VISUAL-ONLY DISCRIMINATION BETWEEN NATIVE AND NON-NATIVE SPEECH

Christos Georgakis[1] *Stavros Petridis*[1] *Maja Pantic*[1,2]

[1]Department of Computing, Imperial College London, London, UK
[2]EEMCS, University of Twente, Enschede, The Netherlands

ABSTRACT

Accent is an important biometric characteristic that is defined by the presence of specific traits in the speaking style of an individual. These are identified by patterns in the speech production system, such as those present in the vocal tract or in lip movements. Evidence from linguistics and speech processing research suggests that visual information enhances speech recognition. Intrigued by these findings, along with the assumption that visually perceivable accent-related patterns are transferred from the mother tongue to a foreign language, we investigate the task of discriminating native from non-native speech in English, employing visual features only. Training and evaluation is performed on segments of continuous visual speech, captured by mobile phones, where all speakers read the same text. We apply various appearance descriptors to represent the mouth region at each video frame. Vocabulary-based histograms, being the final representation of dynamic features for all utterances, are used for recognition. Binary classification experiments, discriminating native and non-native speakers, are conducted in a subject-independent manner. Our results show that this task can be addressed by means of an automated approach that uses visual features only.

Index Terms— Non-Native Speech Identification, Accent Classification, Visual Speech Processing

1. INTRODUCTION

Accent manifests itself in speech through a set of pronunciation, intonation, lexical stress, rhythmic and articulation patterns, present in a common language of a group of people. Determining whether a speaker is native or non-native from a spoken utterance, without knowledge of speech content, could arguably enhance speech recognition, as a preprocessing accent-biasing step, or serve speaker verification purposes, as a biometrics problem in its own right [1]. Moreover, it could provide assistance to intelligent applications that require voice/conversation adaptation [2].

Most related early work has approached accent identification through a classification framework whose goal is to assign a speech example either to the accent of mother tongue or to one of separately modelled foreign-language accents [1, 2, 3]. These approaches use Hidden Markov Models (HMMs) modelling, on the phoneme or word level, trained on cepstral or prosodic acoustic features, and are evaluated on isolated word databases. Trajectory-based classifiers are, alternatively, used to capture accent-sensitive spectral dynamics [4].

More recently, binary discrimination between native and non-native speech has been targeted, as opposed to the multiclass accent identification task [5, 6, 7]. These works borrow inspiration from the similar fields of language and speaker recognition, and perform evaluation in larger speech databases. Shriberg *et al.* [5] employ maximum likelihood regression and phone N-gram features, while in [6] the proposed framework consists of two phonetic-based and two non-phonetic-based subsystems, built on HMMs and Gaussian Mixture Models (GMMs), respectively. Omar and Pelecanos [7] introduce a novel universal background model and use it in conjunction with Support Vector Machines (SVMs) to detect non-native speakers and their native language.

All the above mentioned related research has relied exclusively on acoustic features, ignoring information from the visual cue. However, features derived from the visual modality have been successfully used for speech recognition [8, 9, 10]. Such findings underline the contribution of visual information to speech perception by humans, especially when the auditory stream is noisy. Furthermore, visual speech alone has proved sufficient source of information for human observers to identify the language spoken by a talker. For example, Ronquest *et al.* [11] report several experiments, where the participants perform much better than chance in visual discrimination between English and Spanish. Newman and Cox [12] employ appearance and shape features, along with language phonotactics modelling, and show that automatic language identification is feasible even through visual features only. However, their approach depends on the accurate sub-phonetic or phone-level recognition of visual speech.

Motivated by the aforementioned ideas, we address discrimination between native and non-native speech in English, formulated as a binary visual-only classification problem. To the best of our knowledge, there has been no previous work that addresses this task through the visual modality only. Furthermore, our approach has the advantage of not relying on either speech transcriptions or language-specific modelling, as opposed to other visual-only speech processing

