Probing a Neural Network Model of Sound Change for Perceptual Integration

Cerys Hughes
University of Massachusetts Amherst

1 Introduction

The main cue to a phonological contrast can shift over time, e.g. a voicing contrast ([pa] vs [ba]) becoming a pitch contrast ([pá] vs [pà]). Some contrast shift sound changes are more likely to occur than others, but existing computational models of sound change (e.g. Kirby & Sonderegger, 2015) do not directly[1] focus on contrast shifts or asymmetries in their typology.

This paper builds on Yang (2019)'s proposal that contrast shift is more likely to occur between cues that enhance the same auditory dimension. While supported by a cross-linguistic survey, Yang (2019)'s account remains to be computationally implemented, with a model of these auditory dimensions as a prerequisite. As a first step toward this goal, I tested whether the speech perception component of a neural network model of sound change (Beguš, 2020) exhibits behavior characteristic of the auditory dimension "spectral continuity," relevant to stop voicing, by adapting a relevant human experimental paradigm (Kingston et al., 2008).

Motivated by the theoretical properties of the auditory dimensions appearing similar to the model's representations for processing acoustic data, I tested the hypothesis that the model's perceptual results would show dependence between the same pairs of cues as human listeners have been found to. However, the results of this probing experiment did not support the hypothesis that the model exhibits human-like behavior, indicating that the model's behavior in practice differs from expected and that further modification is needed to model the proposed bias in contrast shift typology.

Furthermore, the model under investigation was a Convolutional Neural Network (CNN), commonly used in speech applications (e.g. Palaz et al., 2013), and these results identified one way that they may process acoustic data differently from human listeners. The probing experiment that I conducted in this paper represents a synthesis of a human experimental paradigm and a neural network interpretation strategy; it combines phonetic experimental paradigms, where human participants' representations of stimuli are estimated with tasks, and neural network probing, which directly accesses the model's internal representations of stimuli.

I extended Ward's (2019) probing method, which was used to compare human visual perception to a CNN trained for visual tasks; I applied it to a speech processing experiment. Burridge & Vaux (2022) also applied a human experimental paradigm to a speech processing CNN, but their method is incompatible with the human experimental paradigm relevant to the auditory dimensions of Yang (2019)'s account. Burridge & Vaux (2022) probed the output of the network rather than the internal representations because their focus was on the misperception of natural speech categories in accordance with the categories in the learning data, rather than discriminating stimuli with arbitrary categories that may not naturally occur, as in the human experiment of interest (Kingston et al., 2008).

2 Background

2.1 *Integrated Perceptual Properties* Rather than perceiving cues such as fundamental frequency (f0), voicing duration, and formant values independently, humans are thought to also perceive auditorily similar acoustic properties as integrated into combined, abstract auditory dimensions called Integrated Perceptual

* I would like to thank Michael Becker, John Kingston, and Kristine Yu for their helpful feedback. I would also like to thank everyone at UMass Amherst's Sound Workshop and AMP.

[1] A potential model of contrast shift is discussed briefly by Wedel (2006). Contrast shifts are also implicitly modeled by Kirby (2010) with a Gaussian mixture models focusing on independently-perceived cues.

Proceedings of *AMP 2022*

Properties (IPPs) (Kingston et al., 2008). The integration of cues into an IPP is not categorical, but instead gradient based on the degree to which the cues can be perceived separately. As a case study, I tested convolutional neural networks for exhibiting behavior characteristic to an IPP relevant to stop voicing contrasts: spectral continuity, which refers to the lack of an abrupt change in the spectrum over time. Voiced stops tend to have higher spectral continuity across their closure onset than voiceless stops (Kingston et al., 2008). A stop closure results in a change from a high-energy preceding vowel to a timespan of mostly silence, with higher frequencies filtered out by the obstruction (Lisker, 1957). However, this change is less dramatic if voicing persists through the closure, which tends to occur for stops phonologically categorized as [voiced] (Figure 3). The duration of voicing during the closure thus contributes to the percept of spectral continuity. Spectral continuity is also enhanced if the low-frequency energy of the closure voice bar (Hogan & Rozsypal, 1980) is preceded by a low first formant (F1) at the offset of the preceding vowel. Additionally, a lower vowel offset fundamental frequency (f0) contributes to spectral continuity because f0 tends to be lower during closure voicing (Kingston et al., 2008). Figure 1 shows example tokens of a high-spectral-continuity voiced stop and low-spectral-continuity voiceless stop.

