JASA ARTICLE

Deep learning assessment of syllable affiliation of intervocalic consonants

Zirui Liu[a] and Yi Xu
Speech, Hearing, and Phonetic Sciences, University College London, London, WC1N 1PJ, United Kingdom

ABSTRACT:
In English, a sentence like "He made out our intentions." could be misperceived as "He may doubt our intentions." because the coda /d/ sounds like it has become the onset of the next syllable. The nature and occurrence condition of this resyllabification phenomenon are unclear, however. Previous empirical studies mainly relied on listener judgment, limited acoustic evidence, such as voice onset time, or average formant values to determine the occurrence of resyllabification. This study tested the hypothesis that resyllabification is a coarticulatory reorganisation that realigns the coda consonant with the vowel of the next syllable. Deep learning in conjunction with dynamic time warping (DTW) was used to assess syllable affiliation of intervocalic consonants. The results suggest that convolutional neural network- and recurrent neural network-based models can detect cases of resyllabification using Mel-frequency spectrograms. DTW analysis shows that neural network inferred resyllabified sequences are acoustically more similar to their onset counterparts than their canonical productions. A binary classifier further suggests that, similar to the genuine onsets, the inferred resyllabified coda consonants are coarticulated with the following vowel. These results are interpreted with an account of resyllabification as a speech-rate-dependent coarticulatory reorganisation mechanism in speech. © *2023 Acoustical Society of America*.
https://doi.org/10.1121/10.0017117
(Received 19 August 2022; revised 11 January 2023; accepted 17 January 2023; published online 3 February 2023)
[Editor: Ewa Jacewicz] Pages: 848–866

I. INTRODUCTION

Despite the wide recognition of the syllable as a speech unit among speakers and researchers (Browman and Goldstein, 1992; Levelt, Roelofs and Meyer, 1999; MacNeilage, 1998), there have been doubts about the role of the syllable due to ambiguity associated with syllable boundaries. One situation where ambiguity is especially severe is in regard to the syllable affiliation of intervocalic consonants. For example, the phrase "escort us" in British English (/ɛ s#kɪɔ t#əs/) can be syllabified as /ɛ s#kɪɔ #tə s/ in connected speech, according to observation of a noisy release during the word final /t/ (Levelt *et al.*, 1999). The phenomenon is more formally known as resyllabification, which usually denotes a shift of syllabification of a coda consonant into the onset of the following vowel-initial syllable (Levelt *et al.*, 1999; Schiller *et al.*, 1997). For English, empirical work examining resyllabification goes back as far as 70 years ago, when Stetson used the kymograph to investigate consonant vowel (CV) and vowel consonant (VC) production at different speech rates (Stetson, 1951). He observed that in a sequence of syllables such as /bi bi bi…/, the CV structure remains stable regardless of speech rate. In contrast, a sequence of VC syllables, such as /ib ib ib…/, becomes very similar to /bi bi bi…/ when repeated at a fast rate, according to kymograph data, indicating that the coda /b/ is resyllabified as an onset consonant. The perceptual finding was consistent with articulatory patterns recorded by the kymograph. Stetson's findings were later replicated by Tuller and Kelso (1990, 1991) with glottal transillumination data, which showed that glottal movements shifted drastically at a critical rate of speech, and perception of the spoken sequences also shifted to be mostly identified as /ip ip ip…/.

In languages such as Spanish and French (Bermúdez-Otero, 2011; Gaskell *et al.*, 2002), resyllabification is recognised as a phonological process, although there are –cross dialect variations according to acoustic evidence such as consonantal duration (Strycharczuk and Kohlberger, 2016). Due to the lack of clear empirical evidence, the existence of resyllabification in English is questioned (Shattuck-Hufnagel, 2011), as mentioned above. Furthermore, the status of the syllable is called into question because of boundary ambiguity due to resyllabification (Blevins, 2003; Steriade, 1999). A major source of the difficulty of determining the syllabification status of segments is that it is mainly based on the subjective judgment of listeners (Ní Chiosáin *et al.*, 2012; Content, 2001; Goslin and Frauenfelder, 2001; Schiller *et al.*, 1997). Even when acoustic measurements are taken, listener judgments are still treated as the "ground truth" (de Jong *et al.*, 2004; Mullooly, 2003). Yet, as found in de Jong *et al.* (2004), listeners agree with each other well only in cases in which a gap between the release of the coda consonant and the beginning of voicing for the next vowel can be easily detected. In the absence of apparent gaps, listener judgments

[a]Electronic mail: zirui.liu.17@ucl.ac.uk

https://doi.org/10.1121/10.0017117

become very diverse. de Jong *et al.* (2004), therefore, suggested that the difference between the coda and onset consonant is more closely related to how they are *motorically optimised* in production in ways that are too subtle for most listeners to detect.

What is needed is an alternative definition of resyllabification which departs from conventional definitions that are based on language-specific phonotactics (what is phonologically legal), perceptual impression, and language-specific acoustic properties (aspiration, voicing, etc.). In this study, we consider an articulatory-acoustic definition that specifies the affiliation of an intervocalic consonant based on an articulatory definition of the syllable. The definition of the syllable, as will be reviewed next, also addresses coarticulation, another essential issue of speech articulation.

A. Resyllabification, coarticulation, and the syllable

Resyllabification is closely related to a well-documented asymmetry between onset and coda consonants in phonology and phonetics. For languages that allow for coda consonants, codas are more vulnerable than their onset counterparts as they are more susceptible to deletion and reduction (Barlow and Gierut, 1999; Xu, 1986, 2020). In contrast, onset consonants are often inserted when the syllable is vowel initial, such as glottal stop insertion (Birgit, 2001; Garellek, 2012), intrusive /r/'s (Gick, 1999; Uffmann, 2007), and vowel hiatus breakers (Mudzingwa, 2013; Smith, 2001). In terms of canonical syllable structures, CV syllables are also more common than VC and consonant vowel consonant (CVC) syllables in many languages (Clements and Keyser, 1983; Levelt *et al.*, 1999; Xu, 2020).

According to articulatory phonology, the vulnerability of codas is likely related to an asymmetry in coarticulation within the syllable. That is, onset consonants are coupled "in-phase" with the vowel, resulting in synchronous activation between the vocalic and onset C gestures (Goldstein *et al.*, 2006). On the other hand, coda consonants are coupled "antiphase" with the vowel, which is a less stable mode of coordination. Resyllabification is, therefore, "analysed as an abrupt transition to a more stable coordination mode" that is likely to occur under increased speaking rate (Goldstein *et al.*, 2006, p. 237).

An alternative account of resyllabification is provided by the synchronisation model of the syllable (Xu, 2020), as shown in Fig. 1, which shares some similarities with articulatory phonology but differs from it in certain critical details. The model assumes that syllable is a mechanism for eliminating most of the temporal degrees of freedom by synchronising consonant, vowel, and glottal movements at syllable onset (vertical lines), whereby each movement (dotted lines) is to approach an underlying target within its allocated time interval. The synchronisation makes the initial consonant fully overlapped, hence, coarticulated with the initial portion of the "following" vowel. In contrast, a coda consonant is articulated sequentially after the vowel because its closing movement directly conflicts with the opening movement of the vowel (Xu and Liu, 2006). There are two differences between this model and articulatory phonology that

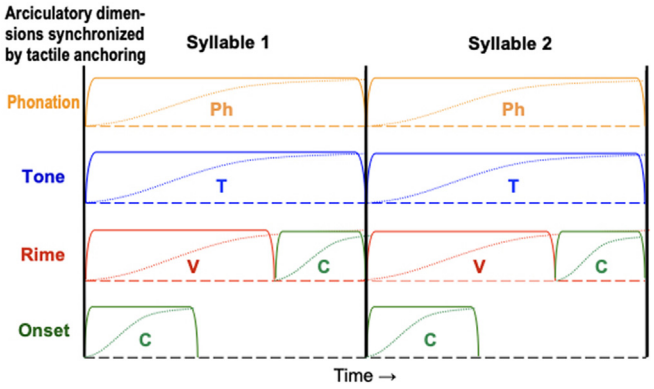

FIG. 1. (Color online) The synchronisation model of the syllable (Xu, 2020).

are directly relevant for the current study. First, synchronisation is assumed to be a fundamental design of the syllable (likely centrally controlled) rather than emerging from the coupling of the gestural planning oscillators as in articulatory phonology (Goldstein *et al.*, 2006). Second, the sequential articulation of coda consonant is not modelled in terms of phase relation between C and V because (a) individual target approximation movements are frequently allocated an insufficient amount of times (Nakatani *et al.*, 1981; Xu and Wang, 2009), thus, disallowing them to form complete movement cycles (Xu and Prom-on, 2019), and (b) syllables constantly vary their duration due to stress, phrasing, and other linguistic factors, which makes it difficult for syllable sequences, together with their constituent segments, to be temporally periodic to make oscillation-based modelling possible.

According to the synchronisation model of the syllable, resyllabification is due to a lack of articulation time, as schematised in Fig. 2 rather than due to transition from antiphase to in-phase articulatory coordination. In Fig. 2(A), the coda consonant (C_2) occupies its own time interval because it is sequentially articulated after the first vowel (V_1). Meanwhile, the second syllable is not articulated as a true

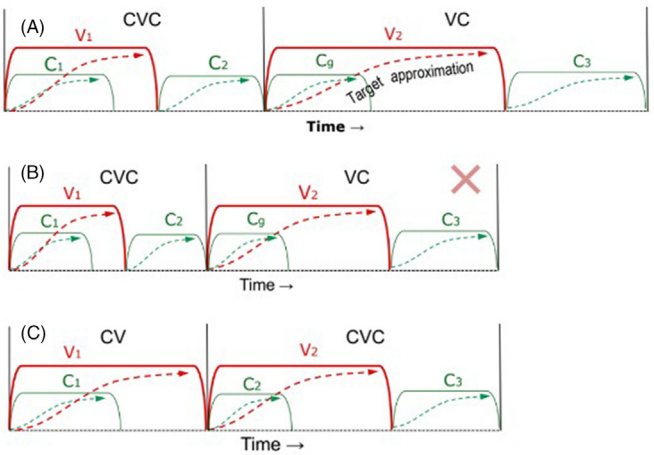

FIG. 2. (Color online) An illustration of articulatory resyllabification based on the synchronisation model of the syllable.

VC because it actually starts with a glottal stop (C_g). Such glottal stops have been reported as frequently occurring at slow speech rate (Birgit, 2001; de Jong, 2001) but would disappear as speech rate reached a certain threshold, leading to a perceptual shift from /VC#VC/ to /CV#CV/ (de Jong, 2001). As illustrated in Fig. 2(B), as speech rate increases, less time is allocated to the syllable, which would require the duration for V_1 and C_2 to be shortened to an implausible extent (as indicated by the red cross). The increased time pressure (Tiffany, 1980; Xu and Prom-on, 2019) may then lead to the replacement of the glottal stop (C_g) with C_2 when speech rate approaches a certain threshold (e.g., 350 ms per syllable; de Jong, 2001). Now, C_2 becomes the initial consonant of the second syllable, as shown in Fig. 2(C). This reorganisation gives V_1 more articulation time while preserving all of the segmental composition of the original syllables.

Based on this account of resyllabification, two predictions can be made. (1) Due to similarity in articulatory structure, resyllabified codas spectrally resemble their onset counterparts more than their canonical form, and the opposite can be observed for the neural network inferred non-resyllabified sequences (correctly classified coda sequences). (2) Because a resyllabified coda is fully coarticulated with the vowel of the following syllable, there is similar amount of vowel information shared between the resyllabified onsets and the canonical onsets but not between canonical codas and canonical onsets. These predictions can be tested on English by applying machine learning models on acoustic data.

B. Using deep neural networks with acoustic data to identify resyllabification

Given the difficulty of subjectively judging the occurrence of reyllabification (de Jong *et al.*, 2004), an alternative is to obtain objective evidence by taking advantage of recent developments in machine learning technology. This study, therefore, aims to determine the occurrences of resyllabification using deep learning models and dynamic time warping (DTW) in combination with continuous acoustic data. The deep learning models used were inspired by state-of-the-art automatic speech recognition (ASR) networks (Amodei *et al.*, 2015). ASR systems without language models are error prone when detecting the canonical structure of resyllabified sequences (Adda-Decker *et al.*, 2002; Mirzaei *et al.*, 2018; Wu *et al.*, 1997). For example, a sequence like "fade out" could be recognised as "fay doubt" if the coda /d/ is resyllabified as the onset of the second syllable. We trained recognition networks on slow speech data with no resyllabification occurrences and used them to classify data from normal rate speech. The reason behind using data from the slow speech rate condition for training is to ensure that there are no resyllabified sequences in the training data. In other words, for the model to be able to misclassify a sequence as its onset counterpart due to resyllabification, it should not be trained with a resyllabified sequence labelled as its canonical version. The misclassified sequences in normal speech rate (i.e., "fade out" as "fay doubt") were further examined to shed some light on the articulatory structure of the syllable.

II. METHODS

We trained a deep neural network classifier to identify word sequences such as "coo part" and "coop art." The utterances in the slow condition were used for training the classifiers. Then, we used the trained classifiers to classify the same utterances spoken in the normal rate recordings. A /CVC#VC/ sequence, such as "coop art," was categorised as resyllabified if the classifier "misclassified" it as its counterpart /CV#CVC/ sequence, i.e., "coo part." These neural network inferred resyllabified sequences are referred to as NN-resyllabified to avoid confusion between the cognitive process of syllable reorganisation and the inferred syllabification status by the classifier. DTW was next used to investigate the spectral similarities between the NN-resyllabified sequences in the normal speaking rate and the sequences in the slow rate (e.g., NN-resyllabified "coop art" vs slow "coo part" or NN-resyllabified "coop art" vs non resyllabified slow "coop art"). Furthermore, to test prediction (2), we built binary neural network classifiers to categorise contrastive pairs, such as "coop art" vs "coop eat," whose training data only consisted of the intervocalic consonantal portions of the acoustic signal (e.g., aspiration for /p/). The closure interval was not included due to very little acoustic energy in the data, as /p/ is a voiceless stop. The results were compared between speech rates and syllable structures.

