Discrimination Between Native and Non-Native Speech Using Visual Features Only

Christos Georgakis, *Student Member, IEEE*, Stavros Petridis, *Member, IEEE*, and Maja Pantic, *Fellow, IEEE*

***Abstract*—Accent is a soft biometric trait that can be inferred from pronunciation and articulation patterns characterizing the speaking style of an individual. Past research has addressed the task of classifying accent, as belonging to a native language speaker or a foreign language speaker, by means of the audio modality only. However, features extracted from the visual stream of speech have been successfully used to extend or substitute audio-only approaches that target speech or language recognition. Motivated by these findings, we investigate to what extent temporal visual speech dynamics attributed to accent can be modeled and identified when the audio stream is missing or noisy, and the speech content is unknown. We present here a fully automated approach to discriminating native from non-native English speech, based exclusively on visual cues. A systematic evaluation of various appearance and shape features for the target problem is conducted, with the former consistently yielding superior performance. Subject-independent cross-validation experiments are conducted on mobile phone recordings of continuous speech and isolated word utterances spoken by 56 subjects from the challenging MOBIO database. High performance is achieved on a text-dependent (TD) protocol, with the best score of 76.5% yielded by fusion of five hidden Markov models trained on appearance features. Our framework is also efficient even when tested on examples of speech unseen in the training phase, although performing less accurately compared to the TD case.**

***Index Terms*—Foreign accent detection, non-native speech, visual accent classification, visual speech processing.**

I. Introduction

ACCENT manifests itself in speech through a set of pronunciation, articulation, intonation, lexical stress, and rhythmic patterns that are common in the speaking style of individuals belonging to a particular language group. Accent classification has attracted growing interest in the speech processing and human language technology research community over the past two decades [1]–[7].

Unveiling *a priori* whether a speech episode is spoken by a second-language (L2) speaker or by a mother tongue (L1) speaker has emerged as a need to overcome limitations posed by accent-sensitive speech recognizers [8]. Language-specific speaking style is intrinsically related to physiological phenomena in the speech production system, such as vocal tract functions and articulatory movements, that are developed while acquiring language skills at a young age [9]. There is evidence that these traits are transferred to any second language learnt [1], [9], [10]. Specifically, L2 speakers "borrow" phonemes from their mother tongue to replace unfamiliar phonemes that they encounter in the foreign language. Such misarticulations and varying pronunciations can lead to substantial divergence from the actual phonetic configuration of the second language, thus resulting in higher word error rates in accented speech recognition. Hence, identifying the accent, and, at a second step, adapting the acoustic, pronunciation, and language models can markedly enhance speech recognition [1], [2].

The knowledge of the accent of a speaker is not useful only as a preprocessing step for speech recognition. Primarily, accent is an important soft biometric trait of an individual [11], such as age and gender, and, as such, can serve for verification purposes [6], [12]. Accent analysis is also essential for applications such as pronunciation modeling [13] and computer-assisted L2 learning [14].

Most related work has viewed accent identification as a multiclass classification problem that aims to assign a speech sample to either the native accent of the target language or one of separately modeled foreign-language accents [1], [2], [15]. These approaches mainly use hidden Markov models (HMMs) [16], on the phoneme or word level, trained on acoustic features, such as prosodic and cepstral features. Alternative methodologies borrow inspiration from language identification (LID) [3] and usually rely on language and phonotactics modeling, with Gaussian mixture models (GMMs) as their basic tool. More recently, discrimination between native and non-native speech has been targeted by means of binary classification frameworks [4]–[6]. These works mainly rely on cepstral, prosodic, speech recognition-based or *N*-gram language features, and employ support vector machines (SVMs) for classification.

All the above works on accent classification and detection have persistently ignored features derived from the visual stream. However, the beneficial role of visual information

Manuscript received May 13, 2014; revised December 12, 2014 and September 1, 2015; accepted September 29, 2015. This work was supported by the European Community Horizon 2020 [H2020/2014-2020] under Grant 645094 (SEWA: Automatic Sentiment Analysis in the Wild). The work of S. Petridis was supported by the European Community 7th Framework Programme [FP7/2007-2013] under Grant 611153 (TERESA: Telepresence Reinforcement-learning Social Agent). This paper was recommended by Associate Editor B. W. Schuller.

C. Georgakis and S. Petridis are with the Department of Computing, Imperial College London, London SW7 2AZ, U.K. (e-mail: christos.georgakis@imperial.ac.uk; stavros.petridis04@imperial.ac.uk).

M. Pantic is with the Department of Computing, Imperial College London, London SW7 2AZ, U.K., and also with Electrical Engineering, Mathematics, and Computer Science, University of Twente, Enschede 7522NB, The Netherlands (e-mail: m.pantic@imperial.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2015.2488592

to speech comprehension has been well documented [17] and experimentally validated [18]–[20]. Furthermore, recent findings indicate that human observers can actually perform language identification through the visual modality only [21]. Automated approaches for visual-only language identification have also been proposed (see [22]). Another study shows that visual identification of accent is a feasible task for human observers [23]. This indicates that visual manifestations of accent in speech suffice for a human observer to identify a speaker as native or non-native, even in the absence of the audio stream.

In this paper, our aim is to show that these accent-related speech dynamics, which are visually discernible by humans, can be efficiently modeled and identified through a fully-automated visual approach. Herein, we present this paper on the task of visual-only discrimination between native and non-native speech in English, which is targeted as a binary classification problem. Visual-only accent classification can prove useful for authentication devices to recognize impostors in the case of noisy crowded environments, where other voice-based biometric tools provide less reliable measurements. It can also be utilized to assist L2 learning in applications that would provide feedback to the learners by assessing their "level of accent" as they pronounce words or sentences in front of a camera.

The twofold contribution of this paper consists of: 1) providing the first basic study about the target problem and 2) introducing a fully-automated approach that could be used to visually discriminate native from non-native speech. The presented study includes a systematic comparative evaluation, in terms of their robustness for the problem investigated, of various appearance- and shape-based visual descriptors, which are coupled with HMMs [16]. Also, different feature normalization alternatives are examined, demonstrating the importance of tailoring the normalization step to each utilized feature.

Speaker-independent accent classification experiments are conducted first on continuous reading speech samples from the MOBIO database [24], all captured by mobile phones. In that experiment, the experimental scenario is text-dependent (TD), i.e., all speakers utter the same three-sentence paragraph. This ensures that our system recognizes distinct accent-related patterns rather than simply differentiating between mouth movements corresponding to different speech content. In a second experiment, the classifiers that have been trained and optimized on the three-sentence speech scenario are tested on visual speech examples corresponding to isolated words. Finally, we also examine the text-independent (TI) experimental scenario, i.e., testing on examples of speech content different from that "seen" in the training phase, and thus perform a fair comparison with the TD counterpart. In all experiments presented herein, multifold cross-validation is performed, so that all data are used for testing.

Our results indicate that a system based on appearance descriptors and a sequential classifier can reliably be used to discriminate native from non-native speech. Shape features yield above-chance-level performance, but they seem to be less informative for the target task, when compared to the appearance-based features. Accent-class predictions are accurate also when our systems are tested on short speech segments containing isolated words. Finally, we show that our framework can also address accent classification in the TI case, though with lower performance.

This paper is organized as follows. Section II provides an outline of previous work on audio-only accent classification and detection, along with highlights of research that has employed visual modality for speech, language, and accent recognition. Section III presents the proposed methodology, while Section IV describes the database which we use in our experiments. Section V explains the appearance and shape features used in this paper, along with the preprocessing procedure, i.e., tracking and mouth region of interest (ROI) extraction, and the normalization schemes applied. Section VI focuses on the experimental protocol and details regarding the normalization schemes, as well as the topology and parameters of the classifiers. Sections VII and VIII report and discuss the results of the TD and TI experiments, respectively. Finally, Section IX concludes this paper.

II. Previous Work

A. Audio-Only Approaches to Accent Classification

There has been considerable research targeting the acoustic characteristics of foreign accent [1], [9], [25]. Accent classification has mostly been addressed as the task of automatically assigning speech in a target language to either the native accent or a language-specific foreign accent. An established common approach in this area is to model each accent separately, based on phonology and phonotactics acoustic features, and subsequently employ a generative classifier, such as HMMs or GMMs, to produce accent-specific probabilistic scores. In [1], energy and Mel-frequency cepstral coefficients [16] are used, in conjunction with accent-specific phone- and word-level HMMs, to classify among four accents in American English (neutral, Chinese, Turkish, and German). Accent classification rate of 93% is achieved on test strings composed of seven to eight isolated words. In an additional experiment, 21 human listeners (12 native and nine non-native English speakers) were asked to characterize accent as native or non-native and also classify non-native accent (as one among the three foreign accent classes) on 48 speech samples, which had been randomly selected from a test set containing utterances of 20 words. The proposed computer algorithm is shown to provide accent classification accuracy that is higher by 8.4%, compared to the best performance among human listeners, when tested on the same evaluation set. Also, the results indicate that listeners who are native speakers of English can classify accent better than non-native speakers. Teixeira *et al.* [15] build an accent classification framework targeting six accents in English, based on competing HMM-based subnets, and they show that speech recognition benefits from the use of multiaccent training data.

To avoid the computational cost entailed by training phoneme-level HMMs for each accent, many approaches have resorted to alternative classification schemes. Deshpande *et al.* [26] extract formant frequencies only on voiced frames and use accent-specific GMMs to distinguish between native and Indian accent in American English. Angkititrakul and Hansen [10] introduce two trajectory-based models on the cepstral space, and conduct accent classification experiments on five English speaker groups. Recently, Biadsy *et al.* [27] use phone-type GMM-supervectors and an *ad hoc* SVM classifier, achieving high performance on a series of accent classification experiments on continuous speech. Zhang and Qin [7] build upon the above framework to develop a semi-supervised accent detector to distinguish among Native, Southern, and Hispanic American English.

Other works address accent classification by borrowing techniques commonly used in the similar field of LID [3], [28]. Torres-Carrasquillo *et al.* [3] approach dialect identification by means of GMMs trained on shifted-delta-cepstral features. In [28], experiments are conducted on 23-way classification of non-native English accents. Their framework, which is based on heteroscedastic linear discriminant analysis (HLDA), maximum mutual information (MMI) training, and Gaussian tokenization is shown to improve the baseline GMM LID-like model.

