Available online at www.sciencedirect.com
ScienceDirect
Computer Speech and Language xxx (2015) xxx–xxx
COMPUTER SPEECH AND LANGUAGE
www.elsevier.com/locate/csl

Directly data-derived articulatory gesture-like representations retain discriminatory information about phone categories☆

Vikram Ramanarayanan*, Maarten Van Segbroeck, Shrikanth S. Narayanan

Signal Analysis and Interpretation Lab, University of Southern California, Los Angeles, CA 90089, United States

Received 27 June 2014; received in revised form 2 March 2015; accepted 15 March 2015

Abstract

How the speech production and perception systems evolved in humans still remains a mystery today. Previous research suggests that human auditory systems are able, and have possibly evolved, to preserve maximal information about the speaker's articulatory gestures. This paper attempts an initial step toward answering the *complementary* question of whether speakers' articulatory mechanisms have also evolved to produce sounds that can be optimally discriminated by the listener's auditory system. To this end we explicitly model, using computational methods, the extent to which derived representations of "primitive movements" of speech articulation can be used to discriminate between broad phone categories. We extract *interpretable* spatio-temporal primitive movements as recurring patterns in a data matrix of human speech articulation, i.e., representing the trajectories of vocal tract articulators over time. To this end, we propose a weakly-supervised learning method that attempts to find a part-based representation of the data in terms of recurring basis trajectory units (or primitives) and their corresponding activations over time. For each phone interval, we then derive a feature representation that captures the co-occurrences between the activations of the various bases over different time-lags. We show that this feature, derived entirely from activations of these primitive movements, is able to achieve a greater discrimination relative to using conventional features on an interval-based phone classification task. We discuss the implications of these findings in furthering our understanding of speech signal representations and the links between speech production and perception systems.
© 2015 Elsevier Ltd. All rights reserved.

Keywords: Speech communication; Movement primitives; Phone classification; Motor theory; Information transfer

1. Introduction

Consider an information-based model of speech communication where the aim is to optimally and robustly convey a piece of information from speaker to listener. Scientists are still unclear about how the speech communication system has evolved in humans to achieve this task. One possibility is that the human auditory system has evolved to perceive speech produced by talkers, while another is that speakers' articulatory mechanisms have evolved to produce sounds that can be perceived by the listener's auditory system. A more likely possibility is that these systems have evolved

☆ This paper has been recommended for acceptance by Roger K. Moore.

* Corresponding author. Tel.: +1 2137403477.
E-mail address: vikram.ramanarayanan@gmail.com (V. Ramanarayanan).

http://dx.doi.org/10.1016/j.csl.2015.03.004
0885-2308/© 2015 Elsevier Ltd. All rights reserved.

Please cite this article in press as: Ramanarayanan, V., et al., Directly data-derived articulatory gesture-like representations retain discriminatory information about phone categories. Comput. Speech Lang. (2015), http://dx.doi.org/10.1016/j.csl.2015.03.004

together, the development of each bootstrapped by the other. This is because speech articulation is not the only action that can be produced by the human vocal organs, and likewise speech is not the only class of sounds that can be perceived by the auditory system. The human speech production system can perform actions other than those required for producing speech sounds (swallowing, chewing, etc.), while the auditory system can perceive natural sounds in the 20–20000 Hz range, including those that have distinct spectro-temporal characteristics not found in human speech. If we assume that the speech production and perception systems co-evolved to jointly optimize their (information encoding/decoding) performance with respect to each other, among other criteria, then this supposition would posit two broad predictions. First, the auditory system in listeners must process speech so as to preserve maximal information about the "intended" speech gestures of the speaker. Second, speakers must encode information – linguistic and/or paralinguistic – into speech gestures (and thereby speech) in such a manner that it can be robustly extracted by listeners.

With respect to the first prediction, researchers have presented evidence suggesting that the objects of speech perception are the intended gestures of the speaker, which could be represented, for instance, as invariant motor commands for linguistically significant movements (Liberman and Mattingly, 1985; Fowler and Galantucci, 2005). Though there is still debate among researchers regarding its validity, this theory, dubbed the Motor Theory of Speech Perception, is one popular theory that explicitly attempts to link speech production and perception. Smith and Lewicki (2006) found that the filterbank model of the cochlea has high coding efficiency for conveying maximal information to the brain for a wide range of natural sounds and, in particular, speech. It was in fact mathematically shown by Silva and Narayanan (2009) that a cochlear-like filterbank provides the Bayes optimal phonetic classification. Further, the research of Ghosh et al. (2011) and Bertrand et al. (2008) has shown that processing speech signals using an auditory cochlea-like filterbank preserves maximal mutual information between articulatory gestures and the processed speech signals. In other words, auditory filterbank-like transformations might improve speech perception/recognition performance because they maximize the articulatory information that speakers transmit. Hence there is some evidence in the literature in favor of the hypothesis that the human auditory system has evolved to maximally and robustly perceive information regarding the talker's speech gestures.

Now the second prediction posits that speech gestures must encapsulate information such that listeners can optimally perceive it. In particular, listeners must be able to derive categorical information regarding the underlying learned phonological structures (such as phonemes or syllables) of the language being spoken. This information must be discriminative such that these discrete constructs or categories can be teased apart from the continuous acoustic signal by listeners. Hitherto there has been little empirical evidence for such a claim for two reasons. For one, it was not until recently that major developments have been made in speech articulation measurement (see Ramanarayanan et al., 2012, for a review of recent developments in this field), which have allowed researchers to better explore hypotheses such as the two predictions mentioned above. Furthermore, speech gestures are theoretically defined in terms of abstract constriction-producing dynamical systems, and it is not clear how to extract these from speech articulation data in a principled manner. However, we recently showed qualitatively and quantitatively that one can robustly extract gesture-like movement primitives from speech articulation data using knowledge-informed machine learning techniques (Ramanarayanan et al., 2013). If these primitive representations are truly gesture-like and our hypothesis is true, they should contain discriminative information regarding underlying linguistic structure, such as phone categories. This leads us to the central question of this paper, that relates specifically on the second of the two predictions we presented earlier: *do directly data-derived "activation functions" of gesture-like movement primitives contain information to robustly discriminate between different phone categories?*

In addition to supporting scientific understanding, answering such a research question is important for speech technology applications such as automatic speech recognition; finding efficient representations is a key building block for such efforts (for a more detailed discussion, please see Ramanarayanan et al., 2012). Some reasons for this include: (i) improved noise robustness (Rose et al., 1996), (ii) better performance on spontaneous speech which exhibits a greater degree of coarticulation due to factored representations (Deng et al., 1997; Farnetani, 1997; Mcdermott and Nakamura, 2006), (iii) better modeling of different sources of variability, e.g., vocal tract morphology (Lammert et al., 2011), (iv) provision of a complementary view of the information captured by acoustic features alone (Arora and Livescu, 2013), and (v) the significantly lower-dimensional space of articulatory-based feature representations (Browman and Goldstein, 1995; King et al., 2007). To motivate the final argument in particular from a linguistics standpoint, Articulatory Phonology (Browman and Goldstein, 1995) theorizes that the act of speaking is decomposable into units of vocal tract action called "gestures" that are essentially low dimensional in nature, and suggests that lexical items are assembled from these dynamic primitive units, i.e., constriction actions of the vocal organs. Furthermore, Atal

Please cite this article in press as: Ramanarayanan, V., et al., Directly data-derived articulatory gesture-like representations retain discriminatory information about phone categories. Comput. Speech Lang. (2015), http://dx.doi.org/10.1016/j.csl.2015.03.004

(1999) observed that the speech signal at the acoustic level has a much higher bit rate (e.g., 64 kbits/s assuming 8 kHz sampling rate and 8 bits/sample encoding) as compared to that of the underlying sound patterns that have an information rate of less than a 100 bits/s. The presence of this large redundancy in the speech signal means that we first need to extract a lower-dimensional representation of the signal that best captures the discriminative information required for a given task at hand. For example, in the case of a phone discrimination task, we would want to extract a representation that is able to capture the differences between various sounds in a language. Extracting such a representation from speech data is not straightforward. There is much work on acoustic-to-articulatory inversion, i.e., extracting low-dimensional articulatory information such as vocal tract (constriction) variables from the acoustics (see Atal et al., 1978; Toda et al., 2004; Ghosh and Narayanan, 2010; Uria et al., 2012). It is not clear whether these articulatory tract variables are an optimal representation of speech that captures phone-specific information; though this is most likely not the case, judging by sub-optimal *standalone* performance of these features on tasks such as speech recognition relative to state-of-the-art acoustic features (Frankel and King, 2001). However, if we are able to extract from speech discriminative information about articulatory gestures (see Browman and Goldstein, 1995), which we know are useful in distinguishing different sounds in a language, we might be better positioned to solve this problem. Mitra et al. (2012) present encouraging work on machine learning techniques to estimate articulatory gestures from speech, but to our knowledge, their model is not completely interpretable and it does not explicitly optimize for phone category information. With that additional motivation, we explore the question of how well low-dimensional "articulatory movement primitives" derived from data can discriminate between broad phone categories. That said, it is important to stress that the goal of this paper is *not* to improve the state of the art in speech classification/recognition, but to *enhance our understanding of the scientific link between speech production and perception* using computational means.

In earlier work (Ramanarayanan et al., 2013, 2011), we formulated and defined articulatory movement primitives (or exemplars) as a dictionary or template set of articulatory movement patterns in space and time. Weighted combinations of the elements of this dictionary can be used to represent the set of coordinated spatio-temporal movements of vocal tract articulators required for speech production. Although this is not a completely validated model for human speech production, we showed that the primitives-extraction method empirically captures articulatory gesture-like components and therefore compositional elements within speech. Moreover such a representation captures information regarding movement synergies, i.e., combinations that simplify the production of movements by reducing the degrees of freedom that need to be specified by the motor control system (Kelso, 2009).

