applied sciences

MDPI

Article

A Preliminary Study of Model-Generated Speech

Man-Ni Chu [1] and Yu-Chun Wang [2,*]

[1] Graduate Institute of Cross-Cultural Studies, Fu Jen University, New Taipei City 242062, Taiwan; mannichu@gmail.com
[2] Department of Buddhist Studies, Dharma Drum Institute of Liberal Arts, New Taipei City 208303, Taiwan
* Correspondence: ycwang@dila.edu.tw

Abstract: The goal of this study was to compare model-generated sounds with the process of sound acquisition in humans. The research utilized two dictionaries of the Chaoshan dialect spanning approximately one century. Identical Chinese characters were selected from each dictionary, and their contemporary pronunciations were documented. Subsequently, inconsistencies in pronunciation were manually rectified, following which three machine learning methods were employed to train the pronunciation of words from one dictionary to another. These methods comprised the attention-based sequence-to-sequence method, DirecTL+, and Sequitur. The accuracy of the model was evaluated using five-fold cross-validation, revealing a maximum accuracy of 68%. Additionally, the study investigated how the probability of a sound's subsequent unit influences the accuracy of the machine learning methods. The attention-based sequence-to-sequence model is not solely influenced by the frequency of input but also by the probability of the subsequent unit.

Keywords: attention-based seq2seq; Chaoshan dictionary; DirecTL+; Sequitur

Citation: Chu, M.-N.; Wang, Y.-C. A Preliminary Study of Model-Generated Speech. *Appl. Sci.* **2024**, *14*, 3104. https://doi.org/10.3390/app14073104

Academic Editor: Douglas O'Shaughnessy

Received: 2 March 2024
Revised: 3 April 2024
Accepted: 4 April 2024
Published: 7 April 2024

1. Introduction

With the rapid development of science and technology, data exploration techniques can now be applied to different levels of knowledge. The increasing sophistication of machine learning models enables the simulation of a growing array of phonetic changes, therefore mirroring real-world linguistic shifts. For example, the study conducted by [1] employs an agent-based model (ABM) methodology to investigate speech stability and change by utilizing authentic speech data from a cohort of speakers. The analysis encompasses two distinct speech databases, with the initial dataset comprising recordings of individuals across different age groups speaking Standard Southern British English (SSBE). By employing an ABM facilitated by unsupervised machine learning algorithms, the research identifies a progressive phonetic shift concerning the /u/ vowel sound. However, no significant phonological reclassification of any vowels was observed. These findings suggest a tendency towards phonetic alteration while maintaining phonological stability. Furthermore, the ABM approach was extended to examine diphthongs in New Zealand English over the preceding five decades, particularly focusing on the merger of /eə/ into /ɪə/. This exploration delves into the principles of exemplar theory to elucidate the dynamics of phonetic change within the context of language evolution. As machine learning methodologies effectively capture these transformations, researchers in phonetics and phonology can leverage such simulations to scrutinize and refine existing theoretical frameworks of phonetic change. This comparative analysis serves to elucidate specific theories pertaining to the mechanisms underlying phonetic evolution. In complement to micro-level analyses of speech evolution, the macroscopic observation of historical linguistic changes represents another realm where machine learning methodologies can offer substantial contributions.

In parallel, the history of sound changes in a language seems to be being investigated, but this has rarely been found in previous studies. One of the related fields is machine transliteration, which involves the phonetic conversion of words from a source

language into the words of a target language. Positioned as a subset of machine translation, it encompasses diverse methodological approaches aimed at addressing the intricacies of transliteration tasks. In that sense, utilizing a novel corpus automatic detection method could potentially bridge existing gaps by facilitating a comparative analysis between model-generated predictions and the derivational sequence of sound development. This parallel examination holds promise for enhancing our understanding of phonetic evolution and the predictive capabilities of computational models in phonological research.

Indeed, a targeted approach involves leveraging established models to translate a given word from one contemporary pronunciation to another. By systematically analyzing the accurately generated words from such models in juxtaposition with the sound acquisition sequence observed in human beings, which typically progresses from unmarked features to marked ones, valuable insights into phonetic evolution and model performance can be gleaned.

On the other hand, the evolution of the Chaoshan dialect within historical linguistic contexts underscores its origins and development into the contemporary dialect spoken today. Scholars such as [2,3] assert that the Chaoshan dialect is part of the Southern Min dialect, which itself belongs to the broader Min dialect family. Historically, Min dialects have exhibited close linguistic ties with other Chinese dialects, yet they have also maintained a distinctiveness, rendering Min among the most conservative of all Chinese dialect groups [4]: p. 216). The Chaoshan dialect's formation can be traced back to around the Song Dynasty. Evidence provided by scholars ([2]: p. 124; [3]: p. 97–99) suggests that during the Tang and Song Dynasties, the Chaozhou and Fujian dialects exhibited only regional differences without any fundamental distinctions. It was not until the Yuan and Ming dynasties that the Chaoshan dialect gradually diverged from the Fujian dialect and began to assimilate various aspects of Cantonese, including pronunciation, vocabulary, and grammar ([5]: p. 17, p. 52–56).

Thus, this section will introduce Chaoshan phonology, with two meticulously preserved Chaoshan dictionaries employed as the case study. Subsequently, we focus on the sequential emergence of vowels, consonants, and tones in children's acquisition processes. The anticipated outcomes predicted by models for the Chaoshan dictionaries include a parallel assessment of children's speech sound acquisition processes. Such analysis is poised to augment our comprehension of the predictive efficacy inherent in computational models within the domain of speech research.

1.1. Chaoshan Phonology

According to [6], the Chaoshan dialect has 18 initials, /p, t, k, p^h, t^h, k^h, b, g, ts, ts^h, s, z, m, n, ŋ, l, h, and ø/; six vowels /i, e, a, o, and u, ɯ/; and eight tones, T1 = 33, T2 = 52, T3 = 213, T4 = 2, T5 = 55, T6 = 35, T7 = 11, and T8 = 4, with respect to the five-scale tonal description of [7]. The syllable structure is reported to be CGVE, where C is a [+consonantal] segment; G is a glide; V is the vowel nucleus; and E is either a nasal or a glide. According to the two dictionaries, Chaoshan has systematically lost the mostly unmarked [t] and [n] at the end of the rhyme after more than 100 years. This evaluation was replicated in the perception and production experiments conducted by [8–10].

1.2. The Acquisition of Vowels, Consonants, and Tones by Children in Terms of Their Order of Emergence

Concerning the development and acquisition of languages by children, several different studies have been undertaken in relation to the Sino-Tibetan language family, including Mandarin Chinese, Taiwan Southern Min (TSM), and Cantonese. Consonant, vowel, and tone acquisition have been discussed based on different Sino-Tibetan languages. In consonant acquisition, with regards to the manner of articulation (MOA), stops have been found to be acquired earlier than fricatives and following affricates [11,12] for Mandarin Chinese in Taiwan, [13], for TSM, [14], and for Cantonese). As for the place of articulation (POA), it has been found that consonants articulated in the anterior part of the mouth are

usually acquired earlier than consonants in the posterior part of the mouth [15]. This is supported by [11,12] for Mandarin Chinese in Taiwan and [13] for TSM. However, ref. [14] assert that coronals are acquired earlier than labials in Cantonese. As for voiced features, voiced consonants are usually acquired earlier in English [16], TSM [13], and Cantonese [14], compared to Mandarin Chinese in Taiwan [11,12]. As for aspirated features, unaspirated consonants are acquired first, followed by their aspirated counterparts in Taiwan Mandarin Chinese [11,12], in TSM, [13], and in Cantonese, [14]. Overall, the order of consonant acquisition in Taiwan Mandarin Chinese, TSM, and Cantonese pretty consistently goes from unmarked to marked features according to Jakobson's definition.

In vowel acquisition, most studies have demonstrated that children acquire vowels earlier than consonants. Regarding the nature of the vowels, [a] seems to be acquired first, then [i] and [u], followed by the mid-vowels and others. In Mandarin Chinese in Taiwan, ref. [11] not only demonstrated the same pattern but also reported that the CV structure is dominant at the first stage, followed by CVC, CVV, and the rest of the complex syllabic structures. Moreover, oral vowels are acquired earlier than nasalized vowels, and single vowels are acquired first, compound vowels second, and triple vowels last in TSM [13]. For vowel acquisition by children in Mandarin, Jakobson's mature theory, i.e., from unmarked to marked feature acquisition, seems to hold.

