

Voice onset time in Assamese-English

A perception experiment and LSTM-based modelling

Supervisor:
Paul Boersma

Second reader:
Titia Benders

Student (15192845):
Arav

Abstract: Listeners of speech are able to adapt to a great variation of acoustic forms that are used for the same underlying sound abstraction. One such adaptation to variation is when listeners hear different sociophonetic varieties (accents). Additionally not all information about a category is neatly present as segments in the acoustic signals, therefore listeners also have to account for sequential cues. This study looks into the perception of varying voice onset times (VOTs) on coronal stops for Assamese-English (AsE) listeners with regard to adaptation to accent differences sequentially carried by following vowel. This is investigated through perception experiment by using an adapted visual analogue scale task intended to elicit responses which target listeners' categorisation more closely than forced-choiced tasks. The experiment results are then modelled on a Long Short Term Memory neural network, which was analysed for temporal predictions, and unit causality. Participant results showed that for the Cru: continuum, AsE listeners on average categorise the same VOT more towards the aspirated category for AsE vowels than for American-English vowels. The LSTM model was found to show higher causality around ~250 ms with unit 38 showing the strongest effect and pushing predictions towards the non-aspirated direction with positive clamps. The findings imply that AsE listeners can adapt coronal stop categorisation to accent variation across categories, and that the LSTM unit activations become more salient prediction drivers before the end of the audio signal.

Bachelor's Thesis, Linguistics

University of Amsterdam

2026-08-13

Contents

1	Introduction	1
1.1	Sequential information is important	1
1.2	Methodological choices	2
1.3	Stop contrast situation	3
2	Methods	5
2.1	Experiment	5
2.1.1	<i>Stimuli</i>	5
2.1.2	<i>Participants</i>	7
2.1.3	<i>Task</i>	7
2.1.4	<i>Analysis</i>	8
2.2	Modelling the experiment	8
2.2.1	<i>Audio preprocessing</i>	9
2.2.2	<i>LSTM unit</i>	10
2.2.3	<i>Training</i>	11
2.2.4	<i>Analysis</i>	12
2.2.4.1	<i>Temporal</i>	12
2.2.4.2	<i>Interventional</i>	12
3	Results	13
3.1	Human participants	13
3.1.1	<i>Response means</i>	13
3.2	Model	16
3.2.1	<i>Trained model</i>	16
3.2.2	<i>Analyses</i>	17
3.2.2.1	<i>Temporal</i>	17
3.2.2.2	<i>Interventional</i>	18
4	Discussion	20
5	Conclusion	21
	References	22
	Appendix	25
A.	Response functions	25
B.	Zoomed C0t Step 1 Stimuli	25
C.	Stimuli	26
D.	Mel-Spectrograms	27
E.	Filtered data summary	28
F.	Participant response plots	29
G.	Loss	30
H.	Temporal Causal Heatmaps	31
I.	Scripts	31

1 Introduction

The process of speech perception in the brain involves both encoding of narrow-band frequency spectro-temporal detail in the primary auditory cortex and higher-order parallel phonetic representations in superior temporal gyrus, suggesting that the brain does not discard lower-level details (Gwilliams et al., 2025). This is supported by studies investigating gradiency in phoneme perception across acoustic continua (Kutlu et al., 2022; McMurray, 2022). These show that perception remains sensitive to changes in within-category acoustic details allowing listeners to adjust to subphonemic variation, recover from ambiguity, and adapt to variation. An example of adaptation to variation is when a listener encounters a sociophonetic profile (accent) distinct from their own or from the one(s) they are used to listening to. As speech is continuous information, it must be processed in sequences and non-linearly transformed into abstractions which lead to meaning. This is interesting for the reason that although encoding of phonological categories can usually be temporally segmented on a phonetic level, part of their details can also be carried by neighbouring sequences. This study looks at a case where adaption to sociophonetic variation for categories (coronal stops) is not purely on the acoustic segment they directly phonetically correspond to (closure, burst, afterburst), but also by those which correspond to the following phonological unit (vowel): **stop categorisation between accent differences for Assamese-English listeners**¹. This is investigated through a perception experiment measuring categorisation through gradient responses, and by modelling the results with neural networks for sequential data processing.

1.1 Sequential information is important

Phonological information encoded in acoustics can be more segmentable for certain categories than others. A stop-to-vowel transition is more apparent than a vowel-to-vowel transition; however, crucial information from the stop is often present on the following vowel in a CV sequence. Formant transitions at the start of the vowel works as a cue for place of articulation, as they encode how the oral cavity shapes as the tongue moves from contact with the palate to the position for the vowel (Ladefoged & Disner, 2012). This underlines why sequential information exists in production and its importance for perception. Sequential information can also be present across longer distances of segments like in gradient metaphony where the target segments close to the harmony-triggering segment are harmonised more than those which are further away (Archangeli & Yip, 2019). Sequential cues can also be the primary cues for contrasting categories. For production, this has been documented for the initial laryngeal contrasts in Korean where the categories lenis and aspirated are realised with reduced differences for voice onset time, and enhanced differences for the f0 and voice quality on the following vowel (Cho et al., 2002; Kim, 2014). This has also been looked at for perception: in Yanbian Korean stops with following creaky voiced vowel where addition of

¹Assamese-English is the non-L1 English variety of L1 Assamese speakers/listeners. The study approaches this population as speaker/listeners of the variety and not merely as non-L1 learners of English, this is to reflect upon the multilingual language ecology of Assam where people acquire English not to communicate with L1 English speakers/listeners but with other non-L1 speakers/listener. And even when they claim to not have any knowledge of English, there is still considerable knowledge of English vocabulary which is present in the L1-Assamese (Nath, 2025).

creak increased listener responses to fortis compared to lenis (Jin, 2025), and for Hindi-Urdu stops with murmured vowel, which increase identification for the voiced-aspirated category even for stimuli without prevoicing or aspiration (Schertz & Khan, 2020).

1.2 Methodological choices

Given that the vowel is also involved in preceding stop perception, categorisation needs to adapt to more variability which arises from having more acoustic features than just the closure, burst and afterburst. This increase in possible variability also means increase in the listener's possible ranges of sensitivity to the variability. Therefore it is important to target categorisation more precisely when investigating the listener's perception by using a task which does not necessarily collapse their graded sensitivity. Following this, the **visual analogue scale (VAS)** task is adapted and used for the participant experiment. The VAS task takes the listener's sensitivity to auditory encoding for categorical identification into consideration, and departs from the classical categorical-perception view (where the listener creates invariance themselves) assumed by alternate-forced-choice (AFC) tasks (Apfelbaum et al., 2022). This is achieved by having trials where instead of discrete response options there is a scale for graded responses between the target categories.

In Figure 1, Apfelbaum et al. (2022) give five examples of representational mapping between auditory perceptual encoding, categorisation and 2AFC responses. (A) is for discrete auditory encoding, which is categorised discretely and results in discrete mapping to 2AFC responses. (B) is where the initial encoding is continuous but this continuity is discarded when mapping into categories; as a result, 2AFC responses are discrete. In (C), both encoding and categori-

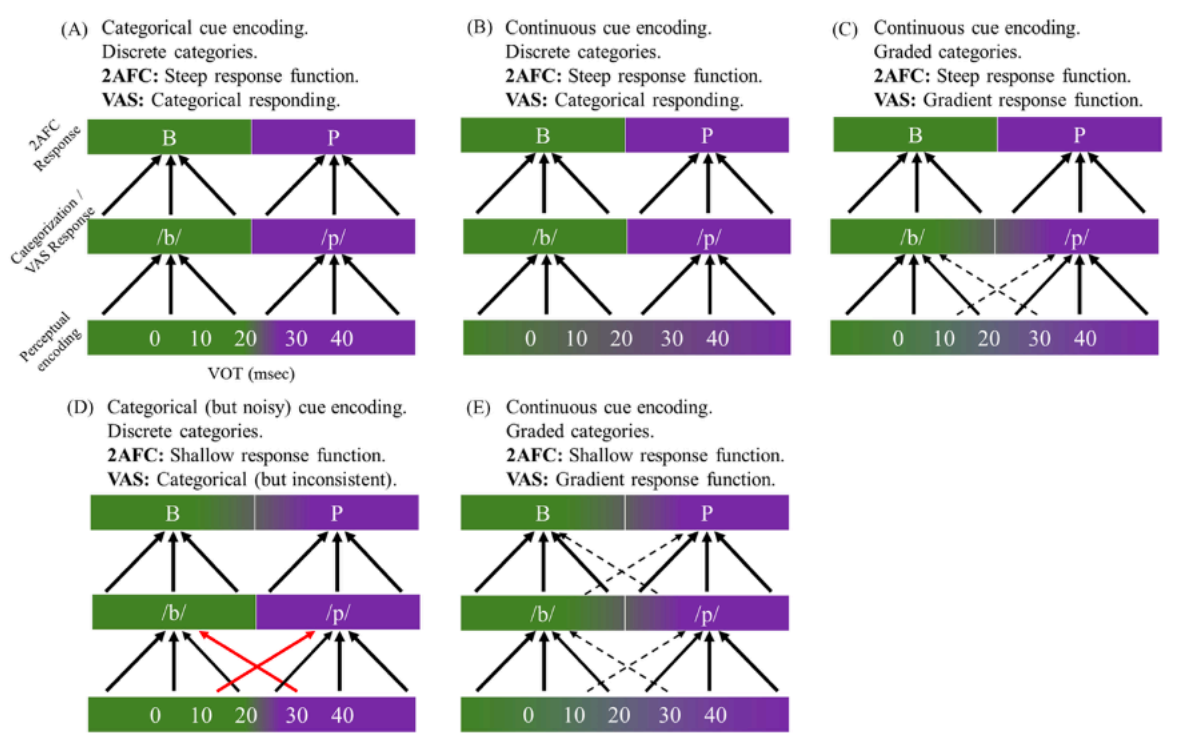


Figure 1: Five possible mappings for speech between the encoding, categorisation, and response stage for the 2AFC identification task (figure sourced from Apfelbaum et al. (2022)).

sation are continuous and this graded mapping is indicated with dotted arrows; however, the 2AFC task having binary options forces response discretisation. (D), like (A), has discrete encoding and categorisation, but with noise for the mapping between the two; these noisy mappings which cause a flip between categories are shown with red arrows. This results in errant responses near the boundary, making it less discrete and appear more graded. (E) shows continuous encoding, categorisation and responses with graded mappings for the case where the listener tries to probabilistically match the continuous nature of the representation onto 2AFC responses. These examples show that for discrete responses, both gradient transitions and step-like transitions (shallow and steep response functions respectively, see Figure 16 in Appendix A.) between response-categories can mean different things for the actual categorisation.

Although the VAS task gives a more precise measure for categorisation, the responses themselves do not tell much about how the auditory signal was processed while the listener takes them as a stream of sequences. One way to understand cognitive processes like perception is by simulating the target phenomena using neural networks. These networks do not have to necessarily be based on the exact neuronal architecture and mechanisms involved in the brain, and still be able to provide valuable insights about the processes, as even these minimalist architectures make use of basic principles of neuronal connectivity/processing such as having layered processing and distributed data representations.

An example of using computer simulations for uncovering speech processes is Boersma et al. (2022), where a Deep Boltzmann Machine based network is trained on auditory and semantic input pairs for modelling bidirectional processing, and shows learnability of emergent and substance-free phonological features by such networks. In the same vein, the present study, although being concerned with only unidirectional processing (perception) of auditory information to categories, tries to find insights about how the acoustic features are integrated from the input stream. An exploratory computational modelling is attempted by using the experiment stimuli responses as training data, and intended as an interpretable proxy for the tested participants' temporally dependent phonetic perception. The **Long Short Term Memory (LSTM)** architecture (which is a type of recurrent neural network with an additional memory dependency adapted for lengthier sequence processing without having to run into the vanishing/exploding gradient issue) is used for modelling the experiment results. The trained model is then analysed using two interpretability techniques (see Section 2.2.4).

1.3 Stop contrast situation

The voice onset time (VOT) of a stop followed by periodic sound is the duration measure for the point in time when periodicity starts relative to the burst. Stop contrasts in languages often use VOT as a cue. VOTs are traditionally classified as i. lead or pre-voicing: periodicity starts before the burst, ii. short-lag or voiceless: periodicity starts right after the burst, and iii. long-lag or aspiration: periodicity starts after the burst with a delay. The phonological voicing for American-English (AmE) has been described to use the contrast for initial stops where /p, t, k/ are realised with long-lag and /b, d, g/ with short-lag and even lead for some speakers and dialects (Davidson, 2016; Lisker & Abramson, 1964). Phonological voicing in (standard) Assamese and Assamese-English (AsE) on the other hand is realised with short-lag for /p, t,

k/ and lead for /b, d, g/² (Wiltshire & Sarmah, 2021). Assamese additionally contrasts stops for phonological aspiration (/p^h, t^h, k^h/) and breathy voice (/b^h, d^h, g^h/) categories, which tend to spirantise (or dentalise in case of /t^h/) finally, and are used for English loanwords with fricatives /f, v, θ/. Dutta & Kenstowicz (2018) give examples:

- f → p^h : telephone [telip^hun], cough [kɔf]
- v → b^h : vote [b^hut], love [lav]
- θ → t^h : thinker [t^hinjkar], eighth [eit]³.

Like these loanwords in Assamese, and other non-South Indian Englishes, the aspirated coronal plosive is also present in AsE and it takes the place of /θ/; however, like AmE coronal plosives the place of articulation is initially alveolar instead of dental (Gargesh, 2008; Wiltshire, 2020)⁴. This language situation results in a mismatch between the VOT-defined phonetic realisations and their phonological representations for the initial coronal stop systems across the two English varieties, where long-lag represents /t^h/ for AsE but /t/ for AmE, and short-lag represents /t/ for AsE but /d/ for AmE (Figure 2).