In Fig. 1 we present a schematic overview of the paper. Input articulatory data is passed to a feature extraction module (features could be articulatory primitives, or traditional features like Mel-Frequency Cepstral Coefficients), following which the resulting features are passed to a classification module to understand how well each feature is able to discriminate between different phone classes. We describe the articulatory data used for experiments in Section 2. In Sections 3 and 4 we present the mathematical formalism used for primitive extraction and a brief quantitative evaluation of the extraction procedure, respectively. Next, in Section 5, we describe an *interval*-based phone classification setup for validation including appropriate feature preprocessing steps. Finally, we present our experimental observations along with a discussion of possible implications for future research in Sections 6 and 7.

We use the following mathematical notation to present the analysis described in this paper. Matrices are represented by bold uppercase letters (e.g., $\mathbf{X}$), vectors are represented using bold lowercase letters (e.g., $\mathbf{x}$), and scalars are represented without any bold case (either upper or lower case). We use the notation $\mathbf{X}^\dagger$ to denote the matrix transpose of $\mathbf{X}$. Further, if $\mathbf{x}$ is an N-dimensional vector, we use the notation $\mathbf{x} \in \mathbb{R}^N$ to denote that $\mathbf{x}$ takes values from the N-dimensional real-valued set. Similarly, $\mathbf{X} \in \mathbb{R}^{M \times N}$ denotes that $\mathbf{X}$ is a real-valued matrix of dimension $M \times N$. Finally, we use the notation $\mathbf{X} = [\mathbf{x}_1 | \mathbf{x}_2 | \ldots | \mathbf{x}_K]$ to denote that matrix $\mathbf{X}$ is formed by collecting the vectors $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_K$ together as its columns.

2. Data

We analyzed ElectroMagnetic Articulography (EMA) data from the Multichannel Articulatory (MOCHA) database (Wrench, 2000), which consists of data from two (British English) speakers – one male and one female. Acoustic and articulatory data were collected while each speaker read a set of 460 phonetically diverse TIMIT sentences. The articulatory channels include EMA sensors directly attached to the upper and lower lips, upper and lower incisors (jaw), tongue tip (5–10 mm from the tip), tongue blade (approximately 2–3 cm posterior to the tongue tip sensor), tongue dorsum (approximately 2–3 cm posterior to the tongue blade sensor) and soft palate (see Table 1). Each articulatory

Please cite this article in press as: Ramanarayanan, V., et al., Directly data-derived articulatory gesture-like representations retain discriminatory information about phone categories. Comput. Speech Lang. (2015), http://dx.doi.org/10.1016/j.csl.2015.03.004

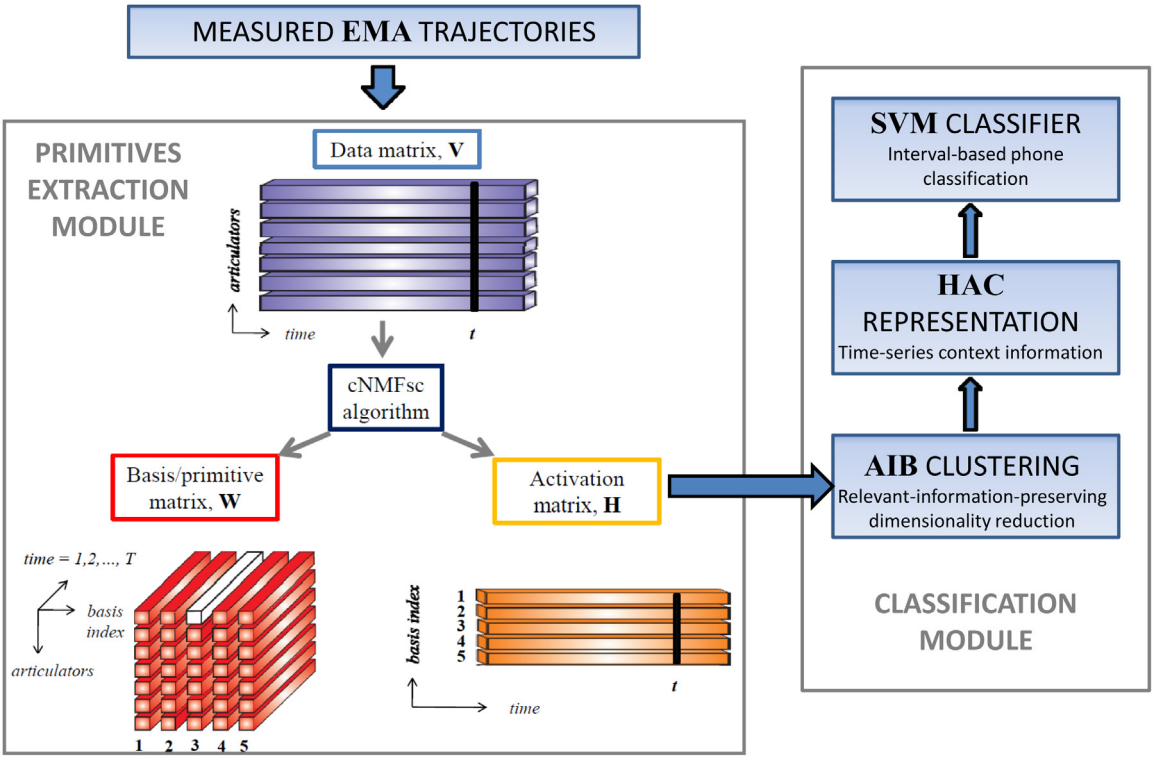

Fig. 1. Schematic of the experimental setup. The input matrix **V** is constructed from real (EMA) articulatory data. In this example, we assume that there are $M = 7$ articulator fleshpoint trajectories. We would like to find $K = 5$ basis functions or articulatory primitives, collectively depicted as the big red cuboid (representing a three-dimensional matrix **W**). Each vertical slab of the cuboid is one primitive (numbered 1–5). For instance, the white tube represents a single component of the 3rd primitive that corresponds to the first articulator (T samples long). The activation of each of these 5 time-varying primitives/basis functions is given by the rows of the activation matrix **H** in the bottom right hand corner. For instance, the 5 values in the tth column of **H** are the weights which multiply each of the 5 primitives at the tth time sample. The activation matrix is used as input to the classification module, which consists of 3 steps – (i) dimensionality reduction using agglomerative information bottleneck (AIB) clustering, (ii) conversion to a histogram of cooccurrence (HAC) representation to capture dependence information across timeseries, and (iii) a final support vector (SVM) classifier.

Table 1
Articulator flesh point variables that comprise the MOCHA-TIMIT electromagnetic articulograph dataset that we use for our experiments.

Symbol	Articulatory parameter
UL(x, y)	Upper lip
LL(x, y)	Lower lip
UI(x, y)	Upper incisor
LI(x, y)	Lower incisor (Jaw marker)
TT(x, y)	Tongue tip
TB(x, y)	Tongue body
TD(x, y)	Tongue dorsum
VEL(x, y)	Velum

channel was sampled at 500 Hz with 16-bit precision and zero-phase low-pass filtered with a cut-off frequency of 35 Hz (Ghosh and Narayanan, 2010). Next, for every utterance, we subtracted the mean value from each articulatory channel (Richmond, 2002; Ghosh and Narayanan, 2010). Then we added the mean value of each channel averaged over all utterances to that corresponding channel. Finally, we downsampled each channel by a factor of 5 to 100 Hz and further normalized data to the range [0,1].

After preprocessing the articulatory trajectories, we performed a phonetic alignment of the audio data corresponding to each set of articulator trajectories (using the Hidden Markov Model toolkit, HTK, Young et al., 2006). This was

Please cite this article in press as: Ramanarayanan, V., et al., Directly data-derived articulatory gesture-like representations retain discriminatory information about phone categories. Comput. Speech Lang. (2015), http://dx.doi.org/10.1016/j.csl.2015.03.004

done in order to facilitate learning of primitives based on a weak supervision step (explained in the following section) as well as for validation, to see how the algorithm performs for different phone classes.

3. Extraction of primitive movements

Modeling data vectors as sparse linear combinations of basis elements is a general computational approach[1] which we will use to solve our problem of modeling articulatory primitives (Lee and Seung, 2001; d'Avella and Bizzi, 2005; Smaragdis, 2007; O'Grady and Pearlmutter, 2008; Kim et al., 2010). Smaragdis (2007) presented a convolutive NMF algorithm to extract "phone"-like elements from speech spectrograms which could be used to characterize different speakers (or audio sources) for speech (or music) separation problems. O'Grady and Pearlmutter (2008) included the notion of sparsity in this formulation and showed that this gave more intuitive results. Note that we can view all these formulations as optimization problems with a cost function that involves (1) a data-fit term (which measures how accurately the model represents the data[2]) and (2) a regularization term (which enforces sparsity and/or smoothness constraints). If $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_M$ are the M time-traces (represented as column vectors of dimension $N \times 1$) of EMA articulator trajectory variables (note that the x- and y-positions of each articulator position are represented as two separate variables or dimensions), then we can design our data matrix $\mathbf{V}$ to be:

$$\mathbf{V} = [\mathbf{x}_1 | \mathbf{x}_2 | \ldots | \mathrm{x_M}]^{\dagger} \in \mathbb{R}^{M \times N} \tag{1}$$

where † is the matrix transpose operator. We will use convolutive nonnegative matrix factorization or cNMF (Smaragdis, 2007) to solve our problem. cNMF aims to find an approximation of the data matrix, $\hat{\mathrm{V}}$ using a basis tensor $\mathbf{W}$ and an activation matrix $\mathbf{H}$:

