects the voiceprint distinctiveness, yet is provided by manufacturers and thus can not be controlled by users. Without loss of generality, we study three representative voiceprint models in the paper: (1) i-vector [20, 25], (2) x-vectors [9, 21, 36], and (3) an advanced end-to-end DNN model, i.e., U-LEVEL [22]. The first two models are most commonly used in speaker verification at present, while the third one represents the state-of-the-art performance. All three models have different feature extraction methods and network structures. We model these impact factors with a global variable 'model type'. Models are trained beforehand (in Sec. 4.1) and demonstrated usabilities. The goal of using multiple models is two-fold: (1) reducing the biases of measurement results, and (2) exploring the consistency and diversity of measurement results across models. Of course, more models can be supported and currently we take the three models as examples to derive the distinctiveness of voiceprints from phones.

3.5 Evaluation Metrics

We define the evaluation metric as false recognition rate, which is the sum of (1) False Acceptance Rate of distinguishability (FAR), and (2) False Rejection Rate of consistency (FRR). The former refers to the probability that any illegitimate person's voiceprint in the test dataset is incorrectly verified as a legitimate one. The latter refers to the probability that any other speaker is classified to be a given registered

Table 1: Experiment Setup

	Model	Training Set	Test Set	Performance (EER)
Model	i-vector	85 speakers in VCTK; over 3,400 utterances	40 speakers in VoxCeleb	5.446%
	x-vector	over 7,300 speakers in VoxCeleb; over 1.2 million utterances		3.13%
	U-LEVEL	over 5,900 speakers in VoxCeleb; over 1 million utterances		3.22%
Test/ Enrolling Dataset	Speakers	Data Source	Enrolling Utterances	Test Utterances
	45	VCTK	5 per speaker	5946 per speaker
	2,412	LibriSpeech		

speaker. In practice, the FAR and FRR are supposed to be relatively balanced. Unbalanced FAR and FRR may lead to the bias of model performance. To avoid it, we adjust the threshold of the verification model during the measurement to keep the two metrics balanced. We hope it can help derive more unbiased measurement results.

4 Experiments

4.1 Experiment Setups

Models. As mentioned in Sec. 3.4, we used three text-independent models in this paper: (1) i-vector [25], (2) x-vectors [21], and (3) U-LEVEL [22]. To get rid of the influence from ASV models, we trained and tested these models with the training and test datasets described in Tab. 1 and the performances of these models are trained to approach the state-of-the-art, with detailed Equal Error Rates (EERs) shown in Tab. 1. During the following experiments, we conduct measurements for the aforementioned test variables on each model respectively.

Enrolling Dataset. The enrolling dataset is extracted from the VCTK and LibriSpeech datasets as well but has no overlap with the test datasets. It consists of 12,285 randomly selected utterances in total, i.e., 5 utterances for each speaker. Each enrolling utterance lasts for more than 2 s, to provide sufficient information for voiceprint extraction [37].

4.2 Experiment Results

4.2.1 Influence of Richness ❶

In this experiment, we investigate how the richness (R) of the speech content affects voiceprint distinctiveness by varying the number of phone types in the test audios.

Setup. Given 40 types of single phones in English, we varied R in the range of $[1,40]$. For any R, the length (L) (i.e., the number of phones) in the test samples is uniformly distributed over the range of $[R,40]$, and we randomly selected $(41-R_0)\times 6$ phone sequences as the test samples, i.e., 6 sequences for each value of L. Thus, the test dataset of each speaker consists of 4,920 audio samples. For each speaker, FAR and FRR are calculated at each value of the variable. The final results are the aggregation of all the speakers and presented in a form of box plots as shown in Fig. 3. As the trend of the results is similar across 3 models, we show the results of U-LEVEL model in Sec. 4.2, and the results of i-vector and x-vector can be found in Appendix. A. We show the results in the same way for Sec. 4.2.2 and Sec. 4.2.4-4.2.7.

Result. From the results, we can see that as the number of phone types increases, the FRR is significantly reduced while the FAR increases slightly. This observation is consistent across models but differs slightly. Specifically, we find: (1) Richness can enhance the distinctiveness of voiceprint logarithmically but the improvement saturates after a threshold, i.e., R_{th}. (2) The saturation thresholds differ across models, i.e., R_{th} is 15 for i-vector, 12 for x-vector, and 7 for U-LEVEL, respectively.

Insights 1: The richness variable has an impact on the voiceprint distinctiveness and the enhancement by the richness improvement is approximately logarithmic.

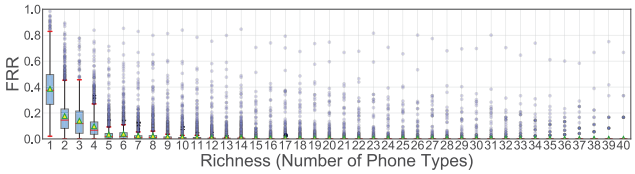
Analysis. We assume it is because the model may fail to obtain enough voiceprint information and reject all verification requests when the phone richness is low. As a result, the FRR approaches 1 while the FAR approaches 0 on the left side of the sub-graphs in Fig. 3. As the number of phones increases, the test samples carry more voiceprint information, and thus the FAR and FRR converge to equilibrium. The logarithmic changes in false recognition rates may be caused by the overlap in the voiceprint information provided by different types of phones. With high richness, adding another phone may provide limited additional information. However, the impact of phone richness will be saturated when the value of richness exceeds a threshold.

4.2.2 Influence of Length ❷

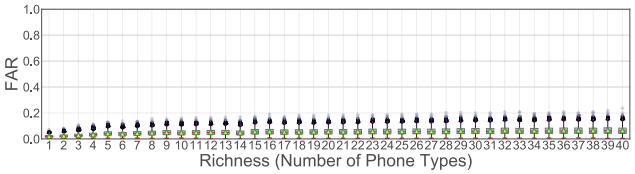
In this experiment, we investigate how the length (L) of the speech content affects voiceprint distinctiveness by varying the number of phones in the test audios.

Setup. The test dataset for length is similar to that for richness, except that the requirements of L and R are interchanged, i.e., for each L, R is uniformly distributed in the range of $[1,L]$, and the number of test samples is $L\times 6$. The results can be found in Fig. 4.

Result. The trends of the FRR and FAR under the length variation show similarity as that under the richness variation. We also find: (1) The increase of phone length improves the distinctiveness of voiceprints logarithmically with a threshold L_{th}, beyond which the distinctiveness stops raising. (2) The

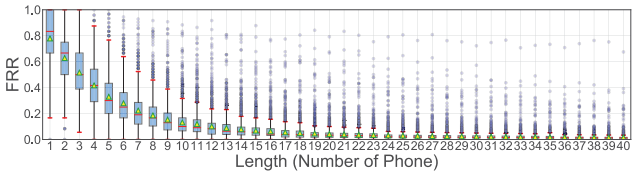


(a) FRR of Consistency (U-LEVEL)

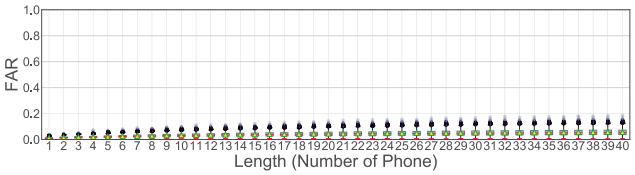


(b) FAR of Distinguishability (U-LEVEL)

Figure 3: False recognition rate vs. Richness ❶. (a) shows the FRR of consistency in U-LEVEL. (b) shows the FAR of distinguishability. FRR decreases logarithmically with an increasing R and FAR has a slight increase. (The lower and upper bound of the box are the values of the first quartile and the last quartile, the black line is the maximum and minimum, the green triangle is the mean, the red line is the median, and the blue circles mean outliers.)



(a) FRR of Consistency (U-LEVEL)



(b) FAR of Distinguishability (U-LEVEL)

Figure 4: False recognition rate vs. Length ❷. (a) shows the FRR of consistency in U-LEVEL. (b) show the FAR of distinguishability. FRR decreases logarithmically with the increase of L and FAR has only a slight increase.

thresholds differ across various models, i.e., the L_{th} is 31 for i-vector, 26 for x-vector, and 24 for U-LEVEL.

Insights 2: The length variable has an impact on the voiceprint distinctiveness and the enhancement by length improvement is approximately logarithmic.

Analysis. The reasons for Insights 2 are similar to those for Insights 1. However, the results of this experiment have higher stability values and slower convergence speed compared to the experiment of richness. These two impacts are further synergistically analyzed in Sec 4.2.5.

4.2.3 Influence of Element ❸

In this experiment, we investigate how the element (E) of the speech content affects voiceprint distinctiveness by changing the types of phones in the test audios.

