ESTIMATION OF THE INVARIANT AND VARIANT CHARACTERISTICS IN SPEECH ARTICULATION AND ITS APPLICATION TO SPEAKER IDENTIFICATION

Abhay Prasad[1], *Vijitha Periyasamy*[2], *Prasanta Kumar Ghosh*[2]

[1]Manipal Institute of Technology, Manipal 576104, India
[2]Department of Electrical Engineering, Indian Institute of Science (IISc), Bangalore 560012, India.
abhayprasad.337@gmail.com, vijithaster@gmail.com, prasantg@ee.iisc.ernet.in

ABSTRACT

Speech articulation varies across speakers for producing a speech sound due to the differences in their vocal tract morphologies, though the speech motor actions are executed in terms of relatively invariant gestures [1]. While the invariant articulatory gestures are driven by the linguistic content of the spoken utterance, the component of speech articulation that varies across speakers reflects speaker-specific and other paralinguistic information. In this work, we present a formulation to decompose the speech articulation from multiple speakers into the variant and invariant aspects when they speak the same sentence. The variant component is found to be a better representation for discriminating speakers compared to the speech articulation which includes the invariant part. Experiments with real-time magnetic resonance imaging (rtMRI) videos of speech production from multiple speakers reveal that the variant component of speech articulation yields a better frame-level speaker identification accuracy compared to the speech articulation as well as acoustic features by 29.9% and 9.4% (absolute) respectively.

Index Terms— speech articulation, invariant gestures, speaker identification

1. INTRODUCTION

Discovering the variant and invariant aspects in speech is fundamental to extract the linguistic message as well as analyze the paralinguistic information including speaker's characteristics from the speech signal [2, 3]. The variant and invariant aspects in speech have been investigated both in the acoustic [4, 5] as well as in the articulation domain [1, 6]. The invariant representation from speech acoustics aims at removing the speaker variability and distortions including spectral shaping, background noise and reverberation [4]. The variant aspects in speech acoustics may arise as a result of a phonological process or the speaker characteristics [7]. Similarly, there is variability in the movement of speech articulators when an utterance is spoken by a speaker multiple times as well as by multiple speakers, although the execution and planning of speech motor actions are in terms of relatively invariant multimovement gestures [1]. Thus the variation in speech articulation across speakers is due to the differences in the morphology of speakers [8, 9] including the vocal tract size [5] and the speaker-specific articulatory dynamics.

In this work, we present a formulation for decomposing the speech articulation of an utterance by multiple speakers into its variant and invariant components. We represent the speech articulation by the articulogram defined as the sequence of vocal tract tube profile (VTTP) estimated from the real-time magnetic resonance imaging (rtMRI) video of articulatory dynamics in the mid-sagittal plane recorded when a speaker utters a sentence. VTTP captures the vocal tract shaping and hence represents speaker's articulation. Since different speakers take different amount of time to utter the same sentence, we first temporally align the articulograms of different speakers using the phonetic boundaries. Then the temporally aligned articulogram of each speaker is assumed to be a spatially (along the vocal tract) warped version of an invariant articulogram. We hypothesize that this warping captures the speaker-specific variation in articulation. Both the invariant articulogram as well as the warping functions are estimated using the articulograms from multiple speakers corresponding to a particular sentence. The variant component is obtained by subtracting the invariant component from the temporally aligned articulogram. Thus the variant component of the articulogram captures the difference between one speaker's articulation and the invariant component. We refer to the variant component as articulation style (ARTS).

In order to understand how well ARTS represents the speaker-specific characteristics, we use ARTS for the speaker identification task. In the literature, most of the work on speaker identification is based on the acoustic speech signals. Cepstral features have been commonly used [10, 11, 12, 13], while other features such as linear prediction (LP) [14, 15, 16] and perceptual LP (PLP) [17, 16] have also been exploited including representations of speech segments by i-vectors [18, 19, 20]. James et al. in [21] use nasal phonation features for speaker identification. A variety of other acoustic features has been proposed for speaker identification including AM-FM information [22], group delay [23], prosodic information [24] and wavelets [25]. Apart from the audio based features, multimodal approaches using audio, lip and facial motion have also been exploited for speaker identification [26, 27, 28, 29, 30]. In contrast to these audio and visual features, speaker identification using ARTS would reflect the amount of speaker-specific information encoded in the speech articulation style. Interestingly, the speaker identification experiments using the rtMRI corpus reveal that the ARTS results in a better frame-level speaker identification accuracy compared to the acoustic features.

