Information theoretic optimal vocal tract region selection from real time magnetic resonance images for broad phonetic class recognition∗

Abhay Prasad, Prasanta Kumar Ghosh *

Department of Electrical Engineering, Indian Institute of Science, Bangalore, Karnataka 560012, India

Received 24 December 2014; received in revised form 26 March 2016; accepted 28 March 2016
Available online 6 April 2016

Abstract

We propose an information theoretic region selection algorithm from the real time magnetic resonance imaging (rtMRI) video frames for a broad phonetic class recognition task. Representations derived from these optimal regions are used as the articulatory features for recognition. A set of connected and arbitrary shaped regions are selected such that the articulatory features computed from such regions provide maximal information about the broad phonetic classes. We also propose a tree-structured greedy region splitting algorithm to further segment these regions so that articulatory features from these split regions enhance the information about the phonetic classes. We find that some of the proposed articulatory features correlate well with the articulatory gestures from the Articulatory Phonology theory of speech production. Broad phonetic class recognition experiment using four rtMRI subjects reveals that the recognition accuracy with optimal split regions is, on average, higher than that using only acoustic features. Combining acoustic and articulatory features further reduces the error-rate by ~8.25% (relative).

© 2016 Elsevier Ltd. All rights reserved.

Keywords: Mutual information; Phonetic recognition; Speech production; Region splitting

1. Introduction

The speech signal is encoded with both linguistic and para-linguistic information including speaker’s characteristics, background noise condition, and emotional state of the speaker. Seeking representations from the speech signal has been one of the main challenges in speech research (Greenberg and Kingsbury, 1997; Paliwal, 1998). The nature of a representation could change depending on the type of task at hand. For example, a representation for speech recognition is expected to be invariant to recording environment, and speaker’s characteristics (Greenberg and Kingsbury, 1997). On the other hand, for speaker recognition, a representation should mostly capture speaker specific information (Perez-Meana, 2007). Representations robust to noise and other channel distortions are also critical for reliable performance of a speech based system in different working conditions.

∗ This paper has been recommended for acceptance by Shrikanth Narayanan.

* Corresponding author. Tel.: +91 80 2293 2694; fax: +91 80 2360 0444.
E-mail address: prasantg@ee.iisc.ernet.in (P.K. Ghosh).

http://dx.doi.org/10.1016/j.csl.2016.03.003
0885-2308/© 2016 Elsevier Ltd. All rights reserved.
Speech is a time-varying signal with complex spectro-temporal characteristics. The temporal and spectral resolutions necessary to analyze different speech sounds often vary over a short period of time. Thus, most of the representations in the literature have been designed assuming quasi-stationarity of the speech signal within a short-time window. A number of representative features are available exploiting spectral-only, temporal-only, and joint spectro-temporal properties. Spectral features include Filter Bank Analysis (Potter et al., 1947; Nadeu et al., 2001), Mel Cepstrum Analysis (Davis and Mermelstein, 1980; Olli and Kari, 1998; Han et al., 2006), Linear Predictive Coding (LPC) (Itakura, 1975), Perceptually Based Linear Predictive Analysis (PLP) (Hermansky, 1990), and spectral subband centroids (Paliwal, 1998). Temporal features such as temporal envelope cues (Shannon et al., 1995), short-time energy and fundamental frequency are used as features for speech recognition. Benefits of joint spectro-temporal features for speech recognition have also been demonstrated (Kleinschmidt, 2003; Xavier et al., 2008).

In spite of several proposed speech signal-based features, obtaining a robust representation still remains a challenge. This has led researchers to look into obtaining features from other modalities that capture the speech production process directly. For example, Krishnamurthy and Childers (1986) have demonstrated the benefit of using representation from electroglottography (EGG) for classification of speech according to different voicing qualities. While EGG captures the glottal vibrations of the speaker, other modalities such as X-ray microbeam (Fujimura et al., 1973) and electromagnetic articulography (EMA) (Schönle et al., 1987) capture the movement of critical points on speech articulators in the vocal tract by placing sensors on the articulators. Representations derived from these sensor data in addition to the speech acoustic have been shown to be useful for speech recognition (Sun et al., 2000; Wrench and Richmond, 2000; Frankel and King, 2001; Ghosh and Narayan, 2011; Markov et al., 2006). Unlike sensor tracking in EMA, ultrasound captures a complete spatial view of the tongue (Hueber et al., 2007b; Stone and Davis, 1995); representations derived from the ultrasound images have been shown to be effective for the continuous-speech recognition as well as speech synthesis (Hueber et al., 2007a; Denby and Stone, 2004). Features from electromyography (EMG) as well as electropalatography (EPG) have also been shown to be effective for speech recognition (Jou et al., 2006; Manabe et al., 2003; Jørgensen et al., 2003; Schultz and Wand, 2010; Soquet et al., 1999); these signals capture the muscle movements in the face while speaking. Representations from EMA and EPG data together have been shown to improve the recognition accuracy when used jointly with spectral features (Wrench and Richmond, 2000).

Unlike ultrasound, in real time magnetic resonance imaging (rtMRI) the complete upper airway of a subject is imaged including the subject’s nose and upper palate in addition to the vocal tract region starting from the lips to the glottis (Narayan et al., 2011). The rtMRI video frames also have regions that are outside the subject’s face in the midsagittal plane. Thus, the rtMRI video directly captures time varying dynamics of the changes in the vocal tract shape. Since the vocal tract shapes cause the production of different sounds, a representation from such vocal tract images would be robust to noise present in the speech signal. Representation from the rtMRI images could also be complementary to the spectro-temporal features derived from the speech signal since the relation between the vocal tract dynamics and speech acoustics is highly non-linear (Deng, 2006).

1.1. Research question

rtMRI captures the air-tissue boundaries along the entire vocal tract region from the glottis to the lips. A commonly used protocol for recording rtMRI video captures MR images at a frame rate of 23.18 frames/s with a resolution of 68 × 68 pixels with simultaneous audio recorded at 20 kHz. Hence the rtMRI has ~106,352 dimensional data captured per second, which is more than that captured in audio (sampling rate of 20 kHz). Thus, the entire image may not be an efficient representation of the corresponding sound due to its large dimensionality. The key information related to the sound may lie only in a few pixels or regions in the image. In this work, we address the problem of automatically obtaining optimal regions from the rtMRI images such that the articulatory features derived from these optimal regions best represent the respective broad phonetic classes. Features derived from the optimal regions in the rtMRI images could potentially remove the redundancy in representing various sounds.