Setup. The test dataset consists of 45 audio samples of the same length for each speaker, which are constructed by repeating 45 types of phones separately. Each audio sample includes 90 phones. To avoid unbalanced FAR and FRR and obtain unbiased results, we adjust the thresholds of the verification models during the experiment. Results are presented in the form of cumulative histograms in Fig. 5, where CFRR refers to the cumulative false recognition rate.

Result. From the results, we find that the distinctiveness of phones differs greatly and the difference is highly related to the model. We observe that: (1) For i-vector, consonants are slightly more distinct than vowels. For U-LEVEL, most vowels are more distinct than consonants. The trend of distinctiveness ranking for x-vector is similar to that of U-LEVEL. (2) For i-vector, plosive and non-sibilant fricative consonants are more distinct, and the difference between voiced and unvoiced consonants is not significant. For x-vector and U-LEVEL, sibilant fricative and tongue coronal consonants lack distinctiveness. For U-LEVEL, unrounded and central vowels perform better in distinctiveness while for x-vector, there is no obvious distinctiveness difference between vowels.

Insights 3: The element variable influences voiceprint distinctiveness differently across models. There are, however, *magic* phones showing consistently good and poor distinctiveness, e.g., [ə] and [o] respectively.

Analysis. The results of x-vector are similar to that of U-LEVEL, which we assume is because they are both neural-network-based models. U-LEVEL shows a better performance for phone-repetition-based utterances and a large difference between phones. We assume it is because of the improved performance for short speeches, where a large number of repetitions of most phones can provide sufficient voiceprint information.

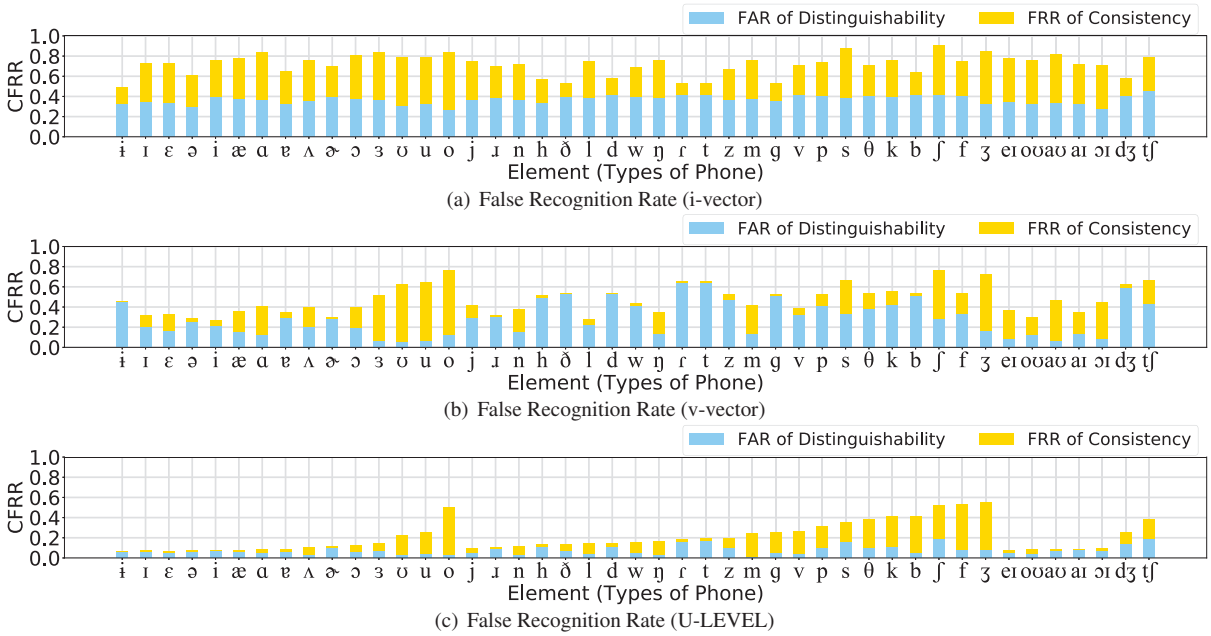
4.2.4 Influence of Order ❹

In this experiment, we investigate whether the order of phones affects voiceprint distinctiveness by changing the sequential order of phones in the speech content.

Setups. We regard the top 10 types of distinct phones observed by Sec. 4.2.3 as Class ***A*** (e.g., [i], [ə]), and the least 10 as Class ***B*** (e.g., [o], [ʃ]). We repeat and reassemble ***A*** and ***B*** into 4 groups of sequences ***ABAB***, ***BABA***, ***AABB***, and ***BBAA*** respectively, each consisting of 10 utterances. The results are as shown in Fig. 6.

Result. We can observe that four groups of utterances with order changes are identical in FAR, FRR, and CFRR, which indicates that order has little effect on the voiceprint distinctiveness.

Insights 4: The order of phones in a speech has little impact on the voiceprint distinctiveness.



(a) False Recognition Rate (i-vector)

(b) False Recognition Rate (v-vector)

(c) False Recognition Rate (U-LEVEL)

Figure 5: False recognition rate vs. Element❸ in i-vector, x-vector and U-LEVEL. Individual phones behave differently across models while there are some magic phones such as [ə] that has low false recognition rate than others for all three models.

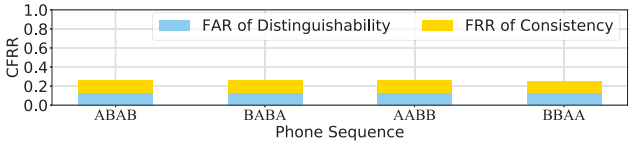


Figure 6: False recognition rate vs. Order ❹ in U-LEVEL. The effects of element and richness are about identical.

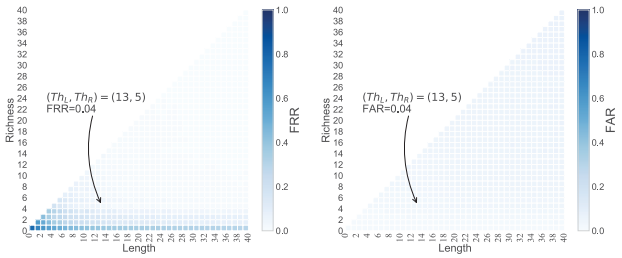
Analysis. The low impact of phone order on the voiceprint distinctiveness may be because ASV systems usually extract features by operations such as sampling, framing and Fourier transform. The change in the order of phone-length audio clips has barely effect on either the frame-length audio or the frequency features.

4.2.5 Influence of Synergies of Richness and Length ❺

In this experiment, we vary the number of phone types and the number of phones simultaneously. These two factors are relevant, and we study both their stand-alone distinctiveness improvement abilities and their synergistic effects.

Setups. We vary L and R of test audios in the range of $[1,40]$, with L greater than or equal to R. For any (L,R) pair, 6 sequences are randomly selected, and the test dataset for each speaker consists of 4,920 audios. We obtain FAR and FRR at each value of the variable pair, and show the averaged results across speakers in Fig. 7.

Result. It can be observed that the false recognition rate varies more in the direction of R. Specifically, it can not be kept at a low level when R is less than or equal to 5 even when L has reached 40. Based on the experimental results in Fig. 4 and Fig. 5, we find that richness shows a better performance in improving the voiceprint distinctiveness compared with



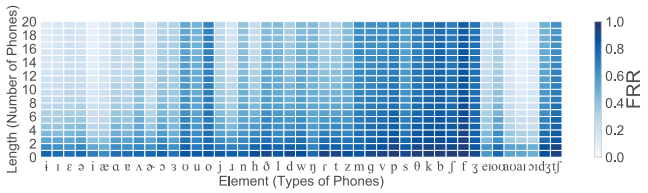
(a) FRR of Consistency (U-LEVEL) (b) FAR of Distinguishability (U-LEVEL)

Figure 7: False recognition rate vs. Synergies of Richness and Length ❺. (a) shows the FRR of consistency and (b) shows the FAR of distinguishability respectively in U-LEVEL. The distinctiveness thresholds i.e., (13,5) with $FRR = 0.04$ for synergistic changes in richness and length, as well as the corresponding FAR and FRR, are marked in the figure.

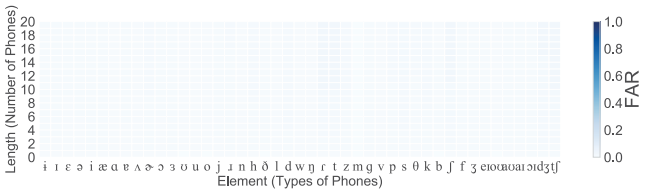
length. In addition, both richness and length have distinctiveness thresholds, beyond which the voiceprint distinctiveness does not improve. The length threshold Th_L, richness threshold Th_R, and the corresponding FAR and FRR are marked in Fig. 7.

In general, we have the following observations: (1) Richness has a greater impact on distinctiveness compared with length, and repetition only has a limited improvement. (2) When the number of phone types increases, the length threshold decreases. (3) When the number of phones increases, the richness threshold first falls and then rises.