$$\mathbf{V} \approx \sum_{t=0}^{T-1} \mathbf{W}(t) \cdot \vec{\mathbf{H}}^{t} = \hat{\mathrm{V}} \tag{2}$$

where each column of $\mathbf{W}(t) \in \mathbb{R}^{\geq 0, M \times K}$ is a time-varying basis vector sequence, each row of $\mathbf{H} \in \mathbb{R}^{\geq 0, K \times N}$ is its corresponding activation vector ($\mathbf{h}_i$ is the ith row of $\mathbf{H}$), T is the temporal length of each basis (e.g., number of data samples or frames), and the $(\vec{\cdot})^{k}$ operator is a shift operator that moves the columns of its argument by k spots to the right, as detailed in (Smaragdis, 2007). Fig. 2 pictorially depicts how weighted and shifted additive combinations of the basis reconstruct the original input data sequence. In order to derive primitives that are maximally discriminative of different phone classes, we augmented the data matrix $\mathbf{V}$ with phone label information (after Van Segbroeck and Van hamme, 2009) to obtain:

$$\mathrm{V_{aug}} = \begin{bmatrix} \mathbf{V} \\ \mathrm{V_{lab}} \end{bmatrix} \approx \sum_{t=0}^{T-1} \begin{bmatrix} \mathbf{W}(t) \\ \mathrm{W_{lab}}(t) \end{bmatrix} \cdot \vec{\mathbf{H}}^{t} = \hat{\mathrm{V}}_{\mathrm{aug}} \tag{3}$$

where $\mathbf{V}_{aug}$ is the augmented data matrix, $\hat{\mathrm{V}}_{\mathrm{aug}}$ its model approximation, and each column of $\mathbf{V}_{lab}$ is a 40×1 vector whose entries are all 0 save for one – we set the entry corresponding to the phone label of the current frame[3] to 1 (there are 40 phone labels in all annotated for this dataset). To force the training algorithm to extract one unique primitive for each phone, we (i) added a (weak) supervision step to the multiplicative update rules of the cNMF training algorithm by forcing the $\mathbf{W}_{lab}$ matrix to be a 40×40 identity matrix, and (ii) set the number of primitives K equal to the number of unique phone classes (40). We further add a sparsity constraint on the rows of the activation matrix to obtain the final minimum-squared-error formulation of our optimization problem, termed cNMF with sparseness constraints (or cNMFsc) (Ramanarayanan et al., 2013, 2011):

$$\min_{\mathbf{W},\mathbf{H}} \left\| \begin{bmatrix} \mathbf{V} \\ \mathrm{V_{lab}} \end{bmatrix} - \sum_{t=0}^{T-1} \begin{bmatrix} \mathbf{W}(t) \\ \mathrm{W_{lab}}(t) \end{bmatrix} \cdot \vec{\mathbf{H}}^{t} \right\|_F^2 \quad \text{s.t.} \mathit{sparseness}(\mathrm{h_i}) = S_h, \forall i. \tag{4}$$

[1] This approach is termed variously as dictionary learning or sparse coding or sparse matrix factorization depending on the exact problem formulation.

[2] As measured by a suitable metric such as a norm distance (e.g., mean squared error) or a divergence metric (e.g., Kullback–Liebler divergence).

[3] Phone labels of each frame were obtained through automatic phonetic alignment of the audio data.

Please cite this article in press as: Ramanarayanan, V., et al., Directly data-derived articulatory gesture-like representations retain discriminatory information about phone categories. Comput. Speech Lang. (2015), http://dx.doi.org/10.1016/j.csl.2015.03.004

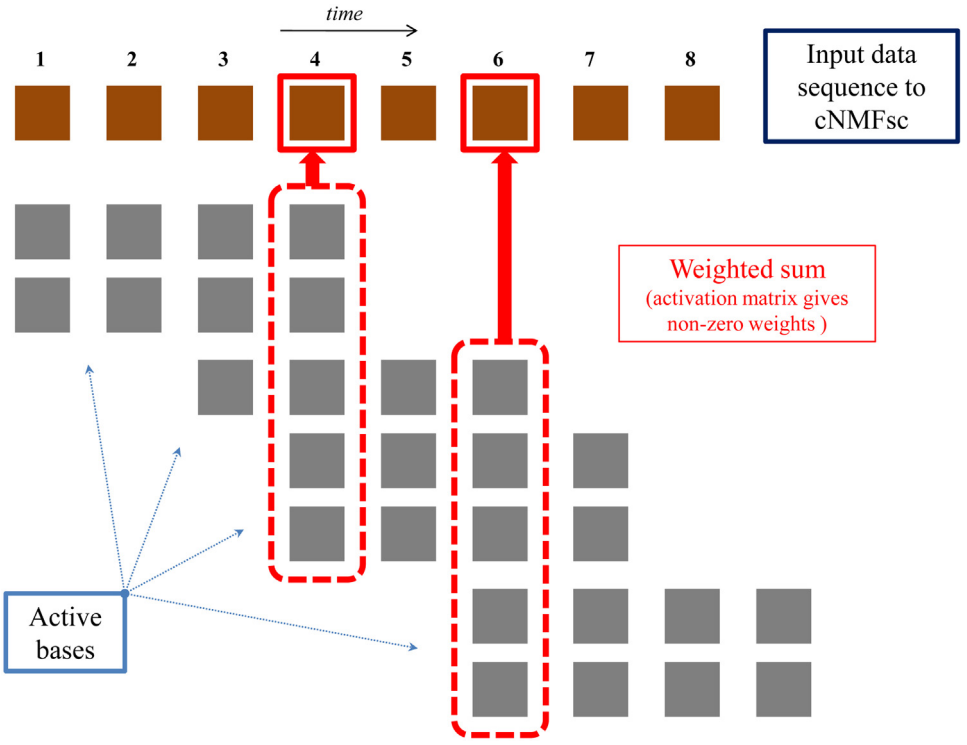

Fig. 2. Schematic illustrating how shifted and scaled primitives can additively reconstruct the original input data sequence. Each gold square in the topmost row represents one column vector of the input data matrix, **V**, corresponding to a single sampling instant in time. Recall that our basis functions/primitives are time-varying. Hence, at any given time instant t, we plot only the basis functions/primitives that have non-zero activation (i.e., the corresponding rows of the activation matrix at time t has non-zero entries). Notice that any given basis function extends $T = 4$ samples long in time, represented by a sequence of 4 silver/gray squares each. Thus, in order to reconstruct say the 4th column of **V**, we need to consider the contributions of all basis functions that are "active" starting anywhere between time instant 1 to 4, as shown.

where $\|\cdot\|_F$ denotes the matrix Frobenius norm. The sparseness metric is based on a relationship between the l_1 and l_2 norms (as proposed by Hoyer, 2004) as follows:

$$sparseness(\mathbf{x}) = \frac{\sqrt{n} - ((\sum_i |x_i|)/\sqrt{\sum_i x_i^2})}{\sqrt{n} - 1} \tag{5}$$

where n is the dimensionality of **x**. This function equals unity iff **x** contains only a single non-zero component and 0 iff all components are equal up to signs and smoothly interpolates between the extremes. Note that the level of sparseness ($0 \leq S_h \leq 1$) is user-defined. We imposed sparseness constraints on only the activation matrix as opposed to the basis matrix for two reasons: first, we want only a few bases to be "active" at any given sampling instant, which requires us to impose constraints on the activation matrix specifically; and second, not imposing any sparsity constaints on the basis matrix allows us to visualize and interpret the basis elements in the space in which the data was originally collected (sparsification of the basis elements will force the basis elements to be nonlinearly transformed). Please see Ramanarayanan et al. (2013) for the details of an algorithm that can be used to solve this problem. Briefly, the algorithm involves alternating between a gradient descent step that tries to move toward a lower objective function value and a projection step, that attempts to satisfy the sparseness constraints.

4. Algorithm performance

4.1. Quantitative metrics

In order to choose model parameters appropriately, we computed the Akaike Information Criterion (AIC; Akaike, 1981) on a subset of speaker *fsew0*'s data, which trades off the objective function value (or log-likelihood value, which

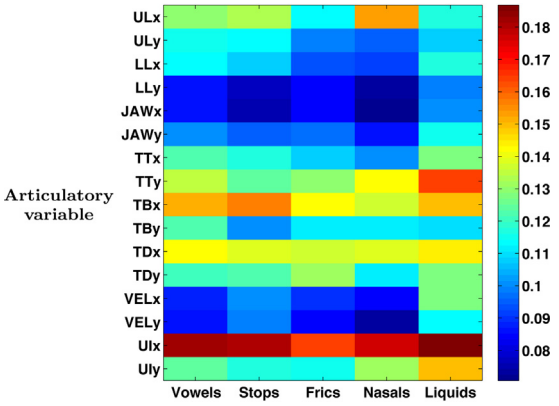

Fig. 3. Root mean squared error (RMSE) for each articulator and broad phone class obtained as a result of running the algorithm on all 460 sentences spoken by male speaker *msak0*.

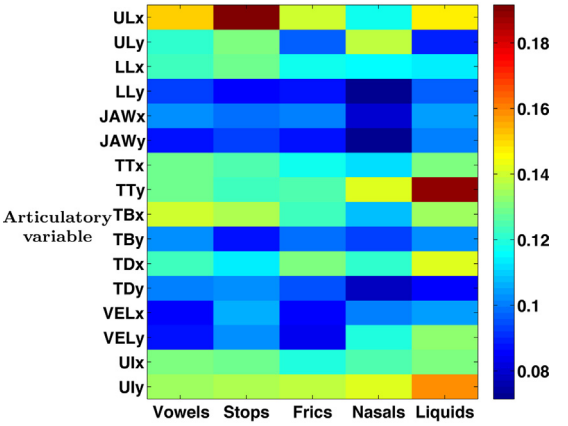

Fig. 4. Root mean squared error (RMSE) for each articulator and broad phone class obtained as a result of running the algorithm on all 460 sentences spoken by male speaker *fsew0*.