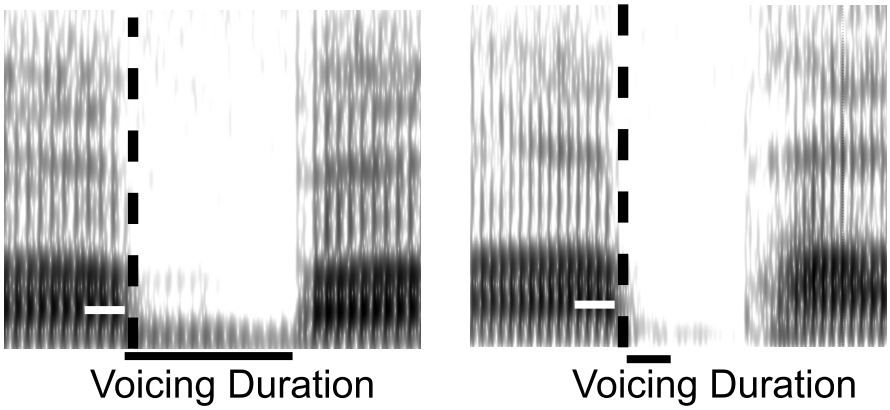


Figure 1: Left: a spectrogram of a voiced English stop with an offset F1 (indicated by a white line) of 480Hz. Right: a spectrogram of a voiceless stop with an offset F1 (indicated by a white line) of 550Hz. The closure onsets are marked by a dashed line, and closure voicing duration is marked by a solid horizontal line. These tokens were collected by (MIT, 2005).

If cues contribute to the same IPP, then they will not be perceived independently of one another, referred to as perceptual integration (Kingston et al., 2008). Perceptual integration for a pair of cues can be diagnosed with the Garner paradigm (Garner, 1953), which tests the perceptual distances between four stimuli that are constructed with all of the logically possible combinations of high versus low values of each cue. For example, if we were testing for the perceptual integration of F1 and voicing duration during the closure, our four stimuli would consist of sounds synthesized to have a vowel-stop-vowel (VCV) sequence with one of these combinations of acoustic properties: high F1 and long voicing, high F1 and short voicing, low F1 and long voicing, or low F1 and short voicing (Fig. 4). The physical distances between each pair of stimuli can thus be described in terms of how many cues they differ on. The perceptual distance between each pair of cues is estimated with a discrimination task (Kingston et al., 2008).

If the cues are perceived independently, the relative perceptual distances will be proportional to the physical distances. However, if the two cues are not perceived independently, the perceptual distances will be warped from the physical ones; two pairs of stimuli might differ on the same cues, but one pair might be further apart than another based on their values. For example, for F1 and voicing in Fig. 2, the pair High/Short vs Low/Long differs on the same number of cues as High/Long vs Low/Short. However, High/Short and Low/Long are perceptually further apart. In other words, the perceptual space is stretched along a diagonal dimension that is a combination of the F1 and voicing duration dimensions. This dimension corresponds with spectral continuity; High/Short corresponds with low spectral continuity, and Low/Long corresponds with high spectral continuity. Pairs of cues that covary in speech but do not contribute to the IPP can be tested as a control (Kingston et al., 2008). Table 1 summarizes the pairs of cues investigated in this paper for the

comparison between model and human perceptual representations.

Cue Pair	Human Behavior
F1 and closure voicing	Integrated
f0 and closure voicing	Integrated
F1 and closure duration	Not integrated
f0 and closure duration	Not integrated

Table 1: Summary of human results found by Kingston et al (2008). Cues enhancing "spectral continuity" are shown in italics.

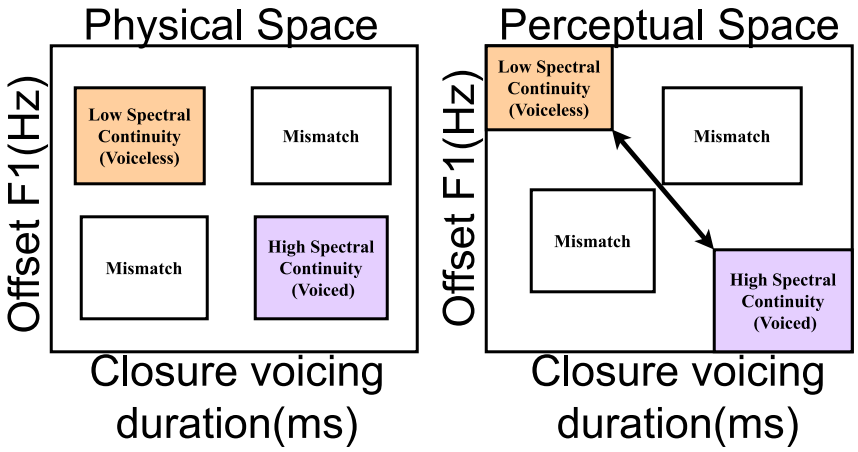


Figure 2: Schematic of the Garner paradigm showing perceptual integration of F1 and closure voicing, adapted from Kingston et al. (2008). Each box represents a stimulus, which is a sound consisting of a vowel-stop-vowel sequence.