As opposed to this multiclass classification formulation, accent identification has been recently addressed as a binary classification problem, usually termed accent detection. In that scenario, the goal is to determine the nativeness or non-nativeness of speech. An important limitation to be battled within this scope is the increased variability that speakers belonging to different language backgrounds inflict on the non-native accent class. Shriberg *et al.* [4] perform accent detection experiments, using features based on maximum likelihood linear regression (MLLR) adaptation transforms, phone *N*-gram, prosodic, and word *N*-gram features, while relying on linear SVMs for classification. Omar and Pelecanos [6] employ GMM supervectors and feed them into SVM classifiers for accent detection. The authors report a relative improvement of 23.4% over previously published results [4] on the English Fisher database [29]. Tan *et al.* [5] utilize four subsystems, two phonetic-based, and the other two nonphonetic-based, built on HMMs and GMMs, respectively, and fusion of them in order to address discrimination between native and non-native American English. Recently, Sam *et al.* [30] use modulation spectrum features and GMM classifiers for non-native accent detection in French.

B. Visual Speech and Accent

All the aforementioned research on accent classification and detection has relied solely on auditory information, thus disregarding the visual modality. Yet the contribution of the visual information to speech comprehension has been well investigated [17]–[19]. Experimental findings show that visual speech features significantly boost the performance of automatic speech recognizers, compared to audio-only approaches, especially when the auditory stream is noisy [18], [31].

Experiments with human observers of visual speech have shown that language identification is feasible through the visual modality only. Specifically, Ronquest *et al.* [21] carry out experiments in which participants are asked to observe, without listening, video clips that show a male or a female speaker talking in English or Spanish. Both speakers appearing in the videos are bilingual in English and Spanish. The task for the observers is to identify from the visual speech frames that contain the whole face of the speaker which language is spoken in each video clip. The authors show through a series of experimental scenarios that observers perform much higher than chance level on visual discrimination between English and Spanish, even when the visual speech segments are presented to them in reverse temporal order. Recently, Newman and Cox [22] present an automated approach for both speaker-dependent and speaker-independent visual-only language identification, based on features that capture phonology and phonotactics characteristics of visual speech. They use active shape models and active appearance models [32] for visual feature extraction, which is carried out frame-wise, and then feature vectors are “tokenized” in visually transcribed phonemes. Bi-gram language models, one for English data and the other for Arabic, are then constructed and SVMs are used for classification. Their results show that visual features can alone be discriminative for the target problem. This paper also suggests that the attained accuracy highly depends on the visually described phoneme recognizers. The data used in this paper have not yet been made publicly available, which made our efforts to reconfirm the findings of this paper impossible. This further motivated our choice to use publicly available data in this paper and make our results repeatable by other researchers in the field.

Recent evidence suggests that speaker variability in terms of accent, even in the form of regional accent, can largely affect speechreading performance. Ellis *et al.* [33] perform speechreading experiments in British English on deaf people, which reveal that visible regional accent is the most important cause resulting in degraded performance. Irwin *et al.* [34] conduct a systematic analysis of British regional accent. In one of their experiments, they ask participants, half of them with Nottingham accent and the other half with Glaswegian accent, to speechread sentences uttered by speakers from both accent groups. Both groups of listeners found it more difficult to visually comprehend speakers with Glaswegian accent. In another experiment, they report that in scenarios where both speaker and observer belonged to the same accent class, speechreading performance was higher.

All the above findings support the assumption that there are highly informative visual cues related to manifestations of different articulation and pronunciation patterns, that could alone serve for accent identification. Furthermore, visual information from the speaker’s mouth region can disclose physiological phenomena, such as range and velocity of lip movements and place of articulation, that are intrinsically related to lexical stress and other accent traits [23].

To the best of our knowledge, there has been no previous extensive study of the task of automatic discrimination between native and non-native speech, based on the visual modality only. The study in [23] reports human observer experiments on visual discrimination of accents, in particular, English versus French. Specifically, 30 participants, all being native English speakers, were asked to identify the accent used by a bilingual English male speaker

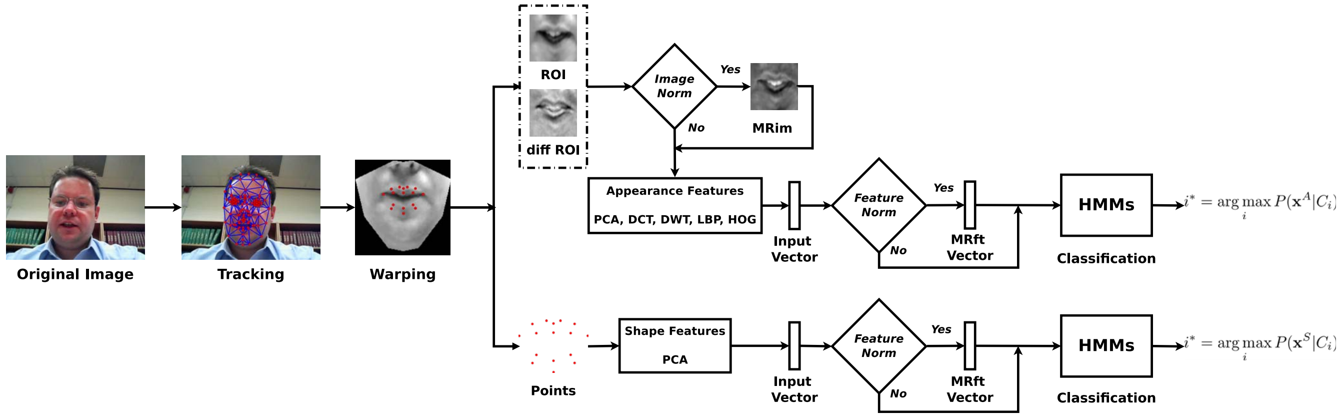

Fig. 1. Illustration of the discrete stages of the proposed framework for visual-only discrimination between native and non-native speech. The upper branch and the lower branch refer to the systems based on appearance- and shape-based features, respectively. The dashed–dotted line denotes units that involve more than one alternatives, i.e., original or difference ROIs.

(native speaker of English and French) in episodes of visual speech. Four experimental scenarios included English speech with an English accent, English speech with a French accent, and viceversa. Note that in none of the four conditions, the participants were aware of the language spoken. Observers were able to identify the accent of the speaker at much higher performance levels than chance level. Better performance was achieved in cases where language and accent were matching, while the authors report a significant accuracy decline in cases where incongruent stimuli were presented.

III. Overview of Proposed Method

The proposed system for visual-only discrimination between native and non-native speech, graphically illustrated in Fig. 1, consists of the following steps.

1) Facial point tracking, image registration, and ROI extraction.
2) Appearance and shape features computation.
3) Classification.

First, facial characteristic landmarks on the speaker's face are tracked throughout each video of an utterance, using the appearance-based tracker [35]. Only the points corresponding to the lower face region are used in further processing. Apart from the coordinates of the actually tracked points, the tracker estimates their position on a pose-free coordinate system. These pose-free points, after undergoing a global alignment, serve to register the appearance of the lower face image (i.e., remove variations due to head movements). This texture warping process yields frontal face images, from which the pixel intensities lying in a rectangular region around the lips are used as mouth ROI.

In the next stage, we use the mouth ROIs and 16 registered mouth points localized in each frame to extract appearance and shape features, respectively. We investigate five different appearance-based descriptors: 1) principal component analysis (PCA) [36]; 2) 2-D discrete cosine transform (DCT) [36]; 3) discrete wavelet transform (DWT) [36]; 4) local binary patterns (LBPs) [37]; and 5) histograms of oriented gradients (HOGs) [38], all calculated on pixel intensities. We choose to investigate these appearance descriptors because they have proven to be highly descriptive and informative of facial expression changes by numerous studies on automatic facial expression recognition (see [39]–[41]). Shape features are also examined, based on a set of 16 registered points on the inner and outer lip contour (see Fig. 1), which are yielded by the previous image registration phase. PCA applied on points is used to capture the top high-variance modes and thus project points onto principal mouth configurations.

For classification, we use HMMs [16]. HMMs have been shown highly suitable for temporal modeling of speech [42] and classification of facial events and patterns including visual speech recognition, facial muscle action recognition, emotion recognition, and social signal recognition (see [31], [43]–[46]). In particular, we forward the static features to the classifier, which undertakes the task of modeling the accent-specific visual speech dynamics. Two continuous-density models, one for each accent class, are utilized to capture the evolution of visual speech along the entire utterance, rather than on a subword or word level. This entails no reliance whatsoever of our framework on speech transcriptions. Finally, each test feature vector is assigned to the class-model yielding the highest likelihood (see Section VI-C).

IV. Data Set

The growing interest of the speech processing community in the effect of accented speech on the performance of speech and speaker recognition systems has led to the development of various data sets that contain also speech from non-native speakers. Raab *et al.* [47] list and succinctly review non-native speech databases. They report that the anxiety can be heard in speech recordings of participants who are asked to speak in a language other than their mother tongue. Furthermore, they note that the variability induced in the non-native class by different levels of L2 proficiency should be also considered.

In this paper, we are aware of these factors. However, the absence of previous work on the target task leaves us with no precedent published results on existing databases. This renders impossible any comparison with respect to a reference protocol. On the other hand, this gives us the opportunity to

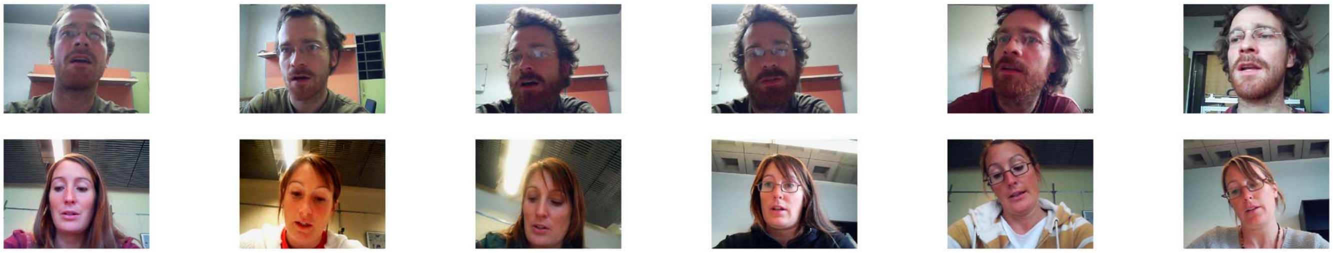

Fig. 2. Example frames, one for each of the six sessions of phase I, for two individuals from the MOBIO database (top row: subject m431 and bottom row: subject f010). Note the high intraspeaker variability across sessions in terms of appearance, clothing, head pose, illumination, and background.

test our system on a recently published challenging data set that satisfies our experimental setting choices. The proposed framework is evaluated in a subject-independent fashion on native and non-native speech episodes of English speech from the MOBIO database [24].