A. Subjects

Eight subjects aged 20–40 years old, whose first language was Southern Standard British English (6 female and 2 males), participated in this study. No speaking or hearing disorders were reported prior to recording. To ensure data quality, all of the potential participants had to submit a short recording on Gorilla (Berlin, Germany). The experimenters then visually inspected the recordings in the computer program Praat (Boersma and Weenink, 2022). Only participants with an external microphone and sufficient recording quality took part in the study.

B. Stimuli and data collection

Table I lists the word sequences used in this study. The stimuli include three groups of four sequences. For each group, the onset pair and coda pair match in terms of segments and differ in syllable structure, e.g., /CVC#VC/ vs /CV#CVC/. This maximises the possibility that if the classifier misclassified a coda sequence as its onset counterpart, it is likely due to the shift in syllable structure, i.e., resyllabification.

TABLE I. Stimuli.

Group	Onset		Coda	
1	Lee steal	Lee stale	Least eel	Least ale
2	Do mart	Do meet	Doom art	Doom eat
3	Coo part	Coo Pete	Coop art	Coop eat

https://doi.org/10.1121/10.0017117

Note that there exist differences other than syllabification between onset and coda sequences, such as lexical, syntactic, or prosodic properties. For example, "doom art" is a noun/verb noun sequence, whereas "do mart" is a verb noun sequence. The neural network classifier could use information such as syllabification, syntactic, and lexical differences between the onset and coda tokens. Therefore, it is important to minimise the *similarities* between items such as "coo part" and "coop art" due to the following: If the classifier misclassified "coop art" as "coo part," it is important to minimise the possibility that the misclassification took place due to prosodic or lexical similarity between the two rather than coarticulation between the intervocalic C and the second V. Therefore, within each onset and coda pair, we use word combinations that differ in their morphosyntactic structure (e.g., "Lee steal" vs "least eel"). However, other unknown factors may still result in similarities between the onset and coda pairs, which could contribute to misclassification. The current design can only assume that when a coda sequence is misclassified as its onset counterpart, it is due to similarity in coarticulation structure rather than other unknown factors.

There is also a vowel minimal contrast in the second syllable for each syllable structure condition in each group. The vowel contrast allows us to examine the amount of coarticulation in the intervening consonant by assessing the performance of a binary classifier at predicting the second vowel identity using only acoustic data from the annotated consonant interval. Previous studies have used a minimal pair design and showed that when a consonant is coarticulated with the upcoming vowel, acoustic information associated with the vowel can be detected during the consonant (Liu and Xu, 2021, Liu *et al.*, 2022). Liu and Xu (2021) also show that the entire cluster in /clusterV/ syllables in British English is coarticulated with the vowel. Thus, a cluster triplet is included in the current study to investigate whether the following vowel is coarticulated from the onset of the consonant cluster.

Participants were instructed to say the word sequences in isolation in two blocks of different speaking rates—first slow, then normal. For the slow block, the speakers were instructed to articulate the words clearly and fluently at a slow pace. In the normal condition, speakers were informed to speak at a faster pace in a colloquial style. There were no instructions on what resyllabification was or whether they should or should not resyllabify anything. The stimuli were read aloud with 20 and 10 repetitions for the randomised slow and normal blocks, respectively, yielding 360 tokens per speaker ($12 \times 20 + 12 \times 10$). Around 3% of the data were excluded as a result of background noise during recording.

The recording took place online over Zoom (San Jose, CA) with the sampling rate of 32 kHz with Zoom's original sound feature turned on, which preserved the original recording quality by minimising the amount of audio enhancement. All of the participants used an external microphone during the experiment, and the recording quality was assessed by the researcher prior to the experiment. For the resyllabification classifiers, the recordings were annotated in either $[C_1V_1\#C_2V_2C_2]$ or $[C_1V_1C_1\#V_2C_2]$ format (subscripts denote syllable position), where the first boundary is the start of acoustic landmark of onset C_1 (e.g., lateral murmur for /l/) and the second boundary is the end of acoustic landmark of the coda C_2. For the binary classifiers, the consonantal intervals were segmented as the plosive aspiration for /p/, nasal murmur for /m/, and frication for /s/. An example is displayed in Fig. 3.

C. Speech rate analysis

As speech tempo can be speaker specific due to difference in speaker characteristic (Jacewicz *et al.*, 2009), participants were free to speak at a rate they deemed appropriate as slow or normal. For the slow and normal rate conditions, participants were instructed to speak fluently (i.e., without spontaneous pausing). No spontaneous pauses were

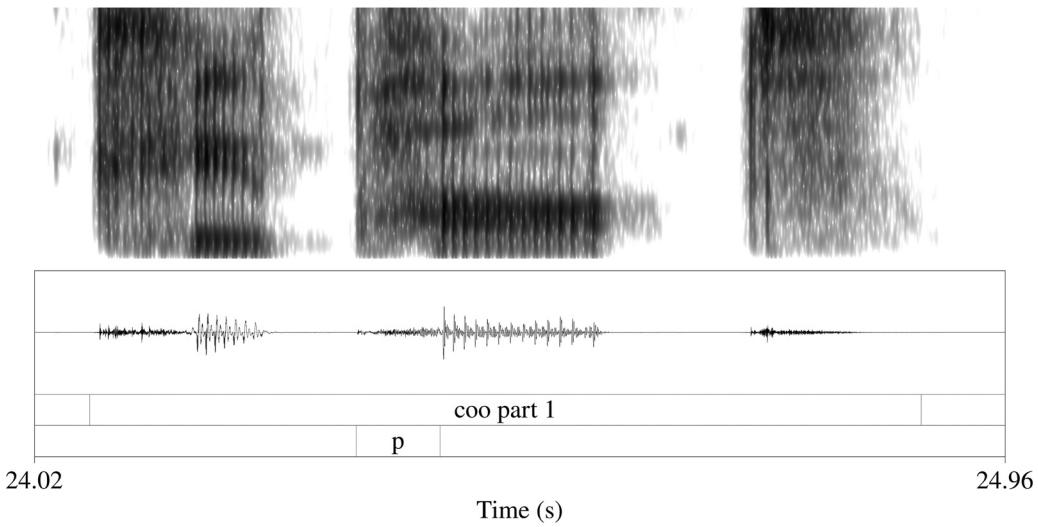

FIG. 3. An annotation example of "coo part" from one speaker, where the vertical lines indicate the segmentations.

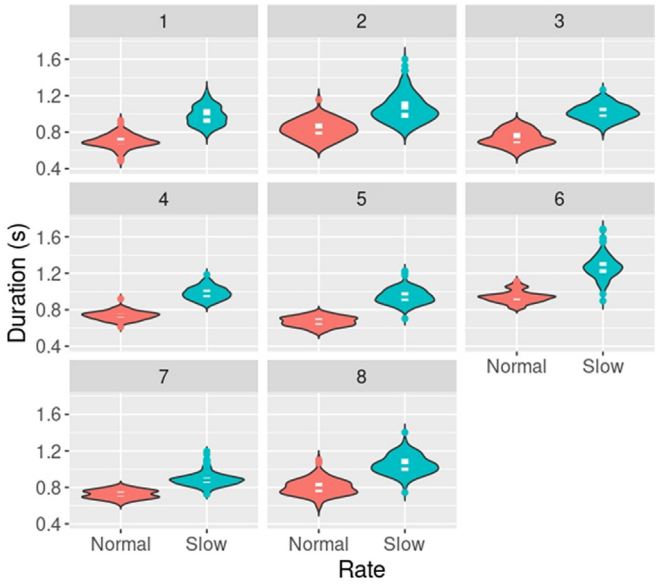

FIG. 4. (Color online) An annotated sequence duration for eight speakers.

identified in the data during the annotation process. Therefore, speech rate in the present study is analogous to articulation rate, which does not include hesitation, pausing, or emotional expressions. The duration values of annotated tokens are presented in Fig. 4. As Fig. 4 shows, speech rate was faster for the normal condition compared to the slow condition for all of the speakers. On average, speakers produced 2.9 syllables per second for the normal rate and two syllables per second for the slow rate. According to de Jong (2001), resyllabification should take place when articulation rate approaches 2.8 syllables per second.

D. Neural network classifier for identifying resyllabification

1. Data preparation

To ensure high accuracy, neural networks were trained for each speaker individually. The segmented word sequences from the slow condition were converted into mel-frequency spectrograms with 40 mel filter-banks with 25 ms as the window length and a hopping interval of 5 ms. We augmented the data to boost the amount of training data by using common augmentation techniques such as speed augmentation, noise addition, and frequency/time masking (Ko *et al.*, 2015; Park *et al.*, 2019). First, half of the tokens from the speaker were selected and sped up randomly between the factors of 0.3 and 0.9 by using the Audacity software with a custom Python script (Audacity Team, 2021). This resulted in 360 samples per speaker. Then, 15% of the resultant dataset was reserved as the testing set ($N = 54$), and 85% was reserved as the training set ($N = 306$).[1] Note that the samples were randomised before data splitting. Since the original data are balanced between word classes, the train and test split should also contain approximately balanced data, resultant of the random sampling process. The training set was then further boosted by augmenting 30% with random Gaussian noise in addition to the raw acoustic signal (Pervaiz *et al.*, 2020) or frequency or time masking to the spectrograms (Park *et al.*, 2019), yielding 398 samples for the training set. Not only does data augmentation improve model generalisation and performance, the sped-up samples also familiarise the model with shorter acoustic signals such as those in the normal speech rate condition. The motivation for doing noise addition and masking boost after the speed boost is to provide the benefit of these augmentation techniques for the sped-up tokens as well rather than just the original slow sequences.

2. Model architecture

The model architecture is shown in Fig. 5,[2] which was inspired by a combination of Deep Speech and ResNet, developed by Baidu (Beijing, China; Amodei *et al.*, 2015) and Microsoft (Redmond, WA; He *et al.*, 2015), respectively. Each model was trained for 120 epochs unless the average accuracy across the last 5 epochs has reached the threshold of 98% for the testing set. For each epoch, the spectrograms were padded to the same duration as the longest sequence in the batch ($N = 32$), and then fed into the neural network. Note that Fig. 5 demonstrates the flow of data through the network by a batch size of one. The spectrogram is first passed through a two-dimensional (2D) convolutional layer [i.e., convolutional neural network (CNN)], which had a 3×3 kernel with a stride of 1 and 32 channels. The output from the 2D convolutional layer is then passed through three residual blocks (He *et al.*, 2015), the convolutional layers in each residual block had a 5×5 kernel with a stride of one. For the 2D convolutional and residual layers, padding was used to retain the shape of the tensors. The motivation behind these two types of convolutional layers is for the model to extract features, such as dynamic information of spectral energy, between frequencies or time steps (e.g., velocity of energy variation between time steps; Luo *et al.*, 2018; Sharma *et al.*, 2020). To preserve as much acoustic information as possible, no pooling was used. The output from the residual layers was reshaped by collapsing the 32 channels, resulting in tensors with the shape of 1280 $\times$ n timesteps, which was further reduced by a fully connected layer with 512 units. Five layers of bidirectional gated recurrent units (GRU) were then used to process the sequential acoustic features. Only the last time step's output was used from the GRU. Finally, the output was fed into two fully connected layers with a final SoftMax activation, which generated the 12-dimensional probability vector, 1 vector for each word sequence in Table I. Due to the complexity of the model, we used dropout as the regularisation technique to combat overfitting (Semeniuta *et al.*, 2016). A dropout rate of 0.1 was used throughout the network (see Fig. 5 for dropout locations). Furthermore, batch normalisation was applied after each mini batch to stabilise learning as well as provide some regularisation effect (Ioffe and Szegedy, 2015). The hyperparameters were tuned by using

https://doi.org/10.1121/10.0017117

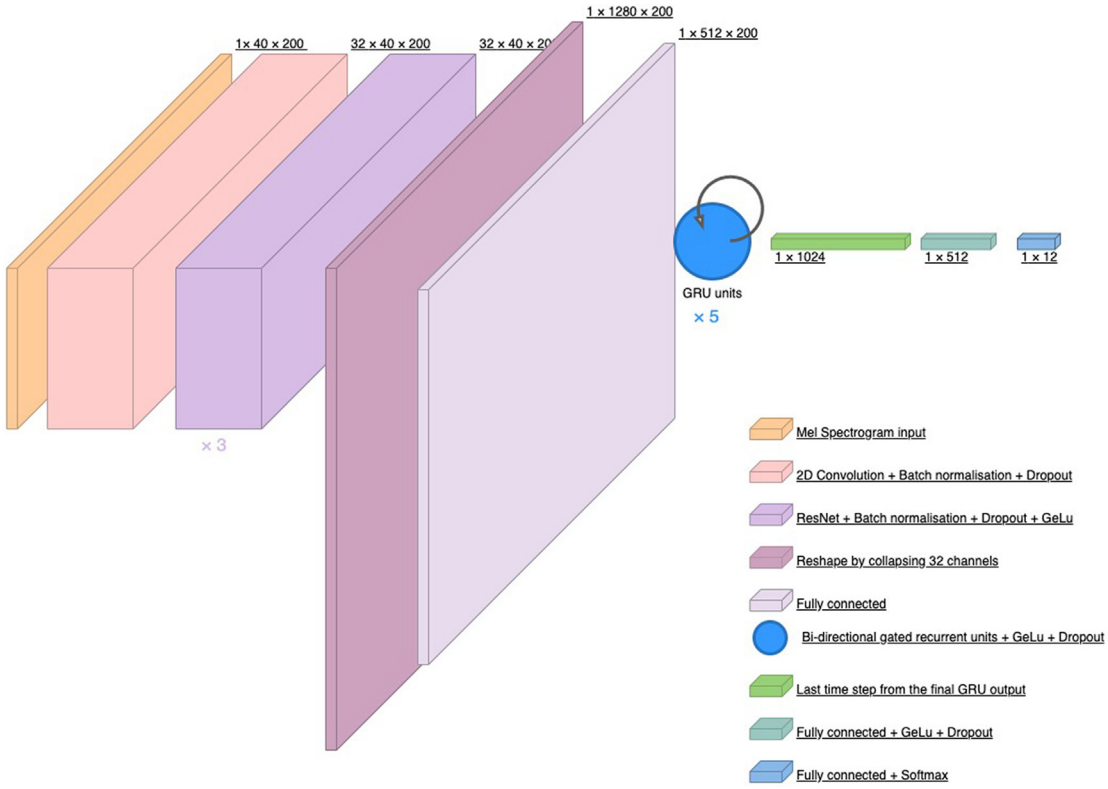

FIG. 5. (Color online) Model architecture for the resyllabification classifier. The tensor dimensions for a batch size of one are shown. The box sizes reflect tensor shapes as annotated above each box. The depth, height, and width of the boxes are not to scale and for illustration purposes only.

grid search with data from the pilot study. The hyperparameters used can be found in Table II in the Appendix.