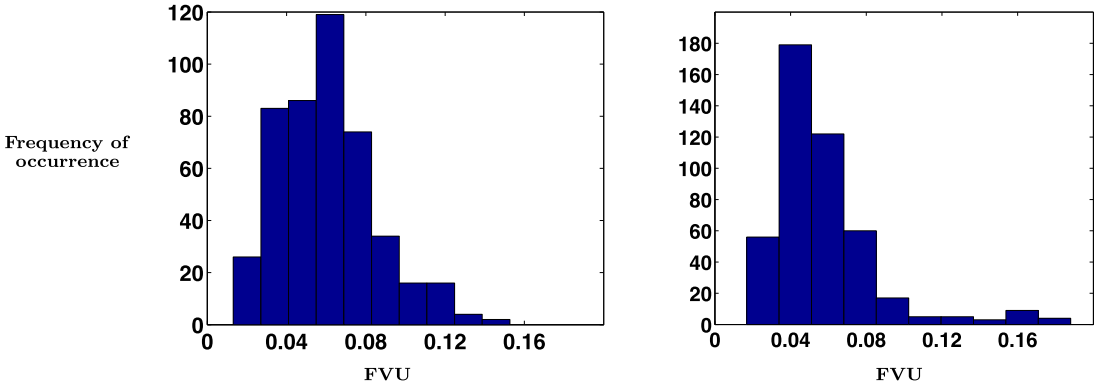

Fig. 5. Histograms of the fraction of variance unexplained (FVU) by the proposed cNMFsc model for MOCHA-TIMIT speakers *msak0* (left) and *fsew0* (right). The samples of the distribution were obtained by computing the FVU for each of the 460 sentences. (The algorithm parameters used in the model were $S_h = 0.65$, $K = 40$ and $T = 10$).

measures how well the model represents the data) against the model complexity (or number of parameters in the model). We found that this criterion overwhelmingly selected parameter values that resulted in low model complexity. Based on this analysis we set the temporal extent of each basis sequence or primitive (T) to 10 samples (since this corresponds to a time period of approximately 100 ms, factoring in a sampling rate of 100 samples per second) to capture effects of the order of the length of a phone on average. As mentioned earlier, we chose the number of bases, K, to be equal to the number of phone classes, i.e., 40, and chose a sparseness of $S_h = 0.65$ based on experiments with synthetic data.[4]

In order to see how the algorithm performs for each phone class, we leveraged the aforementioned phonetic alignment of the audio data to enable association with different phone classes. Figs. 3 and 4 show for MOCHA-TIMIT speakers *msak0* and *fsew0*, respectively, the root mean squared error (RMSE), computed as the Frobenius norm distance between the (*un*augmented) data matrix $\mathbf{V}$ and its cNMFsc-model approximation $\hat{\mathbf{V}}$ (see Eq. 2) for different articulators and broad phone classes (averaged over time and all phones belonging to a particular broad phone class). In other words, we first truncated the augmented data matrices $\mathbf{V}_{aug}$ and $\hat{\mathbf{V}}_{\text{aug}}$ (by removing the last 40 rows), and then computed the Frobenius norms of each submatrix formed by grouping together matrix entries belonging to each articulator and broad phone class (one submatrix each per articulator and broad phone class pair, for $\mathbf{V}_{aug}$ and $\hat{\mathbf{V}}_{\text{aug}}$, respectively). Figs. 3 and 4 display these norms in a matrix-like layout for ease of visualization. Since we have previously normalized

[4] We set the sparseness parameter to approximately equal the proportion of articulatory constriction tasks that were active at any given time instant in data generated by the Task Dynamics Application (TaDA) articulatory synthesizer. For further details, please see (Ramanarayanan et al., 2013).

each row of the original data matrix to the range [0, 1] (and hence each articulator trajectory), the error values in Figs. 3 and 4 can be read on a similar scale. We observe the errors are highest (0.13–0.2) for tongue-related articulator trajectories and the upper incisor variable. On the other hand, trajectories of the lip (*LLx* and *LLy*) and jaw (*JAWx* and *JAWy*) sensors are reconstructed with lower error ($\leq$0.1). We further computed the fraction of variance that was not explained (FVU) by the model for each sentence in the database as the ratio of the mean squared error between the model and the observed data to the variance of the data. The histograms of these distributions are plotted in Fig. 5. The mean and standard deviation of this distribution was 0.079 ± 0.028 for speaker *msak0* (i.e., approx. 7.9% of the original data variance was not accounted for on average). These statistics suggest that the cNMFsc model accounts for more than 90% of the original data variance.

4.2. *Visualization of extracted basis functions*

Figs. 6–8 show exemplar basis functions extracted from MOCHA-TIMIT data from speaker *msak0* corresponding to vowels, stops and fricatives respectively. We observe that the bases are interpretable and capture the salient articulatory movements required to produce each phone. For example, consider Fig. 6. The first column and last column of figure panels depict bases corresponding to *monophthong* vowels. They are arranged in a manner inspired from classical vowel quadrilateral visualizations of vowels (that plot the difference in the first two formants (F2–F1) against the first formant (F1)). Hence the vowel IY is plotted in the top left panel, while UW and AA are plotted in the top and bottom right panels, respectively. Notice that the bases capture the articulatory movements required to make the corresponding sounds – for example, we see the tongue body raising and lowering for IY, while the tongue root retracts into the pharynx for AA. *Diphthongs* are plotted in the middle column of panels. Again, the bases clearly capture the movement from the first vowel comprising the diphthong to the second – for example, in the case of AY, the tongue starts from a retracted position in the pharynx required for the low back vowel, advances, and finally rises toward the palate in order to achieve the high front vowel target.

In Fig. 7, bases for different stop and nasal consonants are arranged in order of manner and place of articulation – the first, second and third *columns* represent labial, coronal, and dorsal consonants, respectively. The first, second and third *rows* depict voiceless stops, voiced stops and nasals, respectively. The fourth and final row depicts approximants (not necessarily arranged in any particular order of place of articulation, however). The formation, achievement and release of constrictions is evident in the bases representing different coronal and dorsal consonants. For labial consonants, we observe evidence of lower lip raising and lowering, while the tongue behavior is more variable, as might be expected.

Fig. 8 employs a similar panel arrangement scheme as in Fig. 7, with one main difference. Each odd row depicts a set of *unvoiced* fricatives, while the even row immediately following depicts the corresponding set of *voiced* fricatives. Again, we see that the bases do an excellent job of capturing the basic articulatory movements required to produce different sounds in English.

5. *Interval-based* phone classification setup

In this section, we describe how activation matrices obtained using the algorithm described above are transformed into features suitable for phone classification experiments. We can hypothesize the sequence of phones corresponding to a given utterance along with their corresponding time-boundaries by phonetically aligning the audio. Therefore in this work, the phone categories are entirely based on categorical information obtained from the audio signal. At this point we would like to reiterate that the main goal of this paper is *not* to improve the state of the art in speech classification/recognition, but to verify whether the proposed "grounded" activation features indeed capture phone-discriminatory information, and thereby enhance our understanding of the scientific link between speech production and perception using computational means.

Since the activation matrices are sparse by formulation, it does not make sense to use columns of the activation matrix (one per frame) as feature prototypes in a frame-based phone classification experiment (since there will be zeros corresponding to time-frames where no basis is activated). Instead, we choose to compute *one* feature per phone interval. This way, we are formulating the classification problem as an *interval*-based phone classification experiment. Therefore, given a segment of activation columns for a given phone interval (i.e., a block subset of columns of the activation matrix), we have to compute a single feature. First, we quantize the space of activation vectors (columns of the activation matrix) to generate a codebook representation of the time-series using an agglomerative information

Please cite this article in press as: Ramanarayanan, V., et al., Directly data-derived articulatory gesture-like representations retain discriminatory information about phone categories. Comput. Speech Lang. (2015), http://dx.doi.org/10.1016/j.csl.2015.03.004