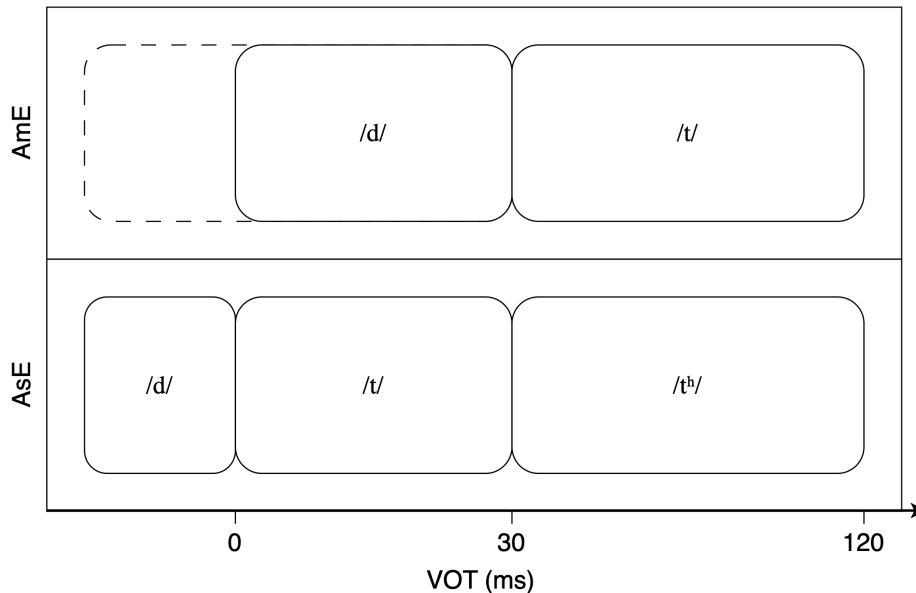


Figure 2: Schematic VOT-defined stop contrasts in AmE and AsE, roughly based on range values from (Lisker & Abramson, 1964) for AmE /d/ and /t/. AsE boundaries are overlaid as per traditionally described VOT: (pre)voiced, voiceless, aspirated along with approximate upper boundaries for /t/ and /t^h/ from (Nath et al., 2014).

²Note about crosslinguistic sameness of phonological voicing between the varieties: one possible explanation Wiltshire & Sarmah (2021) mention for non-aspirated /t/ is to do with founder varieties, where English varieties like Scottish English brought to India had true voicing contrast (lead for voiced, short-lag for voiceless)

³The phonetic representations [b^h] and [r] in the examples are to be read with room for interpretation. For Standard Assamese the described realisation of the first is with lag VOT but not lead VOT in initial position (Nath et al., 2014), and for the second is as an alveolar approximant with optional deletion resulting in preceding vowel lengthening non-initially (Mahanta, 2012).

⁴These studies however have not directly looked at AsE /t^h/ for AmE /θ/.

For an extreme effect where the accent information from the vowels discretely shifts the categorisation, the same consonant prototype would be heard as different phonemes (Table 1). Here the phonetic definition for phonological voicing drastically changes percept strategies from using a prevoicing-based definition (same as Assamese) to an aspirated-based definition (same as AmE). For an accent effect to happen, the category (re-)adaption⁵ is expected to approach a system like this in some capacity.

Table 1: Responses for an extreme accent driven VOT contrast shift. (Re-)adaptation of categories for the investigated accent effect is expected to show up in similar fashion.

Prototypes	Perceived Phonemes	
	When hearing AsE	When hearing AmE
[t] (short-lag VOT)	/t/	/d/
[t ^h] (long-lag VOT)	/t ^h /	/t/
Contrast type used for /t/ perception:	Prevoicing	Aspiration

Now that the necessary background is established, in this study, first the question of whether AsE listeners show an accent effect for coronal stops perception when listening to AsE and AmE will be addressed; then the results are modelled using an LSTM, and the trained model is investigated for how the sounds get sequential processed and arrive to responses.

2 Methods

2.1 Experiment

2.1.1 Stimuli

The stimuli comprised of sound files targeting two VOT continua based on English (specifically AsE) minimal pairs between /d, t, t^h/:

1. **Cot** – /d/ot ~ /t/aught ~ /t^h/ought
2. **Cru**: – /d/rew ~ /t/rue ~ /t^h/rough

Similar to neighbouring Tibeto-Burman Indian Englishes, AsE merges the (without cot-caught merger) AmE corresponding /ɔ/ and /ɑ/ to /ɔ/, therefore making *dot* rhyme with *taught* and *thought* (Sarmah & Wiltshire, 2024). Another peculiarity is with the onset with rhotic in the second continuum. AmE typically devoices the /r/ in this position when following voiceless stops, and although Wiltshire & Sarmah (2021) did not find the same for AsE (a reasoning would be that since AsE voiceless stops are not aspirated, there is not need to overlay the long-lag quality over the rhotic), the AsE speaker recording for *through* had a devoiced rhotic. Additionally high density of <T> responses for the Cru: continua (Appendix F.) gives a contrary observation to Wiltshire & Sarmah (2021)’s findings.

⁵Re-adaption assumes one of them is default, AsE being that would be fair to say as participants are AsE listeners, however in the context of the experiment (covered in Section 2.1.3) where they are listening to the varieties in blocks, the contextual default maybe altered for participants depending on the order of their blocks.

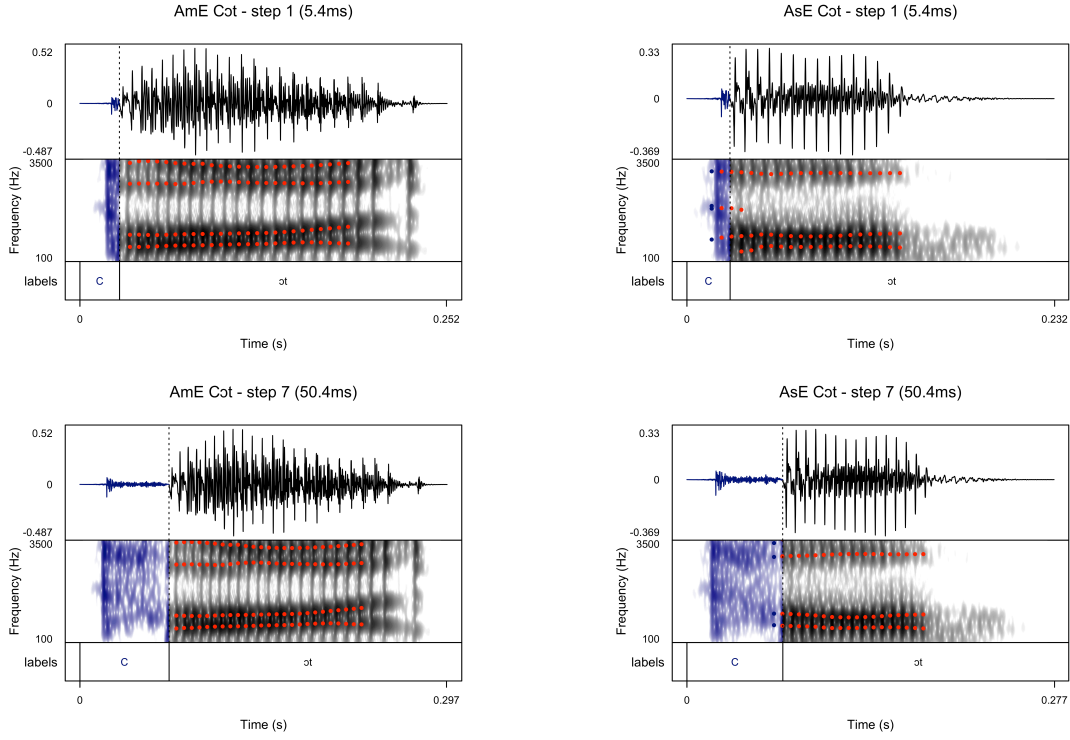


Figure 3: Waves and spectrograms for Step 1 (top) and 7 (bottom) in Cɔt continuum: AmE (left), AsE (right). Onsets are highlighted blue, and formant contours shown with red dots.⁶

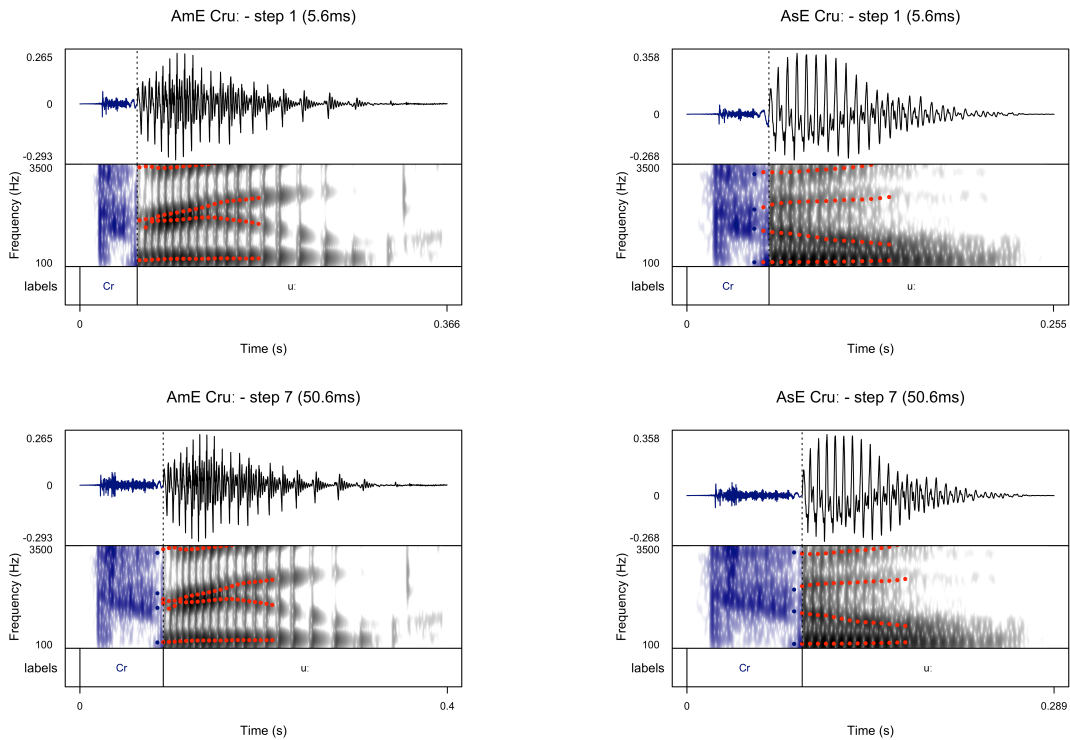


Figure 4: Waves and spectrograms for Step 1 (top) and 7 (bottom) in Cru: continuum: AmE (left), AsE (right). Onsets are highlighted blue, and formant contours shown with red dots.

⁶All highlighted stimuli pictures were made using the praatpicture R package (Puggaard-Rode, 2024).

Stimulus creation: First, the recordings for the words *true* and *taught* were clipped from two uttered sentences by a male AmE speaker from the CMU-Arctic Database (Kominick & Black, 2004). The same sentences but target words replaced with *through* and *thought* were uttered by a male AsE speaker, and the target words were clipped. The recording sample rate was set at 16 kHz (mono) to match the files from CMU-Arctic Database. These source sound files were then further clipped to separate the onsets and rhymes. The onsets of the AsE source files were manipulated for their duration (VOT steps) and then concatenated with the AsE and AmE rhymes recoverably using the a praatscript (see Appendix I.) which performs the manipulations and creates seven equally distanced steps. Onset duration manipulation was done in Praat (Boersma et al., 2026) using the DurationTier in the Manipulation object from the end of the burst till the end of aspiration (for Cɔt) or pre-rhotic aspiration (for Cru:)⁷. The range of the steps were decided (through auditory judgements by the author) so the endpoint steps of the continuum clearly sound as different categories (/t, t^h/) and the steps in between being more ambiguous. This turned out with Cɔt steps ranging from VOTs 5.4 ms till 50.4 ms, and Cru: steps ranging from (pre-rhotic) VOTs. 5.5 ms till 50.5 ms. The praatscript was run four times for the two accents × two continua. Table 2 lists the VOTs for the seven steps (total number of stimuli = 28). Figure 3 and Figure 4 show pictures of the manipulated stimuli for steps 1 and 7 (find all stimuli pictures in Appendix C.).

Table 2: VOT (ms) steps for manipulated stimuli

Steps	1	2	3	4	5	6	7
VOT ±0.1 (ms)	5.5	13.0	20.5	28.0	35.5	43.0	50.5

2.1.2 Participants

12 adult AsE listeners (L1: Assamese, non-L1: English) participated in the perception task. Participants were asked to report their Assamese language background in terms of geographical region; this was included to check if participants with Lower Assamese background showed response patterns towards the <D> side of the scale for lower VOT steps as suggested by Nath et al. (2014)’s finding of voiced stops without prevoicing in Nalbaria (a Lower Assamese variety). Eight participants had Lower Assamese background. As visual inspections on participant means do not show any clear region-based patterns, and as group sizes are unequal, the study does not proceed with region-based analyses and interpretations.

2.1.3 Task

Ratings for stimuli were collected using an adaptation of the Visual Analog Scaling (VAS) task (Apfelbaum et al., 2022; Kutlu et al., 2022) designed on an ExperimentMFC object in Praat (Boersma et al., 2026). The task was accessed by participants remotely from their computer/laptop using Zoom and they were able to control the host’s mouse. The experiment file was operated from the host’s laptop. The trials played a randomly selected stimulus on a blank screen, followed by a screen with three minimal pair text boxes (left, center and right) and a

⁷There is not easy metric to determine where the devoiced rhotic starts during the aspiration, therefore auditory judgements for a boundary was made, ie. sound was played from different starting positions until onset stops sounding like /^hr/ or /t^hr/ and starts sounding like /r/. The duration for the rhotic aspiration was 29.1 ms, therefore all Cru: manipulated onsets are 29.1 ms longer than Cɔt manipulated onsets.

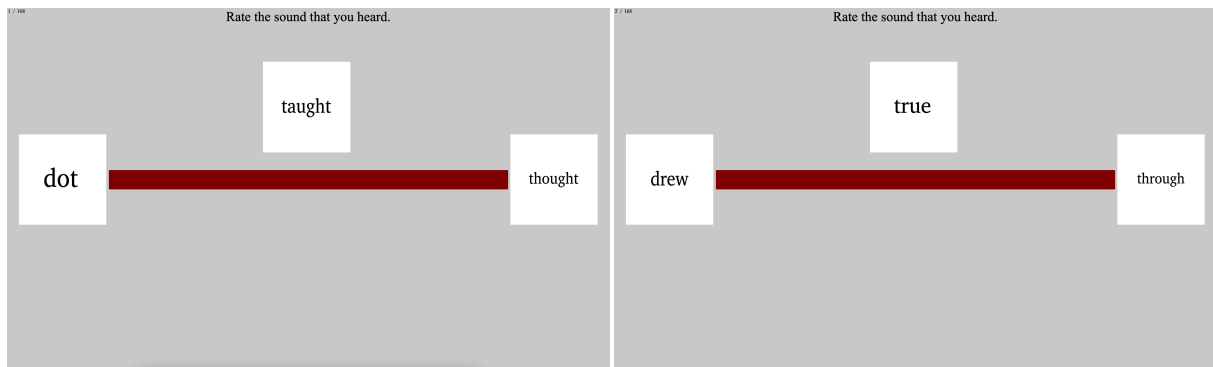


Figure 5: Experiment trial screen for Cɔt (left) and Cru: (right) continua.

scale stretching between the left and right boxes Figure 5. The participants clicked on a region of the scale to indicate where on the continua they perceive the token; the trials proceeded only after clicking a response. One complete test consisted of 336 trials (twelve trials per unique stimulus) grouped in two blocks for accent type (the order of the accents was randomised across participants)⁸ and participants took about 10 to 15 minutes to complete it (excluding time for reading consent forms and going through instructions). Responses on the scale range from -1.000 to 1.000 (left to right), and minimal pairs were ordered with increasing VOT (left to right: <D> to <T> to <TH>).

