Embedding and measurement of vowels using machine perception

James Burridge[1] and Bert Vaux[2]
[1]School of Mathematics and Physics, University of Portsmouth, UK[a]
[2]Faculty of Modern and Medieval Languages and Linguistics, University of Cambridge, UK

(Dated: 11 May 2022)

We present a spatial embedding and measurement method for vowel sounds based on the output of a convolutional neural network (CNN) which has been trained to recognize phonemic categories from spectrograms, and has similar perceptual behaviour to humans. We define the perceptual similarity between two categories as a typical listener's degree of belief that one category was intended by a typical talker given that the other was uttered. In low-dimensional "chart-space" we model utterance distributions, conditional on phonemic category, as Gaussian, with means chosen to match the perceptual characteristics of the CNN as closely as possible. In this way, perceptual similarities between high dimensional spectrogram images encoded by the CNN are given a low dimensional representation in the chart. We then present a likelihood-based mapping from acoustic to chart space, via the output of the CNN. This generates low dimensional measurements of individual vowel spectrograms. Our method can in principle be generalized to measure any form of acoustic variation, provided we are able to first define a set of categories of sounds which describe it. The potential value of our approach is that it provides a means of producing perceptually meaningful measurements of sounds directly from their spectrograms.

[https://doi.org(DOI number)]

[XYZ] Pages: 1–12

I. INTRODUCTION

Our aim is to introduce a method for measuring sounds which combines the simplicity and interpretability of traditional measures such as formants[1–4], with the powerful perceptual capabilities of modern machine learning approaches[5–7]. Our idea is to train a machine learner, in our case a convolutional neural network[8] (CNN), to recognize a given set of sound categories, and then to map the output of this machine to a low dimensional measurement space which we refer to as a "chart". We present an example of the approach in which the distinctive categories are taken to be the phoneme set used to transcribe vowels in the TIMIT corpus[9], but in principle the method can be applied to any set of categories deemed linguistically important. Our mapping uses an idealized model of the distribution of category members in chart space, where each category is modelled as a Gaussian, as in Linear Discriminant Analysis[6]. The means of these distributions are chosen to reproduce, as far as possible, the perceptual characteristics of the machine, and may be viewed as forming a perceptual vowel chart[10,11]. To turn this chart into a measurement system for individual sounds we define a likelihood based mapping from the response of the machine learner to a sound, to a point in chart space. We then explore the characteristics of this measurement system. Before presenting our technique, we review existing measurement methods, and their strengths and weaknesses. We will use this review to motivate and contextualize our own method.

A. Formants

Since the classic work of Peterson and Barney[12], it has been common practice to measure vowels using their formants[1–4]. Formants are defined according to the American National Standards Institute[13] as follows: "[given] a range of frequencies in which there is an absolute or relative maximum in the sound spectrum …the frequency [unit, hertz (Hz)] at the maximum is the formant frequency." This definition is not universal[14]. An alternative is that formants are poles of the transfer function of the supra-glottal vocal tract[14–16]. This decouples them from the glottal source. Neither definition is perfect; the first is ambiguous because the number and locations of peaks depend on the parameters used to compute the spectrum, and on the peak detection method. The second assumes an idealized model of the vocal tract[16] which can never match reality. Obtaining consistent measurements of formants is also not straightforward, however one chooses to define them[2]. Despite these problems, formants have proven useful in many applications, including studies of vowel perception[4,17], dialect comparison[18], measurement of vowel systems[19], speech synthesis[20] and analysis of vocal tract configuration[15,21]. Their effectiveness in applications derives from the resonant nature of vowels, with formants approximating the frequencies of these resonances, and correlating to the traditional phonological parameters of tongue height and backness[21].

[a]james.burridge@port.ac.uk

B. Can we improve on formants?

There is no doubt that formants provide a powerful and easily-interpretable way to measure speech sounds. However, they lack some desirable characteristics of a quantitative measurement system. As noted above, their definition is ambiguous and they are difficult to measure consistently[2]. Moreover, in connected speech they are highly dispersed in the sense that the formant distributions of sounds belonging to linguistically distinct categories (e.g., different phonemes) typically have very large overlap[22] (see Figure 10). Although this is a feature of their definition rather than a fault, it limits their utility as a method to measure differences between speech sounds other than in tightly controlled conditions[19]. Formants also ignore a great deal of spectral information, which is both a strength and a weakness: they are simple to read and interpret, and the information they preserve is particularly relevant linguistically, but they are by definition limited in what they can measure. Vowel nasality, for example, is more effectively measured using other characteristics like formant amplitudes, bandwidth and spectral tilt[23]. To improve on Formants, a measurement system would need to incorporate their simplicity and perceptual relevance, but be able to measure any acoustic distinctions of interest, and do so in a precise way.

C. Full spectral measures

To precisely describe the full complexity of real speech sounds requires a high dimensional representation. A standard technique is to calculate the power spectrum of short time Fourier transforms[24] of the raw speech signal over a sequence of overlapping windows. The sequence of spectra forms a two-dimensional spectrogram which, although it contains no phase information, can be used to reconstruct a signal which is almost indistinguishable from the original[25]. We may therefore view the spectrogram as a more or less complete, but very high dimensional, representation of speech sounds. In comparison, formants represent a low-dimensional approximate encoding of the spectrum for a single window. The fact that formants discard spectral detail has led researchers to explore higher dimensional "whole spectrum measures" as predictors in vowel classification tasks[3,4], discovering that these richer features lead to better classification accuracy and more human-like perceptual confusions.

A lower dimensional alternative to the power spectrum are Mel frequency cepstral coefficients[26] (MFCCs), which are a standard predictive feature used in automatic speech recognition. These are obtained by first filtering the spectrum from each window into overlapping bands (using triangular band filters) which are equally spaced on the Mel scale[27], and then applying a discrete cosine transform which decorrelates the filter bank. The MFCCs are then the first 12 or 13 coefficients of the transform (higher coefficients contain less relevant information). MFCCs have been shown to be very effective for phoneme classification (79.3% accuracy on TIMIT[9,28]) when combined with Boltzmann Machines[28], a precursor to neural networks. While MFCCs are the most commonly used spectral features in automatic speech recognition (ASR)[7,29], in recent work[30,31] convolutional neural networks have been used to learn features directly from the raw audio signal. In particular, training a CNN to generate phone class probabilities directly from the raw signal can increase performance in Hidden Markov Model ASR compared to systems which extract MFCCs as a first step[31]. Interestingly, the filters in the first convolution layer appear to automatically learn formant-like information from the signal[31].

D. Perceptual measurement systems

We have argued that formants are an important and useful feature of the speech signal, but with some weaknesses. They are archetypal hand-crafted low-dimensional features, which are easily interpreted by humans and relate directly to the perceptual differences between sounds. More complete representations such as the spectrogram, MFCCs, and the raw signal avoid feature loss, but are not directly interpretable by humans. Modern machine learning methods are able to take these high dimensional representations and extract human-readable information, most notably orthographic representations ("speech-to-text" systems). This raises a question: is there a way to utilize the capabilities of a machine learner to generate a low-dimensional formant-like representation of sounds which is capable of accurately measuring any acoustic distinction in which we are interested?

To develop such a system we must first decide on the distinctions we wish to capture, and use these to define a set of categories of sounds. An obvious set of distinctions, for which good data exist, is the differences between sound categories used for phonetic transcription in a given language or dialect. In this paper we will focus only on this case, and specifically on vowels, and we will refer to the categories as phonemes. We require a measurement system which is sensitive to the acoustic differences which distinguish these categories, but insensitive to differences which are either imperceptible or ignored at the relevant level of analysis. For example speakers of different ages, genders or body size uttering the same word in a given dialect will generate quite different acoustic outputs, but listeners are still able to recognize the word as being the same, implying that they can discount these variations[32] when interpreting lexical meaning. This does not mean that they lack the sensorineural capabilities to detect the differences they discount[33]. We want a measurement system capable of discounting such variations, if we wish it to.

Suppose we define a $d \in \mathbb{N}$ parameter measurement space (for example, if we use the first three formants to measure sounds then we have $d = 3$). Our measurement system maps sounds to points in $\mathbb{R}^d$. We wish to design the system so that changing sounds in a way which is not perceptually significant to category membership results in only small changes in its parameters. Perceptually significant changes should produce parameter shifts which are proportionately large. Such a system, applied to a large sample of vowel phonemes uttered by a range of speakers, will map sound categories to clusters in $\mathbb{R}^d$, with inter-cluster distances proportional to the perceptual differences between phonemes. We

may view the centroids of these clusters as the locations of vowels in a perceptual chart.