2. DATASET

For the experiments in this paper, we have used the rtMRI corpus [31] consisting of simultaneous recordings of speech and rtMRI video of the upper airways in the midsagittal plane acquired from two female (F1 and F2) and two male (M1 and M2) speakers of American English while they read the same 460 sentences used in the MOCHA-TIMIT corpus [32]. These speakers are referred to as 'sub1', 'sub2', 'sub3' and 'sub4' respectively. A rtMRI video has a frame rate of 23.18 frames/sec. Audio data is simultaneously recorded at a sampling frequency of 20 kHz inside the MRI scanner while speakers are imaged. A specially designed noise cancellation technique is used to remove the scanner noise from the recorded audio [33]. Thus, the denoised rtMRI audio is unlike a speech

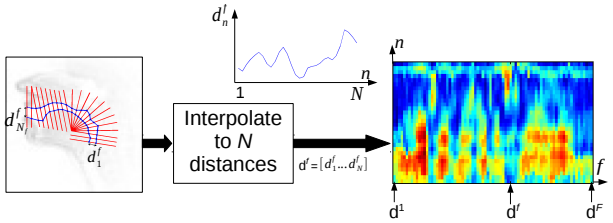

Fig. 1. Steps involved in computing an articulogram.

recording in a clean environment.

We manually annotate the upper and lower vocal tract tube boundaries on each of the upper airway images. This is done using a graphical user interface (GUI) developed in MATLAB. Annotators meticulously mark the air-tissue boundaries in the vocal tract. Since manual annotation is time-consuming and the total number of frames ($\sim 15.28\times10^4$) is large for all 460 sentences from four speakers, for the present study, we select four sentences with highest phonetic richness from each speaker. This is done using a forward sentence selection procedure to maximize the entropy of the phonetic set in the chosen sentences. These sentences are sen1 - "She always jokes about too much garlic in his food", sen2 - "There was a gigantic wasp next to Irving's big top hat", sen3 - "Laugh, dance, and sing if fortune smiles upon you" and sen4 - "Eating spinach nightly increases strength miraculously". A total of 46 different phonemes are present in these four sentences among 51 phonemes used for phonetic transcription. A total of 1518 rtMRI images corresponding to the chosen sentences are manually marked with the vocal tract tube boundaries, with 388, 349, 407 and 374 frames from four speakers respectively. The number of frames for sub1, sub2, sub3 and sub4 are (79, 73, 105, 81) for sen1, (101, 98, 111, 110) for sen2, (95, 82, 97, 87) for sen3 and (113, 96, 94, 96) for sen4 respectively.

3. ARTICULOGRAM

Articulogram is the representation of a time-varying VTTP. This forms the basis for estimation of the articulation style of a speaker. The VTTP at a given time is obtained from the corresponding rtMRI frame as illustrated in Fig. 1. The Maeda's Grid [34] (red lines on a rtMRI frame in Fig. 1) is overlaid on the manually drawn upper and lower vocal tract tube boundaries (blue curves on a rtMRI frame in Fig. 1). We have used additional two grid lines to capture the lip opening. Let the f-th ($1\leq f \leq F$) rtMRI frame has N_f grid lines from glottis to lips as shown in Fig. 1, where F is the total number of frames in an utterance. The intersection points between a grid line and the upper and lower boundaries are found and the distance (in number of pixels) between each pair of points is computed; these distances are denoted by $d^f_1, \cdots, d^f_{N_f}$. N_f distances are linearly interpolated to obtain a fixed set of N distances in each frame and, thus, an N-dimensional vector $\mathbf{d}^f$ in the f-th rtMRI frame is obtained. All these F N-dim vectors are stacked next to each other to form an articulogram as shown in Fig. 1.

4. ESTIMATION OF ARTICULATION STYLE (ARTS)

ARTS of a speaker carries speaker-specific characteristics in the articulogram. Suppose $\mathcal{S}_k$ be the articulogram of the k-th ($1 \leq k \leq K$) speaker where $\tilde{\mathcal{S}}_k(i,j) = d^j_i$, where $1 \leq i \leq N$, $1 \leq j \leq F_k$ and K is the total number of speakers. F_k denotes the length of the utterance in number of frames. We assume that the articulogram of a speaker is the sum of a speaker-invariant component and a speaker-specific component (ARTS) that varies across speakers. We estimate ARTS from the $N \times F_k$-dimensional articulographs of a sentence spoken by K speakers in two steps: 1) temporally aligning all speakers' articulograms by using the phonetic boundaries in the utterance, 2) spatially warping the temporally aligned articulograms of different speakers to remove the speaker-invariant component as shown in Fig. 2.