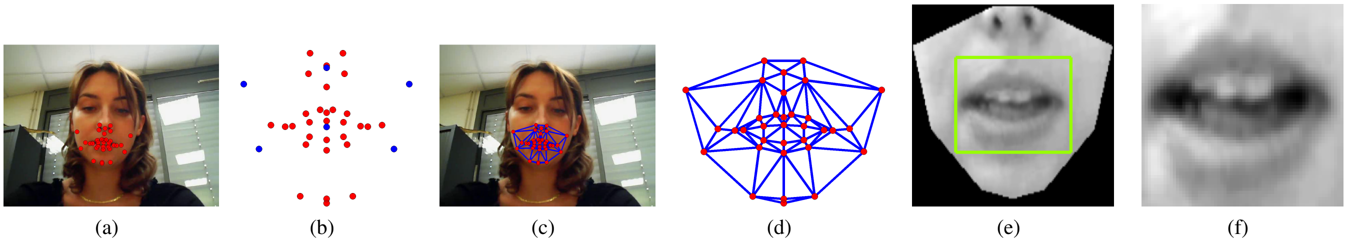

Fig. 1. Instances of the ROI extraction process, illustrated on a non-native speech example frame from the MOBIO Database.
(a) Actually tracked points, (b) Pose-normalised points, including the 6 base points for alignment (shown in blue), (c) Triangulated mesh on the input lower face image, (d) Triangulated mesh of the aligned pose-normalised points, (e) Warped frontal lower face and mouth bounding box, (f) Final rescaled mouth ROI.

works [12]. Appearance descriptors are extracted at each frame from a rectangular region surrounding the speaker's mouth. Dynamic information is encoded through concatenation of static features in a temporal window. Bag-of-features histograms are then used to represent these dynamic features for each sample. Modelling dynamics of visual speech prior to the classification stage allows us to circumvent the need for use of a sequence classifier, such as HMMs, and resort to SVMs for classification. We test the proposed methodology in a challenging setting, that is, on continuous speech examples recorded by mobile devices and uttered by speakers of various nationalities. The results indicate the discriminative power of quantised dynamic appearance features for discrimination between native and non-native speech.

2. DATABASE

The proposed framework is evaluated in a subject-independent fashion on native and non-native speech episodes in English from the MOBIO Database [13]. This bimodal database was recorded at six sites in five countries in two phases, each comprised of six sessions. Each session includes different scenarios, such as short-response questions, pre-defined text, and free-speech questions. In total, 192 audiovisual recordings in English, almost exclusively captured on mobile phones, are available for each of the 150 participants. In all cases, the acquisition device is handheld, which implies high variability in pose and illumination. Additional challenges are posed by the varying appearance of subjects and background, as well as the different recording conditions.

In the current study, we choose to include only the visual speech samples from Phase I in which all subjects read the same pre-defined text, thus establishing a text-dependent experimental scenario. The data used are balanced over the two classes, with 135 samples belonging to 28 native English speakers and 137 to 28 non-native English speakers. The paragraph read in all such recordings is the following:

"I have signed the MOBIO consent form and I understand that my biometric data is 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 behaviour. 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 end of the recordings are removed by applying a voice activity detector on the corresponding audio stream. The mean and standard deviation of duration over all 272 samples used is 22.5 and 3.4 seconds, respectively. The video stream, which is provided in variable frame rate encoding, is converted in a sequence of still frames, corresponding to approximately 15 frames per second, for almost all samples.

3. FEATURES

Experimental findings have shown that appearance-based features, calculated in the area around the speaker's mouth, can efficiently encode valuable information from visual speech [8, 14, 15]. However, their performance largely depends on the accuracy of the *Region Of Interest (ROI)* tracking.

3.1. Region of Interest

Locating and tracking the mouth ROI, as well as registration, i.e., removing variations due to head movements, is an indispensable step that precedes appearance features extraction. This is achieved here in the following steps.

Facial point tracking: We initially track 113 characteristic facial points, using the Appearance-Based Tracker [16]. These are manually annotated in the first frame and tracked for the remaining frames. Out of these points, we only use 34 points that correspond to the lower face region, specifically their 2D spatial coordinates (Fig. 1(a)), along with the coordinates of their pose-free version (Fig. 1(b)), all provided as a part of the tracker's output. Six base points (see blue points in Fig. 1(b))), 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) - and serve to register the face region and calculate the pose-free points. This is done for all 34 pose-free points, which are subsequently used.

Lower Face Texture Warping: Texture warping is performed to acquire lower face images in frontal view. First, for each frame, two 2D meshes (one for actually tracked points and one for aligned pose-normalised points), are triangulated. A piecewise affine warp is defined between the corresponding triangles. This warp is then used to map the texture of the mesh in the input image (Fig. 1(c)), onto the pose-free mesh (Fig. 1(d)). Finally, all warped lower faces are re-sampled to dimension 200×200 (Fig. 1(e)), and the pose-normalised points are accordingly rescaled.