Insights 5: Richness has a larger impact on voiceprint distinctiveness than length.



(a) FRR of Consistency (U-LEVEL)



(b) FAR of Distinguishability (U-LEVEL)

Figure 8: False recognition rate vs. Synergies of Length and Element ❻. (a) shows the FRR of consistency in U-LEVEL, and (b) shows the FAR of distinguishability. Distinctiveness growth rate differ for diverse types of phone and is not consistent with stable value ranking.

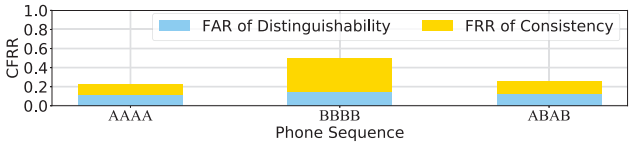


Figure 9: False recognition rate vs. Synergies of Richness and Element ❼ in U-LEVEL. The effect of phone ordering is low.

Analysis. The reason is that the voiceprint information contained in two repeated phones overlaps compared with two different phones, resulting in a less amount of information gain. The results of this experiment can be used to assess the security scores of both richness and length in Sec. 5.1.

4.2.6 Influence of Synergies of Length and Element ❻

In this experiment, we change both the length and element of the speech content to investigate whether distinct phones under saturated length perform equally after being repeated a few times.

Setup. We repeat 45 phones respectively with L varying in the range of $[1,20]$, and the test dataset for each speaker consists of 900 audios. When L of diphthong and affricate is singular, the audio used is identical to that of the adjacent double. The confusion matrices of FRR and FAR across all the speakers for three models are shown in Fig. 8.

Result. From the results, we can see that the effect of repetition on the distinctiveness of various phones is different. Compared with the results of Sec. 4.2.3, we find that the distinctiveness ranking of phones under the sufficient length differs from the ranking when the length changes. Specifically, we find: (1) Element has a greater impact on voiceprint distinctiveness compared with length, and it is difficult to enhance the distinctiveness of an indistinct phone by increasing the length alone. (2) Various phones show different distinctiveness sensitivities to the length.

Insights 6: Element shows a larger impact on the voiceprint distinctiveness compared with length.

Analysis. The distinctiveness improvements caused by phone repetitions can be abstracted as a joint effect of the initial value and the growth rate, neither of which is related to the stable value. The results of this experiment can be used to assess the distinctiveness scores of phone elements in Sec. 5.1.

4.2.7 Influence of Synergies of Richness and Element ❼

In this experiment, we construct audios of the same length but with different phones to compare the distinctiveness improvement abilities of richness and element.

Setup. Similar to Sec. 4.2.4, we repeat and reassemble ***A*** and ***B*** to generate 3 groups of sequences: ***AAAA***, ***BBBB***, and ***ABAB***, each consisting of 10 utterances. The thresholds of the verification models are adjusted during the experiment. The averaged results are shown in Fig. 9.

Result. It can be seen that for U-LEVEL, sequence ***AAAA*** is slightly more distinct than ***ABAB***, and more distinct than ***BBBB***. However, for i-vector and x-vector, ***ABAB*** is more distinct than ***AAAA***. We find (1) When the difference between phones is large, richness is slightly more influential than element for i-vector and x-vector, but it's opposite for U-LEVEL. (2) Element and richness are both significant in general, but the speech content constructed purposefully by selecting phones is more distinct than that generated by randomly increasing the number of phone types.

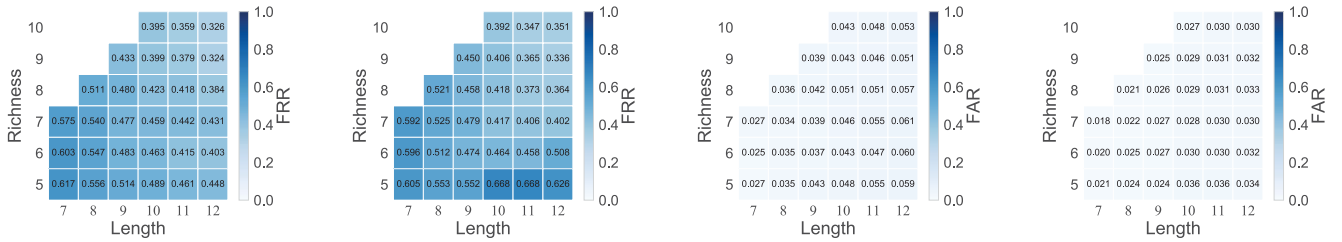
Insights 7: Both element and richness show comparable impact on the voiceprint distinctiveness.

Analysis. The reason is that the voiceprint information derived from repeating high distinct phones is comparable to that of adding low distinct phones.

4.3 Comparison with Real Words

To investigate whether disordered phone combinations and real English words share the same insights and observations, we conduct an additional experiment with real English words whose length is in the range of [7,12] and richness is in the range of [5,10] within which most commercial wake-up words fall. The experiment settings including model training are identical to those in 4.2.5.

The result of i-vector is shown in Fig. 10 while results for the other two models can be found in Appendix. A. Comparing the CFRRs of real words and random phones, the mean of the difference is 0.0009 and the standard deviation is 0.063. From the result, we can see that the absolute value and variation trend of FAR/FRR in the two test cases are similar for the same model. In several squares in Fig. 10, the FRR of real words is substantially high than that of random phones



(a) FRR of Consistency (i-vector, random phones)
(b) FRR of Consistency (i-vector, actual words)
(c) FAR of Distinguishability (i-vector, random phones)
(d) FAR of Distinguishability (i-vector, actual words)

Figure 10: False recognition rate vs. Synergies of Richness and Length ❺. (a), (b) show the FRR of consistency from test cases composed of random phone sequences and real words in i-vector. (c), (d) show the FAR of distinguishability for both. Real English words and generated phone sequences show little difference in terms of FRR under the i-vector model.

because real words are limited in the changes of variables, which brings biases to the element of test utterances. Therefore, we believe that the impact of speech contents is similar in both cases while random phone sequences can provide more unbiased test utterances.

5 PROLE Score Design and Validation

In this section, we first provide the design of PROLE Score based on the insights from Sec. 4. Then, we test 30 wake-up words from commercial products to provide an informative conclusion about the security level as well as alternative secure phrases. Finally, to validate the effectiveness of PROLE Score, we conduct a user study involving 46 users to compare the voiceprint distinctiveness from the user study with that given by PROLE Score.

5.1 PROLE Score Design

In this subsection, we formally provide the definition and the calculation method of the PROLE Score followed by an illustration. We also introduce the meaning of a PROLE Score.

Definition. As phone richness, length, and element have influence upon the voiceprint distinctiveness, we design PROLE Score as a function with richness R, length L, and element E as well as model M as parameters. Specifically, the formula of PROLE Score is:

$$\begin{aligned} S &= f^M(L,R,E) \\ &= 10 * (1 - F^M_{L+R} - \Delta F^M_E) \\ &= 10(1 - F^M_{L+R}(L,R) - \frac{\sum^E_\alpha F^M_{E\_\alpha}(L) \times L(\alpha)}{L} + F^M_{E\_eq}(L)) \end{aligned} \quad (1)$$

where F^M_{L+R} is the score component from length and richness, ΔF^M_E is the deviation value of the score component from element, α is a phone from the input words, $F^M_{E\_\alpha}$ is the score component of α, $L(\alpha)$ is the length of α, and $F^M_{E\_eq}$ is the score component of equally distributed elements.

Parameter calculation. We illustrate the parameter calculation process with 'Hey' as an example:

- **Step 1:** Calculate the richness R, length L, and element E of the input words. For example, 'Hey' with [**heɪ**] has richness 3 and length 3.
- **Step 2:** Obtain the bi-variate function F^M_{L+R} by curving the relationship between the false recognition rates and the synergies of richness and length (❺) based on the results in Sec. 4.2.5. Then obtain $F^M_{L+R}(L,R)$ with the function F^M_{L+R}, richness R, and length L. For example, the estimated false recognition rate derived from length and richness metric is $F_{L+R}(3,3)$ for 'Hey'.
- **Step 3:** Obtain the phone function $F^M_{E\_\alpha}$ by curving the relationship between the false recognition rates and the length of α based on the results in each column in Fig. 8. After that, obtain the equally-distributed element function $F^M_{E\_eq}$ by averaging the phone function $F^M_{E\_\alpha}$ for each phone α. For 'Hey', F_{E_eq} of [**heɪ**] is $F_{E\_eq}(3)$, and another part is $(F_{E\_[\mathbf{h}]}(3) \times 1 + F_{E\_[\mathbf{eɪ}]}(3) \times 2)/3$ [‡].