The MOBIO database was recorded over one and a half years at six sites in five countries, including only English speech produced by both native and non-native speakers. The recordings are temporally distributed into two phases, each comprises six sessions. The full database contains 61 h of audio-visual data corresponding to a total of 150 participants. Each session of phase I includes four different scenarios, that is, five short response questions, one predefined text, ten free speech questions, and five short free speech questions. Phase II does not include the fourth scenario, while participants are prompted with five free speech questions, instead of ten. Audiovisual data are almost exclusively captured by mobile devices. Only the very first out of the 12 sessions is captured also by laptop computers in a separate recording. Considering that the acquisition device is handheld, high variability in pose, and illumination characterizes the data samples, even within the same recording. As a matter of fact, the visual stream includes frames with high variability in head pose, and non-uniform illumination in the region of the face (partial occlusions due to varying shadowing). An additional challenge is posed by appearance fluctuations of the same subject across different sessions (e.g., haircut, glasses, and beard). Finally, visual noise caused by different background and recording conditions varies significantly. Characteristic frames from two individuals can be seen in Fig. 2.

In this paper, we choose to establish a TD scenario for our baseline experiments. Hence, we use only those visual speech samples from the six sessions of phase I in which the same three-sentence text is read. All of the data used in this paper were captured by mobile phones, as we do not include in this paper the laptop recordings. By using different sentences of the same paragraph for our training, validation, and test set, we also evaluate our framework on the TI scenario (see Section VIII). As no information regarding the nationality and mother tongue of the speakers is publicly available, the nativeness/non-nativeness of speakers was examined by four Ph.D. students of our group, all being native English speakers. After watching for each of the 150 speakers two audiovisual samples containing the entire utterances of the three-sentence text, they identified in unison 28 speakers from MOBIO database as being non-native. No other information was used for the accent annotation. With the goal of forming a balanced set with respect to accent class, we randomly selected 28 speakers out of those annotated as native. The only bias involved in this selection was posed by our choice to keep our data balanced over gender as well. Out of 336 samples, corresponding to the six phase I sessions for the 56 subjects used, 64 samples for which ROIs were erroneously extracted, e.g., due to erratic point tracking or inaccurate warping (see Section V-A1), were excluded from the experiments. Therefore, totally 272 samples from 56 subjects—135 samples belonging to 28 native English speakers and 137 to 28 non-native English speakers—are used in the experiments presented in this paper. The paragraph read in all such recordings is the following.

"I have signed the MOBIO consent form and I understand that my biometric data are being captured for a database that might be made publicly available for research purposes. I understand that I am solely responsible for the content of my states and my behavior. I will ensure that when answering a question I do not provide any personal information in response to any question."

Long silence segments in the beginning and the end of the samples are removed by applying the statistical model-based voice activity detector in [48] on the corresponding audio stream, setting the threshold for the likelihood ratio test to 95%. The mean and standard deviation of duration over all 272 samples used is 22.5 and 3.4 s, respectively. The video stream, which is provided in variable frame rate encoding, is converted to a sequence of still frames, corresponding to approximately 15 frames/s, for almost all samples. The frame size for all the samples is 640 × 480 pixels.

V. Feature Extraction

This section describes the appearance and shape features extracted on the visual speech samples from MOBIO database. Aside from the description of each feature and the choice of parameters, the rationale behind using each of them is given.

A. Appearance Features

Various appearance descriptors are examined in this paper, to assess their ability to capture discriminative visual accent-sensitive patterns. Feature extraction is based on either a global transformation or local operations, calculated directly upon pixel intensities.

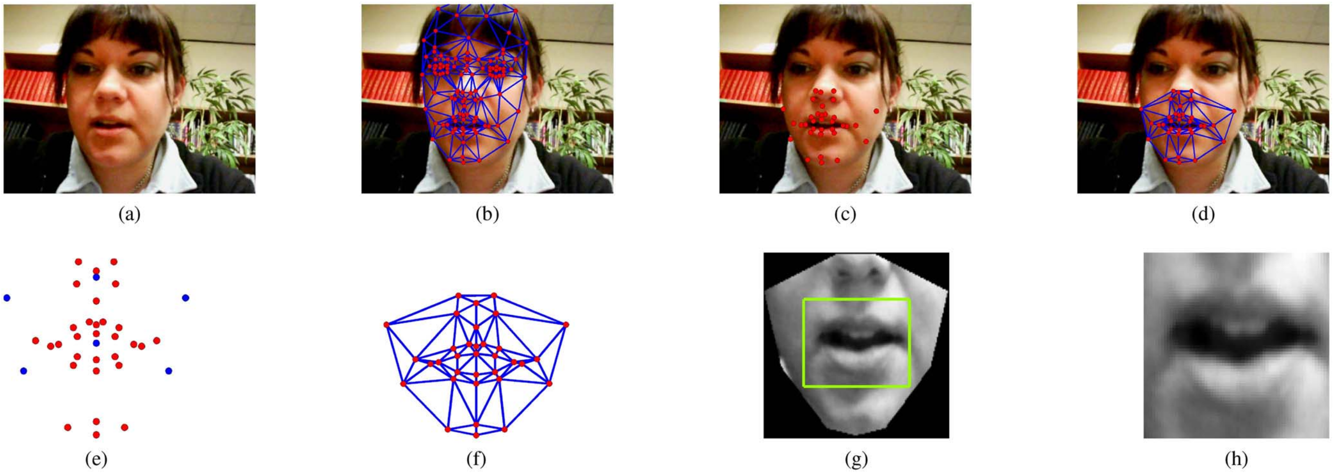

Fig. 3. Instances of the ROI extraction procedure, illustrated on a native speech example frame from the MOBIO database (subject f118). (a) Original frame. (b) Candide wireframe fitted on all 113 facial points produced by the tracker. (c) 34 lower face points. (d) Triangulated mesh of the lower face points, superimposed on the input face image. (e) Pose-normalized lower face points (as yielded by the tracker), including the six base points used for alignment (shown in blue). (f) Triangulated mesh of the registered points. (g) Warped frontal lower face and mouth bounding box. (h) Final mouth ROI, which is obtained after rescaling the initial ROI to 64 × 64 pixels and, subsequently, smoothing it by a spatial Gaussian filter of dimension 3 × 3 pixels.

First, appearance in the lower face has to be registered, i.e., variations owing to head movements have to be removed and face images have to be globally aligned. The mouth ROI extraction procedure is described in detail below.

1) Mouth Region of Interest Extraction: We initially track 113 characteristic points on the face [Fig. 3(b)], using the appearance-based tracker [35]. A Candide wireframe [Fig. 3(b)] is manually fitted on the face at the first frame, thus initializing the position of the points. These are then automatically tracked for the remaining frames. Out of these points, that are estimated for the N frames of a video sample, we only keep the 2-D spatial coordinates of 34 points that correspond to the lower face region [Fig. 3(c)]. The collection of these coordinates for all N frames form the set of actually tracked points $\mathcal{T} = \{T_1, T_2, \ldots, T_N\}$ of dimension $N \times 34 \times 2$. Aside from this set, we use the set $\mathcal{T}^{\text{norm}} = \{T_1^{\text{norm}}, T_2^{\text{norm}}, \ldots, T_N^{\text{norm}}\}$, containing the coordinates of the pose-free version of the 34 points for all frames [Fig. 3(e)], again provided as a part of the tracker's output. Six base points [see blue points in Fig. 3(e)] that are relatively invariant to facial deformations—the two "ear-level" points on the face boundary, the two points where the jaws are attached to one another, the tip of the nose, and the center of the mouth (calculated based on the location of 16 points representing the lips contour)—serve to register the face region, that is, align the set $\mathcal{T}^{\text{norm}}$ of pose-free points. The registration is performed by means of a similarity transformation (translation, scaling, and rotation), defined by the values of these six base points in a reference frame. The reference frame contains a rescaled version of the mean shape computed over all pose-free shapes. The rescaling is such that the interocular distance matches the mean interocular distance computed over all tracked shapes. The similarity transformation is applied to all 34 pose-free points to yield the registered points set $\mathcal{T}^{\text{reg}} = \{T_1^{\text{reg}}, T_2^{\text{reg}}, \ldots, T_N^{\text{reg}}\}$, which is employed in the next stage.

Texture warping is then performed to acquire lower face images in frontal view. First, for each frame, two 2-D meshes (one for actually tracked points $\mathcal{T}$ [Fig. 3(d)] and one for the registered points $\mathcal{T}^{\text{reg}}$ [Fig. 3(f)]) are triangulated. A piecewise affine warp is defined between the corresponding triangles in the meshes. This warp is then used to map the texture of the mesh in the input image [Fig. 3(d)] onto the registered mesh [Fig. 3(f)]. Finally, all warped frontal lower faces are resampled to dimension 200 × 200 pixels [Fig. 3(g)], and the registered points are accordingly rescaled. These are subsequently employed to calculate shape features (see Section V-B).

The mouth ROI is extracted from the warped frontal image as a 94 × 114 pixels bounding box containing the pixel intensities around the mouth [Fig. 3(g)]. This dimension was set on the basis of the 99.99%-percentile of the maximum horizontal and vertical distances of the outer lip contours appearing in half of the speech examples (the maximum dimension was 104 × 118 pixels, and is not used because it corresponds to inaccurate tracking of the outer lip points on a small portion of frames). Finally, all mouth ROIs are downsampled to 64 × 64 pixels and, subsequently, smoothed by a 3 × 3 pixels gaussian filter [Fig. 3(h)].

In order to incorporate dynamic information related to subtle short-term articulation cues, prior to feature extraction, we also examine the technique of ROI frame differencing, similarly to [49]. In particular, the mouth ROI I_n at each frame n is replaced by the difference in pixel intensities between I_n and the ROI at the previous frame, I_{n-1}

$$I_n^{\text{diff}} = I_n - I_{n-1}. \tag{1}$$

These new mouth ROIs for all N frames of the utterance form the set of difference ROIs $\mathcal{I}^{\text{diff}} = \{I_1^{\text{diff}}, I_2^{\text{diff}}, \ldots, I_N^{\text{diff}}\}$. In this way, the dynamics of the recorded visual articulations are captured (i.e., changes in the skin appearance around the mouth) and redundant speaker-specific information that is static in all frames is removed. We evaluate our approach both on the original mouth ROIs and on the difference ROIs.