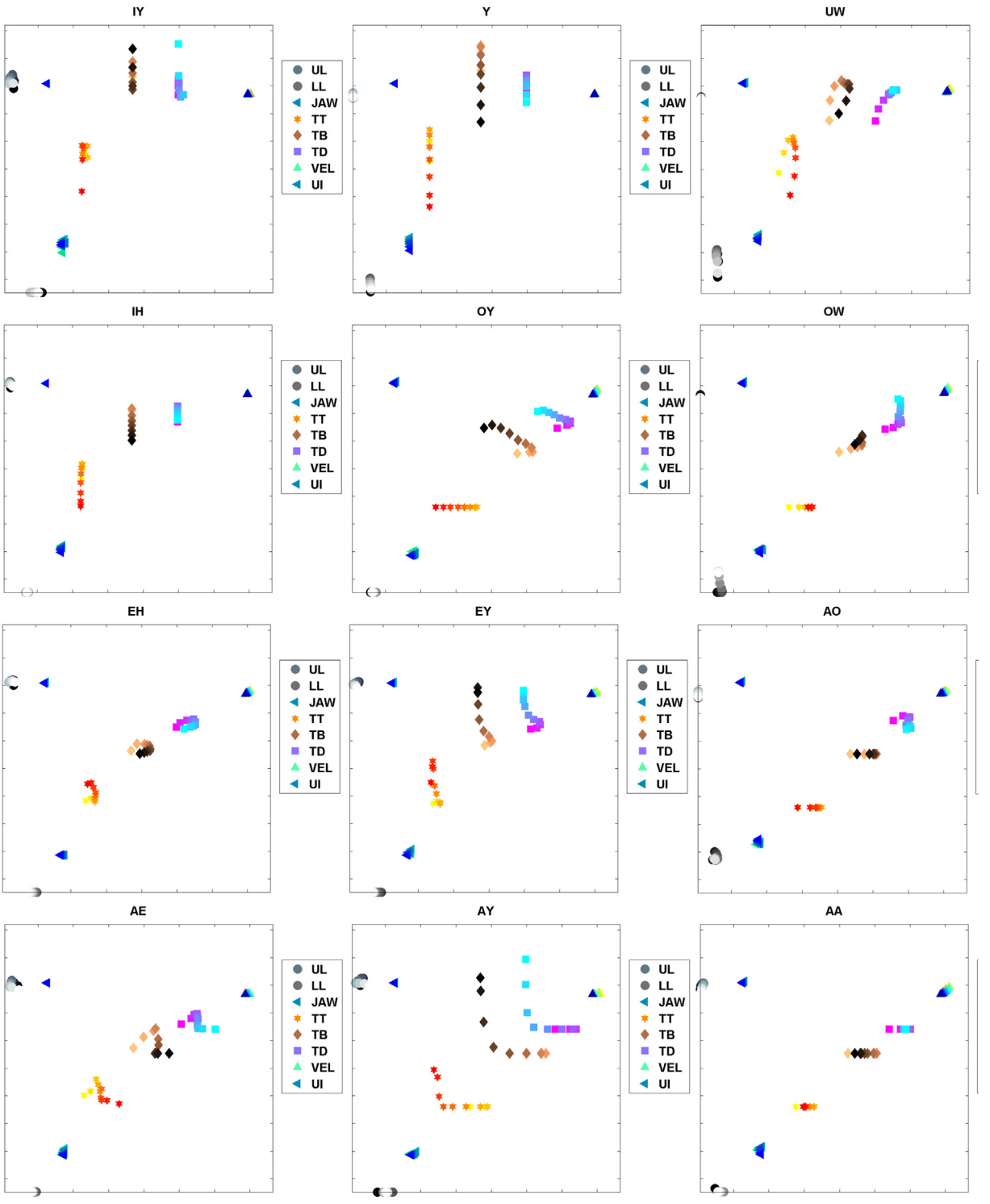

Fig. 6. Spatio-temporal basis functions or primitives extracted from MOCHA-TIMIT data from speaker *msak0* corresponding to different English monophthong (first and third columns) and diphthong (second column) vowels. Each panel is denoted by ARPABET phone symbol. The algorithm parameters used were $S_h = 0.65$, $K = 40$ and $T = 10$. The front of the mouth is located toward the left hand side of each image (and the back of the mouth on the right). Each articulator trajectory is represented as a curve traced out by 10 colored markers (one for each time step) starting from a lighter color and ending in a darker color. The marker used for each trajectory is shown in the legend (see Table 1 for the list of EMA trajectory variables).

Please cite this article in press as: Ramanarayanan, V., et al., Directly data-derived articulatory gesture-like representations retain discriminatory information about phone categories. Comput. Speech Lang. (2015), http://dx.doi.org/10.1016/j.csl.2015.03.004

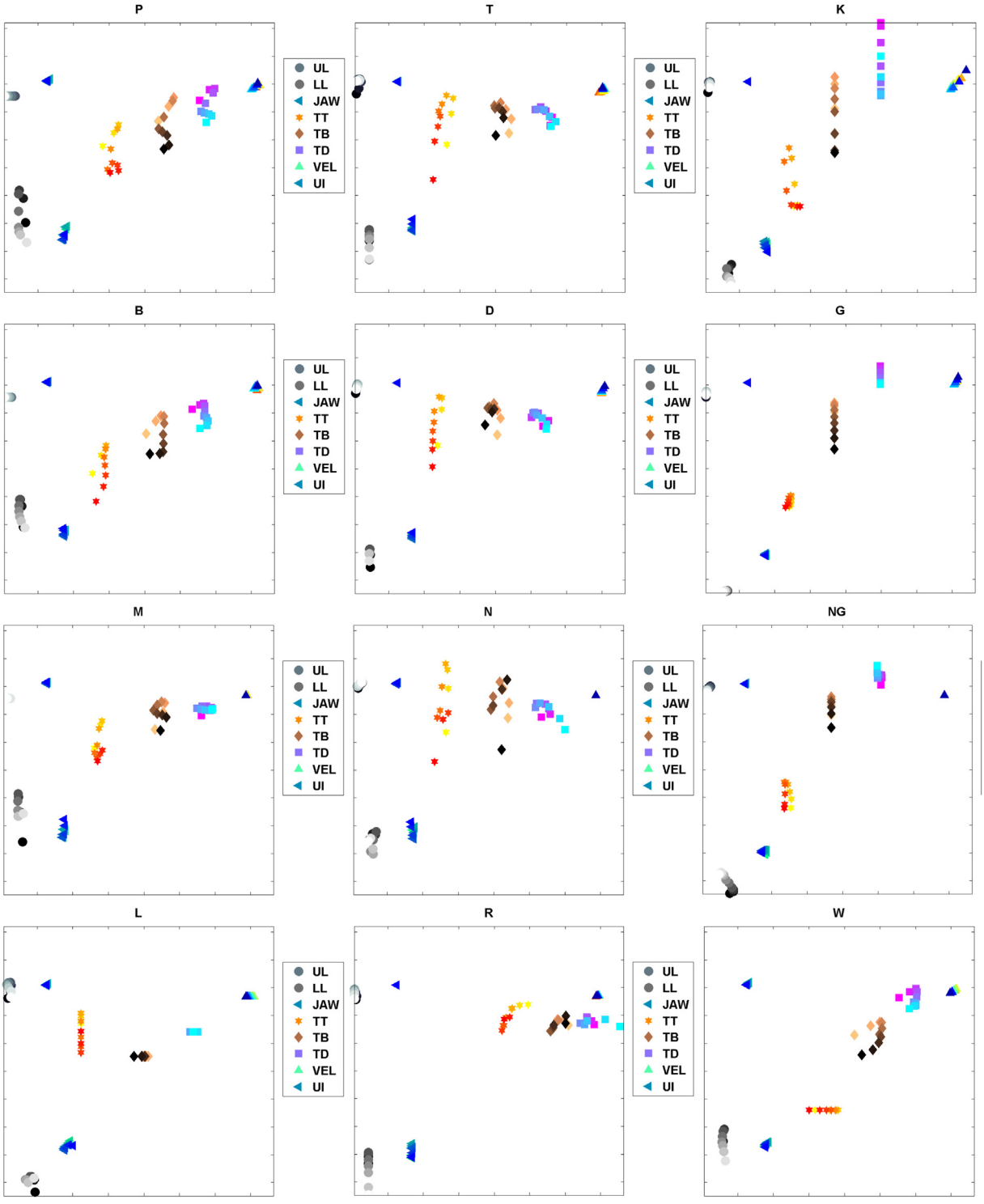

Fig. 7. Spatio-temporal basis functions or primitives extracted from MOCHA-TIMIT data from speaker *msak0* corresponding to stop (first two rows), nasal (third row) and approximant (last row) consonants. All rows except the last are arranged in order of labial, coronal and dorsal consonant, respectively. Each panel is denoted by ARPABET phone symbol. The algorithm parameters used were $S_h = 0.65$, $K = 40$ and $T = 10$. The front of the mouth is located toward the left hand side of each image (and the back of the mouth on the right). Each articulator trajectory is represented as a curve traced out by 10 colored markers (one for each time step) starting from a lighter color and ending in a darker color. The marker used for each trajectory is shown in the legend.

Please cite this article in press as: Ramanarayanan, V., et al., Directly data-derived articulatory gesture-like representations retain discriminatory information about phone categories. Comput. Speech Lang. (2015), http://dx.doi.org/10.1016/j.csl.2015.03.004

