JASA ARTICLE

Comparing human and machine's use of coarticulatory vowel nasalization for linguistic classification[a)]

Georgia Zellou,[1,b)] Lila Kim,[2] and Cédric Gendrot[2]
[1]*Phonetics Lab, Linguistics Department, University of California-Davis, Davis, California 95616, USA*
[2]*Laboratoire de Phonétique et Phonologie, Université Sorbonne Nouvelle, UMR 7018 CNRS, Paris, France*

ABSTRACT:
Anticipatory coarticulation is a highly informative cue to upcoming linguistic information: listeners can identify that the word is *ben* and not *bed* by hearing the vowel alone. The present study compares the relative performances of human listeners and a self-supervised pre-trained speech model (wav2vec 2.0) in the use of nasal coarticulation to classify vowels. Stimuli consisted of nasalized (from CVN words) and non-nasalized (from CVCs) American English vowels produced by 60 humans and generated in 36 TTS voices. wav2vec 2.0 performance is similar to human listener performance, in aggregate. Broken down by vowel type: both wav2vec 2.0 and listeners perform higher for non-nasalized vowels produced naturally by humans. However, wav2vec 2.0 shows higher correct classification performance for nasalized vowels, than for non-nasalized vowels, for TTS voices. Speaker-level patterns reveal that listeners' use of coarticulation is highly variable across talkers. wav2vec 2.0 also shows cross-talker variability in performance. Analyses also reveal differences in the use of multiple acoustic cues in nasalized vowel classifications across listeners and the wav2vec 2.0. Findings have implications for understanding how coarticulatory variation is used in speech perception. Results also can provide insight into how neural systems learn to attend to the unique acoustic features of coarticulation.
https://doi.org/10.1121/10.0027932
(Received 25 September 2023; revised 24 June 2024; accepted 27 June 2024; published online 16 July 2024)
[Editor: Jody Kreiman] Pages: 489–502

I. INTRODUCTION

There is an enormous amount of variation in speech. One large source of variability is due to coarticulation, or the acoustic effects of overlapping articulations from adjacent sounds in the speech signal. Coarticulation is a natural and essential property of speech dynamics because it permits speech to be produced in a fluid and intelligible manner (e.g., Fowler and Saltzman, 1993). For instance, vowels before nasal consonants are produced with some amount of nasality since the velum-lowering gesture from the following sound begins early and overlaps with the vowel. Since coarticulation is a natural and systematic feature of speech, human listeners are highly sensitive to its variation: they use coarticulation to predict upcoming sounds in order to more efficiently comprehend the speaker's message (Fowler, 1984). For instance, looking at the example of anticipatory coarticulatory nasality in English from the side of a perceiver, listeners can identify that a word they are hearing is *bun*, and not *bud* based on hearing nasal coarticulation on the vowel only. This indicates that coarticulatory nasalization provides cues to lexical identity even before the final consonant is pronounced (Ali *et al.*, 1971; Ohala and Ohala, 1995; Beddor *et al.*, 2013).

Yet, there are many open questions about how the use of anticipatory coarticulatory cues present on vowels to categorize speech might vary across different types of contexts, speakers, and comprehenders. The current study asks to what extent machine speech recognition systems use anticipatory coarticulation present on vowels to make classifications of speech in ways similar to, or different from, human listeners. We investigate this by testing classifications by a widely used Self-Supervised Pretrained speech model (wav2vec 2.0 [W2V2]) (Baevski *et al.*, 2020) of American English nasalized and non-nasalized vowels (spliced out of context and presented in isolation) produced by 96 distinct voices. Perceptual responses to these vowels by human listeners are also examined, and we compare listener and machine classification performance.

Since coarticulation is systematic and provides predictive information, one hypothesis is that the machine recognition system will use nasal coarticulatory cues to accurately classify vowels in ways similar to how human listeners perform. Alternatively, this speech model might outperform humans in use of coarticulation to perform linguistic classifications. More specifically, human listeners typically perceive vowel nasalization in the context of a nasal consonant and at least partially attribute the acoustic effects of coarticulation to its source (Beddor and Krakow, 1999; Zellou, 2017); so hearing vowels spliced from their appropriate coda contexts might sound phonetically shifted or otherwise

[a)]This paper is part of a special issue on Acoustic Cue–Based Perception and Production of Speech by Humans and Machines.
[b)]Email: gzellou@ucdavis.edu

odd to listeners, leading to lower performance. Meanwhile, the speech model is more likely to interpret coarticulatory-acoustic cues as inherent to the vowel. Speech models, such as W2V2, have been primarily designed to perform automatic speech recognition (ASR) but it has also proved efficient for other tasks (Kunze *et al.*, 2017; Lian *et al.*, 2018). While global, human-level performance is the goal in developing ASR systems, investigating whether there are context-specific asymmetries in speech recognition system performance is one step to achieving fine-grained sensitivities to speech variation that parallel human listeners.

Another open question is whether speech models handle cross-speaker variation in the same ways that human listeners do. Coarticulation has been shown to vary extensively across individual speakers (Beddor and Krakow, 1999; Zellou, 2017; Yu and Zellou, 2019). Therefore, we also examine cross-speaker variation in the performance of W2V2, relative to human listeners, in classifications of nasal-coarticulated vowels. Investigating how speech models designed for ASR systems classify coarticulated speech is practically important for understanding how deep learning handles within- and across-speaker variation.

Finally, we also compare human vs machine performance in classification of coarticulated vowels across naturally produced speech and synthetic voices (i.e., text-to-speech [TTS] generated from those openly available by several companies). It is a new digital era: the number of spoken language interactions between machines and humans are common and increasing every day due to voice-activated AI devices in the home (Ammari *et al.*, 2019). Thus, it is relevant and apt to ask how speech perception might vary across human–human vs human–device (and, even device–device) interactions.

A. Coarticulation in speech recognition

There is a growing body of work investigating how listeners use coarticulation during speech perception. Coarticulation is a natural and systematic property of speech. Thus, it has been proposed that listeners attend to coarticulatory details to make predictions about, or more efficiently process, upcoming linguistic information (Lahiri and Marslen-Wilson, 1991; Beddor, 2009; Beddor *et al.*, 2013; Scarborough and Zellou, 2013). Since vowel nasality in English always and only occurs in the context of a nasal consonant, the presence of nasalization provides reinforcing evidence about the identity of an upcoming nasal coda. Indeed, many past studies have shown that American English listeners use coarticulatory information as soon as it is available to identify a lexical item, supporting this hypothesis (Beddor *et al.*, 2013; Zellou, 2022; Zellou *et al.*, 2023).

Moreover, people are now regularly talking with machines and rely on them to accurately understand their speech; thus, it is relevant to ask whether technological systems rely on acoustic features to classify words in the same ways that human listeners do. Do ASR systems use coarticulatory information to classify speech similarly to how human listeners use it? Neural networks are machine learning tools that use high-dimensional spaces to perform classifications. Speech recognition systems using artificial neural networks have shown drastic improvements in recent years, particularly when they are designed and trained to take into account coarticulatory patterns (Kanthak and Ney, 2002; Ansary and Salehi, 2004; Mun *et al.*, 2022). One understudied aspect of ASR performance is coarticulatory variation (Liu *et al.*, 2020). While prior work has achieved some success in synthesizing visual output, such as facial animation (Deng *et al.*, 2006) and videos of talking speakers (Liu *et al.*, 2020) by testing ASR performance on speech coarticulation, many open questions remain. We ask whether W2V2 uses coarticulatory cues to correctly identify that a vowel is either nasalized or non-nasalized in ways that parallel how human listeners perform classifying the same set of stimuli (i.e., analogous to a human identifying the word is *bent* from the vowel alone in Beddor *et al.*, 2013).

More specifically, we ask whether there are the same patterns of linguistic classifications across nasalized and non-nasalized vowel types for a machine speech recognition model and human perceivers. Overall comparison with human-level accuracy is the gold standard of evaluating the performance of artificial models. However, such global comparisons might be masking underlying differences in how human perceivers and neural networks categorize different types of speech sounds. Most broadly, understanding what specific types of sound patterns that are difficult for machine recognizers can be informative to improve their performance with further fine-tuning.