Perceptual charts date back to the mid 20th century[10,11]. In particular, Shepard[10] (1972) defined a symmetric similarity metric between phonemes, based on an experimental confusion matrix, C. The element C_{ij} gives the fraction of times that a group of listeners identified phoneme i as phoneme j. Phonemes were then embedded in $\mathbb{R}^d$ so as to minimize the squared error between similarities and an exponential function of their Euclidean separation (in effect defining perceptual distances between phonemes to be the negative logarithm of their perceptual similarity). Embedding of this kind does not define a measurement system because it cannot be used to map individual sounds to points in $\mathbb{R}^d$. However, if the group of experimental listeners who were used to generate the confusion matrix are replaced with a machine trained to have human-like perceptual characteristics, then we can repeatedly present this machine with new sounds, and map them individually to $\mathbb{R}^d$, accounting for the perceptual similarity to the categories on which the machine was trained. Our aim is to define such a system, and explore its properties and potential applications.

E. Outline of paper

Our measurement system consists of three parts. First, in section II A, we train a CNN to recognize the categories (TIMIT monophthongs[9]), and compare its perceptual characteristics to those of human listeners. Second, in section II B, we embed categories in $\mathbb{R}^d$ so that each is represented by a standard position. Third, in section II C, we define a method for mapping individual sounds into the same space so that their proximity to the standard positions is a measure of their similarities to these categories. In section III we present the results. Specifically, the embedding results are presented in section III A and the behaviour of the measurement system in section III B. A simple application to dialect variation is given in section III C. Our findings are summarized, along with further potential applications, in the discussion section IV.

II. METHODOLOGY

A. Data preparation, CNN training and behaviour

We use acoustic data from the TIMIT corpus[9], which contains recordings of 630 speakers representing 8 major dialect divisions of American English, with each speaker uttering 10 sentences. It also includes time-aligned phonetic transcriptions of every sentence into ARPAbet[34]. We will be interested in monophthongal vowels, which are listed in their ARPAbet and IPA representation[35,36], with example words[9], in Table I.

Training and test data are extracted from TIMIT sentences by first scaling the amplitude of all sample points of the raw audio signal (16000 Hz sample rate $\equiv$ 0.0625ms sampling interval) which do not belong to the monophthongal vowels listed in table I, by a factor of 10^{-3}. The modified signal, consisting of a sequence of vowel sounds, is then

TABLE I. ARPAbet monophthongs from TIMIT, IPA representation, the example word (assumed spoken in American English) given in TIMIT documentation[9], and the relative occurrence frequency of the character out of all monophthongs in TIMIT. A more accurate IPA representation of the ARPAbet symbol /ao/ is /ɒ/, but we use /ɔ/ for consistency with the standard conversion[34,37].

ARPAbet	IPA	Example	Freq. (%)
iy	i	beet	12.1
ix	ɨ	debit	15.0
ih	ɪ	bit	8.8
eh	ɛ	bet	6.7
ae	æ	bat	7.0
aa	ɑ	bott	5.3
ao	ɔ	bought	5.1
uh	ʊ	book	0.9
uw	u	boot	1.0
ux	ʉ	toot	3.3
ah	ʌ	but	4.0
ax	ə	about	6.3
ax-h	ə̥	suspéct	0.7
er	ɝ	bird	3.6
axr	ɚ	butter	5.9

transformed to a spectrogram by dividing into overlapping windows of length 32 ms (2^9 sample points), with each successive window shifted forward by 4 ms (2^6 sample points) with respect to the last. The Hamming window[24] is applied to each signal window before taking the fast Fourier transform (FFT), in order to counteract the FFT assumption that the signal is infinitely repeating, and to reduce spectral leakage[24]. A 64-band Mel filter bank is then applied to the Fourier power spectrum, and the lowest 50 bands are retained, corresponding to a maximum frequency of 4210 Hz, which is high enough to contain F_1, F_2 and F_3 for all speakers[19]. The Mel filtered spectrum is then converted to the decibel scale, and 50-window sequences centred on each monophthong are extracted, generating a 50×50 real valued matrix for each vowel. These matrices may be interpreted as monochrome images, and some examples are provided in Figure 1, labelled by the vowel phonemes they represent. The set of matrix-label pairs is divided into training and test sets using the TIMIT recommended split[9] so that no speaker appears in both.

Our aim is to train a machine learner to identify the phonetic label of these vowel spectrograms. Because this learner is standing in for a human listener, we want it to possess typical human-like perceptual characteristics. In this sense we are not attempting to maximize the accuracy of the classification as in previous work on TIMIT[28] ($\approx 80\%$); we will be satisfied with an error rate similar to that of humans ($\approx 70\%$ accuracy for vowels[38]). Our learner of choice is the Convolutional Neural Network (CNN), now the dominant approach

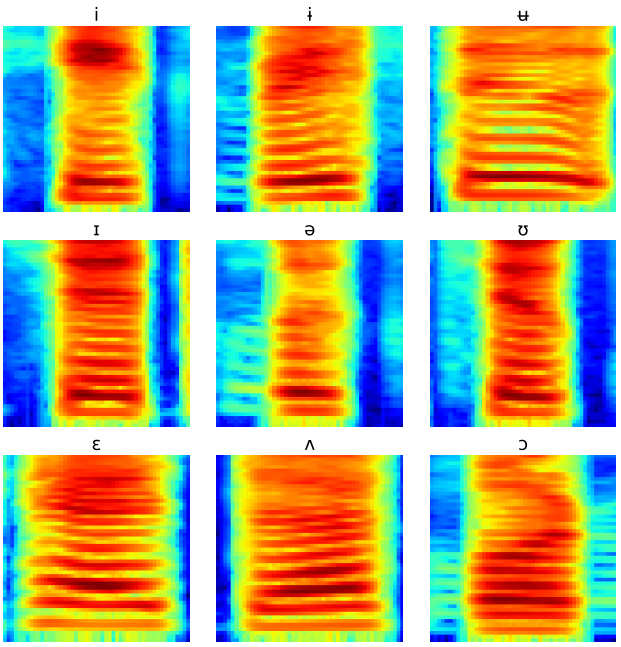

FIG. 1. Spectrogram images (50×50) of a sample of vowel sounds, labelled by their IPA character. Window length 32 ms, window stride 4 ms, duration of each image 200 ms. Mel filter bank (64 bands) applied to Fourier spectrum, and lowest 50 bands retained, maximum frequency 4210 Hz.

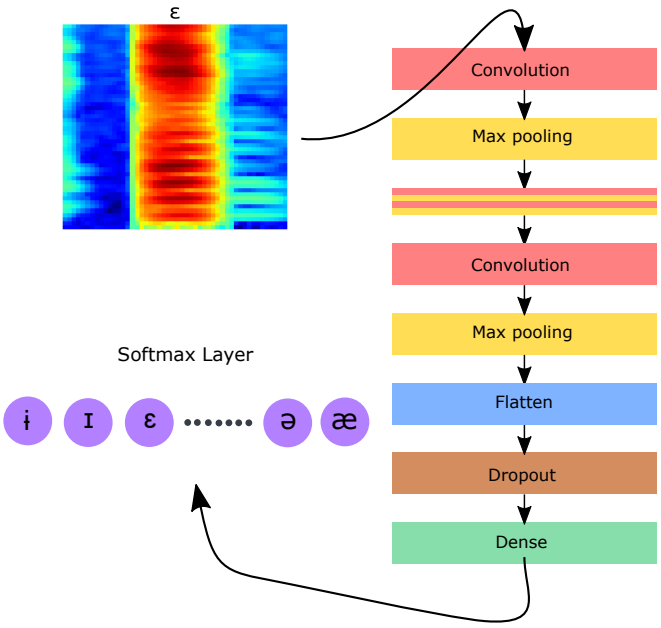

FIG. 2. Diagram of the CNN used in this paper. The input is a 50×50 single channel spectrogram image. The convolution layers filter 3×3 squares of pixels and have output depths 32, 64, 128, and 128 respectively. Each max pooling layer reduces each 2×2 square of pixels into a single pixel. Dropout rate is set to 50%. The dense (fully connected) layer contains 512 nodes, and the output layer consists of K softmax nodes, where K is the number of phonemes. The output of these nodes may be interpreted as probabilities associated with each phoneme, given the input spectrogram.

for image recognition tasks[8]. A CNN is a class of deep learning model inspired by the organization of the animal visual cortex[39,40], the practical utility of which was first demonstrated by LeCun and colleagues[41] (1989). The simple CNN we use is similar in structure to AlexNet[42] (a variant of LeCun's original architecture), which won the 2012 ImageNet Large Scale Visual Recognition Challenge with an error rate 10.8% lower than the runner-up (a significant moment in the history of deep learning[43]). Although CNNs were developed for image recognition, they are increasingly used to analyse speech and other sounds, either by treating the spectrogram as an image[44–47], or by applying convolutional filters to the raw signal[31,48].