The trained models were used to classify tokens from the normal speech rate condition for each speaker. If a coda sequence was misclassified as its onset counterpart (e.g., "coop art" classified as "coo part"), we categorised it as resyllabified.

E. DTW analysis

DTW was used to measure how similar the NN-resyllabified and non NN-resyllabified tokens were in relation to the onset or coda conditions in the slow speech rate condition. DTW has been demonstrated to be effective at measuring similarity between sequences such as acoustic signals. For example, it has been widely used for speech recognition (Sakoe and Chiba, 1978; Zhang *et al.*, 2014), as well as other applications such as bird song recognition (Kogan and Margoliash, 1998), speech segment clustering (Lerato and Niesler, 2019), and accent quantification (Bartelds *et al.*, 2020). The DTW algorithm is illustrated in Fig. 6. First, a cost matrix is computed by measuring the distance between the feature vectors (in this case, we used mel-

TABLE II. Hyperparameters for the multi-class classifiers.

Hyperparameter	Value
Number of residual blocks	3
Number of GRU layers	4
Number of units in the GRU layers	512
Number of units in the linear layers	512
Dropout rate	0.1
Number of channels for the CNN layers	32
Batch size	32
Learning rate	0.0001
Optimiser	RMSprop
Epoch number	120

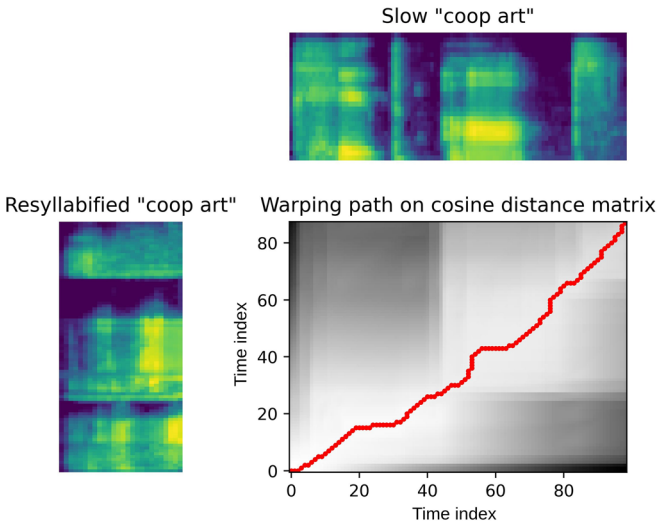

FIG. 6. (Color online) A demonstration of the DTW algorithm. The dotted line shows the dynamic warping path. The spectrograms are mel-spectrograms of the tokens "coop art" (bottom left) and "coop art" (top). The pixel intensity in the lower right heatmap represent feature distances at each time step between the two spectrograms.

spectrograms) between two sequences at each time step. We used cosine similarity for the calculation of distance as it is not affected by the magnitude of spectral energy, i.e., frequency decibels (e.g., the same recording played at different volumes would measure zero in cosine distance but not in Euclidean distance). The lower right heatmap in Fig. 6 shows the cosine distance between the mel-frequency vectors in the two sequences at all time steps. DTW works by finding the path in the distance matrix that results in the lowest cumulative distance (i.e., cost). Therefore, the DTW distance between the two sequences in Fig. 6 is the sum of the distance values through the warping path shown by the red line.

Using DTW, we can compute the similarity between word sequences while minimising the effect of speech tempo. For this study, we calculated the distances between the NN-resyllabified as well the non NN-resyllabified coda sequences and their onset and coda counterparts in the same group from the slow rate condition (e.g., NN-resyllabified "coop art" vs slow "coo part," "coo Pete," or NN-resyllabified "coop art" vs slow "coop art" and "coop eat"). Note that because the vowel contrast is constant between the distance comparisons, it should not confound the analysis.

The DTW analysis was used to compare the similarities between the NN-resyllabified sequences and the onset and coda sequences in the slow condition. In addition, a parallel DTW analysis was conducted for the non NN-resyllabified (correctly classified normal rate coda sequences) to assess whether they are more similar to their canonical form.

F. Detecting V_2 information in the intervocalic consonant

As illustrated in Fig. 3, the researcher manually segmented the canonical acoustic intervals from the intervocalic consonant or the first cluster component (i.e., nasal murmur for /m/, aspiration for /p/, and frication for /s/), which were used to investigate the articulatory alignment of the consonant and the following vowel. The segmented intervals differ in terms of articulatory meaning between groups as aspiration corresponds to the consonantal release gesture and nasal murmur and frication correspond to consonantal closures. This difference should have an impact on the amount of vowel information detected in each group. Similar to methods used in Tilsen (2020), Tilsen *et al.* (2021), and Liu and Xu (2021) to detect vowel information in the segmented intervocalic C, we trained a simple recurrent neural network (RNN) to predict the second vowel identity between contrastive pairs (e.g., NN-resyllabified "coop art" vs NN-resyllabified "coop eat"). Liu and Xu (2021) showed that for tautosyllabic C_nV, binary classifiers are able to detect vowel information in the acoustic intervals of onset C, such as during frication or lateral murmur.

For each minimal pair, tokens from all eight speakers were used. From the normal speech rate condition, only the NN-resyllabified tokens and true onset tokens were examined. According to results from the neural network classifiers, not all of the coda tokens were NN-resyllabified, which gave rise to the possibility of accuracy scores from the onset conditions being higher than the NN-resyllabified codas as a result of having significantly more training data. For example, a speaker resyllabified 5 out of 10 repetitions of "coop art" and "coop eat," which would result in 10 samples in total for the neural network, whereas 20 samples are available for the onset condition (i.e., 10 repetitions of "coo part" and "coo Pete"). Therefore, we balanced the sample sizes between the two conditions by randomly subsampling the onset tokens for each speaker to match the number of NN-resyllabified tokens. For instance, if a speaker resyllabified five out of ten repetitions of "coop eat," only 5 random selections of "coo Pete" were used from this speaker for training the binary classifier.

The classifiers were bidirectional RNNs with long short-term memory (LSTM) units (Soltau *et al.*, 2016). The network details appear in Fig. 7. The hyperparameters were tuned with data from the pilot study using grid search, and details can be found in Table III in the Appendix. The segmented tokens were converted into mel-spectrogams with 26 filter-banks with 0.025 s as the window length and 0.005 s as the hop length. Before training, all of the spectrograms were padded to the same length as the longest one. As Fig. 7 shows, masking was applied in the input layer, which tells the model to ignore the padded duration. Due to the absence of CNN, we included delta coefficients (i.e., first-order differentials) to aid model performance, which resulted in a 52-dimensional vector at each time step. The data were split into training and testing splits with the ratio of

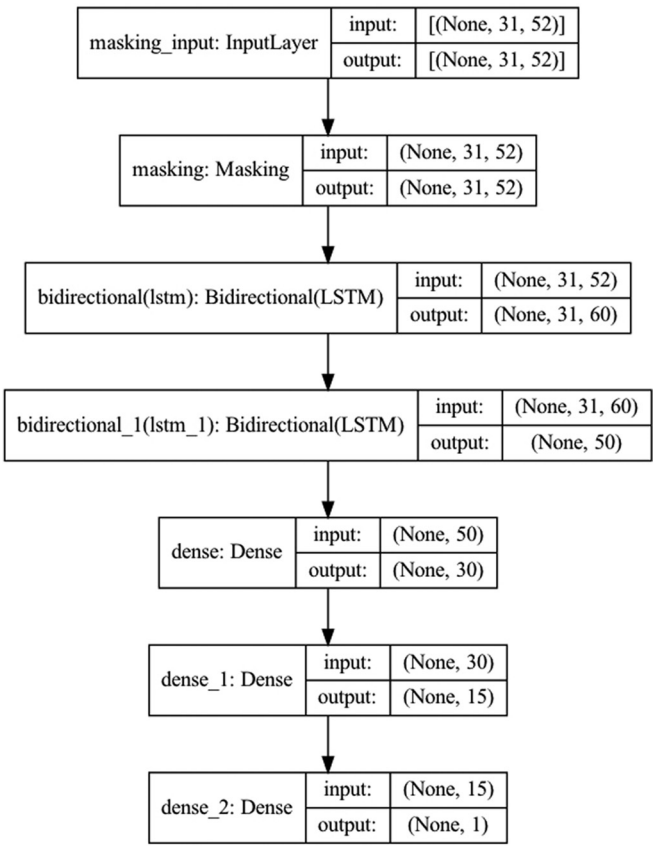

FIG. 7. The model architecture of the binary classifiers. The tensor shapes are denoted on the right of each box.

https://doi.org/10.1121/10.0017117

TABLE III. Hyperparameters for the binary classifiers.

Hyperparameter	Value
Number of units in the first LSTM layer	60
Number of units in the second LSTM layer	30
Dropout rate for the first LSTM layer	0.1
Dropout rate for the second LSTM layer	0.2
Number of units in the linear layer	50
Merge mode	Summation
Batch size	16
Optimiser	Adam
Learning rate	0.001
Epoch number	70

8:2. We randomly shuffled the data for each minimal pair and trained a model from scratch 80 times and reported the accuracy distribution on the testing sets. The motivation behind examining an accuracy distribution is to avoid the issue of accidental above chance performance, which could arise with small datasets (Combrisson and Jerbi, 2015; Ojala and Garriga, 2009).

1. Bayesian analysis

To test the amount of vowel information in the acoustic signal, we used Bayesian analysis with beta likelihood to model the effect of syllable structure (i.e., onset vs coda) on model accuracy. A conventional nonsignificant result cannot be used to validate a null hypothesis as it only suggests a failure to reject it. The advantage of using Bayesian statistics is that it simply tells us which model is more supported by the evidence in the data, and the models do not need to be nested. The motivation behind using beta regression is a result of the nature of accuracy rate being bounded between zero and one. Beta regression assumes that the data generating process can be modelled by a beta distribution (Balakrishnan and Nevzorov, 2003 in which the distribution can be parameterized with the mean-precision (μ-φ) parameters, where φ is analogous to the inverse of data dispersion. Because $Y \sim \text{Beta}(\mu,\varphi)$, beta regression presumes that the mean μ of the response given the predictor X is linear on the logit transformed scale (Douma and Weedon, 2019). In other words, in a beta regression model, the dependent variable can be mapped from the bounded space [0,1] to unbounded real numbers with a link function (most commonly, the logit function), where an ordinary linear regression can be used to model the logit transformed data. During Bayesian estimation of the posterior distribution of the model parameters, the likelihood function with the μ-φ parameterization is

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} \quad (1)$$

and

$$\mu = \text{logit}^{-1}(X\beta). \quad (2)$$

Γ is the gamma function, μ is the inverse logit transformed model prediction, y is the observed data bounded between zero and one, and φ is the precision parameter. Note that model predictions are mapped back to the bounded space with the inverse logit function. Our accuracy data contains values equal to one. Therefore, the one-inflated beta distribution is needed, which produces a mixture density (Ospina and Ferrari, 2012). The likelihood function using the one-inflated beta distribution incorporates a new parameter, α, such that

$$f(y;\alpha,\mu,\phi) = \begin{cases} (1-\alpha)f(y;\mu,\phi) & (0 < y < 1), \\ \alpha & (y = 1). \end{cases} \quad (3)$$

To construct beta regression models with Bayesian analysis with the one-inflated beta distribution for the likelihood function, we defined a custom response distribution with the brms package in R.[3] Weakly informative Gaussian priors [$\beta \sim N(0,5^2)$] were used as the priors for the regression coefficients. The half Cauchy distribution was used for φ {$\varphi \sim \text{Cauchy}[0,5^2)$}, and the beta distribution was used for α [$\alpha \sim \text{Beta}(0.5,8)$]. Note that model coefficients do not need to be bounded in any way as model output is transformed with the inverse logit function into the bounded space.