2.2 *Contrast Shift* Yang (2019) proposed that IPPs aid in explaining the asymmetries in the cross-linguistic frequency of contrast shifts in a typological and experimental study. Yang found that participants' cue reweighting ability was enhanced when the cue values corresponded to an IPP and concluded that this synchronic cue reweighting preference is related to the typological observation that diachronic contrast shifts are more common between cues that contribute to the same IPP.

Yang focused on the cues of voicing, breathiness, pitch, and vowel duration. Voicing, breathiness, and low pitch are hypothesized to enhance the IPP "energy at low frequencies;" breathiness contributes to this IPP because it is characterized by more energy on the first harmonic than higher harmonics. Vowel duration also tends to coincide with these cues, but there is no known shared IPP auditory effect that they contribute to. In a cross-linguistic survey of consonant contrasts involving these cues, Yang (2019) found that the contrast shifts were much more frequent between the cues that enhance the same IPP (voicing, breathiness, pitch) and almost never occurred between between the cues that covary but do not enhance the same IPP (vowel duration and pitch).

Yang (2019) thus proposed a potential mechanism by which IPPs influence contrast shift: experimental participants' ability to change their cue weighting for two artificial categories depends on where the cue values fall on IPP dimensions. Participants received extensive training in an initial phase where the categories were better separated by one cue, then encountered another phase where the more informative cue was switched. Yang (2019) manipulated whether the particular cue values were enhancing, falling along the IPP dimension, or non-enhancing, falling perpendicular to the IPP dimension (Fig. 2). Participants were not able to adapt to the cue change if the values opposed the IPP dimension. Yang (2019) thus proposed participants' preference for IPPs in changing cues is related to IPP cues' frequency in diachronic contrast shifts. However, this account remains to be computationally implemented.

2.3 *Neural Network Model* I propose the application of a neural network model of sound change to the task of implementing Yang (2019)'s account and took the first step of evaluating the behavior of its speech perception component. A neural network maps input vectors to output vectors with multiple layers of weighted sums and nonlinear functions, such as the logistic function (Ito, 1991). It learns by adjusting its weights to optimize its performance on training data (examples of input-output pairs), specifically by minimizing a loss function with one of the many forms of gradient descent (Ruder, 2017). An existing neural network model of sound change (Beguš, 2020) uses learners that are each implemented with a Generative Adversarial Network, where the "discriminator" is roughly analogous to a speech perception component. This component is implemented with a Convolutional Neural Network (CNN), the type of neural network probed in this paper. A CNN does not require predefined cue measurements as its input, but rather detects patterns directly from the raw audio waveform. This detection of patterns, as opposed to simply weighing and transforming different measurements taken by phoneticians for different cues, has the potential to model how properties of the signal we would measure separately (e.g. low voicing duration and F1) may be captured by a single dimension in the auditory system.

CNNs make use of pattern detectors called filters to process their inputs (O'Shea & Nash, 2015). A filter is a vector of numbers that is compared to each section of the input; the more similar they are, the greater the filter's output is for that section. When the filter closely matches the region of the input, there is a larger value in the corresponding region of the output. A dot product is computed between the filter and each section of the input. Crucially, the result of the dot product tends to be higher if the two vectors (in this case, the filter and a region of the input) are more similar; in other words, the result is greater if the input region matches the filter's pattern (O'Shea & Nash, 2015). Figure 3 shows a step-by-step calculation of how a filter's output is produced by computing dot products at different regions of the input. A CNN makes use of many of these filters to detect many different patterns, and it also feeds the results as inputs into more layers of filters, allowing the for detection of longer, higher-level patterns (O'Shea & Nash, 2015).

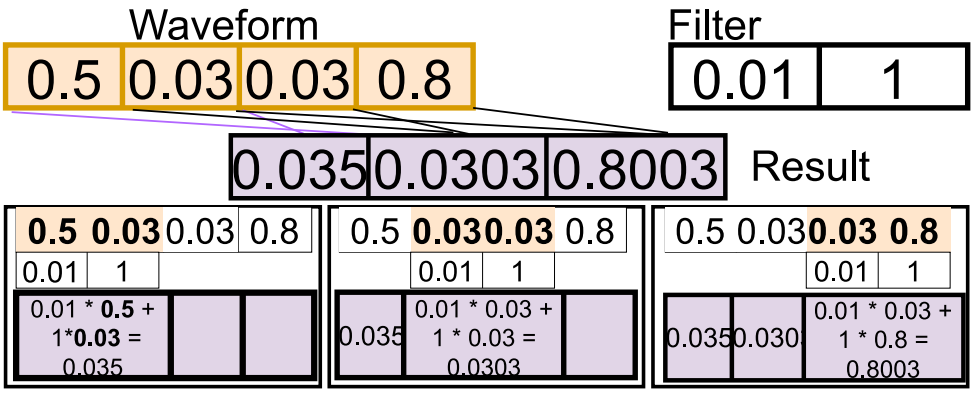


Figure 3: Step-by-step calculation of the output for a toy input signal and filter. The output of the [0.01, 1] filter is greatest at the [0.03, 0.8] region of the input.