1.2. Related works

Several approaches have been used to obtain optimal features from rtMRI. The most commonly used approach is to compute the mean pixel intensity in a defined region of interest (mostly, the region of the articulators) across all the recorded rtMRI frames; this captures the temporal change in pixel intensities of the defined regions. The defined
regions of interest have been placed manually in a few studies (Bresch et al., 2010; Proctor et al., 2011). A commonly used technique for automatically finding the regions of interest (i.e., of the articulators) is by tracking the air-tissue boundaries from the image sequence. Semi-automatic approaches using snakes for automatic contour tracking have been implemented in Avila-Garcí et al. (2004) to automatically capture the contours of the vocal tract. The snake algorithm has been used in Byrd et al. (2009) for tracking the tongue, alveolar ridge, velum and pharyngeal wall. Proctor et al. (2010) has proposed another semi-automatic approach for parametric analysis of vocal tract shape by tracking the air-tissue boundaries, which has been further used for a study of articulation of Mandarin sibilants (Proctor et al., 2012). A technique of image segmentation applied on the rtMRI images using spatial frequency domain representations has been proposed in Bresch and Narayanan (2009), later used by Katsamanis et al. (2011) for monitoring the deformations of the articulators. The temporal dynamic of each feature representing deformations was modeled using HMM for the target application of phoneme recognition (Katsamanis et al., 2011). Unlike tracking the air-tissue boundaries, another technique of segmentation of articulators by directly examining the variation and co-variation of the pixel intensities is proposed by Lammert et al. (2010). By analyzing how pixels vary and co-vary over time, the rtMRI image frames were segmented into spatially localized regions in which pixels were found to be highly correlated. The pixel intensity variation in the localized regions corresponds to the vocal tract constrictions. Multimodal approaches of combining both rtMRI and EPG have also been exploited (Narayanan et al., 1997). Similarly, deep canonical correlation analysis has been used to determine critical articulatory features from the movement data of eight pellets placed on different articulators (Andrew et al., 2013a).

Feature selection for representing an image for classification tasks is a well investigated topic in the area of image processing, especially in the area of audio–visual speech recognition (Jain, 1989; Fisher, 1936; Rao, 1948), where the image is transformed using either the discrete cosine transform (DCT), principal components analysis (PCA), linear discriminant analysis (LDA) or a cascade of these. For example, only the low spatial frequencies from the DCT are kept and LDA is used on the resulting coefficients to reduce the dimensionality (Potamianos et al., 1998). As alternative to LDA, several feature selection approaches have been proposed in the literature including Canonical Correlation Analysis (Sargin et al., 2007), kernel CCA (Wang et al., 2012), deep CCA (Andrew et al., 2013b), locally linear embedding and local discriminant graphs (Fu et al., 2007), and Laplacian eigenmaps (Keller et al., 2009). One of the popular approaches is to use mutual information (MI) as a selection criterion for DCT coefficients. For example, spatial frequency components in the DCT of the image of the speaker’s mouth region are selected by utilizing the MI criterion for speech classification (Scanlon et al., 2004). Similarly, MI criterion has been used to select the most informative subset of eigen-features from PCA (Arsic and Thiran, 2006). Likewise, the distribution of time-frequency features from spectrogram have been estimated using MI for phonetic classification task (Hasegawa-Johnson, 2000; Yang et al., 1999). Visual feature selection using MI has also been proposed where a penalty for redundancy between the features is used (Gurban and Thiran, 2009) for the audio–visual speech recognition. In addition to the MI based feature selection, several works on image segmentation (Kim et al., 2005; Blake et al., 2004; Tsai et al., 2004; Heckemann et al., 2006; Klein et al., 2008; Artechevarria et al., 2009) as well as registration (Wells et al., 1996; Gaens et al., 1998; D’Agostino et al., 2004; Maes et al., 2003) utilize information theoretic criteria.

1.3. Contributions of this paper

In contrast to the feature selection in the transformed domain, in this work, we use an information theoretic measure to automatically obtain optimal regions directly from the rtMRI image domain. We also show the benefit of the articulatory features from the selected regions for broad phonetic class recognition. Since the rtMRI images capture the position of various speech articulators corresponding to different speech sounds, we hypothesize that the average intensity over a set of optimally chosen connected regions would represent the position of different articulators resulting in the representation of the respective sounds. We develop an algorithm for finding a set of regions of any arbitrary shape using MI between the articulatory features and broad phonetic classes as the criterion. In the proposed algorithm, we first quantify how much information each pixel in the rtMRI image provides about various target sound classes. Then we parametrically (threshold) group the neighboring pixels as per their sound specific information bearing potential. This offers flexibility in choosing the number of optimal regions and hence the dimensionality of the representation. Thus, the selection of the optimal regions becomes equivalent to finding the optimal threshold and the optimal number of regions which maximizes the MI between the broad phonetic classes and the articulatory representations derived from these selected regions. We also propose a greedy forward region splitting algorithm which further increases the
MI between the articulatory features from the split region and the broad phonetic classes. It should be noted that region growing and splitting (Tremeau and Borel, 1997; Miˇcuˇšík and Koˇseck, 2009) is a widely-used technique for image segmentation; however, information theoretic criterion based region splitting for feature selection has not been addressed in the literature.

We compute the optimal regions from the videos of the MRI-TIMIT corpus (Narayanan et al., 2011) consisting of four subjects; some of the optimal regions appear similar in general across subjects, while some of them are found to be subject specific. We use the optimal regions for deriving articulatory features and use them for broad phonetic class recognition. Recognition experiments separately on each speakers data reveal that on average the articulatory features result in a lower (2.26% relative) recognition error rate compared to that using acoustic-only features. However, when the articulatory features from the optimal regions are added to the acoustic-only features, the recognition error-rate drops over the articulatory-only recognition error-rate by ~8.25% (relative). In addition to the benefit in recognition, some of the features from the optimal regions are found to correlate well with the articulatory gestures involved in speech production as defined in the Articulatory Phonology (Brown and Goldstein, 1986, 1989, 1990) thereby providing interpretable articulatory representations from rtMRI images.

2. Dataset and pre-processing

We have used a multimodal rtMRI articulatory corpus (Narayanan et al., 2011) for the present study. The rtMRI corpus consists of simultaneous recording of speech and articulatory dynamics in the midsagittal plane acquired from two male (M1 and M2) and two female (F1 and F2) speakers of American English. The average age of the speakers is ~29 years. None of the speakers reported abnormal hearing or speech pathologies (Narayanan et al., 2011). Speakers’ upper airways are imaged in the midsagittal plane while they read the same 460 sentences used in the MOCHA-TIMIT corpus (Wrench and William, 2000). The image resolution in the sagittal plane is 68 × 68 pixels. Image data has a frame rate of 23.18 frames/s. Audio data is simultaneously recorded at a sampling frequency of 20 kHz inside the MRI scanner while the subjects are imaged. A specially designed noise cancellation technique is used to remove scanner noise from the recorded audio (Bresch et al., 2006). The total duration of the recordings are 39.05, 38.07, 38.19, and 37.99 min for the four subjects – M1, M2, F1, and F2, respectively. During the rtMRI recording, the 460 sentences are split into 92 groups of five sentences each. Thus the rtMRI corpus provides 92 video files for each of the four subjects. The subjects read each TIMIT sentence with a pause both before and after each sentence.