For readers unfamiliar with CNNs, we briefly explain the architecture. The network is structured in layers, with the input being a rectangular image. Each pixel is a "feature vector" whose dimension—the number of "channels"—is referred to as its depth. Initially channels represent colors, with greyscale images having only one channel. Each convolution layer slides a filter over the input image which maps each small square of pixels to a new pixel with a new feature vector, typically of different depth to the input (deeper, usually). The filtered squares usually overlap so the resulting image contains a similar number of pixels as the input. The purpose of the filters is to extract features which are relevant to the prediction task at hand. Convolution layers are interspersed with max-pooling layers which downsample by replacing each 2×2 square of pixels with a single pixel equal to the maximum of the square in each channel, producing an image with a quarter the number of pixels. The output of the final convolution layer is flattened into a vector, before "regularizing" by setting a random set of features to zero during training (so-called "dropout"), to prevent over-fitting of the network. This regularized output is then fed into a densely connected layer (every node receiving every output of the previous layer), which then feeds into an output layer. This final layer contains K nodes, where K is the number of different phonemes in the training data. For a given input to the network, the outputs of these nodes represent the probabilities that the input was a spectrogram of each of the K phonemes. The network is trained by adjusting the weights of the filters and the dense layer (using back propagation) to minimize a loss function which measures its performance at classifying a large set of training images. Figure 2 shows the detailed structure of the network we use.

To understand the perceptual behaviour of the CNN, we consider the loss function, which in our case is the categorical cross entropy[49]. We label our phonemes $1, 2, \ldots, K$. Let $X \in \mathbb{R}^N$ denote the acoustic output that a speaker makes when uttering phoneme V. We may view X as a spectrogram for that phoneme. Since different speakers produce different acoustic outputs for the same phoneme, and may use similar acoustic outputs to convey different phonemes, then given X we cannot be certain of V. We can characterize this uncertainly as follows. Define the indicator function that a speaker

intended phoneme k

$$I_k(V) = \begin{cases} 1 & \text{if } V = k \\ 0 & \text{otherwize.} \end{cases} \tag{1}$$

The probability that k was intended given that x was uttered is then

$$p_k(x) = \mathbb{P}(V = k|X = x) = \mathbb{E}(I_k(V)|X = x), \tag{2}$$

where, for notational simplicity, we have assumed the utterance distribution is discrete. Using the above notation, the set of utterance-phoneme pairs used to train our CNN may be written $\mathcal{D} = ((x_1, v_1), (x_2, v_2), \ldots (x_n, v_n))$. We assume pairs are drawn from the joint distribution

$$f_{XV}(x, v) = \mathbb{P}(X = x \cap V = v), \tag{3}$$

with marginals $f_X(x)$ and $f_V(v)$, and conditional distributions $f_{X|V}(x|v)$ and $f_{V|X}(v|x) \equiv p_v(x)$. We balance $\mathcal{D}$ by resampling so that it contains approximately equal numbers of each phoneme. The output of the kth softmax output node of our CNN given input x_n is written $\hat{y}_k(x_n)$, with the softmax function given by

$$\hat{y}_k(x) = \frac{\exp(h_k(x))}{\sum_{i=1}^{K} \exp(h_i(x))} \tag{4}$$

where $h_k(x)$ is the total input received by the kth output node from the dense layer which precedes it. The categorical cross entropy loss function[49] is

$$E = -\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} I_k(v_i) \ln \hat{y}_k(x_i). \tag{5}$$

The above average over the training data $\mathcal{D}$ approximates an average over the joint distribution f_{XV}. Using the fact that for any bivariate function g, $\mathbb{E}(g(X, V)) = \mathbb{E}(\mathbb{E}(g(X, V)|X))$ then, as $n \to \infty$, we have

$$E \sim -\sum_{x,v} f_{XV}(x, v) \sum_{k=1}^{K} I_k(v) \ln \hat{y}_k(x) \tag{6}$$

$$= -\sum_{x} f_X(x) \sum_{k=1}^{K} \mathbb{E}\left(I_k(V) \ln \hat{y}_k(X)|X = x\right) \tag{7}$$

$$= -\sum_{x} f_X(x) \sum_{k=1}^{K} p_k(x) \ln \hat{y}_k(x). \tag{8}$$

The training process seeks to adapt the network weights to find the functions $\hat{y}_x(x)$ which minimize E. Suppose that the network is sufficiently flexible and there are sufficient training data, so that it can come close to finding the $\hat{y}_k(x)$ which minimize the limiting form (8). In this case the outputs must, for any x, minimize

$$-\sum_{k=1}^{K} p_k(x) \ln \hat{y}_k(x), \tag{9}$$

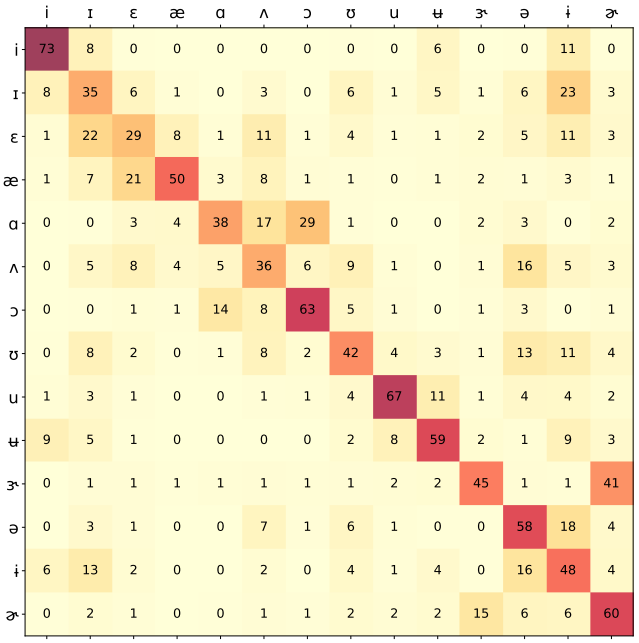

FIG. 3. Similarity matrix S estimated from our trained CNN. Each row i gives the conditional probabilities $\mathbb{P}(V_l = j|V_t = i)$ (as percentages) with $i, j \in \{1, 2, 3, \ldots, K\}$. Numerical phoneme labels have been replaced with IPA symbols for interpretability.

which occurs when $\hat{y}_k(x) = p_k(x)$. The outputs of the optimal CNN, in the limit of large n, given input X, will therefore be the conditional phoneme probabilities $p_k(X)$, defined in equation (2). This is the interpretation we attach to them throughout the paper. Accordingly, the classification rule which predicts the phoneme corresponding to utterance x as

$$v(x) = \text{argmax}_k \hat{y}_k(x) \tag{10}$$

is the Bayes classifier[50], which is optimal in the sense that its error rate is at least as low as any other classifier.

Viewing the CNN as a proxy for a human listener, the quantity $\hat{y}_k(X) \equiv p_k(X)$ represents the listener's degree of belief that acoustic output X was intended to represent phoneme k. This allows us to measure the similarity between two phonemes as the listener's degree of belief that phoneme j was intended by the talker given that i was uttered, averaged over all utterances of i. Writing V_t for the phoneme uttered, and V_l for the phoneme heard (predicted) by the listener, we can write this similarity

$$S_{ij} = \sum_{x} \mathbb{P}(V_l = j|X = x)\mathbb{P}(X = x|V_t = i) \tag{11}$$

$$= \sum_{x} p_j(x) f_{X|V}(x|i) \tag{12}$$

$$= \mathbb{P}(V_l = j|V_t = i). \tag{13}$$

This defines a stochastic (unit row sums) similarity matrix with rows equal to the average output of the CNN for each input phoneme. The result is displayed in Figure 3.