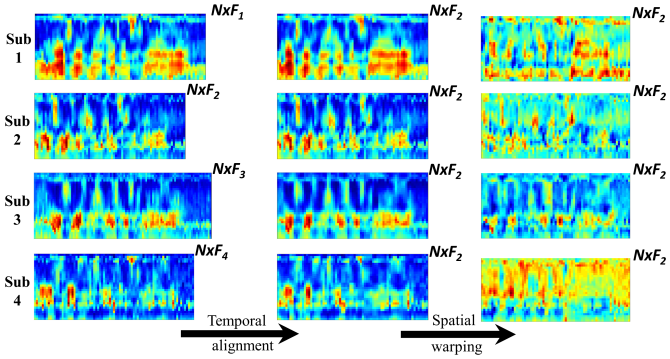

Fig. 2. Steps involved in the estimation of ARTS for sen1.

4.1. Temporal alignment

Let there be P phonemes in the phonetic transcription of a sentence and the corresponding phonetic boundaries are known in terms of the frame indices. For the k-th speaker, let the phonetic boundaries be $\{b^k_i,\ 1 \leq i \leq P+1\}$, where $b^k_1 = 1$ and $b^k_{P+1} = F_k$. Also suppose $\kappa = \arg\min_k F_k$. We then use phonetic boundaries between $\mathcal{S}_\kappa$ and $\mathcal{S}_k$, $\forall k \neq \kappa$ to time-warp $\mathcal{S}_k$ ($N \times F_k$-dimensional) to $\tilde{\mathcal{S}}_k$ ($N \times F_\kappa$-dimensional). We assume that the duration of each phoneme is minimum in the sentence with least duration, to which sentences of other subjects are warped to avoid error due to upsampling. The warping function between b^k_i and b^k_{i+1} is assumed to be linear: $h(\beta) = b^k_i + (b^k_{i+1} - b^k_i) \times (\beta - b^\kappa_i)/(b^\kappa_{i+1} - b^\kappa_i)$, $b^\kappa_i \leq \beta \leq b^\kappa_{i+1}$. Thus, $\tilde{\mathcal{S}}_k(\cdot,\beta)$=$\mathcal{S}_k(\cdot, h(\beta))$, $1 \leq \beta \leq F_\kappa$, where $\tilde{\mathcal{S}}_k(\cdot,\beta)$ denotes the β-th column of $\tilde{\mathcal{S}}_k$. The $h(\beta)$-th column of $\mathcal{S}_k$, i.e., $\mathbf{d}^{h(\beta)}$, is obtained by linearly interpolating columns of $\mathcal{S}_k$, if $h(\beta)$ is not an integer.

4.2. Spatial warping

We assume that $\tilde{\mathcal{S}}_k$ is a spatially warped version of an speaker-invariant component, ($\tilde{\mathcal{S}}$) i.e., $\tilde{\mathcal{S}}_k(l,\cdot) \approx \tilde{S}(f_k(l),\cdot)$, $1 \leq l \leq N$ where the warping (f_k) captures the differences in ARTS among speakers. In other words we assume that

$$\begin{aligned} \tilde{\mathcal{S}}_k(g_k(l),\cdot) &= \tilde{S}(l,\cdot) + E_k(l,\cdot), \forall k, 1 \leq l \leq N \\ \text{or } \tilde{\mathcal{S}}_k(g_k,\cdot) &= \tilde{S} + E_k, \forall k \end{aligned} \quad (1)$$

where $g_k(\cdot) = f^{-1}_k(\cdot)$ (assuming f_k is invertible and E_k is the model error). The goal is to find the speaker-specific warping function $g_k, \forall k$ and $\tilde{S}$ as follows:

$$\{\{g^*_k, \forall k\}, \tilde{S}^*\} = \underset{\{g_k,\forall k\},\tilde{S}}{\arg\min}\ \mathcal{J}(\{g_k, \forall k\}, \tilde{S}) \quad (2)$$