2) Descriptors: Five appearance descriptors are calculated based on the pixel intensities of each mouth ROI for all speech samples processed. These appearance features include PCA, DCT, DWT, LBP, and HOG. The image-transform-based descriptors, i.e., PCA, DCT, and DWT [36], are the most commonly used feature representations for visual speech processing tasks [50], [51]. LBP has been widely used as a robust image compression technique for texture representation [37], and it is one of the most commonly used facial appearance descriptors in face recognition and facial expression recognition [39], [40]. HOG was first applied for human detection in images [38], but has proved successful in a wide range of computer vision problems, like recently in lip activity analysis [52].

Let us first explain each feature descriptor and the corresponding parameter choices in more detail.

PCA performs an eigen-decomposition of global intensity variation over the training set mouth ROIs. The aim herein is to capture the prevalent modes of texture variance in the mouth images, such as the ones that "control" the opening/closing of mouth, the stretching/shrinking of the lip corners, or the degree of visibility of teeth and tongue. Our feature vector corresponds to the principal components, i.e., eigenvectors of the covariance matrix, accounting for the 95% of the total variance. Mouth ROIs are first downsampled to dimension 32×32 to reduce dimensionality, which in the current experiments varies from 19 to 62 (see Section VI-B).

DCT projects intensity values onto real cosine basis functions. We use it with the goal of unveiling the richest global frequency information, in terms of energy of the visual signal of the mouth ROIs. In this way, we can capture fluctuations in the intensity values corresponding to movements of the mouth and the muscles around it, thus discarding homogeneous skin or lip regions. Similarly to [51], we apply 2-D DCT to eight non-overlapping 32×16 blocks of the ROI. Then, the 2-D DCT coefficients that lie in the upper-left corner of each block correspond to the lowest frequencies (equivalently, highest energies). After rearranging these coefficients in a zig-zag manner, we retain the first four for each block and so construct a 32-D vector.

DWT is used to perform frequency decomposition of the mouth images at various resolutions. The main difference to DCT lies in the basis functions which are shifted and scaled versions of some mother wavelets. Specifically, at each level DWT involves consecutive filtering in the horizontal and vertical direction of the image with a high-pass and low-pass filter, thus producing four subband images (LL, LH, HL, and HH). The image LL is used for the computation of the next level. After rescaling the ROIs to dimension 16×16, we use the Daubechies-4 wavelet filter, with three levels of decomposition, to compute the 2-D DWT coefficients. As in [51], the approximation coefficients (those related to the LL subband) of the third level, along with all the detailed coefficients (those related to LH, HL, and HH subbands) of the second and third level, are concatenated in a single 64-D vector.

LBP rely on intensity differences between each pixel of the image and a set of P equally spaced pixels on a circle of radius R around it. It is intuitive to expect that LBP will be capable of capturing finer manifestations of articulation in the mouth ROIs, since it encodes local texture information. Out of the 2^P possible binary patterns, it has been shown that $u2$-uniform patterns, i.e., those in which at most two 0/1 or 1/0 bitwise transitions occur, reveal the most prominent texture structures [37]. In this paper, we use the $\text{LBP}^{u2}_{(P=8,R=1)}$ operator, which acts in a neighbourhood of eight pixels on a circle of one-pixel radius. The final descriptor is a normalized histogram encoding the frequency of occurrence of each of the 59 $u2$-uniform patterns for the entire ROI [37].

HOG is an scale-invariant feature transform-like descriptor that performs local-gradient orientation histogramming. As such, it is expected to model local orientation information associated with meaningful edges, caused by visible effects of speech production and articulation. Each mouth ROI is divided into a fixed number of cells. The gradient at each pixel is discretized into one of four orientation bins, and each pixel contributes to the local histogram of the cell with a "vote" proportional to the gradient magnitude. The histogram of each cell is normalized four times, with respect to the total energy of the four 2×2 blocks of cells that contain that particular cell. Setting the cell size to 32×32 results in a feature vector of length 64 for the whole mouth ROI.

3) Normalization: Feature mean normalization (FMN) is commonly applied in lipreading, with the purpose of removing redundant information related to the speaker and the recording conditions [36]. This is achieved by subtracting the mean feature vector $\bar{f}$, over all N speech frames, from the feature vector f_n of each frame

$$f_n^{\text{FMN}} = f_n - \bar{f} \tag{2}$$

$$\bar{f} = \frac{1}{N}\sum_{n=1}^{N} f_n. \tag{3}$$

Aside from the above scheme, we also use a simple normalization technique introduced in [49] in order to alleviate undesirable illumination effects. This is based on the removal of the mean intensity over the utterance from each ROI, rather than the removal of the mean feature vector. We refer to this scheme as mean removal at the image level. Mean removed ROIs are computed in the same way as in (2) and (3), where the mouth ROIs I_n are used instead of feature vectors f_n.

Both approaches for normalization are used in this paper and are denoted by MRft and MRim, respectively. We evaluate them on both the original ROIs $\mathcal{I}$ and difference mouth ROIs $\mathcal{I}^{\text{diff}}$.

B. Shape Features

Geometric visual features extracted from the mouth region have been examined in the earliest works that targeted lipreading [53]. Experiments on bimodal speech recognition with HMM classifiers show that, even with the aid of a set of few simple geometric features, speech recognition accuracy rises significantly, compared to audio-only systems [54].

Our aim herein is to investigate whether capturing the dynamics of mouth shape, described for each frame by means of a relevant projection onto prototype mouth configurations, can encode accent-specific mouth shape and thus efficiently discriminate between native and non-native accent. We follow an approach based on point distribution models [55] and PCA.

1) PCA on Mouth Shape: The shape features are extracted frame-wise on 16 rescaled registered mouth points derived from the set $\mathcal{T}^{\text{reg}}$, which is produced in the output of the mouth ROI extraction procedure (see Section V-A1). We retain only 16 mouth points, out of the 34 lower face points. Specifically, we accumulate the (x, y) coordinates of these 16 points T_n^{reg} computed at each frame n in a single 32-D vector. Next, we apply PCA on the set of such feature vectors computed for all training speech samples. Again, we only retain the coefficients that are associated with the components collectively accounting for 95% of the total variance. The number of those components for our MOBIO database training set turns out to be 3. In order to enrich the representation and also enclose dynamic information, prior to the classification stage, first- and second-order time derivatives of the coefficients are appended. This results in a shape feature vector of dimensionality equal to 9.

2) Normalization: FMN is applied, in an identical way as in (2) and (3). Both the unnormalized and mean removed (MRft) version of the shape features are examined for the first experiment of this paper.

VI. Experimental Protocol—Implementation Details

A. Formation of Sets

As mentioned above in Section IV, we use totally 272 samples from 56 subjects—135 samples belonging to 28 native English speakers and 137 samples to 28 non-native English speakers. We evaluate the proposed framework by means of fourfold subject-independent experiments, so that collectively all data are used for testing. Each fold contains the visual speech data that correspond to 14 subjects, seven native, and seven non-native speakers (68, 69, 67, and 68 samples for the folds 1, 2, 3, and 4, respectively). For each of the four runs, the training set is composed of two folds, while one fold is used for the validation set and one for the test set. Thus, the three sets consistently include samples from different subjects and are also balanced over the two accent classes. The validation set serves to optimize the number of states for the HMMs for each fold. The best topology, which is tuned according to the mean value of the F1 measure for the two classes, is used for testing. The same distribution of samples/subjects across the four folds is utilized for all the experiments reported in this paper, thus establishing a consistent protocol.

B. Normalization Variants

The use of original or difference (diff) mouth ROIs (see Section V-A1) and the options of using no normalization scheme, mean removal at the feature level (MRft), or at the image level (MRim) (see Section V-A3), imply six different variants for each appearance feature examined. For the unnormalized features at the original ROIs, we use no notation, while for the remaining five combinations, we use the notation MRft, MRim, diff, diffMRft, and diffMRim. For the appearance-based PCA descriptor, since each time we retain the components that account for 95% of the total variance, the dimensionality varies with the normalization scheme used. The resulting sizes of the normalization variants PCA, PCA_{MRft}, PCA_{MRim}, PCA_{diff}, $\text{PCA}_{\text{diffMRft}}$, and $\text{PCA}_{\text{diffMRim}}$ are, respectively, 19, 19, 57, 58, 58, and 62.

C. HMMs Topology and Parameters

HMMs [16] have been widely used for accent analysis, classification, and detection in audio-based approaches [2], [5], [15]. In this paper, we hypothesize that HMM-based modeling will be able to unveil transitions in the speech evolution that could characterize accent-related traits such as effects associated with pronunciation, lexical stress, and articulation. We presume that the temporal aspect is not to be discarded, thereby regarding HMMs as more advantageous than GMMs for our framework. Our HMMs are constructed so that they do not rely on word or subword models. In other words, our models are given the task of modeling the entire speech segments as time-series data. In this way, our system is more generic, as our models are not vocabulary-specific. The idea of using HMMs to directly model time-sequential information as a whole, rather than separately in sub-units, is not new. For instance, continuous HMMs have been successfully employed in this fashion to categorize music clips into music genres [56]. Also, Schuller *et al.* [57] approached emotion recognition by feeding acoustic low-level descriptors, such as adjusted pitch and energy contours along with their first- and second-order derivatives, into emotion-specific continuous HMMs.

Two HMMs are learned, one for each accent class for each experimental fold. As a baseline topology, we use continuous-density left-right HMMs with no state skips and with one GMM mixture component for the observations of each emitting state. Indicative validation set results obtained by using models with skips between states and more than one GMM components showed no significant improvement in performance and thus are not reported. For classification, a Viterbi decoder [58] is used to estimate the average probability $P(\mathbf{x}|C_i)$ that the sequence of feature vectors $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$ corresponding to the whole speech example is produced by each of the two accent models $C_i, i \in \{1, 2\}$ (1: native and 2: non-native). The test utterance is finally assigned to the accent class i^* that yields the highest likelihood, i.e., $i^* = \arg\max_i P(\mathbf{x}|C_i)$.