We now investigate how well the behaviour of our CNN approximates that of a human listener. Human-perceptual similarities between phonemes may be measured experimentally by asking listeners to identify sounds selected at random from some predefined set of vowels or consonants[3,38,51]. When hearing a phoneme, participants must guess what member of the predefined set it represents. The responses of a group of listeners may be summarized by their collective confusion matrix, the rows of which give the probabilities that a single phoneme will be identified as each of the others in the test set. In the studies of Zahorian and Jagharghi[3] and Cutler et al.[38], which determine confusion matrices for American English vowels, the effect of acoustic context (which can include preceding or following phonemes or parts thereof), had a small but statistically significant effect on the error rate ($< 6\%$). Zahorian[3] also compared human confusion matrices to machine classifiers, finding the greatest similarity (correlation coefficient 0.74 between machine and human confusion matrices) when using time series of the discrete cosine of the spectrum as predictors in a simple neural network.

The experimental conditions, phoneme contexts, and phoneme sets used in experiments on human listeners differ between studies, and from the TIMIT data used to train our CNN. For this reason, we cannot directly compare confusion probabilities from different experiments. An alternative is to compare the rankings of confusion probabilities within the intersection of the phoneme sets used in two studies. By definition such comparisons are invariant to factors (e.g. context, noise, phoneme set) which change the magnitudes of probabilities without changing their order. We perform two comparisons: first between our CNN and the larger and more recent human study (Cutler[38]), and then for reference, between the two human studies. Table II lists, for each of the nine monophthongs in both Cutler and TIMIT, the five most common predictions made by human listeners and by our CNN. We also list Spearman's rank correlation coefficient for each phoneme, along with the corresponding p values, which provide strong evidence against the null hypothesis of no correlation. The average correlation over all phonemes is $\bar{\rho} = 0.80$. Inspection of Table II reveals a strong correspondence between the order of phonemes identified by the two methods, with differences in ranking often involving the exchange of phonemes which occupy nearby locations on the IPA chart. To benchmark the value of Spearman's ρ we make the same comparison between Cutler and the steady vowel confusion matrix of Zahorian (this gives the closest correspondence to TIMIT, and the most similar overall identification rate to Cutler). We find $\bar{\rho} = 0.73$. The fact that the correlation between the most detailed human study and the CNN is greater than the correlation between the two human studies indicates that our CNN provides a good approximation to human perceptual characteristics.

To summarize this subsection: we used a well known CNN architecture[41,42] to create a full spectrogram connected-speech phoneme classifier, which we showed to have human-like perceptual characteristics. We also defined a metric of similarity: the average degree of belief among listeners that utterances of phoneme j by talkers were intended as phoneme i. We now explain how this classifier and perceptual similarity metric can be used to embed phonemes in $\mathbb{R}^d$, and then measure individual utterances by mapping them to the same space.

TABLE II. For each of the nine monophthongs occurring in both TIMIT and Cutler[38] (VC pattern), the five most common predicted phonemes from the same set are listed in order of the frequency with which they are predicted by humans and our CNN. Spearman's rank correlation coefficient, ρ, calculated from all nine ranks is also given, with average over all phonemes $\bar{\rho} = 0.80$.

Phoneme	ρ (p-val)	Listener	1	2	3	4	5
ɑ	0.88 (0.002)	Human	ɑ	ɔ	ʌ	æ	ɪ
		CNN	ɑ	ɔ	ʌ	æ	ɛ
æ	0.67 (0.047)	Human	æ	ɛ	ɔ	ɪ	ɑ
		CNN	æ	ɛ	ʌ	ɪ	ɑ
ʌ	0.73 (0.024)	Human	ʌ	ɑ	ɔ	æ	ʊ
		CNN	ʌ	ɔ	ɛ	ɑ	ʊ
ɔ	0.84 (0.005)	Human	ɔ	ɑ	ʌ	ʊ	æ
		CNN	ɔ	ɑ	ʌ	ʊ	u
ɛ	0.90 (0.001)	Human	ɛ	ɪ	æ	ʌ	i
		CNN	ɛ	ɪ	ʌ	æ	ʊ
ɪ	0.81 (0.009)	Human	ɪ	ɛ	i	u	ʌ
		CNN	ɪ	i	ɛ	ʊ	ʌ
i	0.88 (0.002)	Human	i	ɪ	ɛ	u	ɔ
		CNN	i	ɪ	ɛ	ʊ	u
ʊ	0.61 (0.08)	Human	ʊ	ʌ	u	ɑ	ɔ
		CNN	ʊ	ʌ	ɪ	u	ɛ
u	0.89 (0.001)	Human	u	ʊ	i	ʌ	ɔ
		CNN	u	ʊ	ɪ	i	ɔ

B. Embedding method

In section II A we assumed that spectrograms were drawn from the high dimensional conditional distributions $f_{X|V}(x|v)$. Our embedding method uses a low dimensional model of these distributions $\psi_v(z)$, where $z \in \mathbb{R}^d$. Our final measurement procedure is defined as a function $\mathbb{R}^N \to \mathbb{R}^d$ which maps spectrograms to points in measurement space. Here we have used $N(= 2.5 \times 10^3)$ to denote the dimension of acoustic (spectrogram) space. The first step of this process is to determine the locations of the low dimensional distributions. In general we can allow the ψ_v to be arbitrarily complicated functions, but for simplicity we here define them as d dimensional Gaussians

$$\psi_v(z) = \frac{e^{-\frac{1}{2\sigma^2}|z-\mu_v|^2}}{(2\pi)^{\frac{d}{2}}\sigma^d} \tag{14}$$

with means $\mu_1, \mu_2, \ldots, \mu_K$, and variance σ^2. We view these functions as idealized low dimensional models of phoneme distributions, and z as an idealized acoustic variable. We

calculate the perceptual similarities between these idealized phonemes by direct analogy with formula (13)

$$\tilde{S}_{ji} = \int_{\mathbb{R}^d} \tilde{p}_i(z)\psi_j(z)dz = \int_{\mathbb{R}^d} \frac{\psi_i(z)\psi_j(z)}{\sum_k \psi_k(z)}dz. \quad (15)$$

Here $\tilde{p}_i$ and $\tilde{S}_{ji}$ denote conditional probabilities and similarities in the idealized low dimensional model. The integral in (15) is analytically intractable and prohibitively slow to perform using quadrature, so we estimate by monte carlo[24]. Because we chose the distributions $\psi_v(z)$ to all have the same variance, the similarities so defined are symmetric, making $\tilde{S}$ doubly stochastic. To select the means μ_k, which we refer to as the "standard positions" of the phonemes, we minimize the difference between a symmetrized version of the similarity matrix and its idealized form. Letting D_K denote the set of $K \times K$ doubly stochastic matrices, we define the symmetrized similarity matrix as the doubly stochastic matrix with minimum total squared element-wise deviation from S

$$\mathcal{S} = \underset{M \in D_K}{\text{argmin}} \sum_{ij} (S_{ij} - M_{ij})^2. \quad (16)$$

This matrix may be found by numerical optimization.

Starting from some initial condition $\{\mu_1(0), \ldots, \mu_K(0)\}$, we evolve the phoneme locations through time using the following dynamics

$$\dot{\mu}_i = \eta \sum_{k=1}^{K} (\mathcal{S}_{ik} - \tilde{S}_{ik}) \frac{\mu_k - \mu_i}{|\mu_k - \mu_i|} \quad (17)$$

which pulls together phonemes whose idealized similarity is less than the true (symmetrized) similarity, and repels phonemes whose idealized similarity is greater. From Figure 3, we see that the similarities of many phonemes are very small, meaning that the only constraint on their relative position is that they must be separated by a distance significantly greater than σ. This flexibility means that there may be many equilibria of (17) which have similar levels of perceptual accuracy, but lack an overall pattern which can be meaningfully understood in terms of traditional linguistic variables such as height and backness. We return to this point in section III A. We will refer to the final configuration of standard positions, once our embedding dynamics (17) has reached equilibrium, as a vowel chart.

