Deep learning assessment of syllable affiliation of intervocalic consonants

Zirui Liu,[1,a] Yi Xu[1, b]

[1] *Speech, Hearing and Phonetic Sciences, University College London, London, WC1N 1PJ, United Kingdom*

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 the 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 (VOT) or average formant values to determine the occurrence of resyllabification. This study tested the hypothesis that resyllabification is a coarticulatory re-organisation that realigns the coda consonant with the vowel of the next syllable. We used deep learning in conjunction with dynamic time warping (DTW) to assess syllable affiliation of intervocalic consonants. The results suggest that convolutional and recurrent neural network (CNN-RNN) 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

[a] zirui.liu.17@ucl.ac.uk
[b] yi.xu@ucl.ac.uk

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.

I. INTRODUCTION

Despite the wide recognition of the syllable as a speech unit among speakers and researchers (Browman & Goldstein, 1992; Levelt, Roelofs & 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 (/ɛs#kːɔt#əs/) can be syllabified as /ɛs#kːɔ#təs / 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 early as 70 years ago, when Stetson used the kymograph to investigate CV and VC production at different speech rates (Stetson, 1951). He observed that in a sequence of syllables like /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 & 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 & 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). But as found in de Jong et al. (2004), listeners agree with each other well only in cases where 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 become very diverse. Those authors 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, that 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. And 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 both 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 & 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 both VC and CVC syllables in many languages (Clements & 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 “anti-phase” 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., 2006:237).

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 & Liu, 2006). There are two differences between this model and articulatory phonology that 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 insufficient amount of times (Nakatani et al., 1981; Xu & Wang, 2009), thus disallowing them to from complete movement cycles (Xu & Prom-on, 2019), and b) syllables constantly vary their duration, due to stress, phrasing and other linguistic factor, which makes it difficult for syllable sequences, together with their constituent segments, to be temporally periodic to make oscillation-based modelling possible.

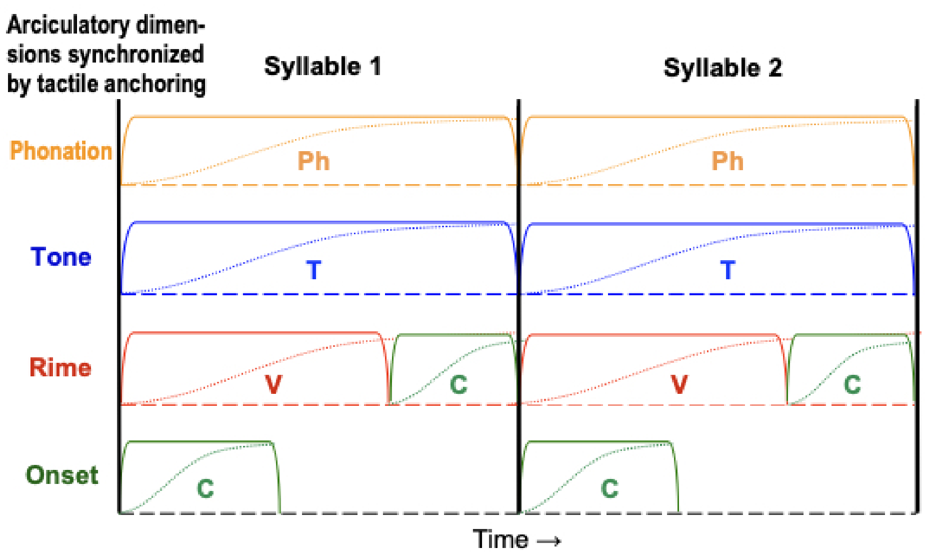

FIG. 1. The synchronisation model of the syllable (Xu, 2020).

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 anti-phase to in-phase articulatory coordination. In Fig. 2A, 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 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. 2B, as speech rate increases, less time is allocated to the syllable, which would require the duration for both 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 & 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)). C_2 now becomes the initial consonant of the second syllable, as shown in Fig. 2C. This reorganisation gives V_1 more articulation time while preserving all the segmental composition of the original syllables.

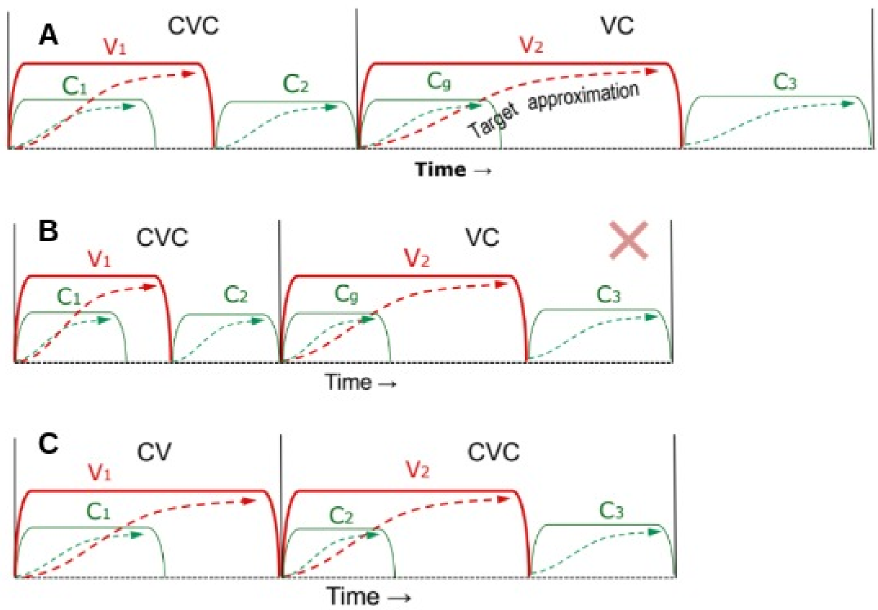

FIG. 2. Illustration of articulatory resyllabification based on the synchronisation model of the syllable.

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 non NN-resyllabified ones. 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 advantages of recent development in machine learning technology. This study therefore aims to determine the occurrences of resyllabification using deep learning models and dynamic time warping 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. Dynamic time warping was then 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 participated in this study, whose first language was Southern Standard British English (6 female and 2 males). No speaking or hearing disorders were reported prior to recording. To ensure data quality, all potential participants had to submit a short recording on Gorilla. The experimenters then visually inspected the recordings in the computer program Praat (Boersma & 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

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, where as “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 & 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 × 20 + 12 × 10). Around 3% of the data were excluded due to background noise during recording.

The recording took place online over Zoom 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 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), with the first boundary being the start of acoustic landmark of onset C_1 (e.g. lateral murmur for /l/), and the second boundary being 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 shown in Fig. 3.

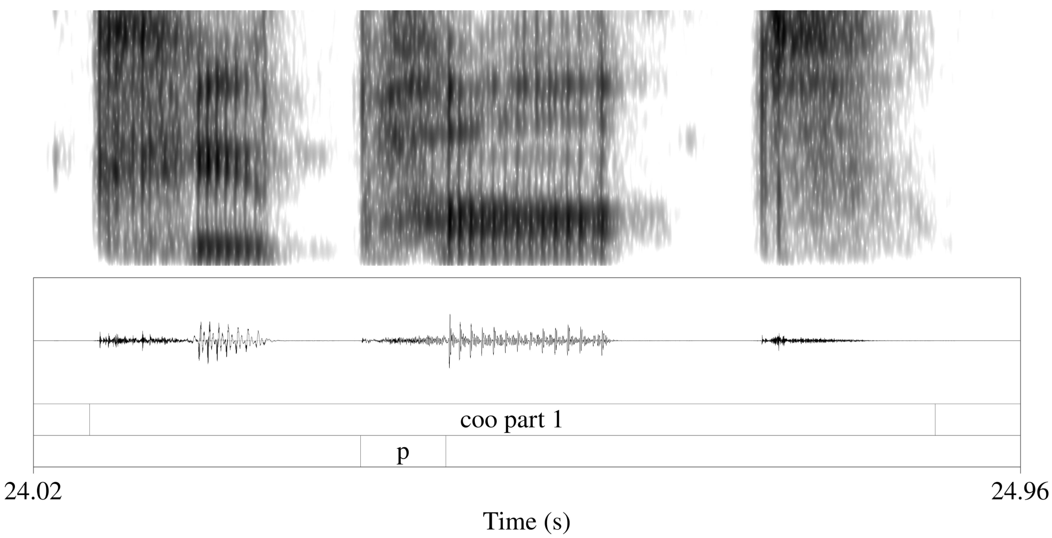

FIG. 3. An annotation example of “coo part” from one speaker, with the vertical lines indicating the segmentations.

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 both the slow and normal rate condition, participants were instructed to speak fluently (i.e. without spontaneous pausing). No spontaneous pauses were 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 the figure shows, speech rate was faster for the normal condition compared to the slow condition for all speakers. On average, speakers produced 2.9 syllables per second for the normal rate and 2 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.