Bayes factors (BFs) were used for model comparison (Dienes, 2016; Liu *et al.*, 2022; Stone, 2013). There is controversy regarding using BF to substitute for null hypothesis testing (Gelman *et al.*, 2013). However, BF is used here to compare which model is more likely given the evidence (i.e., the data) rather than the likelihood of the observed effect being due to chance, as is the case in null hypothesis testing (Morey *et al.*, 2016; Wagenmakers *et al.*, 2016). Other popular methods, such as the Bayes leave-one-out (LOO) analysis, show limitations when the ground truth is consistent with the null hypothesis. Gronau and Wagenmakers (2019) demonstrate that when the number of observations consistent with the simpler model (i.e., H_0) grows larger, LOO's support for it reaches an upper bound, and this bound can sometimes be very modest. It was also demonstrated that depending on the prior distribution, as more H_0 consistent data is added, LOO's support for H_0 can decrease. Therefore, to avoid potential bias toward the more complex model, we use BFs for model comparison.

If BF_0 (the BF indicates evidence for H_0 over H_1) is between zero and 1/10, the data strongly support H_1 over H_0. Conversely, if BF_0 is larger than ten, there is strong evidence for the null hypothesis (Jeffreys, 1961; Biel and Friedrich, 2018; Dienes, 2014; Harms and Lakens, 2018; Lakens *et al.*, 2020; Schönbrodt and Wagenmakers, 2018; Lee and Wagenmakers, 2014).

For each speech rate condition, a full model was constructed with the main effects of syllable structure (onset vs coda for the slow rate and coda vs NN-resyllabified coda for the normal rate) and group. The null model was constructed with group as the only main effect. We also tested whether the effect of syllable structure differed between item groups by including an interaction term.

G. Duration analysis of NN-resyllabified and canonical onset consonants

Although resyllabified sequences may have become similar to their onset counterparts in terms of spectral pattern, there

is evidence that resyllabifed codas retain their underlying coda status through duration (Gao and Xu, 2010; Lehiste, 1960). Specifically, the durations of the resyllabified consonants are shorter compared to those of the canonical onsets. To test whether duration differs between the two, the same acoustic intervals from Sec. II F were used. Bayesian analysis with *linear* regression was used to determine if duration of the acoustic interval was affected by syllable affiliation (i.e., genuine onset vs NN-resyllabified coda). Duration was used as the dependent variable and item group and syllable affiliation were used as the predictor. The likelihood function used the normal Gaussian distribution. For the regression coefficient priors, we used weakly informative Gaussian prior [$\beta \sim N(0,5^2)$], and for the sigma prior, we used the half Cauchy distribution {$\sigma \sim \text{Cauchy}[0,5^2)$}.

III. RESULTS

A. Resyllabification classifiers

Figure 8 shows the model performance of the word sequence classifiers. Since we trained a model for each speaker separately, the result in Fig. 8 was calculated by summing over each speaker's confusion matrix. As shown, the classifiers achieved near ceiling accuracy on the test split for the slow speaking rate, indicating that the models could distinguish the word sequences very well.

Figure 9 shows the model performance on the normal speaking rate by summing over the results from all of the speakers. Table IV lists the accuracy rate for the onset, coda, and all of the sequences. As can be observed, most of the onset sequences were classified correctly. Thus, the classifiers trained on the slow speaking rate data also did well on the onset conditions spoken at a faster rate, such as "Lee steal" or "Lee stale." In the coda condition, the classifiers misclassified a large portion of the sequences as their onset

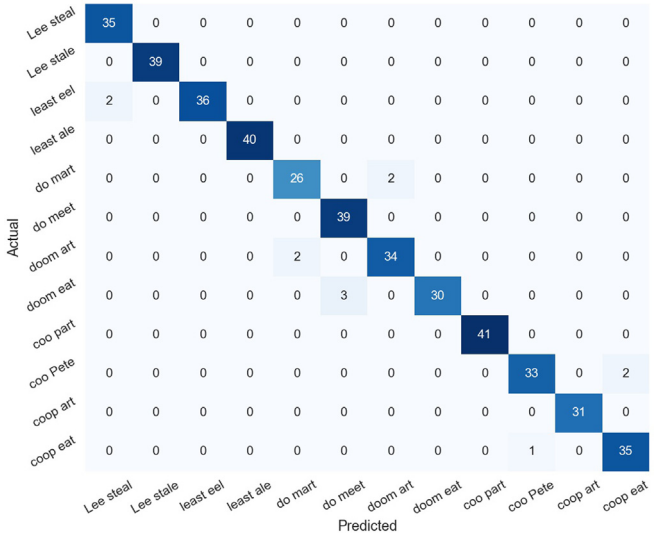

FIG. 8. (Color online) A confusion matrix of model performance on the testing split of the slow speech rate. This is an element wise summation of all of the speakers' confusion matrices. The colour intensity of tiles reflects numeric value.

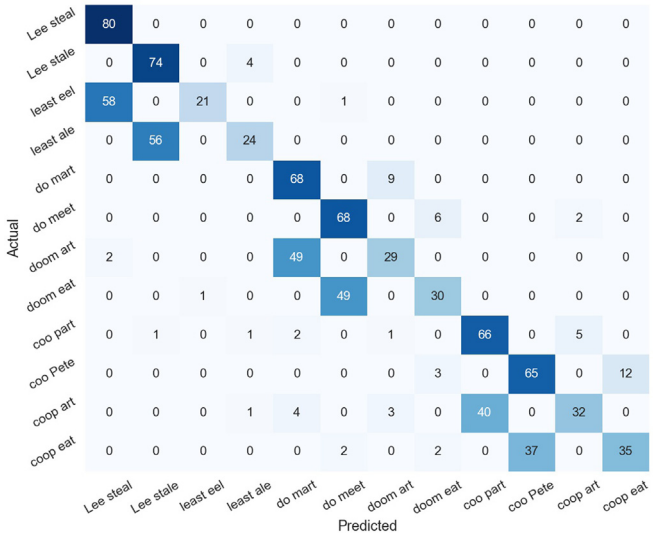

FIG. 9. (Color online) A confusion matrix of model performance on the normal speech rate. This is an element wise summation of all of the speakers' confusion matrices. The colour intensity of tiles reflects numeric value.

counterpart, such as classifying "least eel" as "Lee steal." These misclassified sequences, presumably due to resyllabification, are examined in detail later.

B. DTW analysis

Figure 10 shows a bar graph of the cosine distance between the NN-resyllabified tokens and te slow tokens. The NN-resyllabified sequences were only compared to slow sequences in the same group. Figure 10 shows that when minimising the effect of speech tempo, NN-resyllabified words, such as "least eel," is more similar to its canonical onset counterpart "Lee steal" than to its non-resyllabified version. In other words, when comparing the NN-resyllabified condition with the slow onset condition, the cosine distance is smaller than when comparing with the slow true coda condition.

The result from the DTW analysis can be reflected by the spectrograms in Fig. 11. "Doom art" in the middle of Fig. 11 was classified as "do mart" by the neural network in Sec. II, therefore, we treated it as a resyllabified token. The NN-resyllabified "doom art" appears to be more similar to the canonical onset version "do mart" in the top panel of Fig. 11. The bottom panel of Fig. 11 shows "doom art" spoken in the slow condition, likely with a glottal stop at the beginning of the second syllable "art."

Figure 12 shows the DTW cosine distance between correctly classified normal rate coda tokens and the slow tokens. The opposite trend from Fig. 10 can be observed: the non NN-resyllabified sequences are more similar to their

TABLE IV. Accuracy summary for the normal speech rate tokens.

Coda	0.36
Onset	0.90
Overall	0.63

https://doi.org/10.1121/10.0017117

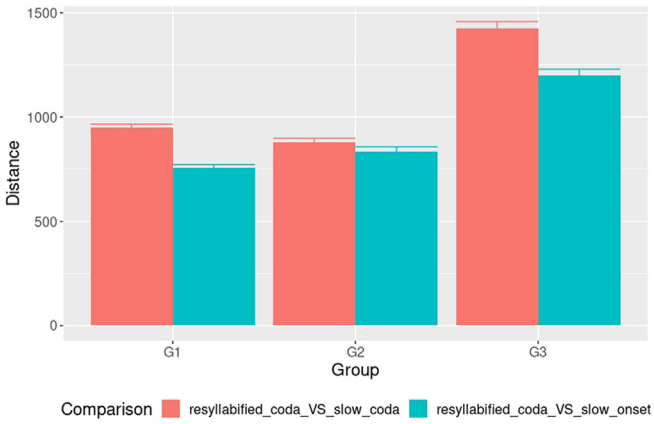

FIG. 10. (Color online) The DTW cosine distance between resyllabifed normal rate sequences and slow sequences. The error bars represent 95% of the confidence interval. G1, "least eel," "least ale," "Lee stale," "Lee steel"; G2, "doom art," "doom eat," "do mart," "do meet"; G3, "coop art," "coop eat," "coo part," "coo Pete."

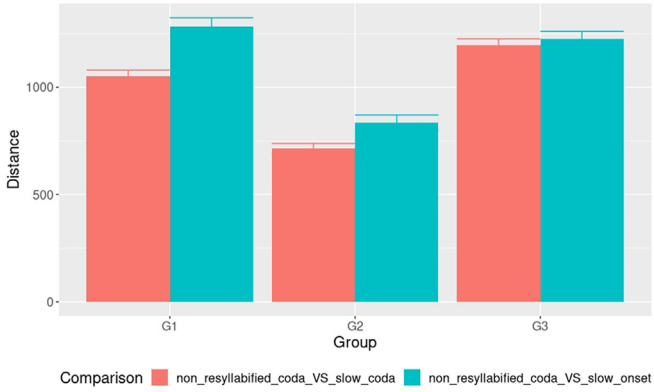

FIG. 12. (Color online) The DTW cosine distance between non NN-resyllabifed normal rate sequences and slow sequences. The error bars represent 95% of the confidence interval. G1, "least eel," "least ale," "Lee stale," "Lee steel"; G2, "doom art," "doom eat," "do mart," "do meet"; G3, "coop art," "coop eat," "coo part," "coo Pete."

canonical coda form in the slow rate condition, which supports the prediction that correctly classified coda tokens likely have not been resyllabified, unlike their misclassified counterparts.

C. Intervocalic consonant alignment analysis

1. Results for slow speech rate

With the consonant intervals described in Sec. II E, we trained 80 neural networks for each vowel minimal pair in Table I and obtained an accuracy distribution from the test set. Figure 13 shows the accuracy rate from the slow speech rate condition. As Fig. 13 depicts, for /s/ frication in the intervocalic cluster (i.e., G1), the vowel classification accuracy is around chance, indicating that little to no vowel information was picked up by the binary classifier in the frication of /s/ for the onset (e.g., "Lee stale") and coda conditions (e.g., "least ale"). For G2, the intervocalic /m/ contains more detectable vowel information as the onset of the second syllable and less so when it is the coda of the first syllable. Similar trends can be observed for G3, although with overall higher accuracy, the binary classifier performs better when /p/ is the onset of the second syllable.

To test the hypothesis via model comparison, we use the BF, which can offer support for a model based on the observed data (Dienes, 2014; Harms and Lakens, 2018). The posterior distributions of the model parameters are not very

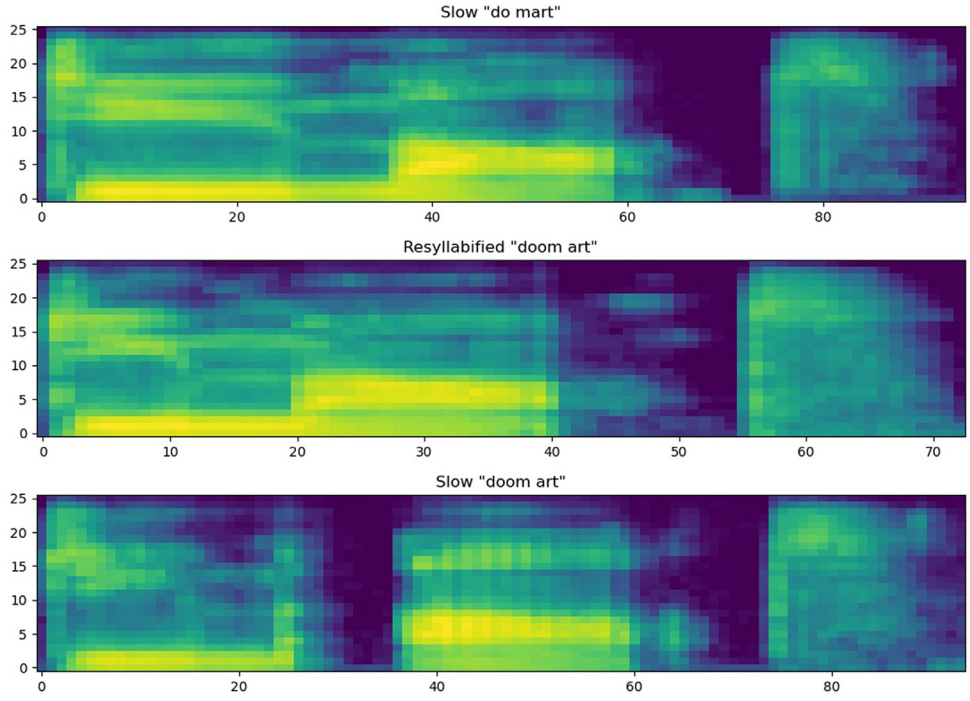

FIG. 11. (Color online) Mel-spectrograms of three word sequences from one speaker.