2.1.4 Analysis

One participant was excluded as they showed noticeable lack of attention during the task. Responses were analysed for a total of 11 participants. Inattentive responses with reaction times less than 0.5 ms (approximately the length of longest stimulus) were filtered out. A total of **3671** stimulus-response pairs were used. Paired t tests for overall, and per-continua stimuli were done to compare response mean (dependent variable) differences between the two English varieties/accents: AsE, AmE (independent variable), and Cohen's d was used to get the direction and magnitude of the effects. Analysis and graph creation was carried out in R (version 4.6.0) (R Core Team, 2026) using the tidyverse (Wickham et al., 2019) and ggpubr (Kassambara, 2026) packages (find code in Appendix I).

2.2 Modelling the experiment

This section covers the creation of dataset: for the stimuli to have approximated auditory representation and be inputable to the network, the LSTM architecture: how it takes the sequential inputs and processes them to outputs, training: how the model was trained and validated, and analyses: temporal and interventional⁹. All processes covered in this section were carried out in Python (version 3.14.4) (Python Software Foundation, 2026) using the packages Torchaudio (Hwang et al., 2023) and Librosa (McFee et al., 2026) for audio extraction, PyTorch (Paszke et al., 2019) and PyTorch Lightning (Falcon & The PyTorch Lightning team,

⁸Pilot testing with a non-participant AsE listener without accent blocks where stimuli across accents were randomised together showed confounded responses, which could be because hearing the 336 trials with randomised order for the accents creates a neuralising effect between the two, or as the stimuli are very short in length, preceding trials effect successive ones and therefore the responses.

⁹See Section 2.2.4.2 for explanation of intervention used in this study.

2024) for model development, NumPy (Harris et al., 2020) and Pandas (The pandas development team, 2024) for data handling, and Matplotlib (Hunter, 2007) and Seaborn (Waskom, 2021) for visualisations.

2.2.1 Audio preprocessing

The filtered stimuli-responses from the experiment were extracted as a CSV file; this was used to access the sound files by their filenames. The manipulated stimuli were saved as WAV files from Praat with every sample stored as a 16-bit integer by default. When loaded with Librosa and PyTorch, these were converted to 32-bit floats for the ease of operations which are implemented in the next steps (mel-spectrogram transformation, band amplitude normalisation and forward pass) requiring continuous numbers; all parameters are stored as 32-bit floats. The stimuli were also fixed to the length of **6394 samples** ($\frac{6394}{16000} \approx 0.4$ s) which is the length of the longest stimulus (AmE Cru: Step 7), and files of lesser length were zero-padded at the right edge.

Humans perceive frequencies logarithmically and as the speech perception processes involve encoding of time-frequency representation, the digital sound files were converted to mel spectrogram to approximate auditory perception. Mel spectrograms use overlapping triangular filters called mel filters which scale frequencies logarithmically so that differences in higher frequencies are perceptually the same as differences in lower frequencies; this is unlike frequency represented in standard spectrograms where higher-frequency differences are perceptually less distanced than lower-frequency differences. Mel spectrograms for stimuli were computed for **128 mel filterbank bands**, on **time windows of sample length 320** (20 ms) with a 1024-samples fast Fourier transform (FFT) (FFT of sample length 1024 instead of 320 increases the frequency resolution), and **sample hop length of 44** (2.75 ms), resulting in a total of **146** ($\lceil \frac{6394}{44} \rceil$) **time windows**. Time window amplitudes were transformed using the Hamming function (raised cosine) to minimise spectral leakage caused by discontinuities at

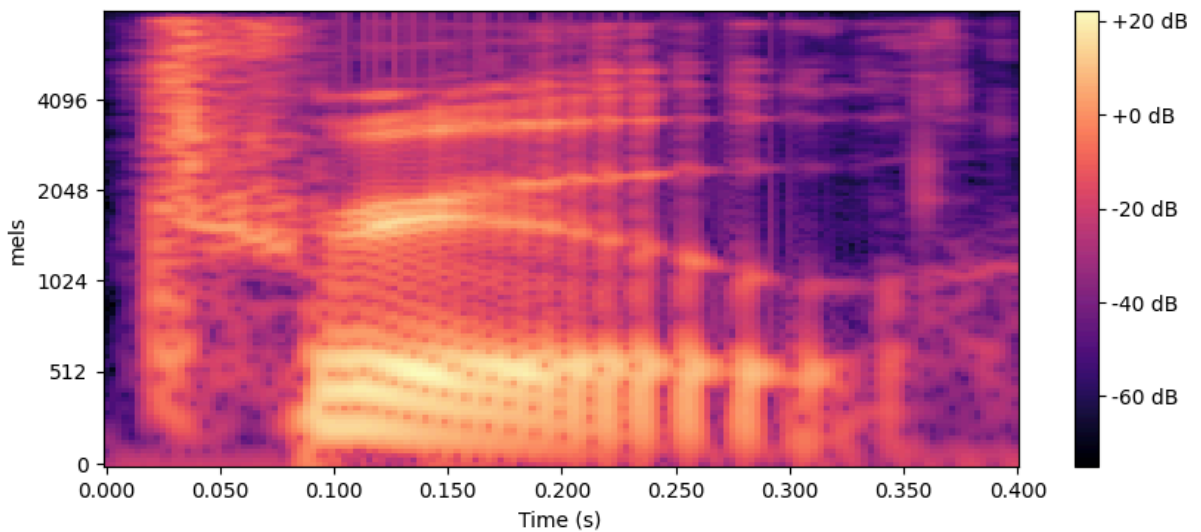


Figure 6: Extracted mel spectrogram (before Z-score normalisation) for AmE Cru: Step-7 with band power values converted to decibels (dB).

the window edges¹⁰.

After the transformation, the band power values were converted to decibels, which is a logarithmic unit for sound power and is a perceptually relevant representation for loudness (for humans). Figure 6 shows the extracted mel spectrogram with power scaled to decibels for the AmE Cru: Step 7 stimulus. Independently for each band across all stimuli, the mel-band decibel values were Z-score normalised (mean set to 0 and standard deviation to 1), so that each band equally contributes as input for the model¹¹.

2.2.2 LSTM unit

The LSTM unit consists of two types of memory states, and uses 64 dimensions (ie. 64 units) for these states in the model: **hidden state** $h_t \in \mathbb{R}^{64 \times 1}$ (short-term) and **cell state** $c_t \in \mathbb{R}^{64 \times 1}$ (long-term). The three gates: forget, input, and output, and the candidate cell state connect the input x to the unit and update memories.

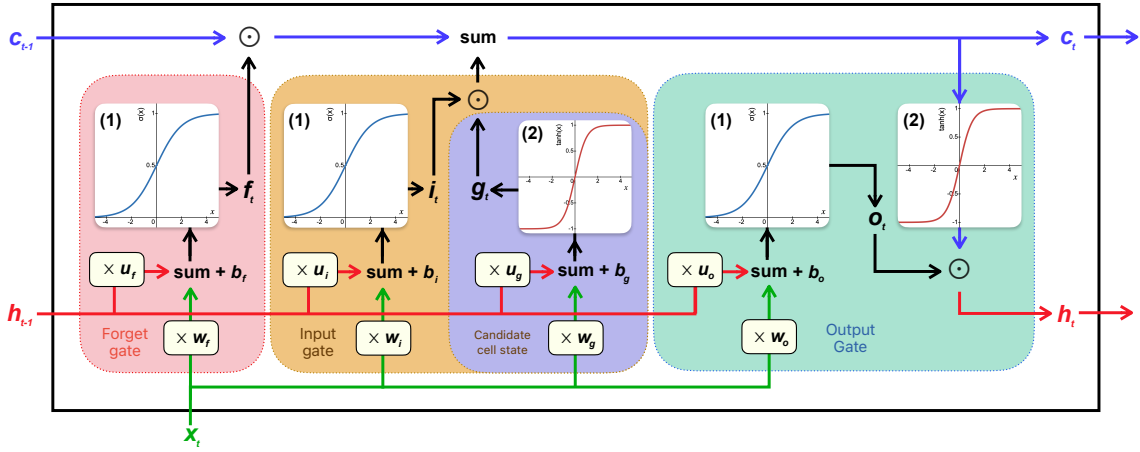


Figure 7: Visual representation of the LSTM unit

Two activation functions are used in different operations in the unit: (i) the **logistic** or **sigmoid function** given in (1), which takes a value and outputs a number between 0 and 1, and (ii) the **hyperbolic tangent function** given in (2), which takes a value and outputs a number between -1 and 1.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (1)$$

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (2)$$

The extracted mel spectrogram is treated as sequences of $T = 146$ time frames and values

¹⁰These exist naturally and are only discontinuities because the FFT expects the window to loop onto itself, i.e. start and endpoints of the window with the same amplitude value.

¹¹This was not a naturalistic choice, but it was adapted as a practical measure for training the model as loss did not improve and model was not able to predict ratings when it was trained without mel-band normalisation and with global min-max mel-band normalisation.

for each frame ($\mathbb{R}^{128 \times 1}$; mel bands dimensions) is passed as input: $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T$. The LSTM (single layer) takes the inputs one sequence at a time and updates the internal states $\mathbf{h}$ and $\mathbf{c}$. First the weights $\mathbf{w}_* \in \mathbb{R}^{64 \times 128}$ (where $*$ is a placeholder for $\mathbf{f}, \mathbf{i}, \mathbf{g}$, or $\mathbf{o}$) are multiplied with the current input $\mathbf{x}_t$, and the weights $\mathbf{u}_* \in \mathbb{R}^{64 \times 64}$ with the previous hidden state $\mathbf{h}_{t-1}$ and then both are summed together along with the bias vectors $\mathbf{b}_* \in \mathbb{R}^{64 \times 1}$. As given in (3), this pre-activation is computed independently with unique weight vectors for each of the three gates (forget, input and output) and the one candidate cell state shown in Figure 7. Then the output is passed through (1) for three gates: $\mathbf{f}_t, \mathbf{i}_t, \mathbf{o}_t$ and through (2) for the candidate cell state $\mathbf{g}_t$, these operations are defined as (4).

$$X_* = w_* \mathbf{x}_t + u_* \mathbf{h}_{t-1} + b_* \quad (3)$$

$$\mathbf{f}_t = \sigma(X_f) \quad (4)$$

$$\mathbf{i}_t = \sigma(X_i)$$

$$\mathbf{o}_t = \sigma(X_o)$$

$$\mathbf{g}_t = \tanh(X_g)$$

The gates can now be defined according to their role for updating the memories. For the long-term memory, the forget gate $\mathbf{f}_t$ controls the percentage of previous long-term memory $\mathbf{c}_{t-1}$ to be kept. The input gate $\mathbf{i}_t$ controls the percentage of the candidate cell state $\mathbf{g}_t$ to be added. The current cell state $\mathbf{c}_t$ then is calculated as (5), where $\odot$ denotes element-wise multiplication.

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t \quad (5)$$

The output $\mathbf{o}_t$ controls the percentage of the the current cell state $\mathbf{c}_t$ passed through (2) that is to be kept for determining the current short-term memory. The hidden state $\mathbf{h}_t$ calculation is given in (6).

$$\mathbf{h}_t = \tanh(\mathbf{c}_t) \odot \mathbf{o}_t \quad (6)$$

Finally, when all sequences are passed through the unit, the end values of the hidden state $\mathbf{h}_T \in \mathbb{R}^{64 \times 1}$ are passed through a fully connected layer with weight $\mathbf{w}_{fc} \in \mathbb{R}^{64 \times 1}$ and scalar bias $\mathbf{b}_{fc}$ to produce the final predicted rating $\hat{\mathbf{y}}$, this is defined in (7).

$$\hat{\mathbf{y}} = w_{fc} \cdot \mathbf{h}_T + b_{fc} \quad (7)$$

2.2.3 Training

The complete dataset was randomly split into a training set (80%) and a validation set (20%). The training set was used to update model weights and the validation set was used by the learning-rate scheduler to monitor improvements in model performance.

The predicted ratings $\hat{\mathbf{y}}$ were obtained after the LSTM passes through a stimulus from the stimulus-rating pair trials; the squared difference between actual ratings $\mathbf{y}$ and $\hat{\mathbf{y}}$ were aver-

aged across the batch (16 trials). This gave the Mean Squared Error loss, which was minimised by finding gradients of the loss function of all trainable parameters using backpropagation through time and by updating the parameters to move in the opposite direction of the gradient. This was handled by the Adam optimiser (Kingma & Ba, 2017) with an initial learning rate of 0.01. Adam maintains the momentum of the gradient descent (by keeping the moving averages of past gradients) and exponential dampening of descent oscillations (by scaling the updates using the variance of the past gradients). This makes the optimisation less susceptible to slowing down on plateaus and less likely to shooting past minima. The optimisation process also made use of the ReduceOnPlateau scheduler, which checks validation improvements after every epoch and reduces learning rate by half if validation remains plateaued for more than 10 successive epochs.

2.2.4 Analysis

Temporal

To examine how the trained model processes auditory information across successive time frames to commit to featural or categorical decisions, the fully-connected layer is run for each frame after the hidden states are updated, giving the prediction trajectory over 146 frames. This modified structure of the trained model was passed with all unique step 1, 4 and 7 stimuli.