Finally, the subject of tone acquisition appears to have drawn the most attention from scholars. It has been established that level tones are acquired earlier than contour tones ([11,12] for Mandarin Chinese in Taiwan; [13] for TSM; [14] for Cantonese). For TSM, Hsu (1989) [13] reported that high initial tones (high-level, high-falling, and high-rising) were acquired first, followed by the low-falling tones, and then the mid-level tones. The low, entering tone, *yin-ru*, was acquired last because the high, entering tone, *yang-ru*, merged into the mid-level tone, yielding a total of six tones. In addition, an unchecked tone is acquired earlier than a checked tone.

To summarize, the development and acquisition of Mandarin Chinese in Taiwan, Taiwan Southern Min (TSM), and Cantonese among children is in line with that proposed by [15]. Regarding acquisition by children, ref. [15] proposed a prototype of 'maturational theory', whereby the biological programming in human beings determines the structure of language acquisition. Based on Jakobson's hierarchy of development at the word level, each component can be assigned as a subset of binary distinctive features (DFs) underlying the phonemes of the world's languages. The higher the hierarchy where the phoneme is located, the easier/earlier it can be produced by a child. This sequence/order of the emergence of sounds usually indicates a shift from unmarked to marked features, which, due to being harder, are acquired later by children. Maturational theory implies that phones may be acquired in the same order all over the world. In addition, quantitatively speaking, the number of sounds with unmarked features is always more than or equal to those with marked features. For example, ref. [17] documented sound distributions in more than 300 languages. They found that when compared to unmarked sounds, infrequent sounds are marked, and the domain could be either within one language, usually referring to the phoneme level, or exist across languages. Furthermore, simple-articulated sounds are more frequent than complex-articulated sounds. The vowel [a] offers a good example, as it appears in all languages, whereas the vowel [ë] is relatively rare. It is well known that [i, a, and u] are canonical vowels, meaning all languages have them. As such, it is no wonder that these canonical vowels are easier models to learn. As for consonants, with respect to the MOA and POA, it is clear that stops are more consonant-like compared to other consonants (and fricatives are more easily articulated than affricates), while coronals represent the default/unmarked POA. As for tones, previous studies on the unchecked tones acquired by children were reviewed because the checked tones are seldom reported. For Mandarin Chinese, ref. [18] proposed that the high-level and falling tones are acquired before the rising and dipping tones, meaning the level tones are likely easier than the contour tones. In this case, the former maintains a consistent frequency in pitch, while the latter needs to change the pitch level at least once. Yip (2001) postulated a similar argument. In addition

to contour vs. level tones, she also proposed that rising tones are more marked than falling tones, while high-level tones are more marked than low-level tones. Her theory of tone thus predicts that all the unmarked tones are acquired earlier than the marked tones.

2. Related Work

In computational linguistics, automated models focusing on language evolution primarily concentrate on detecting cognates among different related languages [19–23] and inferring phonological correspondences [24]. These efforts aim to construct phylogenetic trees representing the relationships between languages. However, the objective is to identify the rules of sound change rather than to actually build a model capable of predicting the historical evolution of language sounds. A similar task is found in machine transliteration, which phonetically converts the words of a source language into the words of a target language. Machine transliteration can be regarded as a subtask of machine translation. Many different approaches to machine translation have been adopted to solve the task of machine transliteration. Early approaches were based on dictionaries or lexicons to map the phonemes between the source and target languages [25,26]. Later, statistical machine learning methods were adopted to learn the mapping from the source to the target languages [27,28].

Statistical machine learning constitutes a foundational framework for the domain of machine learning, deriving its principles from the disciplines of statistics and functional analysis. This theoretical construct is primarily concerned with the challenge of statistical inference, specifically the derivation of a predictive function from a given dataset. The implementation of statistical learning theory has precipitated significant advancements across a variety of domains, including but not limited to computer vision, speech recognition, and bioinformatics. Through its rigorous approach to understanding and modeling the underlying patterns within data, statistical learning theory has played a pivotal role in the development and enhancement of algorithms that facilitate complex decision-making and predictive analyses in these fields.

Ref. [27] proposed a statistical machine learning framework that facilitates direct orthographical mapping (DOM) between two distinct languages via a joint source-channel model, herein referred to as the n-gram Transliteration Model (TM). The n-gram TM model streamlines the orthographic alignment process by automatically generating aligned transliteration units from a bilingual dictionary. Employing the n-gram TM within the DOM framework significantly diminishes the effort required for system development and achieves a substantial enhancement in transliteration accuracy, surpassing the performance of contemporary state-of-the-art machine learning algorithms. The efficacy of this modeling framework is corroborated through a series of experimental validations focusing on the transliteration between the English and Chinese language pair.

Ref. [28] utilized conditional random field (CRF) models to formulate transliteration as a sequence-labeling problem. The many-to-many (m2m) aligner was used to generate character mappings between English and Arabic, and then a CRF model was trained based on the alignment results to label each English input character with a sequence of Arabic characters. CRFs are a class of statistical modeling methods often applied in pattern recognition and machine learning and are used for structured prediction. Diverging from traditional classifiers, which determine labels for individual samples in isolation, CRFs are designed to incorporate contextual information, considering the interdependencies among adjacent samples. This contextual consideration is operationalized through the formulation of a graphical model, encapsulating the dependencies among predictions. The architecture of the graph utilized is contingent upon the specific requirements of the application at hand.

In recent decades, deep learning-based models have been widely adopted for machine transliteration. Deep learning is the subset of machine learning methods based on artificial neural networks (ANNs) with representation learning. The adjective “deep” refers to the use of multiple layers in the network. The methods used can be either supervised, semi-supervised, or unsupervised. Artificial neural networks (ANNs) draw inspiration

from the paradigms of information processing and the distributed communication observed within biological systems. Despite this inspiration, ANNs exhibit several fundamental distinctions from biological brains. Notably, ANNs typically manifest as static and symbolic constructs, which is in contrast to the dynamic (plastic) and analog nature inherent to the biological brains of most living organisms. Consequently, ANNs are often regarded as rudimentary or low-fidelity models when it comes to accurately replicating the complex functionalities and adaptive capabilities of the brain.

Ref. [29] proposed a neural network model combining a convolutional neural network (CNN) and a recurrent neural network (RNN) for English-Chinese transliteration. A CNN is a type of feedforward neural network, the neurons of which can respond to a subset of the surrounding units within a certain coverage range, demonstrating outstanding performance in large-scale image processing. Composed of one or more convolutional layers atop fully connected layers (akin to classical neural networks), CNNs also incorporate pooling layers along with associated weights. This architecture allows CNNs to leverage the two-dimensional structure of input data. Compared to other deep learning structures, CNNs yield superior results in image and speech recognition tasks. Moreover, this model can be trained using the backpropagation algorithm. With fewer parameters to consider compared to other deep, feedforward neural networks, CNNs represent an attractive deep learning architecture. Moreover, RNNs represent one of the primary categories of ANNs, distinguished by the bi-directional flow of information across its layers. Unlike unidirectional feedforward neural networks, RNNs facilitate a feedback loop within their architecture, allowing outputs from certain nodes to influence the subsequent inputs to those same nodes. This distinctive feature endows RNNs with the capacity to maintain an internal state or memory, therefore enabling the processing of sequences of inputs of arbitrary lengths. This capability renders RNNs particularly suited for applications in tasks that involve sequential data, such as unsegmented, connected handwriting recognition or speech recognition. The terminology "recurrent" neural network specifically applies to networks classified by an infinite impulse response, which is in contrast to "convolutional" neural networks, which are characterized by a finite impulse response. Both types of networks demonstrate temporal dynamic behavior, which is crucial for processing time-dependent data. A network with a finite impulse response can be conceptualized as a directed acyclic graph, which permits unfolding into an equivalent strictly feedforward neural network. Conversely, a network with an infinite impulse response, due to its cyclical graph structure, cannot be unfolded in this manner, reflecting its inherent capacity for modeling complex temporal dynamics.