From the rtMRI video, it is observed that the position of the subject’s face in rtMRI frame changes across videos due to head movement. However, no significant change is observed across frames within one video (i.e., recording of each set of five sentences). To compensate for this drift in the position of a subject’s face, the tip of the nose is used as a fixed reference point and the nose tip location of the first frame of each of the 92 video is manually marked. Assuming the drift in the subject’s face is due to translation only in the midsagittal plane, all frames of the recordings were automatically aligned using the tip of the nose as a reference point. This procedure aligns the locations of different articulators across video frames. While a more sophisticated technique of alignment has been proposed (Kumar and Narayanan, 2014), it requires the segmentation of air-tissue boundaries, which are not available along with the rtMRI corpus. Obtaining good quality air-tissue boundaries automatically still remains a challenge; manually obtaining the segment boundaries of all video frames is time consuming and expensive. Hence, we use a simple translation based frame alignment for the present work. All 92 synchronized video and audio files, containing recordings of 5 sentences in each are split into recordings for each sentence. The splitting is done by manually marking the inter utterance pauses between consecutive sentences from the recorded audio.

It should be noted that the subjects in the rtMRI corpus commit lapses while reading the 460 sentences; these include missing words (MW), insertion of words (IW), swapping of words (SW) and replacement of a word (RW) by a different word. Hence, the audio corresponding to each sentence spoken by each subject is carefully compared to the sentence prompts given to the subjects while recording. The numbers of MW, IW, SW and RW is found to be (0, 29, 2, 5), (0, 16, 0, 9), (3, 28, 2, 18), and (2, 9, 0, 9) for F1, F2, M1, and M2, respectively. This results in manual corrections of 31, 16, 25, and 18 sentence prompts to prepare the transcriptions for F1, F2, M1, and M2, respectively. A standard English lexicon is carefully updated with all additional words in the transcriptions. A set of 51 phones is used to construct the phonetic transcriptions of each word in the lexicon. This updated lexicon is used finally in the phonetic forced-alignment using the Hidden Markov Model Toolkit (HTK) (Young, 1994). 256 mixtures are used to model the Gaussian Mixture model (GMM) in each state of the 3-state phonetic HMMs. 39-dimensional Mel frequency cepstral coefficients (MFCCs)
(including velocity and acceleration coefficients) with cepstral mean subtraction are used to construct the acoustic feature vector. A frame length of 20 ms is used as the acoustic feature vector with 10 ms frame shift. Forced-alignment yields the phonetic boundaries in eachutterance, which consists of the start frame and end frame corresponding to every phoneme. Although denoising is performed on the noisy audio recordings of rtMRI, the denoised audio, unlike a clean recording, contains segments of speech distorted due to high MRI noise. This causes errors in the forced-aligned phonetic boundaries. To remove errors in the forced-aligned phonetic boundaries, the forced aligned boundary frames were manually checked and corrected wherever necessary. These manually corrected phonetic boundaries are used as the ground truth for the present study.

3. Information theoretic optimal region selection

A set of arbitrary shaped optimal regions from the rtMRI images is derived following three steps as outlined in Fig. 1. In the first step, we compute the amount of information each pixel’s intensity provides about the broad phonetic categories. We refer to this by pixel specific phonetic information (PSPI). Once PSPI values at all pixels are calculated, we group neighboring pixels with high PSPI to form contiguous regions (Section 3.2). The number of contiguous regions and their areas are controlled parametrically. The parameters are determined to maximize the information obtained by the average intensities from these regions about different broad phonetic classes. Finally, the selected regions are split into sub regions to further increase the information provided about the broad phonetic classes. Pixel intensity as well as the average intensity of a region are random variables at a chosen rtMRI frame. Similarly, the phonetic label at an rtMRI frame is also a random variable. Hence, we quantify the information carried by an individual pixel or a group of pixels about different phonetic classes using mutual information (MI), which captures the statistical dependency between two random variables. The three steps are described in detail in the following subsections.

3.1. Information about phonetic classes by a pixel

Let $\mathcal{M}$ denotes the $68 \times 68$ rtMRI image at any time instant, where $\mathcal{M}(i, j)$ denotes the intensity at the $(i, j)$th pixel. Let $Z \in \{0, 1, 2, \ldots, K_1 - 1\}$ denotes the discrete random variable representing the pixel intensity of an rtMRI video frame, where $K_1$ is the number of different intensity levels of a pixel and $X \in \{1, 2, \ldots, K_2\}$ denotes the corresponding broad phonetic labels, where $K_2$ is the total number of broad phonetic labels. MI between $Z$ and $X$ is calculated using all the rtMRI frames annotated with phonetic labels. To compute MI the probability mass functions (PMF) of $Z$ and $X$ need to be known. Since we have only realizations of $Z$ and $X$, we compute MI by estimating the joint distribution of $Z$ and $X$ in the finite alphabet space ($\mathbb{R}^{K_1 \times K_2}$) using standard maximum likelihood criterion – frequency counts (Duda and Hart, 1983) and finally applying the definition of MI for discrete random variables as follows:

$$ I(Z, X) = \sum_{z=0}^{K_1-1} \sum_{x=1}^{K_2} P(Z = z, X = x) \log \frac{P(Z = z, X = x)}{P(Z = z)P(X = x)} $$

(1)
3.2.2. Thus, and 3.2.1. 3.2. Pixel grouping and optimal region selection

The PSPI matrix provides the information landscape over the rtMRI video frame about different broad phonetic classes. We use PSPI to group pixels and form arbitrary shaped regions which provide maximal information about broad phonetic classes. This is done in two steps – (1) obtaining peaks of PSPI(i, j), 1 ≤ i ≤ 68 and 1 ≤ j ≤ 68 and grouping pixels to form regions associated with each peak, (2) parametrically selecting a subset of these regions to maximize the MI between the articulatory features from the selected regions and the broad phonetic classes.

3.2.1. Pixel grouping

The (i, j)th pixel is declared to be a peak location in the PSPI matrix if

\[(i, j) = \arg \max_{i-1 \leq k \leq i+1, j-1 \leq l \leq j+1} \text{PSPI}(k, l)\]  

and PSPI(i, j) is declared to be the corresponding peak value. Let us assume that there are M peaks in a PSPI matrix; the mth peak is denoted by \(R_m\) with the corresponding peak value \(V_m\). We associate each pixel with one of the M peaks. This is done by performing a gradient ascent (Haykin, 2008) starting from every pixel. Suppose, starting from the (i, j)th pixel, gradient ascent leads to \(R_m\); we denote this by \((i, j) \rightarrow R_m\). A region \(R_m\) corresponding to \(R_m\) is a subset of pixels defined as follows:

\[R_m = \{(i, j) | (i, j) \rightarrow R_m\}\]  

Thus, the rtMRI video frame is segmented into M non-overlapping regions \(R_m, 1 \leq m \leq M\). Pixels in the mth region correspond to the mth peak of the PSPI matrix. It is easy to show that \(R_m\) is a connected region. Fig. 2(a) illustrates an example PSPI matrix. The peaks \(R_m\) are shown by blue dots in Fig. 2(b), which also shows the boundaries of the respective regions \(R_m\) (in red).