Mouth ROI Extraction: The mouth ROI is extracted from the warped frontal image as a 94×114 bounding box around the center of the mouth (Fig. 1(e)). Finally, all mouth ROIs are downsampled to dimension 64×64 and, subsequently, smoothed by a 3×3 gaussian filter (Fig. 1(f)).

In this work, we choose to apply frame differencing before the computation of appearance features. In particular, after extracting the mouth ROIs, we replace them by the *difference ROIs*, as in [17]. These are computed by subtracting the mouth ROI at the previous frame from the mouth ROI at the current frame. In this way, effects caused by local varying illumination are alleviated, while, at the same time, subtle coarticulation transitions are considered before the feature extraction phase. Henceforth, when we refer to mouth ROIs we mean the *difference ROIs*. Note that samples for which ROIs were erroneously extracted, e.g., due to erratic point tracking or inaccurate warping, were excluded from the experiments.

3.2. Appearance Features

In order to describe appearance within the mouth ROIs, we examine five appearance features: PCA, DCT, DWT [14], LBP [18] and HOG [19], all applied to pixel intensities.

Principal Component Analysis (PCA), when applied to pixel intensity values, models the intensity variation over the training mouth ROIs through an optimal linear transformation, in terms of minimum mean squarred error. Our feature vector corresponds to the principal components accounting for the 95$\%$ of the total variance.

Discrete Cosine Transform (DCT) compresses mouth ROI information by decoupling components in the frequency domain. Similarly to [15], we apply 2D DCT to 8 non-overlapping 32×16 blocks of the ROI. Then, the 2D DCT coefficients that lie in the upper-left corner of each block correspond to the lowest frequencies. After arranging these coefficients in a zig-zag manner, we collect the first 4 for each block and thus construct a 32-dimensional vector.

Discrete Wavelet Transform (DWT) is used to perform frequency decomposition of the mouth images at various levels of resolution. After rescaling the ROIs to dimension 16×16, we employ the Daubechies-4 wavelet filter, with 3 levels of decomposition, to compute the 2D DWT coefficients. As in [15], the approximation coefficients of the third level, along with all the detailed coefficients of the 2nd and 3rd level, are retained in a single 64-dimensional vector.

Table 1. Distribution of Native and Non-Native speakers and corresponding samples, over Training, Validation and Test Set (notation *a/b*: *a* for number of speakers, *b* for number of samples).

	Native	**Non-Native**	**All**
Training	14/71	14/65	28/136
Validation	8/32	7/36	15/68
Test	6/32	7/36	13/68
All	**28/135**	**28/137**	**56/272**

Local Binary Operator (LBP) relies on intensity differences between neighbouring pixels to encode patterns in texture. We use the $LBP^{u2}_{(8,1)}$ operator, which acts in a neighbourhood of 8 pixels on a circle of 1-pixel radius. Its output is the discrete occurence histogram of 59 “uniform” LBP patterns, calculated over the entire ROI [18].

Histograms of Oriented Gradients (HOG) is employed to reveal local edge orientation information in the area inside and around the lips. Parameters are chosen so that the dimensionality is comparable to that of the other features. Each mouth ROI is divided into four 32×32 cells. The gradient at each pixel is discretised 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. Each histogram is normalised four times, with respect to the total energy of the four 2×2 blocks of cells that contain that particular cell. This leads to a vector of length 64 for the whole ROI.

A feature-level normalisation scheme is applied on each speech sample, for all utilised image descriptors. In particular, the mean vector over the whole utterance is subtracted from the feature vector corresponding to each frame. This technique is in accordance with the *feature mean subtraction*, commonly used in the lipreading research [8].

3.3. Bag of dynamic features

Dynamics of visual speech carry valuable information related to accent patterns and modes of articulation. In order to capture dynamic characteristics of speech, we concatenate $\pm L$ adjacent feature vectors around the current frame into a large vector. The value for L is set to 2 in this study, i.e., a 5-frame window with a corresponding duration of 1/3 secs is used.

For the final representation of dynamic features by means of a vector of lower dimensionality, we employ the quantisation technique known as *Bag-Of-Features (BOF)* [20]. We perform k-means clustering on all training dynamic feature vectors, belonging to either the native-speaker class or to the non-native-speaker class, to construct a “vocabulary” of visual “words”. K-means is initialised five times. The cluster centroids yielded by the iteration with the best convergence form the final vocabulary. For every speech segment processed, each dynamic feature is assigned to the closest “word”, in terms of the Euclidean distance. The visual speech example is finally represented by a histogram encoding the frequency of occurence of each “word”.