Ref. [30] used an attentional sequence-to-sequence (seq2seq) model for Arabic-English transliteration. The seq2seq model transforms input sequences into output sequences. It avoids the problem of vanishing gradients by utilizing RNNs or, more commonly, networks based on LSTM (long short-term memory) or GRU (gated recurrent unit) architectures. The content of a current item always stems from the output of the previous step. The seq2seq model is primarily composed of an encoder and a decoder. The encoder converts the input into a hidden state vector, which encapsulates the content of the input items. Conversely, the decoder performs the reverse process, transforming the vector back into an output sequence, and uses the output from the previous step as the input for the next step. This model was initially developed to enhance machine translation technology, allowing machines to discover and learn how to map a sentence from one language to its corresponding sentence in another language.

The Named Entity Workshop (NEWS) was established in Singapore in 2009 by the Agency for Science, Technology, and Research (A*STAR) to develop machine transliteration techniques. The NEWS compiles high-quality multilingual datasets for machine transliteration and defines metrics to evaluate performance. In this study, we have attempted to adopt statistical and deep learning machine transliteration methods to train models for sound changes in the Chaoshan dialect and utilize the metrics from the NEWS to evaluate their performance.

Typologically, linguists collected data to extract shared innovative features, which were judged to have the same/different ancient languages. Take the implementation of Chinese languages as an example; Refs. [31–34] have carried out a series of studies on the relationships between Chinese dialects. In [32]'s work, Pearson's correlation coefficients are employed within the statistical analysis to compute the presence of vocabulary and phonetic forms, categorized by the existence of sounds, rhymes, and tones. The significance of quantitative calculation methodologies in language classification resides in their establishment of a framework for gauging dialectical proximity, therefore offering a systematic approach to delineating dialect groupings. Additionally, phonological characteristics were scrutinized across 17 dialects, focusing on initial consonants [p, p^h, and b], tracing their descent from Middle Chinese. This examination encompassed considerations of nucleus (vowels), lips (rounded/unrounded), and mouths (open/closed), elucidating the classification of rimes. Tonal aspects were analyzed based on the four tones of flat, rising, falling, and entering, alongside three distinct categories of initial consonants: voiceless consonants, voiced stops and sibilants, and sonorants. It is acknowledged that the voiced or voiceless nature of the initial consonant influences the tonal pattern, a phenomenon rooted in ancient Chinese phonology. Most phonologists concur that the characteristics of the initial consonant interact with tonal variations. As a result, ref. [32] reported that Min languages exhibit greater mutual affinity compared to Hakka and Gan languages. Concerning the tones, ref. [31] observes a quantitative correlation between the pitch characteristics of yin and yang tones and the phonetic attributes of initial consonants. In an inventory encompassing 3433 dialects within 737 distinct linguistic variations, it was revealed that high tones predominate in most dialects. The falling tones are the most prevalent contour tone, while bi-directional tones are comparatively rare, which is categorized by [35] as a more marked tone. There is an observed trend wherein the dialects situated further south tend to exhibit a greater diversity of tones, whereas those located farther north tend to possess fewer tones [31].

Ref. [34] conducted a comparative analysis of vowel rhyme structures between calculated representations and the contemporary Beijing dialect. This analysis employed the notion of "communication degree" as a criterion for evaluating phonological historical constructs. For two dialects, A and B, the communication degree was determined by examining cognate words in A and B, with A serving as the source and B as the target, utilizing the "one-way communication degree" criterion. Each phonetic element, including initial consonants, glides, vowels, endings, and tones, is assigned a basic weight of one out of five, with positive (informational) and negative (noisy) values attributed accordingly. By multiplying these weights, the one-way communication degree is computed, enabling a comparison of the relative relationships between dialects. The ultimate finding indicates that in terms of communicative proficiency, the Beijing dialect holds the third position, trailing behind the Chengdu and Hankou dialects. Furthermore, ref. [34] inferred that the maximum mutual intelligibility between old and young generations, based on the index derived by [33], was typically 0.92, suggesting a generational discrepancy of up to 0.08. This discrepancy serves as a benchmark for mutual intelligibility. Moreover, the formula is extended to assess intelligibility among different dialects, where a high index indicates seamless communication and a lower index signifies potential communication barriers. For example, when the mutual intelligibility between Chengdu and Hankou reaches the highest level at 0.795, communication between speakers of these dialects poses no significant challenges. However, a decrease in mutual intelligibility to 0.475 between Beijing and Guangzhou dialects indicates potential communication difficulties [34]. By combining the findings of [33,34], and if the life expectancy difference between two generations is about 100 years, the study extrapolates a mutual intelligibility loss of 0.08 per century, offering insights into historical language dynamics. Notably, this research contributes to our understanding of mid-phonological changes and dialectical relationships.

Ref. [36] proposed and implemented a method for the study of historical phonetics. The main idea is that the speech database requires native speakers to provide a corpus,

especially in cases where there has been a gradual decline in native speakers. We are left with a question: Without fieldwork, can a prediction by a model be used to fill in unknown pronunciations? Furthermore, can this help preserve cultural heritage in the future? They applied the method to the seven major Chinese languages of the Sino-Tibetan language family: Mandarin, Wu, Cantonese, Xiang, Hakka, Gan, and Fujian. Ref. [36] translated the pronunciation of each Chinese character into IPA and marked eight phonological characteristics: tonal category, tonal value, initials, mediations, vowels, diphthongs, nasalization, and final rhymes. They proposed a new generative model, which imported hidden random variables and mapped each phonological feature to the hidden random variables. The probability distribution of the random variables produced the various possible characters for each Chinese character using a Markov chain Monte Carlo method. Finally, they used the voice data as reference data to solve the actual value of the random variable in the generative model. This method can simultaneously use material from the middle-ancient rhyme book and the possibly incomplete dialect phonetic data to explore the superlingual rhymes. This research made three contributions: (1) in addition to the rhyme data, the dialect data are very helpful for predicting the phonetics of another dialect; (2) this generative model is more effective at predicting closely related dialect data; and (3) filling out the voice data through the proposed model can improve accuracy in predicting new dialect voices.

Many previous studies have proposed adding missing speech [36] or constructing the phonetic forms of ancestral language [37]. However, while the models constructed by these methods can predict possible voice forms, they find it difficult to interpret and explore phonetic evolution. The objective of this investigation is to simulate the linguistic transformations within the Chaoshan region over the preceding century. This endeavor utilizes two dictionaries, one published in 1883 and the other in 2015, to elucidate the contemporary pronunciation dynamics during these periods. Employing the pronunciation data from 1883 as the source language and those from 2015 as the target language, the model extrapolates predicted pronunciations based on the 1883 dataset. Subsequently, a comparative analysis is conducted between these projected pronunciations and the actual phonetic changes observed in 2015, therefore simulating the linguistic evolution within the Chaoshan area over the past 100 years. As such, the questions of interest in this study were the following:

1. Do the vocalic and consonantal phones and tones generated by the three stated models (attention-based seq2seq, DirecTL+, and Sequitur) have any relationship to the patterns of vowel, consonant, and tone distribution observed in this study?
2. Do the vocalic and consonantal phones and tones generated by the three models resemble the developmental patterns of language learning by infants and, thus, reflect underlying universal constraints?

In the following parts, Section 3 introduces the methods used in this study, and Section 4 presents the primary results. Section 5 compares the results for the distribution of the data and the sequence of sound emergence in human beings. In the concluding remarks, the major findings are summarized, and potential areas of further research are noted.

3. Method

3.1. Dataset and Cleaning

We chose the dictionary edited by [38] to represent the 1883 Chaoshan dialect because each character has a corresponding pronunciation and vocabulary explanation. In the 21st-century Chaoshan dialect, we chose the Chaoshan dictionary, first edited by Zhang Xiaoshan in 2009 (with a second edition in 2015), as representative because it contains the most complete vocabulary that can be found for the Chaoshan area. Pronunciations of the same Chinese characters in each dictionary were selected. Because the literary and colloquial readings were mixed in each word and labeled via a different system, manually checking the two dictionaries was necessary. By guessing the similarities between word combinations in relation to literary and colloquial language, the first author deleted any

inconsistency between the same words with either literary or colloquial pronunciations that were mismatched in the two dictionaries, i.e., if one word with one sound in one dictionary corresponds to a word in the other dictionary, they will have only one sound. A total of 5523 words made up the dataset. Five-fold cross-validation was used for evaluation. Each dictionary was individually labeled with onset, rhyme, and tone.