The properties of CNN filters provide reason to expect that they might provide a model of cues integrated into the same auditory dimension, like an IPP. The filters simply detect patterns in regions of the input, so CNNs are not constrained to learning separate sets of filters for independently detecting acoustic measures such as a vowel offset F1 and voicing. Instead, a CNN could potentially learn a shared set of filters representing the pattern of a low F1 adjacent to substantial closure voicing. Because the cues in an IPP contribute to the same abstract effect, such as spectral continuity, it is not unimaginable that a CNN would learn a shared filter for cues that are in the same IPP, while learning separate, independent filters for pairs of cues that have no such qualitative similarity, such as a low F1 and closure duration. To visualize this distinction between shared and independent filters, Figure 4 shows a hypothetical example for the sounds in Figure 3. The CNN that I probed had the task of categorizing English stops as voiced or voiceless (as discussed in Section 4.1). A sound token's vowel F1, closure voicing duration, and closure duration all help indicate a stop's category, so we could expect the CNN to detect these values in some way. However, out of these cues, only a low F1 and voicing duration contribute to the qualitative pattern of spectral continuity.

As voiced stops tend to have higher spectral continuity, this pattern provides information about whether a token is voiced or voiceless. We would thus expect the CNN to use each of the cues, but only integrate F1 and voicing. However, the complex learning and network structure make it extremely difficult to predict the CNN's behavior, motivating an empirical test.

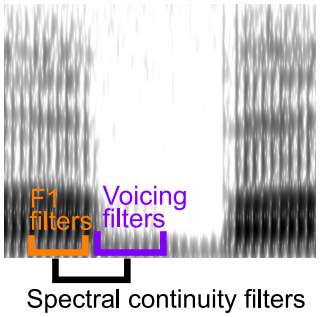


Figure 4: Rough schematic of potential types of filter a CNN would learn in categorizing voiced versus voiceless stops. If the CNN captured human-like IPP behavior, it would learn the spectral continuity filter (perhaps in addition to the F1 and voicing filters). If not, it would learn only the voicing and F1 filters.

3 Experimental Design

Using spectral continuity as a case study, I asked whether CNNs exhibit IPP-like representations, a prerequisite of implementing Yang (2019)'s account. Specifically, do they integrate pairs of cues in the same way that humans do, as found in Kingston et al. (2008)'s spectral continuity experiments? In order to test whether CNNs represent sounds with human-like IPPs, I adapted the Garner paradigm to CNNs. Because the CNN only outputs a probability of voiced vs voiceless category membership, and not all the stimuli in the Garner paradigm are designed to fall into those categories,[2] the CNN is not compatible with the discrimination task used to estimate humans' perceptual distances. Instead, following a visual perception study by Ward (2019), I calculated CNNs' perceptual distance between a pair of stimuli by extracting the network's representation for each stimulus and then computing the distance between them. These representations are the activations (weighted sum values) at the last layer of the network when the stimulus is input to the network, so they are just vectors of numbers. The distance between them could thus be calculated with cosine distance, which is proportional to the angle between them. These perceptual distances could then be interpreted as described for human perceptual distances; CNNs representing spectral continuity would be supported by perceptual warping along the IPP direction when both cues contribute to spectral continuity.

I tested two complementary, opposing hypotheses: 1) The CNN will exhibit the same pattern of human-like integration as shown in Table 1. This outcome is predicted by the CNN learning to represent the input with IPP-like sets of filters. 2) The CNN will exhibit a different, non-human-like pattern of integration from that shown in Table 1. This outcome is predicted by the CNN learning to represent the input without IPP-like sets of filters.

Each hypothesis is composite; to determine whether each cell in Table 1 is integrated or not, I had to determine whether there is a significant difference between the distances along each diagonal across the different random seed starting states. In addition, I tested for the direction of the difference; if the CNNs exhibit IPPs, then the distances will be longer along the IPP diagonal and not the opposite direction for the pairs of cues that contribute to the same IPP.

[2] Burridge & Vaux (2022) also apply a human experimental paradigm to a speech processing CNN, but their focus was on misperception of natural speech rather than discriminating arbitrary stimuli; they probed the output of their network rather than the internal representations.