Interventional

For this study, an intervention on the model is consists of taking the hidden state values at a certain timeframe, up until which the LSTM processes sequences naturally, then performing manipulations on it after which the LSTM is allowed to continue processing the remaining sequences; this is done to examine how the model behaviour changes from its natural one after running said manipulation. In this study the performed manipulation is manually setting (clamping) the natural hidden state values to specific alternate values (clamps). An intervention method inspired by CDEV (causal disentanglement with extreme values) (Beguš et al., 2023), specifically the method for clamping units and measuring treatment effects, is applied here. The core idea for the adapted method is to intervene a hidden state $\mathbf{h}_t$ of a specific unit at a time during sequence processing by clamping the state to a value $\mathbf{v}$ different from its natural one $\mathbf{v}_0$, on a range beyond what the LSTM is used to (-1 to +1) for its hidden states. The effect of the intervention is measured by finding the difference between the new prediction and the natural one. This is done for all N unique stimuli and the difference is averaged. This operation: average treatment effect (ATE), is formulised in (8).

$$\hat{\tau}(\mathbf{v}) := \frac{1}{N} \sum_{i=1}^N \hat{y}_i(\mathbf{v}) - \frac{1}{N} \sum_{i=1}^N \hat{y}_i(\mathbf{v}_0) \quad (8)$$

This procedure is looped over for a set of clamped values $\mathbf{v}$ between -10 to $+10^{12}$, and then this in turn is looped over for all 64 units (unit 0 to unit 63), therefore an ATE curve is produced for each unit in a time frame. Before zooming into one time frame, first a scan across the whole

¹²with 11 equally spaced points for the heatmaps and 31 equally spaced points for the curves

sequence is made by by calculating the absolute ranges of the ATE curves for every fifth frame. The first (0) and last (145) frames were excluded, as the purpose of this method for this study is to quantify interventional effects while the audio is being processed. This results in a (64, 29) matrix which is visualised as a heatmap.

3 Results

3.1 Human participants

3.1.1 Response means

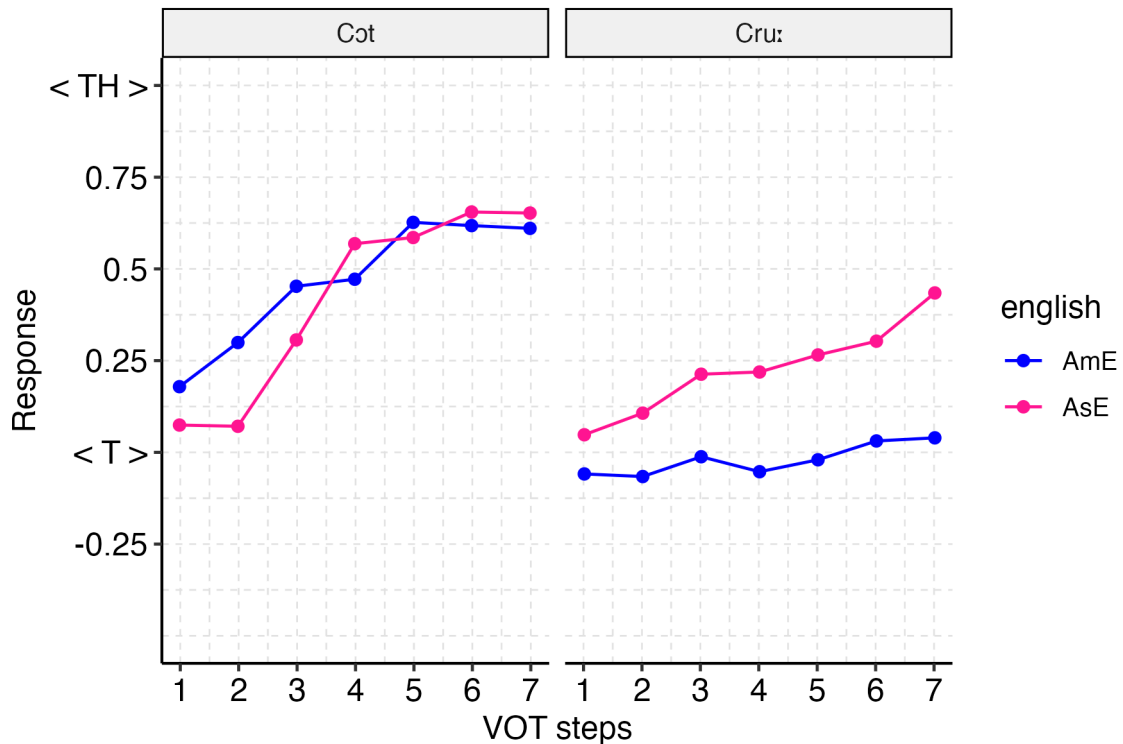


Figure 8: Mean participant response (y-axis) for continua Cɔt and Cru: across the English varieties (blue = AmE, red = AsE), and VOT steps (x-axis)

Figure 8 shows the mean participant responses for the two continua, by steps and English variety. Overall participants had a mean AmE response of 0.22 with a standard deviation of 0.13 and a mean AsE response of 0.32 with standard deviation of 0.14; the difference of the two means was -0.10 ($SD = 0.17$). For the Cɔt continuum, the mean AmE response was 0.46 with SD of 0.17, and the mean AsE response was 0.42 with SD of 0.19 ($\Delta M = 0.05$, $SD = 0.15$). For the Cru: continuum, AmE showed a mean response of -0.02 with SD of 0.15, and AsE showed a mean response of 0.23 with SD of 0.19 ($\Delta M = -0.25$, $SD = 0.26$). The paired sample t-test for the Cru: continuum points to a significant difference between the responses for the two varieties, $t(10) = -3.12$, $p = .011$ (which is smaller than the Bonferroni-corrected criterion of .0167), Cohen's $d = -0.94$, and the 95% Confidence Interval ranged from -0.42 to -0.07 . For the other two inferences, the response differences between AmE and AsE were not found significant. The descriptive and inferential results are summarised in Table 3. With these results, the study

rejects the null-hypothesis that for continuum Cru: AsE listeners do not show differences in categorisation of coronal stops when hearing the same VOT for AsE and AmE. For the continuum Cɔt and the overall grouping, the study fails to reject this null-hypothesis.

Table 3: Statistical results

continuum	AmE	AsE	AmE-AsE	95% Confidence Interval		<i>t</i> (10)	<i>p</i>	Cohen's <i>d</i>
	<i>M</i> (<i>SD</i>)	<i>M</i> (<i>SD</i>)	<i>M</i> (<i>SD</i>)	lower	upper			
grouped	0.22 (0.13)	0.32 (0.14)	-0.10 (0.17)	-0.21	0.01	-1.95	.079	-0.59
Cɔt	0.46 (0.17)	0.42 (0.19)	0.05 (0.15)	-0.05	0.15	1.06	.315	0.32
Cru:	-0.02 (0.15)	0.23 (0.19)	-0.25 (0.26)	-0.42	-0.07	-3.12	.011	-0.94

Interpretation: The slope for lines for Cɔt in Figure 8 seems to be positive, which shows that with higher VOT the responses shift towards the <TH> side of the scale, i.e. categorisation is towards /t^h/. This is not as strongly observed (or very negligibly for AmE) for the lines in the Cru: continuum. Both continua had the lower-end mean responses for AmE close to <T> in the scale, instead of <D> suggested in Table 1. This indicates that adaptation of categories is stronger for one side of the scale (T ↔ TH) than the other (D ↔ T) (however, see participants NL1, NL2 in Figure 9, participant L1 and L2 in Figure 10, and response density for AmE in Appendix F.).

For all steps in Cru: the mean responses for AmE were higher than AsE, suggesting that after the onset, the listeners adapt their VOT-based stop categorisation by hearing the difference in accent from the vowel. For higher VOTs, this adaptation seems to be stronger. The direction of adaptation is consistent with Table 1. The same effects of variety difference were not observed for the Cɔt continuum; instead, for lower VOTs the lines suggest adaptation in the opposite direction, which goes against Table 1 (see Section 4 for further discussion on this).

Grouping the means by participants show existence of individual variability (particularly for the Cɔt continuum) in response patterns (Figure 9, Figure 10). It can be seen that for certain participants the response mean lines are flat in Figure 10, suggesting that the distance between VOT steps were not strong enough for them to show difference in categorisation.

Figure 11 visualises the differences between the two varieties for each step, across participants as a heatmap (find the lineplot version of this in Appendix F.). This captures the adaptation directionality difference between the two continua, where Cɔt appears to have higher means for AmE than AsE, however with smaller effect size (Cohen's *d* = 0.32). Taking this at face value may suggest that Cɔt follows a different category adaptation strategy to Cru:, but this might have also resulted due to low saliency of the vowel (or both vowel and coda¹³) differences

¹³As AmE also allows for final stops to have an audible release, however this is less observed for Assamese (where coda stops are usually unreleased), and by possibly exten AsE.

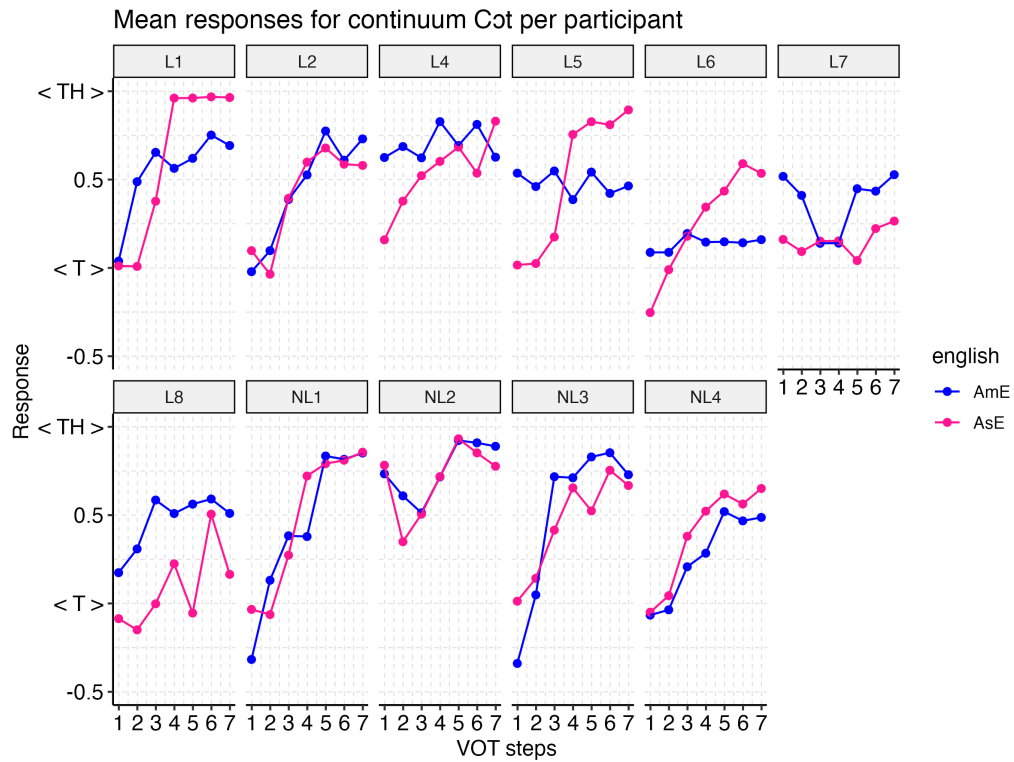


Figure 9: Mean participant response by participants for Cot continuum across the English varieties (blue = AmE, red = AsE), and VOT steps (x-axis).

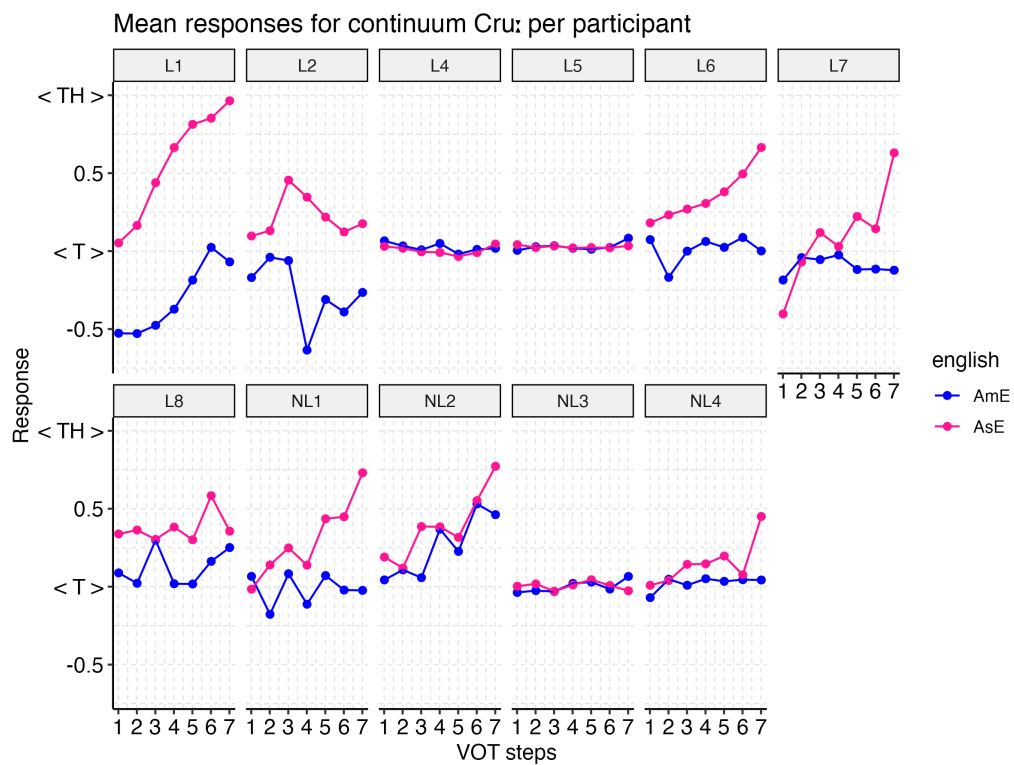


Figure 10: Mean participant response by participants for Cru: continuum across the English varieties (blue = AmE, red = AsE), and VOT steps (x-axis).

between the two varieties (compare formant contours in Figure 3). Another possibility is differences in burst energies relative to vowel onset; this is discussed in Section 4.