The current study focuses specifically on coarticulatory vowel nasality. For instance, while coarticulation is useful for listeners, there is also evidence that orality (or, the absence of nasalization) is a strong cue that listeners use to identify that the upcoming sound is unequivocally not a nasal consonant. Moreover, studies of English, Korean, and Mandarin show that there is a huge amount of variation across and within languages in degree of nasal coarticulation (Scarborough and Zellou, 2013; Cho *et al.*, 2017; Jang *et al.*, 2018; Li *et al.*, 2020). Since nasal coarticulation varies so much, listeners might be sensitive to the fact that orality is a more reliable cue than nasality in a language where vowel nasality occurs only contextually (i.e., due to coarticulation from nasal consonants). Indeed, there is already some evidence that listeners more systematically use orality as a cue, compared to nasality, in classifying isolated vowels. For instance, Zellou (2022) investigated listeners' perception of coarticulatory variation across a range of speakers, using a coda-completion task, the same perception task we use in the current study. Overall, that study found that listeners were above chance at identifying nasal vs oral codas from hearing the vowel alone. However, there is an asymmetry in performance for oral vs nasalized vowels: correct identification of the coda was lower for nasal-coarticulated vowels, compared to oral vowels. Also, relative to oral vowels, nasalized vowels in English are more ambiguous, more

likely to be misperceived in isolation, and less likely to be accurately discriminated (Wright, 1986; Beddor, 1993; Zellou, 2017). Therefore, together, the empirical evidence suggests that listeners can use nasal coarticulation to predict that an upcoming sound is a nasal coda, yet oral vowels provide more reliable cues to coda identity in English. The use of coarticulation supports models of speech perception where listeners use available systematic details in the acoustic signal to efficiently process speech (Beddor, 2009). Yet, the asymmetry between oral and nasalized vowels is also informative. There are two possible explanations for this pattern. For one, it could be because English listeners' experience with nasality as only a coarticulatory cue (i.e., it only occurs in the presence of a nasal consonant) means they are just more sensitive to orality. Another possibility is that coarticulation is just acoustically more variable across words and speakers, since it is non-contrastive. For instance, recent work has found cross-speaker variation in the weighting of multiple acoustic features associated with coarticulatory nasality (Zellou and Cohn, 2024). This might mean that nasalization is a slightly less robust cue than orality for listeners.

Comparing human and machine classifications of nasalized and non-nasalized vowels in American English can tease apart these possibilities. We investigate whether neural network speech classifiers show the same classification patterns as human listeners. On the one hand, if neural networks are trained on natural American English speech input, we might predict that it will mimic the patterns of categorizations displayed by human listeners: it will use coarticulation to categorize vowels as nasalized, but overall performance will be highest for oral vowels. On the other hand, since coarticulation is a systematic property of vowels and the neural network could attend to this type of variation differently than humans, the W2V2 model might show sensitivity to coarticulatory details to a greater extent than human listeners. In either outcome, comparison of W2V2's and listeners' classification of nasalized and non-nasalized vowels can provide insight into the role of coarticulation in speech perception, either as a universal acoustic clue or something that perceivers are sensitive to based on their language-specific experience.

B. Cross-speaker coarticulatory variation

Another question addressed in the current study is whether across-speaker patterns of nasalized vowel classifications are similar for the W2V2 speech model as it is for human listeners. There is a great deal of variation within and across speakers in its realization. For instance, the degree of nasal coarticulation (realized as vowel nasality produced in words with CVN structure) varies greatly across American English speakers (Zellou, 2017, 2022). An examination of the cross-speaker differences in Zellou (2022) revealed a large amount of variability in how individual talkers implement coarticulatory vowel nasalization and thus how their coarticulated vowels provided predictive cues. For instance, some talkers produce coarticulated vowels that were systematically identified as originating from CVN words, while other speakers produced coarticulated vowels that provided little cues to coda nasalization since they were exclusively categorized as coming from CVC items. Oral vowels, on the other hand, provided consistent cues for listeners to coda orality across speakers.

Do machine recognition systems show the same cross-talker variability in speech classifications as human listeners? On the one hand, a major goal in the development of speech recognition technology is a neural network that displays minimal cross-speaker variation (e.g., Huang, 1992). Thus, investigating whether an artificial neural network can achieve talker-independent ability to use coarticulation in speech classification would be one step toward this practical goal. On the other hand, talker-independence is not the reality for speech perception by human listeners. For instance, some speakers are just more intelligible than other speakers, in ways that are systematic (e.g., by dialect in Clopper and Bradlow, 2008) or idiosyncratic (Bradlow *et al.*, 1996). Thus, if W2V2 displays talker-dependent classification of vowels, it would be useful to know if these patterns mirror those displayed by human listeners, or whether they vary in some systematic, but meaningful, way.

1. Comparing naturally produced vs TTS voices

As mentioned above, understanding how speech perception varies across naturally produced speech and TTS is relevant in an era where human-machine conversational interactions are a daily occurrence. Furthermore, ASR systems are increasingly being trained on TTS speech, as well (Fazel *et al.*, 2021). Hence, the design approach of the current study is 2×2: comparing the use of coarticulation for classification in naturally produced speech vs TTS by human listeners and a machine speech recognizer.

There is some prior work investigating acoustic properties of coarticulation in industry-generated TTS and its effect on listener perception. For instance, Ferenc Segedin *et al.* (2019) found that acoustic nasality for TTS voices generated for Amazon's Alexa devices was more phonetically ambiguous than naturally produced human speech. Zellou *et al.* (2021) compared acoustic nasality in CVN contexts across different types of TTS and found a larger difference in acoustic nasality between oral and nasalized vowels in concatenative TTS than neural TTS. They also found that TTS voices containing larger differences in coarticulatory nasalization between oral and nasalized vowels (as in the concatenative TTS voices) correlated with higher listener discrimination. These studies suggest that TTS contains more ambiguous coarticulatory cues but, at the same time, like for naturally produced speech, there is variation across TTS voices and that listeners track and use the patterns of coarticulation in synthesized speech to comprehend the speech signal. There is no prior work, to our knowledge, comparing use of coarticulation for machine classification of naturally produced speech vs TTS speech.

C. Acoustic cue use in perception of vowel nasalization

Does an ASR system use the multiple cues for coarticulatory information in the acoustic signal to categorize speech in the same way as human listeners? There are multiple acoustic properties associated with vowel nasalization. Vowel nasalization manifests itself acoustically as dampening of the amplitude of the first formant spectral peak (A1) and the introduction of a low-frequency nasal resonance, which increases in amplitude as degree of nasalization increases (P0). One acoustic measure for nasalization is quantified by subtracting one measure from the other, A1-P0 (Chen, 1997; Styler, 2017). There are additional acoustic differences across nasalized and non-nasalized vowels. The frequency of F1, which is typically modulated by tongue height, has been shown to be different across nasalized and non-nasalized vowel counterparts (Carignan, 2017). Nasalized vowels also display wider F1 bandwidths than oral vowels (Hawkins and Stevens, 1985; Stevens, 2000; Styler, 2017). Voice quality can also differ across nasalized vowels, with nasalization correlating with more breathy phonation metrics (Garellek *et al.*, 2016) and smaller spectral tilt values (e.g., measured as A3-P0 in Styler, 2017). Much prior work has shown that listeners use these acoustic features to some extent when perceiving vowel nasalization (Beddor *et al.*, 1986; Krakow *et al.*, 1988; Wright, 1986).

Thus, another research question addressed in the current study is whether similar acoustic features are used (and, used to the same extent) by listeners and artificial learning models in correct categorizations of nasal-coarticulated vowels. While comparing overall perception patterns across human perceivers and W2V2 can reveal fundamental similarities or differences in how they categorize coarticulated speech, we can also examine whether there are differences in the use of specific acoustic features in making these classifications across people and machine learning models.

Pruthi and Espy-Wilson (2007) trained a support vector machine classifier on multiple acoustic cues to vowel nasalization, including F1 frequency, F1 bandwidth, A1-P0, and spectral tilt and found high levels of accuracy in categorizations of vowels extracted from natural speech corpora. However, no prior work, to our knowledge, has compared relative acoustic cue weightings across human listeners and W2V2 models trained on speech. We ask whether there are differences in phonetic cue weighting across human listeners and W2V2 by relating categorization patterns to a set of acoustic features. If W2V2 uses the same acoustic features as human listeners to classify nasalized vowels, there will be no difference in relative cue weightings of these properties across perception and classification data for nasalized vowels. On the other hand, differences in how these acoustic properties predict human and machine classifications of nasalized vowels will indicate that the neural network uses spectral features of vowel nasalization differently than listeners.

D. Current study

In the current study, we investigate the role of nasal coarticulation by two different types of comprehenders (human listeners and a W2V2 pre-trained speech model for ASR) in two different types of voices (human speakers and TTS voices). First, we elicited CVN and CVC words by 60 human speakers and generated the same set of words from 36 distinct TTS voices. Next, we played the spliced CV portions of the words to listeners who performed a word completion task (i.e., "is the word *ben* or *bed*?"). Then, we extracted the acoustic features of each stimulus using a W2V2-large model and ran a deep neural network to classify nasalized and non-nasalized vowels. The aim of this work is to address three basic research questions. (1) How do classification patterns of nasal-coarticulated and non-nasalized vowels vary across human listeners and automatic speech classifiers. (2) Do cross-speaker classification patterns, across a range of different talkers including naturally produced human speech and TTS voices, vary in similar ways across listeners and W2V2? (3) Do human listeners and W2V2 use the multiple acoustic features of nasalized vowels in similar ways to classify coarticulated vowels?

We compare human listeners and a state-of-the art speech recognition model on use of coarticulation in linguistic classification. Coarticulation is an important acoustic feature that is quite variable across speakers and contexts. Thus, we ask how good speech models are at using this feature to recognize language (i.e., compared to the standard reference, which is human listener performance). We selected W2V2 because it is a state-of-the art speech model that is used by lots of researchers and language technology engineers. Our findings can be informative about how the most advanced speech recognition models handle variation in coarticulatory cues. They can also shed light on how speech recognition models work (e.g., how it weighs various acoustic cues).