3.2. Three Models

3.2.1. Attention-Based Sequence-to-Sequence Method

The sequence-to-sequence (seq2seq) model is a popular method for natural language processing. The seq2seq model is based on an encoder-decoder architecture constructed by two RNNs. One neural network is an encoder, which adopts the RNN to encode the input sequence into a context vector. This context vector is then passed to the next neural network to decode and generate an output sequence. For the RNN in the seq2qeq model, long short-term memory (LSTM) is usually adopted to avoid the vanishing gradient problem, which often happens in the training process of a general RNN. The objective function of the seq2seq model can be defined as follows:

$$p(y_1, \ldots, y_{T'} | x_1, \ldots, x_T) = \prod_{t=1}^{T'} p(y_t | v, y_1, \ldots, y_{t-1}) \quad (1)$$

where $x_1, x_2, \ldots x_T$ is the input sequence given time step T; v is the encoded context vector generated by the encoder; $y_1, \ldots, y_T$ is the previously generated output sequence; and the function is to maximize the probability of the output sequence with the next token, $y_{T'}$.

The seq2seq model encodes the entire input sequence into a vector, but it is sometimes difficult to capture enough information when the input sequence is long. Furthermore, for some applications, such as machine translation, when we translate a word in a sentence into another language, we usually focus on some important words around it rather than the entire sentence. In order to give the seq2seq model the ability to focus on some important portion of the input sequence while engaged in decoding, an "attention mechanism" is introduced. Attention is a technique that allows the neural network to devote greater focus to small but important parts of the input. In the network, each input token has its own weight. The ith input token has an attention weight, w_i. For each input token, i, the encoder RNN has its hidden state, v_i. The weighted average, $\sum_i w_i v_i$, is the output of the attention mechanism. An overview of the attention-based seq2seq model is shown in Figure 1.

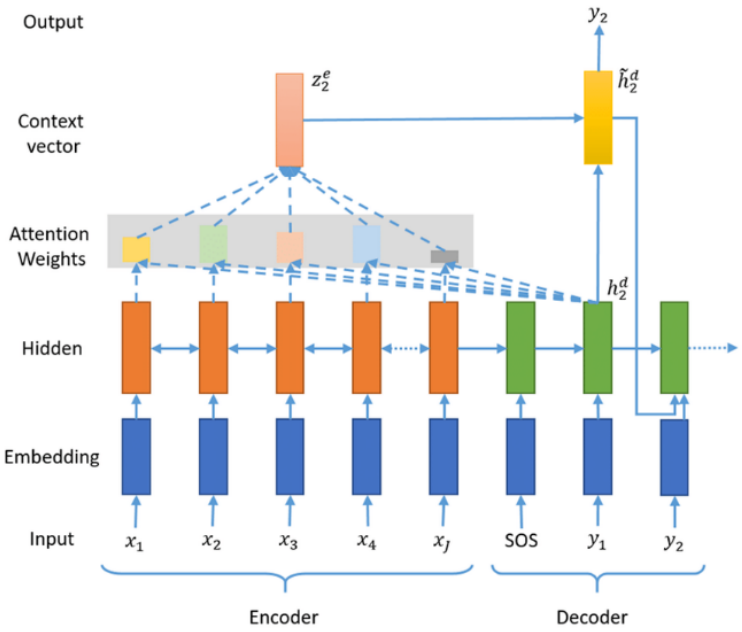

Figure 1. Attention-based sequence-to-sequence (seq2seq) model.

3.2.2. DirecTL+

DirecTL+ is an online discriminative training model for string transduction problems. It was initially developed for the grapheme-to-phoneme conversion problem by Jiampojamarn et al. in 2008. It has also been applied to name transliteration tasks. In order to train the DirecTL+ model, a many-to-many (m2m) aligner is required to align the source and target tokens. The m2m aligner, also proposed by Jiampojamarn, is adopted together with DirecTL+.

The m2m aligner is based on an expectation-maximization (EM) algorithm. The EM algorithm comprises two steps: expectation and maximization. These two steps are performed alternatively until convergence is achieved. The expectation step estimates the maximum likelihood value with the current hidden variables; the maximization step aims to find the parameters that maximize the quantity.

After obtaining the many-to-many alignment results, the DirecTL+ online discriminative training framework can be applied to train a model to convert the source sequence $\mathbf{x}$ into the target sequence $\mathbf{y}$. For each given pair, $(\mathbf{x}, \mathbf{y})$, we can define a feature vector, $\Phi(\mathbf{x}, \mathbf{y})$, representing evidence for the sequence $\mathbf{y}$ found in $\mathbf{x}$, with α as a feature weight vector providing a weight for each component of $\Phi(\mathbf{x}, \mathbf{y})$. Algorithm 1 is the algorithm for DirecTL+.

Algorithm 1 DirecTL+ algorithm pseudocode.

Require: Data $(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \ldots, (\mathbf{x}_m, \mathbf{y}_m)$
Ensure: Learned weights ψ
  $\psi := 0$
  for k iterations **do**
    for $j = 1 \ldots m$ **do**
      $\hat{Y}_j = \{\hat{y_{j1}}, \ldots, \hat{y_{jn}}\} = \underset{y}{\operatorname{argmax}}[\psi \cdot \Phi(\mathbf{x}_j, \mathbf{y})]$
      update ψ according $\hat{Y}_j$ and $\mathbf{y}_j$
    end for
  end for
  return ψ

The algorithm consists of the following three main components: a scoring model, represented by a weighted linear combination of features $(\alpha \cdot \Phi(\mathbf{x}, \mathbf{y}))$; a search for the highest scoring phoneme sequence for a given input word; and an online update equation to move the model away from incorrect outputs and towards the correct output.

For all the training data from the aligned pairs, the algorithm performs k iterations. In each iteration for each input sequence, $\mathbf{x}_j$, the model generates the n-most possible sequence, $\hat{Y}_j$. The parameters of the model are updated by the difference between $\hat{Y}_j$ and the correct answer in each iteration. The updated model is based on the margin-infused relaxed algorithm (MIRA) (Crammer and Singer, 2003).

3.2.3. Sequitur

Sequitur is a data-driven translation tool originally developed for grapheme-to-phoneme conversion by [39]. It is based on the joint source-channel model. The joint source-channel model was first introduced by Li et al. (2004). When given paired sequences X and Y, where x and y are representative of their segment units, the conversion process seeks to find the alignment for the subsequences of the input string X and the output string Y. This can be represented for an n-gram model as the following:

$$\begin{aligned} P(X,Y) &= P(x_1, x_2, \ldots, x_k, y_1, y_2, \ldots, y_k) \\ &= P(< x_1, y_1 >, \ldots, < x_k, y_k >) \\ &= \prod_{i=1}^{k} P(< x, y >_i \mid < x, y >_{i-n+1}^{i-1}) \end{aligned} \tag{2}$$

where k is the number of alignment units, and $P(X, Y)$ is the joint probability of the ith alignment pair, which depends on n previous pairs in the sequence.

In this study, we applied three different models to learn changes in Chaoshan pronunciation. Although these three models are machine learning-based, the design of their architectures is different. For the DirecTL+ model, the alignments between the sources and targets must be completed before training the string transduction model. Thus, the settings of the m2m aligner, such as the maximum number of tokens that can be aligned, affect the performance of the alignments and propagate to the DirecTL+ model. The Sequitur model is different from many translation models, such as DirecTL+, as it is able to train a joint n-gram model from unaligned data. Additionally, the model can be built up cumulatively. The higher order n-grams are trained iteratively from the smaller ones—first, a unigram model is trained, this is then used for a bi-gram model, and so on.

The seq2seq model takes another perspective to deal with this problem. Instead of learning the alignment or correspondence between the subsequences of the input and output strings, it adopts a neural network to encode all of the input strings into a vector, which condenses and embeds all its information. Then, another neural network is used to decode this vector to produce the output. This architecture can consider the whole input string, which means that the seq2seq model can learn more constraints across the entire input sequence, not just the adjacent tokens.

4. Evaluation

4.1. Settings

In the seq2seq model, both the input and output consist of characters from orthographic sequences represented in dictionaries. In the encoder, the maximum length of input was set to 10. The output dimension of the Embedding layer was fixed at 100, and the hidden layer was constructed using LSTM cells, where the latent dimension of each cell is set to 256. During the training of the seq2seq model, the Adam optimizer was employed, with the loss function being categorical entropy. The number of epochs was set to 100, and an early stopping mechanism was utilized.