4 Method

4.1 *Model* I replicated the CNN architecture of the discriminator component of Donahue et al. (2019), which forms the speech perception component of Beguš' (2020) sound change model. However, because I isolated this component from the rest of the model where its function was to compare the learner's outputs to the previous generation's speech, it required a different training task. I defined this to be the simple classification of English stops (in vowel-stop-vowel tokens) as voiced or voiceless. Additional implementation differences from Donahue et al. (2019) were removing phase shuffling (which was only relevant for the original comparison task) and reducing the audio input size, as Donahue et al.'s (2019 code was written for longer audio sequences than VCV tokens. I represented the voiced-vs-voiceless categorization with a one-hot vector (simply meaning 01 represents voiced and 10 represents voiceless) and had a softmax activation function on the final layer of the network so it could be interpreted as a probability distribution over whether the input is voiced or voiceless. The layer used for the extraction of internal representations was the second to last, where the nodes are flattened into one dimension. The code for the model, as well as the input data preprocessing involved in this experiment, is available at https://github.com/ceryshughes/CNN_Integration.

4.2 *Training data* Because my question was not about the characteristics of the input data but instead about the properties of the model and the structure of its internal representations, I controlled the input by synthesizing it with the Klatt synthesizer implementation in Praat (Klatt & Klatt, 1990 and Boersma, 2006). Synthesizing the training data allowed me to exclude any possible cues other than the ones of interest that might occur in a sample of natural speech. Each training token consisted of a sequence of a vowel, stop, and vowel (VCV) such as /ibi/ or /ipi/. There were 500 training tokens of voiced stops and 500 training tokens of voiceless stops.

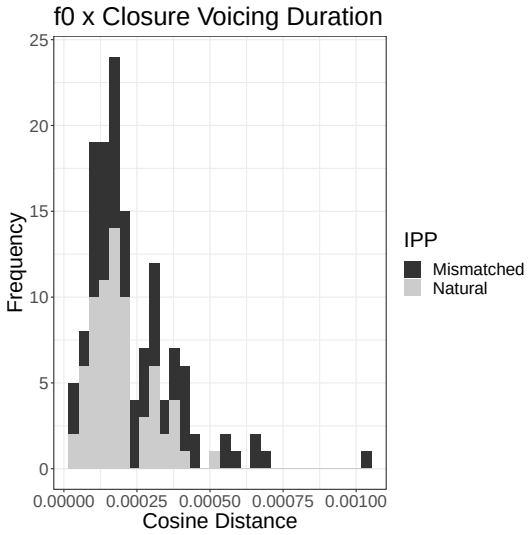
Even though the training tokens are synthesized, I set the synthesis parameters based on 321 real English vowel-stop-vowel productions collected in a lab by the MIT Speech Communication Group with three English speakers, one female and two male (2003). This experiment was conducted on specifically English stops to remain consistent with the human results under comparison, which involved English speakers (Kingston et al., 2008). I extracted f0 and F1 values both during the steady state portion of the vowel and right before the stop closure. I also extracted closure duration and voicing duration. I used the voiced and voiceless distributions from this real speech data to guide the random sampling of measurements for voiced and voiceless synthesized tokens; this way, I could to some extent mimic the distributions of the cues of interest without the noise and potential confounds in the natural speech productions. For each training datum, I synthesized the offset f0, offset F1, and closure voicing of one of the natural tokens, [3] with a small amount of Gaussian noise added so the tokens were not identical. The small sample of speakers in (MIT, 2005) did not reliably use closure duration as a cue, so I sampled the closure duration values from separate Gaussian distributions for voiced and voiceless tokens. All other synthesis parameters were held constant across all tokens.

4.3 *Training and Experimental Procedure* I trained each random seed version of the model for 10 epochs; each epoch consists of an exposure to all training examples. This number of epochs was chosen because by this point, the preliminary models all achieved over 95% accuracy on the training data. Otherwise, I followed Donahue et al.'s (2019) training procedure with Stochastic Gradient Descent, the Adam optimizer, mini batches, batch normalization, and gradient clipping. The experimental stimuli were replicated from Kingston et al. (2008) so that the CNN results would be comparable to the human results. The pairs of cues I investigated are those included in Table 1. To be more certain of the model's behavior in general (Corkery et al., 2019), I repeat the experiment (training a model and extracting its perceptual distances for the Garner paradigm) with 70 different random seeds, which determine the weight initializations.