II. EXPERIMENT 1: PERCEPTION OF HUMAN SPEAKERS AND TTS VOICES

A. Methods

1. Target words

Target words included 5 sets of CVC and CVN minimal pairs containing non-high vowels (/ɑ/, /æ/, /ʌ/, /ɛ/, /ow/): *bod, bon, bad, ban, bud, bun, bed, ben, bode, bone*. We focus on non-high vowels following prior work on nasal coarticulatory patterns in American English (Beddor and Krakow, 1999; Zellou, 2022.)

Each target word was placed in a carrier phrase: "__, the word is __."

2. Human speakers

The speakers were 60 native speakers of American English (49 female, 11 male, age range = 18–33 years old,

average age = 19.6 years old; all reported to be born and raised in California), recruited through the University of California-Davis subject pool. Participants received course credit for their participation. The elicitation was conducted over a single session, lasting approximately 30 min, in a sound-attenuating booth at the University of California-Davis Phonetics Lab. Recordings were made using a Shure WH20 XLR head-mounted microphone and digitally sampled at a 44.1-kHz sampling rate.

3. TTS voices

The same set of target words in carrier sentences was generated in 36 American English TTS voices available through three different platforms. Nine Siri voices' productions were generated using the command line on an Apple computer (OSX 10.13.6) and changing the Siri voice setting in system preferences (7 female: Agnes, Allison, Ava, Princess, Samantha, Susan, Victoria; 2 male: Alex, Fred). Six Amazon Polly voices' productions were generated using the AWS Polly online console (4 female: Joanna, Kendra, Kimberly, Salli; 2 male: Joey, Matthew). While the Amazon Polly voices are only available in 6 adult speakers, we generated all items in the two different TTS generation techniques: neural (generated via neural speech synthesis, the most naturalistic) and "standard" (generation via parametric speech synthesis, the more robotic type), resulting in 12 distinct voices. We also generated the productions in 15 voices available through the Microsoft Azure online console (9 female: Amber, Aria, Ashley, Cora, Elizabeth, Jenny, Michelle, Monica, Sara; 6 male: Brandon, Christopher, Davis, Eric, Guy, Jacob), all generated in the default "general" speaking style and with the speaking speed and pitch values at their default settings. The Microsoft Azure voices are generated via Neural TTS speech synthesis. Apple announced in 2019 that the original version of Siri (i.e., the "Samantha" voice) was originally generated using concatenative speech synthesis via unit selection, but the new version will be generated using Neural TTS. Information about the TTS method used to synthesize the other Apple voices was not publicly available.

4. Stimulus preparation

The words were extracted from the carrier phrases and prepped for presentation in the perception experiment. Stimuli for the perception experiment consisted of CV syllables spliced from the CVC and CVN word productions by the 60 human speakers and the 39 TTS voices. Following the method used by Ohala and Ohala (1995), the syllables were then gated into white noise, at a level 5 dB less than the peak intensity of the vowel. In other words, the last 5 ms of the vowel overlapped with noise, which then continued for 150 ms after the vowel ended; the purpose of this was to avoid a stop-bias that might occur with the syllables abruptly ending in silence.

5. Participants

Participants were 166 native American English speakers (mean age: 19.4 years old), recruited from the University of California-Davis Psychology Subject Pool. They received course credit for participation. None of the participants reported any speech or hearing impairments. This study was approved by the University of California-Davis Institutional Review Board, and all participants completed informed consent.

6. Procedure

The experiment, conducted online using Qualtrics, began with a sound calibration procedure: participants heard one sentence produced by each speaker (not used in experimental trials), presented in silence at 60 dB, and were asked to identify the sentence from three multiple choice options, each containing a phonologically close target word (e.g., they hear "Bill asked about the host" and are given options for sentences ending in host, toast, coast). After, they were instructed to not adjust their sound levels again during the experiment. Participants completed a word completion task. The task was designed following the paradigm used by Ohala and Ohala (1995). On a given trial, listeners heard one of the speakers produce either a CV syllable (truncated from a CVC word) or a CṼ syllable (truncated from a CVN word) gated into noise. Then, listeners select which one of two minimal pair choices (either a CVC or CVN word, corresponding to the minimal pair option for that syllable) the syllable was originally taken from.

Eight experimental lists were generated containing all of the stimulus items from 12 talkers (randomly selected). Each list contained only one type of speaker: 5 lists contained only human speakers, 3 lists contained only TTS speakers. Participants were randomly assigned to one of the lists, in either the human speaker condition ($n = 94$ participants) or the TTS voices condition ($n = 72$ participants). [Note: The perception data on the human speaker stimuli was previously analyzed in Zellou (2022). The present study includes this subset along with novel data from the perception of the TTS voices.]

In order to keep the experiment a reasonable length, only the first production of each word from the frame sentence was used in the perception study.

B. Results

Word identification accuracy on each trial was coded binomially (1 = correct intended word identification, 0 = incorrect). Listeners' global mean performance in identifying the coda correctly based on the vowel information alone is 78.5%. Figure 1 displays the aggregated word identification accuracy results by vowel type and speaker type. For the human speakers, aggregated lexical identification of items across speakers is above chance performance for both non-nasalized (93.1%) and nasalized (71.7%) vowels, with an overall accuracy rate of 82.4%. For the TTS voices, performance is very high for the non-nasalized vowels (91.5%)

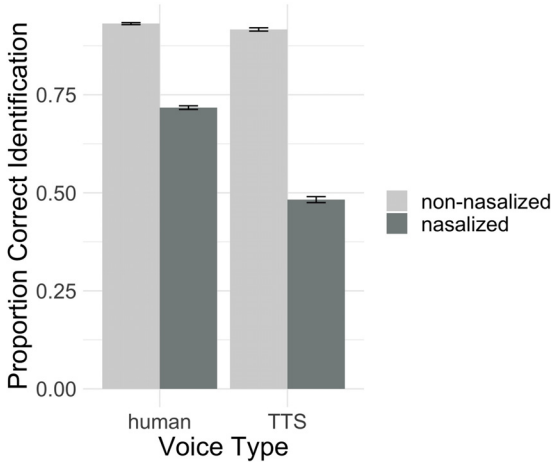

FIG. 1. Accurate word identification rates by listeners for nasalized and non-nasalized vowels for naturally produced stimuli (human speakers) and TTS stimuli.

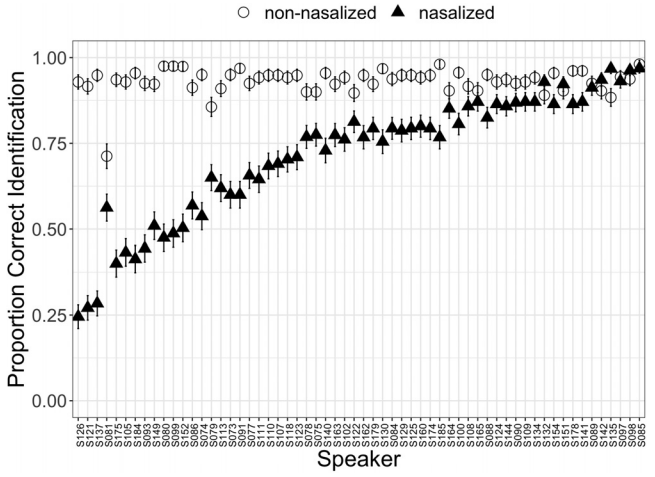

FIG. 2. Mean accurate word identification rates (and standard errors) by listeners for nasalized and non-nasalized vowels for each of the human speakers.

but close to chance for the nasalized vowels (47.6%). The overall correct identification for TTS voices was 69.9%.

The accuracy data were modeled using a mixed-effects logistic regression with *lme4* R package (Bates *et al.*, 2015). Estimates for p values were computed using the *lmerTest* package (Kuznetsova *et al.*, 2015). Fixed effects included vowel type (nasalized vs non-nasalized), speaker type (TTS vs human), and the two-way interaction. Random effects included by-listener and by-speaker random intercepts as well as by-listener random slopes for vowel type and by-speaker random slopes for vowel type. Contrasts were sum-coded.

Table I provides the summary statistics for the model. There was an effect of vowel type: overall, listeners are less accurate identifying the original lexical item for nasalized vowels compared to oral vowels. There was an effect of speaker type indicating that listeners are more accurate at coda-identifications from the vowel when produced by human voices than TTS voices. There was also an interaction between vowel type and speaker type. As seen in Fig. 1, the difference in accuracy across TTS and human voices for nasalized vowels is large, relative to that for oral vowels:

TABLE I. Summary statistics from the logistic mixed effects model run on word identification accuracy.

	Coefficient	SE	z value	p value
Intercept	1.78	0.07	24.2	<0.001
Vowel type (nasalized)	–1.24	0.08	–14.73	<0.001
Speaker type (human)	0.39	0.07	5.42	<0.001
Vowel type × speaker type	0.25	0.08	2.96	<0.01

Random effects	Variance	SD
Listener (intercept)	0.28	0.53
Vowel type	0.38	0.61
Speaker (intercept)	0.27	0.52
Vowel type	0.37	0.60

Number of observations ($n = 27\,440$), listeners ($n = 166$), speakers ($n = 96$).

listeners perform better at word identifications for nasalized vowels for human voices than for TTS voices, while the differences across voice types for non-nasalized vowels were similar.