where $\mathcal{J}(\{g_k, \forall k\}, \tilde{S}) = \sum_k ||\ \tilde{\mathcal{S}}_k(g_k,\cdot) - \tilde{S}\ ||^2_{\mathbb{F}}$, where $||\cdot||_{\mathbb{F}}$ is the Frobenius Norm of a matrix. $\mathcal{J}(\cdot)$ is the objective function which is not convex in the optimization variables. A closed form solution of eqn (2) is also not possible. Interestingly, if $g_k, \forall k$ are known, $\mathcal{J}(\cdot)$ becomes a convex function of $\tilde{S}$ and it can be obtained as follows: $\tilde{S} = (\sum_{k=1}^{K} \tilde{\mathcal{S}}_k(g_k,\cdot))/K$. Similarly if $\tilde{S}$ is known, g_k can be solved by obtaining the best warping function between $\tilde{\mathcal{S}}_k$ and $\tilde{S}$, $\forall k$. Thus we follow an iterative approach in solving eqn (2) such that in each iteration the objective function value decreases. Let $\{g^i_k, \forall k\}, \tilde{S}^i$ be the solution at the i-th iteration then, $\cdots \geq \mathcal{J}(\{g^{i-1}_k, \forall k\}, \tilde{S}^{i-1}) \geq \mathcal{J}(\{g^{i-1}_k, \forall k\}, \tilde{S}^i) \geq \mathcal{J}(\{g^i_k, \forall k\}, \tilde{S}^i) \geq \cdots$. The iterative procedure is continued till the change in the objective function value is smaller than a predefined threshold (ϵ).

We assume that g_k, $\forall k$ is a piece-wise linear function with L piece-wise line segments such that $g_k(1) = 1$ and $g_k(N) = N$. Thus, f_k is also a piecewise linear function and f_k is invertible. The boundaries of the i-th piece-wise line segment are assumed to have integer co-ordinates: (m_k^i, n_k^i) and (m_k^{i+1}, n_k^{i+1}), $1 \leq m_k^i < m_k^{i+1} \leq N$ and $1 \leq n_k^i < n_k^{i+1} \leq N$, where $g_k(m_k^i) = n_k^i$ and $g_k(m_k^{i+1}) = n_k^{i+1}$. Then

$$g_k(m) = n_k^i + \frac{n_k^{i+1} - n_k^i}{m_k^{i+1} - m_k^i}(m - m_k^i), \quad m_k^i < m < m_k^{i+1} \tag{3}$$

$\tilde{S}_k(g_k(m), \cdot)$ is obtained by linear interpolation using $\tilde{S}(n, \cdot)$, $1 \leq n \leq N$. Thus the objective function over the *i-th* piece-wise line segment can be computed as follows

$$\delta(m_k^i, n_k^i, m_k^{i+1}, n_k^{i+1}) = \sum_{m=m_k^i}^{m_k^{i+1}} || \tilde{S}_k(g_k(m), \cdot) - \tilde{S}(m, \cdot) ||_2^2 \tag{4}$$

Note that $|| \tilde{S}_k(g_k, \cdot) - \tilde{S} ||_{\mathbb{F}}^2 = \sum_{i=1}^{L} \delta(m_k^i, n_k^i, m_k^{i+1}, n_k^{i+1})$

When $\tilde{S}$ is known, the best warping function g_k is obtained by minimizing $|| \tilde{S}_k(g_k, \cdot) - \tilde{S} ||_{\mathbb{F}}^2$ through a dynamic programming (DP) approach, the steps of which are explained in Algorithm 1

Algorithm 1 Estimation of g_k

1: **Inputs:** $\tilde{S}_k, \tilde{S}, L$ [$\tilde{S}_k$ and $\tilde{S}$ are $N \times F$ dimensional; for simplicity we drop subscript k]

2: **Initialization:** $\mathcal{D}_1(1,1)$=0,

$$\left.\begin{aligned} \mathcal{D}_2(m,n) &= \delta(1,1,m,n) \\ \zeta_2(m,n) &= (1,\ 1) \end{aligned}\right\} \begin{array}{l} 1 < m \leq (N - (L-1)), \\ 1 < n \leq (N - (L-1)) \end{array}$$

3: **Iteration:**
for $3 \leq i \leq L+1$, and $i < m, n < N - (L - (i-1))$

$$\mathcal{D}_i(m,n) = \min_{i-1 < m' < m, i-1 < n' < n} \{\mathcal{D}(m', n') + \delta(m', n', m, n)\}$$

$$\zeta_i(m,n) = \arg\min_{i-1 < m' < m, i-1 < n' < n} \{\mathcal{D}(m', n') + \delta(m', n', m, n)\}$$

4: **Backtracking:**

$$(m^{L+1}, n^{L+1}) = (N, N)$$
$$(m^i, n^i) = \zeta_{i+1}(m^{i+1}, n^{i+1}), i = L, L-1, \cdots, 1$$

5: **Output:** $g_k(\cdot)$ piece-wise line segments constructed using (m^i, n^i), $i = 1, 2, \cdots, L+1$

Once $\tilde{S}$ is estimated, the ARTS for the k-th speaker is obtained as $\tilde{S}_k - \tilde{S}$.