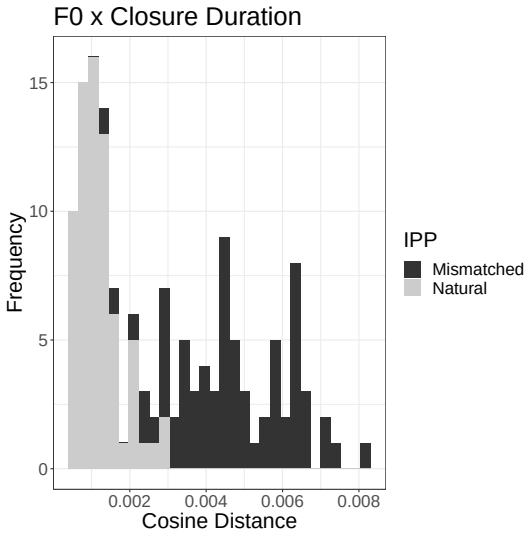
[3] Frequency measurements used Praat's pitch tracker (Boersma, 2006) through the Parselmouth package (Jadoul et al., 2018).

5 Results

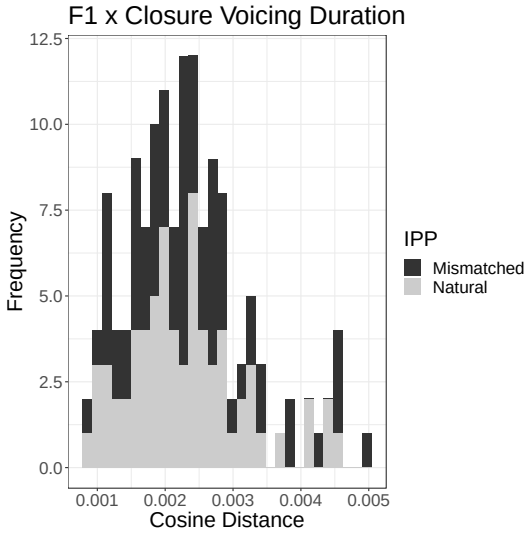
Figure 5 shows the perceptual distances between stimuli for each cue pair, contrasting the distances along the spectral continuity IPP dimension or, for closure duration, co-occurrence in the input (the "natural" dimension") and the opposite, mismatched, perpendicular dimension (the "mismatch" dimension). The overlap between distances for f0×voicing and F1×voicing would suggest a lack of integration. Surprisingly, given the human results and hypothesized IPP, the lack of overlap for f0×duration and F1×duration would suggest integration between these cue pairs.



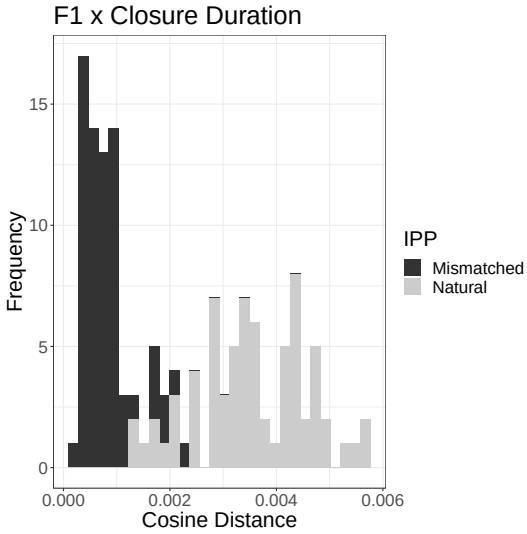
(a) The CNN distances for f0 and voicing overlap, whereas human-like results would show greater distances on the natural diagonal.



(b) The CNN distances for the mismatch diagonal are greater than the natural diagonal, whereas human-like results would show overlapping distances.



(c) The CNN distances for F1 and voicing are overlapping, whereas human-like results would show greater distances on the natural diagonal.



(d) The CNN distances for the natural diagonal are greater than the mismatch diagonal, whereas human-like results would show overlapping distances.

Figure 5: The CNN perceptual distances for each diagonal in the Garner paradigm and for each pair of cues tested.

For each cue pair, I fit a mixed effects linear regression model with R's lme4 (Bates, 2010) and lmerTest

(Kuznetsova et al., 2017) package to test whether stimuli are further apart along the combined, diagonal dimension that corresponds with an IPP or co-occurrence in the input (the "natural" dimension"). Following Kingston et al. (2008), I tested whether the perceptual distances along this diagonal are larger than the perceptual distances along the opposite, mismatched diagonal (Figure 2). The linear regression fits cosine distance (analogous to human perceptual distance) as a function of whether the stimuli fall on the "natural" dimension or the "mismatch" dimension. In other words, I tested whether the "natural" distance is larger than the "mismatch" distance on average, averaging over different random seed versions of the CNN. I also allowed the intercept to vary by random seed to account for differences in their baseline distance values (e.g. in case some random seed versions tend to have larger overall distances). In lmer notation, the regression fits (1) below.