In the DirecTL+ approach, within the m2m aligner, the maximum limit for the correspondence between the input and output sequences was set to 2. This is because, in the orthography of both Chaoshan dictionaries, a single phoneme is represented by, at most, two characters. Regarding the configuration for training the DirecTL+ model, the context size was set to 3, the number of training iterations was established at 10, and the n-Best setting was also fixed at 10.

For the Sequitur model, we utilized the implementation provided in its official GitHub repository (https://github.com/sequitur-g2p/sequitur-g2p, accessed on 22 June 2023). The training of the Sequitur model is incremental; we employed a training procedure that progresses from Uni-gram to 4-gram, resulting in the final model.

4.2. Experimental Results

In this section, the results of the attention-based sequence-to-sequence method, DirecTL+, and Sequitur are reported. In order to further analyze the results, we separated the predicted accuracy of the sounds into three categories: onset, rhyme, and tone.

After manually deleting inconsistencies, we adopted the attention-based sequence-to-sequence method, DirecTL+, and Sequitur to train the model. We employed a dataset consisting of 5523 of the same words with known pronunciations in both dictionaries. The pronunciations from 1883 instances were utilized as input, and the three models were trained accordingly. Subsequently, the predicted words from the models were compared with those also present in the 2005 dataset to assess the accuracy of the models. The accuracy rates of the three models, the attention-based sequence-to-sequence method, DirecTL+, and Sequitur, are presented in Table 1. Additionally, the results for the evaluation were categorized into three groups: accuracy with tones, accuracy without tones, and F1-score.

In the evaluation, we employed the commonly used metrics in machine transliteration: Accuracy, Recall, Precision, and F1-score. The definition of accuracy is as follows:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N}\left\{\begin{array}{ll} 1 & \text{if } r_i = c_i \\ 0 & \text{otherwise} \end{array}\right\} \quad (3)$$

where N represents the number of data in the test dataset, r_i is the phoneme string predicted by the model, and c_i is the correct phoneme string. An $Accuracy = 1$ indicates that all model predictions match the actual dictionary answers perfectly, whereas $Accuracy = 0$ indicates that none of the predictions match the dictionary answers. The definitions of Recall, Precision, and F1-score are estimated based on the longest common subsequence (LCS) of string comparisons. The definition of LCS is as follows:

$$LCS(c, r) = \frac{1}{2}(|c| + |r| - ED(c, r)) \quad (4)$$

where ED is the "string edit distance", which measures how many insertions, deletions, or substitutions are required to transform one string into another. It is a common method for estimating the similarity between two strings. The definitions of Precision, Recall, and F1-score are as follows:

$$Precision = \frac{LCS(c_i, r_i)}{|c_i|} \quad (5)$$

$$Recall = \frac{LCS(c_i, r_i)}{|r_i|} \quad (6)$$

$$F1\text{-score} = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (7)$$

Five-fold cross-validation was used for evaluation, and the accuracy of our model was up to 68%.

Table 1. Results for the evaluation of the three models in relation to the Chaoshan dataset.

Method	**seq2seq**			**DirecTL+**			**Sequitur**		
Fold	**Accuracy (with Tone)**	**Accuracy (w/o Tone)**	**F1-Score**	**Accuracy (with Tone)**	**Accuracy (w/o Tone)**	**F1-Score**	**Accuracy (with Tone)**	**Accuracy (w/o Tone)**	**F1-Score**
0	0.6742	0.819	0.9362	0.6471	0.8072	0.9307	0.6778	0.8118	0.9377
1	0.7023	0.8344	0.9473	0.6624	0.8072	0.9394	0.6950	0.8190	0.9450
2	0.6561	0.8235	0.9362	0.5891	0.7575	0.9237	0.6579	0.8145	0.9370
3	0.7032	0.8244	0.9404	0.6326	0.7846	0.9270	0.6959	0.8190	0.9392
4	0.6706	0.8199	0.9389	0.6308	0.7864	0.9314	0.6715	0.8145	0.9393
Average	0.6813	0.8242	0.9398	0.6324	0.7886	0.9304	0.6796	0.8157	0.9396

Generally speaking, speech flow can be dissected into several syllables, with the segments constituting the fundamental elements. Apart from segments, supra-segmentals provide another perspective on the analysis of speech flow. Upon further examination of the segments, they can be subdivided into vowels and consonants, and supra-segmentals encompass aspects such as tone, duration, stress, and intonation. Consequently, our dataset was scrutinized in terms of syllables, vowels, consonants, and tones.

Syllables, the basic units of speech, typically consist of consonants (C) and vowels (V), with common constructions being CV, CVC, and so forth, though specifics vary across languages. Vowels are characterized by tongue height, tongue backness, and lip rounding. Additionally, consonants exhibit features such as POA, MOA, and voicing (or not). In Chinese, aspiration (or not) can serve as a distinctive feature.

Tones, which are fundamental components of speech in tonal languages such as Chinese, include level tones and contour tones. Elements such as tonal register and contour

are analyzed to understand tone patterns. The interaction between tone and consonants in final positions can lead to checked tones characterized by shorter duration. Accordingly, the subsequent analysis of the results will be conducted based on these delineated elements, encompassing syllables, vowels, consonants, and tones, to provide a comprehensive understanding of the dataset.

Table 2 shows the sounds predicted by the model compared to those found in the 2015 Chaoshan dictionary, using the same characters as the 1883 Chaoshan dictionary as inputs. In terms of accuracy, the following syllable structures were predicted (going from high to low): CVE > CV > CṼ > CṼṼ, with the nature of the vowels playing a role. This means that if the vowel is nasalized, a simple syllable structure, CṼ, is more accurately predicted than a complex one, such as CṼṼ. However, if the vowel is plain, the accuracy of CVE—where E could be either a consonant or vowel—is greater than that of CV. The syllabified consonants [m] and [ŋ] are the least accurate. Thus, we can further examine the distribution of different vowels when following the sonorant sequence of CGVE, where V stands for the peak of the syllable. The results showed that regardless of the vowel's nature, the prediction of nasalized vowels was less accurate than for oral vowels. More specifically, the front vowels were easier to correctly predict than the back vowels. One should also note the performance of ê[ɯ] and ên[ẽ], with approximately the same accuracy of 54%, and [õ], with no accuracy.

Table 2. Accuracy of the three models compared to the 2015 Chaoshan dictionary.

Category	Subcategories/Description	Tendency
Syllable structure	CVE > CV > CṼ > CṼṼ	VS (64%) > VN (61%) > VV(58%) > V(55%) > Ṽ (25%) > ṼṼ(21%) > syllabified consonants (5%)
Vowel/ Nucleus	V > Ṽ; front V > back V	i[i] (65%) > ê[e](64%) > a[a] (61%) > u[u](57%) > e[ɯ] (55%) > o [o] (49%) ên[ẽ] (54%) > in[ĩ] (20%) > an[ã] (17%) > un[ũ] (15%) > syllable consonants (8%) > on[õ](0%)
Consonants		
Onset	manner of articulation (MOA)	lateral (61%) > fricatives (59%), affricates(59%), nasals (59%) > zero onset (57%) > stops (56%)
Onset	place of articulation (POA)	glottal (60%) > velar (59%) > bi-labial (58%) > zero onset (58%) > coronal (57%)
Onset	voiced feature	Voiced (60%) > voiceless (57%)
Onset	aspirated feature	unaspirated (58%) > aspirated (54%)
Tones		
Tone	tonal contour [1]	level tone (T1, T5, T7) (57%) > contour tone (T2, T3, T6) (55%)
Tone	initial tonal register [2]	H initial T value (T2, T5, T8) (65%) > L initial T value (T3, T4, T7) (49%)
Tone	duration	checked tone (T4, T8) (64%) > unchecked tone (T1, T2,T3,T5,T6,T7) (57%)

[1] Checked tones are not included. [2] T1 /33/ and T6 /35/ are not included.

The consonants were analyzed based on their MOA and POA, whether they were voiced or not, and whether they were aspirated or not. Overall, we found no significant difference in terms of accuracy in these four different aspects. Regarding the MOA, the accuracy varied from 61% to 56%. One should note that stops, which are usually considered to be acquired first in children, showed the lowest accuracy. As for the POA, the accuracy

rates varied from 60% to 57%. One should note that coronal unmarked features are considered to be acquired first or with the least restriction on their combination with the lowest rate of accuracy. Voiced consonants were easier to predict correctly compared to voiceless ones; in contrast, the aspirated consonants were harder to predict correctly. The difference between the voiced/voiceless and aspirated/unaspirated consonants was the smallest, with a value of 3–4%.