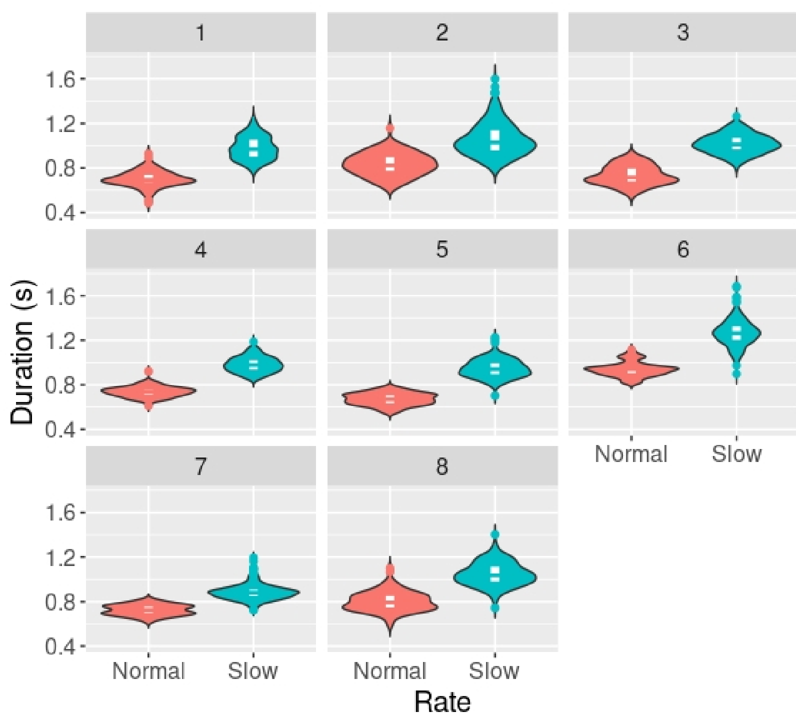

FIG. 4. Annotated sequence duration for 8 speakers.

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 factor of 0.3 to 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 were reserved as the testing set (N = 54), and 85% 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 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 signal 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 (Amodei et al., 2015) and Microsoft (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), then fed into the neural network. Note that Fig. 5 demonstrates the flow of data through the network by a batch size of 1. The spectrogram is first passed through a 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 3 residual blocks (He et al., 2015), the convolutional layers in each residual block had a 5 × 5 kernel with a stride of 1. For both 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 by n timesteps, which was further reduced by a fully connected layer with 512 units. Five layers of bi-directional Gated Recurrent Units (GRU) were then used to process the sequential acoustic features. Only the last timestep'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, one 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 & Szegedy, 2015). The hyperparameters were tuned by using grid search with data from the pilot study. The hyperparameters used can be found in Table II in the Appendix section.

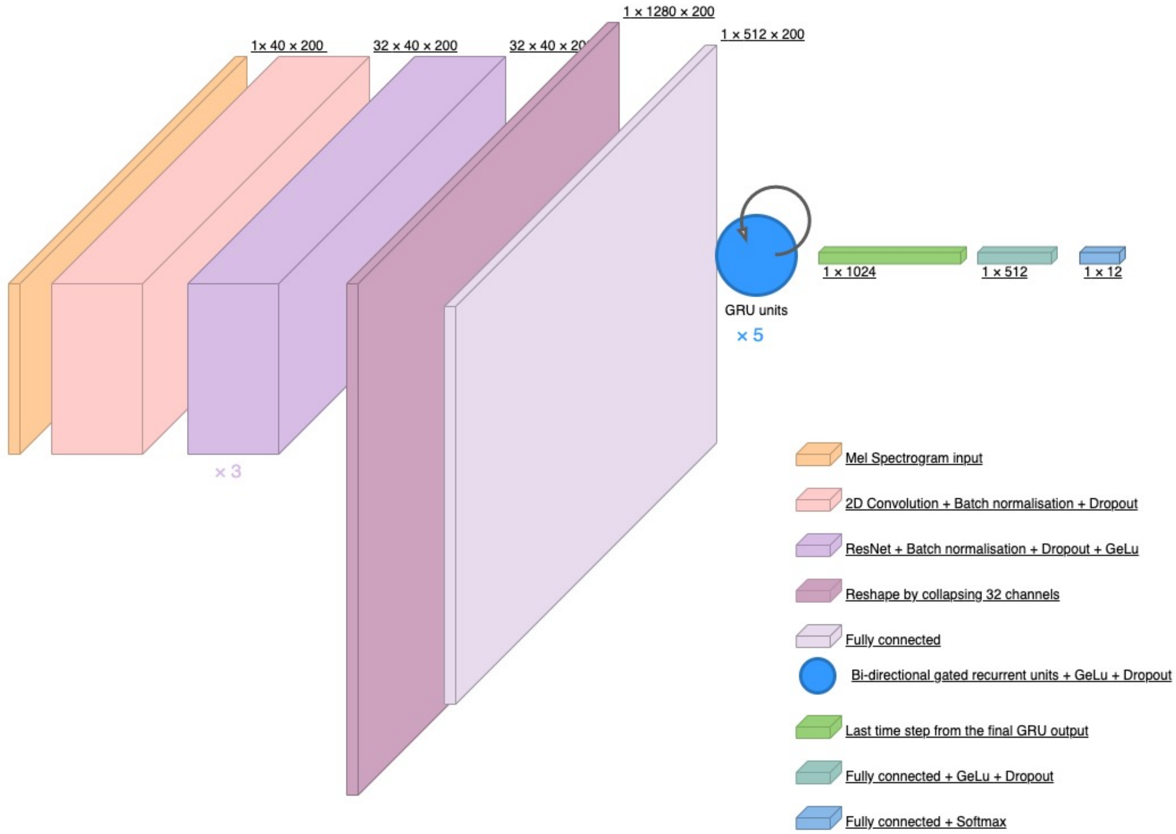

FIG. 5. Model architecture for the resyllabification classifier. The figure shows the tensor dimensions for a batch size of 1. The box sizes reflect tensor shapes, as annotated above each box. The depth, height and width of the boxes are not to scale and is for illustration purposes only.

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. Dynamic time warping analysis

Dynamic time warping (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 & Chiba, 1978; Zhang et al., 2014), as well as other applications such as bird song recognition (Kogan & Margoliash, 1998), speech segment clustering (Lerato & 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-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 0 in cosine distance but not 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 result 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.

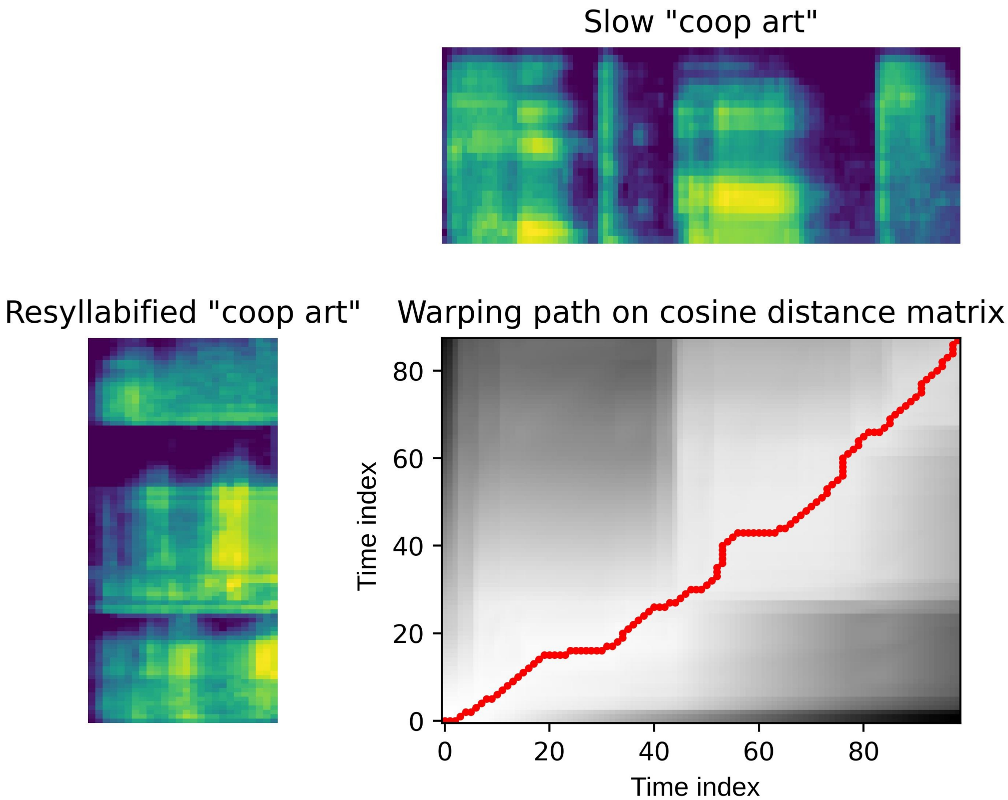

FIG. 6. Demonstration of the DTW algorithm. The dotted line shows the dynamic warping path. The spectrograms are mel-spectrograms of the tokens “coop art” (bottom left) and “coop art” (top). The pixel intensity in the lower right heatmap represent feature distances at each time step between the two spectrograms.

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 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 since 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 correspond to the consonantal release gesture and the other two 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 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 8 speakers were used. From the normal speech rate condition, only the NN-resyllabified tokens and the true onset tokens were examined. According to results from the neural network classifiers, not all 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, due to 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 sub-sampling the onset condition for each speaker to match the number of NN-resyllabified ones. For instance, if a speaker resyllabified 5 out of 10 repetitions of “coop eat”, only 5 random selections of “coo Pete” were used from this speaker for training the binary classifier.

The classifiers were bi-directional recurrent neural networks with Long Short-Term Memory (LSTM) units (Soltau et al., 2016). The network details are shown 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 section. The segmented tokens were converted into melspectrogams with 26 filter banks, with 0.025 s as the window length and 0.005 s as the hop length. Before training, all 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 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 & Jerbi, 2015; Ojala & Garriga, 2009).