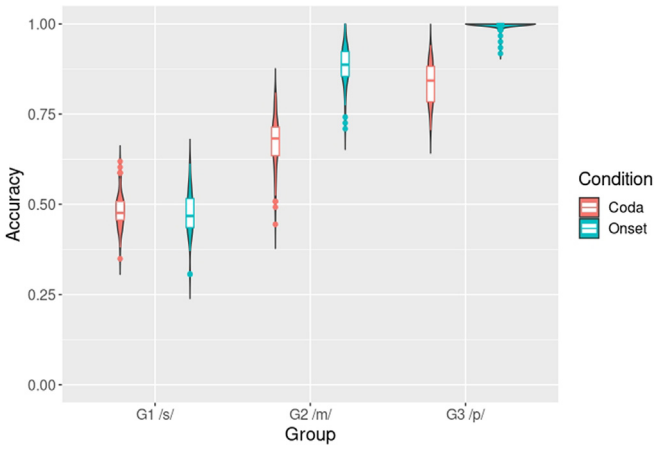

FIG. 13. (Color online) Vowel classification accuracy by group from the slow speech rate condition. G1, "least eel," "least ale," "Lee stale," "Lee steel"; G2, "doom art," "doom eat," "do mart," "do meet"; G3, "coop art," "coop eat," "coo part," "coo Pete."

informative as predictions need to be transformed with the inverse logit function, and their details are included as supplementary material.[4] Therefore, the predicted distribution from 100 random samples is displayed in Fig. 14, which shows that the model with an interaction term exhibits the best predicative power. BF_0 was very close to zero (i.e., BF_1 is larger than ten). Hence, the data indicate that the alternative model, i.e., onset and coda conditions are different, is highly more likely because model accuracy differs greatly. We also constructed a model with an interaction effect between item group and syllable structure. $BF_{interaction}$ (the BF indicating support for the interaction model over the full model) is larger than ten, which provides strong support for the interaction model. To conclude, the data show strong evidence for the effect of syllable structure, which differs greatly between groups. In other words, there is robust effect of syllable structure for G2 and G3 but likely not for G1.

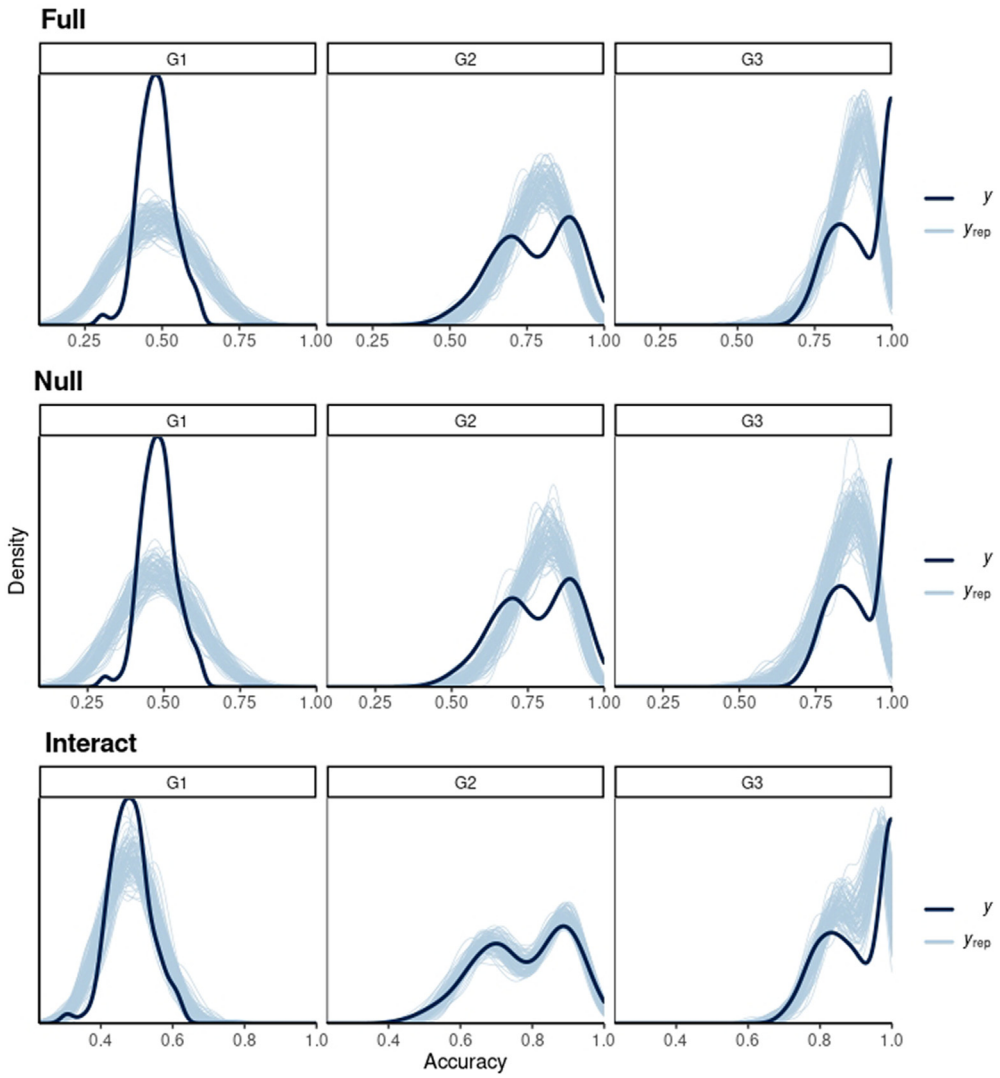

FIG. 14. (Color online) Model predictions against 100 random samples for the slow rate, where y refers to the observed data and y_{rep} refers to predictions. The columns correspond to item groups and the rows correspond to model type.

https://doi.org/10.1121/10.0017117

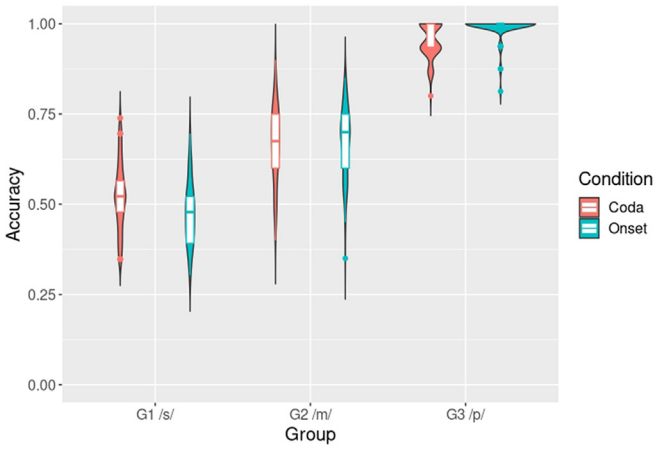

FIG. 15. (Color online) Vowel classification accuracy by group from the normal speech rate condition. The coda condition here refers to the NN-resyllabified coda sequences in the normal speech rate condition. G1, "least eel," "least ale," "Lee stale," "Lee steel"; G2, "doom art," "doom eat," "do mart," "do meet"; G3, "coop art," "coop eat," "coo part," "coo Pete."

2. Results for normal speech rate

The accuracy distributions from the normal speech rate condition are observed in Fig. 15. Note that the coda condition only contained NN-resyllabified sequences. Figure 15 shows that the amount of vowel information detected during the acoustic consonantal intervals (e.g., /s/ frication in "Lee stale") was very similar between the NN-resyllabified coda and onset sequences. The item group wise trends are similar to the slow rate condition in Fig. 13. The aspiration from the plosive onset /p/ contains the most vowel related energy, and the nasal murmur from /m/ contained enough vowel information for the classifier to perform above chance. For /s/ in G1, the accuracy distributions are centered at chance level (i.e., 50%), indicating that little to no vowel information was detected by the binary classifiers during the frication intervals.

The predicted distributions from the Bayesian analysis results are shown in Fig. 16. The posterior distributions of model parameters can be found in the supplementary

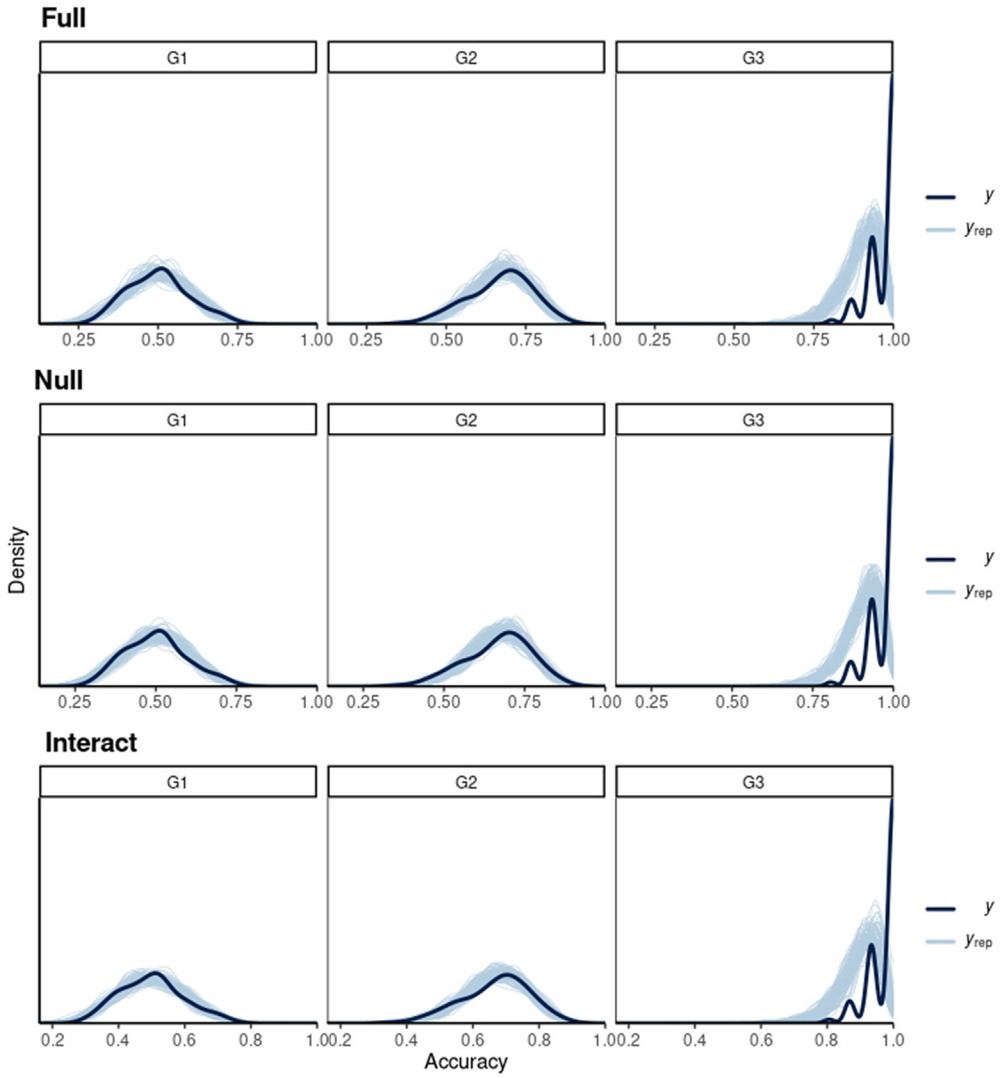

FIG. 16. (Color online) Model predictions against 100 random samples for the normal rate, where y refers to the observed data and y_{rep} refers to predictions. The columns correspond to item groups and the rows correspond to model type.

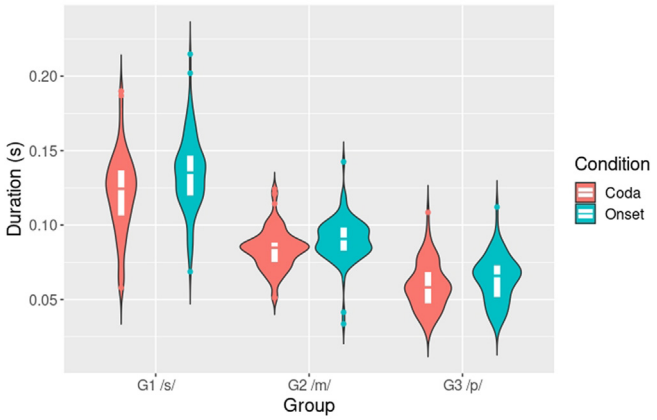

FIG. 17. (Color online) The duration of onset and NN-resyllabified consonants from the normal speaking rate condition.

material.[4] Visually, the predicted distributions do not differ too much from one another. BF_0 was larger than ten, signifying that the data provide more support for the null model. Figure 15 indicates that model accuracy might differ slightly between the NN-resyllabified coda and onset sequences for G1. In other words, there might be an interaction between the effect of syllable structure and group. $BF_{interaction}$ (the BF indicating support for the interaction model over the null model) is smaller than 1/10, therefore, there is little to no evidence suggesting that accuracy differs between onset and NN-resyllabified coda tokens for G1.

D. Duration of intervocalic consonants

The duration of the acoustic intervals for the canonical and NN-resyllabified onsets are shown in Fig. 17. Congruent with previous findings (Gao and Xu, 2010; Lehiste, 1960), NN-resyllabified codas are shorter than the canonical onsets. Predictions of the Bayesian analysis are displayed in Fig. 18, and the parameter posterior distributions are included as supplementary material.[4] The effect of syllable structure was estimated to be around 0.01 ($\mu = 0.008$ [0.005,0.012]). BF_0 is smaller than 1/10, which indicates that duration differs between syllable structures.