Table I also reports the variance and standard deviation for the random effects of the model. Notably, there is a large amount of variation in listeners' ability to correctly identify words across vowel types for individual speakers as there is for speaker types (human vs TTS). To further investigate this, Fig. 2 (for human speakers) and Fig. 3 (for TTS voices) plot mean correct listener identifications by vowel type for each speaker. As seen, there is variation in accurate word identification across speakers for both human and TTS voices. However, as seen in both figures, the cross-speaker variation is largest for nasalized vowels. In other words, correct identification of the word for nasalized vowels, specifically, is highly speaker-dependent.

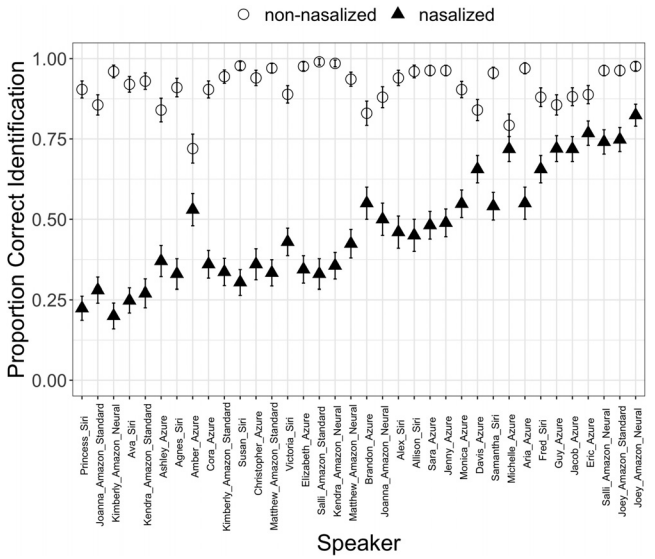

FIG. 3. Mean accurate word identification rates (and standard errors) by listeners for nasalized and non-nasalized vowels for each of the TTS voices.

https://doi.org/10.1121/10.0027932

III. EXPERIMENT 2: W2V2 CLASSIFICATION OF HUMAN AND TTS VOICES

A. Methods: W2V2

An attention mechanism is integrated into the W2V2 model, which is primarily utilized for speech recognition tasks (Hinton *et al.*, 2012). Its architecture comprises three key components: a convolutional encoder for processing raw audio signals and extracting speech representations, a quantization module for discretizing these latent representations, and an attention mechanism known as *Transformer* (Vaswani *et al.*, 2017). The latter allows the extraction of context vectors by capturing information across the entire audio sequence, thereby facilitating interactions among various latent representations. As a self-supervised training model, W2V2 does not require a large volume of manually annotated data to perform a task, unlike other deep learning models (Lee *et al.*, 2021; Radford *et al.*, 2023). It undergoes pre-training on a substantial corpus of unannotated audio (53 000 h) and can subsequently be fine-tuned with less labeled data for specific tasks (Baevski *et al.*, 2020). Our utilization of the W2V2 model follows the "feature probing" approach (Guillaume *et al.*, 2023; Triantafyllopoulos *et al.*, 2022; Shah *et al.*, 2021; Ma *et al.*, 2021), whereby it serves as a feature extractor from audio data, without fine-tuning the model on annotated data. The encoded audio was then used as input to a multi-layer perceptron classifier to evaluate the presence of nasal coarticulation in speech.

The model was first trained on three English corpora: the Buckeye corpus of conversational American English speech (Pitt *et al.*, 2005), the Timit corpus (Garofolo, 1993), and the UCLA corpus (Keating *et al.*, 2021). Two categories were considered: "nasal" with one of the five vowels (/ɑ/, /æ/, /ʌ/, /ɛ/, /ow/; target words provided in II.A.1) followed by a nasal consonant (/m/ or /n/) vs "non-nasal" with one of the five vowels not followed by a nasal consonant. A total of 54 232 occurrences (27 116 nasal and 27 116 non-nasal) were used for training. For the training to take place, CVN and CVC sequences were extracted from natural speech with a Praat script with a rectangular window in order to include phonetic contexts. The extracted raw waveforms were encoded using the open-source pre-trained model W2V2-large. This feature extractor allows the construction of a 1024-dimensional vector for the entire provided audio signal. Pasad *et al.* (2021) looked for linguistic and acoustic information present in the different transformer layers of the W2V2 model and compared this information between layers. They found that phonetic identity is most represented in the 11th, 18th, and 19th layers of the W2V2-large model, hence our choice of the 18th layer in this work. Our multi-layer perceptron classifier consisted of 12 fully connected layers (with respectively 1024, 1024, 512, 512, 256, 256, 128, 128, 64, 64, 32, and 16 neurons) followed by a dropout layer to avoid overfitting, and a ReLu activation function was applied in each layer. At the end of the classifier, the classification layer was initialized with a sigmoïd activation with a binary cross-entropy loss function. The classification algorithm classifies into two distinct categories: nasalized and non-nasalized.

To implement and train a network, we used Keras neural network library. AdamW optimization was applied as a stochastic gradient descent method with a learning rate of 0.000125 and a decay of 0.01. The batch size was 128, and the number of epochs was fixed to 150 with early stopping strategy.

The model was tested using the same stimuli described in Sec. II A. It is important to note that the model was trained on CVC or CVN structures and then evaluated on CV stimuli extracted from CVC/CVN items (without the noise added at the end, as described in Sec. II A 4). The vector representations of these stimuli were generated using the same procedure as the training set, utilizing a W2V2-large model. The evaluation of the classifier was conducted using the same words that were perceived by the participants in the perception test, totaling 600 stimuli (300 nasalized and 300 oral) produced by human voices and 320 stimuli (180 nasalized and 180 oral) generated by TTS voices.

B. Results

We coded each token as classified as the nasal or non-nasal category based on the assigned probability value by W2V2: a probability greater than 0.5 of belonging to the nasal category is coded as a nasal classification; otherwise, the vowel is classified as non-nasal. Classifications for each item were coded binomially for being correct (=1) or incorrect (=0).

The overall performance of the W2V2 in correctly classifying the identity of the coda based on the vowel information alone was 80.8%. The mean percentage correct classification rates for nasal-coarticulated and non-nasal vowels across human and TTS voices are provided in Fig. 4. The overall result for the human voices was 87% accuracy. For the human voices, W2V2 correctly classified 80.7% of nasalized vowels and 93.3% of non-nasalized vowels. The overall result for TTS voices was 70.5%; 89.4% correct

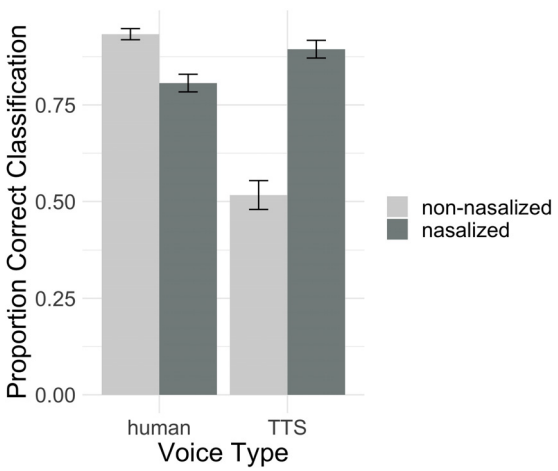

FIG. 4. Correct classification rates by W2V2 for nasalized and non-nasalized vowels for naturally produced stimuli (human speakers) and TTS stimuli.

TABLE II. Summary statistics from the logistic mixed effects model output for correct classification rates by W2V2.

	Coefficient	SE	z value	p value
Intercept	2.11	0.20	10.30	<0.001
Vowel type (nasalized)	0.65	0.21	3.05	<0.01
Speaker type (human)	0.44	0.16	2.74	<0.01
Vowel type × speaker type	–0.94	0.17	–5.55	<0.001
Random effects	Variance			SD
Speaker (intercept)	0.83			0.91
Vowel type	1.12			1.06
Number of observations ($n = 960$), speakers ($n = 96$).				

classification of nasalized vowels and 51.7% of non-nasalized vowels.

The correct classifications data were modeled using a mixed-effects logistic regression with *lme4*. Fixed effects included vowel type (nasalized vs non-nasalized), speaker type (TTS vs human), and their interaction. Random effects structure consisted of by-speaker random intercepts and by-speaker random slopes for vowel type. Contrasts were sum-coded.

Table II provides the summary statistics from the W2V2 classifications logistic model. There was an effect of vowel type: overall, W2V2 is better at classifying the nasalized vowels, compared to the non-nasalized vowels. There was an effect of speaker type: W2V2 performs better for human voices than for TTS voices. There was also an interaction between vowel type and speaker type, which is illustrated in Fig. 4: while W2V2 performs higher for non-nasalized vowels than nasalized vowels in human voices, the lowest rate of correct classifications is for the non-nasalized vowels produced by the TTS voices.