In our experiments, HMMs were trained using the hidden Markov model toolkit (HTK) [58]. First, we manually create prototype models, which have zero mean and unity variance for the output distribution of states and a transition matrix that follows the topology constraints. The models are initialized and, subsequently, reestimated through the iterative Baum–Welch algorithm [58], with a convergence threshold of 10^{-5}. The number of states (referring to the total number of states, i.e., the emitting states plus the first and last state) is varied in the interval $\{5, 6, \ldots, 20\}$, and the value yielding the highest mean F1 measure on the validation set for each feature is used for the experiments on the test set. Thus, for each fold we use a different optimal HMM for testing, according to the performance obtained on the corresponding validation set. The classification accuracy measure reported for the test set results is the one produced by HTK, that is, the percentage of the correctly classified examples over the total number of test examples.

VII. Text-Dependent Experiments

Native/non-native visual-only accent discrimination results for the TD experiments conducted in this paper are reported here. The presentation of results aims also at a comparative evaluation across the different features and normalization schemes examined.

A. *Experiment 1—Training and Test on Whole Utterances*

In this experiment, we evaluate our framework using whole utterances, that is, the feature vectors in the training, validation, and test set for each fold, correspond to different subjects reading the same three-sentence paragraph (see Section IV). Results, in terms of average classification accuracy on the test set over the four folds, are presented in Table I.

TABLE I
Results in Terms of Average Test Set Classification Accuracy (%) Over the Four Folds of Whole MOBIO Database Utterances. The Number in the Subscript Refers to the Standard Deviation (%) for the Four Folds. For Each Feature, the Result Obtained by the Best-Performing Normalization Variant Is Shown in Boldface. Decision-Level Fusion Results Are Also Reported, Corresponding to the Combined Accent Predictions of HMMs Trained With the Three Top-Scoring Appearance Features as Well as All Five of Them

Normalisations/ Features	**PCA**	**DCT**	**DWT**	**LBP**	**HOG**	**Shape**
no norm.	**63.6**$_{(3.5)}$	**68.8**$_{(6.6)}$	**71.0**$_{(9.7)}$	55.9$_{(1.8)}$	**71.6**$_{(10.1)}$	**55.9**$_{(3.1)}$
MRft	53.4$_{(10.6)}$	58.5$_{(13.2)}$	60.3$_{(12.3)}$	52.2$_{(10.1)}$	54.4$_{(7.1)}$	54.4$_{(11.1)}$
MRim	49.6$_{(5.8)}$	58.5$_{(13.2)}$	60.3$_{(12.3)}$	58.9$_{(5.8)}$	65.0$_{(8.4)}$	-
diff	53.4$_{(12.5)}$	63.6$_{(8.8)}$	63.3$_{(9.2)}$	**62.9**$_{(1.6)}$	65.1$_{(16.7)}$	-
diffMRft	55.6$_{(14.3)}$	61.4$_{(7.5)}$	63.6$_{(9.5)}$	57.8$_{(8.7)}$	55.1$_{(6.8)}$	-
diffMRim	55.2$_{(15.2)}$	61.4$_{(7.5)}$	63.6$_{(9.5)}$	60.3$_{(3.5)}$	63.9$_{(16.6)}$	-

DCT + DWT + HOG	**PCA + DCT + DWT + LBP**$_{diff}$ **+ HOG**
73.5$_{(6.8)}$	**76.5**$_{(6.7)}$

First of all, classification scores yielded by the best normalization for each feature reveal the superiority of appearance features over the shape-based features. This conforms to evidence that geometric features are not as capable of encapsulating visual speech information, as appearance features [36]. This behavior is to be expected, as shape is related directly to coarse movement, while appearance reveals fine movement and tale-telling transient features. In other words, the positioning of the mouth and flesh around it mainly accounts for different sounds coming out of mouths with similar shapes. Based on the above, the shape features are not examined in the experiments that ensue in this paper.

Appearance features perform well in modeling accent-sensitive speech dynamics, with the unnormalized HOG descriptor yielding the highest average accuracy of 71.6%. HOG efficiently captures local edge orientation information, which corresponds to bulges and wrinkles in the area around the mouth, as well as lip configurations. These transient features can be manifestations of articulation phenomena in the visual stream, leading to efficient accent modeling by HOG. High mean accuracies of 71.0% and 68.8% are also furnished by the frequency-based descriptors DWT and DCT, respectively. This finding is congruent with lipreading results reported in [50], where frequency-based image transformations prove robust when acting in conjuction with visual-only HMMs. PCA and LBP seem to be less informative for the target problem when combined with HMMs. This behavior could be attributed to the higher susceptibility of PCA and LBP to misalignments and image registration errors.

Feature- and image-based normalization schemes do not seem to be beneficial for the majority of features. Regarding the shape-based PCA, the performances of the unnormalized and normalized features are quite similar (55.9% and 54.4%, respectively). This suggests that, since PCA is calculated on globally aligned pose-free points, where speaker-related information has already been suppressed, the removal of the mean feature vector does not lead to information gain for the target problem. DCT and DWT, which are based on component decoupling in the frequency domain, are not assisted by further normalization in the pixel domain or in the feature space. Similarly, the HOG descriptor, whose default computation has already catered for robust local normalization based on neighboring cells, does not show any performance boost when is further normalized. Appearance-based PCA seems more informative when capturing the global texture variance in the richer visual stream of unnormalized mouth ROIs, rather than in the sparser image domain of the diff images. Finally, LBP is the only descriptor that benefits from the diff ROIs. This is attributed to it being a local texture operator, and thus the influence of misalignments is alleviated when it is calculated on the difference ROIs.

After observing that the various appearance features perform best on the validation set in different folds, and bearing in mind that they can capture complementary information in the visual stream, we decide to combine the outputs of the corresponding HMMs. Therefore, decision-level fusion is also examined, that is, accent prediction on each test sample is provided as the mode of the predictions of three and five separately trained models. In other words, each test example is assigned to the class predicted by the majority of three and five appearance-based classifiers, respectively. For the three-model combination, the top-scoring appearance features in average on the validation set are picked, namely DCT, DWT, and HOG. Fusion results shown in Table I corroborate our assumption that the synergy of efficient frequency decoupling by DCT and DWT, and local edge orientation information by HOG, can lead to more accurate accent classification by the combined models (73.5% accuracy). Moreover, when global texture variance and local texture information are incorporated through the PCA- and LBP-trained HMMs, respectively, the five-model combination reaches even higher accuracy, amounting to 76.5%. Note also that the standard deviations of 6.8% and 6.7%, occurring for the corresponding fusion results,

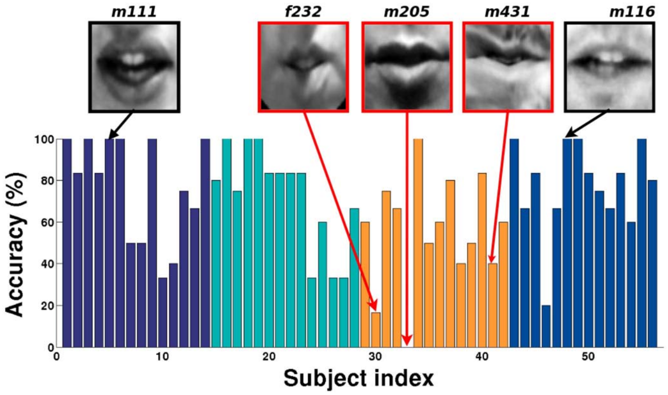

Fig. 4. Test set classification accuracy (%) obtained for each MOBIO subject by the DWT-trained HMM. The plot bars corresponding to subjects that belong to the same test fold (14 subjects per fold) are shown in the same color. Five-mouth ROI images accompany the graph. Three ROIs correspond to subjects of the third fold (the one shown in orange), on which the accuracies reached are among the worst. The remaining two ROIs belong to subjects m111 and m116, whose samples are tested on the first and fourth fold, respectively, with resulting overall accuracies of 100%.

are relatively lower than the values characterizing the distributions of results for most of the single-feature schemes. This highlights that the fusion framework is more robust, with its performance fluctuating less across the folds, since predictions made by more than one HMMs collectively are more confident.

The high standard deviation observed for most of the classification results (reported in the subscript of accuracy percentages in Table I) can be largely attributed to bad performance achieved on samples of certain MOBIO subjects. Accent misclassification occurs in the outputs of nearly all HMMs on most of the samples of those subjects, the majority of which belongs to the third test fold of our experiments. In Fig. 4, we show the test set classification accuracy, as produced by the DWT-trained HMM, separately for the samples of each of the 56 MOBIO subjects. Each fold contains 14 subjects and is shown in different color in the graph. One can easily notice the presence of "bad-performing" subjects in the third fold (16.7% for subject f232, 0% for subject m205, and 40% for subject m431, respectively). By observing the characteristic mouth ROIs, which are visualized above the graph for these subjects, one will easily discern that the ROIs of subjects f232 and m431 are the result of inaccurate frontal warping, whereas for subject m205 the illumination is bad and largely varying across the ROI. We deduce that these factors, which are detrimental to the quality of mouth ROIs for these subjects, make it impossible for any feature scheme to capture the accent-related information correctly. Instead, on subjects m111 and m116, which belong to the first and fourth fold, respectively, an accuracy of 100% is achieved by the DWT descriptor, as well as all the remaining features. As can be seen in Fig. 4, the corresponding mouth ROIs are much more reliable in terms of warping, alignment, and illumination.

B. Experiment 2—Test on Isolated Words

In this experiment, we test the HMMs, which have been previously trained and validated on utterances of the same three-sentence paragraph, on speech segments containing isolated words, again retrieved from the utterances of the same paragraph by the same test set subjects for each of the four folds. Our aim is to investigate whether the proposed system can still yield accurate accent predictions based on short segments of speech, corresponding to word utterances.