(1) `distance ~ dimension.type +(1|Seed)`

where contrasts were defined for dimension.type as -0.5 for mismatch and 0.5 for natural. A significantly larger distance along the natural dimension, and thus integration, was found for f0×duration and F1×duration ($p < 0.001$). A more weakly significant difference was found for F1×voicing ($p = 0.0305$), but in the wrong dimension; stimuli are closer, not further apart, along the IPP diagonal than the opposite diagonal. The contrast with human behavior for one of these cue pairs, F1×voicing, is visualized in a Garner paradigm in Figure 6, which shows a perceptual stretching along the IPP dimension for humans but not the CNN. In summary, the network does not integrate the cue pairs in the same way that humans do.

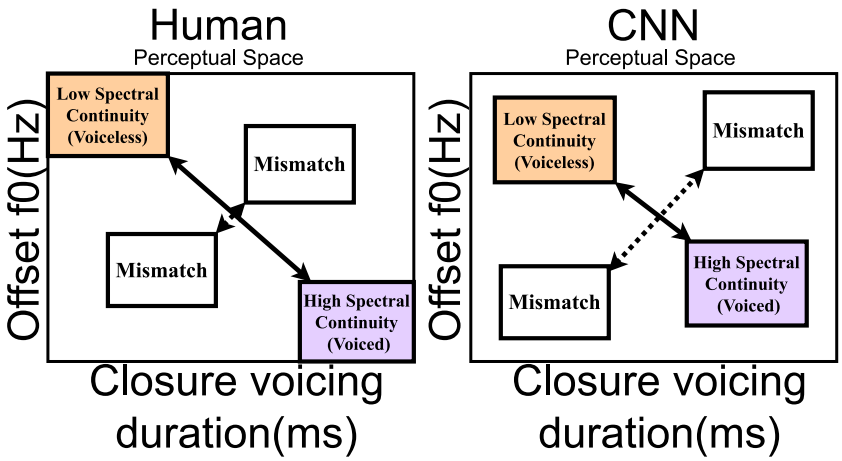


Figure 6: Visual interpretation of human and CNN perceptual spaces for one pair of cues, f0 and closure voicing.

6 Discussion and Conclusion

Given that the CNN does not exhibit the same perceptual integration behavior as humans, these results highlight a property of CNNs to be aware of when using them in models of speech. This property is an issue with using them to model Yang (2019)'s account as part of a larger neural network of sound change, indicating a direction for further research on their modification or replacement for this specific purpose. The near inverse of human-like perceptual integration behavior raises several puzzling questions about the CNN's representation of input sounds. For example, for f0 and closure duration, the CNN considers the mismatched stimuli more distant from each other, indicating it does not perceive these cues independently and perhaps has an auditory dimension corresponding to the mismatched diagonal dimension. Why would the CNN form this representation when the data it was exposed to during training followed the opposite pattern? Because its training task was to separate stops as voiced or voiceless, we could imagine error would be reduced more if the representations reflected the cue values voiced and voiceless stops actually differ on, making the network more likely to develop weights for these representations. Furthermore, F1 and closure duration was the only cue pairing the CNN integrated in accordance with the voiced/voiceless cue value pattern. Humans do not

integrate this pair of cues as they do not contribute to the same IPP, spectral continuity (Kingston et al., 2008). Perhaps CNNs' integration of cues is not driven by the joint qualitative effect of spectral continuity, but solely by the frequency of co-occurrence of cue values with each other or the categories to be learned. In designing the training data, the closure duration distribution was artificially set to be very different for voiced and voiceless stops, perhaps motivating its integration with other voiced vs voiceless cues even if it does not contribute to spectral continuity.

Although the properties of CNNs suggest they might integrate cues like humans do and provide a model of IPPs, this empirical test of their actual behavior did not support that hypothesis, at least given the particular hyperparameters (filter size and number of filters, number of layers, training procedure, etc.) examined here based on a component of Beguš' (2020) sound change model. Future work should continue the investigations further with other probing methods, such as examining feature maps (the outputs of filters for a given input, e.g. Beguš & Zhou, 2022), and examining the representations at different layers of the network. Even so, these results indicate that the task of formalizing and simulating Yang (2019)'s proposed IPP bias on contrast shift remains. Although building particular filters explicitly into a network is not common practice for CNNs, one possible standard modification to CNNs that might induce IPP-like behavior is pooling. Pooling applies some operation, such as summing or averaging, to the output of a filter over multiple regions of the input; for example, in the computer vision literature, Babenko & Lempitsky (2015) find that sum pooling is beneficial for aggregating features of images. For the spectral continuity cue, voiced stops (higher spectral continuity) may have a larger total sum of low-frequency energy before and after the closure than voiceless stops (lower spectral continuity).