Meaning of a score. The meaning of a PROLE Score is to reflect the security level of a word/phrase under a specific ASV model, in terms of voiceprint distinctiveness. For a specific word, its PROLE Score is linearly inversely proportional to the estimated false recognition rate of the word. For example, ignoring bias from other factors, a phrase with the PROLE Score of 9 is expected to has a false recognition rate of 0.1 in a given ASV model. For example, from the distinctiveness threshold in Fig. 7 we regard 0.08 as a good false recognition rate in U-LEVEL, so we can define 9.2 as a good score of U-LEVEL.

5.2 Assessment of Wake-up Words from Commercial Products using PROLE Score

To test the behavior of PROLE Score, we collect the wake-up words in both Chinese and English of commercial ASV systems and calculate their PROLE Score. The wake-up words and ASV models with their scores are shown in Tab. 2 and listed alphabetically according to their brands. For each wake-up word, we calculate its PROLE Score under all three mod-

[‡] [**eɪ**] is a diphthong with the length of 2, where [**e**] cannot exist alone in English, so it is calculated as a whole.

Table 2: Commercial Wake-up Words distinctiveness Scoring

English Wake-up Words Scores					Chinese Wake-up Words Scores				
Developer	Wake-up Words	i-v[1]	x-v[2]	U-L[3]	Developer	Wake-up Words	i-v	x-v	U-L
Amazon	Alexa	3.28	2.89	8.16	Alibaba	TianMaoJingLing	6.80	7.32	9.60
Amazon	Amazon	3.29	2.89	8.39	Baidu	XiaoDuXiaoDu	5.30	5.52	9.60
Amazon	Computer	4.61	4.38	8.68	Huawei	NiHaoXiaoE	4.90	5.35	9.93
Amazon	Echo	1.50	1.26	6.06	Huawei	NiHaoYoYo	4.72	4.63	9.27
Apple	Hey Siri	4.20	4.43	9.32	Huawei	XiaoEXiaoE	4.76	4.81	9.92
Google	Hey Google	4.33	4.56	8.88	JD	DingDongDingDong	4.61	4.38	8.55
Google	Ok Google	5.55	6.04	9.00	JD	Hey XiaoJingYu	6.86	7.49	9.82
Huawei	Hey Celia	4.44	4.21	9.24	Lenovo	NiHaoLianXiang	6.42	7.17	9.82
Microsoft	Hey Cortana	5.60	6.00	9.42	MeiZu	NiHaoMeiZu	6.06	6.25	9.52
Multiverse	Extreme	4.02	3.72	8.76	Microsoft	NiHaoXiaoNa	6.26	6.81	9.83
MyCroft	Hey Mycroft	5.57	6.01	9.02	Mobvoi	NiHaoWenWen	5.01	5.44	9.58
Nuance	Hello Dragon	6.01	6.76	9.46	OPPO	XiaoBuXiaoBu	5.17	5.39	9.43
OPPO	Hey Breeno	5.19	5.46	9.20	OPPO	XiaoOuXiaoOu	4.62	4.57	9.75
Samsung	Hey Bixby	4.85	4.79	9.05	Tencent	XiaoWeiXiaoWei	5.84	6.66	9.90
SoundHound	OK Hound	5.41	5.80	9.31	XiaoMi	XiaoAiTongXue	6.85	7.36	9.87

[1] Abbreviation for i-vector model. [2] Abbreviation for x-vector model. [3] Abbreviation for U-LEVEL model.

Table 3: The Effect of Interjection Words on Distinctiveness

Suffix Words		Alexa	Bixby	Breeno	Celia	Cortana	Google	Siri
Average	Initial	4.77	4.89	4.74	4.16	5.53	3.55	3.70
PROLE Score	+Hey	6.92,21.43%[1]	6.74,18.51%	6.88,21.36%	6.31,21.60%	7.28,17.51%	6.25,27.03%	6.33,26.30%
and Growth	+OK	7.32,25.41%	6.78,18.92%	7.09,23.52%	7.09,29.39%	7.60,20.73%	7.07,35.29%	7.13,34.27%
Ratio	+Hello	7.00,22.30%	7.36,24.65%	6.96,22.21%	6.53,23.76%	7.50,19.69%	6.42,28.73%	6.62,29.13%

[1] The last three lines of each suffix word are the PROLE Score after adding interjections and score growth rate.

els, i.e., i-vector, x-vector and U-LEVEL. The left side of Tab. 2 shows the scores of English wake-up words while the rights side shows the Chinese ones. Moreover, we provid a tool website (https://sites.google.com/view/voiceprint-sec) which can generate PROLE Score for any words.

Models matter in PROLE Score. From Tab. 2, we can find that both English and Chinese wake-up words show variances and the scores can be greatly different, e.g., 1.5 (Amazon Echo, i-vector) vs. 6.06 (Amazon Echo, U-LEVEL) across models and words. Among the three models, U-LEVEL, as a representative DNN-based ASV model, behaves the best for all the wake-up words, regardless of languages. Note that we did not directly compare the scores of English and Chinese wake-up words since there are four types of phones that are different between the two languages.

Why current wake-up words work? From Tab. 2 we can also answer that most of the current wake-up words are effective to represent voiceprints. Take "Hey Cortana" as an example whose phone sequence is [**heɪ kɔːɹtɑːnə**] in IPA. First, the Richness (number of types) and Length (number of phones) are all 10, which is close to the distinctiveness thresholds for i-vector and x-vector and larger than the distinctiveness thresholds for U-LEVEL, indicating that the influences from the Richness and Length are nearly saturated. While for the Element factor, [**ɪ**], [**ɹ**], [**n**], [**ə**] can be ranked in the top half for distinctiveness on all three models, and [**h**], [**t**], [**ɑ**] are ranked in the top third on at least one model, thus phone types score is also high enough.

In addition to the above findings, we especially highlight the following three interesting observations for the readers, as they can reveal what is a good/bad choice of a wake-up word.

Observation 1: Both English and Chinese wake-up words favor a prefix word such as "OK", "Hey" or "Nihao"[¶] while Chinese ones prefer repeating.

Analysis. Among the commonly used wake-up words, 10/15 and 7/15 have prefix words in English and Chinese respectively. For the Chinese wake-up words, 6/15 has a repeating style, such as "DingDongDingDong" or "XiaoDuXiaoDu". This is mainly because of the cultural difference between the two nations.

Observation 2: Interjections like 'OK', 'Hey', 'Hello' prefixed to the content can significantly improve the voiceprint distinctiveness while repeating does not necessarily.

Analysis. To investigate the effectiveness of a prefix word such as the interjection words, we list the prefix as well as the suffix words (e.g., "Google" or "XiaoNa") in Tab. 3. We comprehensively compare the derived scores with and without the prefix words for 7 ASV systems and three interjection words, i.e., "Hey", "OK" and "Hello". It is interesting to find that interjections can increase the PROLE Score by around 24%. The reason is that an interjection word increases the phone length and phone richness of the wake-up word. Although

[¶]Nihao means hello in Chinese.

Table 4: Recommended Words and Scores

	In i-v[1]	Scores			In x-v[1]	Scores			In U-L[1]	Scores		
		i-v	x-v	U-L		i-v	x-v	U-L		i-v	x-v	U-L
High Scoring Words	unfortunately	6.53	7.33	9.28	unfortunately	6.53	7.33	9.28	surprisingly	5.89	6.62	9.53
	surprisingly	5.89	6.62	9.53	surprisingly	5.89	6.62	9.53	realize	4.00	3.70	9.43
	frustration	5.84	6.53	9.19	frustration	5.84	6.53	9.19	cafeteria	4.57	4.29	9.43
	uncomfortable	5.80	6.41	8.80	particularly	5.77	6.46	9.15	immediately	4.95	4.89	9.41
	conversation	5.77	6.41	9.12	conversation	5.77	6.41	9.12	cafeteria	5.34	5.38	9.38
	In i-v[1]	**Scores**			**In x-v[1]**	**Scores**			**In U-L[1]**	**Scores**		
		i-v	x-v	U-L		i-v	x-v	U-L		i-v	x-v	U-L
Self-created and Similar Words	[tivədɛgðəz] towards ers	5.86	6.43	9.16	[gɛzithəgwɑ] jes dog wa	5.81	6.47	9.20	[wijəʊziəʊd] we ears out	4.86	5.09	9.74
	[dædtəgiðɛm] dad together	5.84	6.35	9.21	[itzɑgbæθgə] it third best	5.80	6.43	8.99	[təhəʊziəʊt] ter house out	4.98	5.16	9.70
	[dibutgɛhəŋ] double hang	5.83	6.31	9.22	[əglizʤuti] a glass duty	5.77	6.43	9.11	[həŋidætnəu] her that now	5.67	6.08	9.65
	[mæthʊdðəgi] master key	5.83	6.30	9.25	[lɛgæbgiʤə] Le ga bridge	5.77	6.42	9.00	[jəʊhædnitɔ] your had little	5.62	6.08	9.65
	[bədɑgtiðəm] product them	5.82	6.36	9.05	[gitizəbkwɛ] git the booker	5.77	6.42	9.06	[nəizhəʊlæt] nice outlet	5.63	6.09	9.64

[1] The abbreviations are the same as in Tab. 2.

most of the suffix words only include around 8 phones, which is far from the distinctiveness threshold, the increase by interjections can help the phonetic factors of the combined word approach the threshold.