Finally, tonal accuracy was analyzed in terms of tone contour, register, and duration. The accuracy of the prediction of level tones was found to be higher than for contour tones, indicating no significant difference. However, the predicted value for the H initial tone, at 3 on the five-point Chao scale, was higher than that of the L initial tonal value (65% vs. 49%). The checked tone accuracy (64%) was higher than the unchecked tone accuracy (57%); the former is always acquired later by children.

To sum up, the predicted sounds made by the machine learning models displayed very different phenomena compared to human language learning. The differences between the model-generated sounds and the human-acquired sounds were, thus, compared.

In machine learning tasks, training a complicated neural network model such as seq2seq requires a substantial amount of data to establish an effective model. However, due to the limited number of entries in the two Chaoshan dictionaries currently available, only 5523 data items were available for training and testing. Although this quantity may seem somewhat insufficient for training a seq2seq model, it is quite adequate for traditional machine learning models such as DirecTL+ and Sequitur, especially for machine transliteration tasks. The experimental results indicate that the performance of seq2seq models still surpasses that of DirecTL+ and Sequitur. Therefore, the issue of having fewer data should be considered negligible.

5. Discussion

5.1. The Sounds Generated by the Models Are Governed by the Sequence

We found that the accuracy of the Chaoshan dictionary, as generated by the three models—attention-based seq2seq, DirecTL+, and Sequitur—highlights, to some degree, what machine learning can achieve. However, their performance was not found to be at the same level as the process of human language learning, i.e., proceeding from unmarked features to marked ones.

More specifically, strong consonants are defined as more obstruent when airflow passes the oral cavity, i.e., they are more consonant-like. In this definition, aspirated, stop, and voiceless consonants are strong compared to their unaspirated, lateral, and voiced counterparts. According to this definition, the consonants generated by the models are weak consonants rather than being more consonant-like.

Table 3 shows the accuracy of the model-generated sounds and the distribution of the vowels following different consonants with respect to the MOA. For example, the distribution of vowels occurring after the stops was about 39% among the 5523 items, ranking first. As such, because there was a greater chance for different vowels to follow stops, the prediction made by the three models offered the lowest accuracy, with a value of 56%, ranking sixth. On the contrary, if the MOA was lateral, the prediction made by the three models achieved its highest value at 61%. However, the possibility of candidate vowels following a lateral gave the lowest value at 7%. Even though the ranking order varies according to the model accuracy and the distribution of vowels following each subcategory's consonants, the tendency is clear: the more following candidate vowels, the lower the chance that the three models can predict the consonants according to their MOA.

Table 3. Comparison between the model-generated speech and consonants with respect to the MOA and their following vowel distribution.

Subcategory	Model_acc	Rank	Distribution	Rank
Stops	56%	6	39%	1
Fricatives	59%	2	20%	2
Affricates	59%	2	16%	3
Nasals	59%	2	8%	5
Lateral	61%	1	7%	6
Zero onset	57%	5	10%	4

Table 4 shows the accuracy of the model-generated speech and the distribution of the vowels following different consonants with respect to the POA. For example, the distribution of vowels occurring after a coronal is about 47% (5523 items), ranking first. As such, because there was a greater chance of different vowels following the coronal, the prediction of the three models gave the lowest value at 56%, ranking fifth. On the contrary, if the POA was glottal, the prediction made by the three models achieved its highest value at 60%. However, the prediction of the possible candidate vowels following a glottal gave the lowest value at 10%. Even though the ranking order varies between the model accuracy and the distribution of the vowels following each subcategory's consonants, the tendency is not particularly clear. For zero onset, both the prediction of the three models (57%) and the distribution of the following vowels (10%) ranked last. The argument that the more following vowel candidates there are, the less chance the three models can predict the consonants when divided with respect to their place of articulation only holds for the first subcategory.

Table 4. Comparison between the model-generated speech and consonants with respect to the POA and their following vowel distribution.

Subcategory	Model_acc	Rank	Distribution	Rank
Labial	58%	3	13%	3
Coronal	56%	5	47%	1
Velar	60%	1	19%	2
Glottal	60%	1	10%	4
Zero onset	57%	4	10%	4

Table 5 shows the accuracy of the model-generated speech and the distribution of vowels following different consonants with respect to being voiced. For example, the distribution of vowels occurring after a voiceless consonant was about 71% (5523 items), ranking first. As such, because there is a greater chance for different vowels to follow voiceless consonants, the prediction made by the three models gave the lowest value at 57%, ranking second. In contrast, if the consonant was specified as being voiced, the prediction made by the three models achieved its highest value at 60%. However, the prediction of possible vowel candidates following a voiced consonant gave the lowest value at 19%. Even though the ranking order varied between the model accuracy and the distribution of vowels following each subcategory's consonant, the tendency is clear. For zero onset, the prediction made by the three models (57%) and the distribution of following vowels (10%) both ranked last. The argument that the more following vowel candidates there are, the smaller the chance the three models can predict the consonant when divided with respect to their voiced feature still holds.

Table 5. Comparison between the model-generated speech and consonants with respect to the voiced feature and their following vowel distribution.

Subcategory	Model_acc	Rank	Distribution	Rank
Voiceless	57%	2	71%	1
Voiced	60%	1	19%	2
Zero onset	57%	2	10%	3

Table 6 shows the accuracy of the model-generated speech and the distribution of vowels following different consonants with respect to the aspired feature. For example, the distribution of vowels occurring after an unaspirated consonant was about 30% (5523 items), ranking first. In contrast to the MOA, POA, and voiced feature, the greater the chance of different vowels following an unaspirated consonant, the higher the prediction value made by the three models, with 58%, also ranking first. This means the same pattern was found both for the model accuracy and the following vowel distributions.

Table 6. Comparison between the model-generated speech and consonants with respect to the aspirated feature and their following vowel distribution.

Subcategory	Model_acc	Rank	Distribution	Rank
Unaspirated	58%	1	30%	1
Aspirated	54%	2	21%	2

To sum up, a similar pattern was revealed both in the prediction made by the three models and the distribution of the following vowels, showing that the behaviors of the major features—the MOA, POA, and voicing—differ from those of the minor and aspirated features.

One may speculate that the onset prediction of the three models and the distribution of the following vowels may be related to the sonority sequence. If we postulate that the sonority hierarchy is essentially uniform across languages, once a word has a legal syllable structure in one particular language, meaning that its minimal sonority distance in this language is met, it partially fulfills the phonotactic constraint. If so, the reason why some syllable combinations are less common is probably due to a preference for making the sonority distance as large as possible. Thus, when selecting the onset in Chaoshan, for example, a stop is preferred to a nasal and then a liquid because the following component is either a glide or a vowel, which are always ranked as the most sonorant. This regulation of phonotactic constraint in Chaoshan determines the word combination.

Concerning feature geometry, if the aspirated consonant is described as [+/− spread glottis] under the supervision of the laryngeal and [+/− consonant; +/− sonorant], we may wonder about how the behaviors of consonants with aspirated features fit with the MOA, POA, and voicing. This brings us to the next question: Why are aspirated consonants synchronized with the distribution of the following? However, only MOA, POA, and voiced features manifest a similar pattern. One might postulate that POA, MOA, and voicing are major features with which to describe consonants, whereas aspirated features are not. Of course, more research is needed to investigate this issue more fully.

Table 7 shows the accuracy of the model-generated speech and the distribution of the consonants preceding different types of vowels. We may take VN as an example. The consonants preceding the VN type constitute about 34% (5523 items), ranking first, whereas the accuracy of VN is 59%, ranking second. In contrast, in relation to the accuracy of consonants and the distribution of the following vowels, consonants (the left-to-right sequence) indicate sufficient prediction accuracy; vowels (the right-to-left sequence) do not have such prediction. The sequence determines the accuracy of the models and the one (consonants in this study) on the left-hand side with fewer possible following sounds, where we can see the higher accuracy of this consonant (L → R). However, if the vowel is the pivot, no such phenomenon can be observed.

Table 7. Comparison between the model-generated speech and different types of vowels and their preceding consonant distribution.