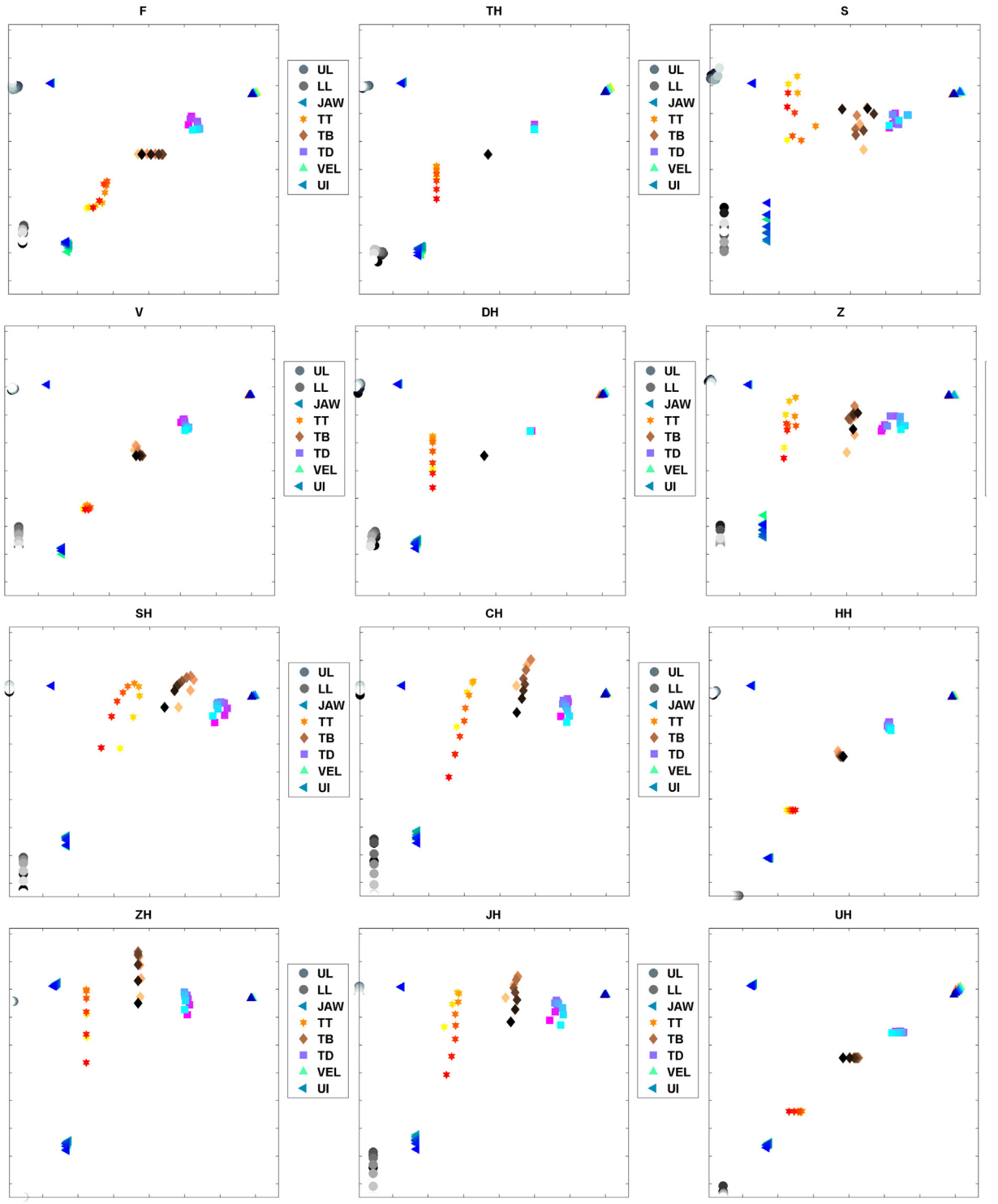

Fig. 8. Spatio-temporal basis functions or primitives extracted from MOCHA-TIMIT data from speaker *msak0* corresponding to fricatives and affricate consonants. Each panel is denoted by ARPABET phone symbol. The algorithm parameters used were $S_h = 0.65$, $K = 8$ and $T = 10$. The front of the mouth is located toward the left hand side of each image (and the back of the mouth on the right). Each articulator trajectory is represented as a curve traced out by 10 colored markers (one for each time step) starting from a lighter color and ending in a darker color. The marker used for each trajectory is shown in the legend.

Please cite this article in press as: Ramanarayanan, V., et al., Directly data-derived articulatory gesture-like representations retain discriminatory information about phone categories. Comput. Speech Lang. (2015), http://dx.doi.org/10.1016/j.csl.2015.03.004

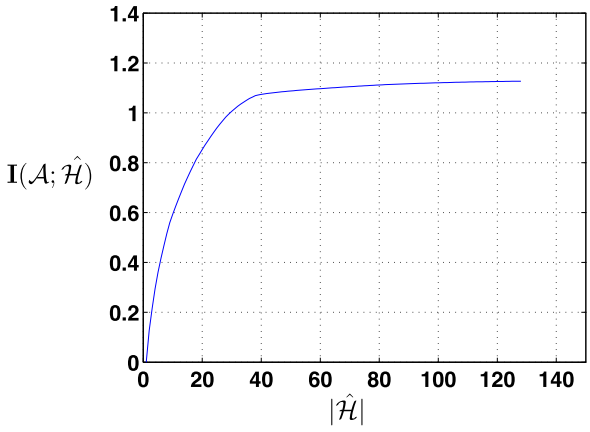

Fig. 9. Mutual information $\mathbf{I}(\mathcal{A};\hat{\mathcal{H}})$ between quantized activation space $\hat{\mathcal{H}}$ and the space of acoustic features $\mathcal{A}$ as a function of the cardinality of $\hat{\mathcal{H}}$ (in other words, the number of quantization levels).

bottleneck-based clustering technique; second, we compute histograms of co-occurrences (denoted HAC (Van hamme, 2008)) of the codebook indices over the time-series. Notice that the initial quantization step is needed because the column entries are not discrete-valued, making it impractical to compute meaningful co-occurrences directly. HAC representations in particular are useful since they explicitly model cooccurrences of articulatory feature instances over time. We describe the procedure in more detail below.

5.1. Codebook generation

We perform vector quantization (VQ) of the columns of the activation matrices in two steps: first, we clustered them using a k-means clustering algorithm with 64 clusters.[5] Second, we applied the agglomerative information bottleneck (AIB) principle (Slonim and Tishby, 1999) to formulate the problem as that of finding a quantization or a compressed representation $\hat{\mathcal{H}}$ of this pre-clustered activation matrix $\mathcal{H}$ that minimizes the mutual information $\mathbf{I}(\mathcal{H};\hat{\mathcal{H}})$ between them, while simultaneously maximizing the mutual information $\mathbf{I}(\mathcal{A};\hat{\mathcal{H}})$ between $\hat{\mathcal{H}}$ and a matrix of acoustically-derived phonetic labels $\mathcal{A}$. In other words, we would like to find that quantization of the (pre-clustered) activation space that achieves maximal compression while retaining as much discriminative information as possible about phonetic labels.[6] We use the VLFeat software (Vedaldi and Fulkerson, 2008) to perform this clustering. Fig. 9 plots the mutual information $\mathbf{I}(\mathcal{A};\hat{\mathcal{H}})$ as a function of the number of clusters/codebook entries. We observe a rapid drop in mutual information as the number of clusters drops below 30. Based on empirical observation of this graph, we choose a codebook size of 32 clusters for our experiments.

5.2. Computing histograms of articulatory co-occurrences (HAC)

We first replace each frame of the activation matrix $\mathbf{H}$ with the best matching centroid of the codebook. This way, the activation matrix is now represented by a single row vector of VQ-labels, $\mathbf{H}_{quant}$. A HAC-representation of *lag* τ is then defined as a vector where each entry corresponds to the number of times all pairs of VQ-labels are observed τ frames apart. In other words, we construct a vector of lag-τ co-occurrences where each entry (m, n) signifies the number of times that the input sequence of activation frames is encoded into a VQ-label m at time t (in the row vector

[5] We would ultimately like to compare the performance of primitive activation features to other standard speech features such as MFCCs or raw EMA features, which may be of a different dimensionality than the primitive activation features. In addition, agglomerative information bottleneck (AIB) clustering requires an estimate of probabilities of different activation values, which is difficult to estimate on a continuous dataset. Hence, prior to performing agglomerative information bottleneck-based clustering, we applied k-means clustering (with a larger number of clusters) to estimate these probabilities more easily while ensuring that we were making a fair comparison between the different feature-sets being compared, as far as possible. We empirically experimented with different k-means cluster-size values and found that 64 clusters performed optimally.

[6] Note that we aren't *adding* any extra information from acoustics to the activation features obtained from articulatory data. We are just clustering it differently using acoustically-derived phone label information.

Please cite this article in press as: Ramanarayanan, V., et al., Directly data-derived articulatory gesture-like representations retain discriminatory information about phone categories. Comput. Speech Lang. (2015), http://dx.doi.org/10.1016/j.csl.2015.03.004

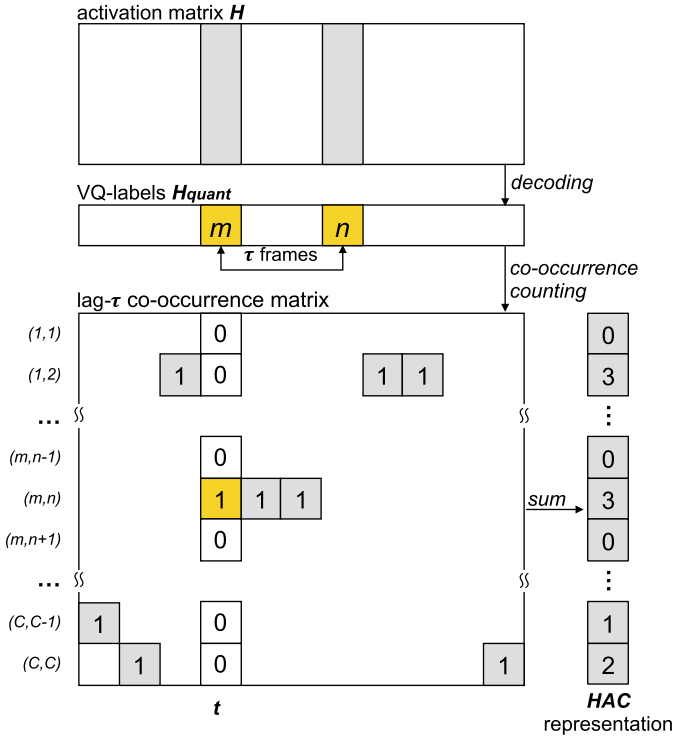

Fig. 10. Schematic depiction of the computation of histograms of articulatory cooccurrences (HAC) representations. For a chosen lag value, τ, and a time-step t, if we find labels m and n occurring τ time-steps apart (marked in gold), we mark the entry of the lag-τ cooccurrence matrix corresponding to row (m, n) and the tth column with a 1 (corresponding entry also marked in gold). We sum across the columns of this matrix (across time) to obtain the lag-τ HAC representation. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

H_{quant}), while encoded into VQ-label n at time $t+\tau$ (Van Segbroeck and Van hamme, 2009). By stacking all (m, n) combinations, each phone interval can be represented by a single column vector where the elements express the sum of all C^2 possible lag-τ co-occurrences (where C is the number of VQ clusters; in our case, 32). See Fig. 10. We can repeat the procedure for different values of τ, and stack the results into one "supervector". Note however, that the dimensionality of the HAC feature increases by a factor of C^2 for each lag value τ that we want to consider. In our case, we empirically found that choosing four lag values of 2, 3, 4 and 5 frames gave an optimal classification performance.