Observation 3: The improvement of voiceprint distinctiveness depends on both the prefix words as well as the suffix words and prefix words should have as little overlap as possible with the suffix one.

Analysis. The tone words that have little overlap with the suffix word can maximize the phone types, richness or length. For example, 'OK Google' is better than 'Go Google' because 'OK' ([**oʊkeɪ**]) brings more benefits than 'Go' ([**goʊ**]) does as it overlaps with 'Google' ([**gugəl**]) by the phone of [**g**].

We selected 10 words with high scores from 2,000 English common words dictionary [38] as the recommended recognition words to manufacturers. Words and scores for 3 models are shown in Tab. 4, and the scores are for reference only because problems such as inaccurate pronunciation and awkward-sounding may degrade distinctiveness. Moreover, we listed some phone sequences that we created based on the formation rules of English words and gave words with similar pronunciation, as a reference for manufacturers to create wake-up words.

5.3 User Study of `PROLE Score`

To validate the effectiveness of `PROLE Score` in practice, we conduct a user study.

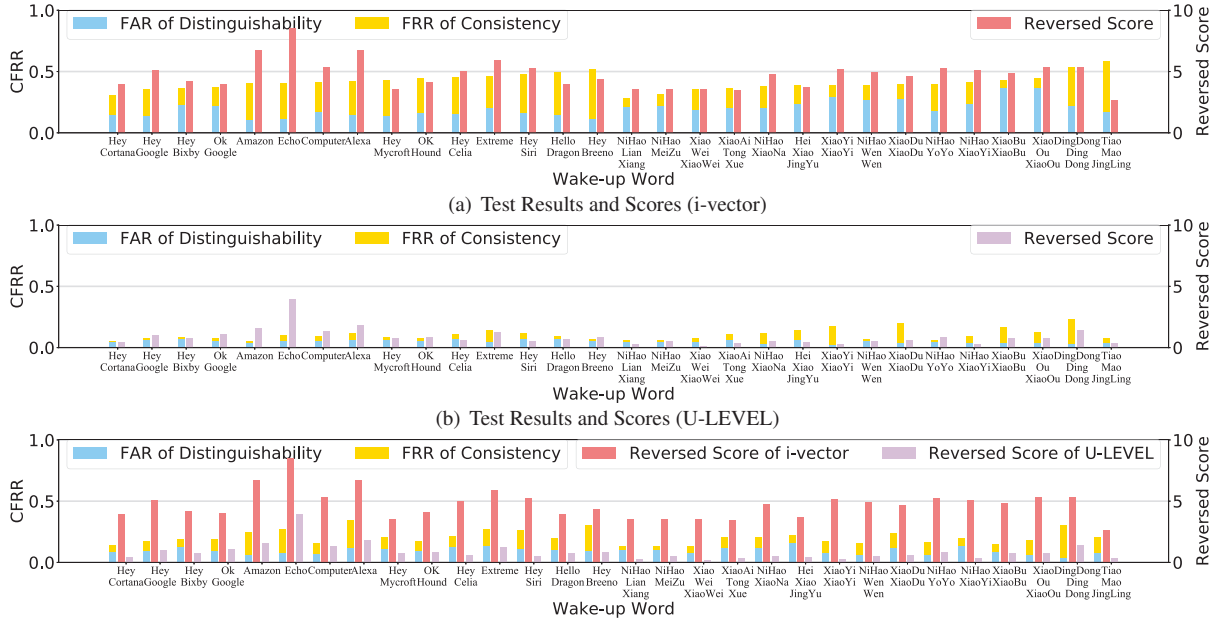
Setup. We recruited 46 volunteers aged between 19 and 50 years old from our campus, including both native and non-native speakers, with 23 females and 23 males [§]. During the user study, the volunteers are required to read five randomly selected sentences from the VCTK as the enrolling set, and the 30 wake-up words three times from Tab. 2 as test set in a quiet room. We recorded their speeches with microphones from 4 phones to prevent bias resulting from recording devices. i-vector and U-LEVEL are used as the verification models. For comparison, we also used a commercial voiceprint API, i.e., iFlytek [39]. The sum of FRR and FAR, i.e., CFRR for each wake-up word were averaged among repetition times and speakers. For a better illustration, we show the value of *reversed score* as 10-`PROLE Score` and compare it with the CFRR under each wake-up word.

How `PROLE Score` behaves. The results are shown in Fig. 11. We used the distance correlation method[40] to evaluate the correlation coefficient between the CFRR and the reserved score. The correlation coefficient ranges from 0 to 1, with a larger value more correlated. The correlation coefficient between the CFRR and the reversed score is 0.781 for i-vector, 0.748 for U-LEVEL. What is more, we also find that the correlation coefficients between CFRR under iFlytek and the reversed value under both i-vector and U-LEVEL are high, e.g., > 0.78.

In summary, CFRR and `PROLE Score` show a strong correlation, indicating that our `PROLE Score` indeed can reveal the security level of a chosen word/phrase.

English vs. Chinese. Besides, We try to understand whether `PROLE Score` from the test in English can evaluate the distinctiveness of speech content in other languages, so we evaluate Chinese wake-up words. The user study indicates that the CFRR of Chinese wake-up words has relevance with `PROLE Score` derived from English test samples. Through the statistical analysis, we find that the distance correlation between Chinese speech content and `PROLE Score` is 0.752 averaged in models, while between English speech content and `PROLE Score` is 0.783. From the results we can find that

[§] We followed the local regulations to protect the rights of human participants despite the absence of Institutional Review Board (IRB).



(a) Test Results and Scores (i-vector)

(b) Test Results and Scores (U-LEVEL)

(c) Test Results (iFlytek) and Scores

Figure 11: CFRR values vs. reversed score (i.e., 10-PROLE Score) for the 30 wake-up words in the user study for three models. CFRR of user study results highly correlate with the PROLE Score, indicating PROLE Score is a effective scoring system representing the security level of a given wake-up word.

the phonetic factors do have a similar influence on both Chinese and English. Theoretically, the types of single phones of Chinese and English are 29 and 40 respectively, among which 22 are the same. Most of the Chinese phones can be evaluated by PROLE Score derived from English.

Text-dependent vs. Text-independent. To improve the comprehensiveness, we repeat the user study by using the same wake-up words as enrolling and test utterances (i.e., Method 1), compared with the results from random enrolling sentences and wake-up words as test samples (i.e., Method 2), and present the results of iFlytek in Fig. 12 while other models in Appendix. A. We can find that the CFRR in Method 1 is obviously lower than in Method 2, while the difference of distinctiveness between wake-up words is similar for the two methods. The distance correlation between CFRRs for both Methods are 0.750 (i-vector), 0.729 (U-LEVEL) and 0.826 (iFlytek). The decrease of false recognition rate is because text-independent models perform better when the enrollment and test phrases are the same.

6 Suggestions

We propose suggestions on how to improve the security of voiceprint from both manufacturer and user sides.

Wake-up Words and Commands For manufacturers, we suggest they can repeat the tests in our paper for their models, calculate the PROLE Score for all the candidate wake-up words and avoid choosing low-PROLE Score ones. Similar to wake-up words, when specifying commands with high relevance to the user's personal and property safety, manufacturers should select phrases that achieve the distinctiveness threshold in length and richness, and include elements with high distinctiveness. If users can set their own wake-up words or voice commands, we suggest that manufacturers can develop an evaluation system which returns the PROLE Score for the user-selected speech contents to help users choose high-PROLE Score ones.

Models. From the model perspective, voiceprint distinctiveness can be improved in a targeted way. If the wake-up words or commands must be simple for usability reasons, the model threshold on specific speech contents can be modified to ensure higher and more balanced performance for wake-up words or commands with insufficient information.

Users. The distinctiveness of voiceprint needs improvement at present. Users should be careful in employing voiceprint recognition or verification in sensitive applications or scenarios, e.g., electric payment, if only short and monotonous verification utterances are supported.