Subcategory	Model_acc	Rank	Distribution	Rank
V	52%	4	21%	3
Ṽ	24%	5	1.5%	6
syllable consonants	0%	7	0.2%	7
VV	57%	3	25%	2
ṼṼ	19%	6	2.8%	5
VS	66%	1	16%	4
VN	59%	2	34%	1

Table 8 shows the accuracy of the model-generated speech and the distribution of the consonants preceding vowels of different natures. For example, the consonants preceding the V-i type account for about 35% (5523 items), ranking first, whereas the accuracy of V-i is 62%, ranking third. If the vowel is the pivot, the prediction is insufficiently accurate. This means that consonants (the left-to-right sequence) work, but this is not the case for vowels (the right-to-left sequence).

Table 8. Comparison between the model-generated speech and different natures of vowels and their preceding consonant distribution.

Subcategory	Model_acc	Rank	Distribution	Rank
V-a	73%	1	17%	3
V-e	55%	4	4%	6
V-ê	64%	2	9%	5
V-i	62%	3	35%	1
V-o	47%	6	13%	4
V-u	55%	4	22%	2
Syllabified consonants	8%	7	0.2%	7

To sum up, the accuracy of the prediction of consonants is negatively correlated with the distribution of the following vowels, whereas that of vowels is not.

As for tones, Table 9 shows the tone accuracy of the models and the distribution of each tone (5523 items). Taking T5 (55) as an example, the tone distribution was about 20%, ranking first, whereas the accuracy of T5 (55) was 67%, ranking first as well. We generalize this by saying that there is a partially positive correlation between the tones and their distribution, meaning the greater the distribution of the tone, the better chance it can be correctly predicted. The odd one out is T8 (4), for which the distribution was pretty low, ranking seventh, but the accuracy of which was high, ranking second.

Table 9. Comparison between the accuracy of each tone generated by the models and tone distribution.

Subcategory	Model_acc	Rank	Distribution	Rank
T1 (33)	59%	5	20%	1
T2 (52)	63%	3	16%	3
T3 (213)	50%	7	13%	4
T4 (2)	63%	3	10%	5
T5 (55)	67%	1	20%	1
T6 (35)	51%	6	10%	5
T7 (11)	20%	8	4%	8
T8 (4)	66%	2	6%	7

Table 10 shows the accuracy of the tones generated by the models in terms of tone contour, register, duration, and their distributions (5523 items). When looking at the level

tones, including T1 (33), T5 (55), and T7 (11), the distribution of this type of tone was about 45%, ranking first, and the accuracy of its prediction was 57%, ranking first as well. We can generalize that there is a partially positive correlation between the prediction of the tones and their distribution, meaning the greater the distribution of the tone, the better the chance of it being correctly predicted. In relation to different tonal durations for unchecked vs. checked tones, even though the checked tones have far smaller distributions than the unchecked versions (16% vs. 84%), the accuracy of tone prediction for the checked tones was greater at 64%.

Table 10. Comparison between tone accuracy for tone contour, register, and duration, generated by the models and tone distribution.

Tone Contour	Model_acc	Rank	Distribution	Rank
Level tone (T1 (33)+ T5 (55)+ T7 (11))	57%	1	45%	1
Contour tone (T2 (52)+ T3 (213)+ T6 (35))	55%	2	38%	2
Tone register				
H initial T value (T2 (52)+ T5 (55)+T8 (4))	65%	1	42%	1
L initial T value (T3 (213)+ T7 (11)+T4 (2))	50%	2	27%	2
Tone duration				
Unchecked tone (T1 (33)+ T2 (52)+ T3 (213)+T5 (55)+ T6 (35)+ T7 (11))	57%	2	84%	1
Checked tone (T4 (2)+ T8 (4))	64%	1	16%	2

5.2. Different Mechanism for the Sounds Generated by the Models and for Those Acquired by Human Beings

[15] claims that "the relative chronological order of development remains everywhere and at all times the same". This means that the pace of development in children may vary, but the order in which the sounds are acquired seems universal. Our study reveals a general pattern for the order in which sounds emerge in children: stops are acquired before affricates, and velars are acquired later, while level tones are acquired earlier than contour tones. However, the sounds correctly generated by the three models display approximately the opposite relationship. Ref. [15] also claims that "only those sounds which are common to all the languages of the world, while those phonemes which distinguish the mother tongue from the other languages of the world appear only later". He also suggests that identical laws operate in the phonological development of language in children and the synchronic structure of the world's languages. As such, unmarked sounds are acquired earlier than marked sounds since unmarked sounds appear in most languages of the world, and the acquisition of some sounds presupposes the acquisition of other sounds. Ref. [15] also provides evidence to show the mirror image of the developmental sequence of language acquisition in the study of aphasia, which is, to some degree, in line with our model-generated sounds. In this sense, we would further argue that model-generated sounds are determined by the sequence of word combinations from left to right. Only the onsets generated by the models present a mirror image of the sequence of children's language acquisition because the phonotactic constraints of word combinations restrict their appearance. The more common this onset is, the more variety is found in its following unit, which is either a glide or a vowel. This means that the more unmarked the onset, the better the chance/probability that it can be preceded by other units. If this is the case, it shows why the prediction achieved by our model is less accurate, resulting in lower accuracy in relation to onset production.

At first glance, hearing-impaired children seem to be affected by their physiological disorder, for example, with reduced ability to hear high-frequency sounds such as stops. However, Ref. [40] found that the English allophone [z] may have three different meanings when combined with other morphemes. Take the word "cook" as an example, where

(a) 'cooks' is the plural form, (b) 'cook's hat' is the possessive case, and (c) 'mummy cooks' is the inflectional form of the verb, 'cook'. The acquisition sequence in children goes from (a) to (c). In the case of a patient with aphasia, the sequence of loss was as follows: first, the suffix of the verb, (c), then the suffix of the possessive, (b), and finally, the suffix of the plural noun, (a). The order of aphasic loss went from the sentence level (c) to the phrasal level (b) and then the word level (a), i.e., from the most complex form to a simple form. The results generated by the three models, to some degree, reveal a similar phenomenon.

Therefore, this means that, under the legal syllable structure/word condition, the more specific the feature the onset contains, the more easily the models can predict it. This is the same as the ability of the aphasia patient. By considering the constraints on markedness, the sounds that are acquired last by humans (because of their markedness) parallel the accuracy of the predictions made by the models, i.e., these are also predicted last due to the large number of potential follow-up units.

To sum up, while the acquisition process in children is determined by frequency, we did find some parallels between the machine learning models and the human acquisition process regarding minor features, such as aspiration. However, the major features MOA and POA, demonstrate the opposite. When taken together, the machine-generated sounds, governed by the distribution of following units from left to right, are applied to onsets only. In addition, part of the results parallel the human acquisition of sounds, primarily determined by the input frequency as a holistic lexicon learning process.

Notably, not all the sounds generated by the models reveal the same marked hypothesis. The nasalized vowels provide an example. They are rarer than oral vowels, and their presence within a system presupposes that of oral vowels. According to [17], the set of nasal vowels is never larger than the set of oral vowels, fulfilling the definition of the markedness category.

6. Conclusions and Further Research

We compared the sounds predicted by the models and found that the accuracy rate (from high to low) regarding the model-generated sounds and the process of learning sounds in human beings is very different. The accuracy of the model-generated sounds is based on the amount of input, i.e., the greater the input, the better the performance. The successive emergence of sounds reported by different scholars is quite different across languages. In language acquisition, the process of moving from unmarked sounds to marked sounds in terms of their features, POA, MOA, and level tones is not some descriptive linguistic phenomenon. This reflects the real-world functions of this mechanism, where frequency is the key to acquiring a language. However, the opposite phenomenon related to the sounds generated by the models when moving from the highest to the lowest accuracy is clear. This means the mechanism used by the models is totally different from the one used by human beings. The distribution or input frequency really plays a role in the models. However, this distribution may not be the primary key to language acquisition in human beings. People present another set of mechanisms: learning by unmarked vs. marked incidence of a single tone (slot), so it will be the opposite.

More specifically, the left-to-right sequence (consonants) indicates sufficient prediction accuracy, whereas the right-to-left sequence (vowels) will not yield such accurate predictions. The accuracy is partially related to the distribution, and only consonants can indicate what follows. Vowels manifest such a pattern, whereas the sequence of the words has nothing to do with children's language acquisition. As such, there are significant differences between the machine-generated sounds and the human-acquired sounds.