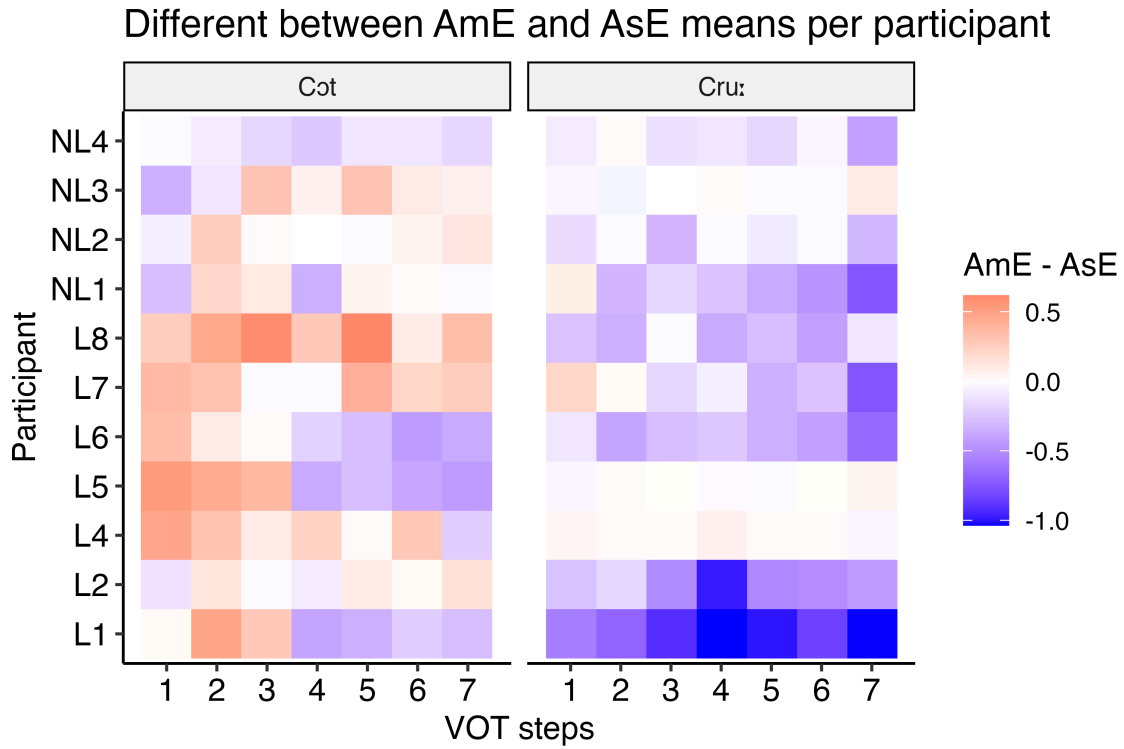


Figure 11: Difference of means between AmE and AsE across VOT steps by participants. Blue = AsE mean > AmE mean, red = AsE mean < AmE mean

3.2 Model

3.2.1 Trained model

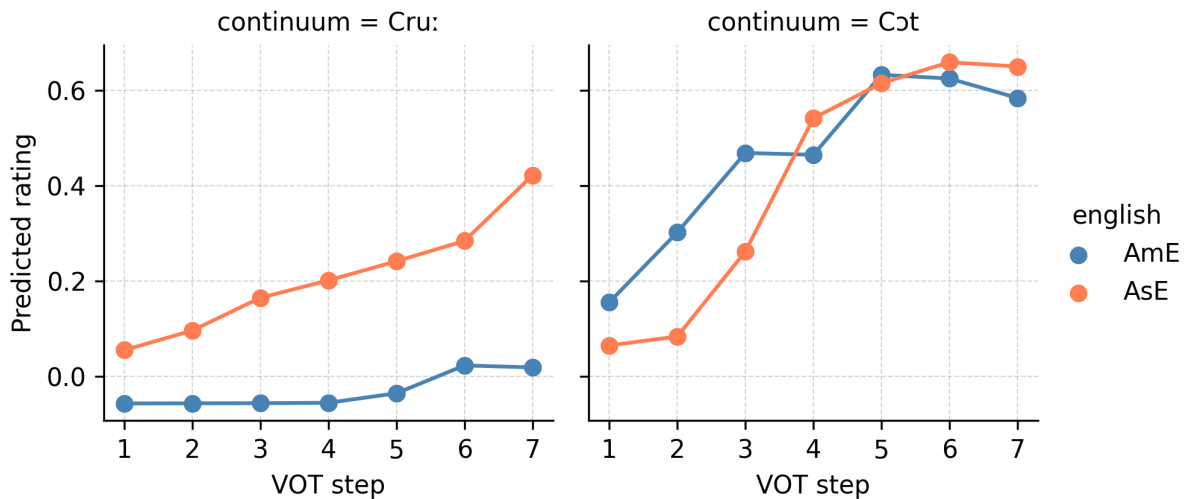


Figure 12: Predicted ratings by the trained model for each unique stimuli in Cru: and Cot continua. Stimuli are grouped by their VOT steps (x-axis) and English variety (colour).

The model was trained for 200 epochs (one epoch corresponds to $\lceil \frac{\text{training samples}}{\text{batch trials}} \rceil$) over four training checkpoints giving a final training loss of 0.1662 and final validation loss of 0.1779 (less than the variance of the participant responses, 0.235, which indicates a good fit over the mean responses). After around epoch 75 the loss dropped only slightly and remained plateaued for the rest of the training (find loss plots in Appendix G.). The best validation was measured by the end of the second checkpoint (50 epochs) with a loss of 0.1668, but the 200 epochs version was chosen for the analyses as its predictions patterned with the experiment response mean lines (Figure 12).

3.2.2 Analyses

Temporal

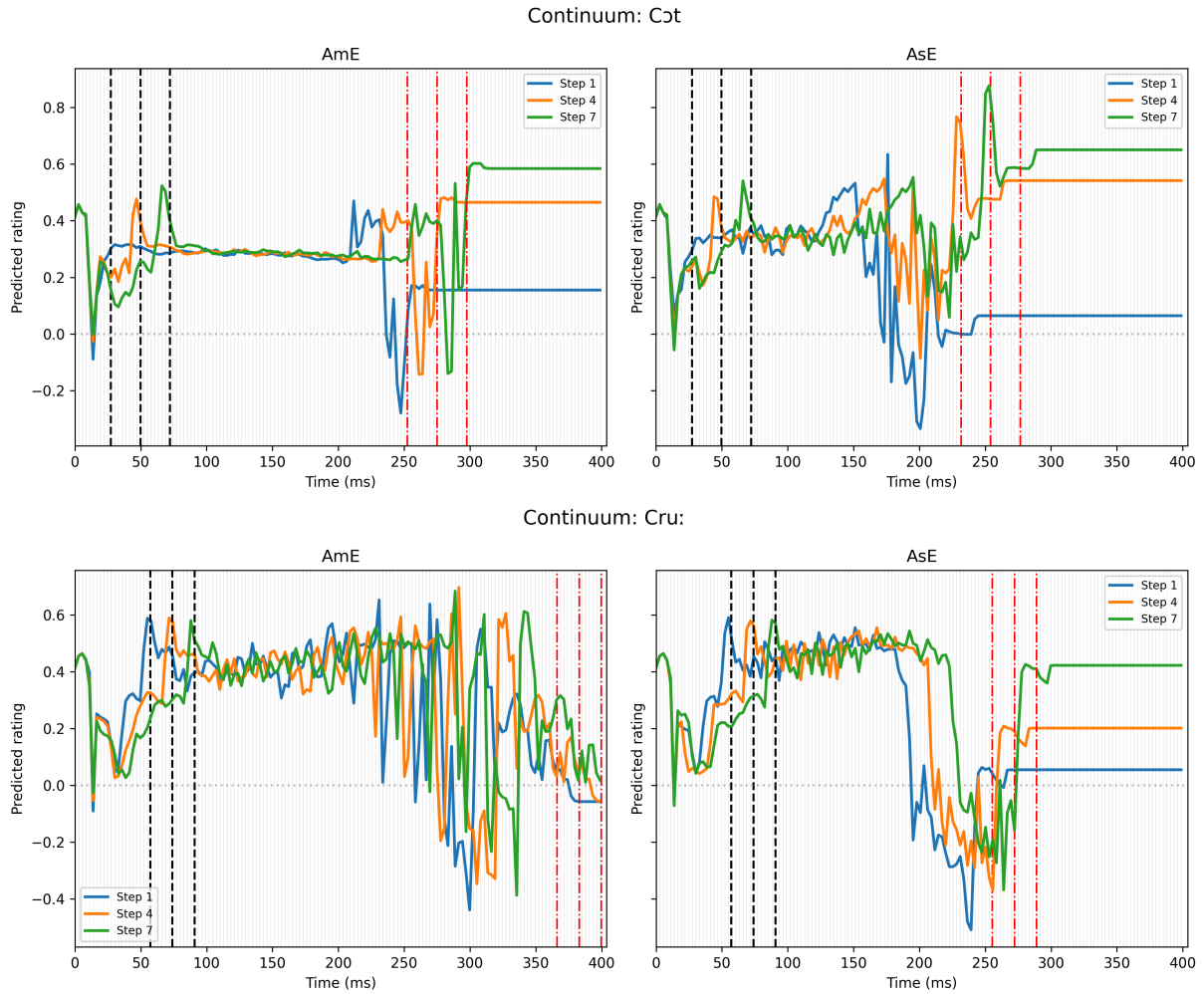


Figure 13: Model prediction trajectories across all time frames for steps 1, 4 and 7 for the two continua. Black dashed lines show the VOT endpoints, and red dashed lines show the end of audio (before padding) for the three steps.

Figure 13 visualises the outputs from the temporal analysis. For all four plots, similar trajectory patterns can be seen on the onset: for the burst and till the end of aspiration. For the three steps the pattern with the burst coincides, but the rest of the curve is phased ahead as step VOT increases, suggesting that the model tracks when the relevant pointers in the aspiration (which were manipulated for duration by speeding up or slowing down) show up. The trajectories

until the end of aspiration are similar within continua¹⁴, and show bigger changes once the vowel starts. This is expected as the onsets within continua comes from the same source.

It can also be noticed that by the end of aspiration, continuum C_{ot} steps seem to have different values (especially between step 1 and the rest), while this is not seen for continuum C_{ru}. This may suggest that the model learns to differentiate between VOTs better for the continuum with larger mean response differences between the steps (larger slope); however, a better quantitative measure is needed to answer if that is the case for this model.

The predictions during the vowel have different patterns for each of the continua-accent intersections. Although the oscillatory nature of these make it hard to interpret what they tell about vowel information encoding of the model, it can be seen that prediction decisions appear to be finalised right before the end of the audio (and flatlines for the padded region). And because three of the intersections' audio end at similar time frames (around 250-300 ms), changes in this region appear to be more salient drivers for final prediction; this is also found for the interventional analysis.

Interventional

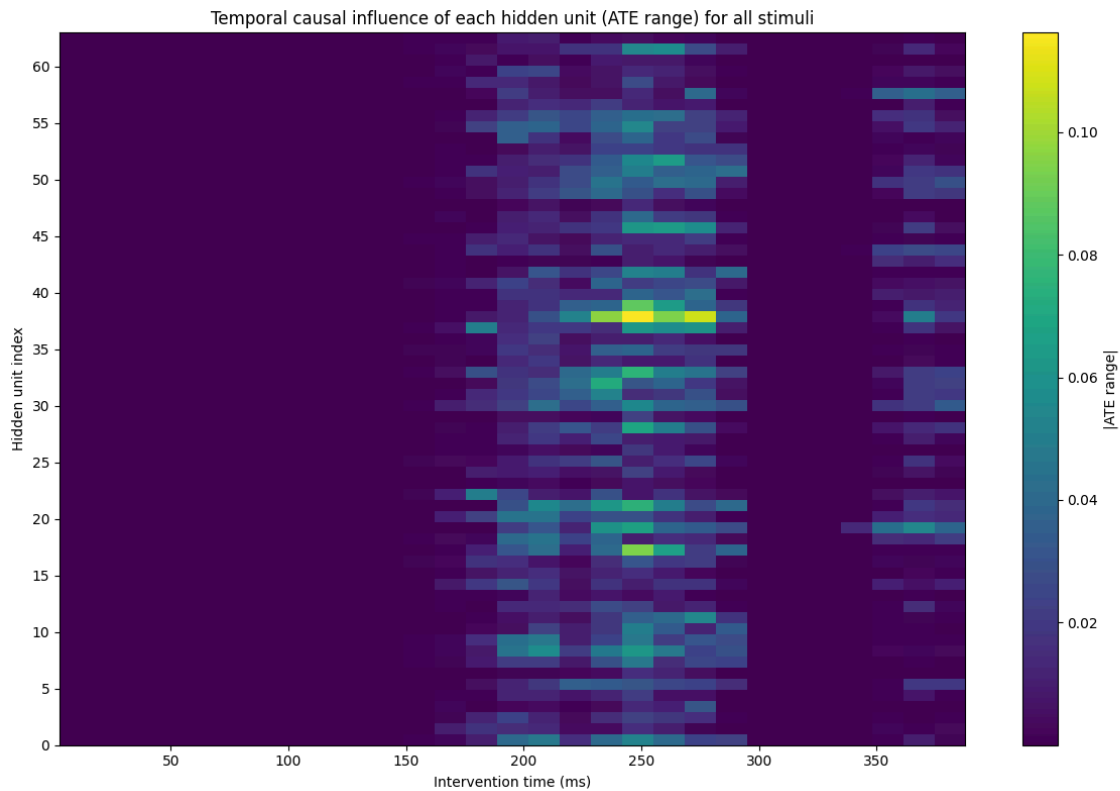


Figure 14: Heatmap showing population-level absolute ATE ranges (on a colour gradient) for all 64 units (y-axis) across every fifth time frame between (and excluding) the first and last frames (x-axis; scale shown in time (ms) instead of frame number).

¹⁴Because the FFT window length can span across the aspiration-vowel boundary, some information from the vowel is carried over to the spectral representation of the preceding onset windows, and this is likely why the within continua onset trajectories are not exactly the same.

Figure 14 is a heatmap showing the population-level (measure for all stimuli) absolute ATE ranges for units across sequences. Continuing the point about certain time frames being more important prediction drivers from the temporal plots, the absolute ATE range heatmap shows higher unit effects between 200-300 ms relative to rest of the other temporal regions. Unit 38 seems to show the highest range at time 250 ms (and its causal strength is also spread out to neighbouring frames), so the time frame 90 ($\frac{44 \times 90}{16000} = 248$ ms) was zoomed in for the ATE curves (Figure 15). Unit 38 shows strongest causality for the population at frame 90 (250ms) with ATE values going beyond -0.1 , the curve shows that unit 38 has an inverse relation with positive clamping, which shows that clamping the unit to positive values results in increasingly negative ratings (towards the opposite direction of <TH> in the response scale¹⁵); for negative clampings, the unit remains stable.