3.2.2. Optimal region selection

Suppose D regions need to be selected from an rtMRI video frame. This is done by selecting a subset of \(\{R_m, 1 \leq m \leq M\}\) such that these selected regions provide maximal information about the broad phonetic classes. Note that selecting too few regions would provide too little information about phonetic classes; selecting all M regions results in the inclusion of uncorrelated imaging noise since an rtMRI video frame contains regions outside the vocal tract which provide no cues about the different sounds a subject produces while speaking. As a part of the optimal region selection, we also select a subset of pixels in each region such that the PSPI values at these pixels are high; this helps in pruning

---

1 We do not show peaks and respective regions with peak value \(V_m < 0.001\), which are mostly outside the subject’s face.
noisy pixels in each region which do not provide a significant amount of information about broad phonetic classes. A subset of pixels $\overline{K}_m$ is selected parametrically from $\overline{K}_m$ using a parameter $\eta \in (0, 1)$ as follows:

$$\overline{K}_m = \{(i, j)| (i, j) \in \overline{K}_m \text{ and PSPI } (i, j) \geq \eta V_m\}$$  \hspace{1cm} (4)

Note that because of this thresholding, it is easy to show that $\overline{K}_m$ will be a connected region. Given values of $D$ and $\eta$, a $D$-dimensional articulatory features vector $z = [z_1, z_2, \ldots, z_d, \ldots, z_D]^T$ (‘T’ denotes the transpose operator) is computed from an rtMRI image as follows:

$$z_d = \frac{1}{|\overline{K}_d|} \sum_{(i, j) \in \overline{K}_d} M(i, j)$$  \hspace{1cm} (5)

where $|\overline{K}_d|$ denotes the cardinality of the set $\overline{K}_d$, i.e., the number of pixels in $\overline{K}_d$. Thus, $z_d$ captures the average intensity over the $d$th region. Note that $z_d$ depends on the choice of $\eta$ while the dimension of $z$ depends on the number ($D$) of selected regions. Because of the movement of articulators during speech production, the average intensity in a region may reflect the presence or absence of an articulator which in turn could indicate the type of sound produced.

The optimal values of $D$ and $\eta$ are computed such that MI between $z$ and $X$ (broad phonetic classes) is maximized as follows:

$$\{D^*, \eta^*\} = \arg\max_{D, \eta} I(z, X)$$  \hspace{1cm} (6)

Note that unlike integer values of $Z$ in Eq. (1), elements of $z$ can take real values depending on the cardinality of $\overline{K}_d$. Since the probability density function of $z$ is unknown, we quantize the space of $z$ (denoted by $Q(z)$) using a pool of rtMRI frames with a finite number ($K_1$) of quantization bins using $K$-means vector quantization (Duda and Hart, 1983; Cover and Thomas, 1991) and then apply Eq. (1) to compute $I(Q(z), X)$. It is well known that $I(Q(z), X)$ is a lower bound of the MI between $z$ and $X$ and with increasing quantization resolution ($K_1$), $I(Q(z), X)$ converges to actual MI. We have chosen $K_1 = 256$ since increasing $K_1$ further does not change the results significantly. We perform a grid search over $D$ and $\eta$ jointly to obtain their optimal values denoted by $D^*$ and $\eta^*$, the corresponding optimal regions and articulatory features are denoted by $\overline{K}_d^*$ ($1 \leq d \leq D^*$) and $z^*$, respectively; the respective value of MI is denoted by $I^* = I(Q(z^*), X)$. Fig. 2(c) shows the optimal regions for $D^* = 8$ and $\eta^* = 0.3$ for the PSPI matrix shown in Fig. 2(a). The optimal regions’ boundaries are shown by solid red lines and the corresponding original $D^*$ regions’ boundaries are shown in dashed red lines (equivalent to $\eta = 0$).

### 3.3. Region splitting

The articulatory features from the optimally selected regions are computed as the average pixel intensity over each connected region (Eq. (5)). However, one connected region may correspond to multiple articulators, which are spatially located next to each other. Thus the average intensity over the entire region may reflect the combined movement of multiple articulators. In contrast to this, capturing dynamics of an individual articulator may provide more information about the different broad phonetic categories. This could be achieved by splitting each of the optimally chosen regions.

Given $D^*$ optimally selected regions ($\overline{K}_d^*$, $1 \leq d \leq D^*$), we use MI as the criterion to develop an algorithm which automatically determines the required splits of each region such that the articulatory features computed from the split regions provide more information about broad phonetic classes compared to that from $D^*$ regions. If splitting a region reduces the MI, we do not split the region. Thus, for region splitting, we address two problems here – (1) determining the optimal splits for a given region, (2) determining the optimal number of splits for all regions.

#### 3.3.1. Optimal split of a region

Let $\rho = \{\rho_1, \rho_2, \ldots, \rho_d, \ldots, \rho_D\}$ be the set of $D$ arbitrary shaped connected regions, where $\rho_d$ is the $d$th connected region. Suppose the goal is to optimally split $\rho_d$. Let $F(\rho_d)$ denotes the feature computed from $\rho_d$ using Eq. (5). Let $F(\rho)$ be the $D$-dimensional feature vector whose $d$th element is $F(\rho_d)$, i.e., $F$ is a mapping from a set of regions to a vector of features.

A tree-structured partition of $\rho_d$ is performed for region splitting where, in each level of the partition, MI between the feature vector from the split region and the broad phonetic classes is computed. The partitioning is stopped when
the MI does not increase any more. Suppose the $d$th region $\rho_d$ has the bounding rectangular box with minimum and maximum row indices $i_{\text{min}}$ and $i_{\text{max}}$, respectively and similarly, minimum and maximum column indices $j_{\text{min}}$ and $j_{\text{max}}$, respectively as illustrated in Fig. 3. In each level of the split, the best partition is found by searching over all possible horizontal and vertical splits in the regions of the respective levels, i.e., row indices from $i_{\text{min}}$ to $i_{\text{max}}$ and column indices from $j_{\text{min}}$ to $j_{\text{max}}$. The steps for optimally splitting a region are summarized in Algorithm 1 (Appendix A).

In an illustrative example shown in Fig. 3, the best split of $\rho_d$ in the first level turns out to be horizontal resulting in two split regions $\rho_{d,1}$ and $\rho_{d,2}$. In the next level of split, a vertical split of $\rho_{d,2}$ improves the MI further resulting in split regions $\rho_{d,21}$ and $\rho_{d,22}$. But there is no improvement in MI by splitting $\rho_{d,1}$. Thus, $\{\rho_{d,1}, \rho_{d,21}, \rho_{d,22}\}$ becomes the final split region.

3.3.2. Optimal number of splits for each region

From Section 3.3.1, it is clear that the optimal split of a region ($R_d^*$) depends on the remaining regions because the split is determined such that the articulatory features computed from all regions including the split ones have maximal MI with the broad phonetic classes. Thus, the optimal number of splits for each region depends on the sequence of the regions during splitting. Therefore, determining the optimal number splits for all $D^*$ regions jointly is a combinatorial problem. However, performing a full search for finding the optimal number of splits becomes computationally prohibitive, particularly, due to the clustering of the features required for computing MI. To circumvent this problem, we propose a forward region splitting algorithm which finds the optimal number of splits of $\rho_d$ based on the optimal number of splits of all previous regions, $\{\rho_1, \ldots, \rho_{d-1}\}$, where $2 \leq d \leq D^*$ is the region index. The steps of the forward region splitting algorithm are summarized in Algorithm 3 (Appendix A). It is clear from Algorithm 3 that the forward region splitting would provide a sub-optimal solution compared to the solution obtained from a full search. We denote the total number of regions after split by $D^*_s$ and the corresponding MI by $I^*_s$.