5. EXPERIMENTS AND RESULTS

5.1. Experimental setup

Articulograms of the four sentences spoken by all four speakers (K=4) are computed from the manually drawn upper and lower vocal tract boundaries using steps outlined in Section 3. As the number of Maeda's grid lines varies from lips to glottis for various speakers in different rtMRI frames, we linearly interpolate N_f distances to N=30 distances since $N_f \leq 30$, $\forall f$. Four articulograms corresponding to one sentence is used to estimate the ARTS and the invariant component. These are repeated for four sentences separately. Phonetic boundaries required for the temporal alignment (Section 4.1) are obtained by running a forced-alignment on the rtMRI audio using the available transcripts. For this purpose, 39-dim Mel frequency cepstral coefficients (MFCCs) are used as the acoustic feature and 3-state left-to-right phonetic hidden Markov models (HMMs) are trained separately for each speaker using the entire recordings of 460 sentences. The phonetic boundaries of all sentences are manually checked and corrected whenever required. It should be noted that there was no difference in the pronunciation among speakers for these four sentences leading to an identical phonetic transcription.

In case of the spatial alignment, the number of line segments L in the speaker-specific warping function g_k is varied as 4, 7, 10 and 13. This is done to examine if more line segments lead to a better estimate of the speaker-specific warping function. The predefined threshold (ϵ) for stopping the iterative optimization (section 4.2) is chosen as 10^{-3}. In the first iteration, we initialize the warping function by a linear function, i.e., $g_k(m) = m,\ 1 \leq m \leq N$.

We use the estimated ARTS for speaker identification, in which each column of the ARTS matrix is used as the feature vector. The speaker identification is done using the four speakers in the rtMRI corpus in a four-fold cross-validation setup. In every fold, one sentence from all four speakers are used as the test set and the remaining three are used for training. This is repeated four times. A four-class support vector machine (SVM) with (Gaussian) Radial Basis Function (RBF) with $\gamma = 0.1$ as the kernel is used for the speaker identification experiment. We also use VTTP for speaker identification. This is done to check the potential of the speaker-specific ARTS feature over the VTTP feature, which contains both the variant and invariant components. We also use acoustic features, MFCCs, for the speaker identification task to examine the difference between the performances of the acoustic and articulatory features. Since it is known that the difference in acoustics of different speakers primarily results from the difference in their vocal tract length (VTL), we directly estimate the VTL from the rtMRI images in each frame and use VTL for speaker identification. VTL is estimated by finding the midpoints between two intersection points in each of the Maeda's grid lines and then summing all inter-midpoint distances from the lips to the glottis. The average VTLs for sub1, sub2, sub3 and sub4 are 11.38 cm, 13.76 cm, 14.25 cm and 13.32 cm respectively. Pitch estimated from the acoustic speech signal is also used as a feature for speaker identification. Since pitch is an acoustic feature and varies across speakers, pitch also provides cues for identifying a speaker [24].

Given a test sentence, a frame-wise four-class classification is performed and a speaker is identified by using a majority rule, i.e., the speaker for which most of the frames got classified. In addition to the accuracy of identifying a speaker from a sentence, we also report the percentage of frames in a test sentence correctly classified. This percentage reflects the robustness of the feature for the speaker identification task.

5.2. Results and discussion

Fig. 3 shows the articulogram of each speaker, the estimated ARTS ($\tilde{S}_k - \tilde{S}$) and the invariant ($\tilde{S}$) component including the speaker-specific warping functions (g_k) for L=4 and sen2. From the figure, it appears that there is similarity among the articulograms $\tilde{S}_k$, $\forall k$ which is due to the phonetic content of sen2. In spite of the gross similarity, there are speaker-specific characteristics in each articulogram. These are captured in the estimated ARTS, shown in the right column (Fig. 3). Speaker-specific articulation styles are also captured in the estimated warping functions. Different warping functions indicate how differently an individual speaker articulates due to his/her morphological variations.