Similar to what was observed for the perception data, the random effects for W2V2 data's model reveal large variance for the by-speaker random slope for vowel type. Figures 5 and 6 display mean correct classifications by

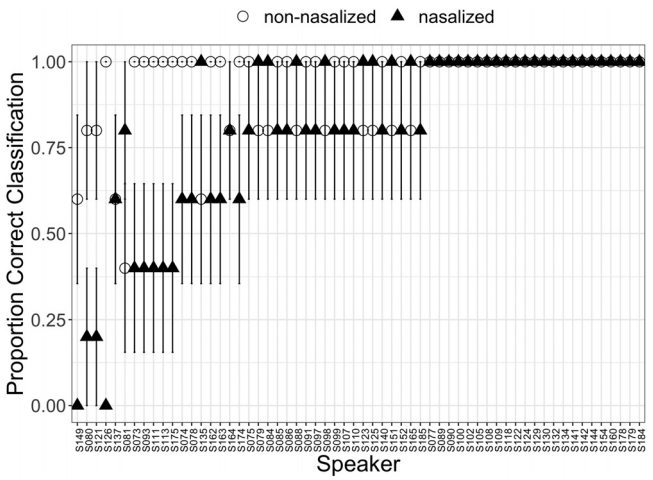

FIG. 5. Mean correct classification rates by W2V2 for nasalized and non-nasalized vowels for each of the human speakers.

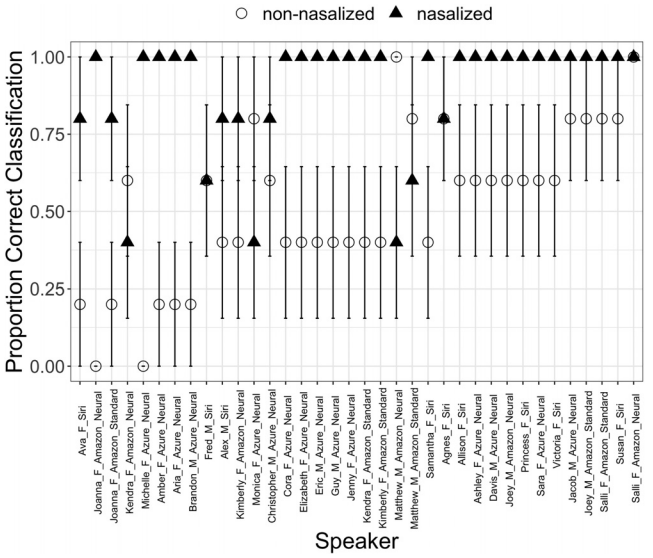

FIG. 6. Mean correct classification rates by W2V2 for nasalized and non-nasalized vowels for each of the TTS voices.

W2V2 for each of the human speakers and each of the TTS voices, respectively. As seen, there is wide variation in overall classification accuracy across speakers. For human speakers, like that seen for human listeners, the variation is largest for the nasalized vowels; yet, in contrast, the variation for the TTS voices varies greatly across the non-nasalized vowels.

IV. INVESTIGATING USE OF ACOUSTIC CUES BY LISTENERS AND W2V2

Our final research question is whether human listeners and W2V2 use acoustic features of nasalized vowels in similar ways when making linguistic classifications. We focus on nasalized vowels only in this analysis since we are interested in the phonetic features associated with vowel nasality, in particular.

In order to assess differences in use of acoustic cues in correctly categorizing nasalized vowels, we measured several acoustic properties of the vowels used in the perception and classification studies (non–noise-masked). First, words and phonemes were segmented using the Montreal Forced Aligner (McAuliffe *et al.*, 2017). Following automatic force-alignment, all of the phoneme boundaries in the target words were hand verified, and corrected where necessary by phonetically trained researchers.

For the nasalized vowels only, we then measured A1-P0 (Chen, 1997): this is a spectral measure of vowel nasalization reflecting the difference between the amplitude of the low-frequency nasal peak, P0 (found around 250 Hz) whose amplitude increases with increased nasality, and the amplitude of the first formant peak, A1, whose amplitude decreases with increased nasality. A smaller A1-P0 value indicates greater acoustic nasality. Since all of the target words used in the present study contained non-high vowels, A1-P0 is an appropriate measure. We measured A1-P0 at 8 equidistant points across each nasalized vowel, and the

https://doi.org/10.1121/10.0027932

average value across points for each vowel was used as our acoustic nasality value for each item. Also, following prior work indicating that F1 frequency, F1 bandwidth, and spectral tilt (measured as A3-P0 in Styler, 2017) are also spectral features that correlate with vowel nasalization, we took these acoustic measurements at 8 equidistant points across each nasalized vowel and the average value across points for each vowel was used. Finally, vowel duration is a feature of stimuli that will likely influence the ability of any human perceiver to correctly classify vowels, especially for the non-high vowels used in the current study (Hillenbrand *et al.*, 2000). Therefore, vowel duration was also measured.

We ran two separate mixed effects logistic regressions: one on the perception study coda identification data and one on the W2V2 classification data. Only the responses for the nasalized vowels were run in each model, again since we are interested in the effect of acoustic features systematically associated with nasalized vowels in the identification of nasality. The fixed effects structure for both models consisted of main effects of F1 frequency, F1 bandwidth, A1-P0, and spectral tilt. The model run on listener perception responses included vowel duration; since the classifier is run on time-normalized information, we did not include duration in the W2V2 model. F1 frequency and vowel duration were logged. All variables were centered and scaled. The model runs on the perception responses included by-listener and by-speaker random intercepts. The W2V2 model included by-speaker random intercepts.

Since our analysis examines multiple acoustic cues associated with nasalization (particularly A1-P0 and the features of F1, bandwidth, and spectral tilt), collinearity of factors is a potential problem since it violates the assumption of orthogonality of predictors. However, collinearity between variables can be handled by orthogonalizing predictor variables through residualization (Gorman, 2010; Zellou and Tamminga, 2014; Wurm and Fisicaro, 2014). In this method, one member of a pair of collinear predictors is taken as a baseline (here, A1-P0) and the other predictor (e.g., F1) is regressed linearly on the values of the baseline. The values of the second predictor are then replaced in the model by the residuals of this regression, which are by definition strictly orthogonal to the baseline predictor values. Therefore, we regressed F1, F1 bandwidth, and spectral tilt each individually by A1-P0 and included the residuals of those models as the predictors for F1, F1 bandwidth, and spectral tilt in each model.

The summary statistics for the perception model are provided in Table III. All of the fixed effects of the model were significant, except for spectral tilt, indicating that human listeners showed use of multiple acoustic cues shown to be reliably present on nasalized vowels in making coda identifications. A lower A1-P0 value on the vowel increased the likelihood of nasal coda identifications, in line with work showing that listeners are more likely to identify a vowel as nasalized in stimuli where A1-P0 decreases (Scarborough and Zellou, 2013; Zellou and Dahan, 2019). A negative coefficient for F1 indicates that listeners are more likely to identify a nasalized vowel as indicating an upcoming nasal coda as the first formant frequency decreases, consistent with prior work that the perceptual effect of vowel nasalization is a raising of the vowel quality (e.g., Beddor *et al.*, 1986). The model reported that smaller F1 bandwidth values increased nasal coda identifications, but spectral tilt does not predict listener responses. Finally, longer nasalized vowels increased the likelihood of correct coda identifications. Note that pre-nasal vowels are shorter in duration than vowels before oral codas (Peterson and Lehiste, 1960; Zellou and Scarborough, 2019), so the positive relationship between vowel duration and response accuracy within nasalized vowels cannot be related to use of the temporal property as a feature signaling greater likelihood of a nasal coda. Rather, this indicates that more acoustic information helps listeners make linguistic decisions.

TABLE III. Logistic mixed effects model output on acoustic features for correct identifications of nasalized vowels by human listeners.

	Coefficient	SE	z value	p value
Intercept	0.66	0.20	3.4	<0.001
A1-P0	–0.86	0.08	–11.28	<0.001
F1 frequency	–0.40	0.04	–9.22	<0.001
F1 bandwidth	–0.14	0.04	–3.75	<0.001
Spectral tilt	–0.04	0.04	–0.94	0.35
Vowel duration	0.67	0.03	22.78	<0.001
Random effects		Variance		SD
Speaker (intercept)		3.37		1.84
Listener (intercept)		0.44		0.67

Number of observations ($n = 13\,670$), speakers ($n = 96$), listeners ($n = 166$).

Table IV provides the output of the model run on W2V2 categorizations for nasalized vowels. The only feature that was significant in the model was spectral tilt: higher values correlated with more accurate nasal classifications. While our human listeners did not show sensitivity to spectral tilt, Styler (2017) showed that nasalized vowels contain higher spectral tilt values than non-nasal vowels.

V. DISCUSSION

The current study compared how human listeners and a state-of-the-art speech model (W2V2) use nasal coarticulatory

TABLE IV. Logistic mixed effects model output on acoustic features for correct classification rates of nasalized vowels by W2V2.