4. Experiments and results

4.1. Experimental setup

The broad phonetic class recognition experiment is performed separately for each subject of the rtMRI corpus. Since there are recordings of only 460 utterances for each subject, the number of frames for several fine class phonemes
turns out to be small, which is insufficient for building a good statistical model for each phoneme required for a fine phonetic class recognition. Hence, following the work by Sainath et al. (2007), we combine the fine class phonemes into five broad classes, namely VOWELS, STOPS, FRICATIVES, NASALS and SILENCE as shown in Table 1. The broad phonetic class recognition experiment is performed in a 5-fold cross validation setup – four folds are used for training and remaining one for testing in a round-robin fashion.

To match with the rate of the acoustic features, we up-sample the rtMRI frames to 100 Hz. This is done by separately up-sampling the temporal trajectory of each pixel intensity over time. The optimal regions (both with and without region splitting) are obtained from the training set of each fold for every subject separately. \( D \) and \( \eta \) are varied over the sets \( \{1, 2, 3, \ldots, 15\} \) and \( \{0.1, 0.2, 0.3, \ldots, 0.9\} \), respectively to find the optimal set of regions.

In order to compare with the proposed region selection method, we use three baseline approaches – one of these is based on the selection of pixels using inter-pixel correlation (denoted by BAF_C), remaining two are based on the well-known dimensionality reduction techniques, namely PCA (denoted by BAF_P) and information gain followed by PCA (denoted by BAF_IP). It should be noted that BAF_C is based on the analysis of correlated image regions in rtMRI, while BAF_P and BAF_IP are based on the dimensionality reduction techniques. In BAF_C, we obtain highly correlated spatially localized regions following the assumption that the pixels which are highly correlated correspond to the regions of the moving articulators (Lammert et al., 2010). We find that many of these regions are outside the area of speech articulators. We use connected pixels to form contiguous regions. It is found that pixels outside the largest seven regions do not fall on the speech articulators. Thus, the top seven highly correlated regions are chosen as the optimal regions for BAF_C. In BAF_P, each rtMRI image frame is vectorized into \( 68 \times 68 \) (=4624) dimensional vector. PCA is used to reduce the dimensionality of this vector to 19. Increasing or decreasing the number of principal components does not improve the recognition accuracy. In BAF_IP, we first use information gain criterion (Dhir et al., 2007; Wu and Xu, 2015) to select the top 250 most informative pixels. These 250 dimensional vectors are further reduced using PCA. On average, the recognition performance is found to be maximum by using top 19 principal components. For all three baseline approaches, we have used the velocity and acceleration features resulting in 21, 57, and 57 dimensions for BAF_C, BAF_P, BAF_IP, respectively.

The articulatory features derived from the proposed optimal regions are used for training the broad phonetic class recognizer. The articulatory features (along with their velocity and acceleration components) obtained from the optimal regions (without split) are referred to as optimal articulatory features (OPAF). OPAF_S refers to the features from the optimal regions after split. For the recognition, a 3-state left-to-right HMM is used for each phonetic class with a 256-component GMM to model the emission probability density at each state. No phonetic language model was used for the recognition experiment. The recognition results obtained using the OPAF and OPAF_S are compared with that using BAF_C, BAF_P, and BAF_IP as well as using the acoustic features (ACF). Finally the benefit of combining the acoustic and articulatory features (ACF+OPAF_S) in recognition is also analyzed. It should be noted that when the dimensionality of the feature vector is greater than 39, we perform the heteroscedastic linear discriminant analysis (HLDA) (Kumar and Andreou, 1998) to reduce the feature vector dimension to 39 in order to match with the dimension of the ACF. Remaining setups for the recognition experiment are kept identical.

---

2 The number of pixels obtained from the proposed optimal regions is lower than 250 for almost all folds of four subjects; to match with the size of the optimal regions we choose 250 in BAF_IP.
4.2. Results and discussions

We first present the optimal regions selected with and without splitting and compare the same with the regions selected from the baseline approaches. Following this, we present the recognition results using acoustic, articulatory as well as acoustic + articulatory features.

4.2.1. Comparison of regions selected by proposed and baseline approaches

Four illustrative PSPI matrices (one from a randomly chosen fold of each subject in the rtMRI corpus) are shown in Fig. 4. It is clear that a very low value of PSPI is observed outside the subject’s face. In general, for all subjects, the regions corresponding to high value of MI fall on velum, glottis, tongue tip and jaw. However, the regions with high MI turn out to be subject specific. For example, in the case of F2, the pixels near upper lip and lower lip yield high value of PSPI. Similarly in the case of F1, the pixels near lower lip also show high PSPI. However, this is not the case with M1 and M2. It is also seen that regions near the throat of F2 have high PSPI. This could be because there is significant throat movement near the laryngeal prominence while F2 speaks. We observe two high PSPI regions near the velum consistently in all folds of four subjects. Interestingly, they correspond to the closing and opening positions of the velum.

Pixel grouping and optimal region selection are performed using the PSPI matrix separately for each fold of all four subjects as described in Section 3.2. The MI obtained by varying $D$ and $\eta$ are shown in Fig. 5 for all five folds for the four subjects. Red dot on each plot indicates the best combination of $D$ and $\eta$ (i.e., $D^*$ and $\eta^*$) which yields the highest MI. It is clear that the best choice of $D$ and $\eta$ is fold and subject specific. However, we observe a common pattern in the values of $D$ for all folds of a subject. For example, for F1 and M1 the values of $D^*$ are $\{5, 11, 7, 6, 5\}$ and $\{9, 8, 6, 8, 5\}$, respectively for all five folds. These are lower compared to the $D^*$ in the case of F2 and M2, $\{14, 15, 13, 15, 15\}$ and $\{14, 14, 14, 14, 10\}$, respectively. Similarly, the best values of $\eta$ for all folds of F1 are on the lower side ($\{0.2, 0.2, 0.2, 0.2\}$) while those for M2 are on the higher side ($\{0.7, 0.8, 0.7, 0.9, 0.9\}$). However, no such consistent pattern is observed in the case of F2 ($\{0.4, 0.6, 0.8, 0.8, 0.7\}$) and M1 ($\{0.2, 0.9, 0.6, 0.6, 0.8\}$). This could be because
of the variation in the anatomy and articulatory dynamics across subjects. Thus the optimal number of regions could vary reflecting the articulatory regions involved in speech production.