5.3. Classification experiments

We used support vector machine (SVM) classifiers to perform classification experiments (Chang and Lin, 2011) with 10-fold cross-validation. We experimented with both linear as well as radial basis function (RBF) kernels and empirically found that the former gave better classification accuracy. This could be due to the large dimensionality of the HAC feature space. Hyperparameters were tuned using a grid-search method.

6. Observations and results

To recap thus far, we have extracted a feature representation that captures phone-discriminative information (by virtue of the grounding procedure that we described earlier). The purpose of the following section is to verify that this claim is true, along with understanding why this might be. Note that it may not be entirely surprising to notice that our primitive features perform well on the classification task we will use for verification, since we have biased the feature representation by supplying class label information during training. However, what we aim to demonstrate is what *other* properties a feature might need to possess in order to provide optimal phone discriminability, and ultimately, to be an optimal representation of speech – namely (i) capturing information about the corresponding articulatory gestures

Please cite this article in press as: Ramanarayanan, V., et al., Directly data-derived articulatory gesture-like representations retain discriminatory information about phone categories. Comput. Speech Lang. (2015), http://dx.doi.org/10.1016/j.csl.2015.03.004

Table 2
Performance of various feature sets on a interval-based phone classification experiment (after appropriate transformation to HAC-representations). For clarity of understanding we also show the entropy of the feature set $\mathcal{X}$, denoted by $\mathbf{H}(\mathcal{X})$, along with the mutual information between the feature set and phone labels $\mathcal{L}$ (40 classes), denoted by $\mathbf{I}(\mathcal{X};\mathcal{L})$, in each case. We also performed classification experiments on a 5-class broad phoneset $\mathcal{L}_{broad}$ (where each of the phones were categorized as either vowel, stops, fricatives, nasals or approximants).

Feature set $\mathcal{X}$	Speaker	Classification accuracy		$\mathbf{H}(\mathcal{X})$	$\mathbf{I}(\mathcal{X};\mathcal{L})$	$\mathbf{I}(\mathcal{X};\mathcal{L}_{broad})$
		Full	Broad			
MFCC + Δ + ΔΔ	msak0	46.46%	71%	6.9	1.68	0.39
	fsew0	49.85%	65.63%	6.9	1.78	0.43
Raw EMA pellets	msak0	36.87%	61.78%	6.9	1.59	0.375
	fsew0	40.23%	57.20%	6.9	1.66	0.40
Primitive activations (EMA)	msak0	80.59%	85.01%	6.5	1.63	0.40
	fsew0	84.16%	87.77%	6.56	1.93	0.51
MFCC + Δ + ΔΔ +	msak0	47.58%	64.67%	6.93	1.87	0.40
Primitive activations (EMA)	fsew0	53.28%	67.46%	6.92	1.87	0.48
Phone labels $\mathcal{L}$		100%	–	4.9	4.9	–
Broad phone labels $\mathcal{L}_{broad}$		-	100%	1.96	-	1.96

and (ii) retaining maximal mutual information with the different phone classes of interest while maintaining a minimal bit rate.

Table 2 shows the performance of the activation features (after appropriate HAC-feature transformation) on an interval-based phone classification task. Also shown for reference purposes are the performances of the raw EMA pellets themselves (16-dimensional), as well as Mel-Frequency Cepstral Coefficient (MFCC) features along with their delta and double-delta coefficients (39-dimensional) on the same task. Each of these feature sets, i.e., primitive activation features, MFCCs and raw EMA features, were each passed through the same classification module (box to the right in Fig. 1), in order to ensure as fair point of reference as possible. Thus the only difference between each of these classification setups lies in the way the features themselves were extracted. We report results on both a full-blown 40-class classification task as well as a simpler broad-phone task consisting of 5 classes (namely vowels, stops, fricatives and affricates, nasals and approximants). As might be expected, the performance for all features on the latter task is significantly better than that on the full-blown classification task. Our experiments suggest that the activation features learnt by the cNMFsc algorithm significantly outperform both raw MFCC and raw EMA features in terms of classification accuracy, which as mentioned earlier, is not entirely surprising. However, another cause for this may be that the classification process (AIB clustering, followed by HAC extraction) might not be as well suited to MFCC and EMA features as they are to the sparse activation features. Further note that results on *interval*-based classification task described here are different from the state-of-the-art performance numbers reported on traditional *frame*-based classification tasks, where MFCCs perform competently (for further details on these numbers, please see for e.g., Lopes and Perdig ao, 2011; Karsmakers et al., 2007; Gunawardana et al., 2005). Finally, we also observe that a concatenated feature set of both primitives (computed from EMA data) *and* MFCC features returned lower numbers on the SVM classification task as compared to standalone primitive features, but performed marginally better than standalone MFCC features. The first observation suggests that: (i) the higher dimensionality of the augmented MFCC + primitives feature space could make the classification task harder relative to the standalone primitives case; (ii) this particular feature fusion might have a destructive as opposed to synergistic effect on the linear separability of different classes in the augmented feature space. However, note that both the entropy of the augmented features as well as the mutual information between the augmented features and the phone labels is generally higher than the case of standalone MFCC features. That we observe the augmented feature performing better than standalone MFCCs is in agreement with previous literature suggesting a benefit of including complementary articulatory information to MFCCs during classification tasks.

We would now like to turn to the more expository information-theoretic numbers in Table 2. For a deeper understanding of what the classification accuracy numbers in Table 2 actually mean, we computed the entropy of each feature

set and mutual information[7] (MI) between each feature set and the phone labels. The last two lines of Table 2 show the estimated entropy of the phone labels $\mathbf{H}(\mathcal{L})$ (which is also equal to the mutual information with itself, $\mathbf{I}(\mathcal{L};\mathcal{L})$), and provide a context to interpret the rest of the information-theoretic numbers in the table. The entropy of the full phone label set was estimated to be around 4.9 bits, which gives us an idea of a *lower* bound on the entropy and an *upper* bound on the mutual information of a given feature vector. What we would like ideally is a feature that achieves the upper bound on mutual information (4.9 bits) while possessing an entropy that is as close to 4.9 bits as possible. Keeping this in mind, we observe that although the entropy (and consequently bit rate assuming a fixed encoding scheme) of primitive activation features is lower than that of the raw MFCC or EMA features, the mutual information between the phone labels and the different features considered is still comparable. This, along with the weak supervision during the learning process, suggests that primitive activations are a useful, low-dimensional representation capable of discriminating phone classes. In addition, we can see that although the MFCC and EMA features have a similar entropy value, the former has a higher MI. This is in agreement with the observation of a higher classification accuracy. The continuing challenge for future work will be finding representations that push the classification accuracy envelope while minimizing the required bitrate.

7. Discussion and outlook

What makes a speech feature representation optimal for recognition by humans or machines in general? Although this question is still a hot topic of debate among experts in the field, and we are still far from a concrete answer to this question, researchers have long agreed upon phone-discriminability as one criterion, and methods such as Linear/Generalized Discriminant Analysis (LDA/GDA; Duda et al., 2012) and related techniques have been extensively explored to get at an optimal feature representation. Our findings reinforce prior work in the literature (e.g., Liberman and Mattingly, 1985; Browman and Goldstein, 1995; Rose et al., 1996; Atal, 1999; Smith and Lewicki, 2006; Ghosh et al., 2011, among others) suggesting that two other properties might be important in our search toward such a feature: first, they should contain information regarding the underlying articulatory gestures used to produce the speech; second, they must possess as low a bit-rate as possible while retaining maximal mutual information with relevant categories/classes of interest.

We find that articulatory movement primitive representations in particular are useful features in this regard. It is important to note that the performance of these features is contingent upon the way they are extracted, and therefore, algorithmic choices, such as the sparseness value S_h, number of primitives, and the temporal extent of each primitive, will greatly influence the outcome of subsequent classification experiments. The primitives we extract will depend on the cost function that we formulate and optimize. Development of better problem formulations and algorithms to extract primitives is an exciting area of ongoing and future research. It is, however, encouraging to observe that the computationally estimated lower-dimensional primitive representations of speech articulation contain useful information to distinguish between broad phone categories. Note also that the (EMA) articulatory data used in the experiments offers only a limited view of the complex articulatory mechanisms. Although they do encode information about phonetic categories, these movements represent only a part of the picture with respect to the phonetic categories.

The results presented here allow us to re-examine the problem of optimal feature extraction specifically for automatic phone classification/recognition in a new light. In other words, good features for phone classification and recognition would be those that retain maximal mutual information with the different phone classes of interest as well as the speech articulation process. We have shown that appropriately postprocessed activation features of articulatory primitives exhibiting this property perform well on classification experiments while retaining linguistic interpretability. In other words, the answer to the central question posed in the paper, "does directly data-derived "activation functions" of gesture-like movement primitives contain information to robustly discriminate between different phone categories?", does indeed seem to be yes.

The challenge for future work is to develop procedures to extract features that possess the properties described above from unlabeled test data where the phone categories are unknown, and further, to extend the ideas presented

[7] To estimate the probability of a given feature vector: (i) we clustered the data (as mentioned previously in Section 5.1) using k-means clustering ($K = 64$) and assigned each feature to a cluster. (ii) We set the probability of occurrence of a feature to be equal to the maximum likelihood estimate of the probability of occurrence of its corresponding cluster.