IV. DISCUSSION

Previous debates on the phenomenon of resyllabification have mainly relied on phonotactic analysis, listener judgment, or phonetic properties such as voicing and aspiration. In this study, we tested an alternative approach that examines articulatory coordination and coarticulation, as reflected in the spectral patterns, using machine learning models with acoustic data. The findings have offered a new perspective on the nature of resyllabification.

A. Overall findings

The results of computational analysis have largely confirmed the two predictions laid out in the Introduction. The deep learning models trained on slow speech rate data misidentified coda sequences by classifying them as their onset counterparts, and DTW analysis showed that for all three consonants (i.e., /st/, /p/, and /m/), the sequences identified as resyllabified were more similar to their onset versions than the original coda versions. Moreover, the correctly classified sequences are more similar to their canonical coda version, which indicates that they likely have not undergone resyllabification. Therefore, the first prediction—codas in

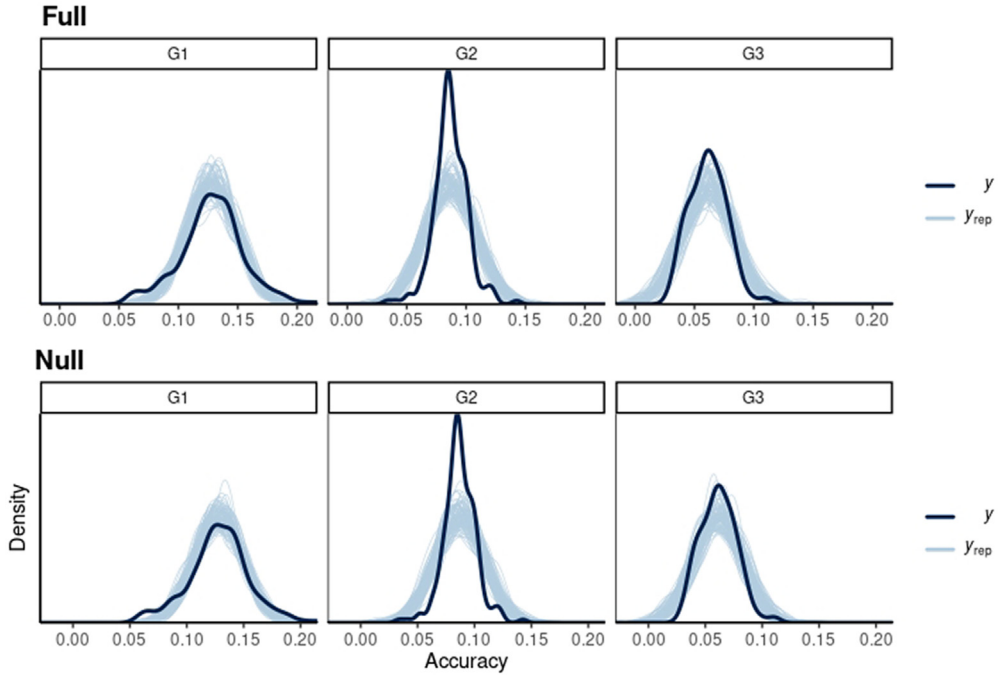

FIG. 18. (Color online) Model predictions against 100 random samples for the duration results, where y refers to the observed data and y_{rep} refers to predictions. The columns correspond to item groups and the rows correspond to model type.

the NN-resyllabified sequences spectrally resemble canonical onsets more than their canonical coda version—was supported. The results from the binary classifiers confirm the second prediction by showing that there was a similar amount of vowel information detected in the NN-resyllabified onsets and canonical onsets but not between the true codas and onsets from the slow condition. This suggests that the underlying articulation was alike between the NN-resyllabified and canonical onsets. Therefore, the results confirm previous findings of resyllabification in English (de Jong, 2001; Gao and Xu, 2010; Stetson, 1951). In connected speech, resyllabification can happen when a coda consonant is followed by a vowel-initial syllable, and it applies to singleton consonants and consonant clusters. The coda status of the NN-resyllabified consonants, however, seems to be partially retained through duration: Resyllabified codas are shorter compared to canonical onsets. This is consistent with the findings of Lehiste (1960) and, more recently, Gao and Xu (2010). Whether or not listeners can perceive the durational cues, though, needs to be tested in future studies. Furthermore, future studies can investigate the effect of resyllabification and syllable position on consonant duration by examining NN-resyllabified and non NN-resyllabified consonants.

It is also interesting to note the relation between resyllabification and speech rate. When syllable duration is around 350 ms in the current study, the rate of inferred resyllabification already reaches above 50%. At 2.86 syllables per second, this speech rate is rather slow compared to the typical normal articulation rate of 5–7 syllables per second in connected speech (Eriksson, 2012; Tiffany, 1980). Yet, this is consistent with the finding by de Jong (2001) that resyllabification starts to take place as speech rate increases to around 350 ms per syllable, and resyllabification rate approaches 100% at 150 ms per syllable. The implication is that the tendency for resyllabification must be very strong so that it would be difficult to avoid at normal speech rate.

The finding of resyllabification aligns with the syllable model shown in Fig. 1, which is based on how the predictions illustrated in Fig. 2 were derived. That is, once a coda consonant is resyllabified as the onset of the next syllable, as determined by the deep learning model and DTW analysis, its articulation is overlapped with the vowel of the next syllable, as determined by the binary classifiers. This is consistent with the recent finding that the movements toward the vowel and onset C are synchronised at syllable onset (Liu *et al.*, 2022; Liu and Xu, 2021; Xu *et al.*, 2019), which is denoted by the rime and onset tiers in Fig. 1.

B. Coarticulation resistance and dimension-specific sequential target approximation (DSSTA)

CV synchronisation does not mean that vowel information is always detectable from the syllable onset or at the same time point, however, which is partly due to *coarticulation resistance*, i.e., the ability of a segment to restrain coarticulatory effects from adjacent segments (Bladon and Al-Bamerni, 1976; Recasens, 1984). Recasens (1984) proposes that the degree of coarticulation resistance is dependent on the amount of constraint that a consonant or vowel places on the tongue body. Xu (2020) further proposes that the phenomenon is a mechanism that resolves the articulatory conflicts between consonants and vowels when they both involve the same articulator while being coproducedd to achieve C-V co-onset (Fig. 1). According to this mechanism, namely, the *dimension-specific sequential target approximation (DSSTA) mechanism*, different (e.g., vertical or horizontal) dimensions of an articulator can be engaged in executing only a single target, which is either consonantal or vocalic, during C-V coproduction. This mechanism maximises the degree of C-V synchronisation while allowing individual articulator dimensions to be engaged in only sequential target approximation movements, i.e., without gestural blending (Saltzman and Munhall, 1989) given its computational difficulty (Tilsen, 2019). The following discussion will offer an account of the differences in the detected vowel information in the present results that include DSSTA as a critical mechanism.

The amount of detectable vowel information in the consonant interval follows the order of group 1 (/s/) < group 2 (/m/) < group 3 (/p/). This order may result from two different sources. The first source, which is more obvious, is the differences in their relative timing. The frication in group 1 and nasal murmur in group 2 correspond to the articulatory closure of the consonants, whereas the aspiration in group 3 corresponds to the articulatory release, which occurs after the closure. This could partially explain why more vowel information was detected in group 3 than in the other two groups. The second source is coarticulation resistance due to DSSTA. The consonant /s/ in group 1 involves the tongue body to form a groove needed to direct the airflow toward the front teeth (Borden *et al.*, 2003). The involvement of the tongue body would generate serious coarticulation resistance in /s/ in group 1 because the horizontal and vertical dimensions of the tongue body are likely involved in approaching the target of the sibilant (Recasens *et al.*, 1997). In contrast, the articulation of /m/ in group 2 requires only lip closure without constraints on the tongue. This would account for the greater amount of detectable vowel information in group 2 than in group 1. The lack of tongue involvement in labial consonants is true of /p/ in group 3 as well. Yet, there, it is added on top of the fact that aspiration, where the binary classification was performed, occurs after the stop closure, thus, giving rise to the maximal vowel information detected by the classifier. Note that had one of the syllables in group 1 contained a rounded vowel, such as /u/, DSSTA would predict that vowel information would be better detected because lip movements are not in direct conflict with the articulation of /s/. This possibility can be tested in future research.

C. Chance level performance of the binary classifier for G1 sequences

The lack of detectable vowel information in /st/ even in normal speech rate may seem to contradict the recent

finding that vowel articulation could be detected at the same time as the onset of a consonant cluster (Liu and Xu, 2021). That study found that for a minimal triplet, such as "slit" vs "slot" vs "flot," the difference between "slit" and "slot" could be detected around the same time as "slot" and "flot," which is *before* the frication onset. However, we have noted three major differences between Liu and Xu (2021) and the current study. First, Liu and Xu (2021) only looked at clusters such as /sp/ and /sl/, but did not consider /st/ as in the current study. /p/ does not require any tongue movement, thus, is less coarticulation resistant than /l/ and /t/. In terms of /l/ and /t/, both are alveolars, and Iskarous *et al.* (2013) found that /t/ is more coarticulation resistant than /l/ in the vertical dimension for the jaw and tongue blade. This could be a result of the requirement of a full closer for /t/ as a plosive but not for the approximant /l/. /t/ being more coarticulation resistant means that it may have delayed much of the vowel movements. Second, much larger vowel contrasts were involved in Liu and Xu (2021)—/slit/ vs /slot/—than those in the present study—/steal/ vs /stale/. The greater the vowel contrast, the greater the magnitude of tongue movement in the articulatory dimensions, which are not essential for the consonant articulation, and the more detectable the vowel information is during the frication interval. Third, the target words were produced with a carrier in Liu and Xu (2021), which made the speech more fluent than the isolated word sequences spoken in the present study. The average speech rate in Liu and Xu (2021) was about 140 ms per syllable compared to 350 ms per syllable in this study. It is hard to tell, however, if any of these factors are decisive or if all of them jointly contribute to blocking the vowel information from being present in the /s/ frication.

D. Above chance performance of the binary classifier for the slow coda sequence in G2

One of the most surprising results of this study, as shown in Fig. 13, is the finding that for the slow speaking rate, there is information of the upcoming vowel in the intervocalic consonants when they are in the coda position of the first syllable (e.g., "doom art"; "coop art"), albeit less than when they are in the onset position. The detection of vowel information in a non resyllabified coda may seem particularly striking given the clear temporal gap or glottalisation between the two syllables, as can be seen in Figs. 19 and 20. However, the glottal component, as can be judged auditorily and spectrally, corresponds to a glottal stop or glotallisation (which is also a form of glottal stop, Redi and Shattuck-Hufnagel, 2001; Garellek, 2013), which serves as the onset of the syllable /art/. A glottal stop, just like that for other stops, such as /b, d, g/, would be fully coarticulated with the following vowel (Xu, 2020), as illustrated in Fig. 2. This means that the target approximation of /a/ must have started some time well before the glottal closure (Liu *et al.*, 2022; Xu and Liu, 2007). This can indeed be observed in Fig. 19, i.e., the brief yet clearly visible labial release after the nasal murmur of /m/ and the F2 transition from "doom" to "eat" during and right before the glottalised interval in Fig. 20. The high vowel detection rate of around 80% for /p/ and 65% for /m/ means that the vowel target approximation may have started during (though probably not before) the closure of the coda, but exactly when during the closure, however, awaits future investigations.

E. Broader implications

The finding of a clear tendency toward resyllabification in this study provides further support for the synchronisation model of the syllable (Xu, 2020) beyond recent findings (Liu *et al.*, 2022; Liu and Xu, 2021). According to the model, there is a strong demand for onset consonants to synchronise (i.e., fully overlap) with the vowel, and a high time pressure against the preservation of coda consonants. This is partially consistent with the maximum onset principle (Pulgram, 1970; Selkirk, 1982) but offers specific articulatory details that can be tested in the acoustic signals as performed in the present study. Because the syllable is essential and highly controversial for theoretical models in linguistics as well as psycholinguistics, the current results may have implications for many broader issues about speech production, but here we focus only on two major issues. The first is about the influential psycholinguistic model of speech

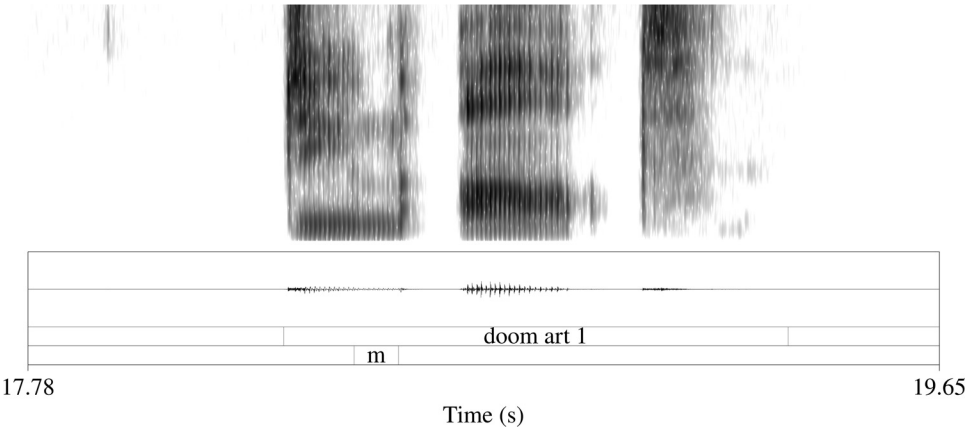

FIG. 19. The spectrogram of "doom art" from a male speaker.

https://doi.org/10.1121/10.0017117

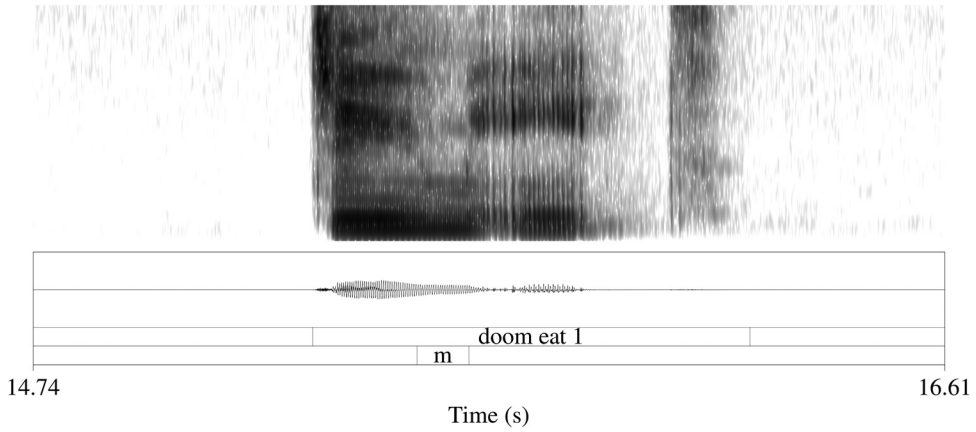

FIG. 20. The spectrogram of "doom eat" from a female speaker.

production (Levelt *et al.*, 1999), which proposes a step-by-step model of how speech production proceeds from lexical selection to articulation. The results of the present study are relevant for the phonological encoding to articulation stages in the model. The most relevant result is probably the corroboration of previous findings that resyllabification is contingent on local articulation rate: highly likely at normal rate but optional at slow rate (de Jong, 2001; Stetson, 1951). This means that until local speech rate is known, the articulatory affiliation of coda consonant is undetermined, which would suggest that either syllables retrieved from memory (during phonological encoding) are incomplete in terms of segment affiliation or the retrieved syllables are reorganised by resyllabification, and this reorganisation would occur *after* the phonetic encoding stage, just before articulation.