7 Discussions

Accuracy of Test Dataset Construction. We consider most of the error in test dataset construction derives from calculating timestamps and durations of phone audios by MFA tool. We took 100 samples randomly from phone audios and recruited 5 people to evaluate whether the timestamps of phones are accurate. By statistical analysis, the accuracy of MFA aligner is 92.4%. We believe the error of MFA may lead to more severe errors on elements while having a slight effect on length and richness. Because the error of segmentation is

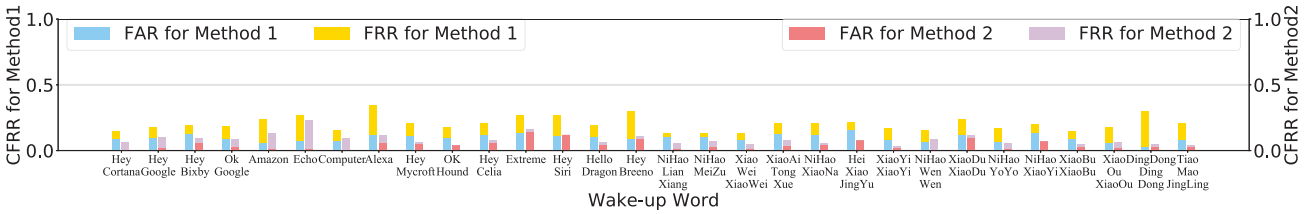


Figure 12: False recognition rates of wake-up words for Method 1 and Method 2 in iFlytek. In Method 1 the enrollment utterances are different from the test, while in Method 2 are the same. There is similar regularity in distinctiveness gap between wake-up words when using Method 1 and Method 2.

averaged over a large amount of data, so that the length and richness of phone sequences approach the desired values. But the phone audios may be mixed with other phones so that it may lead to a smaller score difference among elements. We compensated for the reductive differences among elements by increasing the weight of element metric in `PROLE Score`.

Impact of Pronunciation. In practice, words with higher distinctiveness scores may not necessarily show lower false recognition rates if they are difficult in pronunciation. For instance, "particularly appreciate" has a higher distinctiveness score but "particularly" shows a lower false recognition rate. We assume it is because phrases composed of long and complex words are difficult to pronounce especially for those non-native speakers, resulting in inaccurate and slur pronunciation and thus the distinctiveness decrease.

Transferability across Models. We tested several typical and popular voiceprint verification models in this paper for the sake of exploring common insights and observation across models. We note that it requires further study on more models and larger datasets to obtain more accurate and general conclusions. We remain it as one direction of the further work.

8 Related Work

Voiceprint Security Affecting Factors. Several researches have studied the influencing factors of voiceprint security, most of which focus on the model optimization, the subjective factors related to speakers and speech content. In terms of models, existing work [41] mainly focuses on improving the model structure and features, defending spoofing and imitation attacks, and reducing the effects of short speeches and noises to improve the performance of the voiceprint model, while this work [42] studies the impact of the training dataset size on the model performance. In terms of speakers, some studies focus on the impact of the recording environmental noise [43] and health state [44]. In terms of speech contents, existing studies focus on the model optimization against the negative effects of short speech contents [45], the difference of users' preference and performance between numbers and complete sentences [46], and the information of voiceprint in the feature space for speeches of different lengths [47]. These studies are most related to our work but we analyze from an aspect of phonetic factors, which can model any contents and explore the impact of speech contents on voiceprint distinctiveness in a finer-granularity manner.

Impact of Phone in Speech Content. For the impact of phonemes or phones on speaker recognition, existing studies mainly focus on the performance comparison [48] and model fusion [12–14] of individual phoneme-trained ASV models. Specifically, Alsulaiman et al. [48] investigated the effect of phonemes in Arabic on the performance of ASV systems. Fatima et al. [12] defined the vowel category formed by the combination of English and Chinese, and used them to train a universal background phoneme model based on the conventional GMM-UBM system. Fatima et al. [13] discussed the importance of phones for speaker recognition and showed that vowels represent a large amount of speaker-specific information. Zhang et al. [14] proposed individual phoneme-trained ASV models to solve the problem of poor performance in short speech testing. Different from these studies, our work focuses on the speech contents with phonetic factors, and analyzes the impact of speech contents by adjusting phones.

9 Conclusion

In this paper, we investigate the impact of speech contents on the distinctiveness of voiceprint with 2457 speakers and 14,600,000 test samples. We experimentally obtain the correlation between the false recognition rates and the richness, the length, the order, and the elements of phones. We define `PROLE Score` that can be calculated based on speech content yet can reflect the voice distinctiveness. Under the guidance of `PROLE Score`, we test 30 wake-up words of 19 commercial voice assistants and provide recommendations for both users and manufacturers on selecting secure wake-up words.

Acknowledgments

We thank the iFlytek Co.,Ltd. for its support on the ASV model and the anonymous reviewers for valuable comments. This work is supported by China NSFC Grant 61941120, 61925109, 62071428 and CPSF Grant BX2021158.

References

[1] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li. Spoofing and countermeasures for speaker verification: A survey. *speech communication*, 66:130–153, 2015.

[2] Nilu Singh, RA Khan, and Raj Shree. Applications of speaker recognition. *Procedia engineering*, 38:3122–3126, 2012.

[3] Zhijian Xu, Guoming Zhang, Xiaoyu Ji, and Wenyuan Xu. Evaluation and defense of light commands attacks against voice controllable systems in smart cars. *Noise & Vibration Worldwide*, 52(4-5):113–123, 2021.

[4] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. Dolphinattack: Inaudible voice commands. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 103–117, 2017.

[5] Wiki. Wiktionary:amazon echo. https://en.wikipedia.org/wiki/Amazon_Echo, 2018.

[6] TDBank. Td voiceprint. https://www.tdbank.com/bank/tdvoiceprint.html, 2019.

[7] Rupinder Saini and Narinder Rana. Comparison of various biometric methods. *International Journal of Advances in Science and Technology*, 2(1):24–30, 2014.

[8] Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan, and Zhenyao Zhu. Deep speaker: an end-to-end neural speaker embedding system. *arXiv preprint arXiv:1705.02304*, 650, 2017.

[9] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333. IEEE, 2018.

[10] David Snyder, Pegah Ghahremani, Daniel Povey, Daniel Garcia-Romero, Yishay Carmiel, and Sanjeev Khudanpur. Deep neural network-based speaker embeddings for end-to-end speaker verification. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 165–170. IEEE, 2016.

[11] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4879–4883. IEEE, 2018.

[12] Nakhat Fatima, Xiaojun Wu, Thomas Fang Zheng, Chenhao Zhang, and Gang Wang. A universal phoneme-set based language independent short utterance speaker recognition. In *11th National Conference on Man-Machine Speech Communication (NCMMSC'11), Xi'an, China*, pages 16–18. Citeseer, 2011.

[13] Nakhat Fatima and Thomas Fang Zheng. Vowel-category based short utterance speaker recognition. In *2012 International Conference on Systems and Informatics (ICSAI2012)*, pages 1774–1778. IEEE, 2012.

[14] Chenhao Zhang, Xiaojun Wu, Thomas Fang Zheng, Linlin Wang, and Cong Yin. A k-phoneme-class based multi-model method for short utterance speaker recognition. In *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–4. IEEE, 2012.

[15] Rosa González Hautamäki, Tomi Kinnunen, Ville Hautamäki, and Anne-Maria Laukkanen. Automatic versus human speaker verification: The case of voice mimicry. *Speech Communication*, 72:13–31, 2015.

[16] Zhizheng Wu and Haizhou Li. On the study of replay and voice conversion attacks to text-dependent speaker verification. *Multimedia Tools and Applications*, 75(9):5311–5327, 2016.

[17] Chen Yan, Yan Long, Xiaoyu Ji, and Wenyuan Xu. The catcher in the field: A fieldprint based spoofing detection for text-independent speaker verification. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 1215–1229, 2019.

[18] Chen Yan, Guoming Zhang, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. The feasibility of injecting inaudible voice commands to voice assistants. *IEEE Transactions on Dependable and Secure Computing*, 2019.

[19] Thomas Fang Zheng, Qin Jin, Lantian Li, Jun Wang, and Fanhu Bie. An overview of robustness related issues in speaker recognition. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, pages 1–10. IEEE, 2014.

[20] Suwon Shon. voxceleb-ivector. https://github.com/swshon/voxceleb-ivector, 2018.

[21] Ahilan Kanagasundaram, Sridha Sridharan, Sriram Ganapathy, Prachi Singh, and Clinton Fookes. A study

of x-vector based speaker recognition on short utterances. In *Proceedings of the 20th Annual Conference of the International Speech Communication Association, INTERSPEECH 2019. Vol. 2019-September.*, pages 2943–2947. ISCA (International Speech Communication Association), 2019.

[22] Weidi Xie, Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Utterance-level aggregation for speaker recognition in the wild. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5791–5795. IEEE, 2019.

[23] Siri Team. Personalized hey siri. https://machinelearning.apple.com/research/personalized-hey-siri, 2018.

[24] Chunlei Zhang and Kazuhito Koishida. End-to-end text-independent speaker verification with triplet loss on short utterances. In *Interspeech*, pages 1487–1491, 2017.

[25] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2010.

[26] Daniel Garcia-Romero and Carol Y Espy-Wilson. Analysis of i-vector length normalization in speaker recognition systems. In *Twelfth annual conference of the international speech communication association*, 2011.