The optimal regions ($\mathcal{R}_d$, $1 \leq d \leq D^*$) corresponding to the PSPI matrices of Fig. 4 are shown in Fig. 6. Each optimal region is marked by one color to distinguish from other optimal regions. 7, 13, 8, and 14 optimal regions consisting of a total of 153, 59, 86, and 74 pixels (for F1, F2, M1, M2), respectively are obtained by the optimal region selection process. The highest number of pixels of optimal regions among all folds of four subjects is less than 6% of the total number of 4624 (68 × 68) pixels. This suggests that very few pixels, in fact, carry maximal information about the broad phonetic classes. The optimal regions, while remaining similar across different folds of a subject, vary across different subjects. However, some of the optimal regions fall on similar articulators across different subjects. For example, portions of the tongue tip, velum and glottis fall on the optimal regions for all folds of four subjects. Two regions in the velum are consistently selected as the optimal regions; they correspond to the open and closed position of the velum. These regions could help in classifying nasal vs. non-nasal sounds. Similarly, regions near the glottis could be useful for classifying vowel vs. non-vowel sounds. The optimal values of MI ($\mathcal{I}^*$) corresponding to the optimal regions in case of these illustrative examples are 0.56, 0.42, 0.53, and 0.55 (Fig. 6). These are higher compared to the highest MI obtained using any individual pixel in the respective cases (Fig. 4). This suggests that, by grouping pixels with high PSPI, the MI with the broad phonetic classes increases.

The regions obtained by splitting the optimal regions are illustrated in Fig. 7. Every split region is also marked by one color to distinguish from other split regions within an optimal region as well as remaining optimal regions. It is interesting to note that all optimal regions are split except for small sized regions. For examples, all of 7 optimal regions are split in the case of F1, 6, 2, and 5 regions among 13, 8, 14 regions are not split for F2, M1, and M2, respectively. All these regions have area of 4 pixels or less. From the values of $\mathcal{I}_s$ shown in Fig. 7, it is clear that, due to splitting, the MI increases compared to that using optimal regions without splitting (Fig. 6). This suggests that splitting regions indeed provides more MI about the broad phonetic classes. However, the improvement in MI due to splitting is lower than the improvement in MI due to pixel grouping from PSPI. Splitting increases the dimension of the articulatory feature vector. This in turn can capture more details of articulatory dynamics as opposed to fewer regions, which could...
Fig. 6. Optimal regions obtained using the PSPI matrices of from Fig. 4. The corresponding best choices of $D^*$, $\eta^*$ are shown on the respective plots. The values of $I^*$ corresponding to the optimal regions are also indicated. Each selected region for every subject is coded with one color. For example, in (a) seven colors are used to represent $D^*$ ($=7$) selected regions.

only capture an average effect of multiple articulators’ dynamics. For example, in the case of F1 and M2 the optimal region on the tongue has been split into two regions which can be associated with tongue tip and tongue body or tongue dorsum. These articulators correspond to the production of different phonemes. This in turn may increase the MI with the broad phonetic classes.

Fig. 8(a) and (b) shows the seven regions obtained in BAF_C and the top 250 informative pixels in BAF_IP (before applying PCA), respectively. Note that in BAF_P, all pixels in the image are used in dimensionality reduction and hence no such region-like interpretation is possible for the features obtained in BAF_P. It is clear that the selected pixels in both BAF_C and BAF_IP fall mostly on the speech articulators including the nose, upperlip, lowerlip, jaw, parts of the tongue, glottis and velum. It should be noted that the number of pixels in the regions from BAF_C varies across folds and subjects similar to the OPAF and OPAF_S while that from BAF_IP is fixed (250 pixels). However, the MI values obtained using the baseline approaches are lower than that using OPAF_S. As shown in Table 2, MI between the articulatory features and the broad phonetic classes is highest in the case of OPAF_S for all of the four subjects considered in this work. This suggests that optimal regions selected using the proposed approach are more

![Fig. 6. Optimal regions obtained using the PSPI matrices of from Fig. 4. The corresponding best choices of $D^*$, $\eta^*$ are shown on the respective plots. The values of $I^*$ corresponding to the optimal regions are also indicated. Each selected region for every subject is coded with one color. For example, in (a) seven colors are used to represent $D^*$ ($=7$) selected regions.](image)

![Fig. 8(a) and (b) shows the seven regions obtained in BAF_C and the top 250 informative pixels in BAF_IP (before applying PCA), respectively. Note that in BAF_P, all pixels in the image are used in dimensionality reduction and hence no such region-like interpretation is possible for the features obtained in BAF_P. It is clear that the selected pixels in both BAF_C and BAF_IP fall mostly on the speech articulators including the nose, upperlip, lowerlip, jaw, parts of the tongue, glottis and velum. It should be noted that the number of pixels in the regions from BAF_C varies across folds and subjects similar to the OPAF and OPAF_S while that from BAF_IP is fixed (250 pixels). However, the MI values obtained using the baseline approaches are lower than that using OPAF_S. As shown in Table 2, MI between the articulatory features and the broad phonetic classes is highest in the case of OPAF_S for all of the four subjects considered in this work. This suggests that optimal regions selected using the proposed approach are more](image)

<table>
<thead>
<tr>
<th>Table 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average MI between different articulatory features and the broad phonetic classes in case of each subject. The entries in the bracket indicate the standard deviation (SD) across 5 folds. Bold entries in each column correspond to the highest MI across different articulatory features for each subject.</td>
</tr>
<tr>
<td>Feature</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>OPAF</td>
</tr>
<tr>
<td>OPAF_S</td>
</tr>
<tr>
<td>BAF_C</td>
</tr>
<tr>
<td>BAF_P</td>
</tr>
<tr>
<td>BAF_IP</td>
</tr>
</tbody>
</table>
Fig. 7. Regions obtained by splitting the optimal regions from Fig. 6. The number of regions (and the corresponding MI) thus obtained are indicated in the respective plots. For example, \( D^* (=7) \) regions in Fig. 6(a) are split into \( D^* (=16) \) regions, each of which is indicated by one color.

Fig. 8. Regions obtained by the BAF_C and BAF_IP for a randomly chosen fold of each subject (as in Fig. 6).

representative of the broad phonetic classes compared to those using inter-pixel correlation and generic dimensionality reduction techniques.

The proposed information theoretic approach results in the subject specific optimal regions. This is mainly because the vocal tract morphology changes across subjects. Also the physical size and shape of the head are different for
different subjects. In addition to that, different subjects’ faces are not placed identically in the same location of the rTMR image frame in the mid-sagittal plane during the recording. In that sense, optimal regions obtained for one subject cannot be directly used to compute articulatory features for another subject. Hence, multiple subjects’ data may not be useful to improve optimal region selection. Although we have used only four subjects for the present study, the proposed approach can readily be used to obtain optimal regions from a new subject’s rTMR video frames. Thus, the application of the proposed approach for finding optimal regions is not limited to the subjects used for the present study.