4. EXPERIMENTAL STUDIES

Evaluation of the proposed method is conducted through a subject-independent experiment. Half of our data is used for Training, while 25% of the speech samples is held out as a Validation Set, that serves for parameter tuning. The remaining 25% is used for testing. All sets are balanced in terms of gender. The distribution of speakers and samples for each class across the three sets is shown in Table 1.

For binary classification, we rely on RBF (Gaussian) Support Vector Machines. The Gaussian kernel width and the soft-margin cost of the SVM decision function are optimised on the Validation Set through a single-resolution gridsearch, based on the mean value of the F1 measure over the two classes. The number of clusters-"words" of the BOF vocabulary (or, equivalently, the dimension of the histograms that are input to the classifier) is tuned separately for each appearance descriptor. The values examined lie in $\{16, 32, 64, 128, 256\}$. The BOF histograms used on the Test Set for each feature are those corresponding to the best-performing vocabulary dimension on the Validation Set.

The diagram in Figure 2 illustrates how mean F1 varies on the Validation Set with increasing values of vocabulary size, for each different appearance feature. As can be seen, the optimal number of clusters for all features, except for LBP, is 16. This finding suggests that broader clusters of the vector space are more suitable to quantise dynamic appearance of mouth configurations. This is intuitive, taking into account that most practical definitions of *visemes*, i.e., the equivalent in the visual domain of audio phonemes, agree on a total number of *visemes* in the range 11-16 [21].

Results on the Test Set are reported in Table 2. Our results show that dynamic appearance features address quite accurately the task of discriminating native from non-native speech. This behaviour is to be expected, as appearance information from the mouth ROI reveals fine movement and tale-telling transient features, such as bulges and wrinkles. In other words, different pronunciation and articulation patterns can be visually identified by the positioning and configuration of the mouth and flesh around it.

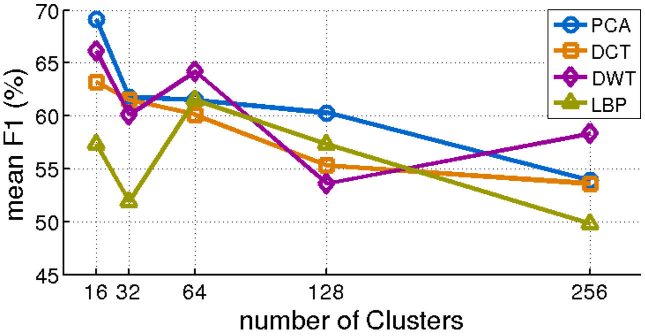

Fig. 2. Performance on the Validation Set, in terms of mean F1, varying with vocabulary size, for each appearance feature.

Table 2. Results on the Test Set, reported as percentages: Class-wise F1 measure (F1-N for Native, F1-NN for Non-Native), mean F1 over the two classes, Unweighted Average Recall (UAR), and classification accuracy (acc.), for each appearance feature.

Features	**#clusters**	**F1-N**	**F1-NN**	**F1$_{\text{mean}}$**	**UAR**	**acc.**
PCA	16	72.1	77.3	74.7	74.7	75.0
DCT	16	78.1	80.6	79.3	79.3	79.4
DWT	16	72.7	74.3	73.5	73.6	73.5
LBP	64	67.7	67.7	67.7	67.9	67.7
HOG	16	60.0	57.6	58.8	59.2	58.8

Image transform-based features, namely DCT, PCA and DWT, stand out as the best-performing, with DCT achieving the highest scores in terms of all measures. Thus, frequency decomposition as well as eigen-decomposition of global texture variation provides discriminative information for the target task. This conforms to results stemming from the lipreading research [22], which similarly show the robustness of the above descriptors in capturing essential information in visual speech. The high accuracy achieved by block-based DCT could be attributed to its ability to encompass the richest local frequency information, that corresponds to the most prominent texture and edge structures in the mouth region. Lower performance is yielded by LBP and HOG, which capture local texture and edge orientation information, respectively. These are less informative as they do not capture buldges and folds representing the speaking patterns.

5. CONCLUSIONS

We have presented a novel approach for visual discrimination between native and non-native speech in English. To the best of our knowledge, this work is the first attempt to address this binary accent classification problem by means of visual features only. We have shown that dynamic appearance descriptors, after undergoing vocabulary-based quantisation, are sufficiently informative to render the target task feasible. Our experiments on a challenging corpus, that includes visual speech samples captured by mobile devices, illustrate the efficiency of the proposed fully-automatic method even in such challenging in-the-wild scenarios.