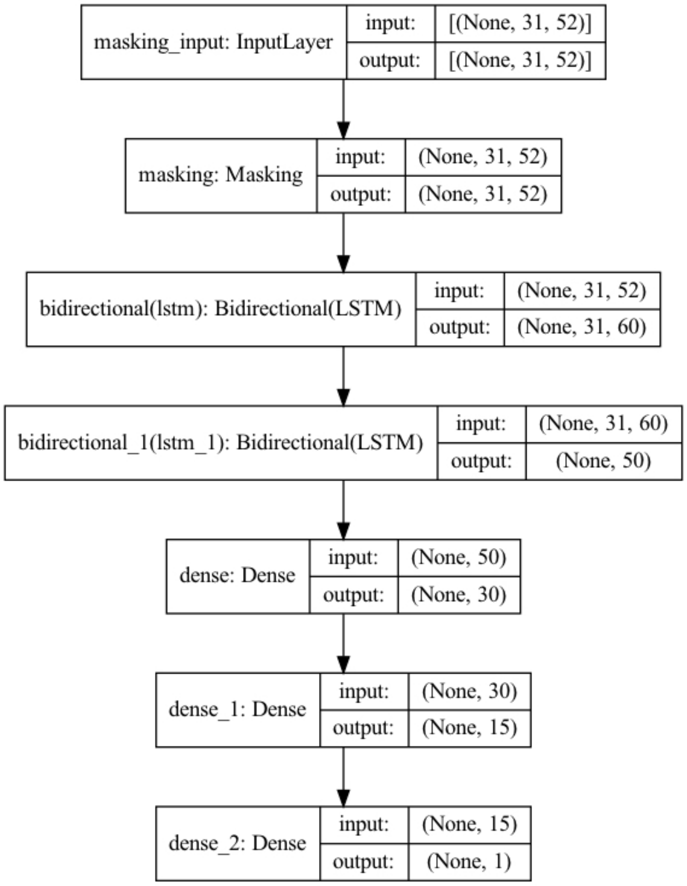

FIG. 7. Model architecture of the binary classifiers. The tensor shapes are denoted on the right of each box.

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 non-significant 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 due to the nature of accuracy rate being bounded between 0 and 1. Beta regression assumes that the data generating process can be modelled by a beta distribution (Balakrishnan & Nevzorov, 2003), where the distribution can be parameterised with the mean-precision (μ-φ) parameters, where φ is analogous to the inverse of data dispersion. Since $Y \sim \text{Beta}(\mu, \varphi)$, beta regression presumes that the mean μ of the response given the predictor X is linear on the logit transformed scale (Douma & 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 parameterisation is:

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

and:

$$\mu=logit^{-1}(X\beta) \tag{2}$$

Γ is the Gamma function, μ is the inverse logit transformed model prediction, y is the observed data bounded between 0 and 1, 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 & Ferrari, 2012). The likelihood function using the one-inflated beta distribution incorporates a new parameter α:

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

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 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 (BF) 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) demonstrates 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 shown 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 towards the more complex model, we use BF for model comparison.

If BF_0 (the BF indicating evidence for H_0 over H_1) is between 0 and 1/10, the data strongly supports H_1 over H_0. Conversely, if BF_0 is larger than 10, there is strong evidence for the null hypothesis (Jeffreys, 1961; Biel & Friedrich, 2018; Dienes, 2014; Harms & Lakens, 2018; Lakens et al., 2020; Schönbrodt & Wagenmakers, 2018; Lee & 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 onset 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 & Xu, 2010; Lehiste, 1960). Specifically, the durations of the resyllabified consonants are shorter compared to the canonical onsets. To test whether duration differs between the two, the same acoustic intervals from the previous section 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

Fig. 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.

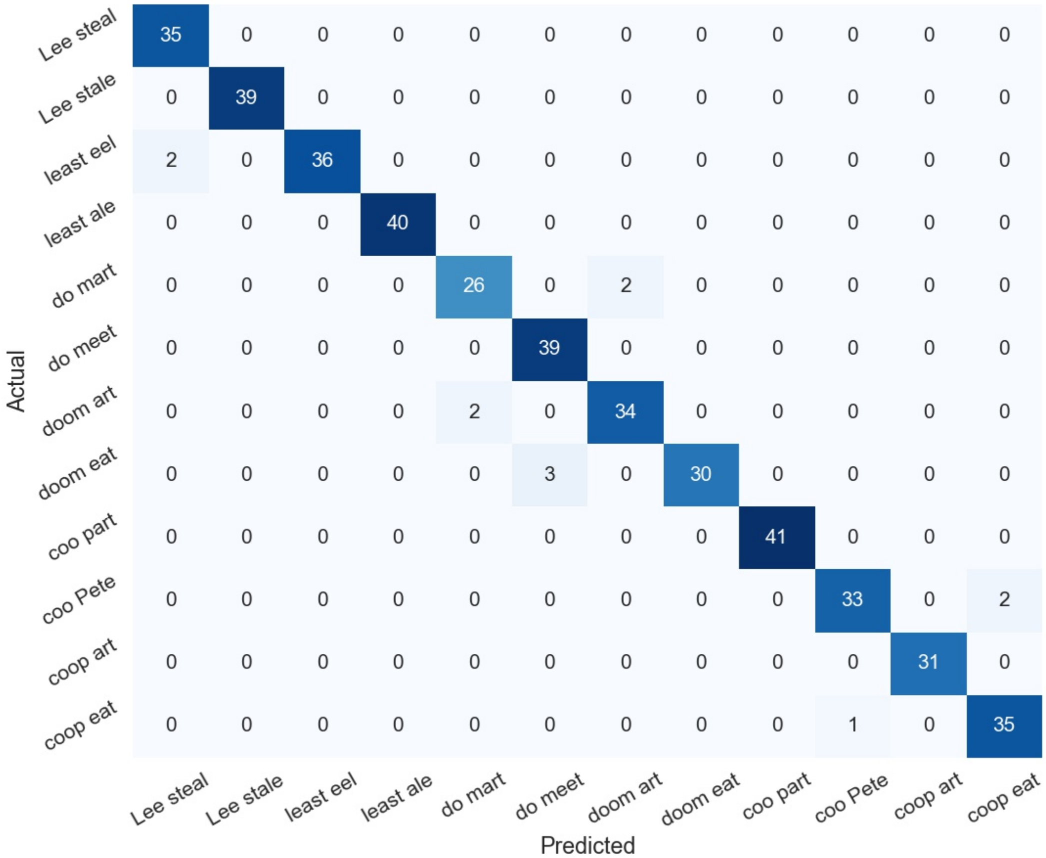

FIG. 8. Confusion matrix of model performance on the testing split of the slow speech rate. This is an element wise summation of all the speakers' confusion matrices. The colour intensity of tiles reflects numeric value.

Fig. 9 shows the model performance on the normal speaking rate by summing over the results from all the speakers. Table IV list the accuracy rate for the onset, coda and all sequences. As can be seen, 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 counterpart, such as classifying “least eel” as “Lee steal”. These misclassified sequences, presumably due to resyllabification, are examined in detail later.

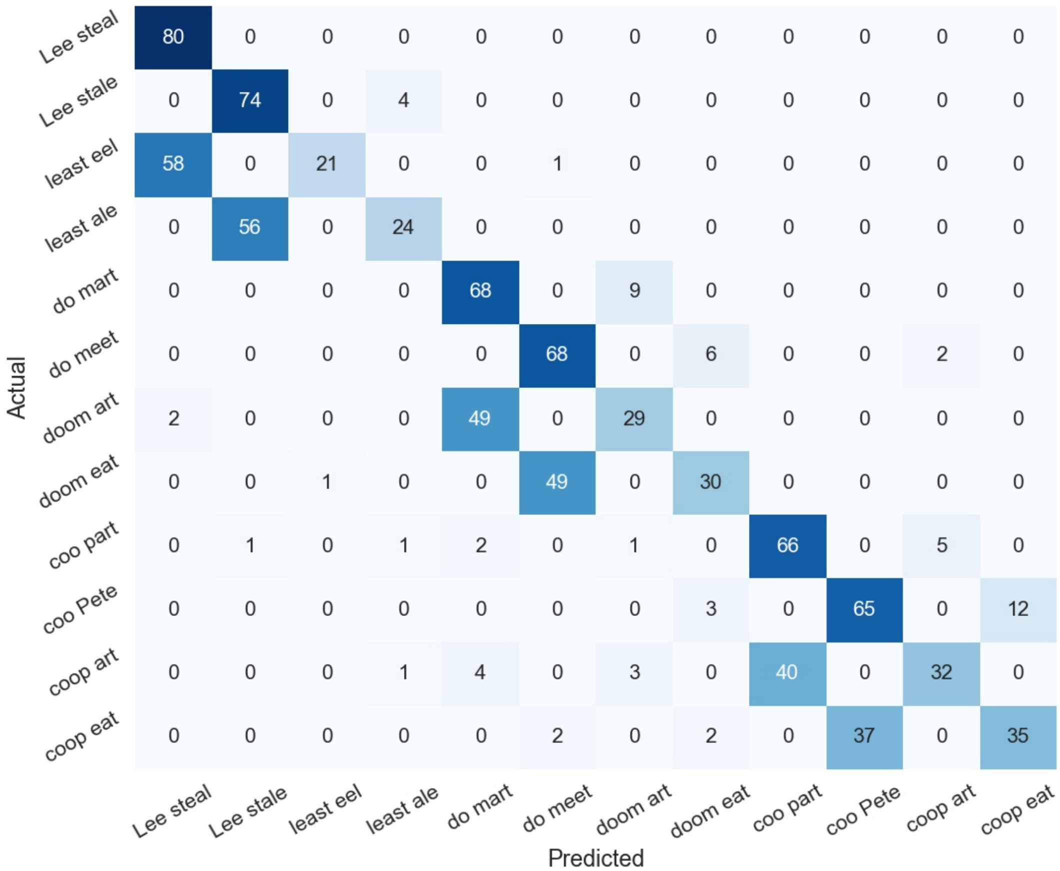

FIG. 9. Confusion matrix of model performance on the normal speech rate. This is an element wise summation of all the speakers’ confusion matrices. The colour intensity of tiles reflects numeric value.