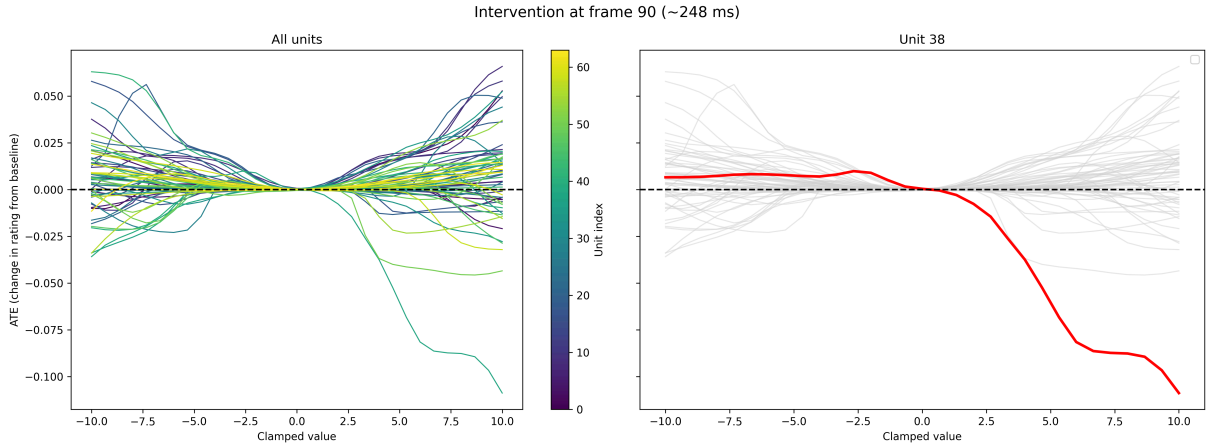


Figure 15: Average treatment effect shown by all hidden units (left) and unit 38 (right) at timeframe = 90, across clamped values ranging -10.0 to $+10.0$.

When reapplying the ATE procedure for within accent/continuum and within accent-continuum pair stimuli, unit effects across timeframes show up differently for each accent-continuum pair, and these combine to give the trend shown by within accent and within continuum ranges (see heatmaps in Table 5 in Appendix H.). For example, through visual inspection it appears that temporal causal effects for Cru: are combination of temporal causal effects for AmE-Cru: and AsE-Cru:, and effects for AsE are the combination from AsE-Cot and AsE-Cru:. This indicates that population-level unit effects are formed through distributed effects for accent and continuum. The population-split heatmaps also tell us about the origin of the strongest population-level causal unit: 38, which shows higher ranges relative to other units for AmE-Cot and AsE-Cot but not for within Cru: stimuli. The within-Cru: heatmaps show more distanced temporal locations for higher unit-ranges between the two accents compared to Cot. This clear separation is possibly the reason which is driving the bigger prediction gap between the two accents compared to the Cot, where prediction gaps are not as spaced out.

¹⁵This can be towards <T> or <D> depending on the stimulus step

4 Discussion

The results of the experiment show that AsE listeners can have different coronal stop categorisations to the same VOTs depending on the accent information carried by the following vowel. This adaptation was found to match the expected direction for the overall comparison and within the Cru: continuum; however, the opposite direction was seen for Cɔt. The findings for the Cru: pattern suggest that AsE listeners were adapting their definitions of the categories: voiceless and aspirated according to the accent they were listening, where the AmE vowel made long-lag stimuli (steps with higher VOT) categorise closer to the voiceless, but the AsE vowel made the same stimuli categorise further towards the aspirated. Assuming the AsE pattern is the default for the listeners, the results suggest that AmE vowel can make an aspirated coronal stop be perceived as voiceless for AsE listeners.

Apart from the possible explanations for the Cɔt pattern provided in Section 3.1.1, here another one is followed which relates to a possible stimuli inconsistency not accounted for in the methods. The results suggest that for lower VOTs of the AmE Cɔt, the stimuli made participants respond more towards <TH> compared to the AsE. When reviewing the Cɔt continua of AsE and AmE, a difference between the two continua was audibly apparent, with the AmE stimuli for lower-VOT steps seeming to lack salient bursts. This becomes more clear when comparing the spectrograms between the two accents for Cɔt step 1 (Figure 17), which shows visual differences between the two burst energies relative to their following vowel energies (in the form of spectrogram color intensity and relative wave amplitudes). The reasoning for why the stimuli manipulation resulted in this difference is beyond the scope of the study (however, it maybe possibly related to FFT window spanning across acoustic feature boundaries¹⁴). The presence of a less audibly salient burst creates more room for the listener for them to perceive it as the debuccalised¹⁶ form (/h/), especially in the presence of noise (Narayan, 2025); for the experiment's case, the <TH> side of the scale maybe a closer approximation to /h/ than <T>. Although participants were advised to do the task in a quiet environment, there was no stringent inspection for it.

Another problem related to the experiment might be the adapted VAS scale for the means observed in the results which mostly range between the <T> and <TH>; having the <T> in the center allows participant to place clicks close to it, whereas this is more difficult when the point is only connected to the scale from one side like it is for <TH> and <D>. This could indicate why means for higher VOT steps do not reach scale values close to <TH> but for lower VOT steps the means reach values close to <T>.

The exploratory theme of the modelling in the study makes it lack quantitative metrics for the model analyses. Additionally, inherent architectural differences between the original method from Beguš et al. (2023) and the LSTM used here create separate implications on the ATE results. The LSTM does not have a hidden state which can be clamped independently from the encoding of stimuli. The hidden states h of the LSTM are inherently dependent on the stimuli population X as they are produced by the stimuli through learned recurrent weights.

¹⁶This is not used here in the strict sense where manner is kept while the non-laryngeal place is lost, but for non-laryngeal to laryngeal place shifts regardless of manner.

Therefore, the causal effects shown by the units should be understood within the context of the model itself, and it does not allow this study to extend them further into claims about human perception. Nonetheless, this method makes it possible for locating the influential units while stimuli are processed on the go (temporally) and determining their effect and directionality. The ATE procedure done on this study found unit 38 to be a strong population-level driver in pulling the predictions to lower scale values at ~250 ms.

The next possible steps in the same direction as this study's exploratory modelling would be to incorporate more behaviourally grounded temporal data, for example: eye-tracking, and to use a sequential model with human-based learning algorithm, along with more robust quantitative reporting for the model analyses. Similar to the model, the adapted VAS task also requires more powerful statistical analysis which allows confounds to be accounted for. An important future direction adjacent to the subject of the experiment is to have more empirically grounded (beyond VOT) production/perception descriptions of the AsE and Assamese coronal stops; this will come in help to establish stimuli creation on better auditory prototypes.

5 Conclusion

This study investigated the VOT-based coronal stop perception of AsE listeners between two English accents (AmE, AsE) for shifts in categorisation measured using an adapted VAS task. Results showed that participants on average show significant difference in responses between the two accents for the Cru: continuum, i.e. they shift their categorisation according to the accent they hear (the directionality of the shift tells that the listeners adapt to the accent variation of the speakers in order to hear towards the speakers' intended categories). The experiment was modelled with an LSTM and analysed for temporal prediction trajectories and unit causality to explore how the learned behaviour is represented within the model. It was found that VOT differences were encoded in a way which shifted the phase of the temporal prediction trajectories, and the ATE procedure revealed unit 38 with strongest effects which was localised ~250 ms (frame 90), where positive clamps pushes prediction towards the opposite direction of <TH> in the scale.

References

- Apfelbaum, K. S., Kutlu, E., McMurray, B., & Kapnoula, E. C. (2022). Don't force it! Gradient speech categorization calls for continuous categorization tasks. *The Journal of the Acoustical Society of America*, 152(6), 3728–3745. <https://doi.org/10.1121/10.0015201>
- Archangeli, D. B., & Yip, J. (2019). Assamese vowels and vowel harmony. *Journal of South Asian Languages and Linguistics*, 6(2), 151–183. <https://doi.org/10.1515/jsall-2019-2010>
- Beguš, G., Leban, A., & Gero, S. (2023). *Approaching an unknown communication system by latent space exploration and causal inference*. <https://arxiv.org/abs/2303.10931>
- Boersma, P., Chládková, K., & Benders, T. (2022). Phonological features emerge substance-freely from the phonetics and the morphology. *Canadian Journal of Linguistics/revue Canadienne De Linguistique*, 67(4), 611–669.
- Boersma, P., Weenink, D., & Shchupak, A. (2026, May). *Praat: doing phonetics by computer*. <https://praat.org/>
- Cho, T., Jun, S.-A., & Ladefoged, P. (2002). Acoustic and aerodynamic correlates of Korean stops and fricatives. *Journal of Phonetics*, 30(2), 193–228. <https://doi.org/10.1006/jpho.2001.0153>
- Davidson, L. (2016). Variability in the implementation of voicing in American English obstruents. *Journal of Phonetics*, 54, 35–50. <https://doi.org/10.1016/j.wocn.2015.09.003>
- Dutta, H., & Kenstowicz, M. (2018). The phonology and phonetics of laryngeal stop contrasts in Assamese. In R. Petrosino, P. Cerrone, & H. van der Hulst (Eds.), *Beyond the Veil of Maya* (pp. 30–64). De Gruyter Mouton. <https://doi.org/10.1515/9781501506734-002>
- Falcon, W., & The PyTorch Lightning team. (2024, August). *PyTorch Lightning*. Zenodo. <https://doi.org/10.5281/zenodo.13254264>
- Gargesh, R. (2008). Indian English: phonology. In R. Mesthrie, B. Kortmann, & E. W. Schneider (Eds.), *Africa, South and Southeast Asia* (pp. 231–243). De Gruyter Mouton. <https://doi.org/10.1515/9783110208429.1.231>
- Gwilliams, L., Bhaya-Grossman, I., Zhang, Y., Scott, T., Harper, S., & Levy, D. (2025). Computational architecture of speech comprehension in the human brain. *Annual Review of Linguistics*, 11, 209–226. <https://doi.org/10.1146/annurev-linguistics-031120-111245>
- Harris, C. R., Millman, K. J., Walt, S. J. van der, Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Kerkwijk, M. H. van, Brett, M., Haldane, A., Río, J. F. del, Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825), 357–362. <https://doi.org/10.1038/s41586-020-2649-2>
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3), 90–95. <https://doi.org/10.1109/MCSE.2007.55>
- Hwang, J., Hira, M., Chen, C., Zhang, X., Ni, Z., Sun, G., Ma, P., Huang, R., Pratap, V., Zhang, Y., Kumar, A., Yu, C.-Y., Zhu, C., Liu, C., Kahn, J., Ravanelli, M., Sun, P., Watanabe, S., Shi, Y.,

- ... Petridis, S. (2023). *TorchAudio 2.1: Advancing speech recognition, self-supervised learning, and audio processing components for PyTorch*.
- Jin, L. (2025). *A perception study on the function of creak in Yanbian Korean* [Master's thesis]. <https://www.fon.hum.uva.nl/archive/2025/2025-MA-LiJin.pdf>
- Kassambara, A. (2026). *ggpubr: 'ggplot2' Based Publication Ready Plots*. <https://doi.org/10.32614/CRAN.package.ggpubr>
- Kim, M.-R. (2014). Ongoing sound change in the stop system of Korean: A three- to two-way categorization. *Studies in Phonetics, Phonology, And Morphology*, 20, 51–82. <https://doi.org/10.17959/sppm.2014.20.1.51>
- Kingma, D. P., & Ba, J. (2017). *Adam: a method for stochastic optimization*. <https://arxiv.org/abs/1412.6980>
- Kominek, J., & Black, A. W. (2004). The CMU Arctic speech databases. *5th ISCA Workshop on Speech Synthesis (SSW 5)*, 223–224. http://www.festvox.org/cmu_arctic/index.html
- Kutlu, E., Chiu, S., & McMurray, B. (2022). Moving away from deficiency models: Gradiency in bilingual speech categorization. *Frontiers in Psychology*, 13. <https://doi.org/10.3389/fpsyg.2022.1033825>
- Ladefoged, P., & Disner, S. F. (2012). *Vowels and consonants*. John Wiley & Sons.
- Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: acoustical measurements. *WORD*, 20(3), 384–422. <https://doi.org/10.1080/00437956.1964.11659830>
- Mahanta, S. (2012). Assamese. *Journal of the International Phonetic Association*, 42(2), 217–224. <http://www.jstor.org/stable/26351864>
- McFee, B., McVicar, M., Gover, M., Hammudoglu, J., Balke, S., Seyfarth, S., Lostanlen, V., Malek, A., Raffel, C., Niekirk, B. van, Lee, D., B, D. H., Cwitkowitz, F., Zalkow, F., Ellis, D., Nieto, O., Fernandes, D., Mason, J., Rabinovich, E., ... AD. (2026, July). *librosa*. Zenodo. <https://doi.org/10.5281/zenodo.21726086>
- McMurray, B. (2022). The myth of categorical perception. *The Journal of the Acoustical Society of America*, 152(6), 3819–3842. <https://doi.org/10.1121/10.0016614>
- Narayan, C. R. (2025, January 11). *An acoustic-perceptual explanation for debuccalization in 10th century Dravidian*. <https://ch-narayan.github.io/LSA2025-Narayan.pdf>
- Nath, S., Sarma, H., & Sharma, U. (2014). A preliminary study on the VOT patterns of the Assamese language and its Nalbaria variety. *Computational Linguistics and Intelligent Text Processing*, 542–552.
- Nath, S. K. (2025). Contrast, context, and contact. *Asia-Pacific Language Variation*, 11(1–2), 187–222. <https://doi.org/10.1075/aplv.25002.nat>
- The pandas development team. (2024, January). *pandas-dev/pandas: Pandas*. Zenodo. <https://doi.org/10.5281/zenodo.10537285>

- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). *PyTorch: an imperative style, high-performance deep learning library*. <https://arxiv.org/abs/1912.01703>
- Puggaard-Rode, R. (2024). praatpicture. A library for making flexible Praat Picture-style figures in R. *Proceedings of the International Seminar on Speech Production*, 13, 145–148.
- Python Software Foundation. (2026). *Python Language Reference, Version 3.14*. <https://www.python.org/>
- R Core Team. (2026). *R: A Language and Environment for Statistical Computing*. <https://doi.org/10.32614/R.manuals>
- Sarmah, P., & Wiltshire, C. R. (2024). Acoustic properties of the monophthongs of Assamese Indian English speakers. *World Englishes*, 43(1), 161–181. <https://doi.org/10.1111/weng.12637>
- Schertz, J., & Khan, S. (2020). Acoustic cues in production and perception of the four-way stop laryngeal contrast in Hindi and Urdu. *Journal of Phonetics*, 81, 100979. <https://doi.org/10.1016/j.wocn.2020.100979>
- Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60), 3021. <https://doi.org/10.21105/joss.03021>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>
- Wiltshire, C. R. (2020). *Uniformity and Variability in the Indian English Accent*. Cambridge University Press.
- Wiltshire, C., & Sarmah, P. (2021). Voicing contrasts in the stops of Indian English produced by Assamese speakers. *Proceedings of Meetings on Acoustics*, 42(1), 60003. <https://doi.org/10.1121/2.0001389>