C. Measurement method

Having determined the standard positions of the idealized phonemes, we now define a mapping $\mathbb{R}^N \to \mathbb{R}^d$ which allows us to measure individual utterances from their spectrograms. Given a spectrogram $X \in \mathbb{R}^N$ we first pass it to our CNN, which generates degrees of belief that X belongs to each possible phoneme. We write the vector of these degrees of belief as $p(X) \in \Delta^K$, where Δ^K is the K dimensional simplex. Our aim is now to find a point $z \in \mathbb{R}^d$ for which the corresponding vector of idealized conditional probabilities $\tilde{p}(z)$ best approximates $p(X)$. This optimal fit is obtained by viewing z as the parameter of a probability model for $p(X)$. That is, we view the probability mass function $p(X) \in \Delta^K$ as having been drawn from a probability distribution parameterized by z. In doing so we are thinking of $p(X)$ as the random variable, rather than the spectrogram X. The Dirichlet distribution[52], which has K parameters and the simplex Δ^K as its support, is a natural choice for the distribution of $p(X)$ (it is a probability distribution of probability distributions). Its density function is

$$f(p;\alpha) = \frac{\prod_{i=1}^K p_i^{\alpha_i - 1}}{B(\alpha)} \quad (18)$$

where $\alpha = (\alpha_1, \ldots, \alpha_K)$ and B is the multivariate Beta function

$$B(\alpha) = \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma(\alpha_1 + \ldots + \alpha_K)}. \quad (19)$$

We define

$$\alpha_i = A\tilde{p}_i(z) \quad (20)$$

or equivalently $\alpha = A\tilde{p}(z)$, where $A > 0$ is a constant which controls how sharply peaked the distribution is. By this definition, $\mathbb{E}(p(X)|X) = \tilde{p}(z)$, and the peak of the distribution lies at

$$p_i = \frac{\tilde{p}_i(z) - \frac{1}{A}}{1 - \frac{K}{A}} = \tilde{p}_i(z) + \frac{K\tilde{p}_i(z) - 1}{A} + O(A^{-2}) \quad (21)$$

in the ith dimension, as $A \to \infty$.

We show in section III B that our measurement method is maximally self consistent, in the sense that the distribution of measurements is as close as possible to the idealized phoneme distributions, when $A \approx 10^3$. In this case the peak of the Dirichlet distribution is located very close to $\tilde{p}(z)$.

The likelihood of the data $p(X)$ given the Dirichlet model is $\mathcal{L}(z|p(X)) = f(p(X); A\tilde{p}(z))$, and the point in $\mathbb{R}^d$ to which X is mapped is determined by maximum likelihood

$$\hat{z} = \text{argmax}_z \mathcal{L}(z|p(X)). \quad (22)$$

Intuitively the optimization process operates as follows. There is a subset, C, of the simplex which is accessible to the idealized model in the sense that for any point in $u \in C$ there is a value z for which $\tilde{p}(z) = u$. Formally,

$$C = \{u \in \Delta^K | u = \tilde{p}(z) \text{ for some } z \in \mathbb{R}^d\}. \quad (23)$$

The point $p(X) \in \Delta^K$ may or may not lie in C. By maximizing the Dirichlet likelihood, we are finding the point $\tilde{p}(\hat{z})$ in C for which the Dirichlet distribution $f(p; A\tilde{p}(\hat{z}))$ has maximum probability weight at $p = p(X)$. Since the peak of the distribution lies close to $\tilde{p}(z)$, this will occur when $\tilde{p}(z)$ is close to $p(X)$.

III. RESULTS

We apply our embedding and measurement method to the following set of 12 TIMIT monophthongs

$$\mathcal{P} = \{\text{i}, \text{ɨ}, \text{ɪ}, \text{ɛ}, \text{æ}, \text{ɑ}, \text{ɔ}, \text{ʊ}, \text{u}, \text{ʉ}, \text{ʌ}, \text{ə}\}. \quad (24)$$

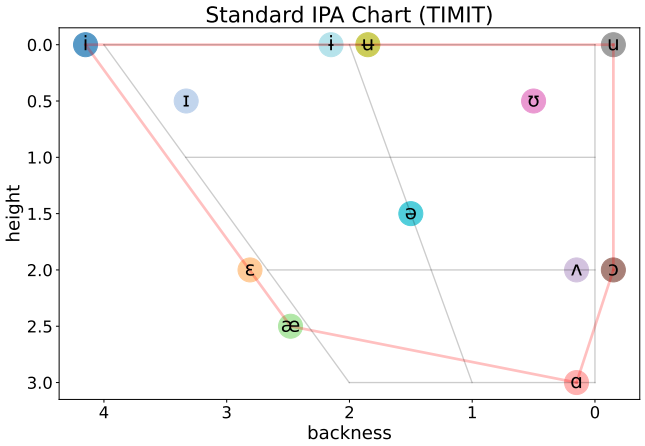

FIG. 4. Standard IPA chart showing the monophthongs in TIMIT, excluding ɚ, ɝ and ᵊ. These positions are used as the initial condition for the perceptual chart dynamics. The red polygon shows the convex hull[53] of these positions. The hull area is 12.73. Total variation distance between chart and CNN: $\delta(\tilde{S}, \mathcal{S}) = 0.276$, occurring when $\sigma = 0.52$.

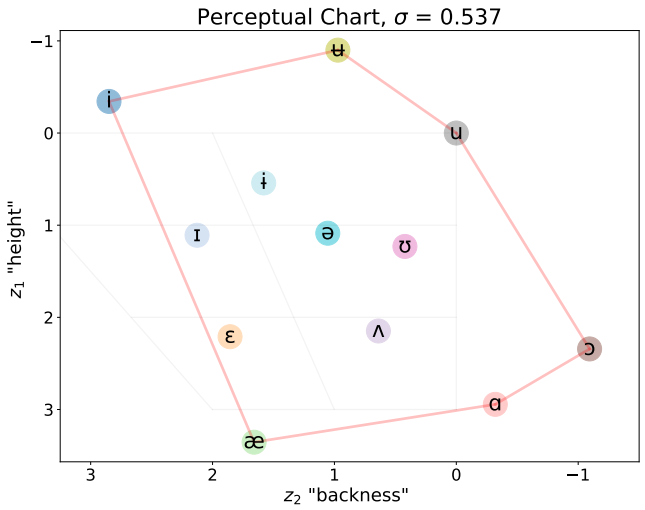

FIG. 5. Perceptual IPA chart showing the monophthongs in TIMIT, excluding ɚ, ɝ and ᵊ. These positions are equilibria of equation (17). The hull area is 12.77. Total variation distance to CNN $\delta(\tilde{S}, \mathcal{S}) = 0.11$.

We have excluded the r-coloured forms of ə and ɜ because capturing this type of variation simultaneously with the acoustic variations which distinguish the members of $\mathcal{P}$ is not easily achieved in two dimensions.

A. Perceptual vowel chart

To make our chart intuitive to interpret, we wish to find an equilibrium of equation (17) which, as far as possible, mimics the pattern of sounds in the standard IPA chart[35] (Figure 4). We write the IPA positions of the ith phoneme as μ_i^{IPA}. To find an equilibrium configuration of points with a similar pattern to the IPA chart, we initialize the standard positions at their IPA values. To maintain an approximate equivalence between the units of measurement used in the IPA and perceptual charts we adjust the variance (σ^2) of the idealized phoneme distributions during the embedding process, so that the area of the perceptual chart's convex hull (the smallest convex polygon which encloses all points) remains equal to the hull area of the IPA. In previous work the convex hull has been used as a measure of acoustic variation in vowel production by individual speakers[54], and is similar to the notion of vowel space area[55]. The result of the embedding process is shown in Figure 5, where we have performed a final constant coordinate shift so that [u] is located at the origin of the coordinate system.