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

Coda	0.36
Onset	0.90
Overall	0.63

B. DTW analysis

Fig. 10 shows a bar graph of the cosine distance between the NN-resyllabified tokens and the slow tokens. The NN-resyllabified sequences were only compared to slow sequences in the same group. The figure 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.

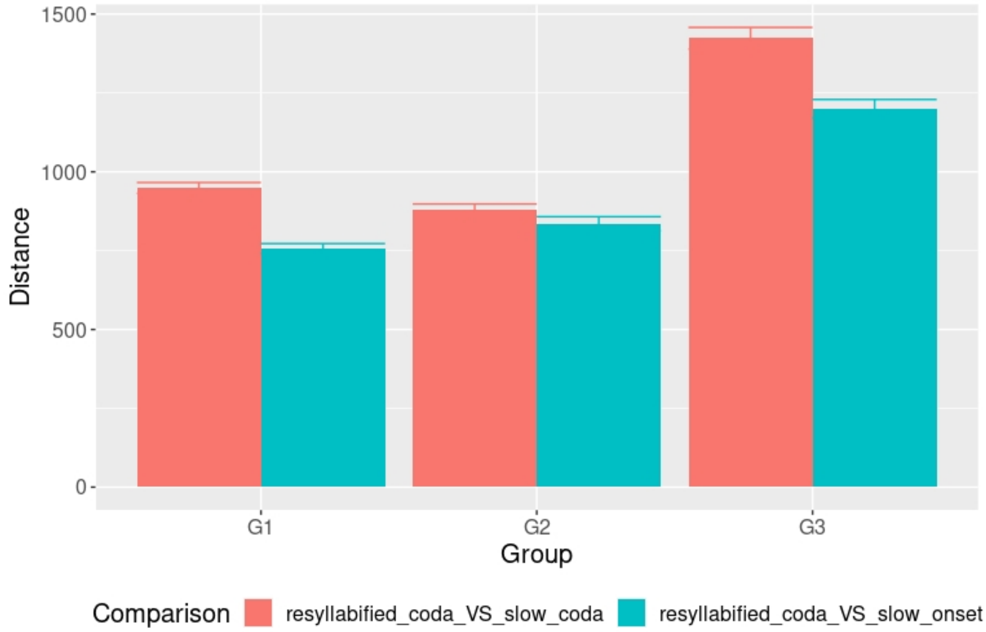

FIG. 10. DTW cosine distance between resyllabifed normal rate sequences and slow sequences. The error bars represent 95% of the confidence interval. G1 – “least eel”, “least ale”, “Lee stale”, “Lee steel”; G2 – “doom art”, “doom eat”, “do mart”, “do meet”; G3 – “coop art”, “coop eat”, “coo part”, “coo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

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 the previous section, 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. The bottom panel shows “doom art” spoken in the slow condition, likely with a glottal stop at the beginning of the second syllable “art”.

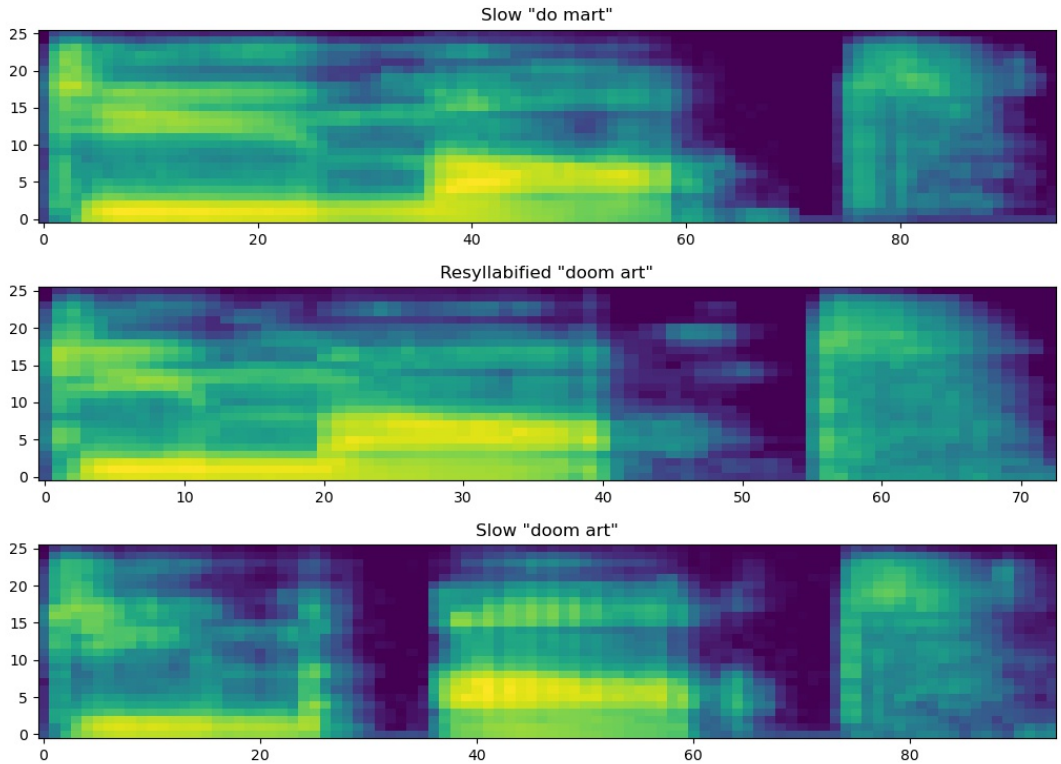

FIG. 11. Mel-spectrograms of three word sequences from one speaker.

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

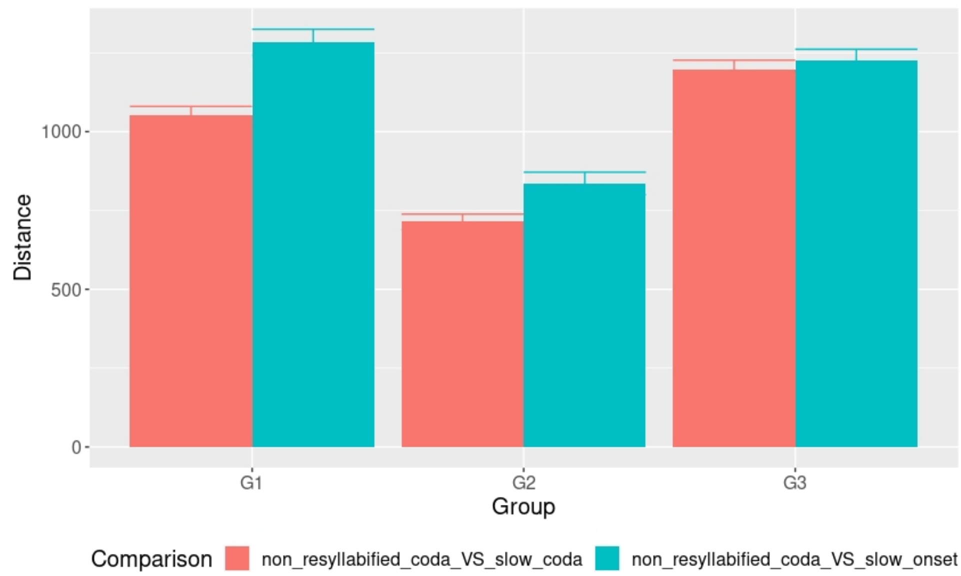

FIG. 12. DTW cosine distance between non NN-resyllabifed normal rate sequences and slow sequences. The error bars represent 95% of the confidence interval. G1 – “least eel”, “least ale”, “Lee stale”, “Lee steel”; G2 – “doom art”, “doom eat”, “do mart”, “do meet”; G3 – “coop art”, “coop eat”, “coo part”,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

C. Intervocalic consonant alignment analysis

1. Results for slow speech rate

With the consonant intervals described in Section II E, we trained 80 neural networks for each vowel minimal pair in Table I and obtained an accuracy distribution from the test set. Fig. 13 shows the accuracy rate from the slow speech rate condition. As the figure shows, 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 both 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.

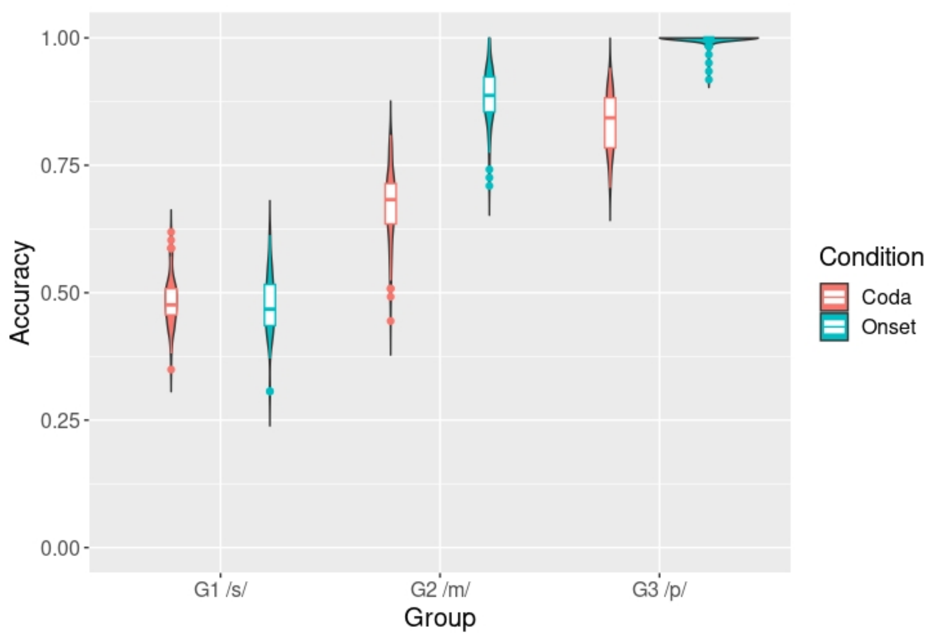

FIG. 13. Vowel classification accuracy by group from the slow speech rate condition. G1 – “least eel”, “least ale”, “Lee stale”, “Lee steel”; G2 – “doom art”, “doom eat”, “do mart”, “do meet”; G3 – “coop art”, “coop eat”, “coo part”, “coo Pete”.