The finding of rate dependency of resyllabification is further relevant to any psycholinguistic model of speech production given the known extensive use of speech timing by linguistic functions. Specifically, local articulation rate, which is jointly determined by syllable duration and pause duration, is used to encode multiple levels of boundary strength (Lehiste, 1972; Klatt, 1976; Nakatani *et al.*, 1981; Wagner, 2005; Wang *et al.*, 2017). Thus, resyllabification is likely a regular variable of connected speech beyond word-level phonetics. In fact, it is likely part of the process of producing connected speech that involves many other phonetic reorganisations, including deletion of intervocalic coda [as opposed to resyllabification in some languages; e.g., tone sandhi (Chen, 2000), intrusive /r/ (Gick, 1999), and vowel hiatus breakers (Mudzingwa, 2013), etc]. There is already evidence that some of these reorganisations may be cognitively real, at least in the case of tone sandhi (Zhang *et al.*, 2015). These phonetic reorganisation tactics could, therefore, be included in an enhanced psycholinguistic model of speech production, and their cognitive reality could be experimentally investigated.

The second broad issue is whether the present results can be interpreted in terms of ambisyllabicity. The original proposal of ambisyllabicity was motivated by the lack of phonetic means to clearly determine syllable boundaries, hence, the affiliation of intervocalic segments had to rely on phonotactic well-formedness, and for cases where *ill-formed* syllables would occur if an intervocalic consonant can only have a single affiliation, e.g., *happy*, *attic*, *hobby*, the solution is ambisyllabicity, i.e., simultaneous affiliation to both adjacent syllables (Kahn, 1976). Exactly how such double association is realised phonetically, however, has remained unclear. Gick (2003) has proposed that some intervocalic segments, e.g., /l/ and /w/, actually consist of a C-gesture and a V-gesture, which are simultaneously phased to the surrounding syllables and, therefore, ambisyllabified. The phonetic evidence is in terms of different time delays in the *achievement of the respective C* and *V gestural goals, which differs from the onset alignment that the current study has examined*. Although this study is not designed for examining ambisyllabicity, at least one phonetic cue is shown to have the potential to indicate the original coda status of a consonant, namely, the shorter duration of NN-resyllabified coda than the original onset consonant (also cf. Lehiste, 1960). However, if CV onset coarticulation is considered as the sole indicator, the NN-resyllabified codas are unambiguously overlapped with the following vowel according to the present data.

F. Caveats

Two of the resyllabification classifiers satisfied the early stopping criteria, which meant that their training epochs were determined with the test split rather than the pilot data. This could have slightly inflated the overall accuracy reported for the slow condition in Sec. III A. However, the use of the classifier is to classify normal rate sequences, which is the focus of the study, and their accuracy has not been inflated as the normal rate data were not used in any way during training.

The possibility of false negatives cannot be completely ruled out regarding the chance level performance of the binary classifier for G1. Providing that upcoming vowel related acoustic information exists during frication, two scenarios could result in false negative detections:

(1) Chance performance due to chance, and

(2) the neural networks are not powerful enough to detect the subtle difference.

The first scenario refers to the opposite of what is described in Combrisson and Jerbi (2015), namely, the model achieved chance performance by chance. This could be due to the randomised nature of the data split and/or model parameter initialisation (not hyperparameters). However, this possibility is accounted for in the current study by repeatedly training 80 classifiers on randomised train and test data and analysing the resultant accuracy distributions. For the second scenario, despite tuning the hyperparameters with data from pilot recordings, the neural network was not tuned for each speaker and consonant type separately. In practice, it is very difficult to construct a perfect network regardless of the type of data in question. Therefore, there is a small possibility that the binary classifier could not detect a difference between groups in G1 due to the lack of robustness. Future studies could incorporate articulatory data as it might provide more detailed information than acoustic data in the current study (Tilsen, 2020).

On the other hand, the possibility of false positives cannot be ruled out either. Providing that the test dataset is large enough, machine learning models cannot always achieve 100% accuracy. The same applies to the word sequence classifiers in this study. This is evident in the results from the slow speech rate in Sec. III A. Although overall accuracy is high, there were still coda sequences classified as their onset counterpart, as well as cases where onset sequences were classified as their coda counterpart. At the slow speech rate (two syllables per second, on average), is it unlikely that resyllabification occurred, hence, these misclassifications are likely genuine incorrect classifications (i.e., not due to syllabification). As for the normal rate results, there should also exist genuine misidentifications like those in the slow rate, which is likely why there are onset sequences classified as their coda counterparts. This means that a small number of the NN-resyllabified sequences might be genuine misidentification as well. However, the normal rate results show that onset sequences reached an accuracy rate of 90% and only 36% was achieved for the coda sequences. Therefore, a large portion of the NN-resyllabified tokens are likely due to syllabification structure and not just simple false positives.

Also, the study did not conduct a parallel analysis of V_2 binary classification for the correctly classified coda tokens. Unlike the DTW analysis, there are too few correctly classified coda sequences in the normal rate for training neural network classifiers, especially for G1 and G2. This issue is exacerbated by the imbalance of speakers in the data, i.e., some speakers had zero or a very small number of correctly classified tokens in certain item groups. Future study can potentially avoid this issue by increasing the number of repetitions in the normal rate condition.

Finally, as noted in Sec. IV B, the lack of detectable vowel information in group 1 might have been avoided had one of the syllables in each pair contained a rounded vowel. This is because despite its involvement of the tongue body, the articulation of /s/ is not in direct conflict with the lip movements of the coproduced vowel. This possibility can be investigated in future research.

V. CONCLUSION

We used deep learning models with acoustic data to investigate the phenomenon of resyllabification. The models trained on slow speech data can be used to infer resyllabified sequences in normal speech rate data. This was verified by DTW analysis, which revealed that compared to slow speech, NN-resyllabified sequences were more similar to the true onset sequences than their original coda productions. The acoustic intervals of intervocalic consonants were examined with bidirectional RNN models. We found that a similar amount of vowel information was detected in the intervocalic consonants between the NN-resyllabified codas and genuine onsets, suggesting that the coarticulation structure of the former resembles that of the latter. For slow speech rate, the results show that the articulatory structures likely differed between the onset and coda sequences. Surprisingly, however, vowel information can still be detected from the closure and release of labial coda consonants, indicating that the articulation of the vowel has started during the acoustic interval of a coda consonant even when it is not resyllabified.

APPENDIX

The hyperparameter details for the multiclass classifier and the binary classifiers are shown in Tables II and III, respectively.

[1]During data splitting, correlated samples resulting from augmentation were not included in the same dataset, e.g., the original "coo part" and its augmented version always ended up in the same split.

[2]The full detail of models and data processing can be found at https://github.com/Clara-liu/deep_speech_resyllabification (Last viewed January 10, 2023).

[3]The details of implementation of custom one-inflated-beta-distribution are available at https://github.com/Clara-liu/deep_speech_resyllabification/blob/main/one_inflated_beta.R (Last viewed January 10, 2023).

[4]See supplementary material at https://www.scitation.org/doi/suppl/10.1121/10.0017117 for details on the posterior distributions for the slow rate condition, the posterior distributions for the normal rate condition, and the posterior distributions for the duration analysis.

Adda-Decker, M., de Mareüil, P. B., Adda, G., and Lamel, L. (**2002**). "Investigating syllabic structure and its variation in speech," in *Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology*, p. 6.

Amodei, D., Anubhai, R., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Chen, J., Chrzanowski, M., Coates, A., Diamos, G., Elsen, E., Engel, J., Fan, L., Fougner, C., Han, T., Hannun, A., Jun, B., LeGresley, P., Lin, L., Narang, S., Ng, A., Ozair, S., Prenger, R., Raiman, J., Satheesh, S., Seetapun, D., Sengupta, S., Wang, Y., Wang, Z., Wang, C., Xiao, B., Yogatama, D., Zhan, J., and Zhu, Z. (**2015**). "Deep Speech 2: End-to-end speech recognition in English and Mandarin," arXiv:1512.02595.

Audacity Team (**2021**). "Audacity," available at https://audacityteam.org/ (Last viewed August 1, 2022).

Balakrishnan, N., and Nevzorov, V. (**2003**). *A Primer on Statistical Distributions* (Wiley, Hoboken, NJ).