To quantify the perceptual accuracy of different charts we compute the average total variation distance[56] between the rows of the idealized perceptual similarity matrix $\tilde{S}$, calculated from the standard positions $\{\mu_k\}$, and the symmetrized CNN similarity matrix $\mathcal{S}$. In general the total variation distance between two probability densities or mass functions is defined

$$\delta(f,g) = \begin{cases} \frac{1}{2}\sum_x |f(x) - g(x)| & \text{discrete case} \\ \frac{1}{2}\int |f(x) - g(x)| dx & \text{continuous case,} \end{cases} \tag{25}$$

and satisfies $\delta(f,g) \in [0,1]$, with $\delta(f,f) = 0$. Our average measure is defined

$$\bar{\delta}(\tilde{S}, \mathcal{S}) = \frac{1}{K}\sum_{i=1}^{K} \delta(\tilde{S}_{i*}, \mathcal{S}_{i*}) \tag{26}$$

where S_{i*} denotes the ith row of S (the conditional mass function $\mathbb{P}(V_l = *|V_t = i)$). As a baseline measure, we generated a large ensemble of random charts, with phoneme positions drawn from the standard bivariate normal density. For each chart, $\bar{\delta}(\tilde{S}, \mathcal{S})$ was minimized with respect to the variance, σ^2, of the idealized phonemes. The resulting distance, $\bar{\delta} = 0.37$, is a measure of the perceptual accuracy of charts which are constructed without regard for perceptual accuracy. By comparison, for the standard IPA chart (Figure 4), $\bar{\delta}_{\text{IPA}} = 0.28$ and for our embedding (Figure 5), $\bar{\delta}_{\text{percept.}} = 0.11$. Figure 6 shows the evolution of $\bar{\delta}$ during the embedding process, plotted against the mean euclidean distance of the standard positions from their IPA counterparts. The improvement of the perceptual chart with respect to the IPA is substantially greater than the improvement of the IPA with respect to the random baseline.

B. Measurement behaviour

To define our measurement processes we must set the free parameter A (equation (20)), which controls the sharpness of the peak of our probability model for the output $p(X)$ of our CNN. Given the test data $\mathcal{D}_{\text{test}} =$

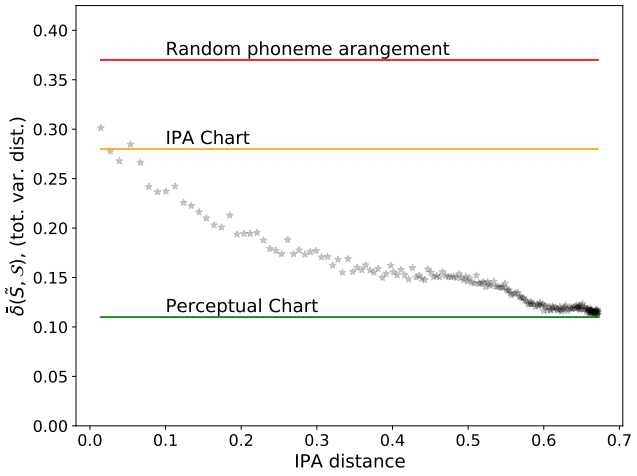

FIG. 6. Evolution of total variation distance between CNN and perceptual chart as the mean distance from IPA positions increases. Horizontal lines show total variation distance for (i) randomly distributed points $\bar{\delta} = 0.37$ [red] (ii) the IPA chart $\bar{\delta} = 0.28$ [yellow] and (iii) the perceptual chart $\bar{\delta} = 0.11$ [green].

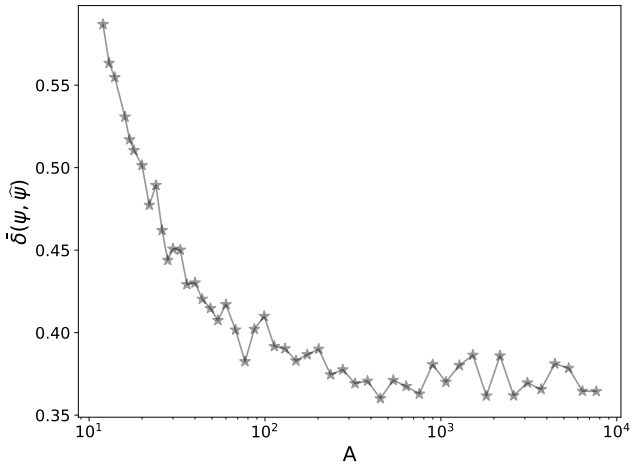

FIG. 7. Dependence on statistical model parameter A (see equation (20)) of the mean total variation distance been idealized and measured phoneme distributions.

$((x_1, v_1), (x_2, v_2), \ldots)$, and a value of A, we can compute the corresponding set of chart positions $\mathcal{D}_z = (z_1, z_2, \ldots)$ using the measurement process defined in section II C. For each phoneme label, v, in the test data, we can use the corresponding positions in $\mathcal{D}_z$ to generate a kernel density estimate[6] of the probability density function, written $\widehat{\psi}_v(z; A)$, of that phoneme on the perceptual chart. The parameter A has been included as an explicit argument because of its effect on the measurement process. We can then calculate the total variation distance between this density and the equivalent idealized phoneme distribution $\psi_v(z)$

$$\delta_v(A) := \delta(\psi_v(*), \widehat{\psi}_v(*; A)) \tag{27}$$

$$= \frac{1}{2} \int_{\mathbb{R}^2} |\psi_v(z) - \widehat{\psi}_v(z; A)| dz. \tag{28}$$

Since our embedding and measurement method is based on the approximation of high dimensional acoustic distributions with an idealized low dimensional model, we desire the resulting phoneme distributions to be as close to this model as possible. We therefore choose A to minimize the average of $\delta_v(A)$ over all phonemes. The optimal A-value therefore satisfies

$$A^* = \underset{A>0}{\text{argmin}} \frac{1}{K} \sum_{v=1}^{K} \delta_v(A) \tag{29}$$

$$= \underset{A>0}{\text{argmin}}\, \bar{\delta}(A) \tag{30}$$

and the measured phoneme density is defined

$$\widehat{\psi}_v(z) := \widehat{\psi}_v(z; A^*). \tag{31}$$

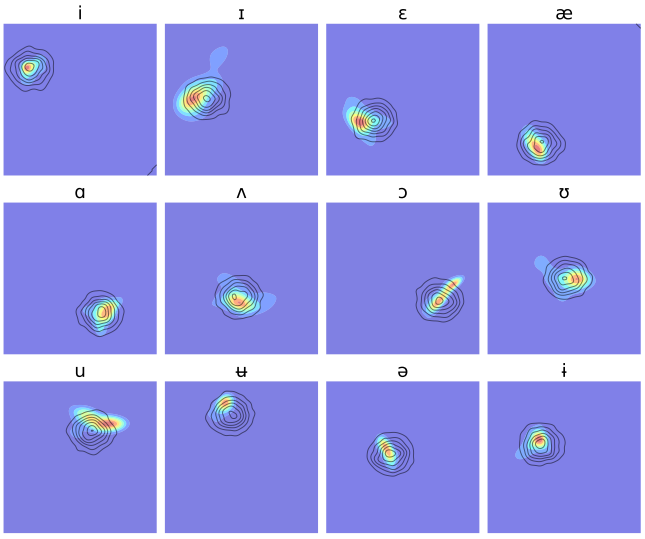

FIG. 8. Heat maps showing kernel density estimates of each phoneme measured from the TIMIT test data using $A^* = 10^3$. Contour plots show idealized phoneme densities (generated by forming kernel density estimates from 2000 samples from each $\psi_v(z)$).

The dependence of $\bar{\delta}(A)$ on A is shown in Figure 7, where we see that the performance of the chart increases until $A \approx 10^3$. We set $A^* = 10^3$ henceforth. The idealized and measured phoneme distributions are shown in Figure 8.

Figure 9 is a scatter plot of the individual measured positions of the TIMIT phonemes, coloured according to the phonemes which generated them. Mean measured positions for each phoneme are typically shifted toward the centre of the chart compared to their modes. When the CNN attaches a high probability to the wrong phoneme, a measurement is generated which has a large deviation (compared to σ) from

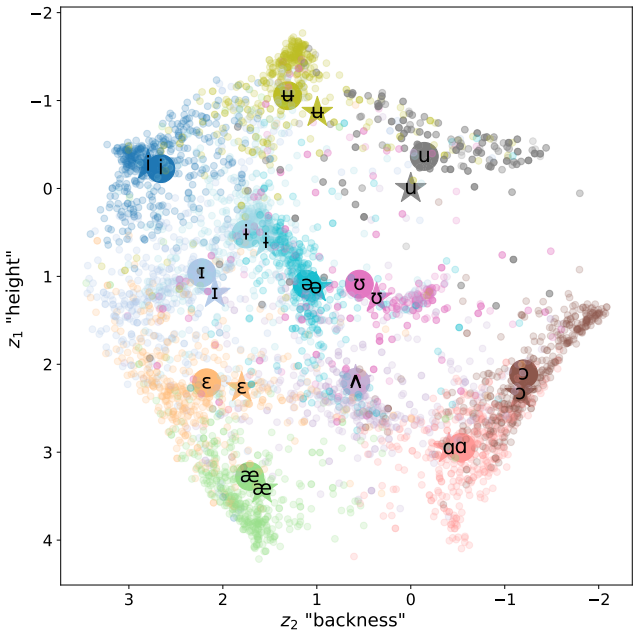

FIG. 9. Scatter plot of phonemes measured from TIMIT test data. Phonetically labelled dots show average positions of each measured phoneme. Labelled stars show corresponding standard positions.