To test hypothesis via model comparison, we use the Bayes Factor, which can offer support for a model based on the observed data (Dienes, 2014, Harm & Lakens, 2018). The posterior distributions of the model parameters are not very informative as predictions need to be transformed with the inverse logit function, and their details are included as supplementary materials[4]. Therefore, the predicted distribution from 100 random samples is shown in Fig. 14. As the figure shows, the model with an interaction term shows the best predicative power. BF_0 was very close to zero (i.e. BF_1 is larger than 10). Therefore, 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 10, which provides strong support for the interaction model. To conclude, the data shows 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.

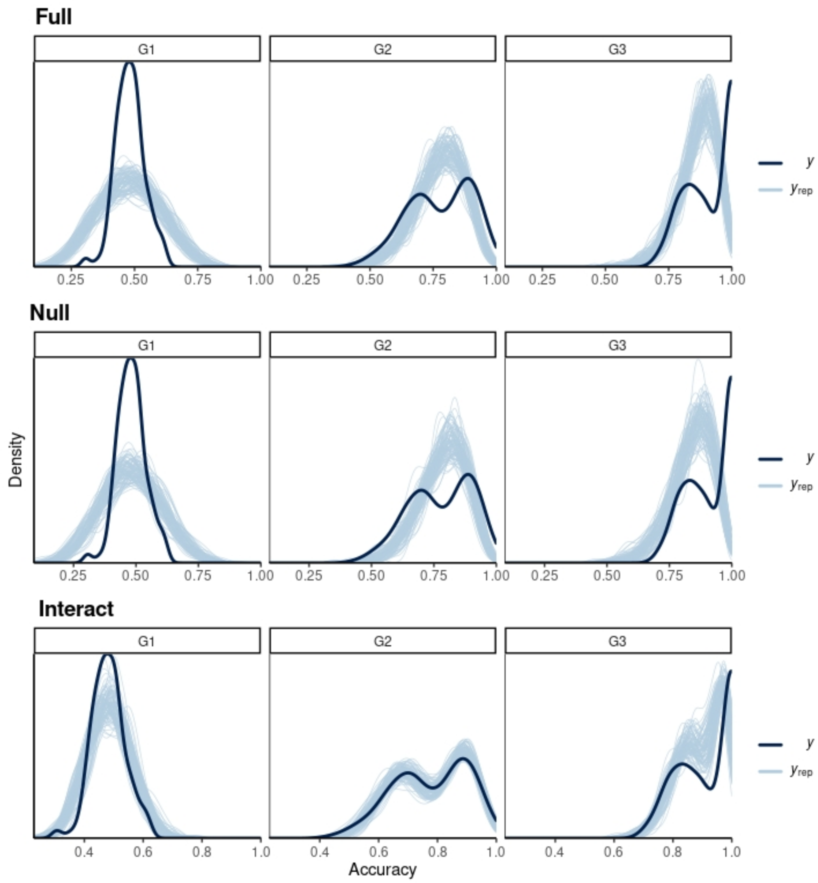

FIG. 14. Model predictions against 100 random samples for the slow rate, where y refers to the observed data and y_{rep} refers to predictions. The columns correspond to item groups and the rows correspond to model type.

2. Results for normal speech rate

The accuracy distributions from the normal speech rate condition are shown in Fig. 15. Note that the coda condition only contained NN-resyllabified sequences. Fig. 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.

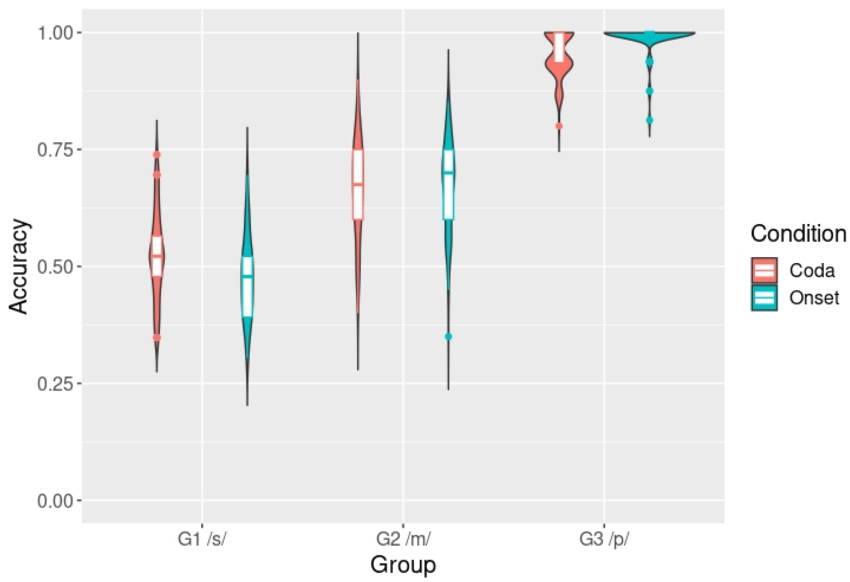

FIG. 15. Vowel classification accuracy by group from the normal speech rate condition. The coda condition here refers to the NN-resyllabified coda sequences in the normal speech rate condition. G1 – “least eel”, “least ale”, “Lee stale”, “Lee steel”; G2 – “doom art”, “doom eat”, “do mart”, “do meet”;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

The predicted distributions from the Bayesian analysis results are shown in Fig. 16. The posterior distributions of model parameters can be found in the supplemented materials[5]. Visually, the predicted distributions do not differ too much from one another. BF_0 was larger than 10, signifying that the data provides more support for the null model. Fig. 15 indicates that model accuracy might differ slightly between the NN-resyllabified coda and the 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.

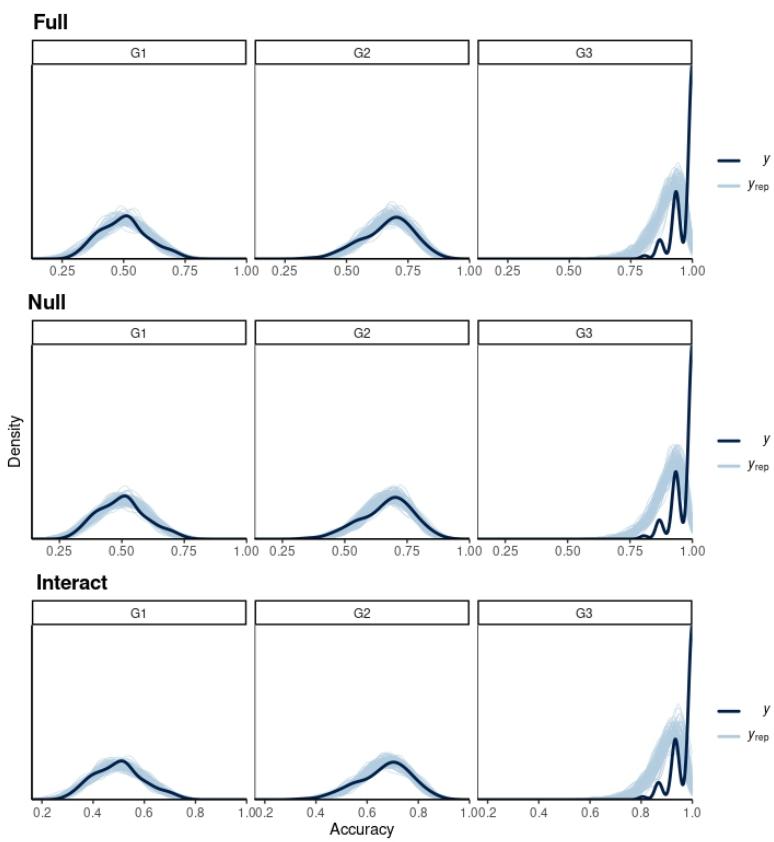

FIG. 16. Model predictions against 100 random samples for the normal rate, where y refers to the observed data and y_{rep} refers to predictions. The columns correspond to item groups and the rows correspond to model type.

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 & Xu, 2010; Lehiste, 1960), NN-resyllabified codas are shorter than the canonical onsets. Predictions of the Bayesian analysis are shown in Fig. 18, and the parameter posterior distributions are included as supplemented materials[6]. 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.

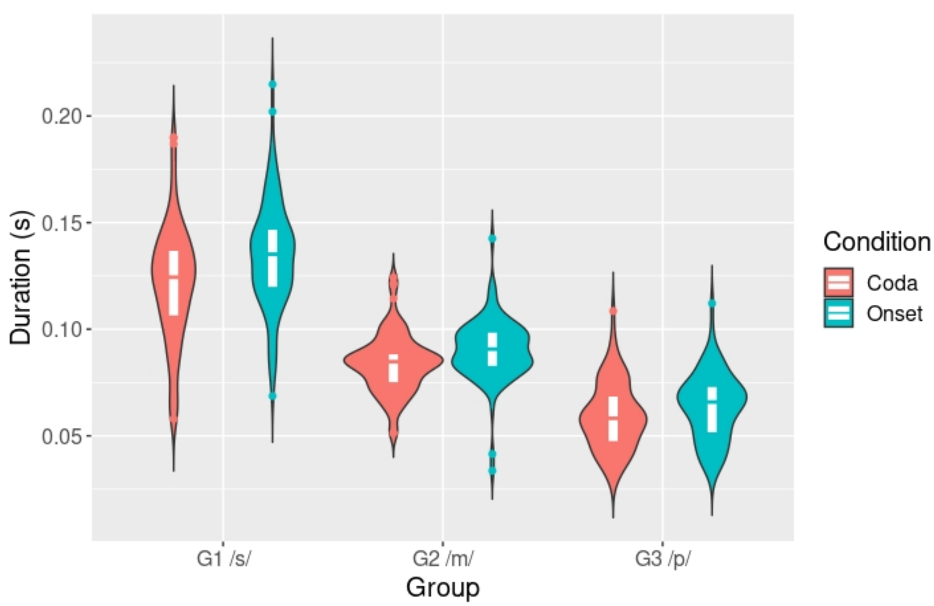

FIG. 17. Duration of onset and NN-resyllabified consonants from the normal speaking rate condition.