	Coefficient	SE	z value	p value
Intercept	2.71	0.36	7.44	<0.001
A1-P0	–0.36	0.25	–1.45	0.15
F1 frequency	–0.41	0.26	–1.61	0.11
F1 bandwidth	0.25	0.26	0.94	0.35
Spectral tilt	0.92	0.28	3.32	<0.001
Random effects		Variance		SD
Speaker (intercept)		3.31		1.82

Number of observations ($n = 480$), speakers ($n = 96$).

cues in speech to classify vowels as either nasalized or non-nasalized. Our stimuli were produced by 60 human talkers and generated by 36 TTS voices (96 total voices). Half of the vowels were extracted from pre-nasal contexts (i.e., from words with CVN structure), which contain nasalization due to coarticulation with the final nasal coda. The other half of the vowels were taken from an oral coda context (i.e., CVC words), and thus were non-nasal coarticulated. Human listeners completed a coda identification task, deciding whether each vowel was extracted from a word ending in a nasal or oral coda. We had a W2V2 perform classifications of vowels as either nasalized or non-nasalized.

A. Listener and W2V2 categorization of nasalized and non-nasalized vowels

Our first aim was to investigate how the patterns of W2V2 and human listener classifications differ. Overall, listeners and W2V2 both show similarly overall high performance in classifying vowels (78.5% accuracy for listeners vs 80.8% correct classifications by W2V2). This demonstrates that acoustic information from coarticulation is sufficient for the W2V2 speech model to categorize English vowels as either nasalized or non-nasalized when performing this restricted vowel classification task. Another similarity is that both listeners and W2V2 perform better for the natural human voices compared to the TTS voices overall. Although TTS voices have become increasingly naturalistic in recent years (e.g., van den Oord *et al.*, 2016), even the most advanced TTS speech (such as those used for Alexa or Siri-enabled devices) is less intelligible than naturally produced speech (e.g., Simantiraki *et al.*, 2018; Aoki *et al.*, 2022).

We also find that listeners hearing nasalized vowels in natural voices do perform well at identifying the originally intended nasal coda. This replicates prior work demonstrating that listeners use coarticulatory cues to predict upcoming linguistic information (Ohala and Ohala, 1995; Beddor *et al.*, 2013). However, listeners' coda identification accuracy is higher for non-nasalized vowels, than for nasalized vowels. Moreover, nasalized vowels generated by TTS speech are the most difficult for listeners to correctly categorize (performance was at-chance level). Thus, listeners' reduced intelligibility for TTS speech only applied to nasalized vowels (oral vowels are identified equally well across human and TTS voices). This is consistent with prior work which found that acoustic nasality for TTS voices was more phonetically ambiguous than nasalized vowels in naturally produced human speech (Ferenc Segedin *et al.*, 2019). From a practical perspective, these results, and future work examining asymmetries in intelligibility across word contexts, can be used to inform TTS synthesis techniques in order to improve speech comprehension of generated speech.

We also found systematic differences in W2V2 classification across vowel and speaker type. For the human voices, W2V2 displays patterns of performance across vowel types that parallel human listener performance (i.e., non-nasalized > nasalized). Yet, W2V2 shows the reverse pattern of classification accuracy for TTS voices: higher correct classification rates for nasalized vowels and low accuracy for non-nasalized vowels. That W2V2 is highly accurate at classifying nasalized vowels in TTS suggests that the acoustic properties of vowel nasalization present in synthesized voices are not ambiguous for W2V2 as they are for human listeners. This does not support a claim that vowel nasalization in synthetic speech results in greater acoustic ambiguity, rather that vowel nasalization in TTS is ambiguous particularly for human (American English-speaking) listeners.

We also examined whether the W2V2 model's behavior parallels human perception patterns with respect to variation across speakers. We found a great deal of cross-speaker variation in listeners' ability to identify the coda from the vowel information alone, yet this variation is largest for nasalized vowels compared to non-nasalized vowels for both human and TTS voices. This is consistent with past work that American English speakers vary greatly in their coarticulatory cues and that listeners are sensitive to this variation (Zellou, 2017). W2V2 correct classifications are also highly speaker-dependent. However, again, the speaker-dependent patterns for the TTS voices are different from those displayed by the real listeners. W2V2 showed the largest variation in identifying non-nasalized vowels across TTS voices.

B. Acoustic cues to nasalized vowel categorization

We also examined how the multidimensional acoustic properties of vowel nasalization are used by real listeners and a speech model designed primarily for ASR in making nasal categorizations of coarticulated vowels. Those analyses also reveal differences across listeners and W2V2 in use of acoustic cues. As expected, the listeners rely on several of the unique acoustic features that have been found to distinguish a nasalized vowel from a non-nasalized vowel when determining whether there is an upcoming nasal coda, such as a spectral measure of relative prominence of nasal formants and the frequency of F1. This confirms that listeners are highly sensitive to the unique coarticulatory features in the acoustic signal when making linguistic decisions. One surprising observation is that listeners did not show the expected pattern for F1 bandwidth: smaller F1 bandwidth values correlated with more nasal identifications, which is opposite from what is found in production. One possible explanation comes from recent work that has shown a changeover apparent time in California listeners' use of acoustic cues; specifically, younger listeners (of the same age group of listeners used in the present study) rely less on F1 bandwidth as a cue for nasal coda identifications than older adults (Zellou and Cohn, 2024). Thus, the reversal from the expected pattern could be part of a trend of changing cue weights in a speech community over time. We also acknowledge that it is possible that our formant bandwidth measurements had greater measurement error, since formant bandwidths are often underestimated when using linear

prediction coding analysis, particularly when fundamental frequency is high (e.g., Atal, 1975).

W2V2's performance also correlates with a well-known acoustic feature: spectral tilt. Prior work has been successful in training a Support Vector Machine classifier to use the acoustic features of vowel nasalization to accurately classify nasalized vowels (Pruthi and Espy-Wilson, 2007). Here, we find that W2V2 uses acoustic cues that are correlated with vowel nasality, but in different ways from human listeners. Investigating why W2V2 uses this particular acoustic cue to nasalization over others is a question for future work.

Coarticulation results in multidimensional phonetic variation. This yields a robust and richly informative acoustic signal where there are multiple cues to segment identities distributed over time. The different patterns of variation in coda identifications within and across humans and machines also has implications for models of speech perception. Just as it is observed that human listeners vary in their attention and use of the multiple coarticulatory cues, we find that machines show distinct patterns as well. More specifically, there is information in the signal that humans attend to that machines do not, and vice versa. Thus, learning how to map a given acoustic signal to a linguistic message can take many different pathways.

This observation that W2V2 is using acoustic cues to nasalization in American English differently than human listeners has implications for the wider use of natural language understanding and ASR systems, in particular during human–computer interactions. One possible way to think about the differences in acoustic properties is that they reflect a "misalignment" in the use of the phonetic cues during speech recognition across listeners and W2V2. In human–human interactions, systematic differences in "cue weighting" of acoustic features during speech perception can result in greater misunderstandings during communication. One example of the comes from work on L1 vs L2 speech perception: learners of English, for instance, show different use of acoustic cues during speech perception than native speakers (e.g., Escudero *et al.*, 2009), which can affect how phonemes are perceived (Kong and Yoon, 2013). There is potential for the same type of misalignment across human users and ASR systems to lead to misunderstanding (although, of course, the sources of the misalignment in use of cues across humans and machines are very different). For instance, in cases where the ASR misunderstands a human user, the user might adapt their speech patterns and produce "clear speech" adjustments to try to be better understood, and hyperarticulate the phonetic cues they use themselves to better understand speech (e.g., Buz *et al.*, 2016). If machine classifiers have different acoustic weightings of cues to perform word identification than speakers assume based on human–human interaction, this can lead to even greater misunderstandings (e.g., "spiraling errors") (Oviatt *et al.*, 1998). This is another line for future work to investigate. Fine-tuning ASR systems to match the cue weightings of human listeners can be one way to lead to better alignment between how users adapt their speech when talking to machines and what the machines rely on to understand speech.

C. Limitations and future directions

There were several limitations of the present study which open even more avenues for future research. For one, while the present study only focused on comparing nasalized and non-nasalized vowels, looking at other types of word structures and coarticulatory patterns presents interesting directions for future research. For instance, next steps could be to compare human listener and W2V2 performance on use of anticipatory labial, liquid, or vowel-to-vowel coarticulation in predicting upcoming linguistic information, since each of these properties has been shown to be used in speech perception (e.g., Krueger and Noiray, 2022; Redford *et al.*, 2018; West, 1999). We also acknowledge that coarticulation can vary in vowel-specific ways (e.g., Zellou and Scarborough, 2019). Investigating vowel-based asymmetries in the use of coarticulatory cues across human listener and machine classifiers is a ripe avenue for future research. The current study also used commercially available TTS voices; future work can explore machine and human classification of parametrically synthesized speech where acoustic cues are systematically varied.