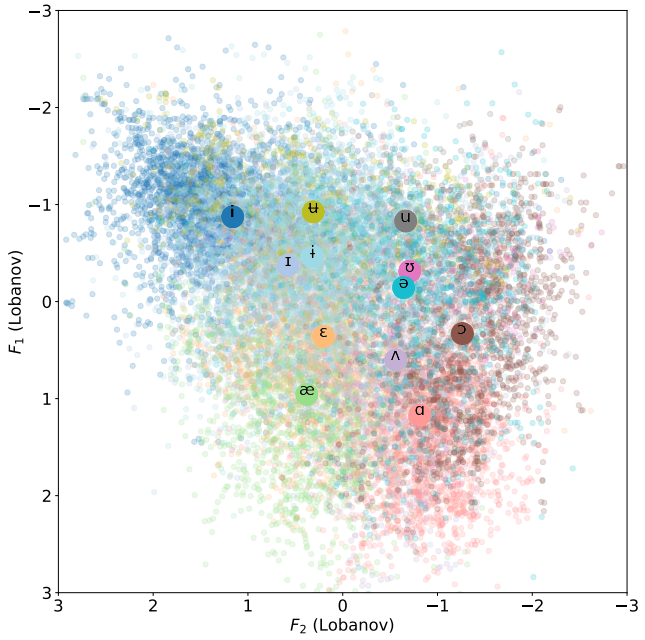

FIG. 10. Lobanov normalized formant measurements (32) from all time points in the VTR database[22] which lie within the phoneme utterances from the set $\mathcal{P}$. Formant points are coloured according to their phoneme, and phonetically labelled circles show mean formant locations for each phoneme.

the standard position. Such deviations skew the measured distributions toward the chart's centre of mass.

It is instructive to compare the results of our measurement procedure with the VTR Formants Database[22]. This is a manually annotated set of formant measurements from TIMIT which serves as a "ground-truth" benchmark for the development and testing of formant measurement methods[57]. Figure 10 shows the distribution of F_1 and F_2 values, standardized by speaker (Lobanov normalized), for the same set of phonemes used in our study. Letting $\bar{F}_i(s)$ and $\sigma_i^2(s)$ be the mean and variance of all measurements of the ith formant of speaker s within the set of 12 phonemes in $\mathcal{P}$, the Lobanov normalized formants for this speaker are defined

$$F_i^{\text{lob}} = \frac{F_i - \bar{F}_i(s)}{\sigma_i(s)}. \tag{32}$$

The mean standardized formant positions for each phoneme in Figure 10 form a similar pattern to the phonemes in our perceptual chart. That the two measurement systems achieve such similar average results is remarkable (perhaps), and emphasizes the effectiveness of formants in capturing the perceptual characteristics of vowels. However, even after speaker normalization, the formant distributions are more dispersed than the results of our measurement procedure. We quantify this effect using the ratio of the average area occupied by each phoneme to the area, A_H, of the convex hull surrounding their mean locations. Phoneme area is defined

$$A_P = \pi \left(\mathbb{V}(F_1^{\text{lob}}) + \mathbb{V}(F_2^{\text{lob}}) \right) \tag{33}$$

where $\mathbb{V}$ denotes sample variance. That is, A_P is the area of the circle with radius equal to the mean squared displacement of the formants of a phoneme from its mean formant position. The phoneme area ratio for phoneme v is then defined

$$\phi(v) = \frac{A_P(v)}{A_H}. \tag{34}$$

This measures the fraction of chart area taken up by phoneme v. In Figure 10 the average area ratio is $\bar{\phi}_F = 0.466$, compared to $\bar{\phi}_Z = 0.210$ in our measurement system (Figure 9). The ability to make accurate phoneme measurements in terms of continuous variables is one of the potentially useful properties of the procedure we have defined.

C. Application

The generation of perceptually accurate low dimensional measurements from high dimensional acoustic representations has potential applications. For example, if we trained our CNN to recognize words from single vowel phonemes, then we would expect our technique to map sets of words sharing a common phoneme to nearby points on our chart. In this way, our mapping could be used as a means of automatic phoneme discovery and cataloguing. By expanding the number of sound categories on which our CNN was trained so that it was more granular and comprehensive, we could increase the ability of our measurement system to pick up subtle differences in utterances. This would provide a means to measure

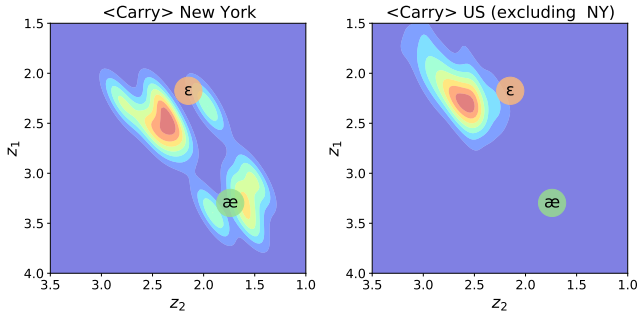

FIG. 11. Kernel density estimates of the probability distribution of vowel phonemes within utterances of the word ⟨Carry⟩ in TIMIT test data. Left plot shows the distribution for the New York dialect area, and right map shows the distribution for all other dialect areas combined.

dialect or accent variation, detect and measure speech disorders, or to allow speakers to train their voices using visual feedback.

As a concrete example we illustrate how our measurements can be used to analyse dialect variation within TIMIT. Within the Cambridge online survey of world Englishes (COSWE)[58] (with $\approx 6 \times 10^4$ geographically tagged respondents from the USA), answers to the question: "How do you pronounce Mary/merry/marry?" show that the first vowel phonemes in these words have merged for most US mainland speakers, except for a distinctive hold-out area in the Northeast, centred on Boston, New York City, and Philadelphia. As a result, words like ⟨Carry⟩, which is relatively common in TIMIT, have retained the original non-rhoticized æ in this geographical area. Figure 11 shows kernel density estimates calculated from all measurements of phonemes (from $\mathcal{P}$) occurring in utterances of ⟨Carry⟩ in the TIMIT test data, separated into the New York dialect area, and the rest of the USA. The presence of a cluster of sounds centred on the standard mean location of æ is consistent with the results of COSWE. The accurate detection of dialect variations from spectrograms of connected speech would be useful in the study of language variation and change, and in situations where known social-geographical variations in speech are used to infer information about the speaker.

IV. DISCUSSION

Over the last century our scientific understanding of speech has been advanced by the development of methods which allow speech sounds to be quantitatively measured. In the study of vowels, formants have been a particularly useful measurement tool due to their fundamental connection to the resonant nature of these sounds, and their intuitive (albeit imprecise) definition. Whereas formants are used primarily as a means for humans to gain insight into speech and language, when training a machine to make predictions from speech, full spectrum measures such as spectrograms or MFCCs are more effective as inputs. However, these measures cannot be usefully interpreted by humans. Modern machine learning methods have made it possible to map raw and complex inputs onto human interpretable outputs.

Our aim was to exploit this new capability to develop a perceptual measurement system for vowel sounds. By training a convolutional network to recognize phonemes, and then mimicking its behaviour with a low dimensional model, we have been able to define a perceptually meaningful mapping from high-dimensional acoustic space to a human-readable measurement chart. Our system has the desirable properties of being perceptually realistic, structurally similar to traditional charts, automatically speaker normalized, and less dispersed than formant measures. The system is also easy to generalize by training the CNN to recognize more and different categories of sounds, and to search for phonemes within a language. An ambitious goal might therefore be to simultaneously identify the distinctive units of sounds in the union of many languages, train a CNN to recognize them, and then define a cross linguistic measurement system following the same steps that we have presented in our simple case.