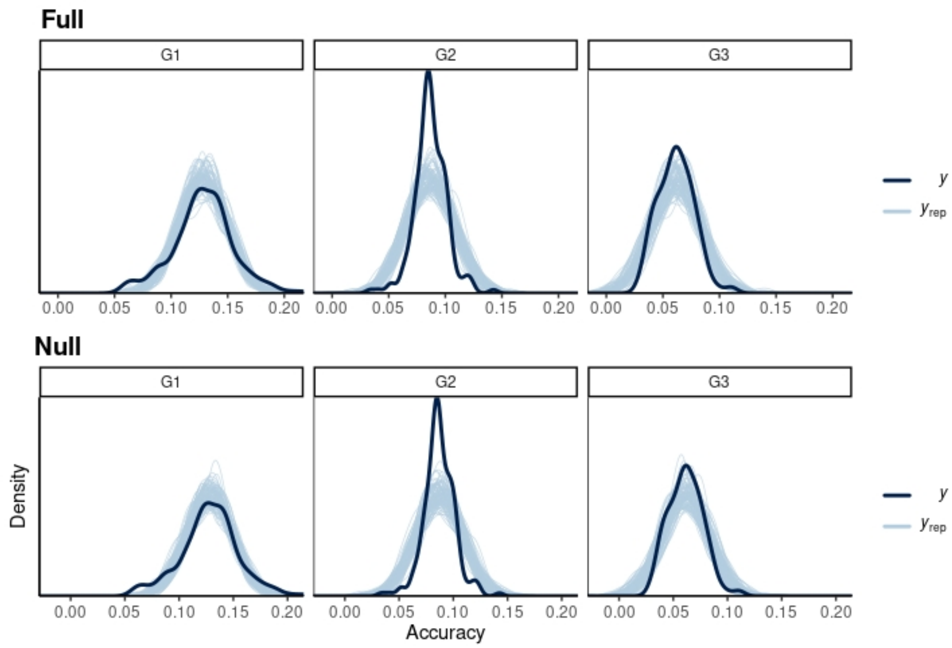

Fig. 18. Model predictions against 100 random samples for the duration results, where y refers to the observed data and y_{rep} refers to predictions. The columns correspond to item groups and the rows correspond to model type.

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 indicate that they likely have not undergone resyllabification. Therefore, the first prediction — codas in 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 & 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 both singleton consonants and consonant clusters. The coda status of the NN-resyllabified consonants, however, seem 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, however, need to be tested in future studies. Furthermore, future studies can investigate the effect of resyllabification and syllable position on consonant duration by examining both 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). But this is consistent with the finding of de Jong (2001) that resyllabification start 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 avoidat normal speech rate.

The finding of resyllabification align with the syllable model shown in Fig. 1 based on which 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 towards the vowel and onset C are synchronised at syllable onset (Liu et al., 2022; Liu & 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, partly due to *coarticulation resistance*, i.e. the ability of a segment to restrain coarticulatory effects from adjacent segments (Bladon & 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 co-produced to achieve C-V co-onset (Fig. 1). According to this mechanism, namely, *dimension-specific sequential target approximation 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 &

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 includes 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, which is more obvious, is the differences in their relative timing. The frication in Group 1 and nasal murmur in Group 2 both 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 Groups 1 involves the tongue body to form a groove needed to direct the airflow toward the front teeth (Borden, Harris and Raphael, 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 both involved in approaching the target of the sibilant (Recasens & Espinosa, 2009). 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. But 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 & 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”, *before* the frication onset. But 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 not /st/ as in the current study. /p/ does not require any tongue movement, thus is less coarticulation resistant than both /l/ and /t/. In terms of /l/ and /t/, both being alveolars, Iskarous et al. (2013) found that /t/ is more coarticulation resistant than /l/ in the vertical dimension for the jaw and the tongue blade. This could be due to 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 not essential for the consonant articulation, and the more detectable the vowel information 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 said 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 is decisive, or 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 is the finding that, as shown in Fig. 13, 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 Fig. 19 and Fig. 20. But the glottal component, as can be judged both auditorily and spectrally, corresponds to a glottal stop or glotallisation (which is also a form of glottal stop: Redi & Shattuck-Hufnagel, 2001; Garellek, 2013), that serves as the onset of the syllable /art/. A glottal stop, just like 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 seen 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. Exactly when during the closure, however, awaits future investigations.

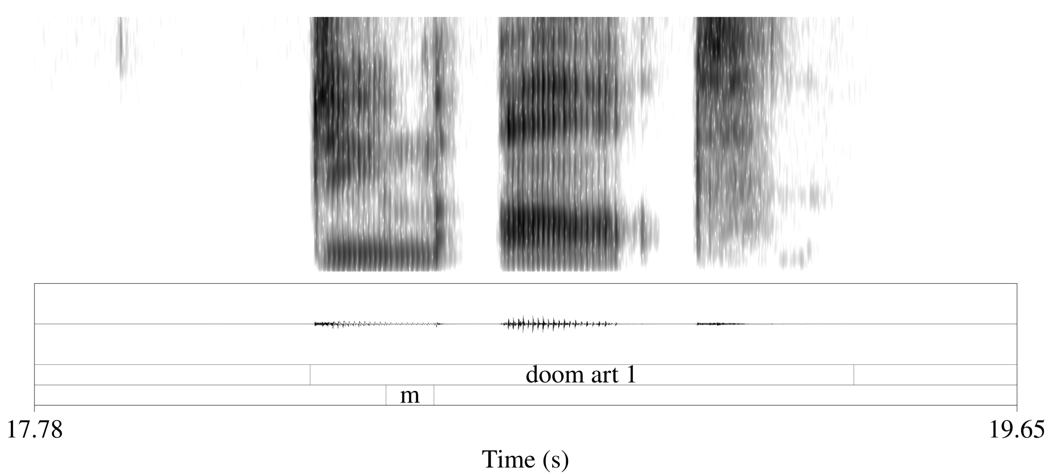

FIG. 19. Spectrogram of “doom art” from a male speaker.

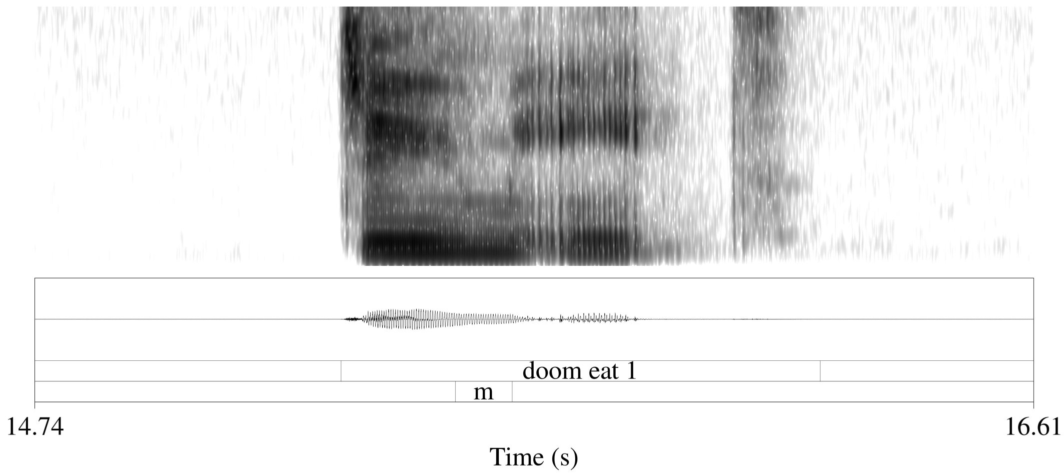

FIG. 20. Spectrogram of “doom eat” from a female speaker.

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 & Xu et al., 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 done in the present study. Because the syllable is both 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 ones. The first is about the influential psycholinguistic model of speech 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 that 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, O'Connor & Aston, 1981; Wagner, 2005; Wang & Xu, 2019). 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, Xia & Peng, 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, so 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., simultaneously 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, 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 the current study has examined.* Although the present 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 c.f. 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 section 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 exist during frication, two scenarios could result in false negative detections:

1. Chance performance due to chance.

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 section 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 (2 syllables per second on average), is it unlikely that resyllabification occurred, so 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 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 ones. 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 section 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 co-produced 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 bi-directional recurrent neural network models. We found that similar amount of vowel information was detected in the intervocalic consonants between the NN-resyllabified codas and the 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 multi-class classifier and the binary classifiers are shown in Table II and Table III, respectively.

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

TABLE III. Hyperparameters for the binary classifiers.

Hyperparameter	Value
Number of units in the first LSTM	60

layer	
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

[1]During data splitting, correlated samples due to 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

[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

[4]See supplementary materials at [URL] for details on the posterior distributions for the slow rate condition.

[5]See supplementary materials at [URL] for details on the posterior distributions for the normal rate condition.

[6]See supplementary materials at [URL] for details on the posterior distributions for the duration analysis.