Appendix

A. Response functions

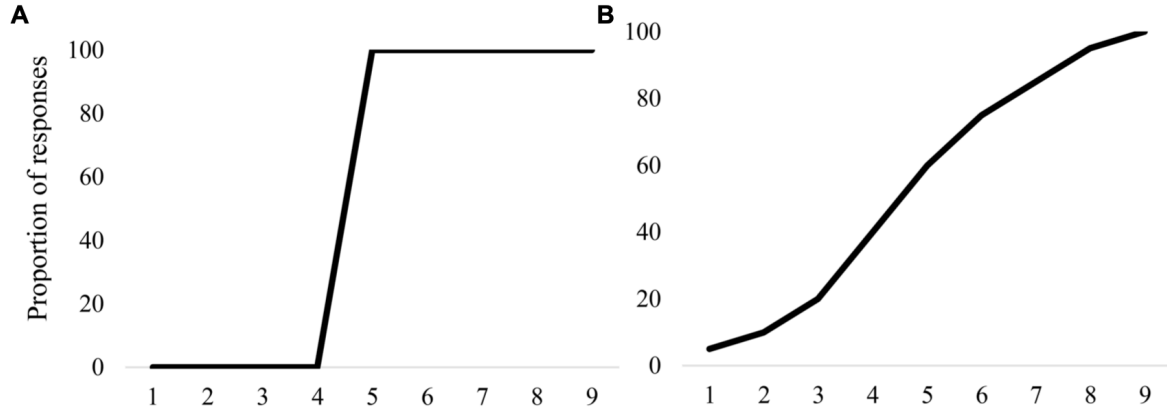


Figure 16: Examples of steep (left) and shallow (right) response functions (proportion of responses by stimuli steps on a sound continuum) for the 2AFC task. (figure sourced from Kutlu et al. (2022))

B. Zoomed Cst Step 1 Stimuli

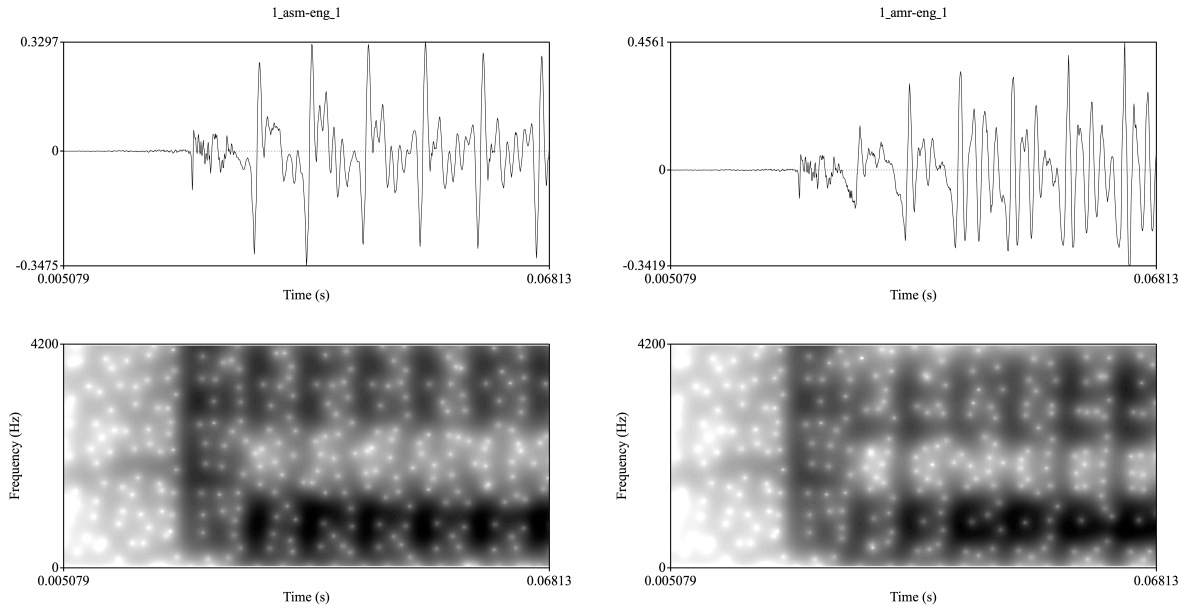


Figure 17: Stimulus Cru: step 1 for AsE (left) and AmE (right) zoomed in to show the first 681.3 ms

C. Stimuli

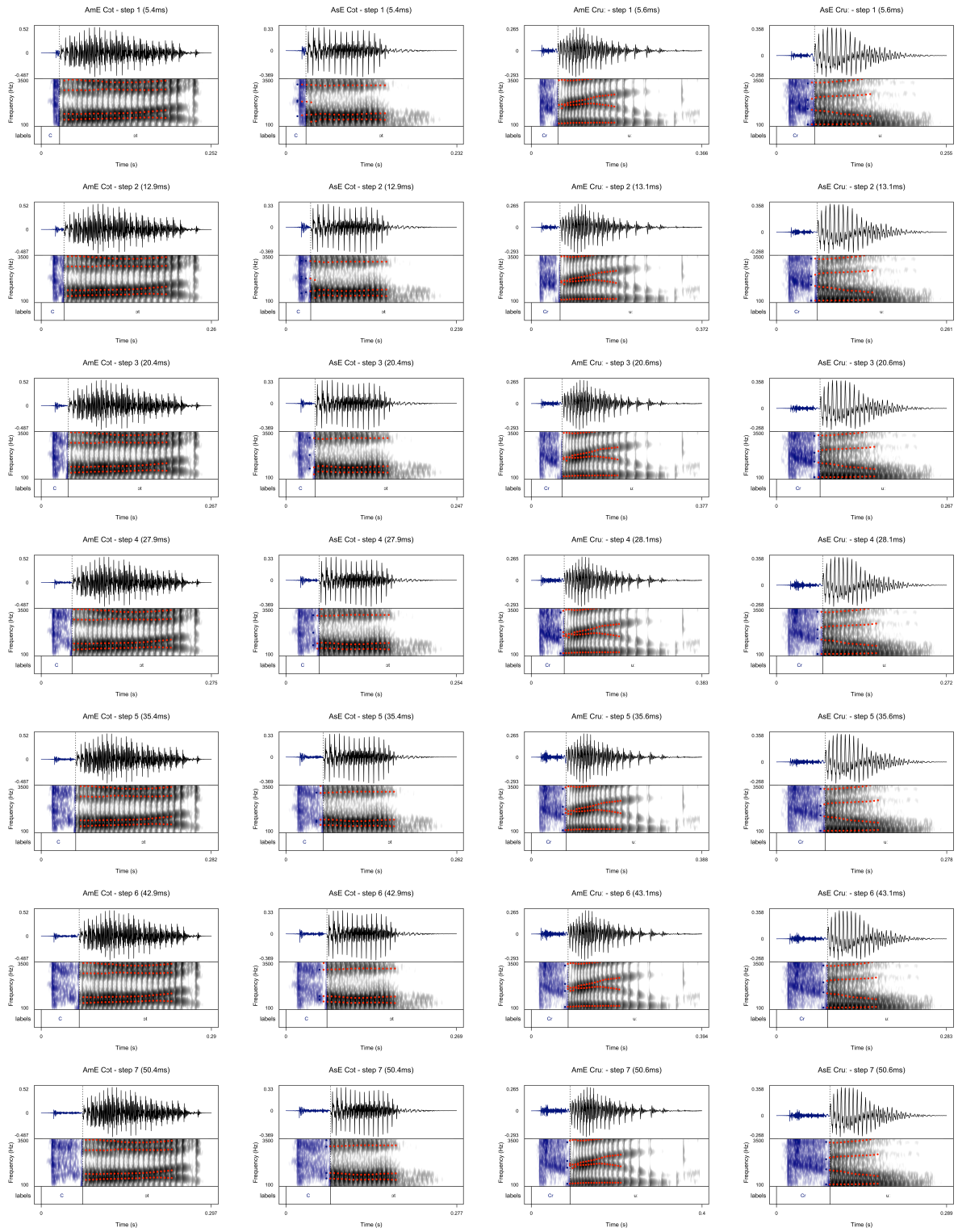


Figure 18: All stimuli pictures: AmE-Cot, AsE-Cot, AmE-Cru, AsE-Cru: (vertically) and steps 1 to 7 (horizontally)

D. Mel-Spectrograms

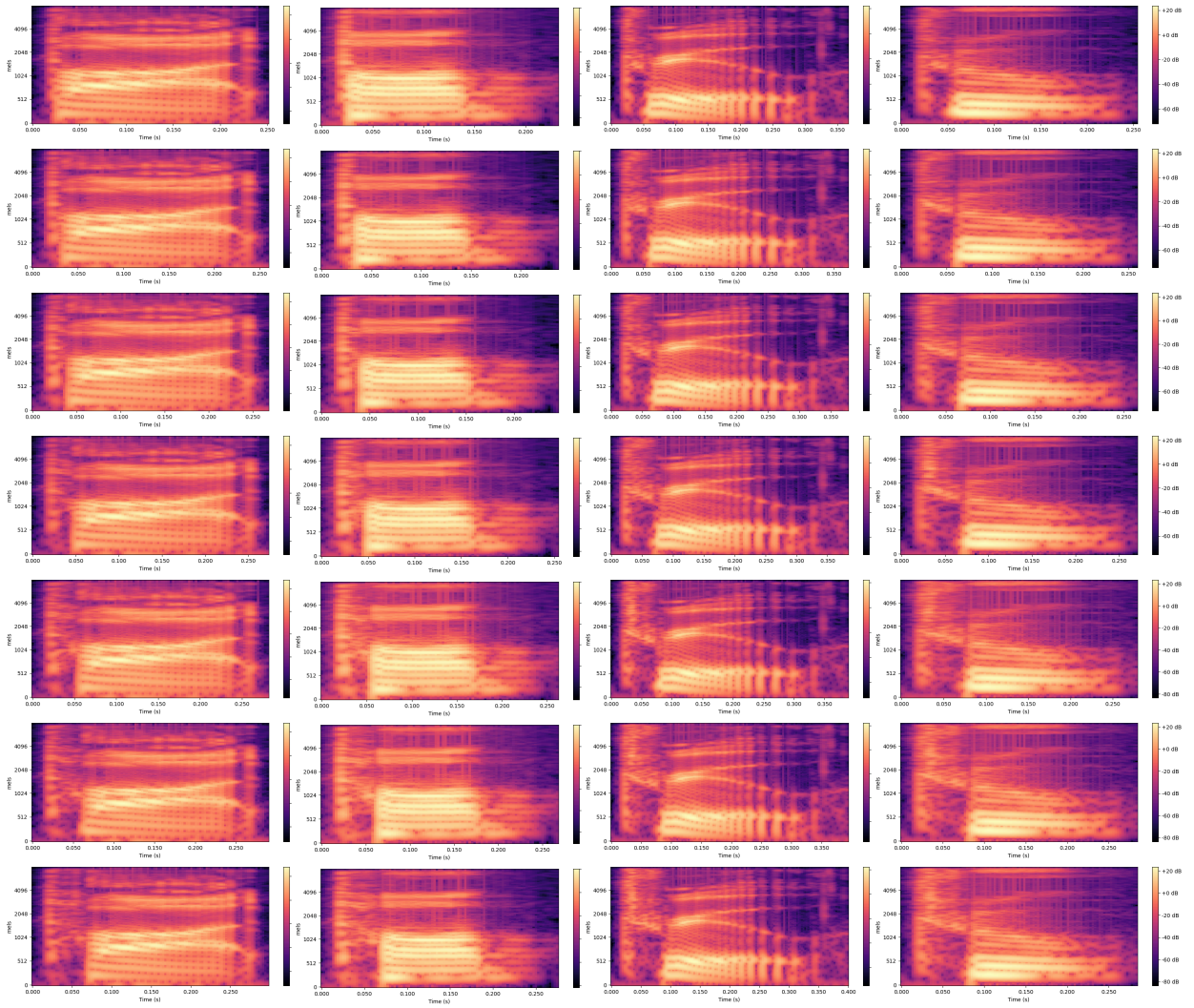


Figure 19: All stimuli extracted mel-spectrograms: AmE-Cot, AsE-Cot, AmE-Cru:, AsE-Cru: (vertically) and steps 1 to 7 (horizontally)

E. Filtered data summary

subject	english	region
Length : 3671 N.unique : 11 N.blank : 0 Min.nchar : 2 Max.nchar : 3	Length : 3671 N.unique : 2 N.blank : 0 Min.nchar : 3 Max.nchar : 3	Length : 3671 N.unique : 2 N.blank : 0 Min.nchar : 1 Max.nchar : 2
step	vot	continuum
Length : 3671 N.unique : 7 N.blank : 0 Min.nchar : 1 Max.nchar : 1	Min. : 5.40 1st Qu. : 13.10 Median : 27.90 Mean : 27.97 3rd Qu. : 42.90 Max. : 50.60	Length : 3671 N.unique : 2 N.blank : 0 Min.nchar : 3 Max.nchar : 4
response	reactionTime	stimulus
Min. : -0.9920 1st Qu. : -0.0170 Median : 0.0980 Mean : 0.2723 3rd Qu. : 0.8100 Max. : 0.9980	Min. : 0.502 1st Qu. : 1.194 Median : 1.550 Mean : 1.882 3rd Qu. : 2.121 Max. : 79.822 ¹⁷	Length : 3671 N.unique : 28 N.blank : 0 Min.nchar : 77 Max.nchar : 78

¹⁷In a few cases the participant(s) wanted could pause for some time due to distraction (eg: phone call), they were asked to place the cursor where they want to respond for the present trial and click on it only after they are ready to continue, thus the high max. reactionTime value.

F. Participant response plots

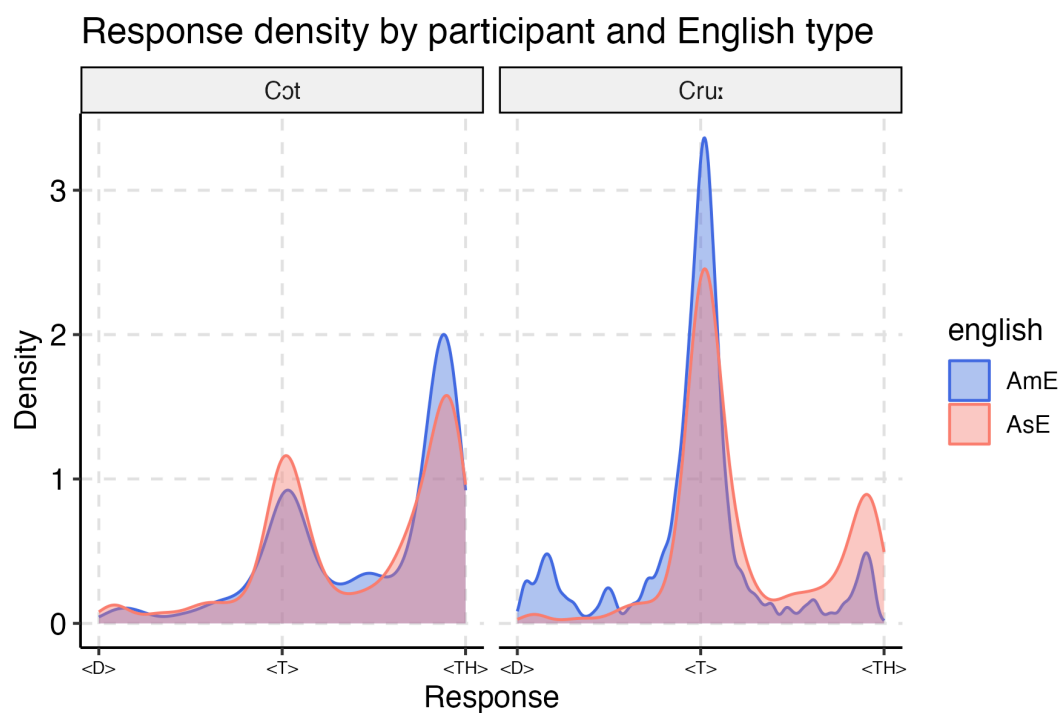


Figure 20: Density plot for responses on the VAS scale by English variety.