ACKNOWLEDGMENTS

This research was supported by a Royal Society APEX award APX\R1\180117 (funded by the Leverhulme Trust).

[1]T. Chiba and M. Kajiyama, The Vowel: Its Nature and Structure, 115–154 (Tokyo-Kaiseikan, Tokyo).

[2]C. H. Shadle, H. Nam, and D. H. Whalen, "Comparing measurement errors for formants in synthetic and natural vowels," The Journal of the Acoustical Society of America 139(2), 712–727 (2016).

[3]S. A. Zahorian and A. J. Jagharghi, "Spectral-shape features versus formants as acoustic correlates for vowels," The Journal of the Acoustical Society of America 94(4), 1966–1982 (1993).

[4]M. Molis, "Evaluating models of vowel perception," The Journal of the Acoustical Society of America 118(2), 1062–1071 (2005).

[5]I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning (MIT Press, 2016) http://www.deeplearningbook.org.

[6]T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. (Springer, New York, 2008).

[7]U. Kamath, J. Liu, and J. Whitaker, Deep Learning for NLP and Speech Recognition (Springer, Switzerland, 2019).

[8]Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature 521, 436–444 (2015).

[9]J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continuous speech corpus cdrom. nist speech disc 1-1.1," NASA STI/Recon Technical Report No. 93 (1993).

[10]R. N. Shepard, Psychological representation of speech sounds, Chap. 4, 67–113 (McGraw-Hill,, New York).

[11]L. C. W. Pols, L. J. T. van der Kamp, and R. Plomp, "Perceptual and physical space of vowels sounds," The Journal of the Acoustical Society of America 46(2), 458–467 (1969).

[12]G. E. Peterson and H. Barney, "Control methods used in a study of the vowels," The Journal of the Acoustical Society of America 24(2), 175–184 (1952).

[13]ANSI, ANSI/ASA (Acoustical Society of America, Melville, NY, 2013).

[14]I. R. Titze et al., "Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization," The Journal of the Acoustical Society of America 137(5), 3005–3007 (2015).

[15]G. Fant, The Acoustic Theory of Speech Production (Mouton, The Hague, 1960).

[16]B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," The Journal of the Acoustical Society of America 50(2), 637–655 (1971).

[17]F. Nenadic, P. Coulter, T. M. Nearey, and M. Kiefte, "Perception of vowels with missing formant peaks," The Journal of the Acoustical Society of America 148(4), 1911–1921 (2020).

[18]C. G. Clopper, D. B. Pisoni, and K. de Jong, "Acoustic characteristics of the vowel systems of six regional varieties of american english," The Journal of the Acoustical Society of America 118(3), 1661–1676 (2005).

[19]J. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of american english vowels," The Journal of the Acoustical Society of America 97(5), 3099–3111 (1995).

[20]J. Allen, M. S. Hunnicutt, and D. H. Klatt, From Text to Speech, 108–122 (Cambridge University Press, Cambridge).

[21]J. Lee, S. Shaiman, and G. Weismer, "Relationship between tongue positions and formant frequencies in female speakers," The Journal of the Acoustical Society of America 139(1), 426–440 (2016).

[22]L. Deng, X. Cui, R. Pruvenok, Y. Chen, S. Momen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," Proceedings of 2006 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1 (2006).

[23]W. Styler, "On the acoustical features of vowel nasality in english and french," The Journal of the Acoustical Society of America 142(4), 2469–2482 (2017).

[24]W. H. Press, S. A. Teukolsky, and W. T. Vetterling, Numerical Recipes: The Art of Scientific Computing, 3rd ed. (Cambridge University Press, Cambridge, UK, 2007).

[25]Z. Prusa and P. Rajmic, "Toward high-quality real-time signal reconstruction from stft magnitude," IEEE Signal Processing Letters 24(6), 892–896 (2017).

[26]K. S. Rao and K. E. Manjunath, Speech Recognition Using Articulatory and Excitation Source Features (Springer, New York, 2017).

[27]S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," The Journal of the Acoustical Society of America 8, 185–190 (1937).

[28]A. Mohamed and G. Hinton, "Phone recognition using restricted boltzmann machines," IEEE International Conference on Acoustics, Speech and Signal Processing 4354–4357 (2010).

[29]A. Hannun et al., "Deep speech: Scaling up end-to-end speech recognition" (2014), arXiv:1412.5567.

[30]D. Palaz, R. Collobert, and M. Magimai.-Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," in Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH) (2013).

[31]D. Palaz, M. Magimai-Doss, and R. Collobert, "End-to-end acoustic modeling using convolutional neural networks for hmm-based automatic speech recognition," Speech Communication 108, 15–32 (2019).

[32]K. Johnson, Speaker Normalization in Speech Perception, Chap. 15, 363–389 (Blackwell Publishing, Malden, MA).

[33]J. Werker and R. Tees, "Phonemic and phonetic factors in adult cross-language speech perception," The Journal of the Acoustical Society of America 75(6), 1866–1878 (1984).

[34]ARPAbet description available at https://en.wikipedia.org/wiki/ARPABET.

[35]IPA chart available at http://www.internationalphoneticassociation.org/content/ipa-chart.

[36]IPA, Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet (Cambridge University Press, Cambridge, UK, 1999).

[37]D. Jurafsky and J. H. Martin, Speech and Language Processing: Second Edition (Pearson, Harlow, Essex, UK, 2013).

[38]A. Cutler, A. Weber, R. Smits, and N. Cooper, "Patterns of english phoneme confusions by native and non-native listeners," The Journal of the Acoustical Society of America 116(6), 3668–3678 (2004).

[39]D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," The Journal of Physiology 195, 215–243 (1968).

[40]K. Fukushima, "Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," Biological Cybernetics 36, 193–202 (1980).

[41]Y. LeCun et al., "Backpropagation applied to handwritten zip code recognition," Neural Computation 1(4), 541–551 (1989).

[42]A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," Communications of the ACM 60(6), 84–90 (2017).

[43]M. Mitchell, Artificial Intelligence: A Guide for Thinking Humans (Pelican, London, 2019).

[44]M. Algabri, H. Mathkour, M. A. Bencherif, M. Alsulaiman, and M. A. Mekhtiche, "Towards deep object detection techniques for phoneme recognition," IEEE Access 8, 54663 – 54680 (2020).

[45]M. Algabri, H. Mathkour, M. Alsulaiman, and M. A. Bencherif, "Deep learning-based detection of articulatory features in arabic and english speech," Sensors 21(4), 1205 (2021).

[46]C. Glackin, J. Wall, G. Chollet, N. Dugan, and N. Cannings, "Convolutional neural networks for phoneme recognition," in Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2018), (2018), pp. 190 – 195.

[47]Q. Zhou, J. Shan, W. Ding, C. Wang, S. Yuan, F. Sun, H. Li, and B. Fang, "Cough recognition based on mel-spectrogram and convolutional neural network," Frontiers in Robotics and AI 8, 580080 (2021).

[48]S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition" (2019), arXiv:1904.05862.

[49]K. P. Murphy, Machine Learning: A Probabilistic Perspective (MIT Press, 2012).

[50]L. Wasserman, All of Statistics: A Concise Course in Statistical Inference (Springer, New York, 2010).

[51]G. A. Miller and P. E. Nicely, "An analysis of perceptual confusions among some english consonants," The Journal of the Acoustical Society of America 27(2), 338–352 (1955).

[52]A. Gelman et al., Bayesian Data Analysis, 3rd ed. (CRC Press, Boca Raton, FL, 2014).

[53]M. de Berg et al., Computational Geometry, 3rd ed. (Springer, New York, 2008).

[54]B. H. Story and K. Bunton, "Vowel space density as an indicator of speech performance," JASA Express Letters 141(5), 458–464 (2017).

[55]G. Fant, Speech Sounds and Features (MIT Press, Cambridge, 1973).

[56]I. Csiszar and P. Shields, Information Theory and Statistics: A Tutorial (Now, Boston, MA, 2004).

[57]Y. Dissen, J. Goldberger, and J. Keshet, "Formant estimation and tracking: A deep learning approach," The Journal of the Acoustical Society of America 145(2) (2019).

[58]B. Vaux and M. Jøhndal, "Cambridge online survey of world englishes" available at http://www.tekstlab.uio.no/cambridge_survey.

+