[27] Jinghua Zhong, Wenping Hu, Frank K Soong, and Helen Meng. Dnn i-vector speaker verification with short, text-constrained test utterances. In *Interspeech*, pages 1507–1511, 2017.

[28] Yun Tang, Guohong Ding, Jing Huang, Xiaodong He, and Bowen Zhou. Deep speaker embedding learning with multi-level pooling for text-independent speaker verification. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6116–6120. IEEE, 2019.

[29] Shi-Xiong Zhang, Zhuo Chen, Yong Zhao, Jinyu Li, and Yifan Gong. End-to-end attention based text-dependent speaker verification. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 171–178. IEEE, 2016.

[30] Alonzo Church. David crystal. linguistics. penguin books ltd., harmondsworth, middlesex, and penguin books, inc., baltimore, maryland, 1971, 267 pp.-frank palmer. grammar. penguin books ltd., harmondsworth, middlesex, and penguin books, inc., baltimore, maryland1971, 200 pp. *The Journal of Symbolic Logic*, 37(2):420–420, 1972.

[31] Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92), 2019.

[32] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5206–5210. IEEE, 2015.

[33] CoML. phonemizer. https://github.com/bootphon/phonemizer, 2020.

[34] Jongseok Park, Kyubyong & Kim. g2pe. https://github.com/Kyubyong/g2p, 2019.

[35] Michael McAuliffe, Muhammad Rifqi Fatchurrahman Putra Danar, Christophe Veaux, Paweł Potrykus, Arlie Coles, and Harsh Mishra. Montrealcorpustools. https://github.com/MontrealCorpusTools/Montreal-Forced-Aligner, 2015.

[36] Manoj Kumar, Tae Jin-Park, Somer Bishop, Catherine Lord, and Shrikanth Narayanan. Designing neural speaker embeddings with meta learning, 2020.

[37] Arnab Poddar, Md Sahidullah, and Goutam Saha. Performance comparison of speaker recognition systems in presence of duration variability. In *2015 Annual IEEE India Conference (INDICON)*, pages 1–6. IEEE, 2015.

[38] Wiki. Wiktionary:frequency lists/contemporary fiction. https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/Contemporary_fiction, 2018.

[39] Jun Du, Yan-Hui Tu, Lei Sun, Feng Ma, Hai-Kun Wang, Jia Pan, Cong Liu, Jing-Dong Chen, and Chin-Hui Lee. The ustc-iflytek system for chime-4 challenge. *Proc. CHiME*, 4:36–38, 2016.

[40] Michael R Kosorok et al. Discussion of: Brownian distance covariance. *The Annals of Applied Statistics*, 3(4):1270–1278, 2009.

[41] Achintya Kumar Sarkar, Driss Matrouf, Pierre Michel Bousquet, and Jean-François Bonastre. Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification. In

Thirteenth Annual Conference of the International Speech Communication Association, 2012.

[42] Ruirui Li, Jyun-Yu Jiang, Jiahao Liu Li, Chu-Cheng Hsieh, and Wei Wang. Automatic speaker recognition with limited data. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 340–348. ACM, 2020.

[43] Thomas Fang Zheng, Qin Jin, Lantian Li, Jun Wang, and Fanhu Bie. An overview of robustness related issues in speaker recognition. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, pages 1–10. IEEE, 2014.

[44] Mohammed Usman. On the performance degradation of speaker recognition system due to variation in speech characteristics caused by physiological changes. *International Journal of Computing and Digital Systems*, 6(03):119–126, 2017.

[45] Arnab Poddar, Md Sahidullah, and Goutam Saha. Speaker verification with short utterances: a review of challenges, trends and opportunities. *IET Biometrics*, 7(2):91–101, 2017.

[46] Nancie Gunson, Diarmid Marshall, Fergus McInnes, and Mervyn Jack. Usability evaluation of voiceprint authentication in automated telephone banking: Sentences versus digits. *Interacting with Computers*, 23(1):57–69, 2011.

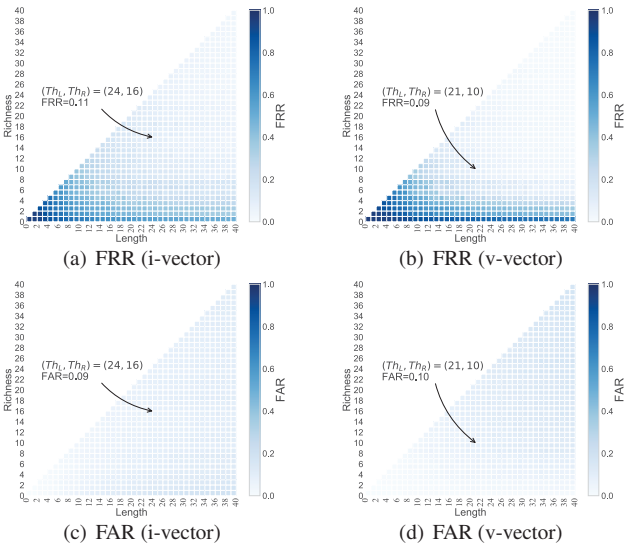
[47] Andreas Nautsch, Christian Rathgeb, Rahim Saeidi, and Christoph Busch. Entropy analysis of i-vector feature spaces in duration-sensitive speaker recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4674–4678. IEEE, 2015.

[48] Mansour Alsulaiman, Awais Mahmood, and Ghulam Muhammad. Speaker recognition based on arabic phonemes. *Speech Communication*, 86:42–51, 2017.

Appendix

A Supplementary Experiment Results

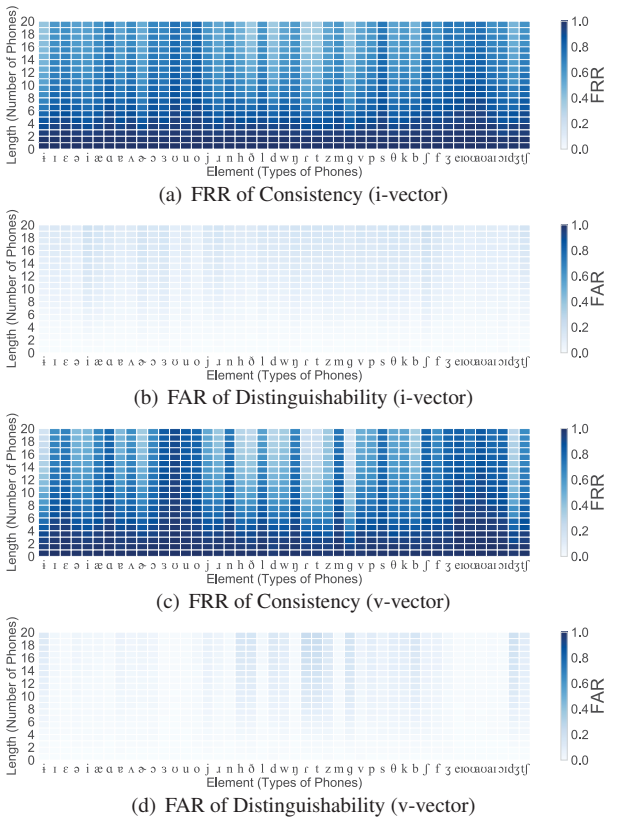
It is the figure of results in Sec. 4.2. The results in i-vector and x-vector of 4.2.1, 4.2.2, 4.2.4, 4.2.5, 4.2.6, 4.2.7 are shown in Fig. 19, Fig. 20, Fig. 15, Fig. 13, Fig. 14, Fig. 16, separately. The results in x-vector and U-LEVEL of 4.3 isc shown in Fig. 17. The results in i-vector and U-LEVEL of 5.3 is shown in Fig. 18.



(a) FRR (i-vector) (b) FRR (v-vector)

(c) FAR (i-vector) (d) FAR (v-vector)

Figure 13: False recognition rate vs. Synergies of Richness and Length ❺. (a), (b) show the FRR in i-vector and x-vector. (c), (d) show the FAR of distinguishability.



(a) FRR of Consistency (i-vector)

(b) FAR of Distinguishability (i-vector)

(c) FRR of Consistency (v-vector)

(d) FAR of Distinguishability (v-vector)

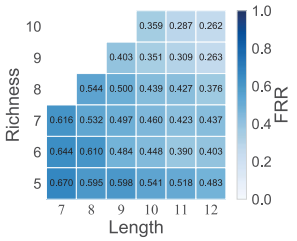
Figure 14: False recognition rate vs. Synergies of Length and Element ❻. (a), (c) show the FRR in i-vector and x-vector. (b), (d) show the FAR.

(a) False Recognition Rate (i-vector) (b) False Recognition Rate (v-vector)

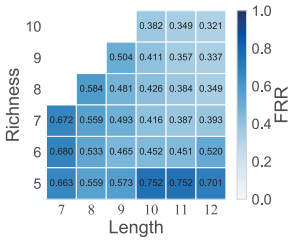
Figure 15: False recognition rate vs. Order ❹.