In order to highlight this speaker-specific characteristics in VTTP, we consider a frame corresponding to the phoneme /z/ (from the word 'was' in sen2) from the time aligned articulograms and the respective variant and invariant components as shown in Fig. 4. It is clear that although all speakers place the tongue tip behind the

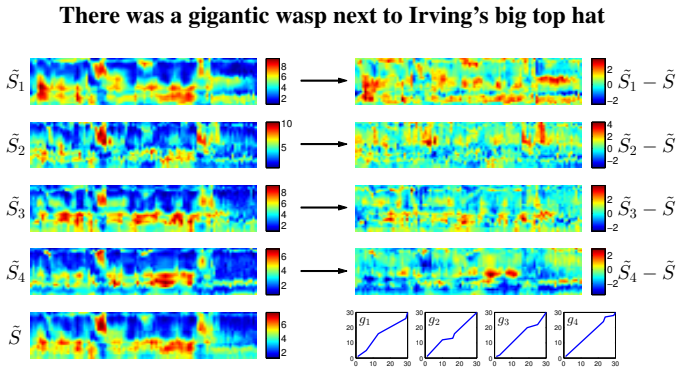

Fig. 3. Estimated ARTS for sen2 - first four rows in the left column show the individual articulogram while the fifth one is the invariant component. First four rows on the right column shows the estimated ARTS while the fifth row shows the estimated warping functions.

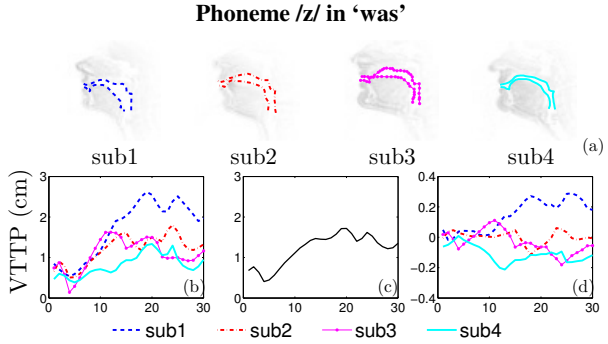

Fig. 4. VTTP for phoneme /z/ (in word 'was' in sen2) from four different speakers: (a) the vocal tract shapes on rtMRI frames. The bottom row shows the profile (corresponding to /z/ phoneme) from (b) $\tilde{S}_k$, (c) $\tilde{S}$ and (d) $\tilde{S}_k - \tilde{S}$.

upper front teeth for producing /z/, the tongue shapes are different for different speakers resulting in different ARTS profiles, which are shown in Fig. 4(d). For example, it is clear that sub3 and sub4 have in general lower vocal tract opening compared to sub1 and sub2. sub1 has the highest vocal tract tube opening. These are the features which could be useful for identifying a speaker.

When the warping functions are approximated with an increasing number (L) of piece-wise line segments, the optimized objective function value ($\mathcal{J}$ in eqn (2)) decreases (Fig. 5(a)) indicating that a better approximation of the warping function leads to a lower mismatch between the warped invariant component and the individual speaker's articulogram in terms of the Frobenius norm. However increasing L does not always increase the frame level speaker identification accuracy as shown in Fig. 5(b). The frame level accuracy averaged across four sentences (Fig. 5(b)) shows that the highest frame level accuracy is obtained using L=4. The difference in the trend of the objective function value and the identification accuracy with increasing L suggests that a lower value of the objective function does not imply a lower identification value as they are two different measures. As L=4 yields the highest identification accuracy, we choose L=4 for the speaker identification experiment.

Speaker identification using five different features VTTP, ARTS, Pitch, MFCC and VTL results in 100% sentence level identification accuracy across all folds. However, it does not mean that all frames of the test utterance are identified correctly since the sentence level accuracy is determined by a majority rule. Hence, in Table 1, we report the percentage of frames correctly classified in an utterance

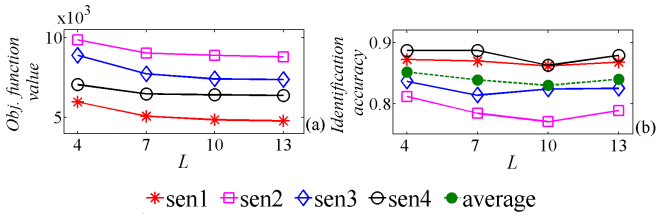

Fig. 5. The change in (a) the objective function value and (b) the frame level identification accuracy with the number of line segments (L) in the warping function.