REFERENCES (BIBLIOGRAPHIC)

Adda-Decker, M., de Mareüil, P. B., Adda, G., & 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., … Zhu, Z. (2015). *Deep Speech 2: End-to-End Speech Recognition in English and Mandarin*. https://doi.org/10.48550/ARXIV.1512.02595

Audacity Team. (2021). *Audacity*. https://audacityteam.org/

Balakrishnan, N., & Nevzorov, V. (2003). A primer on statistical distributions. Hoboken, NJ: John Wiley and Sons.

Barlow, J. A., & Gierut, J. A. (1999). Optimality Theory in Phonological Acquisition. *Journal of Speech, Language, and Hearing Research*, *42*(6), 1482–1498. https://doi.org/10.1044/jslhr.4206.1482

Bartelds, M., Richter, C., Liberman, M., & Wieling, M. (2020). A New Acoustic-Based Pronunciation Distance Measure. *Frontiers in Artificial Intelligence*, *3*, 39. https://doi.org/10.3389/frai.2020.00039

Bermúdez-Otero, R. 2011. Cyclicity. In van Oostendorp, M., Ewen, C J., Hume, E., & Rice, K. (Eds.), Blackwell companion to phonology. Chichester: Wiley-Blackwell, Vol. 4, pp. 2019–2048.

Biel, A. L., & Friedrich, E. V. C. (2018). Why You Should Report Bayes Factors in Your Transcranial Brain Stimulation Studies. *Frontiers in Psychology*, *9*, 1125. https://doi.org/10.3389/fpsyg.2018.01125

Birgit A. (2001). Regional Variation and Edges: Glottal Stop Epenthesis and Dissimilation in Standard and Southern Varieties of German. *Zeitschrift Für Sprachwissenschaft*, *20*(1), 3–41. https://doi.org/doi:10.1515/zfsw.2001.20.1.3

Bladon, R. A. W., & Al-Bamerni, A. (1976). Coarticulation resistance in English /l/. *Journal of Phonetics*, *4*(2), 137–150. https://doi.org/10.1016/S0095-4470(19)31234-3

Borden, G. J., Harris, K. S., and Raphael, L. J. (2003). *Speech Science Primer: Physiology, Acoustics, and Perception of Speech, 4th Edition* (Williams & Wilkins, Baltimore).

Blevins, J. (2003). *Evolutionary phonology: The emergence of sound patterns*. Cambridge University Press.

Boersma, P., & Weenink, D. (2022). *Praat: Doing phonetics by computer* (6.2.14) [Computer software]. http://www.praat.org/

Browman, C. P., & Goldstein, L. (1992). Articulatory Phonology: An Overview. *Phonetica*, *49*(3–4), 155–180. https://doi.org/10.1159/000261913

Chen, M. Y. (2000). *Tone Sandhi: Patterns across Chinese Dialects* (Cambridge University Press, Cambridge, UK).

Clements, G. N., & Keyser, S. J. (1983). CV phonology. A generative theory of the syllable. *Linguistic Inquiry Monographs Cambridge*, *9*, 1–191.

Combrisson, E., & Jerbi, K. (2015). Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *Journal of Neuroscience*

Methods, *250*, 126–136. https://doi.org/10.1016/j.jneumeth.2015.01.010

Content, A., Kearns, R. K., and Frauenfelder, U. H. (2001). Boundaries versus Onsets in Syllabic Segmentation. *Journal of Memory and Language* **45**, 177-199.

de Jong, K. J. (2001). Rate-Induced Resyllabification Revisited. *Language and Speech*, *44*(2), 197–216. https://doi.org/10.1177/00238309010440020401

de Jong, K. J., Lim, B., & Nagao, K. (2004). The Perception of Syllable Affiliation of Singleton Stops in Repetitive Speech. Language and Speech, 47(3), 241–266.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00781

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, *72*, 78–89. https://doi.org/10.1016/j.jmp.2015.10.003

Douma, JC & Weedon, JT. (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), pp. 41-69.

Gao, H., & Xu, Y. (2010). Ambisyllabicity in English: How real is it? *In Proceeding of the 9*th *Phonetic Conference of China*, Tianjin.

Gao, M. (2009). Gestural coordination among vowel, consonant and tone gestures in Mandarin Chinese. *Chinese Journal of Phonetics*, *2*, 43–50.

Garellek, M. (2012). Glottal stops before word-initial vowels in American English: distribution and acoustic characteristics. *UCLA Working Papers in Phonetics*, 110, 1-23.

Garellek, M. (2013). Production and perception of glottal stops (Doctoral dissertation, UCLA). Retrieved from https://escholarship.org/uc/item/7zk830cm

Gaskell, M. G., Spinelli, E., & Meunier, F. (2002). Perception of resyllabification in French. *Memory & Cognition*, 30(5), 798–810. https://doi.org/10.3758/BF03196435

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., & Rubin, D.B. (2013). Bayesian Data Analysis (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b16018

Gick, B. (1999). A gesture-based account of intrusive consonants in English. *Phonology*, *16*(1), 29–54. https://doi.org/10.1017/S0952675799003693

Gick, B. (2003). Articulatory correlates of ambisyllabicity in English glides and liquids. Phonetic interpretation: Papers in laboratory phonology, 6, 222-236.

Goldstein, L., Byrd, D., & 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* (pp. 215–249). https://doi.org/10.1017/CBO9780511541599.008

Goslin, J., & Frauenfelder, U. H. (2001). A Comparison of Theoretical and Human Syllabification. *Language and Speech*, *44* (4), 409–436. https://doi.org/10.1177/00238309010440040101

Gronau, Q.F., Wagenmakers, EJ. Limitations of Bayesian Leave-One-Out Cross-Validation for Model Selection. *Comput Brain Behav,* 2, 1–11 (2019). https://doi.org/10.1007/s42113-018-0011-7

Harms, C., & Lakens, D. (2018). Making ‘Null Effects’ Informative: Statistical Techniques and Inferential Frameworks. *Journal of Clinical and Translational Research*, 24.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition* (arXiv:1512.03385). arXiv. http://arxiv.org/abs/1512.03385

Ioffe, S., & Szegedy, C. (2015). *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift* (arXiv:1502.03167). arXiv. http://arxiv.org/abs/1502.03167

Iskarous, K., Mooshammer, C., Hoole, P., Recasens, D., Shadle, C. H., Saltzman, E., & Whalen, D. H. (2013). The coarticulation/invariance scale: Mutual information as a measure of coarticulation resistance, motor synergy, and articulatory invariance. *The Journal of the Acoustical Society of America*, *134*(2), 1271–1282. https://doi.org/10.1121/1.4812855

Jacewicz, E., Fox, R. A., O'Neill, C., & Salmons, J. (2009). Articulation rate across dialect, age, and gender. *Language variation and change*, *21*(2), 233–256. https://doi.org/10.1017/S0954394509990093

Jeffreys, H. (1961). *The theory of probability (3rd ed.*). Oxford University Press.

Kahn, D. (1976). Syllable-based generalizations in English phonology. ((Doctoral dissertation, MIT).

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., & Khudanpur, S. (2015). *Audio augmentation for speech recognition*. Interspeech 2015, 3586–3589. https://doi.org/10.21437/Interspeech.2015-711

Kogan, J. A., & Margoliash, D. (1998). Automated recognition of bird song elements from continuous recordings using dynamic time warping and hidden Markov models: A comparative study. *The Journal of the Acoustical Society of America, 103*(4), 2185–2196. https://doi.org/10.1121/1.421364

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2020). Improving Inferences About Null Effects With Bayes Factors and Equivalence Tests. *The Journals of Gerontology: Series B, 75*(1), 45–57. https://doi.org/10.1093/geronb/gby065

Lee, M., & Wagenmakers, E. (2014). Bayesian Cognitive Modeling: A Practical Course. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139087759

Lehiste, I. (1960). An acoustic-phonetic study of internal open juncture. *Phonetica Supplement*.

Lehiste, I. (1972). The timing of utterances and linguistic boundaries. *J. Acoust. Soc. Am.* 51, 2018-2024.

Lerato, L., & Niesler, T. (2019). Feature trajectory dynamic time warping for clustering of speech segments. *EURASIP Journal on Audio, Speech, and Music Processing, 2019* (1), 6. https://doi.org/10.1186/s13636-019-0149-9

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(01). https://doi.org/10.1017/S0140525X99001776

Liu, Z., & Xu, Y. (2021). *Segmental Alignment of English Syllables with Singleton and Cluster Onsets*. Interspeech 2021, 3969–3973. https://doi.org/10.21437/Interspeech.2021-187

Liu, Z., Xu, Y., & Hsieh, F. (2022). Coarticulation as synchronised CV co-onset – Parallel evidence from articulation and acoustics. *Journal of Phonetics*, *90*, 101116. https://doi.org/10.1016/j.wocn.2021.101116

Luo, D., Zou, Y., & Huang, D. (2018). *Investigation on Joint Representation Learning for Robust Feature Extraction in Speech Emotion Recognition*. Interspeech *2018*, 152–156. https://doi.org/10.21437/Interspeech.2018-1832

MacNeilage, P. F. (1998). The frame/content theory of evolution of speech production. *Behavial and Brain Sciences* 21, 499–546.

Marin, S., & Pouplier, M. (2014). Articulatory synergies in the temporal organization of liquid clusters in Romanian. *Journal of Phonetics*, *42*, 24–36. https://doi.org/10.1016/j.wocn.2013.11.001

Mirzaei, M. S., Meshgi, K., & Kawahara, T. (2018). Exploiting automatic speech recognition errors to enhance partial and synchronized caption for facilitating second language listening. *Computer Speech & Language*, *49*, 17–36. https://doi.org/10.1016/j.csl.2017.11.001

Mok, P. P. K. (2012). Effects of consonant cluster syllabification on vowel-to-vowel coarticulation in English. *Speech Communication*, *54*(8), 946–956.