Barlow, J. A., and Gierut, J. A. (**1999**). "Optimality theory in phonological acquisition," J. Speech. Lang. Hear. Res. **42**(6), 1482–1498.
Bartelds, M., Richter, C., Liberman, M., and Wieling, M. (**2020**). "A new acoustic-based pronunciation distance measure," Front. Artif. Intell. **3**, 39.
Bermúdez-Otero, R. (**2011**). "Cyclicity," in *Blackwell Companion to Phonology*, edited by M. van Oostendorp, C J. Ewen, E. Hume, and K. Rice (Wiley-Blackwell, Chichester), Vol. 4, pp. 2019–2048.
Biel, A. L., and Friedrich, E. V. C. (**2018**). "Why you should report Bayes factors in your transcranial brain stimulation studies," Front. Psychol. **9**, 1125.
Birgit, A. (**2001**). "Regional variation and edges: Glottal stop epenthesis and dissimilation in standard and southern varieties of German," Z. Sprachwiss. **20**(1), 3–41.
Bladon, R. A. W., and Al-Bamerni, A. (**1976**). "Coarticulation resistance in English /l/," J. Phonetics **4**(2), 137–150.
Blevins, J. (**2003**). *Evolutionary Phonology: The Emergence of Sound Patterns* (Cambridge University Press, Cambridge, UK).
Boersma, P., and Weenink, D. (**2022**). "Praat: Doing phonetics by computer (version 6.2.14) [computer program]," available at http://www.praat.org/ (Last viewed November 9, 2022).
Borden, G. J., Harris, K. S., and Raphael, L. J. (**2003**). *Speech Science Primer: Physiology, Acoustics, and Perception of Speech*, 4th ed. (Williams and Wilkins, Baltimore).
Browman, C. P., and Goldstein, L. (**1992**). "Articulatory phonology: An overview," Phonetica **49**(3–4), 155–180.
Chen, M. Y. (**2000**). *Tone Sandhi: Patterns across Chinese Dialects* (Cambridge University Press, Cambridge, UK).
Clements, G. N., and Keyser, S. J. (**1983**). *CV Phonology. A Generative Theory of the Syllable (Linguistic Inquiry Monographs Cambridge)* (Cambridge, MA), Vol. 9, pp. 1–191.
Combrisson, E., and Jerbi, K. (**2015**). "Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy," J. Neurosci. Methods **250**, 126–136.
Content, A., Kearns, R. K., and Frauenfelder, U. H. (**2001**). "Boundaries versus onsets in syllabic segmentation," J. Mem. Lang. **45**, 177–199.
de Jong, K. J. (**2001**). "Rate-induced resyllabification revisited," Lang. Speech **44**(2), 197–216.
de Jong, K. J., Lim, B., and Nagao, K. (**2004**). "The perception of syllable affiliation of singleton stops in repetitive speech," Lang. Speech **47**(3), 241–266.
Dienes, Z. (**2014**). "Using Bayes to get the most out of non-significant results," Front. Psychol. **5**, 781.
Dienes, Z. (**2016**). "How Bayes factors change scientific practice," J. Math. Psychol. **72**, 78–89.
Douma, J. C., and Weedon, J. T. (**2019**). "Analysing continuous proportions in ecology and evolution: A practical introduction to beta and Dirichlet regression," Methods Ecol. Evol. **10**, 1412–1430.
Eriksson, A. (**2012**). "Aural/acoustic vs automatic methods in forensic phonetic case work," in *Forensic Speaker Recognition* (Springer, New York), pp. 41–69.
Gao, H., and Xu, Y. (**2010**). "Ambisyllabicity in English: How real is it?," in *Proceeding of the 9th Phonetic Conference of China*, Tianjin.
Garellek, M. (**2012**). "Glottal stops before word-initial vowels in American English: Distribution and acoustic characteristics," UCLA Work. Pap. Phonetics **110**, 1–23.
Garellek, M. (**2013**). "Production and perception of glottal stops," Doctoral dissertation, UCLA, available at https://escholarship.org/uc/item/7zk830cm (Last viewed January 5, 2023).
Gaskell, M. G., Spinelli, E., and Meunier, F. (**2002**). "Perception of resyllabification in French," Mem. Cognit. **30**(5), 798–810.
Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (**2013**). *Bayesian Data Analysis*, 3rd ed. (Chapman and Hall/CRC, New York).
Gick, B. (**1999**). "A gesture-based account of intrusive consonants in English," Phonology **16**(1), 29–54.
Gick, B. (**2003**). "Articulatory correlates of ambisyllabicity in English glides and liquids. Phonetic interpretation," Papers Lab. Phonol. **6**, 222–236.
Goldstein, L., Byrd, D., and Saltzman, E. (**2006**). "The role of vocal tract gestural action units in understanding the evolution of phonology," in *Action to Language via the Mirror Neuron System* (Cambridge University Press, Cambridge, UK), pp. 215–249.
Goslin, J., and Frauenfelder, U. H. (**2001**). "A comparison of theoretical and human syllabification," Lang. Speech **44**(4), 409–436.
Gronau, Q. F., and Wagenmakers, E. J. (**2019**). "Limitations of Bayesian leave-one-out cross-validation for model selection," Comput. Brain Behav. **2**, 1–11.
Harms, C., and Lakens, D. (**2018**). "Making 'null effects' informative: Statistical techniques and inferential frameworks," J. Clin. Transl. Res. **24**, 382–393.
He, K., Zhang, X., Ren, S., and Sun, J. (**2015**). "Deep residual learning for image recognition," arXiv:1512.03385.
Ioffe, S., and Szegedy, C. (**2015**). "Batch normalization: Accelerating deep network training by reducing internal covariate shift," arXiv:1502.03167.
Iskarous, K., Mooshammer, C., Hoole, P., Recasens, D., Shadle, C. H., Saltzman, E., and Whalen, D. H. (**2013**). "The coarticulation/invariance scale: Mutual information as a measure of coarticulation resistance, motor synergy, and articulatory invariance," J. Acoust. Soc. Am. **134**(2), 1271–1282.
Jacewicz, E., Fox, R. A., O'Neill, C., and Salmons, J. (**2009**). "Articulation rate across dialect, age, and gender," Lang. Var. Change **21**(2), 233–256.
Jeffreys, H. (**1961**). *The Theory of Probability*, 3rd ed. (Oxford University Press, Oxford, UK).
Kahn, D. (**1976**). "Syllable-based generalizations in English phonology," Doctoral dissertation, MIT, Cambridge, MA.
Klatt, D. H. (**1976**). "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," J. Acoust. Soc. Am. **59**, 1208–1221.
Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (**2015**). "Audio augmentation for speech recognition," in *Interspeech 2015*, pp. 3586–3589.
Kogan, J. A., and Margoliash, D. (**1998**). "Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study," J. Acoust. Soc. Am. **103**(4), 2185–2196.
Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., and Dienes, Z. (**2020**). "Improving inferences about null effects with Bayes factors and equivalence tests," J. Gerontol., Ser. B **75**(1), 45–57.
Lee, M., and Wagenmakers, E. (**2014**). *Bayesian Cognitive Modeling: A Practical Course* (Cambridge University Press, Cambridge, UK).
Lehiste, I. (**1960**). "An acoustic-phonetic study of internal open juncture," Phonetica Suppl. **5**, 5–54.
Lehiste, I. (**1972**). "The timing of utterances and linguistic boundaries," J. Acoust. Soc. Am. **51**, 2018–2024.
Lerato, L., and Niesler, T. (**2019**). "Feature trajectory dynamic time warping for clustering of speech segments," EURASIP J. Audio, Speech, Music Proc. **2019**(1), 6.
Levelt, W. J. M., Roelofs, A., and Meyer, A. S. (**1999**). "A theory of lexical access in speech production," Behav. Brain Sci. **22**(1), 1–38.
Liu, Z., and Xu, Y. (**2021**). "Segmental alignment of English syllables with singleton and cluster onsets," in *Interspeech 2021*, pp. 3969–3973.
Liu, Z., Xu, Y., and Hsieh, F. (**2022**). "Coarticulation as synchronised CV co-onset—Parallel evidence from articulation and acoustics," J. Phonetics **90**, 101116.
Luo, D., Zou, Y., and Huang, D. (**2018**). "Investigation on joint representation learning for robust feature extraction in speech emotion recognition," in *Interspeech 2018*, pp. 152–156.
MacNeilage, P. F. (**1998**). "The frame/content theory of evolution of speech production," Behav. Brain Sci. **21**, 499–546.
Mirzaei, M. S., Meshgi, K., and Kawahara, T. (**2018**). "Exploiting automatic speech recognition errors to enhance partial and synchronized caption for facilitating second language listening," Comput. Speech Lang. **49**, 17–36.
Morey, R. D., Romeijn, J. W., and Rouder, J. N. (**2016**). "The philosophy of Bayes factors and the quantification of statistical evidence," J. Math. Psychol. **72**, 6–18.
Mudzingwa, C. (**2013**). "Hiatus resolution strategies in Karanga (Shona)," Southern Afr. Linguist. Appl. Lang. Stud. **31**(1), 1–24.
Mullooly, R. (**2003**). "An electromagnetic articulography study of resyllabification of rhotic consonants in English," in *International Conference of Phonetic Sciences*, Barcelona.
Nakatani, L. H., O'Connor, K. D., and Aston, C. H. (**1981**). "Prosodic aspects of American English speech rhythm," Phonetica **38**, 84–106.

Ní Chiosáin, M. N., Welby, P., and Espesser, R. (**2012**). "Is the syllabification of Irish a typological exception? An experimental study," Speech Commun. **54**(1), 68–91.
Ojala, M., and Garriga, G. C. (**2009**). "Permutation tests for studying classifier performance," in *2009 Ninth IEEE International Conference on Data Mining*, pp. 908–913.
Ospina, R., and Ferrari, S. L. (**2012**). "A general class of zero-or-one inflated beta regression models," Comput. Stat. Data Anal. **56**(6), 1609–1623.
Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. (**2019**). "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Interspeech 2019*, pp. 2613–2617.
Pervaiz, A., Hussain, F., Israr, H., Tahir, M. A., Raja, F. R., Baloch, N. K., Ishmanov, F., and Zikria, Y. B. (**2020**). "Incorporating noise robustness in speech command recognition by noise augmentation of training data," Sensors **20**(8), 2326.
Pulgram, E. (**1970**). *Syllable, Word, Nexus, Cursus* (Mouton, The Hague).
Recasens, D. (**1984**). "Vowel-to-vowel coarticulation in Catalan VCV sequences," J. Acoust. Soc. Am. **76**, 1624–1635.
Recasens, D., Pallarès, M. D., and Fontdevila, J. (**1997**). "A model of lingual coarticulation based on articulatory constraints," J. Acoust. Soc. Am. **102**, 544–561.
Redi, L., and Shattuck-Hufnagel, S. (**2001**). "Variation in the realization of glottalization in normal speakers," J. Phonetics **29**, 407–429.
Sakoe, H., and Chiba, S. (**1978**). "Dynamic programming algorithm optimization for spoken word recognition," IEEE Trans. Acoust, Speech, Signal Process. **26**(1), 43–49.
Saltzman, E. L., and Munhall, K. G. (**1989**). "A dynamical approach to gestural patterning in speech production," Ecol. Psychol. **1**(4), 333–382.
Schiller, N. O., Mever, A. S., and Levelt, W. J. (**1997**). "The syllabic structure of spoken words: Evidence from the syllabification of intervocalic consonants," Lang. Speech **40**(2), 103–140.
Schönbrodt, F. D., and Wagenmakers, E.-J. (**2018**). "Bayes factor design analysis: Planning for compelling evidence," Psychon. Bull. Rev. **25**, 128–142.
Selkirk, E. O. (**1982**). "The syllable," in *The Structure of Phonological Representations, Part II*, edited by H. v. d. Hulst, and N. Smith (Foris Publications, Dordrecht, The Netherlands), pp. 337–383.
Semeniuta, S., Severyn, A., and Barth, E. (**2016**). "Recurrent dropout without memory loss," arXiv:1603.05118.
Sharma, J., Granmo, O.-C., and Goodwin, M. (**2020**). "Environment sound classification using multiple feature channels and attention based deep convolutional neural network," in *Interspeech 2020*, pp. 1186–1190.
Shattuck-Hufnagel, S. (**2011**). "The role of the syllable in speech production in American English: A fresh consideration of the evidence," in *Handbook of the Syllable*, edited by C. E. Cairns and E. Raimy (Brill, published online), pp. 197–224.
Smith, J. L. (**2001**). "Lexical category and phonological contrast," in *Workshop on the Lexicon*, pp. 61–72.
Soltau, H., Liao, H., and Sak, H. (**2016**). "Neural speech recognizer: Acoustic-to-word LSTM model for large vocabulary speech recognition," arXiv:1610.09975.
Steriade, D. (**1999**). "Alternatives to syllable-based accounts of consonantal phonotactics," in *Item Order in Language and Speech*, pp. 205–245.
Stetson, R. H. (**1951**). *Motor Phonetics:A Study of Speech Movements in Action* (North Holland, Amsterdam).
Stone, J. V. (**2013**). *Bayes' Rule: A Tutorial Introduction to Bayesian Analysis* (Sebtel).
Strycharczuk, P., and Kohlberger, M. (**2016**). "Resyllabification reconsidered: On the durational properties of word-final /s/ in Spanish," Lab. Phonology **7**, 1–24.
Strycharczuk, P., and Scobbie, J. M. (**2017**). "Fronting of Southern British English high-back vowels in articulation and acoustics," J. Acoust. Soc. Am. **142**(1), 322–331.
Tiffany, W. R. (**1980**). "The effects of syllable structure on diadochokinetic and reading rates," J. Speech. Lang. Hear. Res. **23**, 894–908.
Tilsen, S. (**2019**). "Motoric mechanisms for the emergence of non-local phonological patterns," Front. Psychol. **10**, 2143.
Tilsen, S. (**2020**). "Detecting anticipatory information in speech with signal chopping," J. Phonetics **82**, 100996.
Tilsen, S., Kim, S. E., and Wang, C. (**2021**). "Localizing category-related information in speech with multi-scale analyses," PLoS One **16**(10), e0258178.
Tuller, B., and Kelso, J. A. S. (**1990**). "Phase transitions in speech production and their perceptual consequences," in *Attention and Performance*, edited by M. Jeannerod (Erlbaum, Hillsdale, NJ), Vol. 13.
Tuller, B., and Kelso, J. A. S. (**1991**). "The production and perception of syllable structure," J. Speech. Lang. Hear. Res. **34**, 501–508.
Uffmann, C. (**2007**). "Intrusive [r] and optimal epenthetic consonants," Lang. Sci. **29**(2–3), 451–476.
Wagenmakers, E.-J., Morey, R. D., and Lee, M. D. (**2016**). "Bayesian benefits for the pragmatic researcher," Curr. Dir. Psychol. Sci. **25**(3), 169–176.
Wagner, M. (**2005**). *Prosody and Recursion* (MIT, Cambridge, MA).
Wang, B., Xu, Y., and Ding, Q. (**2017**). "Interactive prosodic marking of focus, boundary and newness in Mandarin," Phonetica **75**(1), 24–56.
Wu, S.-L., Shire, M. L., Greenberg, S., and Morgan, N. (**1997**). "Integrating syllable boundary information into speech recognition," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 987–990.
Xu, A., Birkholz, P., and Xu, Y. (**2019**). Coarticulation as Synchronized dimension-Specific Sequential Target Approximation: An Articulatory Synthesis Simulation, pp. 205–209.
Xu, Y. (**1986**). "Acoustic-phonetic characteristics of junctures in Mandarin Chinese," J. Chin. Linguist. **4**, 353–360.
Xu, Y. (**2020**). "Syllable is a synchronization mechanism that makes human speech possible," PsyArXiv, available at https://doi.org/10.31234/osf.io/9v4hr (Last viewed January 9, 2023).
Xu, Y., and Liu, F. (**2006**). "Tonal alignment, syllable structure and coarticulation: Toward an integrated model," Ital. J. Linguist. **18**, 125–159.
Xu, Y., and Liu, F. (**2007**). "Determining the temporal interval of segments with the help of *F*0 contours," J. Phonetics **35**(3), 398–420.
Xu, Y., and Prom-on, S. (**2019**). "Economy of effort or maximum rate of information? Exploring basic principles of articulatory dynamics," Front. Psychol. **10**, 2469–420.
Xu, Y., and Wang, M. (**2009**). "Organizing syllables into groups — Evidence from F0 and duration patterns in Mandarin," J. Phonetics **37**, 502–520.
Zhang, C., Xia, Q., and Peng, G. (**2015**). "Mandarin third tone sandhi requires more effortful phonological encoding in speech production: Evidence from an ERP study," J. Neurolinguist. **33**, 149–162.
Zhang, X., Sun, J., and Luo, Z. (**2014**). "One-against-all weighted dynamic time warping for language-independent and speaker-dependent speech recognition in adverse conditions," PLoS One **9**(2), e85458.