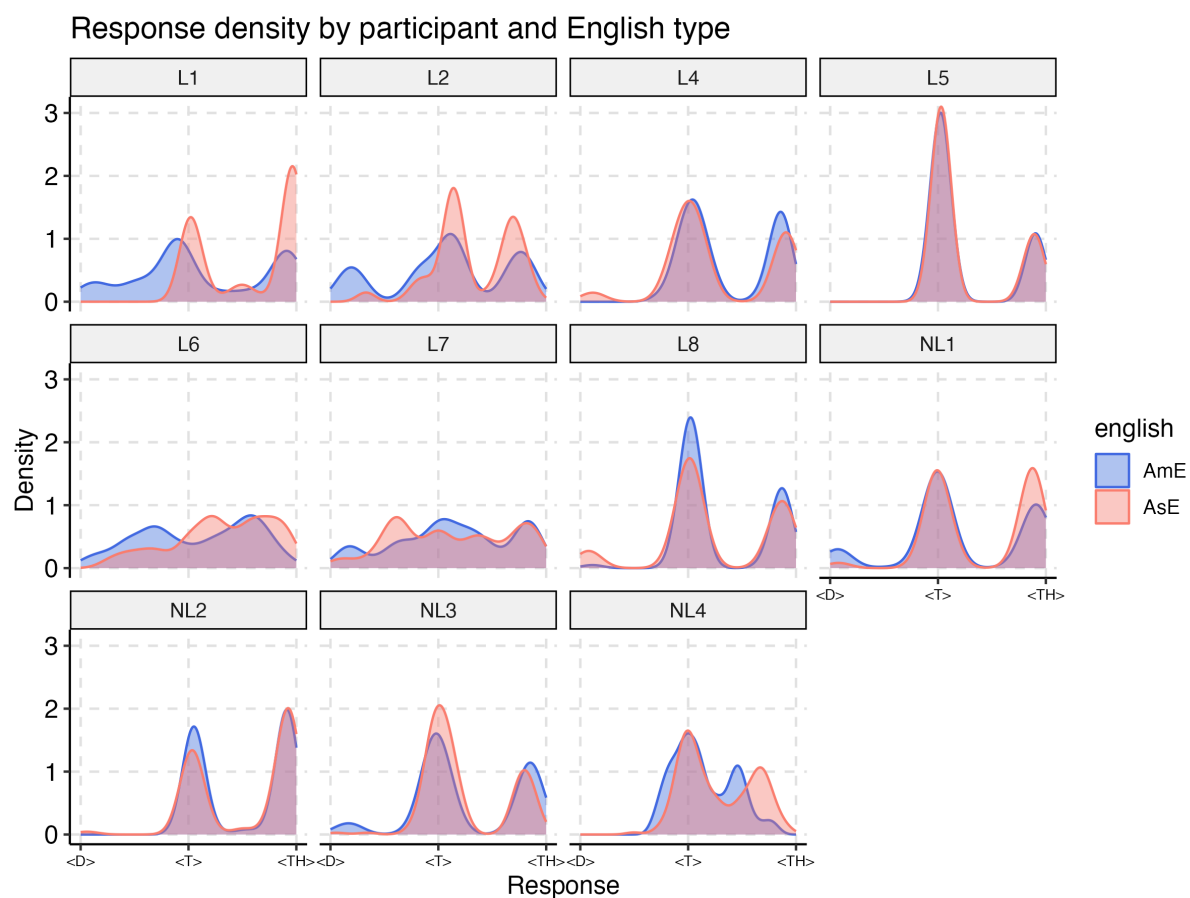


Figure 21: Per participant density plots for responses on the VAS scale by English variety.

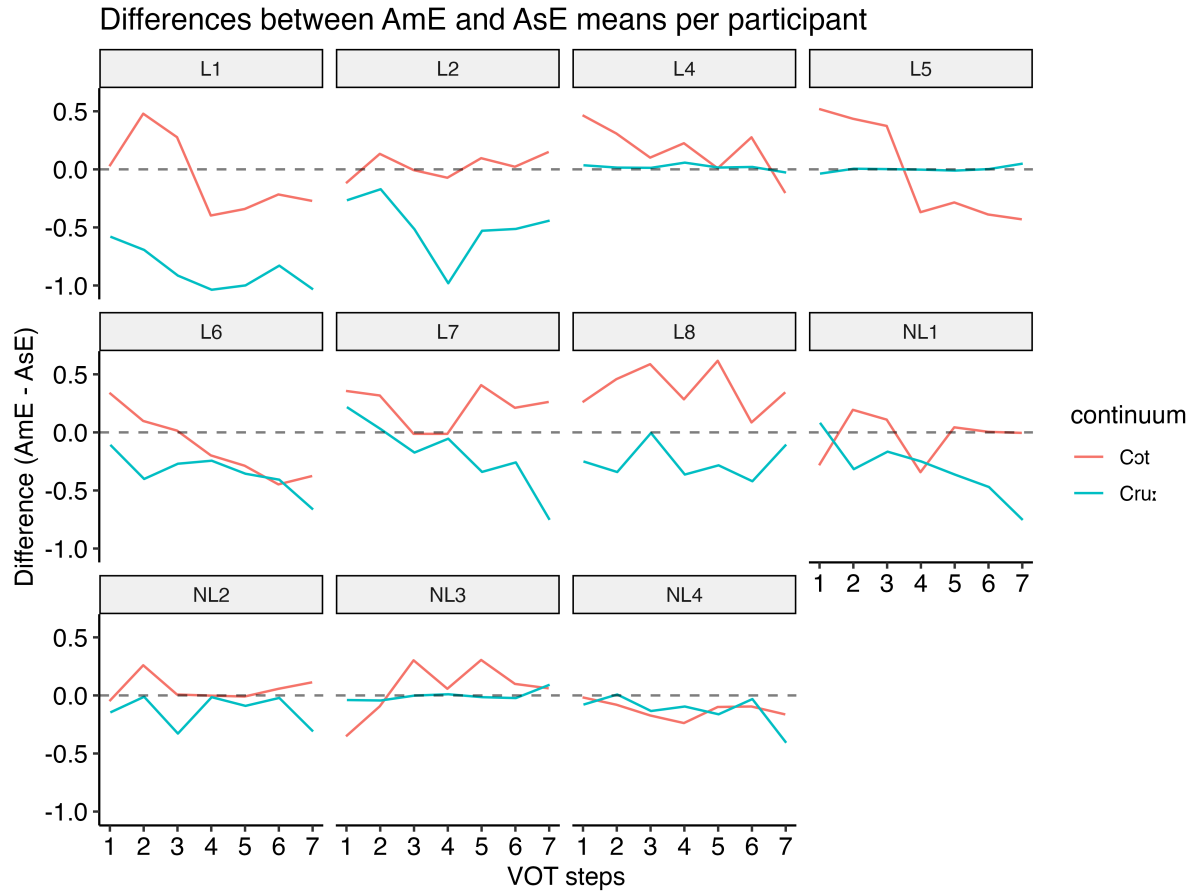


Figure 22: Difference of means between AmE and AsE for each step per participant.

G. Loss

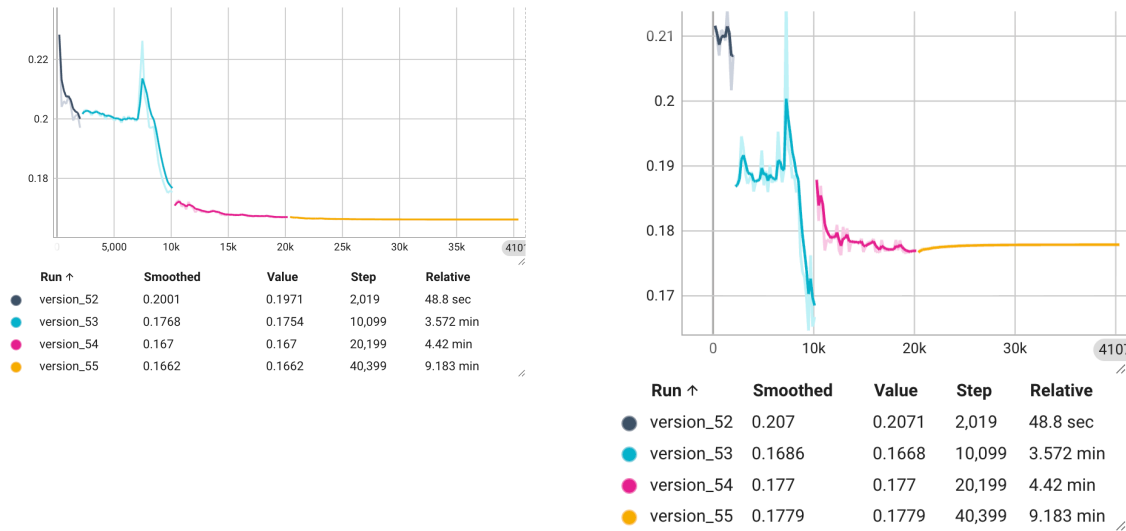
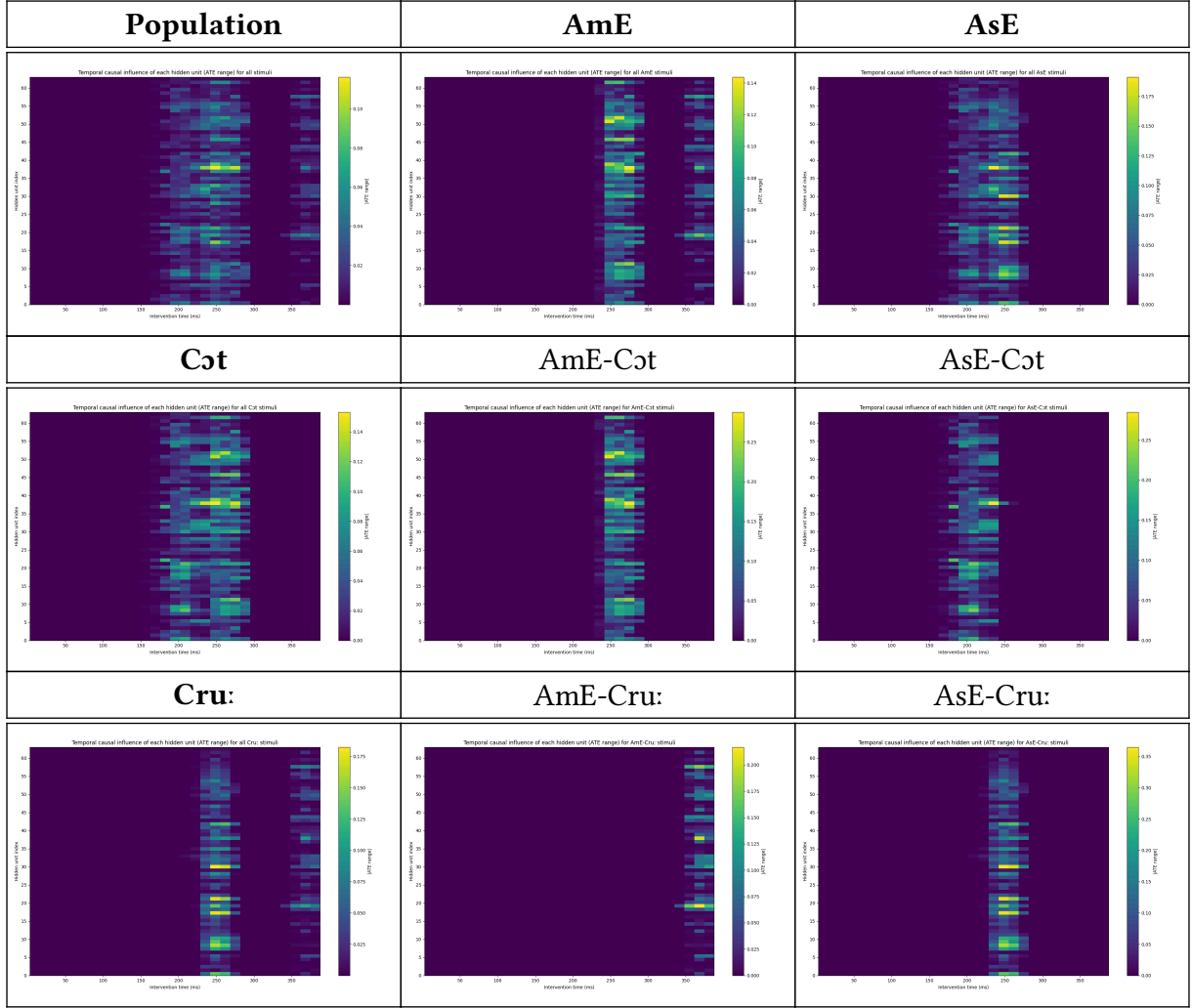


Figure 23: Training (left) and validation (right) losses for 200 epochs from TensorBoard¹⁸ (curves coloured for four training checkpoints).

¹⁸The tensorboard step counter may include additional validation steps.

H. Temporal Causal Heatmaps

Table 5: Temporal ATE range heatmaps for population stimuli, groups and intersection splits



I. Scripts

Find all scripts used in this study at https://github.com/elempenyi/thesis_26 or at <https://fon.hum.uva.nl/archive>.

Repository structure:

```
-- thesis_26/
| |-- doc/ (figures, pdf and typst code for this document)
| |-- src/
| | |-- experiment/ (vot manipulation, ExperimentMFCs, analysis code)
| | |-- modelling/ (model.py, loss logs, melspec-extraction, modelling code)
```