4.2.2. Broad phonetic class recognition results

Broad phonetic class recognition is performed using acoustic, articulatory features as well as their combinations. Fig. 9 shows the errorbar plot of the recognition accuracies averaged across different folds of four subjects using articulatory features from BAF_C, BAF_P, BAF.IP, OPAF, and OPAF_S. Fig. 9 also compares different approaches using the recognition performance averaged across four subjects. The errorbars indicate one standard deviation (SD). It is clear that the recognition accuracies using different features vary across subjects. We also observe that BAF.IP outperforms two other baseline approaches in terms of the average recognition accuracy for all four subjects. BAF.IP yields the highest recognition accuracy in the case of M2. Recognition accuracy using OPAF_S is significantly better than that using OPAF for all subjects consistently. This is also consistent with the fact that MI improves by splitting the optimal regions for all folds of four subjects (Table 2). Improvement in both MI and recognition accuracy due to splitting suggests that MI is a metric that correlates well with the recognition accuracy. From the recognition accuracies averaged across subjects (Fig. 9) it is clear that BAF.IP is the best performing (56.07%) baseline scheme, which yields more recognition accuracy than that provided by OPAF. However, OPAF.S provides a recognition accuracy of 58.03%, which is not only higher over OPAF but also over BAF.IP indicating its superiority over various articulatory features considered in this work. This pattern in the recognition accuracy is also consistent with the MI values in Table 2, where BAF.IP has the highest MI among three baseline schemes and OPAF.S has higher MI compared to BAF.IP.

In Fig. 10, we compare the recognition accuracy using ACF and that using the best baseline (BAF.IP) and the proposed (OPAF.S) schemes separately for each subject as well as averaged across subjects. The highest recognition accuracy using ACF turns out to be 59.62% for F2. Sainath et al. (2007) reported ∼80% broad phonetic class recognition accuracy using acoustic features from clean audio of the TIMIT corpus (DARPA-TIMIT, 1990). Poor recognition accuracy using ACF in the present study is mainly because the audio recording is done in presence of the high MRI noise. Although the audio recordings are denoised, the denoised audio still suffers from distortion unlike clean audio recording. For F1 and M2, the recognition using the best articulatory feature (OPAF.S) is higher than that using ACF, while that is not the case for F2 and M1. However, on average, the highest recognition accuracy among different articulatory features is higher than that using ACF. This indicates the potential of the articulatory features for broad phonetic class recognition. In Fig. 10, we also investigate the benefit of the best baseline and proposed articulatory features (BAF.IP and OPAF.S) further by combining them with acoustic features for recognition. On average, ACF + BAF.IP results in a recognition accuracy of 57.90%, which is higher than that using ACF (57.06%) and BAF.IP (56.70%) separately.
(1.96% and 2.77% relative error-rate reduction). This suggests that ACF and BAF_IP carry complementary information about the broad phonetic classes. On the other hand, ACF + OPAF_S yields an average recognition accuracy of 61.50%, which is higher than the individual ones (10.33% and 8.25% relative error-rate reduction); this is true even for each subject separately. This suggests that ACF and OPAF_S are more complementary in nature compared to ACF and BAF_IP. It is also clear that the recognition error-rate using ACF + OPAF_S is 8.53% (relative) lower than that using ACF + BAF_IP. In fact, the recognition accuracy using ACF + OPAF_S is the highest among recognition accuracies obtained using different acoustic and articulatory features.

4.2.3. Discussions

Both MI (Table 2) and recognition accuracy (Figs. 9 and 10) reveal that the articulatory features obtained from the optimal regions are efficient representations of the articulatory dynamics for discriminating broad phonetic classes. We find that the articulatory features from some of the optimal regions closely match with some of the articulatory gestures, which are defined as the discrete action units in the Articulatory Phonology theory of speech production (Browman and Goldstein, 1986, 1989). The activation of the gestures results in constriction formation or release by five distinct constrictors (lips, tongue tip, tongue body, velum and glottis) along the vocal tract. The kinematic state of each constrictor is defined by its corresponding constriction degree and location coordinates, which are called tract variables (TVs). The constriction degree TVs are lip aperture (LA), tongue tip constriction degree (TTCD), tongue body constriction degree (TBCD), velum (VEL), glottis (GLO). Similarly, the constriction location TVs are lip protrusion (LP), tongue tip constriction location (TTCL) and tongue body constriction location (TBCL). Since the average intensity over a selected region is considered as the articulatory feature, we observe that some of them correlate well with the constriction degree TVs.

For example, Fig. 11 compares the GLO, TTCD, and VEL with the articulatory features computed from a few optimal regions for all four subjects for the word ‘thin’. The TV trajectories corresponding to ‘thin’ are generated using the Haskins Laboratories Task Dynamics Application (Nam et al., 2004). In this model, gestural scores are generated from orthographic or ARPABET transcription inputs, according to the principles of Browman and Goldstein’s Articulatory Phonology (Browman and Goldstein, 1990); TVs are computed using the Task Dynamic (TaDA) model of gestural pattern dynamics (Saltzman and Munhull, 1989). The black dashed plots correspond to the TVs. The high value of the GLO in the initial portion indicates the wide glottis corresponding to the unvoiced phoneme /θ/ in ‘thin’. For the remaining part of the word, low value of GLO indicates glottis constriction or vibration. Low value of TTCD also indicates tongue tip constriction for the phoneme /θ/. High value of TTCD following that corresponds to phoneme /I/ which is followed by a low value of TTCD again corresponding to the tongue constriction for /n/. Towards the end of the word, the VEL becomes high indicating the lowering of the velum leading to the production of nasal sound /n/. GLO, TTCD and VEL (in black dashed lines) are identical across columns corresponding to each subject and are used as references to compare against the articulatory features from the proposed region selection algorithm. Red, blue and violet curves are the plots of the three articulatory features trajectories corresponding to the word ‘thin’ from three optimal regions indicated by the respective colors in the first row in Fig. 11, which also shows a template rtMRI image.
where darker region corresponds to higher pixel intensity. Note that the range of values of TV is in general different from that of the articulatory feature values computed from the optimal regions. Hence, they are scaled to the same range before comparison in Fig. 11. Articulatory features are inverted in order to match with the activation profile of the GLO, TTCD and VEL. This is because the articulatory features, being average pixel intensity, increase in the presence of articulators during constriction whereas the TVs have low values during constriction.

It is clear from Fig. 11 that the articulatory feature trajectories (in red, blue and violet) correlate well with the corresponding gestural score profiles. For example, we observe high value of the articulatory features (red curve) when the gestural score for GLO is high and the feature value is low towards the end of the word similar to the drop in the gesture score. Similarly, the articulatory feature (blue curve) is high during phoneme /I/ with low values before and after, which matches with the TTCD score profile. On the same note, the articulatory feature (violet curve) increases during the word following the VEL score pattern. Hence, the proposed articulatory features can be used to analyze the subject specific constriction patterns while uttering the same word. For example, in the case of GLO, there are subtle differences in the articulatory feature trajectories across subjects although they are computed using regions that fall on the glottis of all subjects. This could be due to the difference in the exact location, shape and size of the optimal regions near the glottis of different subjects. This could also be due to the anatomical differences, differences in the speaking rate and articulation style of the subjects. Similar differences are observed for TTCD and VEL across four subjects as well.

Fig. 12 compares the two gestural scores, namely VEL and TTCD with the articulatory features when subjects speak ‘ended’. GLO is not shown for comparison since all phonemes in this words are voiced and, hence, there is no change in the activation pattern of GLO. Similar to Fig. 11, the optimal regions corresponding to TTCD, while all falling near the tongue tip, vary in shape and size. This is true even for the optimal region corresponding to VEL. The articulatory features from the optimal regions correlate well with the gestural score profile. However, there is subject specific variation in the articulatory feature trajectories similar to Fig. 11. Thus the gestural interpretation of
the articulatory features is a key characteristic of the proposed optimal regions based representation as opposed to the representation obtained by traditional dimensionality reduction technique.