(a) False Recognition Rate (i-vector) (b) False Recognition Rate (v-vector)

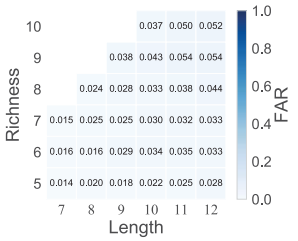
Figure 16: False recognition rate vs. Synergies of Richness and Element ❼.



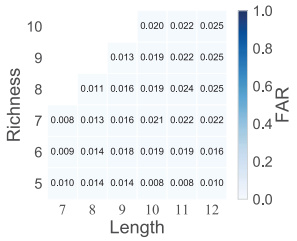
(a) FRR of Consistency (v-vector, random phones)



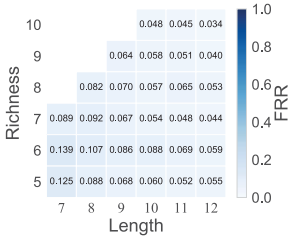
(b) FRR of Consistency (v-vector, actual words)



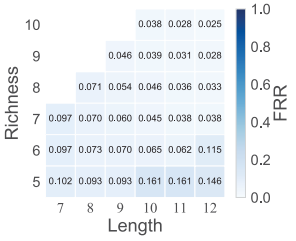
(c) FAR of Distinguishability (v-vector, random phones)



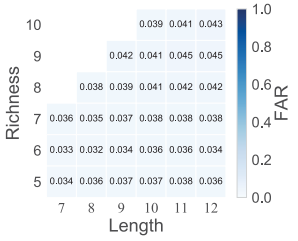
(d) FAR of Distinguishability (v-vector, actual words)



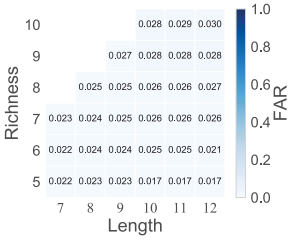
(e) FRR of Consistency (U-LEVEL, random phones)



(f) FRR of Consistency (U-LEVEL, actual words)

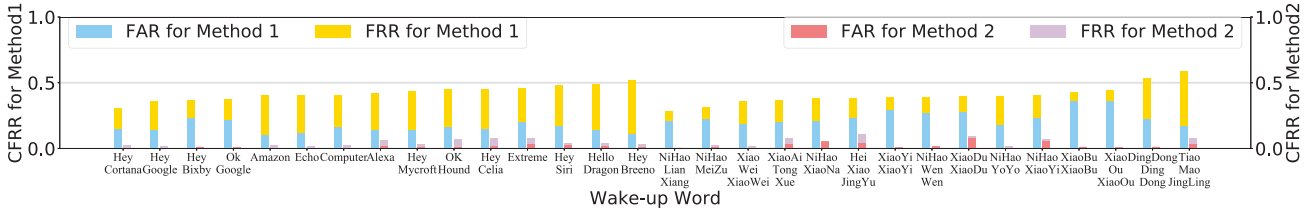


(g) FAR of Distinguishability (U-LEVEL, random phones)

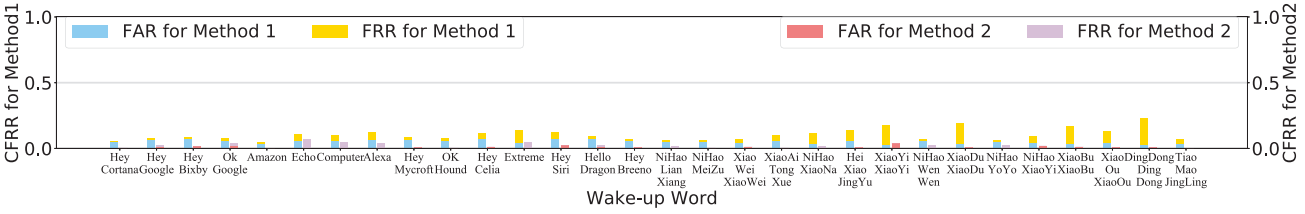


(h) FAR of Distinguishability (U-LEVEL, actual words)

Figure 17: False recognition rate vs. Synergies of Richness and Length ❺. (a), (e) show the FRR of consistency from test cases composed of random combinations of phones in x-vector and U-LEVEL. (b), (f) show the FRR of consistency from test cases composed of actual words in x-vector and U-LEVEL. (c), (g) show the FAR from test cases composed of random combinations of phones in x-vector and U-LEVEL. (d), (h) show the FAR from test cases composed of actual words in x-vector and U-LEVEL.

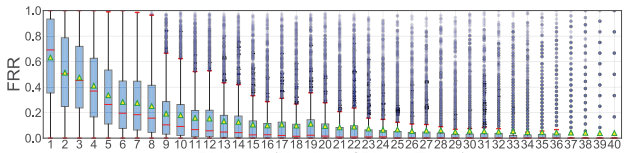


(a) Test Results for Method 1 and Method 2 (i-vector)

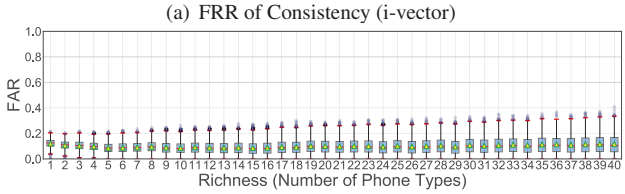


(b) Test Results for Method 1 and Method 2 (U-LEVEL)

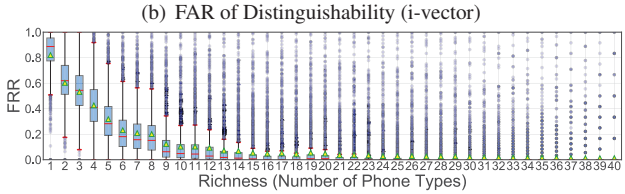
Figure 18: False recognition rates and reversed score (i.e., 10-PROLE Score) from the 30 wake-up words for Method 1 and Method 2.



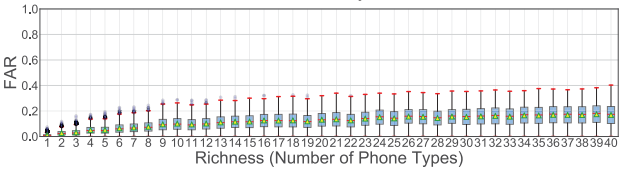
(a) FRR of Consistency (i-vector)



(b) FAR of Distinguishability (i-vector)

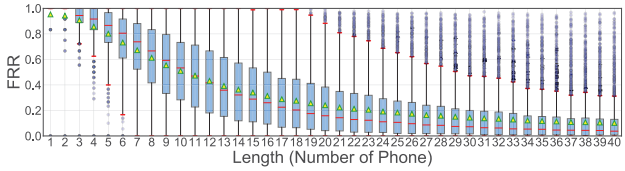


(c) FRR of Consistency (v-vector)

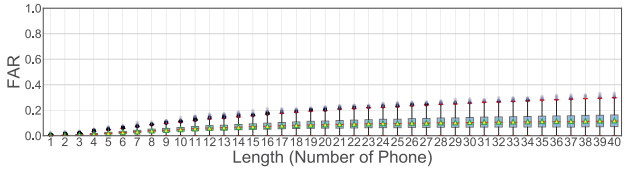


(d) FAR of Distinguishability (v-vector)

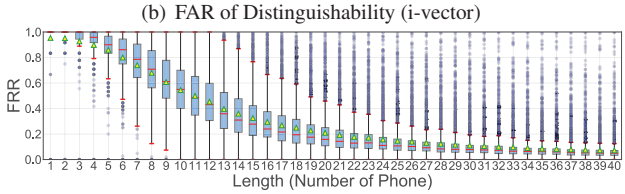
Figure 19: False recognition rate vs. Richness ❶. (a), (c) show the FRR of consistency in i-vector and x-vector. (b), (d) show the FAR of distinguishability in the two models.



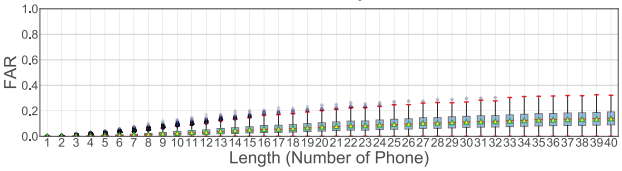
(a) FRR of Consistency (i-vector)



(b) FAR of Distinguishability (i-vector)



(c) FRR of Consistency (v-vector)



(d) FAR of Distinguishability (v-vector)

Figure 20: False recognition rate vs. Length ❷. (a) show the FRR of consistency in i-vector and x-vector. (b), (d) show the FAR of distinguishability.