This pilot study of adapting computational methods to sound model generation and human acquisition did not successfully demonstrate a positive correlation. Certainly, advancing our comprehension of speech models remains a significant endeavor worthy of continued pursuit. The simulation of speech models may not always align perfectly with the process of language acquisition in humans. Therefore, striving for a more ef-

fective interpretation of the speech models in production remains a crucial objective for future efforts.

On the other hand, we are aware that the literary and colloquial readings are not inevitable, especially in the Min languages. The stratum between the two readings is not clear-cut, and understanding how to label a one-to-one correspondence for literary and colloquial readings of the same character is the key to this study. Uncovering the use of the super-stratum and the substratum is the goal of further research.

Author Contributions: Conceptualization, M.-N.C. and Y.-C.W.; methodology, Y.-C.W.; software, Y.-C.W.; validation, M.-N.C. and Y.-C.W.; formal analysis, M.-N.C.; investigation, M.-N.C.; data curation, M.-N.C.; writing—original draft preparation, M.-N.C. and Y.-C.W.; writing—review and editing, M.-N.C. and Y.-C.W.; project administration, M.-N.C.; funding acquisition, M.-N.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by National Science and Technology Council, Taiwan grant number NSTC 111-2410-H-030-0414- and MOST110-2410-H-030-035.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to the copyright issues.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gubian, M.; Cronenberg, J.; Harrington, J. Phonetic and phonological sound changes in an agent-based model. *Speech Commun.* **2023**, *147*, 93–115. [CrossRef]
2. Chang, G. *History of Min and Hakka Dialects*; Nantian Bookstore: Taipei, Taiwan, 1996.
3. Zhou, C. *The Formation, Development and Spread of Southern Min in Taiwan*; Taili Publishing House: Taipei, Taiwan, 1996.
4. Karlgren, B. *Compendium of Phonetics in Ancient and Archaic Chinese; The Bulletin of the Museum of Far Eastern Antiquities, Stockholm.no. 26*; The Museum of Far Eastern Antiquities: Stockholm, Sweden, 1954.
5. Li, R.; Yao, R. *Southern Min Chinese Dialect*; Fujian People's Publishing House: Fuzhou, China, 2008.
6. Lin, L.; Chen, X. *A Study on the Phonetics of the Min Dialect in Guangdong;* Shantou University Press: Shantou, China, 1996.
7. Chao, Y. *Language Problems;* Taiwan Commercial Press: Taipei, Taiwan, 1968.
8. Chu, M. Motivating the Change of Stop Codas in ChaoShan A Perceptual Study. Ph.D. Thesis, National Tsing Hua University, Hsinchu, Taiwan, 2009.
9. Xu, Y. The Teochew Dialect Phonology in the Nineteenth Century. *J. Inst. Chin. Cult.* **2013**, *57*, 223–244.
10. Xu, Y. A study on the evolution of the finals in Chaozhou dialect over the past 100 year. *Philos. Linguist.* **2016**, *1*, 241–260.
11. Hsu, J.H. *A Study of the Stages of Development and Acquisition of Mandarin Chinese by Children in Taiwan*; Crane Publishing: Taipei, Taiwan, 1996.
12. Jeng, J.Y. The Speech Acquisition of Mandarin-Speaking Preschool Children. *J. Chin. Lang. Teach.* **2017**, *14*, 109–136.
13. Hsu, H.C. Phonological Acquisition of Taiwanese a Longitudinal Case Study. Ph.D. Thesis, National Tsing Hua University, Hsinchu, Taiwan, 1989.
14. So, L.K.; Dodd, B.J. The acquisition of phonology by Cantonese-speaking children. *J. Child Lang.* **1995**, *22*, 473–495. [CrossRef] [PubMed]
15. Jakobson, R. *Child Language, Aphasia and Phonological Universals*; Mouton: The Hague, The Netherlands, 1968.
16. Smit, A.B.; Hand, L.; Freilinger, J.J.; Bernthal, J.E.; Bird, A. The Iowa articulation norms project and its Nebraska replication. *J. Speech Hearth Disord.* **1990**, *55*, 779–798. [CrossRef] [PubMed]
17. Ladefoged, P.; Maddieson, I. *The Sounds of the World's Languages*; Blackwell: Oxford, UK, 1996.
18. Li, C.N.; Thompson, S.A. The acquisition of tone in Mandarin-speaking children. *J. Child Lang.* **1977**, *4*, 185–199. [CrossRef]
19. Inkpen, D.; Frunza, O.; Kondrak, G. Automatic identification of cognates and false friends in French and English. In Proceedings of the International Conference Recent Advances in Natural Language Processing, Hissar, Bulgaria, 5–11 September 2005; Volume 9, pp. 251–257.
20. List, J.M. LexStat: Automatic detection of cognates in multilingual wordlists. In Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH, Avignon, France, 23 April 2012; pp. 117–125.
21. Rama, T. Siamese convolutional networks for cognate identification. In Proceedings of the COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, 11–16 December 2016; pp. 1018–1027.

22. Jäger, G.; List, J.M.; Sofroniev, P. Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Valencia, Spain, 3–7 April 2017; pp. 1205–1216.
23. Dellert, J. Combining information-weighted sequence alignment and sound correspondence models for improved cognate detection. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 3123–3133.
24. Hruschka, D.J.; Branford, S.; Smith, E.D.; Wilkins, J.; Meade, A.; Pagel, M.; Bhattacharya, T. Detecting regular sound changes in linguistics as events of concerted evolution. *Curr. Biol.* **2015**, *25*, 1–9. [CrossRef] [PubMed]
25. Knight, K.; Graehl, J. Machine transliteration. *arXiv* **1997**, arXiv:cmp-lg/9704003.
26. Stalls, B.G.; Knight, K. Translating names and technical terms in Arabic text. In Proceedings of the Computational Approaches to Semitic Languages, Montreal, QC, Canada, 16 August 1998.
27. Li, H.; Zhang, M.; Su, J. A joint source-channel model for machine transliteration. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04), Barcelona, Spain, 21–26 July 2004; pp. 159–166.
28. Ammar, W.; Dyer, C.; Smith, N.A. Transliteration by sequence labeling with lattice encodings and reranking. In Proceedings of the 4th Named Entity Workshop (NEWS) 2012, Jeju, Republic of Korea, 12 July 2012; pp. 66–70.
29. Shao, Y.; Nivre, J. Applying neural networks to English-Chinese named entity transliteration. In Proceedings of the Sixth Named Entity Workshop, Berlin, Germany, 12 August 2016; pp. 73–77.
30. Rosca, M.; Breuel, T. Sequence-to-sequence neural network models for transliteration. *arXiv* **2016**, arXiv:1610.09565.
31. Cheng, C.C. A quantitative study of Chinese tones. *J. Chin. Linguist.* **1973**, *1*, 93–110.
32. Cheng, C.C. Quantifying affinity among Chinese dialects. *J. Chin. Linguist. Monogr. Ser.* **1991**, 3, 76–110.
33. Cheng, C.C. Phonological generation gap. In Proceedings of the The Ninth Annual Conference of the International Association of Chinese Linguistics, Singapore, 26–28 June 2000.
34. Cheng, C.C. Language Intelligibility As a Constraint on Phonological Change. *Dialect Var. Chin.* **2003**, *1*, 81–95.
35. Wang, S.-Y. Phonological features of tone. *Int. J. Am. Linguist.* **1967**, *33*, 93–105.
36. Lin, C.C.; Tsai, R.T.H. A Generative Data Augmentation Model for Enhancing Chinese Dialect Pronunciation Prediction. *IEEE Trans. Audio Speech Lang. Process.* **2011**, *20*, 1109–1117. [CrossRef]
37. Bouchard-Côté, A.; Hall, D.; Griffiths, T.L.; Klein, D. Automated reconstruction of ancient languages using probabilistic models of sound change. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 4224–4229. [CrossRef] [PubMed]
38. Goddard, J. *A Chinese and English Vocabulary, in the Tie-Chiu Dialect*; American Presbyterian Mission Press: Shanghai, China, 1883.
39. Bisani, M.; Ney, H. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Commun.* **2008**, *50*, 434–451. [CrossRef]
40. Jakobson, R. *Studies on Child Language and Aphasia*; Mouton: The Hague, The Netherlands, 1971.