5. Conclusions

In this work, we have proposed an optimal region selection algorithm for broad phonetic class recognition using MI as the criterion. The proposed optimal regions are used to compute a set of articulatory features from rtMRI video frames. As shown in Fig. 10, it is found that on average the broad phonetic class recognition accuracy obtained using the proposed features is higher than that using acoustic features computed from the rtMRI audio. Combining these two features improves the recognition accuracy over those using the individual ones. Thus, the proposed region selection based features are not only equivalent but also complementary to the rtMRI acoustic features. In addition to the recognition benefit, some of the articulatory features using the optimal regions are found to correlate well with the articulatory gestures typically used to describe the constriction actions in the Articulatory Phonology theory (Brownman and Goldstein, 1990) of speech production.

Although the proposed optimal region selection algorithm does not use any a priori information about the location of subject’s articulators in the rtMRI image, it is interesting to find that the optimal regions fall on different articulators that are actively used for speech production such as lips, jaw, tongue, velum, and glottis. From Fig. 6, it is interesting to find that for one female subject (F2) from the rtMRI corpus, the optimal regions fall on the laryngeal prominence indicating the movement of the Adam’s apple while speaking. The proposed region selection can also be used to automatically identify parts of different speech articulators in the rtMRI video. While there are several pixels in rtMRI images which correspond to different speech articulators, less than ~6% of the total number of pixels in the image were found to be part of the optimal regions. This suggests that the phonetic representations are localized primarily few pixels of 68 × 68 rtMRI image corresponding to the regions of critical articulators.

It would be interesting to compare the proposed optimal region based articulatory features with those from other modalities such as EMA. EMA directly measures the dynamics of critical articulators, but rtMRI provides a low-resolution image of the entire vocal tract in the midsagittal plane. Comparing representations from these modalities would reflect their relative potential for designing features for speech applications. It would also be interesting to investigate other techniques such as active shape model (Luettin et al., 1996) and broad statistical pixel measure (Racesy et al., 2013) for selecting optimal regions in the rtMRI image.
Appendix A. Algorithms required for region splitting

Algorithm 1. Optimal split of $\rho_d$ (OPTIMALSPLITREGION).
1: \textbf{Inputs:} $\rho$, $d$, and broad phonetic labels $X$ of multiple rtMRI frames
2: $\rho' = \{\rho\} \cup \rho^{ES} \setminus \{\rho_d\}$, and $\rho_T = \rho \setminus \{\rho_d\}$
3: $I^* = \{Q \left\{ F \left( \{ \rho_T, \rho_e', \rho_e'' \} \right) \right\} , X \}$
4: $(\rho^3, I^*) = \text{SPLITONELEVEL} \left( \{ \rho_T, \rho_e', \rho_e'' \} , \rho^{ES}, I^* \right)$
5: if $I^* = I'$ then
6: \hspace{1em} return $\rho^3$
7: end if
8: while $I^* > I'$ do
9: $I' = I^*$
10: $\rho^3_1 \leftarrow$ first element (region) of $\rho^3$
11: $\rho^3_2 \leftarrow$ second element (region) of $\rho^3$
12: $(\rho^{3,1}, I^{*1}_1) = \text{SPLITONELEVEL} \left( \{ \rho_T, \rho_e', \rho_e'' \} \right)$
13: $(\rho^{3,2}, I^{*1}_2) = \text{SPLITONELEVEL} \left( \{ \rho_T, \rho_e', \rho_e'' \} \right)$
14: if $I^{*1}_1 > I^{*1}_2$ then
15: $I^{*1}_3 = I^{*1}_1$
16: $\rho^3 = \rho^3_1$
17: $\rho^t = \rho^t \cup \{ \rho^3_2 \}$
18: else
19: $I^{*1}_3 = I^{*1}_2$
20: $\rho^3 = \rho^3_2$
21: $\rho^t = \rho^t \cup \{ \rho^3_1 \}$
22: end if
23: end while
24: $\rho^T = \rho^t \cup \{ \rho^3 \}$
25: return $\rho^T$ and $I^*$. 

Algorithm 2. Optimal split for one time (SPLITONELEVEL).
1: \textbf{Inputs:} $\rho'$, $\rho''$, $I$
2: $i_{\text{min}} \leftarrow \min \{i | (i, j) \in \rho''\}$
3: $i_{\text{max}} \leftarrow \max \{i | (i, j) \in \rho''\}$
4: $j_{\text{min}} \leftarrow \min \{j | (i, j) \in \rho''\}$
5: $j_{\text{max}} \leftarrow \max \{j | (i, j) \in \rho''\}$
6: $I_{\text{max}} = I$, $\rho_{\text{best}} = \{\rho''\}$
7: for $k = i_{\text{min}} + 1$ to $i_{\text{max}} - 1$ do
8: $\rho_1 = \{(i, j) | (i, j) \in \rho'' \text{ and } i \leq k\}$
9: $\rho_2 = \{(i, j) | (i, j) \in \rho'' \text{ and } i > k\}$
10: $I = \{Q \left\{ F \left( \{ \rho', \rho_1, \rho_2 \} \right) \right\} , X \}$
11: if $I > I_{\text{max}}$ then
12: $I_{\text{max}} = I$
13: $\rho_{\text{best}} = \{\rho_1, \rho_2\}$
14: end if
15: end for
16: for $k = j_{\text{min}} + 1$ to $j_{\text{max}} - 1$ do
17: $\rho_1 = \{(i, j) | (i, j) \in \rho'' \text{ and } j \leq k\}$
18: $\rho_2 = \{(i, j) | (i, j) \in \rho'' \text{ and } j > k\}$
19: $I = \{Q \left\{ F \left( \{ \rho', \rho_1, \rho_2 \} \right) \right\} , X \}$
20: if $I > I_{\text{max}}$ then
21: $I_{\text{max}} = I$
22: $\rho_{\text{best}} = \{\rho_1, \rho_2\}$
23: end if
24: end for
25: return $\rho_{\text{best}}$ and $I_{\text{max}}$. 
Algorithm 3. Optimal number of splits of all region in $\rho$.

1: Inputs: $\rho$ and broad phonetic labels $X$ of multiple rtMRI frames
2: $\rho^{\text{split}} \leftarrow \emptyset$
3: for $i=1$ to $D^n$ do
4: $I_{\text{max}} = \text{OPTIMALSPLITREGION}(\rho, \rho^{\text{split}})$
5: for $j=1$ to $|\rho|$ do
6: $(\rho^{S_j}, I) = \text{OPTIMALSPLITREGION}(\rho_j, \rho^{\text{split}})$
7: if $I > I_{\text{max}}$ then
8: $I_{\text{max}} = I$
9: $\rho^{*} = \rho^{S_j}$
10: end if
11: end for
12: $\rho^{\text{split}} = \rho^{\text{split}} \cup \rho^{*}$
13: $\rho = \rho \setminus \rho_i$
14: end for
15: return $\rho^{\text{split}}$ and $I_{\text{max}}$

References