In this paper, I have piloted a novel approach to better understanding neural network speech models by adapting the Garner paradigm to the internal representations of Convolutional Neural Networks (CNNs). Using this methodology, I explored CNNs' viability as a model of the combined auditory dimensions as part of an implementation of Yang (2019)'s proposal that Integrated Perceptual Properties (IPPs) bias which contrast shifts occur. I did not find support for CNNs exhibiting human-like behavior with regard to spectral continuity, highlighting the need for further research in formalizing and simulating accounts of IPPs' influence on contrast shift.

References

Babenko, Artem & Victor Lempitsky (2015). Aggregating local deep features for image retrieval. *Proceedings of the IEEE international conference on computer vision*, 1269–1277.

Bates, Douglas M. (2010). lme4: Mixed-effects modeling with r.

Beguš, Gašper & Alan Zhou (2022). Interpreting intermediate convolutional layers in unsupervised acoustic word classification. *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 8207–8211.

Beguš, Gašper (2020). Artificial sound change: Language change and deep convolutional neural networks in iterative learning. *arXiv:2011.05463 [cs, eess]* URL `http://arxiv.org/abs/2011.05463`. ArXiv: 2011.05463.

Boersma, Paul (2006). Praat: doing phonetics by computer. *http://www. praat. org/* .

Burridge, James & Bert Vaux (2022). Embedding and measurement of vowels using machine perception URL `https://lingbuzz.net/lingbuzz/006611`. LingBuzz Published In:.

Corkery, Maria, Yevgen Matusevych & Sharon Goldwater (2019). Are we there yet? encoder-decoder neural networks as cognitive models of english past tense inflection. *arXiv preprint arXiv:1906.01280* .

Donahue, Chris, Julian McAuley & Miller Puckette (2019). Adversarial audio synthesis. *arXiv:1802.04208 [cs]* URL `http://arxiv.org/abs/1802.04208`. ArXiv: 1802.04208.

Garner, WR (1953). An informational analysis of absolute judgments of loudness. *Journal of experimental psychology* 46:5, p. 373.

Hogan, John T. & Anton J. Rozsypal (1980). Evaluation of vowel duration as a cue for the voicing distinction in the following word-final consonant. *The Journal of the Acoustical Society of America* 67:5, p. 1764–1771.

Ito, Yoshifusa (1991). Representation of functions by superpositions of a step or sigmoid function and their applications to neural network theory. *Neural Networks* 4:3, p. 385–394.

Jadoul, Yannick, Bill Thompson & Bart De Boer (2018). Introducing parselmouth: A python interface to praat. *Journal of Phonetics* 71, 1–15.

Kingston, John, Randy L. Diehl, Cecilia J. Kirk & Wendy A. Castleman (2008). On the internal perceptual structure of distinctive features: The [voice] contrast. *Journal of Phonetics* 36:1, p. 28–54.

Kirby, James P (2010). *Cue selection and category restructuring in sound change*. Ph.D. thesis, The University of Chicago.

Kirby, James & Morgan Sonderegger (2015). Bias and population structure in the actuation of sound change. *arXiv:1507.04420 [physics]* URL http://arxiv.org/abs/1507.04420. ArXiv: 1507.04420.

Klatt, Dennis H. & Laura C. Klatt (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America* 87:2, p. 820–857.

Kuznetsova, Alexandra, Per B Brockhoff & Rune HB Christensen (2017). lmertest package: tests in linear mixed effects models. *Journal of statistical software* 82, 1–26.

Lisker, Leigh (1957). Closure duration and the intervocalic voiced-voiceless distinction in english. *Language* 33:1, p. 42–49.

MIT, Speech Communication Group (2005). Laff vcv syllables, vcv syllables recorded from two male and one female speakers.

O'Shea, Keiron & Ryan Nash (2015). An introduction to convolutional neural networks :arXiv:1511.08458, URL http://arxiv.org/abs/1511.08458. Number: arXiv:1511.08458 arXiv:1511.08458 [cs].

Palaz, Dimitri, Ronan Collobert & Mathew Magimai Doss (2013). Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks :arXiv:1304.1018, URL http://arxiv.org/abs/1304.1018. Number: arXiv:1304.1018 arXiv:1304.1018 [cs].

Ruder, Sebastian (2017). An overview of gradient descent optimization algorithms :arXiv:1609.04747, URL http://arxiv.org/abs/1609.04747. Number: arXiv:1609.04747 arXiv:1609.04747 [cs].

Ward, Emily J (2019). Exploring perceptual illusions in deep neural networks. *bioRxiv* p. 687905.

Wedel, Andrew B (2006). Exemplar models, evolution and language change. *The Linguistic Review* 23:3, URL https://www.degruyter.com/document/doi/10.1515/TLR.2006.010/html.

Yang, Meng (2019). *Cue Integration and Contrast Shifts: Experimental and Typological Studies*. University of California, Los Angeles.