Another approach for future work is to compare human and automatic machine classifications of coarticulation across different languages. Specifically, there is a large amount of work demonstrating that patterns of coarticulation are language-specific (e.g., Beddor *et al.*, 2002; Keating and Cohn, 1988) and even dialect-specific (Zellou and Tamminga, 2014; Bongiovanni, 2021). Examining whether perception and neural network classifications are more or less similar in languages or dialects that vary in their coarticulatory patterns could be both revealing to understanding the practical consequences of cross-language phonetic variation and also important for developing accurate natural language understanding methods for many different languages. For instance, a recent study used a similar approach to that in the current paper to develop a classifier for vowel nasality in French, where vowel nasality is lexically contrastive (Elmerich *et al.*, 2023). In that study, the testing data involved recordings made with an aerodynamic mask to measure nasal airflow. Their study compared the results obtained with a spectrogram-based convolutional neural network model to the aerodynamic measurements, specifically the nasal airflow. The findings demonstrated that the NN model achieved better nasality detection for French nasal vowels compared to the current results for English, achieving an overall performance of 88% accurate nasal vowel detection. The predictions made by the NN model were found to be correlated with nasal airflow, and with variations observed across vowels and speakers, thereby validating the use of convolutional neural network in this context (Elmerich *et al.*, 2023). Thus, it appears that neural networks are sensitive to language-specific patterns of vowel nasalization. Investigating this question further is a ripe direction for future work.

Another direction for future work is to create a deep learning system that attends to acoustic cues in the same

ways as human listeners. The current study found differences in the use of some acoustic features. Examining context-specific asymmetries in speech recognition system performance and adjusting the weighting of acoustic features, such as those used by listeners, might lead to a more "fine-tuned" deep learning system that can mimic human perceptual patterns (see also Gu *et al.*, 2018, who discuss this as a future approach in neural network research) and provide insight into how neural systems learn to attend to the unique acoustic features of coarticulation. Future work examining the use of coarticulatory cues by different types of listeners (for instance, human learners of English) can also be used to understand language acquisition. It is also worth noting that there are lots of other cues that are used by listeners during spoken language comprehension that are not used by speech recognition systems, such as visual cues from the speaker, socio-pragmatic information, and much more. This is another explanation for why human and machine comprehenders perform speech recognition differently during spontaneous spoken language interactions.

VI. CONCLUSION

Overall, the current study provides insight into how coarticulatory variation in vowels produced before nasal codas in American English are categorized by human listeners and artificial neural networks. The deep learning model yields global correct classification rates of vowels close to human-level accuracy. There are also systematic similarities in the patterns of classifications of nasalized vs non-nasalized vowels for natural human voices (but not for synthetic voices). Acoustic analyses reveal differences in how the deep learning model uses acoustic cues for nasal coarticulation to classify vowels from human perceivers. More broadly, these results demonstrate that comparing listener speech perception and deep learning-based classifications of phonetic variation is one approach that can inform how the models perform like, or differ from, the human perceptual system.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation Grant No. 2140183 to G.Z.

AUTHOR DECLARATIONS

Conflict of Interest

The authors report funding from the National Science Foundation. No other conflicts of interest or competing interests are reported.

DATA AVAILABILITY

The raw data supporting the conclusions of this article will be made available by the authors upon reasonable request.

Ali, L., Gallagher, T., Goldstein, J., and Daniloff, R. (**1971**). "Perception of coarticulated nasality," J. Acoust. Soc. Am. **49**, 538–540.

Ammari, T., Kaye, J., Tsai, J. Y., and Bentley, F. (**2019**). "Music, search, and IoT: How people (really) use voice assistants," ACM Trans. Comput-Hum. Interact. **26**(3), 1–28.

Ansary, L., and Salehi, S. A. S. (**2004**). "Modeling phones coarticulation effects in a neural network based speech recognition system," in *Proceedings of the Eighth International Conference on Spoken Language Processing*, pp. 1–4.

Aoki, N. B., Cohn, M., and Zellou, G. (**2022**). "The clear speech intelligibility benefit for text-to-speech voices: Effects of speaking style and visual guise," JASA Express Lett. **2**(4), 045204.

Atal, B. S. (**1975**). "Linear prediction of speech: Recent advances with applications to speech analysis," in *Speech Recognition*, edited by R. D. Reddy (Elsevier, New York), pp. 221–230.

Baevski, A., Zhou, Y., and Mohamed, A. (**2020**). "wav2vec 2.0: A framework for self-supervised learning of speech representations," Adv. Neural Info. Pro. Syst. **33**, 12449–12460.

Bates, D., Kliegl, R., Vasishth, S., and Baayen, H. (**2015**). "Parsimonious mixed models," arXiv:1506.04967.

Beddor, P. S. (**1993**). "The perception of nasal vowels," in *Nasals, Nasalization, and the Velum*, edited by M. Huffman and R. Krakow (Academic Press, San Diego), pp. 171–196.

Beddor, P. S. (**2009**). "A coarticulatory path to sound change," Language **85**(4), 785–821.

Beddor, P. S., Harnsberger, J. D., and Lindemann, S. (**2002**). "Language-specific patterns of vowel-to-vowel coarticulation: Acoustic structures and their perceptual correlates," J. Phon. **30**(4), 591–627.

Beddor, P. S., and Krakow, R. A. (**1999**). "Perception of coarticulatory nasalization by speakers of English and Thai: Evidence for partial compensation," J. Acoust. Soc. Am. **106**(5), 2868–2887.

Beddor, P. S., Krakow, R. A., and Goldstein, L. M. (**1986**). "Perceptual constraints and phonological change: A study of nasal vowel height," Phonol. Yearb. **3**, 197–217.

Beddor, P. S., McGowan, K. B., Boland, J. E., Coetzee, A. W., and Brasher, A. (**2013**). "The time course of perception of coarticulation," J. Acoust. Soc. Am. **133**(4), 2350–2366.

Bongiovanni, S. (**2021**). "Acoustic investigation of anticipatory vowel nasalization in a Caribbean and a non-Caribbean dialect of Spanish," Ling. Vang **7**(1), 20200008.

Bradlow, A. R., Torretta, G. M., and Pisoni, D. B. (**1996**). "Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics," Speech Commun. **20**(3), 255–272.

Buz, E., Tanenhaus, M. K., and Jaeger, T. F. (**2016**). "Dynamically adapted context-specific hyper-articulation: Feedback from interlocutors affects speakers' subsequent pronunciations," J. Mem. Lang. **89**, 68–86.

Carignan, C. (**2017**). "Covariation of nasalization, tongue height, and breathiness in the realization of F1 of Southern French nasal vowels," J. Phon. **63**, 87–105.

Chen, M. Y. (**1997**). "Acoustic correlates of English and French nasalized vowels," J. Acoust. Soc. Am. **102**(4), 2360–2370.

Cho, T., Kim, D., and Kim, S. (**2017**). "Prosodically-conditioned fine-tuning of coarticulatory vowel nasalization in English," J. Phon. **64**, 71–89.

Clopper, C. G., and Bradlow, A. R. (**2008**). "Perception of dialect variation in noise: Intelligibility and classification," Lang. Speech **51**(3), 175–198.

Deng, Z., Neumann, U., Lewis, J. P., Kim, T. Y., Bulut, M., and Narayanan, S. (**2006**). "Expressive facial animation synthesis by learning speech coarticulation and expression spaces," IEEE Trans. Vis. Comp. Graph. **12**(6), 1523–1534.

Elmerich, A., Kim, L., Gendrot, C., Amelot, A., Crevier-Buchman, L., and Maeda, S. (**2023**). "Nasality detection from acoustic data with a convolutional neural network and comparison with aerodynamic data," *Proceedings of International Congress of the Phonetic Sciences*.

Escudero, P., Benders, T., and Lipski, S. C. (**2009**). "Native, non-native and L2 perceptual cue weighting for Dutch vowels: The case of Dutch, German, and Spanish listeners," J. Phon. **37**(4), 452–465.

Fazel, A., Yang, W., Liu, Y., Barra-Chicote, R., Meng, Y., Maas, R., and Droppo, J. (**2021**). "Synthasr: Unlocking synthetic data for speech recognition," in *Proceedings of Interspeech*, arXiv:2106.077803.

Ferenc Segedin, B., Cohn, M., and Zellou, G. (**2019**). "Perceptual adaptation to device and human voices: Learning and generalization of a phonetic shift across real and voice-AI talkers," in *Proceedings of Interspeech*, pp. 2310–2314.