6. ACKNOWLEDGEMENTS

This work has been funded by the European Community 7th Framework Programme [FP7/2007-2013] under grant agreement no. 611153 (TERESA). The work of S. Petridis is also supported in part by EPSRC grant EP/H016988/1: Pain rehabilitation: E/Motion-based automated coaching.

7. REFERENCES

[1] Levent M Arslan and John HL Hansen, "Language accent classification in American English," *Speech Communication*, vol. 18, no. 4, pp. 353–367, 1996.

[2] Carlos Teixeira, Isabel Trancoso, and António Serralheiro, "Accent identification," in *IEEE ICSLP*, 1996, pp. 1784–1787.

[3] Liu Wai Kat and Pascale Fung, "Fast accent identification and accented speech recognition," in *IEEE ICASSP*, 1999, pp. 221–224.

[4] Pongtep Angkititrakul and John HL Hansen, "Advances in phone-based modeling for automatic accent classification," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 634–646, 2006.

[5] Elizabeth Shriberg, Luciana Ferrer, Sachin Kajarekar, Nicolas Scheffer, Andreas Stolcke, and Murat Akbacak, "Detecting nonnative speech using speaker recognition approaches," in *Proc. IEEE Odyssey Speaker and Language Recognition Workshop*, 2008.

[6] Bozhao Tan, Qi Li, and Robert Foresta, "An automatic non-native speaker recognition system," in *IEEE Int'l. Conf. on Technologies for Homeland Security*, 2010, pp. 77–83.

[7] Mohamed Kamal Omar and Jason Pelecanos, "A novel approach to detecting non-native speakers and their native language," in *IEEE ICASSP*, 2010, pp. 4398–4401.

[8] Gerasimos Potamianos, Hans Peter Graf, and Eric Cosatto, "An image transform approach for HMM based automatic lipreading," in *IEEE ICIP*, 1998, pp. 173–177.

[9] Stéphane Dupont and Juergen Luettin, "Audio-visual speech modeling for continuous speech recognition," *IEEE Trans. on Multimedia*, vol. 2, no. 3, pp. 141–151, 2000.

[10] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proc. of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[11] Rebecca E Ronquest, Susannah V Levi, and David B Pisoni, "Language identification from visual-only speech signals," *Attention, Perception, & Psychophysics*, vol. 72, no. 6, pp. 1601–1613, 2010.

[12] Jacob L Newman and Stephen J Cox, "Language Identification Using Visual Features," *IEEE Trans. on Audio, Speech, and Language Processing,*, vol. 20, no. 7, pp. 1936–1947, 2012.

[13] Christopher McCool, Sebastien Marcel, Abdenour Hadid, Matti Pietikainen, Pavel Matejka, Norman Poh, Josef Kittler, Anthony Larcher, Christophe Levy, Driss Matrouf, et al., "Bi-modal person recognition on a mobile phone: using mobile phone data," in *IEEE ICMEW*, 2012, pp. 635–640.

[14] Gerasimos Potamianos, Ashish Verma, Chalapathy Neti, Giridharan Iyengar, and Sankar Basu, "A cascade image transform for speaker independent automatic speechreading," in *IEEE ICME*, 2000, pp. 1097–1100.

[15] Rowan Seymour, Darryl Stewart, and Ji Ming, "Comparison of image transform-based features for visual speech recognition in clean and corrupted videos," *Image and Video Processing*, 2008.

[16] 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 and Vision Computing*, February 2013.

[17] Patrick Joseph Lucey, *Lipreading across multiple views*, Ph.D. thesis, 2007.

[18] Timo Ojala, Matti Pietikainen, and Topi Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. on PAMI*, vol. 24, no. 7, pp. 971–987, 2002.

[19] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *IEEE CVPR*, 2005, vol. 1, pp. 886–893.

[20] J. Willamowski, D. Arregui, G. Csurka, C.R. Dance, and L. Fan, "Categorizing Nine Visual Classes Using Local Appearance Descriptors," in *ICPR Workshop on Learning for Adaptable Visual Systems*, 2004.

[21] Luca Cappelletta and Naomi Harte, "Viseme definitions comparison for visual-only speech recognition," in *EUSIPCO*, 2011.

[22] Iain Matthews, Gerasimos Potamianos, Chalapathy Neti, and Juergen Luettin, "A Comparison Of Model And Transform-Based Visual Features For Audio-Visual LVCSR.," in *ICME*, 2001.