Features	Sub 1	Sub 2	Sub 3	Sub 4
VTTP	0.85(0.01)	0.35(0.05)	0.62(0.21)	0.4(0.14)
ARTS	**0.94(0.04)**	**0.75(0.11)**	0.75(0.05)	**0.97(0.03)**
Pitch	0.59(0.08)	0.29(0.03)	0.64(0.04)	0.92(0.04)
MFCC	0.80(0.13)	0.69(0.09)	**0.8(0.06)**	0.74(0.08)
VTL	0.86(0.26)	0.59(0.32)	0.77(0.23)	0.51(0.24)

Table 1. Frame level speaker identification accuracy averaged across all folds for each of four speakers. Entries in the brackets indicate the SD. Bold entry in each column indicates the best performing feature for each speaker.

of a test speaker. Each column in the table corresponds to one test speaker and the entries in the table are the frame level accuracies averaged (with standard deviation (SD)) across all folds. It is clear that except sub3, ARTS achieves the highest frame level identification accuracy for all other speakers. ARTS performs consistently better than VTTP (by 29.89% absolute averaged across all speakers) indicating that removing the invariant component enhances the representation capability of the speech articulation for identifying a speaker. Superior performance (by 9.41% absolute averaged across all speakers) of ARTS over acoustic features suggests that estimated speaker-specific articulation characteristics from the directly measured articulatory movement is better representative of a speaker compared to the MFCC. This could be also due to poor audio quality from the denoised rtMRI recording. Lower accuracy using VTL suggests that more information about a speaker is present in his/her articulation than in the VTL.

6. CONCLUSIONS

We propose an automatic algorithm to estimate the invariant component of articulation when a sentence is spoken by multiple speakers. The corresponding variant component is found to carry more speaker-specific information compared to the speaker's articulation. When used for the speaker identification, the variant component is found to be better than the acoustic features typically used for speaker identification. These results indicate that reliable speaker-specific information is present in the speech articulation style, which can be estimated by subtracting the invariant component from the speech articulation. Estimation of ARTS in this work requires each sentence to be spoken by all subjects in the corpus and, hence, can not be generalized to apply to an unseen test data. However, this could be avoided by estimating warping functions for different phonemes specific to each speaker. Similarly, articulogram requires the knowledge of the VT morphology which may not be directly available causing the implementation of the proposed speaker identification to be impractical; however, indirect methods of estimating VTTP (such as from the speech signal [35]) could be used. The proposed approach of estimating speaker-specific ARTS could also be used to analyze articulatory settings [36].

7. REFERENCES

[1] Vincent L Gracco and James H Abbs, "Variant and invariant characteristics of speech movements," *Experimental Brain Research*, vol. 65, no. 1, pp. 156–166, 1986.

[2] Joseph S Perkell and Dennis H Klatt, *Invariance and variability in speech processes*, Psychology Press, 2014.

[3] Leigh Lisker, "The pursuit of invariance in speech signals," *The Journal of the Acoustical Society of America*, vol. 77, no. 3, pp. 1199–1202, 1985.

[4] Steven Greenberg and Brian E D Kingsbury, "The modulation spectrogram: In pursuit of an invariant representation of speech," in *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1997, vol. 3, pp. 1647–1650.

[5] Kenneth N Stevens, "Invariance and variability in speech: Interpreting acoustic evidence,mit," *From sound to sense: 50+ years of discoveries in speech communication*, pp. 77–85, 2004.

[6] Anja Geumann, "Invariance and variability in articulation and acoustics of natural perturbed speech," *Diss. Inst. für Phonetik und Sprachliche Kommunikation der Univ. München*, 2001.

[7] Julle Carson-Berndsen, "Phonological processing of speech variants," in *Proceedings of the 13th conference on Computational linguistics*. Association for Computational Linguistics, 1990, vol. 3, pp. 21–24.

[8] Mark K Tiede, Vincent L Gracco, Douglas M Shiller, Carol Espy-Wilson, and Suzanne E Boyce, "Perturbed palatal shape and north american english /r/ production," *The Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2568–2569, 2005.

[9] Adam Lammert, Michael Proctor, and Shrikanth Narayanan, "Morphological variation in the adult hard palate and posterior pharyngeal wall," *The Journal of Speech, Language, and Hearing Research*, vol. 56, no. 2, pp. 521–530, 2013.

[10] Sheng-Yu Sun, C-L Tseng, Y H Chen, S C Chuang, and H C Fu, "Cluster-based support vector machines in text-independent speaker identification," in *International Joint Conference on Neural Networks*. IEEE, 2004, vol. 1.

[11] Daniel J Mashao and Marshalleno Skosan, "Combining classifier decisions for robust speaker identification," *Pattern Recognition*, vol. 39, no. 1, pp. 147–155, 2006.

[12] Ali Zulfiqar, Aslam Muhammad, and A M Martinez Enriquez, "A speaker identification system using MFCC features with VQ technique," in *Third International Symposium on Intelligent Information Technology Application*. IEEE, 2009, vol. 3, pp. 115–118.

[13] Douglas A Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech communication*, vol. 17, no. 1, pp. 91–108, 1995.