TABLE II
RESULTS IN TERMS OF AVERAGE TEST SET CLASSIFICATION ACCURACY (%) OVER THE FOUR FOLDS OF ISOLATED WORD UTTERANCES FROM THE MOBIO SET-SPEECH PARAGRAPH. FOR EACH FEATURE, ONLY THE BEST ACCURACY AMONG NORMALIZATION VARIANTS IS REPORTED, AND THE STANDARD DEVIATION (%) FOR THE FOUR FOLDS IS SHOWN IN THE SUBSCRIPT. DECISION-LEVEL FUSION RESULTS ARE ALSO REPORTED, CORRESPONDING TO THE COMBINED ACCENT PREDICTIONS OF ALL HMMS TRAINED WITH APPEARANCE FEATURES. THE HIGHEST SCORE ACHIEVED FOR EACH WORD AMONG THE SINGLE-FEATURE FRAMEWORKS IS SHOWN IN BOLDFACE

Words/ Features	**PCA**	**DCT**	**DWT**	**LBP*diff***	**HOG**	**All**
BEHAVIOR	$64.7_{(6.8)}$	$\mathbf{70.3}_{(7.1)}$	$66.2_{(10.5)}$	$62.1_{(5.6)}$	$61.7_{(10.3)}$	$70.6_{(8.5)}$
BIOMETRIC	$66.6_{(3.1)}$	$\mathbf{67.7}_{(10.1)}$	$67.0_{(11.0)}$	$60.7_{(4.9)}$	$59.9_{(4.3)}$	$72.1_{(4.6)}$
DATABASE	$63.6_{(7.8)}$	$\mathbf{68.8}_{(10.0)}$	$64.7_{(7.0)}$	$59.6_{(2.4)}$	$64.6_{(12.1)}$	$71.7_{(3.8)}$
INFORMATION	$64.4_{(8.4)}$	$\mathbf{69.2}_{(9.5)}$	$64.7_{(4.3)}$	$62.1_{(4.0)}$	$63.9_{(10.5)}$	$74.7_{(7.5)}$
PUBLICLY	$62.9_{(3.6)}$	$\mathbf{67.7}_{(9.4)}$	$65.9_{(9.4)}$	$62.4_{(6.5)}$	$61.3_{(5.7)}$	$71.7_{(4.6)}$
PURPOSES	$66.2_{(4.5)}$	$67.0_{(7.6)}$	$\mathbf{67.3}_{(5.9)}$	$63.5_{(8.6)}$	$60.6_{(9.2)}$	$72.8_{(2.6)}$
RESPONSIBLE	$61.8_{(5.3)}$	$\mathbf{68.8}_{(10.4)}$	$67.0_{(7.4)}$	$59.5_{(8.2)}$	$62.4_{(12.4)}$	$68.0_{(4.2)}$
STATES	$60.3_{(9.4)}$	$\mathbf{66.6}_{(9.5)}$	$63.3_{(8.4)}$	$56.6_{(2.1)}$	$58.3_{(17.6)}$	$64.7_{(6.8)}$
UNDERSTAND	$65.1_{(6.0)}$	$\mathbf{67.4}_{(9.7)}$	$65.5_{(8.9)}$	$59.2_{(3.2)}$	$63.9_{(8.8)}$	$69.9_{(6.7)}$

As there were no available word-level transcriptions, the HTK Toolkit [58] was used to segment the entire paragraph into words, through monophone-based Viterbi alignment. Nine words, whose duration in all corresponding utterances exceeds 300 ms, were selected for the current experiment. Prior to feature extraction, mouth ROIs were upsampled from 15 to 100 frames/s, in order to obtain a sufficient number of feature vectors for each word. For the word UNDERSTAND, which appears twice in the three-sentence paragraph, only the first token is used. Since the same normalization schemes were again the best-performing in average on the test set of isolated words like in the previous experiment, we choose to present only the corresponding results obtained by those (namely, PCA, DCT, DWT, LBP_{diff}, and HOG). The mean value of test set classification accuracy over the four folds, for each word and feature, is reported in Table II. Results obtained by decision-level fusion of all five models are shown in the right-most column. Since the five-model fusion consistently outperformed the three-model combinations across all words examined, we choose to report only the results of the former.

Our single-feature frameworks perform quite accurately on predicting the accent class of the test word utterances. The frequency-based descriptors DCT and DWT prove again highly informative for the target task, accounting for the top two highest classifications scores for all nine words examined (DCT always yields better accuracy, except for the word PURPOSES). Another holistic descriptor, namely PCA,

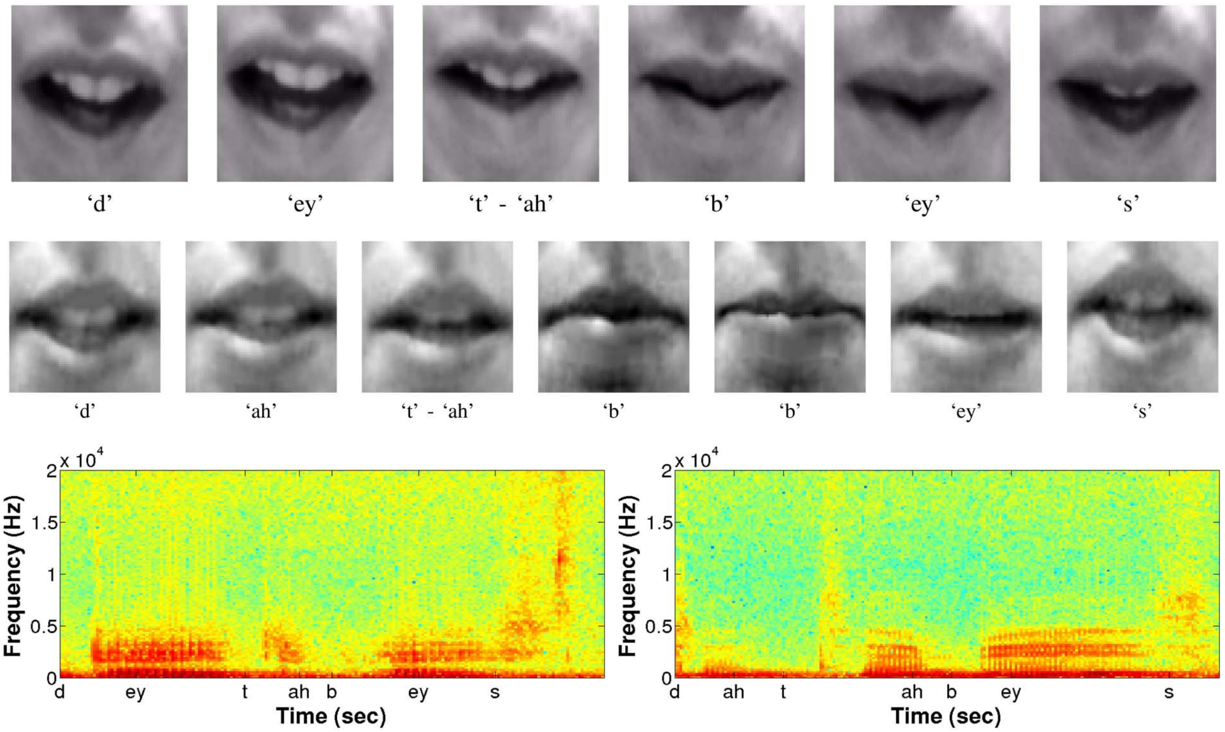

Fig. 5. Mouth ROIs extracted from visual speech frames corresponding to the word DATABASE, as produced by one native speaker (first row: subject m109, phase I, session 05, and frames 78–83) and one non-native speaker (second row: subject f502, phase I, session 06, and frames 88–94) from the MOBIO database. Both samples were correctly classified by our system based on the corresponding top-scoring appearance feature DCT. In the third row, one can see the spectrograms for the corresponding audio files (left: native speaker/utterance in the first row and right: non-native speaker/utterance in the second row). For better visualization, the temporal location of the phonemes pronounced by the speakers is shown in the captions of the mouth ROI images for the video stream, and in the time axes of the spectrograms for the audio stream, respectively.

accounts for the third best performance on seven out of the nine words. These three global transformations are known for their discriminative power in visual speech processing, even for the case of short utterances of isolated words [36]. On the other hand, the local edge orientation-based descriptor HOG accounts for lower performance in the current experiment, compared to the experiment of the previous section. This happens presumably due to the effect of interpolation artifacts and image registration discontinuities induced on mouth ROIs by the upsampling procedure. Finally, again many misclassifications occur for the LBP-trained HMMs, showing a pattern similar to that observed in the previous experiment (see Section VII-A).

The best accuracy across words, produced by a single HMM, is 70.3% and is reached by means of the descriptor DCT on the word BEHAVIOR. The high accuracy achieved specifically on this word might be partly attributed to the difficulty in pronounciation that the voiceless fricative "h" poses to non-native speakers of English [59]. It is worth noting that higher accuracy is obtained for relatively longer words, such as INFORMATION and RESPONSIBLE, as opposed to shorter words, such as STATES and PURPOSES. However, we believe that the longer duration of the former is not the sole factor that could explain the higher performance. As a matter of fact, these longer words involve more complicated rhythmic patterns in their pronunciation, which differs markedly between the two accent classes. The high percentages furnished for specific words as opposed to others might be also due to phoneme substitutions occurring frequently when they are uttered by non-native speakers. In Fig. 5, one can see the mouth ROIs corresponding to an utterance of the word DATABASE, as pronounced by one native and one non-native speaker. Note that in the case of the native speaker the first "a" in the word DATABASE is pronounced as "ey," whereas the same vowel is pronounced as "ah" by the non-native speaker (more evident from the inspection of the spectrograms). Also, one can easily notice from the visual speech frames that in the case of the non-native speaker the labial consonant "b" is pronounced much more intensely.

The framework that combines the accent predictions of five HMMs through majority voting accounts for boost in classification accuracy over the single-feature schemes, for seven out of nine words examined. This behavior is in accordance with the results of the previous experiment, where again fusion was beneficial. Only for the words RESPONSIBLE and STATES, on which there is large deviation from the accuracy achieved by the top-scoring DCT compared to that of the remaining four features, does fusion result in performance drop. The varying discriminativeness of each word's pronunciation across the two accent classes is again depicted in the fusion results. Words involving more complicated rhythmic patterns and being more likely to give rise to phoneme substitutions when pronounced by L2 speakers (e.g., INFORMATION and BEHAVIOR), correspond to higher accuracies.