Fowler, C. A. (**1984**). "Segmentation of coarticulated speech in perception," Percept. Psychophys. **36**(4), 359–368.
Fowler, C. A., and Saltzman, E. (**1993**). "Coordination and coarticulation in speech production," Lang. Speech **36**(2), 171–195.
Garellek, M., Ritchart, A., and Kuang, J. (**2016**). "Breathy voice during nasality: A cross-linguistic study," J. Phon. **59**, 110–121.
Garofolo, J. S. (**1993**). "Timit acoustic phonetic continuous speech corpus," in *LDC93S1. Web Download*. Philadelphia: Linguistic Data Consortium.
Gorman, K. (**2010**). "The consequences of multicollinearity among socio-economic predictors of negative concord in Philadelphia," U. Penn. Work. Papers Linguistics **16**(2), 66–75.
Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., and Chen, T. (**2018**). "Recent advances in convolutional neural networks," Pattern Recognit. **77**, 354–377.
Guillaume, S., Wisniewski, G., and Michaud, A. (**2023**). "From 'snippetlects' to doculects and dialects: Leveraging neural representations of speech for placing audio signals in a language landscape," arXiv:2305.18602.
Hawkins, S., and Stevens, K. N. (**1985**). "Acoustic and perceptual correlates of the non-nasal–nasal distinction for vowels," J. Acoust. Soc. Am. **77**(4), 1560–1575.
Hillenbrand, J. M., Clark, M. J., and Houde, R. A. (**2000**). "The role of duration in the perception of vowel quality," J. Acoust. Soc. Am. **107**(5), 2917–2918.
Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., and Kingsbury, B. (**2012**). "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal Process. Mag. **29**(6), 82–97.
Huang, X. (**1992**). "Minimizing speaker variation effects for speaker-independent speech recognition," in *Speech and Natural Language: Proceedings of a Workshop*, Hairman, NY.
Jang, J., Kim, S., and Cho, T. (**2018**). "Focus and boundary effects on coarticulatory vowel nasalization in Korean with implications for cross-linguistic similarities and differences," J. Acoust. Soc. Am. **144**(1), EL33–EL39.
Kanthak, S., and Ney, H. (**2002**). "Context-dependent acoustic modeling using graphemes for large vocabulary speech recognition," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1–4.
Keating, P. A., and Cohn, A. C. (**1988**). "Cross-language effects of vowels on consonant onsets," J. Acoust. Soc. Am. **84**(S1), S84.
Keating, P. A., Kreiman, J., Alwan, A., Chong, A., and Lee, Y. (**2021**). "The UCLA speaker variability database," LDC2021S09. Web Download (Linguistic Data Consortium, Philadelphia, PA).
Kong, E. J., and Yoon, I. H. (**2013**). "L2 proficiency effect on the acoustic cue-weighting pattern by Korean L2 learners of English: Production and perception of English stops," Phon. Speech Sci. **5**(4), 81–90.
Krakow, R. A., Beddor, P. S., Goldstein, L. M., and Fowler, C. A. (**1988**). "Coarticulatory influences on the perceived height of nasal vowels," J. Acoust. Soc. Am. **83**(3), 1146–1158.
Krueger, S., and Noiray, A. (**2022**). "Developmental differences in perceptual anticipation underlie different sensitivities to coarticulatory dynamics," J. Child Lang. **49**(5), 959–978.
Kunze, J., Kirsch, L., Kurenkov, I., Krug, A., Johannsmeier, J., and Stober, S. (**2017**). "Transfer learning for speech recognition on a budget," arXiv:1706.00290.
Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (**2015**). "Package 'lmertest,' " R package **2**(0), 734.
Lahiri, A., and Marslen-Wilson, W. (**1991**). "The mental representation of lexical form: A phonological approach to the recognition lexicon," Cognition **38**(3), 245–294.
Lee, A., Gong, H., Duquenne, P. A., Schwenk, H., Chen, P. J., Wang, C., Potpourri, S., Adi, Y., Pino, J., Gu, J., and Hsu, W. N. (**2021**). "Textless speech-to-speech translation on real data," arXiv:2112.08352.
Li, H., Kim, S., and Cho, T. (**2020**). "Prosodic structurally conditioned variation of coarticulatory vowel nasalization in Mandarin Chinese: Its language specificity and cross-linguistic generalizability," J. Acoust. Soc. Am. **148**(3), EL240–EL246.
Lian, Z., Li, Y., Tao, J., and Huang, J. (**2018**). "Improving speech emotion recognition via transformer-based predictive coding through transfer learning," arXiv:1811.
Liu, N., Zhou, T., Ji, Y., Zhao, Z., and Wan, L. (**2020**). "Synthesizing talking faces from text and audio: An autoencoder and sequence-to-sequence convolutional neural network," Pattern Recognit. **102**, 107231.
Ma, D., Ryant, N., and Liberman, M. (**2021**). "Probing acoustic representations for phonetic properties," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing* (IEEE, New York), pp. 311–315.
McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., and Sonderegger, M. (**2017**). "Montreal forced aligner: Trainable text-speech alignment using Kaldi," in *Proceedings of Interspeech*, pp. 498–502.
Mun, S. I., Han, C. J., and Hong, H. S. (**2022**). "Exploiting variable length segments with coarticulation effect in online speech recognition based on deep bidirectional recurrent neural network and context-sensitive segment," Int. J. Speech Technol. **25**(1), 135–146.
Ohala, J. J., and Ohala, M. (**1995**). "Speech perception and lexical representation: The role of vowel nasalization in Hindi and English. Phonology and phonetic evidence," Lab. Phon. **IV**, 41–60.
Oviatt, S., Levow, G. A., Moreton, E., and MacEachern, M. (**1998**). "Modeling global and focal hyperarticulation during human–computer error resolution," J. Acoust. Soc. Am. **104**(5), 3080–3098.
Pasad, A., Chou, J. C., and Livescu, K. (**2021**). "Layer-wise analysis of a self-supervised speech representation model," in *IEEE ASRU Workshop*, pp. 914–921.
Peterson, G. E., and Lehiste, I. (**1960**). "Duration of syllable nuclei in English," J. Acoust. Soc. Am. **32**(6), 693–703.
Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., and Raymond, W. (**2005**). "The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability," Speech Commun. **45**(1), 89–95.
Pruthi, T., and Espy-Wilson, C. Y. (**2007**). "Acoustic parameters for the automatic detection of vowel nasalization," in *Proceedings of the Eighth Annual Conference of the ISCA*, pp. 1924–1928.
Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (**2023**). "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning* (PMLR), pp. 28492–28518.
Redford, M. A., Kallay, J. E., Bogdanov, S. V., and Vatikiotis-Bateson, E. (**2018**). "Leveraging audiovisual speech perception to measure anticipatory coarticulation," J. Acoust. Soc. Am. **144**(4), 2447–2461.
Scarborough, R., and Zellou, G. (**2013**). "Clarity in communication: 'Clear' speech authenticity and lexical neighborhood density effects in speech production and perception," J. Acoust. Soc. Am. **134**(5), 3793–3807.
Shah, J., Singla, Y. K., Chen, C., and Shah, R. R. (**2021**). "What all do audio transformer models hear? Probing acoustic representations for language delivery and its structure," arXiv:2101.00387.
Simantiraki, O., Cooke M., and King, S. (**2018**). "Impact of different speech types on listening effort," in *Proceedings of Interspeech*, pp. 2267–2271.
Stevens, K. N. (**2000**). *Acoustic Phonetics* (MIT Press, Cambridge, MA), Vol. 30.
Styler, W. (**2017**). "On the acoustical features of vowel nasality in English and French," J. Acoust. Soc. Am. **142**(4), 2469–2482.
Triantafyllopoulos, A., Wagner, J., Wierstorf, H., Schmitt, M., Reichel, U., Eyben, F., Burkhardt, F., and Schuller, B. W. (**2022**). "Probing speech emotion recognition transformers for linguistic knowledge," arXiv:2204.00400.
van den Oord, D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, A. S., and Kavukcuoglu, K. (**2016**). "Wavenet: A generative model for raw audio," arXiv:1609.03499.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (**2017**). "Attention is all you need," Adv. Neural Info. Pro. Syst. **30**, 1–11.
West, P. (**1999**). "The extent of coarticulation of English liquids: An acoustic and articulatory study," in *Proceedings of International Congress of the Phonetic Sciences*, pp. 1901–1904.

Wright, J. T. (**1986**). "The behavior of nasalized vowels in the perceptual vowel space," Exp. Phon. **1**, 45–67.
Wurm, L. H., and Fisicaro, S. A. (**2014**). "What residualizing predictors in regression analyses does (and what it does not do)," J. Mem. Lang. **72**, 37–48.
Yu, A. C., and Zellou, G. (**2019**). "Individual differences in language processing: Phonology," Annu. Rev. Linguist. **5**, 131–150.
Zellou, G. (**2017**). "Individual differences in the production of nasal coarticulation and perceptual compensation," J. Phon. **61**, 13–29.
Zellou, G. (**2022**). *Coarticulation in Phonology* (Cambridge University Press, Cambridge, UK).
Zellou, G., and Cohn, M. (**2024**). "Apparent-time variation in the use of multiple cues for perception of anticipatory nasal coarticulation in California English," Glossa **9**(1), 1–29.
Zellou, G., Cohn, M., and Block, A. (**2021**). "Partial compensation for coarticulatory vowel nasalization across concatenative and neural text-to-speech," J. Acoust. Soc. Am. **149**(5), 3424–3436.
Zellou, G., and Dahan, D. (**2019**). "Listeners maintain phonological uncertainty over time and across words: The case of vowel nasality in English," J. Phon. **76**, 100910.
Zellou, G., Pycha, A., and Chitoran, I. (**2023**). "Use of gradient anticipatory nasal coarticulatory cues for lexical perception in French," Lab. Phon. **14**(1), 1–27.
Zellou, G., and Scarborough, R. (**2019**). "Neighborhood-conditioned phonetic enhancement of an allophonic vowel split," J. Acoust. Soc. Am. **145**(6), 3675–3685.
Zellou, G., and Tamminga, M. (**2014**). "Nasal coarticulation changes over time in Philadelphia English," J. Phon. **47**, 18–35.