[14] Bishnu S Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *The Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, 1974.

[15] Mihailo S Zilovic, Ravi P Ramachandran, and Richard J Mammone, "Speaker identification based on the use of robust cepstral features obtained from pole-zero transfer functions," *Transactions on Speech and Audio Processing*, vol. 6, no. 3, pp. 260–267, 1998.

[16] Douglas A Reynolds, "Experimental evaluation of features for robust speaker identification," *Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 639–643, 1994.

[17] Jiahong Yuan and Mark Liberman, "Speaker identification on the SCOTUS corpus," *The Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 5687–5690, 2008.

[18] Taufiq Hasan, Rahim Saeidi, John H L Hansen, and David A van Leeuwen, "Duration mismatch compensation for i-vector based speaker recognition systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7663–7667.

[19] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions onAudio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[20] Sandro Cumani, Niko Brummer, Lukas Burget, and Pietro Laface, "Fast discriminative speaker verification in the i-vector space," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2011, pp. 4852–4855.

[21] James W Glenn and Norbert Kleiner, "Speaker identification based on nasal phonation," *The Journal of the Acoustical Society of America*, vol. 43, no. 2, pp. 368–372, 1968.

[22] C R Jankowski Jr, T F Quatieri, and D A Reynolds, "Measuring fine structure in speech: Application to speaker identification," in *International Conference on Acoustics, Speech, and Signal Processing,ICASSP-95*. IEEE, 1995, vol. 1, pp. 325–328.

[23] Rajesh M Hegde, Hema A Murthy, and G V Ramana Rao, "Application of the modified group delay function to speaker identification and discrimination," in *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2004, vol. 1, pp. 517–520.

[24] Michael J Carey, Eluned S Parris, Harvey Lloyd-Thomas, and Stephen Bennett, "Robust prosodic features for speaker identification," in *International Conference on Spoken Language Processing*. IEEE, 1996, vol. 3, pp. 1800–1803.

[25] Jian-Da Wu and Bing-Fu Lin, "Speaker identification using discrete wavelet packet transform technique with irregular decomposition," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3136–3143, 2009.

[26] Juergen Luettin, Neil A Thacker, and Steve W Beet, "Speaker identification by lipreading," in *International Conference on Spoken Language Processing*. IEEE, 1996, vol. 1, pp. 62–65.

[27] Tim Wark and Sridha Sridharan, "Adaptive fusion of speech and lip information for robust speaker identification," *Digital Signal Processing*, vol. 11, no. 3, pp. 169–186, 2001.

[28] H Ertan Cetingul, Yücel Yemez, Engin Erzin, and A Murat Tekalp, "Discriminative analysis of lip motion features for speaker identification and speech-reading," *IEEE Transactions on Image Processing*, vol. 15, no. 10, pp. 2879–2891, 2006.

[29] T Wark, S Sridharan, and V Chandran, "An approach to statistical lip modelling for speaker identification via chromatic feature extraction," in *Fourteenth International Conference on Pattern Recognition*. IEEE, 1998, vol. 1, pp. 123–125.

[30] Alper Kanak, Engin Erzin, Yücel Yemez, and A Murat Tekalp, "Joint audio-video processing for biometric speaker identification," in *International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2003, vol. 2, pp. 377–380.

[31] Shrikanth Narayanan, Erik Bresch, Prasanta Kumar Ghosh, Louis Goldstein, Athanasios Katsamanis, Yoon Kim, Adam C Lammert, Michael I Proctor, Vikram Ramanarayanan, and Yinghua Zhu, "A multimodal real-time MRI articulatory corpus for speech research," in *INTERSPEECH*, 2011, pp. 837–840.

[32] W J Hardcastle Alan A Wrench, "A multichannel articulatory database and its application for automatic speech recognition," in *5th Seminar of Speech Production*, 2000, pp. 305–308.

[33] Erik Bresch, Jon Nielsen, Krishna Nayak, and Shrikanth Narayanan, "Synchronized and noise-robust audio recordings during realtime magnetic resonance imaging scans," *The Journal of the Acoustical Society of America*, vol. 120, no. 4, pp. 1791–1794, 2006.

[34] S Maeda, "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal tract shapes using an articulatory model," *Speech production and speech modelling*, pp. 131–149, 1990.

[35] Hisashi Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *Transactions on Audio and Electroacoustics*, vol. 21, no. 5, pp. 417–427, 1973.

[36] John D M Laver, "Voice quality and indexical information," *International Journal of Language & Communication Disorders*, vol. 3, no. 1, pp. 43–54, 1968.