TABLE III

RESULTS IN TERMS OF MEAN CLASSIFICATION ACCURACY (%) OVER THE 12 FOLDS ON THE TEST SETS OF THE MOBIO SENTENCES. RESULTS FOR BOTH TD AND TI SCENARIOS, ARE PRESENTED IN THE SUBTABLES BELOW. THE NUMBER IN THE SUBSCRIPT REFERS TO THE STANDARD DEVIATION (%) OF THE ACCURACY VALUES OVER THE 12 FOLDS FOR EACH FEATURE-NORMALIZATION COMBINATION. (a) TD. (b) TI

(a)

Normalisations/ Features	**PCA**	**DCT**	**DWT**	**LBP**	**HOG**
no norm.	**63.1**$_{(10.9)}$	**68.0**$_{(9.5)}$	62.5$_{(7.8)}$	57.4$_{(6.4)}$	**65.4**$_{(12.7)}$
MRft	55.8$_{(11.5)}$	62.5$_{(6.5)}$	62.3$_{(8.6)}$	49.6$_{(9.0)}$	53.0$_{(7.8)}$
MRim	56.1$_{(10.1)}$	62.5$_{(6.5)}$	**62.9**$_{(9.4)}$	58.5$_{(10.2)}$	60.6$_{(14.6)}$
diff	59.5$_{(8.2)}$	62.2$_{(9.7)}$	61.6$_{(9.0)}$	62.3$_{(9.7)}$	63.8$_{(11.5)}$
diffMRft	60.0$_{(8.2)}$	61.3$_{(7.0)}$	62.4$_{(9.0)}$	61.8$_{(10.2)}$	51.3$_{(8.7)}$
diffMRim	56.9$_{(10.8)}$	61.3$_{(7.0)}$	62.4$_{(9.0)}$	**62.7**$_{(9.4)}$	64.0$_{(11.3)}$

(b)

Normalisations/ Features	**PCA**	**DCT**	**DWT**	**LBP**	**HOG**
no norm.	**65.5**$_{(7.3)}$	**62.6**$_{(16.6)}$	**63.5**$_{(6.4)}$	55.6$_{(4.9)}$	**62.7**$_{(13.9)}$
MRft	55.8$_{(11.4)}$	60.6$_{(7.7)}$	59.2$_{(8.6)}$	51.0$_{(10.5)}$	53.5$_{(6.8)}$
MRim	50.4$_{(7.1)}$	60.6$_{(7.7)}$	59.2$_{(8.6)}$	56.9$_{(7.4)}$	61.7$_{(16.0)}$
diff	57.9$_{(10.3)}$	60.7$_{(7.5)}$	61.8$_{(7.6)}$	**62.9**$_{(10.7)}$	60.9$_{(12.7)}$
diffMRft	58.1$_{(9.4)}$	60.6$_{(6.4)}$	60.7$_{(6.8)}$	60.2$_{(8.1)}$	47.9$_{(8.9)}$
diffMRim	57.6$_{(8.9)}$	60.7$_{(6.4)}$	60.7$_{(6.8)}$	62.9$_{(9.6)}$	61.7$_{(12.7)}$

VIII. TEXT-INDEPENDENT EXPERIMENTS

In all the subject-independent experiments reported above, the speech content of the test examples was identical to some or the whole part of the training speech utterances (Sections VII-B and VII-A, respectively). In the experiment of this section, our goal is to examine the efficacy of our method in a TI experimental scenario.

To conduct a fair comparison between the TD and TI case, we do the following: for each of the four folds, we keep the training set fixed, and test on examples with matching (non-matching) speech content for the TD (TI) case. We perform the TD and TI experiments for each of the three sentences of the MOBIO paragraph (see Section IV). First, we train our models on the first sentence for both TI and TD, validate them (i.e., optimize the number of HMM states) on the second (first) sentence for the TI (TD) case, and test them on the third (first) sentence for the TI (TD) case. For the second and third sentence, the two experiments are performed in a similar fashion. In other words, in the TD scenario the speech content is fixed across all three sets, while in the TI scenario each set contains examples that correspond to a different sentence of the MOBIO paragraph. For each of the aforementioned three experiments (one for each sentence, both TD and TI), we perform the same fourfold cross-validation, by using 28 subjects for the training set, 14 for the validation set, and the remaining 14 for the test set, in each run. The combination of one experiment per sentence and four folds per experiment results in 12 accuracy values for each of the TD and TI cases. The mean of the classification accuracies over the 12 folds for each feature-normalization combination is reported in separate subtables for the TD and TI scenario in Table III.

From the comparative inspection of the results in Table III for the two cases, it is evident that there is a trend for higher accuracies for the text-dependent scenario, compared to the text-independent counterparts. One characteristic example of TD superiority is the unnormalized DCT that yields 68.0% for the TD case, as opposed to 62.6% in the TI case. This is quite intuitive, as in the TD case, the modeling for the two accent classes is not affected by speech dynamics entailed by the presence of different words, hence articulation transitions. Frequency-based appearance representations, i.e., DCT and DWT, again work well, furnishing accuracies consistently higher than 60% for the TD case [Table III(a)]. Nonetheless, it is worth noting that, even in the more demanding TI case, our framework classifies the speaker's accent at a performance much higher than chance level, with PCA achieving the best performance of 65.5% [Table III(b)].

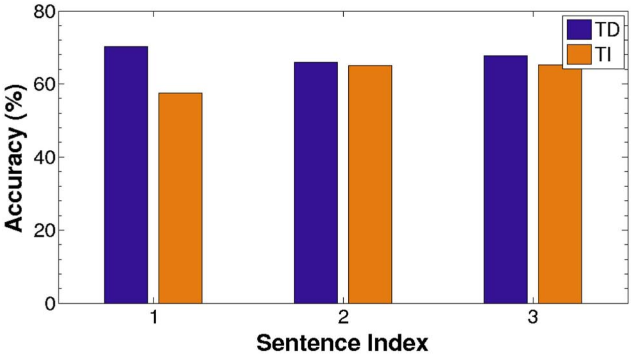

Fig. 6. Classification accuracies (%) yielded by the unnormalized DCT appearance feature for each MOBIO sentence, both for the TD and TI scenarios. The height of each bar in the graph corresponds to the mean value over the four folds of each experiment shown.

Comparing the TD results [Table III(a)] with the corresponding accuracies achieved in our first text-dependent experiment (Table I), one can notice that performance drops. This divergence can be attributed to the fact that in the former case longer utterances (three-sentence paragraph) are used to train the models, whereas in the latter case the training set contains shorter (one-sentence) speech examples. The same assumption holds for the test data as well. This behavior is in conformity with evidence from research in visual-only language identification, according to which performance increases with the length of speech data [22].

In Fig. 6, we show the performance of the unnormalized DCT descriptor for both the TD and TI cases, separately for each MOBIO sentence. In the bar graph, a sentence index of value 1 denotes that training has been performed on the first sentence for both cases, while the test sentence is the third (first) one for the TI (TD) case. The notation is similar for the remaining two sentence indices. In the TD case, best performance is achieved when the HMMs are trained and tested on the first sentence. This could be due to the longer duration of the first sentence. Performance is almost identical for the second and third sentence, which have similar durations. Instead, in the TI case, the lowest accuracy is achieved when sentence index equals 1, i.e., when the model is trained on the first sentence, validated on the second and tested on the third. One speculation is that when training HMMs on the first sentence, which is composite and contains also longer words and pauses, they unavoidably learn these prominent dynamics related to speech content. Thus, when these are not encountered in the remaining two sentences, which are simpler in structure, performance in accent prediction drops.

IX. Conclusion

In this paper, we presented a fully automated approach to discriminating native from non-native speech in English, based on visual features only. Overall, this paper aims to provide a basic study of visual-discrimination between native and non-native speech, thus introducing a research area which can be extremely useful in biometric applications. We demonstrate that useful information for discrimination between native and non-native speech is present in the visual stream and thus is expected to improve performance when combined with audio-only methods, especially in noisy environments.

Subject-independent cross-validation experiments were conducted on continuous fixed-content speech and isolated word utterances captured by mobile devices. Our framework was also comparatively evaluated on both the text-dependent and text-independent scenarios.

Various appearance- and shape-based features were examined. A comparative evaluation of features was performed, revealing the superiority of appearance-based features over shape features. The best accuracy score of 71.6% is achieved by the HOG descriptor. Decision-level fusion consistently provides performance boost over single HMMs, with the fusion of five HMMs yielding accuracy of 76.5%. Our framework classifies accent accurately also in short speech segments of isolated words. Finally, even in the case of text-independent experimental scenario, performance remains much higher than chance level.

In this paper, classifiers were assigned the task to model the evolution of visual speech belonging to each accent class as a time-series input and produce a single accent label for the whole test utterance. No constraints were imposed on these articulation transitions, apart from the inherent assumptions of the classifier (e.g., the Markov assumption and left-right topology of HMMs). We plan to investigate alternative ways to aid the sequential modeling task by reinforcing meaningful abstracting, quantization, or segmentation preprocessing on the speech utterance. We also aim to examine audiovisual approaches for accent classification.

References

[1] L. M. Arslan and J. H. L. Hansen, "Language accent classification in American English," *Speech Commun.*, vol. 18, no. 4, pp. 353–367, 1996.

[2] L. W. Kat and P. Fung, "Fast accent identification and accented speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, vol. 1. Phoenix, AZ, USA, 1999, pp. 221–224.

[3] P. A. Torres-Carrasquillo, T. P. Gleason, and D. A. Reynolds, "Dialect identification using Gaussian mixture models," in *Proc. ODYSSEY Speaker Lang. Recognit. Workshop*, Toledo, Spain, 2004, pp. 757–760.

[4] E. Shriberg *et al.*, "Detecting nonnative speech using speaker recognition approaches," in *Proc. IEEE Odyssey Speaker Lang. Recognit. Workshop*, Stellenbosch, South Africa, 2008, p. 26.

[5] B. Tan, Q. Li, and R. Foresta, "An automatic non-native speaker recognition system," in *Proc. IEEE Int. Conf. Technol. Homel. Security*, Waltham, MA, USA, 2010, pp. 77–83.

[6] M. K. Omar and J. Pelecanos, "A novel approach to detecting non-native speakers and their native language," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Dallas, TX, USA, 2010, pp. 4398–4401.

[7] S. Zhang and Y. Qin, "Semi-supervised accent detection and modeling," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Vancouver, BC, Canada, 2013, pp. 7175–7179.

[8] V. Gupta and P. Mermelstein, "Effects of speaker accent on the performance of a speaker-independent, isolated-word recognizer," *J. Acoust. Soc. Amer.*, vol. 71, no. 6, pp. 1581–1587, 1982.

[9] J. H. L. Hansen and L. M. Arslan, "Foreign accent classification using source generator based prosodic features," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, vol. 1. Detroit, MI, USA, 1995, pp. 836–839.