Morey, R. D., Romeijn, J. W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, 72, 6-18.

Mudzingwa, C. (2013). Hiatus resolution strategies in Karanga (Shona). *Southern African Linguistics and Applied Language Studies*, *31*(1), 1–24. https://doi.org/10.2989/16073614.2013.793953

Mullooly, R. (2003). *An Electromagnetic Articulography study of resyllabification of rhotic consonants in English*. 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.

Nam, H. (2007). *Articulatory modeling of consonant release gesture*. International Conference of Phonetic Sciences, Saarbrücken.

Nam, H., Goldstein, L., & Saltzman, E. (2009). Self-organization of syllable structure: A coupled oscillator model. In F. Pellegrino, E. Marsico, I.

Chitoran, & C. Coupé (Eds.), *Approaches to Phonological Complexity* (pp. 297–328). Walter de Gruyter. https://doi.org/10.1515/9783110223958.297

Nam, H., Mitra, V., Tiede, M., Hasegawa-Johnson, M., Espy-Wilson, C., Saltzman, E., & Goldstein, L. (2012). A procedure for estimating gestural scores from speech acoustics. *The Journal of the Acoustical Society of America*, *132*(6), 3980–3989. https://doi.org/10.1121/1.4763545

Ni Chiosáin, M. N., Welby, P., & Espesser, R. (2012). Is the syllabification of Irish a typological exception? An experimental study. *Speech Communication*, *54*(1), 68–91. https://doi.org/10.1016/j.specom.2011.07.002

Ospina, R., Ferrari, S L.P. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis,* 56(6), 1609-1623.

Ojala, M., & Garriga, G. C. (2009). Permutation Tests for Studying Classifier Performance. *2009 Ninth IEEE International Conference on Data Mining*, 908–913. https://doi.org/10.1109/ICDM.2009.108

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). *SpecAugment: A Simple Data Augmentation Method for*

Automatic Speech Recognition. Interspeech 2019, 2613–2617. https://doi.org/10.21437/Interspeech.2019-2680

Pastätter, M., & Pouplier, M. (2014). *The articulatory modeling of German coronal consonants using TADA*. International seminar on speech production, Cologne, Germany.

Perkell, J., & Chiang, C. M. (1986). *Preliminary support for a 'hybrid' model of anticipatory coarticulation*. International congress of acoustics, Toronto.

Pervaiz, A., Hussain, F., Israr, H., Tahir, M. A., Raja, F. R., Baloch, N. K., Ishmanov, F., & Zikria, Y. B. (2020). Incorporating Noise Robustness in Speech Command Recognition by Noise Augmentation of Training Data. *Sensors*, *20*(8), 2326. https://doi.org/10.3390/s20082326

Prom-on, S., Xu, Y., & Thipakorn, B. (2009). Modeling tone and intonation in Mandarin and English as a process of target approximation. *The Journal of the Acoustical Society of America*, *125*(1), 405–424. https://doi.org/10.1121/1.3037222

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., & Espinosa, A. (2005). Articulatory, positional and coarticulatory characteristics for clear /l/ and dark /l/: Evidence from two Catalan dialects. *Journal of the International Phonetic Association*, *35*(1), 1–25. https://doi.org/10.1017/S0025100305001878

Recasens, D., & Espinosa, A. (2009). An articulatory investigation of lingual coarticulatory resistance and aggressiveness for consonants and vowels in Catalan. *The Journal of the Acoustical Society of America*, *125*(4), 2288–2298. https://doi.org/10.1121/1.3089222

Redi, L., and Shattuck-Hufnagel, S. (2001). Variation in the realization of glottalization in normal speakers. *J. Phonetics,* 29, 407-429.

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *26*(1), 43–49. https://doi.org/10.1109/TASSP.1978.1163055

Saltzman, E., & Byrd, D. (2000). Task-dynamics of gestural timing: Phase windows and multifrequency rhythms. *Human Movement Science*, *19*(4), 499–526. https://doi.org/10.1016/S0167-9457(00)00030-0

Saltzman, E. L., & Munhall, K. G. (1989). A Dynamical Approach to Gestural Patterning in Speech Production. *Ecological Psychology*, *1*(4), 333–382. https://doi.org/10.1207/s15326969eco0104_2

Schiller, N. O., Mever, A. S., & Levelt, W.J. (1997). The syllabic structure of spoken words: Evidence from the syllabification of intervocalic consonants. *Language and Speech*, *40*(2), 103–140.

Schönbrodt, F. D. & Wagenmakers, E.-J. Bayes factor design analysis: planning for compelling evidence. *Psychon. Bull. Rev.* 25, 128–142 (2018).

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., & Barth, E. (2016). *Recurrent Dropout without Memory Loss* (arXiv:1603.05118). arXiv. http://arxiv.org/abs/1603.05118

Sharma, J., Granmo, O.-C., & Goodwin, M. (2020). *Environment Sound Classification Using Multiple Feature Channels and Attention Based Deep Convolutional Neural Network*. Interspeech 2020, 1186–1190. https://doi.org/10.21437/Interspeech.2020-1303

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* (pp. 197–224). Brill.

Shaw, J. A., Gafos, A. I., Hoole, P., & Zeroual, C. (2011). Dynamic invariance in the phonetic expression of syllable structure: A case study of

Moroccan Arabic consonant clusters. *Phonology*, *28*(3), 455–490. https://doi.org/10.1017/S0952675711000224

Smith, J. L. (2001). Lexical Category and Phonological Contrast. *Workshop on the Lexicon*, 61–72.

Soltau, H., Liao, H., & Sak, H. (2016). *Neural Speech Recognizer: Acoustic-to-Word LSTM Model for Large Vocabulary Speech Recognition* (arXiv:1610.09975). arXiv. http://arxiv.org/abs/1610.09975

Steriade, D. (1999). Alternatives to syllable-based accounts of consonantal phonotactics. *Item Order in Language and Speech*, 205–245.

Stetson, R. H. (1951). *Motor Phonetics:A study of Speech Movements in Action*. North Holland.

Stone, J. V. (2013). *Bayes' rule: A tutorial introduction to Bayesian analysis*. Sebtel press.

Strycharczuk, P., & Kohlberger, M. (2016). Resyllabification Reconsidered: On the Durational Properties of Word-Final /s/ in Spanish. *Laboratory Phonology*, *7*, 1–24. https://doi.org/10.5334/labphon.5

Strycharczuk, P., & Scobbie, J. M. (2017). Fronting of Southern British English high-back vowels in articulation and acoustics. *The Journal of the Acoustical Society of America*, *142*(1), 322–331. https://doi.org/10.1121/1.4991010

Tiffany, W. R. (1980). "The effects of syllable structure on diadochokinetic and reading rates," *J. Speech Hear. Res.* **23**, 894-908.

Tilsen, S. (2017). Exertive modulation of speech and articulatory phasing. *Journal of Phonetics*, *64*, 34–50. https://doi.org/10.1016/j.wocn.2017.03.001

Tilsen, S. (2019). Motoric Mechanisms for the Emergence of Non-local Phonological Patterns. *Frontiers in Psychology*, *10*, 2143. https://doi.org/10.3389/fpsyg.2019.02143

Tilsen, S. (2020). Detecting anticipatory information in speech with signal chopping. *Journal of Phonetics*, *82*, 100996. https://doi.org/10.1016/j.wocn.2020.100996

Tilsen, S. (2022). *An informal logic of feedback-based temporal control*. https://doi.org/10.13140/RG.2.2.14017.28003/1

Tilsen, S., Kim, S. E., & Wang, C. (2021). Localizing category-related information in speech with multi-scale analyses. *PloS one*, *16*(10), e0258178. https://doi.org/10.1371/journal.pone.0258178

Tuller, B., & Kelso, J. A. . S. (1990). Phase transitions in speech production and their perceptual consequences. In M. Jeannerod (Ed.), *Attention and performance* (Vol. 13). Erlbaum.

Tuller, B., & Kelso, J. A. . S. (1991). The production and perception of syllable structure. *Journal of Speech and Hearing Research*, *34*, 501–508.

Uffmann, C. (2007). Intrusive [r] and optimal epenthetic consonants. *Language Sciences*, *29*(2–3), 451–476.

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, 25(3), 169–176.

Wagner, M. (2005). Prosody and Recursion. (Massachusetts Institute of Technology).

Wang, C., Xu, Y., and Zhang, J. (2019). Mandarin and English use different temporal means to mark major prosodic boundaries. in *The 19th International Congress of Phonetic Sciences* (Melbourne, Australia).

Wu, S.-L., Shire, M. L., Greenberg, S., & Morgan, N. (1997). Integrating syllable boundary information into speech recognition. *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, *2*, 987–990. https://doi.org/10.1109/ICASSP.1997.596105

Xu, A., Birkholz, P., & Xu, Y. (2019). *Coarticulation as synchronized dimension-specific sequential target approximation: An articulatory synthesis simulation*. 205–109.

Xu, Y. (1986). Acoustic-phonetic characteristics of junctures in Mandarin Chinese. *Journal of Chinese Linguistics*, *4*, 353–360.

Xu, Y. (2020). *Syllable is a synchronization mechanism that makes human speech possible* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/9v4hr

Xu, Y., & Liu, F. (2006). Tonal alignment, syllable structure and coarticulation: Toward an integrated model. *Italian Journal of Linguistics*, *18*, 125–159.

Xu, Y., & Liu, F. (2007). Determining the temporal interval of segments with the help of F0 contours. *Journal of Phonetics*, *35*(3), 398–420. https://doi.org/10.1016/j.wocn.2006.06.002

Zhang, X., Sun, J., & 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. https://doi.org/10.1371/journal.pone.0085458