Please cite this article in press as: Ramanarayanan, V., et al., Directly data-derived articulatory gesture-like representations retain discriminatory information about phone categories. Comput. Speech Lang. (2015), http://dx.doi.org/10.1016/j.csl.2015.03.004

here from phone classification to full-blown phone recognition. Moreover, we would also like to extract and visualize primitives from MFCC or other acoustic-based features to examine the similarities and differences between acoustic and articulatory-based primitive representations and how complementary they are for phone classification/recognition. Also, since we have presented results for only two speakers (as the MOCHA-TIMIT public release is limited to data from two speakers), future work will also look at examining the robustness of these results with articulatory and acoustic data from more speakers. Furthermore, extending these results to a speaker-independent setting (as opposed to the speaker-specific setting described here) would be a useful and important research direction.

We can now return to our predictions regarding the joint optimization and co-evolution of the speech production and perception systems in light of our results. The first prediction was that the auditory system in listeners must process speech so as to preserve maximal information regarding the "intended" speech gestures of the speaker. This position is supported by both theoretical work (Liberman and Mattingly, 1985; Fowler and Galantucci, 2005) as well as recent empirical results (Ghosh et al., 2011; Bertrand et al., 2008). The second prediction was that speakers must encode information (linguistic or paralinguistic) into speech gestures (and thereby speech) in such a manner that it can be robustly extracted by listeners. This paper has shown empirical results supporting this position. The Motor Theory of speech perception (Liberman and Mattingly, 1985) also predicts that the human speech production system must produce just the right maneuvers to fit the demands of the categories imposed by the auditory system. Assuming that articulatory movement primitives can be considered as a surrogate for at least a subset of these maneuvers, our results are in agreement with the theory. This is because we find that experimentally-derived articulatory primitives are not only good surrogate representations of articulatory gestures (as observed in Figs. 6–8), but also contain discriminatory information regarding different phone categories (as seen in Table 2). This work presents a first step toward a more complete understanding of whether information transfer during speech production is performed so as to effect efficient perception of auditory categories.

Acknowledgements

We gratefully acknowledge the support of NIH Grant R01 DC007124-01. We would also like to thank Louis Goldstein for useful discussions and Prasanta Ghosh for help with the EMA data processing.

References

Akaike, H., 1981. Likelihood of a model and information criteria. Journal of Econometrics 16 (1), 3–14.
Arora, R., Livescu, K., 2013. Multi-view CCA-based acoustic features for phonetic recognition across speakers and domains. In: Int. Conf. on Acoustics, Speech, and Signal Processing.
Atal, B., 1999. Automatic speech recognition: a communication perspective. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing – Proceedings, vol. 1, pp. 457–460.
Atal, B.S., Chang, J., Mathews, M.V., Tukey, J.W., 1978. Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. J. Acoust. Soc. Am. 63 (5), 1535–1555.
Bertrand, A., Demuynck, K., Stouten, V., Van hamme, H., 2008. Unsupervised learning of auditory filter banks using non-negative matrix factorisation. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4713–4716.
Browman, C., Goldstein, L., 1995. Dynamics and articulatory phonology. In: van Gelder, T., Port, B. (Eds.), Mind as Motion: Explorations in the Dynamics of Cognition. , pp. 175–193.
Chang, C., Lin, C., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2 (3), 27.
d'Avella, A., Bizzi, E., 2005. Shared and specific muscle synergies in natural motor behaviors. Proc. Natl. Acad. Sci. U. S. A. 102 (8), 3076.
Deng, L., Ramsay, G., Sun, D., 1997. Production models as a structural basis for automatic speech recognition. Speech Commun. 22 (2), 93–111.
Duda, R.O., Hart, P.E., Stork, D.G., 2012. Pattern Classification. John Wiley & Sons, New York, NY.
Farnetani, E., 1997. Coarticulation and connected speech processes. In: The Handbook of Phonetic Sciences., pp. 371–404.
Fowler, C.A., Galantucci, B., 2005. The relation of speech perception and speech production. In: The Handbook of Speech Perception., pp. 632–652.
Frankel, J., King, S., 2001. ASR – articulatory speech recognition. In: Seventh European Conference on Speech Communication and Technology.
Ghosh, P., Goldstein, L., Narayanan, S., 2011. Processing speech signal using auditory-like filterbank provides least uncertainty about articulatory gestures. J. Acoust. Soc. Am. 129, 4014.
Ghosh, P., Narayanan, S., 2010. A generalized smoothness criterion for acoustic-to-articulatory inversion. J. Acoust. Soc. Am. 128, 2162–2172.
Gunawardana, A., Mahajan, M., Acero, A., Platt, J.C., 2005. Hidden conditional random fields for phone classification. In: INTERSPEECH, pp. 1117–1120.
Hoyer, P., 2004. Non-negative matrix factorization with sparseness constraints. J. Mach. Learn. Res. 5, 1457–1469.
Karsmakers, P., Pelckmans, K., Suykens, J.A., Hamme, H.V., 2007. Fixed-size kernel logistic regression for phoneme classification. In: INTERSPEECH, pp. 78–81.

Kelso, J., 2009. Synergies: atoms of brain and behavior. Prog. Motor Control, 83–91.
Kim, T., Shakhnarovich, G., Urtasun, R., 2010. Sparse coding for learning interpretable spatio-temporal primitives. Adv. Neural Inform. Process. Syst., 22.
King, S., Frankel, J., Livescu, K., McDermott, E., Richmond, K., Wester, M., 2007. Speech production knowledge in automatic speech recognition. J. Acoust. Soc. Am. 121, 723–742.
Lammert, A., Proctor, M., Narayanan, S., 2011. Morphological variation in the adult vocal tract: a study using RTMRI. In: Proc. 9th ISSP.
Lee, D., Seung, H., 2001. Algorithms for non-negative matrix factorization. Adv. Neural Inform. Process. Syst., 13.
Liberman, A., Mattingly, I., 1985. The motor theory of speech perception revised. Cognition 21 (1), 1–36.
Lopes, C., Perdig ao, F., 2011. Phone recognition on the TIMIT database. Speech Technol. 1, 285–302.
Mcdermott, E., Nakamura, A., 2006. Production-oriented models for speech recognition. IEICE Trans. Inform. Syst. 89 (3), 1006–1014.
Mitra, V., Nam, H., Espy-Wilson, C., Saltzman, E., Goldstein, L., 2012. Recognizing articulatory gestures from speech for robust speech recognition. J. Acoust. Soc. Am. 131 (3), 2270–2287.
O'Grady, P., Pearlmutter, B., 2008. Discovering speech phones using convolutive non-negative matrix factorisation with a sparseness constraint. Neurocomputing 72 (1–3), 88–101.
Ramanarayanan, V., Ghosh, P., Lammert, A., Narayanan, S., 2012. Exploiting speech production information for automatic speech and speaker modeling and recognition – possibilities and new opportunities. In: Fourth Annual Conference of the Asia-Pacific Signal and Information Processing Association.
Ramanarayanan, V., Goldstein, L., Narayanan, S.S., 2013. Spatio-temporal articulatory movement primitives during speech production: Extraction, interpretation, and validation. J. Acoust. Soc. Am. 134 (2), 1378–1394.
Ramanarayanan, V., Katsamanis, A., Narayanan, S., 2011. Automatic data-driven learning of articulatory primitives from real-time MRI data using convolutive NMF with sparseness constraints. In: Twelfth Annual Conference of the International Speech Communication Association.
Richmond, K., Ph.D. thesis 2002. Estimating Articulatory Parameters from the Acoustic Speech Signal. University of Edinburgh.
Rose, R., Schroeter, J., Sondhi, M., 1996. The potential role of speech production models in automatic speech recognition. J. Acoust. Soc. Am. 99, 1699–1709.
Silva, J., Narayanan, S.S., 2009. Discriminative wavelet packet filter bank selection for pattern recognition. IEEE Trans. Signal Process. 57 (5), 1796–1810.
Slonim, N., Tishby, N., 1999. Agglomerative information bottleneck. Adv. Neural Inform. Process. Syst. 12, 617–623.
Smaragdis, P., 2007. Convolutive speech bases and their application to supervised speech separation. IEEE Trans. Audio Speech Lang. Process. 15 (1), 1–12.
Smith, E.C., Lewicki, M.S., 2006. Efficient auditory coding. Nature 439 (7079), 978–982.
Toda, T., Black, A.W., Tokuda, K., 2004. Acoustic-to-articulatory inversion mapping with Gaussian mixture model. In: INTERSPEECH.
Uria, B., Murray, I., Renals, S., Richmond, K., 2012. Deep architectures for articulatory inversion. In: INTERSPEECH.
Van hamme, H., 2008. HAC-models: a novel approach to continuous speech recognition. In: INTERSPEECH.
Van Segbroeck, M., Van hamme, H., 2009. Unsupervised learning of time-frequency patches as a noise-robust representation of speech. Speech Commun. 51 (11), 1124–1138.
Vedaldi, A., Fulkerson, B., 2008. VLFeat: An Open and Portable Library of Computer Vision Algorithms. http://www.vlfeat.org/
Wrench, A., 2000. A multi-channel/multi-speaker articulatory database for continuous speech recognition research. In: Workshop on Phonetics and Phonology in ASR.
Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., et al., 2006. The HTK Book (for HTK Version 3.4).

Please cite this article in press as: Ramanarayanan, V., et al., Directly data-derived articulatory gesture-like representations retain discriminatory information about phone categories. Comput. Speech Lang. (2015), http://dx.doi.org/10.1016/j.csl.2015.03.004