[10] P. Angkititrakul and J. H. L. Hansen, "Advances in phone-based modeling for automatic accent classification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 2, pp. 634–646, Mar. 2006.

[11] A. K. Jain, S. C. Dass, and K. Nandakumar, "Soft biometric traits for personal recognition systems," in *Biometric Authentication*. Berlin, Germany: Springer, 2004, pp. 731–738.

[12] S. Mangayyagari, T. Islam, and R. Sankar, "Enhanced speaker recognition based on intra-modal fusion and accent modeling," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, Tampa, FL, USA, 2008, pp. 1–4.

[13] I. Amdal, F. Korkmazskiy, and A. C. Surendran, "Joint pronunciation modelling of non-native speakers using data-driven methods," in *Proc. INTERSPEECH*, Beijing, China, 2000, pp. 622–625.

[14] F. William, A. Sangwan, and J. H. L. Hansen, "Automatic accent assessment using phonetic mismatch and human perception," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 9, pp. 1818–1829, Sep. 2013.

[15] C. Teixeira, I. Trancoso, and A. J. Serralheiro, "Accent identification," in *Proc. IEEE Int. Conf. Spoken Language (ICSLP)*, vol. 3. Philadelphia, PA, USA, 1996, pp. 1784–1787.

[16] L. R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, vol. 14. Englewood Cliffs, NJ, USA: Prentice Hall, 1993.

[17] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, Dec. 1976.

[18] C. Neti *et al.*, "Audio-visual speech recognition," Dept. CLSP, Johns Hopkins University, Baltimore, MD, USA, Tech. Rep. WS00AVSR, 2000.

[19] A. W.-C. Liew, S. Wang, and I. Global, *Visual Speech Recognition: Lip Segmentation and Mapping*. Hershey, PA, USA: Med. Inf. Sci. Ref., 2009.

[20] D. Stewart, R. Seymour, A. Pass, and J. Ming, "Robust audio-visual speech recognition under noisy audio-video conditions," *IEEE Trans. Cybern.*, vol. 44, no. 2, pp. 175–184, Feb. 2014.

[21] R. E. Ronquest, S. V. Levi, and D. B. Pisoni, "Language identification from visual-only speech signals," *Atten. Percept. Psycho.*, vol. 72, no. 6, pp. 1601–1613, 2010.

[22] J. L. Newman and S. J. Cox, "Language identification using visual features," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 7, pp. 1936–1947, Sep. 2012.

[23] A. Irwin, S. Thomas, and M. Pilling, "Identification of language and accent through visual speech," in *Proc. Speech Prosody*, 2006.

[24] C. McCool *et al.*, "Bi-modal person recognition on a mobile phone: Using mobile phone data," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Melbourne, VIC, Australia, 2012, pp. 635–640.

[25] M. Munro, "Foreign accent and speech intelligibility," in *Phonology and Second Language Acquisition*, J. G. H. Edward and M. L. Zampini, Eds. Amsterdam, The Netherlands: John Benjamins, 2008, pp. 193–218.

[26] S. Deshpande, S. Chikkerur, and V. Govindaraju, "Accent classification in speech," in *Proc. IEEE Workshop Autom. Identif. Adv. Technol.*, Buffalo, NY, USA, 2005, pp. 139–143.

[27] F. Biadsy, J. Hirschberg, and D. P. W. Ellis, "Dialect and accent recognition using phonetic-segmentation supervectors," in *Proc. INTERSPEECH*, Florence, Italy, 2011, pp. 745–748.

[28] G. Choueiter, G. Zweig, and P. Nguyen, "An empirical study of automatic accent classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Las Vegas, NV, USA, 2008, pp. 4265–4268.

[29] C. D. Cieri, D. Miller, and K. Walker, "The Fisher corpus: A resource for the next generations of speech-to-text," in *Proc. LREC*, vol. 4. Lisboa, Portugal, 2004, pp. 69–71.

[30] S. Sam *et al.*, "Speech modulation features for robust nonnative speech accent detection," in *Proc. INTERSPEECH*, Florence, Italy, 2011, pp. 2417–2420.

[31] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proc. IEEE*, vol. 91, no. 9, pp. 1306–1326, Sep. 2003.

[32] I. Matthews, T. F. Cootes, J. A. Bangham, S. Cox, and R. Harvey, "Extraction of visual features for lipreading," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 2, pp. 198–213, Feb. 2002.

[33] T. Ellis, M. MacSweeney, B. Dodd, and R. Campbell, "TAS: A new test of adult speechreading—Deaf people really can be better speechreaders," in *Proc. Int. Conf. Aud. Vis. Speech Process. (AVSP)*, Aalborg, Denmark, 2001, pp. 13–17.

[34] A. Irwin, M. Pilling, and S. M. Thomas, "An analysis of British regional accent and contextual cue effects on speechreading performance," *Speech Commun.*, vol. 53, no. 6, pp. 807–817, 2011.

[35] J. Orozco, O. Rudovic, J. Gonzàlez, and M. Pantic, "Hierarchical on-line appearance-based tracking for 3D head pose, eyebrows, lips, eyelids and irises," *Image Vis. Comput.*, vol. 31, no. 4, pp. 322–340, Feb. 2013.

[36] G. Potamianos, H. P. Graf, and E. Cosatto, "An image transform approach for HMM based automatic lipreading," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Chicago, IL, USA, 1998, pp. 173–177.

[37] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[38] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1. San Diego, CA, USA, 2005, pp. 886–893.

[39] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image Vis. Comput.*, vol. 27, no. 6, pp. 803–816, 2009.

[40] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin, "Static and dynamic 3D facial expression recognition: A comprehensive survey," *Image Vis. Comput.*, vol. 30, no. 10, pp. 683–697, 2012.

[41] M. Valstar, S. Zafeiriou, and M. Pantic, "Facial action recognition in 2D and 3D," in *Face Recognition Adverse Conditions*. Hershey, PA, USA: IGI Global, 2014, pp. 167–186.

[42] L.-M. Lee and F.-R. Jean, "Adaptation of hidden Markov models for recognizing speech of reduced frame rate," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 2114–2121, Dec. 2013.

[43] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[44] A. Vinciarelli *et al.*, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 69–87, Jan./Mar. 2012.

[45] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image Vis. Comput.*, vol. 31, no. 2, pp. 120–136, 2013.

[46] H. Meng and N. Bianchi-Berthouze, "Affective state level recognition in naturalistic facial and vocal expressions," *IEEE Trans. Cybern.*, vol. 44, no. 3, pp. 315–328, Mar. 2014.

[47] M. Raab, R. Gruhn, and E. Noeth, "Non-native speech databases," in *Proc. IEEE Workshop Autom. Speech Recognit. Und.*, Kyoto, Japan, 2007, pp. 413–418.

[48] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[49] P. J. Lucey, "Lipreading across multiple views," Ph.D. dissertation, Faculty Built Environ. Eng., Queensland Univ. Technol., Brisbane, QLD, Australia, 2007.

[50] I. Matthews, G. Potamianos, C. Neti, and J. Luettin, "A comparison of model and transform-based visual features for audio-visual LVCSR," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Tokyo, Japan, 2001, pp. 825–828.

[51] R. Seymour, D. Stewart, and J. Ming, "Comparison of image transform-based features for visual speech recognition in clean and corrupted videos," *Image Video Process.*, vol. 2008, Jan. 2008, Art. ID 810362.

[52] E. Zavesky, "LipActs: Efficient representations for visual speakers," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Barcelona, Spain, 2011, pp. 1–4.

[53] E. Petajan, B. Bischoff, D. Bodoff, and N. M. Brooke, "An improved automatic lipreading system to enhance speech recognition," in *Proc. SIGCHI Conf. Human Factors Comput. Syst.*, Montreal, QC, Canada, 1988, pp. 19–25.

[54] M. N. Kaynak *et al.*, "Lip geometric features for human–computer interaction using bimodal speech recognition: Comparison and analysis," *Speech Commun.*, vol. 43, nos. 1–2, pp. 1–16, 2004.

[55] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models—Their training and application," *Comput. Vis. Image Und.*, vol. 61, no. 1, pp. 38–59, 1995.

[56] X. Shao, C. Xu, and M. S. Kankanhalli, "Unsupervised classification of music genre using hidden Markov model," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, vol. 3. Taipei, Taiwan, 2004, pp. 2023–2026.

[57] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, vol. 2. Hong Kong, 2003, pp. 1–4.

[58] S. Young *et al.*, *The HTK Book*, Dept. Eng., Cambridge Univ., Cambridge, U.K., 2002, p. 175.

[59] M. L. Castelo, "The phonetic and phonemic treatment of English /h/," *Phonetica*, vol. 11, no. 2, pp. 116–123, 1964.

Christos Georgakis (S'13) received the Diploma degree in electrical and computer engineering from the National Technical University of Athens, Athens, Greece, in 2011. He is currently pursuing the Ph.D. degree from the Department of Computing, Imperial College London, London, U.K., under the supervision of Prof. M. Pantic.

He is a member of the iBUG Group, Department of Computing, Imperial College London. His current research interests include human behavior analysis, computer vision, and statistical machine learning.

Stavros Petridis (S'07–M'10) received the B.Sc. degree in electrical and computer engineering from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2004, and the M.Sc. degree in advanced computing and the Ph.D. degree in computing from Imperial College London, London, U.K., in 2005 and 2012, respectively.

He is a Research Associate with the Department of Computing, Imperial College London. He has been a Research Intern with the Image Processing Group, University College London, London, and the Field Robotics Centre, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA, and a Visiting Researcher with the Affect Analysis Group, University of Pittsburgh, Pittsburgh. He is currently researching on deep learning approaches for audiovisual fusion. His current research interests include pattern recognition and machine learning and their application to multimodal recognition of human nonverbal behavior and nonlinguistic vocalizations.

Maja Pantic (M'98–SM'06–F'12) received the M.Sc. and Ph.D. degrees in computer science from the Delft University of Technology, Delft, The Netherlands, in 1997 and 2001.

She is a Professor of affective and behavioral computing with the Department of Computing, Imperial College London, London, U.K., and the Department of Computer Science, University of Twente, Enschede, The Netherlands.

Prof. Pantic was a recipient of the European Research Council Starting Grant Fellowship 2008, the Roger Needham Award 2011, and various awards for her work on automatic analysis of human behavior. She currently serves as the Editor-in-Chief of *Image and Vision Computing Journal* and as an Associate Editor for both the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING.